Technical Physics, Vol. 48, No. 10, 2003, pp. 1221–1225. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 1–5.
Original Russian Text Copyright © 2003 by Sinkevich, Semenov.

THEORETICAL AND MATHEMATICAL PHYSICS

Solution of the Boltzmann Equation by Expanding the Distribution Function with Several Time and Coordinate Scales in the Enskog Series in Knudsen Parameter

O. A. Sinkevich and A. M. Semenov

Moscow Power Institute (Technical University), ul. Krasnokazarmennaya 17, Moscow, 116250 Russia
e-mail: oleg.sinkevich@itf.mpei.ac.ru; SinkevichOA@mpei.ru

Received November 11, 2002; in final form, February 12, 2003

Abstract—A method to solve the Boltzmann equation is analyzed in the case when the distribution function depends on slow and fast time and coordinate scales. Basic relationships for calculating the nonequilibrium multiscale distribution function are shown to differ substantially from those found in the framework of the Chapman–Enskog method: the transfer equations are complemented by the contributions of relaxation processes. The heat and momentum transfer equations derived from the general solution to the Boltzmann equation involve additional terms accounting for relaxation effects. The relaxation effects included in the energy equation result in both a hyperbolic heat conduction equation and a finite rate of heat transfer. In the viscous stress tensor, the Newtonian term of the transfer equation turns out to be supplemented by relaxation terms. © *2003 MAIK “Nauka/Interperiodica”.*

(1) As is known [1], the Chapman–Enskog method for solving the Boltzmann equation is based on the assumption that the distribution function $f(t, \mathbf{r}, \mathbf{c})$ of gas molecules depends on time t and coordinates $\mathbf{r}$ only via local macroscopic parameters of a nonequilibrium gas: the molecule concentration (gas density) $n(t, \mathbf{r})$, temperature $T(t, \mathbf{r})$, and flow velocity $\mathbf{v}(t, \mathbf{r})$. This approach implies that the space–time behavior of the molecule nonequilibrium distribution function is defined only by the spatial and temporal scales of macroscopic processes.

However, the actual nonequilibrium distribution function involves implicitly such characteristic microscopic variables as the collision radius (interaction), the mean free path of molecules, the impact time, and the mean time between collisions. Accomodating to the space–time variations of the macroscopic parameters, the distribution function varies (relaxes) with a rate that depends on the type of molecular interaction and molecular collision frequency. Therefore, one may expect that the dependence of the molecule distribution function $f(t, \mathbf{r}, \mathbf{c})$ on time t and coordinates $\mathbf{r}$ exhibit, generally speaking, more complicated (multiscale) behavior.

Below, we consider an approach to solving the Boltzmann equation based on the above considerations. It is shown that an algorithm for calculating the nonequilibrium distribution function differs noticeably from that used in the Chapman–Enskog method and that the transfer equations are complemented by the contributions of relaxation processes.

(2) In order to solve the Boltzmann equation (in the standard notation, see [1–3])

$$\frac{Df}{Dt} = J\{f, f\}; \tag{1}$$

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \mathbf{c}\cdot\frac{\partial f}{\partial \mathbf{r}} + \frac{\mathbf{F}}{m}\cdot\frac{\partial f}{\partial \mathbf{c}}; \tag{2}$$

$$J\{f, f\} = \int_{(\infty)} d^3u \int_0^{2\pi} d\varphi \times \int_0^{\infty} bdbg[f(t, \mathbf{r}, \mathbf{c}')f(t, \mathbf{r}, \mathbf{u}') - f(t, \mathbf{r}, \mathbf{c})f(t, \mathbf{r}, \mathbf{u})], \tag{3}$$

we will use the method of multiple scales from the excitation theory [4], which is applied to the problems of physical kinetics in [2, 3].

Let the dependence of the distribution function $f(t, \mathbf{r}, \mathbf{c})$ on time t and coordinates $\mathbf{r}$ be defined by two pairs of scales: slow (t_1 and l_1) and fast (t_2 and l_2), with $t_2 \ll t_1$ and $l_2 \ll l_1$. Then,

$$f(t, \mathbf{r}, \mathbf{c}) = f(t/t_1, \mathbf{r}/l_1, t/t_2, \mathbf{r}/l_2, \mathbf{c}/c). \tag{4}$$

Slow scales are usually comparable to the characteristic time of a macroscopic process and to the size of a spatial inhomogeneity in a nonequilibrium system. Fast scales may be associated with the radius and time of molecular collisions, the mean free path and mean time between collisions, or other variables of the same order of magnitude.

The dependence of the distribution function on the molecule velocity **c** is assumed to be one-scale, with the scale depending on the molecular mean thermal velocity **c**. The spatial and temporal scales may be considered, without loss of generality, to be related by the evident relationship

$$l_1/t_1 = l_2/t_2 = \bar{c}.$$

Let us introduce the slow, $\tilde{t}_1 = t/t_1$ and $\tilde{\mathbf{r}}_1 = \mathbf{r}/l_1$, and fast, $\tilde{t}_2 = t/t_2$ and $\tilde{\mathbf{r}}_2 = \mathbf{r}/l_2$, dimensional variables and the molecule dimensionless velocity $\tilde{\mathbf{c}} = \mathbf{c}/\bar{c}$. The indirect differentiation of the distribution function, in view of (4), yields Liouville term (2) of Boltzmann equation (1):

$$\frac{Df}{Dt} = \frac{1}{t_1}\frac{\partial f}{\partial \tilde{t}_1} + \frac{1}{l_1}\mathbf{c}\cdot\frac{\partial f}{\partial \tilde{\mathbf{r}}_1} + \frac{1}{t_2}\frac{\partial f}{\partial \tilde{t}_2} + \frac{1}{l_2}\mathbf{c}\cdot\frac{\partial f}{\partial \tilde{\mathbf{r}}_2} + \frac{1}{c}\frac{\mathbf{F}}{m}\cdot\frac{\partial f}{\partial \tilde{\mathbf{c}}}.$$

In order to write the Boltzmann equation in dimensionless form, we introduce the dimensionless distribution function

$$\tilde{f}(\tilde{t}_1, \tilde{\mathbf{r}}_1, \tilde{t}_2, \tilde{\mathbf{r}}_2, \tilde{\mathbf{c}}) = f(t, \mathbf{r}, \mathbf{c})\bar{c}^3/n,$$

where n is the numerical gas density. Collision integral (3) is reduced to dimensionless form with another linear scale, which is collision (interaction) radius r_c:

$$\tilde{J}\{\tilde{f}, \tilde{f}\} = J\{f, f\}/[\bar{c}^4(n\bar{c}^{-3})^2 r_c^2].$$

In addition, we use the scale F_0 to make the external force dimensionless, $\tilde{\mathbf{F}} = \mathbf{F}/F_0$.

On transformations, the dimensionless Boltzmann equation takes the form

$$\mathrm{Kn}\left[\frac{\partial \tilde{f}}{\partial \tilde{t}_1} + \mathbf{c}\cdot\frac{\partial \tilde{f}}{\partial \tilde{\mathbf{r}}_1} + \delta\tilde{\mathbf{F}}\cdot\frac{\partial \tilde{f}}{\partial \tilde{\mathbf{c}}} + \frac{1}{\varepsilon}\left(\frac{\partial \tilde{f}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}}\cdot\frac{\partial \tilde{f}}{\partial \tilde{\mathbf{r}}_2}\right)\right] = \tilde{J}\{\tilde{f}, \tilde{f}\}. \tag{5}$$

Here,

$$\mathrm{Kn} = \frac{\bar{l}}{l_1}, \tag{6}$$

where

$$\bar{l} = \frac{1}{nr_c^2}$$

is the Knudsen number;

$$\delta = \frac{F_0 l_1}{m\bar{c}^2}$$

is the Froude number; and

$$\varepsilon = t_2/t_1 = l_2/l_1 \tag{7}$$

is the scale ratio.

(3) Further, we will formally consider two kinds of dependences of distribution function (4) on fast variables: (i) the parameter ε given by (7) is not equal to Knudsen number (6) and is independent of it, in which case the scale l_2 can, for example, be equal to the collision (interaction) radius r_c; and (ii) the parameter ε is equal to Knudsen number (6) and the scale l_2 is equal to the molecule mean free path $\bar{l}$.

It is clear that generally one may deal with several parameters of form (7). A more detailed discussion of physical conditions for particular situations to occur is beyond the scope of this paper.

(i) The parameter ε is not equal to Knudsen number (6) and is independent of it. From dimensionless Boltzmann equation (5) we find at Kn = 0

$$\tilde{J}\{\tilde{f}, \tilde{f}\} = 0. \tag{8}$$

A solution to (8) is known to be a locally equilibrium distribution function $f^{[0]}(t, \mathbf{r}, \mathbf{c})$. Hence, one may look for a solution to dimensionless Boltzmann equation (5) in the form of Enskog series

$$\tilde{f} = \tilde{f}^{[0]} + \tilde{f}^{[1]}\mathrm{Kn} + \tilde{f}^{[2]}\mathrm{Kn}^2 + \ldots. \tag{9}$$

Substituting (9) into (5) and equating the coefficients by the same powers of Kn, we arrive at linear integral equations for the coefficients of this series:

$$\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}}\cdot\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{r}}_1} + \delta\tilde{\mathbf{F}}\cdot\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{c}}} + \frac{1}{\varepsilon}\left(\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}}\cdot\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{r}_2}\right) = \tilde{J}\{\tilde{f}^{[0]}, \tilde{f}^{[1]}\} + \tilde{J}\{\tilde{f}^{[1]}, \tilde{f}^{[0]}\}; \tag{10}$$

$$\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}}\cdot\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_1} + \delta\tilde{\mathbf{F}}\cdot\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{c}}} + \frac{1}{\varepsilon}\left(\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}}\cdot\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_2}\right) = \tilde{J}\{\tilde{f}^{[0]}, \tilde{f}^{[2]}\} + \tilde{J}\{\tilde{f}^{[1]}, \tilde{f}^{[1]}\} + \tilde{J}\{\tilde{f}^{[2]}, \tilde{f}^{[0]}\}; \tag{11}$$

etc. As is easily seen, the derivatives in the parentheses by ε^{-1} in (10) vanish. Indeed, the locally equilibrium distribution function $f^{[10]}(t, \mathbf{r}, \mathbf{c})$ depends on time and coordinates only through field parameters, such as the molecular concentration (gas density) $n(t, \mathbf{r})$, temperature $T(t, \mathbf{r})$, and flow velocity $\mathbf{v}(t, \mathbf{r})$. Hence, it depends on the slow variables $\tilde{t}_1$ and $\tilde{\mathbf{r}}_1$ rather than on the fast $\tilde{t}_2$ and $\tilde{\mathbf{r}}_2$. Therefore, Eq. (10) and its solution $f^{[1]}(t, \mathbf{r}, \mathbf{c})$ actually have the same form as the standard first-order solution to the Boltzmann equation obtained by the Chapman–Enskog method. Thus, effects caused by the presence of two scales may appear only in solving Eq. (11) for $f^{[2]}(t, \mathbf{r}, \mathbf{c})$.

In the Bhatnagar–Gross–Krook (BGK) model, the collision integral is given by (see, e.g., [2])

$$J\{f, f\} = -\frac{f - f^{[0]}}{\tau}, \tag{12}$$

where $\tau = l/\bar{c}$ is the mean time between two collisions (Kn = τ/t_1). In this case, instead of (10) and (11), we find

$$\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \mathbf{r}_1} + \delta \tilde{F} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{c}}} = -\tilde{f}^{[1]}, \tag{13}$$

$$\begin{gathered}\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_1} + \delta \tilde{\mathbf{F}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{c}}} \\ + \frac{1}{\varepsilon}\left(\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_2}\right) = -\tilde{f}^{[2]},\end{gathered} \tag{14}$$

etc.

From (14), we obtain

$$\tilde{f}^{[1]} = -\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{t}_1} - \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \mathbf{r}_1} - \delta \tilde{\mathbf{F}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{c}}}. \tag{15}$$

Substituting (15) into (14), we calculate $\tilde{f}^{[2]}$, etc.

Equations (11) and (14) can be solved by expanding into the power series in ε. In the zero order approximation, a solution will coincide with that in the Chapman–Enskog method. However, the study of this issue is beyond the framework of this paper.

(ii) The parameter ε is equal to the Knudsen parameter. At Kn = 0, dimensionless Boltzmann equation (5) yields

$$\frac{\partial \tilde{f}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}}{\partial \tilde{\mathbf{r}}_2} = \tilde{J}\{\tilde{f}, \tilde{f}\}, \tag{16}$$

where

$$\tilde{t}_2 = t/\tau \text{ and } \tilde{\mathbf{r}}_2 = \mathbf{r}/\bar{l}.$$

The locally equilibrium distribution function $f^{[0]}(t, \mathbf{r}, \mathbf{c})$ is a solution to both Eqs. (16) and (8): it reduces both left- and right-hand sides of (16) to zero, since it is independent of the fast variables.

Thus, in the case under consideration, one can look for the distribution function in the form of Enskog series (9). However, in this case, the coefficients of these series are solutions not to integral but to linear integro-differential equations

$$\begin{gathered}\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{r}}_1} + \delta \tilde{\mathbf{F}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{c}}} + \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_2} \\ = \tilde{J}\{\tilde{f}^{[0]}, \tilde{f}^{[1]}\} + \tilde{J}\{\tilde{f}^{[1]}, \tilde{f}^{[0]}\},\end{gathered}$$

$$\begin{gathered}\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_1} + \delta \tilde{\mathbf{F}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{c}}} + \frac{\partial \tilde{f}^{[2]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[2]}}{\partial \tilde{\mathbf{r}}_2} \\ = \tilde{J}\{\tilde{f}^{[0]}, \tilde{f}^{[2]}\} + \tilde{J}\{\tilde{f}^{[1]}, \tilde{f}^{[1]}\} + \tilde{J}\{\tilde{f}^{[2]}, \tilde{f}^{[0]}\},\end{gathered}$$

etc.

In the BGK approximation, we find the first-order partial differential equations for the same functions

$$\frac{\partial \tilde{f}^{[0]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \mathbf{r}_1} + \delta \tilde{\mathbf{F}} \cdot \frac{\partial \tilde{f}^{[0]}}{\partial \tilde{\mathbf{c}}} + \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \mathbf{r}_2} = -\tilde{f}^{[1]};$$

$$\frac{\partial \tilde{f}^{[1]}}{\partial \tilde{t}_1} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{r}}_1} + \delta \tilde{\mathbf{F}} \cdot \frac{\partial \tilde{f}^{[1]}}{\partial \tilde{\mathbf{c}}} + \frac{\partial \tilde{f}^{[2]}}{\partial \tilde{t}_2} + \tilde{\mathbf{c}} \cdot \frac{\partial \tilde{f}^{[2]}}{\partial \tilde{\mathbf{r}}_2} = -\tilde{f}^{[2]}.$$

Passing in these equations back to the dimensional variables in view of the fact that $f^{[1]} = \mathrm{Kn}\,\tilde{f}^{[1]}$ and $f^{[2]} = \mathrm{Kn}^2\tilde{f}^{[2]}$, we have for $f^{[1]}$

$$\frac{\partial f^{[1]}}{\partial t} + \mathbf{c} \cdot \frac{\partial f^{[1]}}{\partial \mathbf{r}} = -\frac{f^{[1]}}{\tau} - \frac{Df^{[0]}}{Dt} \tag{17}$$

and similarly for $f^{[2]}$.

(4) In the general case, the coefficients of Enskog series (3) can be found from the above integro-differential equations, which is an intricate problem.

However, the form of the transfer equations can be found in the BGK approximation without solving Eq. (17).

The expression for the second term on the right of (17) is well known [1]. It is found by calculating the required derivatives of the locally equilibrium distribution function and eliminating the time derivatives of the molecular concentration, temperature, and flow velocity with the help of the local balance equations for mass, energy, and momentum that follow from the solution to the Boltzmann equation in the zero-order (locally equilibrium) approximation of the Chapman–Enskog method (i.e., from the equations of Euler nondissipative dynamics):

$$\begin{gathered}\frac{Df^{[0]}}{Dt} = f^{[0]}\Big[2\Big(\mathbf{UU} - \frac{1}{3}U^2\mathbf{1}\Big) : \frac{\partial \mathbf{v}}{\partial \mathbf{r}} \\ - \Big(\frac{5}{2} - U^2\Big)\mathbf{U}\Big(\frac{2k_\mathrm{B}T}{m}\Big)^{1/2}\frac{\partial \ln T}{\partial \mathbf{r}}\Big].\end{gathered} \tag{18}$$

Here, **U** is the molecule dimensionless thermal velocity [with respect to the local flow velocity $\mathbf{v}(t, \mathbf{r})$] $\mathbf{C} = \mathbf{c} - \mathbf{v}(t, \mathbf{r})$:

$$\mathbf{U} = \mathbf{C}\Big(\frac{m}{2k_\mathrm{B}T}\Big)^{1/2};$$

UU is the tensor of the dyadic product of the vector **U** by itself; **1** is the unit vector; and $\partial\mathbf{v}/\partial\mathbf{r}$ is the strain rate tensor.

In view of (17) and (18), the heat flux density is

$$\mathbf{q} = \frac{m}{2}\int\limits_{(\infty)} \mathbf{C}C^2 f^{[1]} d^2c.$$

Assuming, for the sake of simplicity, that τ is independent of the molecule velocity, we obtain after simple calculations

$$q_\alpha = -\lambda\frac{\partial T}{\partial x_\alpha} - \tau\left(\frac{\partial q_\alpha}{\partial t} + \sum_{\beta=1}^{3}\frac{\partial}{\partial x_\beta}(q_\alpha v_\beta)\right) - \tau\sum_{\beta=1}^{3}\frac{\partial \Lambda_{\alpha\beta}}{\partial x_\beta}$$

or

$$\mathbf{q} = -\lambda\frac{\partial T}{\partial \mathbf{r}} - \tau\left(\frac{\partial \mathbf{q}}{\partial t} + \frac{\partial}{\partial \mathbf{r}}(\mathbf{q}\cdot\mathbf{v})\right) - \tau\frac{\partial}{\partial \mathbf{r}}\Lambda. \qquad (19)$$

Here,

$$\lambda = -\tau\frac{m}{6}\left(\frac{2k_\mathrm{B}T}{m}\right)^2\frac{1}{T}\int\limits_{(\infty)} f^{[0]}\left(\frac{5}{2} - U^2\right)U^4 d^3c$$

$$= \frac{15}{4}\frac{nk_\mathrm{B}^2T}{m}\tau = \frac{3}{2}mk_BTc_p\tau$$

is the thermal conductivity coefficient; c_p is the specific heat at constant pressure; and Λ is the tensor describing the dissipative part of the flux of the dynamic variable $\mathbf{C}C^2$:

$$\Lambda_{\alpha\beta} = \frac{m}{2}\int\limits_{(\infty)} f^{[1]}C_\alpha C_\beta C^2 d^3c.$$

This tensor does not appear in the known relationships of physical kinetics.

The first term on the right of (19) corresponds to the conventional Fourier law of heat conduction; the second one, to the relaxation effect that provides a finite heat transfer rate, which is described by the well-known hyperbolic heat conduction equation (see below); and the third term, to an as yet unstudied relaxation contribution to the heat flux.

A similar relationship can be found for the viscous stress tensor: the Newtonian term of the corresponding transfer equation turns out to be supplemented by additional relaxation terms

$$\Pi_{\alpha\beta} = -2\eta S_{\alpha\beta}$$

$$-\tau\left(\frac{\partial \Pi_{\alpha\beta}}{\partial t} + \sum_{\gamma=1}^{3}\frac{\partial}{\partial x_\gamma}(\Pi_{\alpha\beta}v_\gamma)\right) - \tau\sum_{\gamma=1}^{3}\frac{\partial \Delta_{\alpha\beta\gamma}}{\partial x_\gamma}. \qquad (20)$$

Here,

$$S_{\alpha\beta} = \frac{1}{2}\left(\frac{\partial v_\alpha}{\partial x_\beta} + \frac{\partial v_\beta}{\partial x_\alpha}\right) - \frac{1}{3}\delta_{\alpha\beta}\frac{\partial}{\partial \mathbf{r}}\mathbf{v}$$

is the shear rate tensor,

$$\eta = \frac{2}{15}\tau m\int\limits_{(\infty)} f^{[0]}U^4 d^3c = \tau nk_\mathrm{B}T$$

is the dynamic viscosity, and

$$\Delta_{\alpha\beta\gamma} = \tau m\int\limits_{(\infty)} f^{[1]}C_\alpha C_\beta C_\gamma d^3c$$

is the tensor describing the dissipative part of another flux that so far has not appeared in physical kinetics either.

The first term on the right of (19) corresponds to the standard Newton law of shear viscosity; the second one, to the relaxation effect responsible for a finite heat transfer rate; and the third term, to a relaxation contribution of unknown nature to the heat flux. As far as the authors know, the relaxation effects mentioned above have not been described in the literature.

(5) Let us assume that the gas is at rest and the contribution of the last term in (19) can be neglected:

$$\mathbf{q} + \tau\frac{\partial \mathbf{q}}{\partial t} = -\lambda\frac{\partial T}{\partial \mathbf{r}}. \qquad (21)$$

Below, expression (21) will be referred to as the generalized Fourier law in the relaxation approximation.

If $\mathrm{Kn} \ll 1$ or $\tau \ll t_1$ (slow process), the relaxation term in (21) makes an infinitesimal contribution to heat transfer, and (21) is the usual Fourier law.

Let us now derive a heat conduction equation that takes into account heat relaxation given by (21). We will use a local equation of enthalpy balance that rigorously follows from Boltzmann equation (1) and assume the gas is at rest and the pressure is constant:

$$\rho c_p\frac{\partial T}{\partial t} = -\frac{\partial}{\partial \mathbf{r}}\mathbf{q}, \qquad (22)$$

where $\rho = mn$ is the density and c_p, the specific heat at constant pressure.

Differentiating (22) with respect to time, multiplying the result by τ, and adding to (22) termwise, we arrive at

$$\rho c_p\frac{\partial T}{\partial t} + \tau\frac{\partial}{\partial t}\left(\rho c_p\frac{\partial T}{\partial t}\right) = -\frac{\partial}{\partial \mathbf{r}}\left(\mathbf{q} + \tau\frac{\partial \mathbf{q}}{\partial t}\right).$$

Substituting generalized Fourier law (21) into this relationship, we find the desired heat conduction equation

$$\tau\frac{\partial}{\partial t}\left(\rho c_p\frac{\partial T}{\partial t}\right) + \rho c_p\frac{\partial T}{\partial t} = \frac{\partial}{\partial \mathbf{r}}\left(\lambda\frac{\partial T}{\partial \mathbf{r}}\right). \qquad (23)$$

If both the specific volumetric heat ρc_p and the thermal conductivity λ are independent of temperature, Eq. (23) takes the form

$$\tau\frac{\partial^2 T}{\partial t^2} + \frac{\partial T}{\partial t} = a\Delta T, \qquad (24)$$

where $a = \lambda/\rho c_p$ is the thermal diffusivity.

This equation is hyperbolic and describes the propagation of thermal excitations with a finite velocity

$$c_q = (a/\tau)^{1/2} = \left(\frac{3}{2}\frac{k_B T}{m}\right)^{1/2}.$$

This value is smaller than the speed of sound in a rarefied monatomic gas

$$c_s = \left(\frac{5}{3}\frac{k_B T}{m}\right)^{1/2}.$$

This fact is physically meaningful and agrees with experimental data.

If Kn $\ll 1$, i.e., $\tau \ll t_1$ (slow process), the relaxation term in (24) makes an infinitesimal contribution to heat transfer and (24) turns into the usual hyperbolic heat conduction equation with an infinite heat transfer rate in a heat-conducting medium.

(6) The propagation of temperature excitations in a material has long been the subject of theoretical debates. The hyperbolic heat conduction equation describing the propagation of excitations with a finite velocity has been discussed at length (see, e.g., [5] and Ref. therein). To our knowledge, however, this equation is proposed without any substantiation and the derivation of excitation propagation velocity is lacking.

Although the type of the equation under consideration is defined by the presence of the second time derivative of temperature, the term involving the first time derivative of temperature in (24) also is of importance. The absence of this term means that the excitation, when propagating along the characteristic of the equation, keeps the shape. On the contrary, its presence in (24) strongly distorts the excitation shape during signal propagation because of velocity dispersion.

It should be noted that these effects show up not only in fast processes but also under normal conditions at the front of the wave that describes the propagation of a temperature excitation. While away from the leading edge, the excitation evolution is described well by the parabolic heat conduction equation, the vicinity of the leading edge can be described only with hyperbolic equation (24).

Note that hyperbolic heat conduction equation (24) follows from the parabolic Boltzmann equation, which involves only the first time derivative of the distribution function, and has nothing to do with the so-called generalized Boltzmann equation, which contains the second time derivative of the distribution function [3].

(7) Analysis of nonstationary heat transfer processes with more general equation (19) and momentum with Eq. (20) without the above simplifications is beyond the scope of this paper. We will briefly note the problems for which the effects considered are essential.

First of all, analysis of heat transfer processes in terms of hyperbolic heat conduction equations (23) and (24) is of crucial importance for fast processes, e.g., the irradiation of a material by nano- and femtosecond laser pulses, the explosion of cathodic whiskers, etc. In slower processes, such as heat transfer in composite materials, where the approximations of effective thermal diffusivity a_{eff} and effective relaxation time τ_{eff} are often used (especially in engineering calculations), the effects mentioned above may show up over times $t \approx \tau_{\mathrm{eff}}$ that are much longer than nanosecond intervals.

However, in the classical problem of heat transfer from a temperature excitation localized at the initial instant of time, the finiteness of heat transfer rate is of importance near the leading edge of a wave moving with the velocity of sound. The problem of heat transfer should be solved in this case by the method of multiscale expansions; the general solution consisting of inner and outer expansions will be different in different regions [4]. From Eq. (24) written in the dimensionless form, it follows that, near the leading edge of excitation, one has to solve the hyperbolic equation, which describes finite-velocity signal propagation. Away from the leading edge, Eq. (24) becomes the conventional parabolic heat conduction equation and the temperature variation is described by well-known relationships available from textbooks. In the method of multiscale expansions, the solutions are joined together by means of asymptotic series (as in the problem of boundary layer).

An even more complicated problem is to abandon the BGK approximation in Boltzmann equation (12) and develop a method to solve integro-differential equations for nonequilibrium corrections to a locally equilibrium molecular distribution function in view of different time and coordinate scales.

REFERENCES

1. J. H. Ferziger and H. G. Kaper, *Mathematical Theory of Transport in Gases* (North-Holland, Amsterdam, 1972; Mir, Moscow, 1976).
2. M. N. Kogan, *Dynamics of Rarefied Gas: Kinetic Theory* (Nauka, Moscow, 1967).
3. B. V. Alekseev, Usp. Fiz. Nauk **170**, 649 (2000) [Phys. Usp. **43**, 601 (2000)].
4. J. Colle, *Perturbation Methods in Applied Mathematics* (Blaisdell, Waltham, 1968; Mir, Moscow, 1972).
5. A. V. Lykov, *Heat and Mass Transfer* (Énergiya, Moscow, 1978).

Translated by M. Fofanov

Technical Physics, Vol. 48, No. 10, 2003, pp. 1226–1232. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 6–12.
Original Russian Text Copyright © 2003 by Aleksandrova, Aleksandrov.

THEORETICAL AND MATHEMATICAL PHYSICS

Method of Integral Equations in Solar Magnetohydrodynamics

A. A. Aleksandrova and Yu. N. Aleksandrov
Kharkov Military University, Klochkovskaya ul. 228, Kharkov, 61043 Ukraine
Kharkov National University of Radio Electronics, Kharkov, 61726 Ukraine
e-mail: et@kture.kharkov.ua

Received March 5, 2003

Abstract—The method of integral equations is generalized for the case of solar magnetohydrodynamics. In the new statement, it is not only based on the use of Green’s functions in an MHD medium but also involves the principle of cancellation as applied to the same medium. The application of these equations is demonstrated with a model nonstationary diffraction problem. *© 2003 MAIK “Nauka/Interperiodica”.*

As is known, matter in the Universe is mostly in the fourth state (the terrestrial ionosphere, the solar atmosphere, etc.). As Priest figuratively writes [1], “… we on the Earth represent a tiny enclave of solid, liquid, and gas immersed in the outflow of solar wind plasma, like a pebble in a stream of water.” Like other stars, the Sun is a massive plasma ball which retains its shape and is compressed under the action of its own gravity. The latter half of the 20th century is called “the golden era” of solar physics. In fact, over this period of time, many impressive fundamental discoveries in this field of science were made. In particular, the fine structure and basic importance of the solar magnetic field was found, the heating of the solar corona was shown to be associated with a magnetic field, the decisive role of a magnetic field in various forms of solar activity was proved, etc. This turned out to be possible because the solar plasma was treated as a unified hydrodynamic medium and MHD approaches were invoked to describe processes of interest, which culminated in the creation of a new domain of science, solar magnetohydrodynamics.

A plasma is a medium where oscillations and waves are readily excited. Wave processes are also observed in the Sun. The Sun is an extremely dynamic system: its structural constituents are in continuous motion and the space–time scales of this motion lie over a wide range.

Mathematically, a wave motion in a homogeneous medium is considered following the standard scheme. First, the equilibrium state with density ρ, velocity U, pressure p, and magnetic field B is analyzed. Then, a small perturbation is introduced and the problem as to whether the resulting perturbation propagates as a wave is solved.

In considering waves, one usually uses the continuity equations for mass, momentum, and energy, as well as the equation for induction:

$$\begin{gathered}
\frac{D\rho}{Dt} + \rho\nabla\cdot\mathbf{U} = 0,\\
\rho\frac{DU}{Dt} = -\nabla p + (\nabla\times\mathbf{B})\times\mathbf{B}/\mu - \rho g\mathbf{z}^2 - 2\rho(\Omega\times\mathbf{U}),\\
\frac{D}{Dt}\left(\frac{p}{\rho^{\nu}}\right) = 0,\\
\frac{\partial B}{\partial t} = \nabla\times(\mathbf{U}\times\mathbf{B}),\quad \nabla\cdot\mathbf{B} = 0.
\end{gathered}\tag{1}$$

Here,

$$\frac{D}{Dt} \equiv \frac{\partial}{\partial t} + \mathbf{U}\cdot\nabla$$

is the material derivative related to the motion of the medium and z^0 is the unit vector in the z direction, which is normal to the solar surface. Equations (1) are written in the coordinate system related to the Sun and rotating (relative to the inertial coordinate system) with a constant angular velocity Ω.

Introducing small perturbations of the parameters listed above, $\rho = \rho_0 + \tilde{\rho}$, $\mathbf{U} = \mathbf{U}_0 + \mathbf{u}$, $p = p_0 + \tilde{p}$, and $\mathbf{B} = \mathbf{B}_0 + \mathbf{b}$, and linearizing Eqs. (1) yields a set of linear differential equations for $\tilde{\rho}$, $\mathbf{u}$, $\tilde{p}$, and $\mathbf{b}$:

$$\begin{gathered}
\frac{\partial\tilde{\rho}}{\partial t} + \rho_0\nabla\cdot\mathbf{u} = 0,\\
\rho_0\frac{\partial\mathbf{u}}{\partial t} = \tilde{\nabla}p + (\nabla\times\mathbf{b})\times\mathbf{B}/\mu - \tilde{\rho}g\mathbf{z}^0 - 2\rho_0(\Omega\times\mathbf{u}),\\
\frac{\partial p}{\partial t} + (\mathbf{u}\cdot\nabla)p_0 - V_s^2\left(\frac{\partial\tilde{p}}{\partial t} + (\mathbf{u}\cdot\nabla)\rho_0\right) = 0,
\end{gathered}\tag{2}$$

1063-7842/03/4810-1226$24.00 © 2003 MAIK “Nauka/Interperiodica”

$$\frac{\partial \mathbf{b}}{\partial t} = \nabla\times(\mathbf{U}_0\times\mathbf{b}) + \nabla\times(\mathbf{u}\times\mathbf{B}_0), \quad \nabla\cdot\mathbf{B} = 0.$$

After straightforward transformations, we arrive at a general wave equation for the velocity **u**:

$$\frac{\partial^2\mathbf{u}}{\partial t^2} = V_s^2\nabla(\nabla\cdot\mathbf{u}) - (\gamma-1)g\mathbf{z}^0(\nabla\cdot\mathbf{u}) - g\nabla u_z$$
$$-2\Omega\times\frac{\partial\mathbf{u}}{\partial t} + [\nabla\times(\nabla\times(\mathbf{u}\times\mathbf{B}_0))]/\mu\rho_0. \tag{3}$$

The difficulty encountered in this situation is that there exist two factors, the gravity force and the magnetic force, that contribute to the inhomogeneity of the Sun. The former increases the pressure toward the center of the Sun, while the latter and the associated Lorentz force often produce the plasma pressure (i) in the direction normal to the magnetic field and (ii) away from the regions where the magnetic flux concentrates. This causes inhomogeneities in the Sun; accordingly, boundaries between these inhomogeneities appear (such as magnetic tubes of force, plane-parallel interfaces, etc.). On the other hand, under certain physical conditions, for example, when the solar wind meets the terrestrial magnetic field or two gaseous objects collide with each other, not only the spatial and time derivatives of MHD parameters but also the parameters themselves become highly discontinuous. The jumps of MHD parameters at the surface of discontinuity are found from the integral laws of conservation or integral balance equations, which imply the continuity of energy flux ($\Delta W_n = 0$), mass ($\Delta\rho U_n = 0$), momentum ($\Delta\pi = 0$, as well as tangential ($\Delta\mathbf{E}_t = 0$) and normal ($\Delta B_n = 0$) components of the magnetic field.

In this situation, differential equations of solar magnetohydrodynamics (2) do not hold because of the nondifferentiability of the quantities at the surface of discontinuity. However, they can be represented in integral form, which is totally equivalent to differential equations (2) and the laws of conservation listed above. When linearized, these laws become the edge conditions for the field parameters desired.

It should be noted that the specific nature of waves in the solar plasma, the possibility of their interconversion, and the need for taking into account ways of forming the movable boundary (or, in other terms, the surface of discontinuity of MHD parameters) make the study of MHD wave scattering by various obstacles difficult, especially in view of the nonstationarity of the process. The study of nonstationary processes requires initial conditions to be set; that is, one must know the zero time of nonstationarity. This is particularly true for the solar plasma, since the nonstationarity is its basic property. Therefore, the integral statement of boundary-value problems, which automatically involves initial conditions (and not only edge conditions mentioned above), becomes of special importance.

Statement of the Boundary-Value Problem

Let an inhomogeneity (a geometrically uniform domain) be characterized by the following parameters: $\mathbf{B}_2$, the undisturbed magnetic field; V_{A2} and V_{S2}, the Alfvén velocity and the velocity of sound; and ρ_2, the density. The inhomogeneity has a volume $V(t)$ with a mobile (generally time-dependent) boundary. This object is placed in an infinite MHD medium, which is characterized (in the undisturbed state) by $\mathbf{B}_1$, V_{A1}, V_{S1}, and ρ_1. A disturbing field is defined by the vector of state $\mathbf{\Psi}_0(\mathbf{x}) = \{\mathbf{u}_0(\mathbf{r}, t), \mathbf{b}_0(\mathbf{r}, t), \rho_0(\mathbf{r}, t), p_0(\mathbf{r}, t)\}$. The state vector of the MHD medium, $\mathbf{\Psi}(\mathbf{x}) = \{\mathbf{u}(\mathbf{r}, t), \mathbf{b}(\mathbf{r}, t), \rho(\mathbf{r}, t), p(\mathbf{r}, t)\}$, represents a set of the deviations of the velocity, $\mathbf{u}(\mathbf{r}, t)$; magnetic field, $\mathbf{b}(\mathbf{r}, t)$; density, $\rho(\mathbf{r}, t)$; and pressure, $p(\mathbf{r}, t)$, from their undisturbed values, which describe the MHD state of the solar plasma.

Omitting the derivation of equations involved in the boundary-value problem posed (for details, see [2, 3]), we give the final form of the integro-differential equation for the vector $\mathbf{\Psi}(\mathbf{r}, t)$:

$$\mathbf{\Psi}(\mathbf{x}) = \mathbf{\Psi}_0(\mathbf{x}) + G_0(\mathbf{x}-\mathbf{x}') * \mathbf{W}(\mathbf{x}'). \tag{4}$$

Here, $\mathbf{x} = \{\mathbf{r}, t\}$ is the four-dimensional space–time vector; the symbol * means convolution, that is, the integral operator

$$(f * g)(\mathbf{x}) = \int_{-\infty}^{\infty} dt' \int_{\infty} f(\mathbf{x}-\mathbf{x}')g(\mathbf{x}')d\mathbf{r}';$$

$\mathbf{W}(\mathbf{x})$ is a discontinuous function (written in the class of generalized function) that uniquely describes the MHD medium inside and outside the inhomogeneity and includes all the parameters of the inhomogeneity; $\mathbf{G}_0(\mathbf{x}-\mathbf{x}')$ is the Green's function for the infinite MHD medium (free space), which is defined by the parameters $\mathbf{B}_1$, V_{A1}, V_{S1}, and ρ_1 (or, in other terms, a fundamental solution to wave equation (3) of solar magnetohydrodynamics with the delta-like right-hand side):

$$\frac{\partial^2 G_0}{\partial t^2} - V_{S1}^2\nabla(\nabla\cdot\mathbf{G}_0) + (\gamma-1)\mathbf{g}\mathbf{z}^0(\nabla\cdot\mathbf{G}_0)$$
$$+ g\nabla\mathbf{G}_0 + 2\Omega\times\frac{\partial\mathbf{G}_0}{\partial t} \tag{5}$$
$$-[\nabla\times(\nabla\times(\mathbf{G}_0\times\mathbf{B}_1))]/\mu\rho_1 = \varepsilon\delta(\mathbf{x}-\mathbf{x}');$$

and ε is the affinor.

Hence, it follows that the Green's function, which is a tensor function of the positions of two points (the point of observation, $\mathbf{x} = \{\mathbf{r}, t\}$, and the source point, $\mathbf{x}' = \{\mathbf{r}', t'\}$) has the form

$$\mathbf{G}_0(\mathbf{x}-\mathbf{x}') = \mathbf{G}_0(\mathbf{r}-\mathbf{r}', t-t') = \mathbf{G}\cdot I(\mathbf{r}-\mathbf{r}', t-t'),$$

where $G = \|G_{ij}\|_{i,j=1,2,3}$ is the differential operator that is given in matrix form in a specially selected basis and

$$I(\mathbf{r}-\mathbf{r}', t-t') = \int\limits_{-\infty+i\sigma_0}^{\infty+i\sigma_0} \exp[-iq(t-t')]dq$$
$$\times \iiint\limits_{\infty} \frac{\exp[i\mathbf{p}(\mathbf{r}-\mathbf{r}')]}{\delta(q,\mathbf{p})\Delta(q,\mathbf{p})\iota(q,\mathbf{p})} d\mathbf{p} \qquad (6)$$

is the Fourier–Laplace transform depending on the shift of the space, $\mathbf{r}-\mathbf{r}'$, and time, $t-t'$, variables.

From (6), it is seen that all the integrals in the space–time representation of the Green's function are taken in infinite terms. To do this, it is necessary to get around singular points. These points (usually poles) may occur if $\mathbf{p}$ and q meet the dispersion relations for waves actually existing in the solar plasma:

$$\delta(q,\mathbf{p}) = q^2 - V_{\mathrm{A}1}^2(\mathbf{s}_1\mathbf{p})^2 = 0,$$

$$\Delta(q,\mathbf{p}) = q^4 - q^2(V_{\mathrm{A}1}^2 + V_{\mathrm{S}1}^2)\mathbf{p}^2$$
$$+ V_{\mathrm{S}1}^2 N^2 \sin^2\Theta_g \mathbf{p}^2 + V_{\mathrm{A}1}^2 V_{\mathrm{S}1}^2 (\mathbf{s}_1\mathbf{p})^2\mathbf{p}^2 = 0,$$

$$\iota(q,\mathbf{p}) = q^2 \pm 2(\mathbf{p}\Omega)^2/p = 0.$$

In particular, the dispersion relation $\delta(q, \mathbf{p}) = 0$ is related to Alfvén waves; $\Delta(q, \mathbf{p}) = 0$, to magnetic atmospheric waves; and $\iota(q, \mathbf{p}) = 0$, to inertial waves.

Since convolution in (4) is an integral operator, expression (4) is the integro-differential equation for the state vector $\mathbf{\Psi}$. The very existence of convolution allows one, owing to its properties, to transfer differentiation from the second factors to the Green's function; then, (4) can be written in operator form as

$$\mathbf{\Psi}(\mathbf{x}) = \mathbf{\Xi}(\mathbf{x}) + (\hat{K}\mathbf{\Psi})(\mathbf{x}), \qquad (7)$$

where the operator $\hat{K}$ is the Volterra integral operator.

The above approach is possible because the kernel of the integral operator, when written as the matrix $\hat{K} = \|K_{ij}\|_{i,j=1,2,3}$, is integrable in four-dimensional space. In space–time representation, it is treated as the intersection of the domain of inhomogeneity, which is bounded by the function

$$\chi(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in V(t) \\ 0, & \mathbf{x} \notin V(t), \end{cases}$$

and the characteristic cone with the vertex at the point of observation (in the electrodynamic problem [4], the characteristic cone is a light cone).

Since generalized wave equation (3) of solar magnetohydrodynamics is a differential equation of hyperbolic type, one can construct for it a cone,

$$(t-t_0)^2 - \sum_{i,j=1}^{3} A_{ij}(x_i - x_i^0)(x_j - x_j^0) = 0,$$

at any point in the space. This cone divides the space into two inner and one outer region. From the general theory of differential equations [5], it follows that a second-order partial differential equation that is linear in higher derivatives may be linearly transformed into canonical form at any point in the space; accordingly, the canonical equation of a characteristic cone in the four-dimensional space of the zero integrals (x_1, x_2, x_3, it) with the vertex at a point $(x_1^0, x_2^0, x_3^0, it_0)$ has the form

$$(t-t_0)^2 - \sum_{i,j=1}^{3} a_{ij}(x_i - x_i^0)^2 = 0.$$

Thus, the characteristic surface is a cone with the vertex at the point of observation. However, the "directrix" of this cone is not necessarily a sphere as in electrodynamics [4]. It may have a more intricate shape because of the triple anisotropy of the problem, which is related to the preferential directions of the undisturbed magnetic field, angular velocity, and the force of gravity of the Sun.

Formally, Eq. (7) may be considered as a linearized equation of solar magnetohydrodynamics with nonlocal boundary conditions, which are written in the laboratory coordinate system in the general case. It should be noted that, when the boundary-value problem is solved in the differential statement, the uncertainty arises as to whether one or several wave modes will satisfy local boundary conditions. In the integral statement, this uncertainty is eliminated automatically because of the physics of the phenomenon. Scattered waves appear in the medium when the fundamental mode with a state vector $\mathbf{\Psi}_0$ induces secondary wave sources. The interference of the secondary waves provides desired oscillation modes. This is an analogue of the theorem of cancellation as applied to solar magnetohydrodynamics. Mathematically, the secondary waves are described by the integral terms on the right of (7).

There are two variants of selecting the initial condition. One of them is based on the simplifying assumption that nonstationarity originates at an infinitely distant time instant; that is, adiabatic switching at infinity takes place. In this case, the initial condition to Eq. (7) is not needed. As for Eq. (7), it is defined in the entire range (space) of field definition but its meaning depends on the position of the point of observation. Integration in this case is always limited by the characteristic function $\chi(\mathbf{x})$, which is related to the domain of inhomogeneity. Hence, relationship (7) is a proper integral equation only at points in the domain $V(t)$. Outside this domain, relationship (7) is a quadrature that allows one to calculate the external field from the internal field determined earlier. Hence, an algorithm for the solution of the boundary-value problem of diffraction is split into two steps: (i) finding the internal field by solving the integral equation and (ii) finding the external field from the internal field already found. With this algorithm, boundary-value problems of magnetohydrodynamics were solved in [6].

In the second variant of selecting the initial condition, time is counted from the (finite) time instant $t = 0$. Let us consider the model problem that follows in terms of work [4].

Assume that the inhomogeneity at $t < 0$ is described by a variety of parameters H_1. At the time instant $t = 0$, the state of the inhomogeneity changes and is described by a variety of parameters H_2.

Now we turn back to relationship (7). As before, it depends on the position of the point of observation $\mathbf{x} = (t, \mathbf{r})$. However, this point is now defined in the four-dimensional domain of the space of events. As in the stationary problem (the previous case), (7) is an integral equation if this point belongs to the inhomogeneity. It is known that the solution of an integral equation depends considerably on the equation kernel. As was noted above, the kernel here is defined as the intersection of the characteristic cone with the domain bounded by the function

$$\chi(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in V(t) \\ 0, & \mathbf{x} \notin V(t). \end{cases}$$

Therefore, it is reasonable to consider the projection of this intersection onto the coordinate plane (x, t) in greater detail (Fig. 1).

Note that the solution algorithm for the boundary-value problem remains the same and also consists of two steps. However, the first step (finding the internal field) now implies the solution of a chain of integral equations that are related to each other in an evolutionary way for three time intervals: $t < 0$, $0 \le t \le t_1$, and $t \ge t_1$.

In the first one ($t < 0$), the four-dimensional (space–time) domain of integration in the integral terms of (7) is infinite; that is, prior to the zero time, an incident field $\boldsymbol{\Psi}_0(\mathbf{r}, t)$ generates an internal field $\boldsymbol{\Psi}_1(\mathbf{r}, t)$, where

$$\boldsymbol{\Psi}_1(\mathbf{x}) = \{\mathbf{u}_1(\mathbf{r}, t), \mathbf{b}_1(\mathbf{r}, t), \rho_1(\mathbf{r}, t), p_1(\mathbf{r}, t)\}$$

is a state vector (with an appropriate set of parameters) that describes a solution to the undisturbed problem, which may be viewed as stationary.

Of greatest interest is the interval $t > 0$, where the inhomogeneity is in a new state. In this case, the domain of integration in the same integral terms of (7) is bounded by the hyperplane $t' = 0$. The analysis of intersections of that part of the characteristic cone corresponding to positive t' with the domain bounded by the surface $\chi(\mathbf{x}) = 1$ of the inhomogeneity makes it possible to distinguish two qualitatively different domains of integration Γ_1 and Γ_2 (Fig. 1a). These are the intervals $0 \le t \le d(\mathbf{r})/u$ and $t > d(\mathbf{r})/u$, where $d(\mathbf{r})$ is the least distance between the point of observation $\mathbf{r}$ and the boundary of the volume $V(t)$ at $t = 0$. In the former (Γ_1), the section of the cone by the plane $t' = \text{const}$ $(0 \le t' \le t)$ does not go beyond the surface $\chi(\mathbf{x}) = 1$ and remains entirely within the inhomogeneity-containing four-dimensional domain of the space of events. In this interval, the inhomogeneity boundaries have no effect, so that the "pure" variation of the properties of the medium is taken into account. When extended to the entire four-dimensional space, expression (7) describes the field in an infinite medium with the same properties as the medium inside the object. This is a consequence of the natural time delay because of the wave velocity finiteness and formally appears as if the MHD field interacts with the infinite medium in the domain $V(t)$. Hence, the problem of finding the MHD field is reduced to the initial problem for an infinite medium. Note that we are dealing with the case where the velocity of the boundary does not exceed u; that is, the slope of the world lines of the domain boundary is smaller than that of the characteristic cone generatrices.

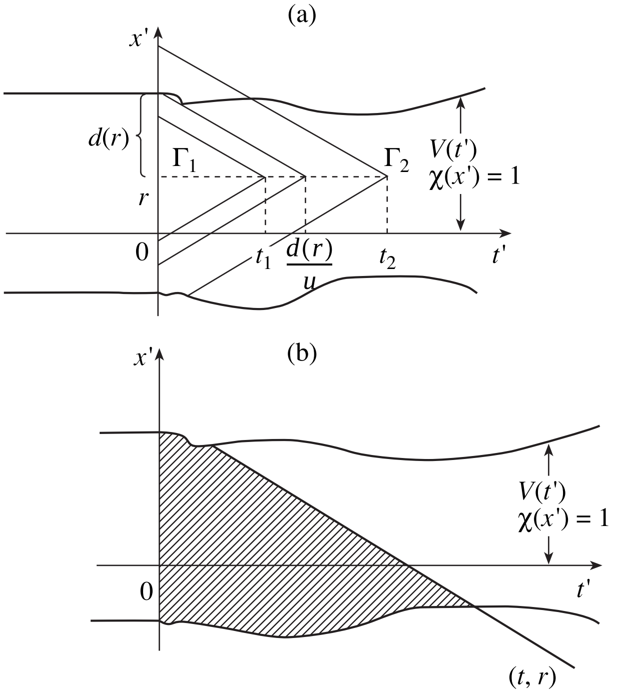


Fig. 1. Intersection of the characteristic cone and domain bounded by the function

$$\chi(x) = \begin{cases} 1, & x \in V(t) \\ 0, & x \notin V(t). \end{cases}$$

(a) Inner and (b) outer problem.

Thus, once the inhomogeneity has changed the state at zero time, its boundary has no effect and the incident field does not contribute directly to the internal field. However, it is "memorized" in the medium; that is, the field $\boldsymbol{\Psi}_2(\mathbf{r}, t)$ sought at the second step is directly generated by the field $\boldsymbol{\Psi}_1(\mathbf{r}, t)$ inside the inhomogeneity before its state has changed and indirectly by the field $\boldsymbol{\Psi}_0(\mathbf{r}, t)$. In other words, a solution of the disturbed problem without including the edge effect is sought at the second step.

Starting from the time instant $t_1 = d(\mathbf{r})/u$ (domain Γ_2), the cone crosses the boundary of the four-dimensional domain $\chi(\mathbf{x}) = 1$ and the effect of the boundary of the domain $V(t)$ cannot be rejected; that is, we come directly to the boundary-value problem for the field $\mathbf{\Psi}_3(\mathbf{r}, t)$. This field is generated by the field $\mathbf{\Psi}_1(\mathbf{r}, t)$, incident field $\mathbf{\Psi}_0(\mathbf{r}, t)$ passed through the boundary, and the field $\mathbf{\Psi}_2(\mathbf{r}, t)$. Thus, at the third step, the field is determined in view of the edge effects and a solution can be constructed by only using solutions found in the previous intervals. It is the last-found solution that contains all information on the variation of the medium inside the domain $V(t)$.

If the point of observation is beyond the domain $V(t)$, expression (7) is a quadrature with which the external field is found from the known internal field (Fig. 1b).

While the problem of wave packet scattering by any interface between two MHD media in the solar plasma is theoretically understood and can be considered in steps, the mathematical description of the phenomenon is a challenge. First of all, wave interconversion at the interface, which is typical of the solar plasma, greatly complicates the problem. The difficulty shows up even in considering the 3D problem of diffraction by the simplest inhomogeneity, namely, half-space. When nonstationarity is adiabatically switched off at infinity, the interconversion is elegantly analyzed in terms of integral equations. In [7], this problem was considered in the pure MHD statement. The integral equations were not separated into those describing particular waves, for example, Alfvén or magnetoacoustic: waves as though "mixed up" in integral terms. Taking into account the finiteness of zero time makes the problem much more difficult. To eliminate wave interconversion completely, it is reasonable to begin with the 1D problem for each wave mode.

The analysis of the transmitted characteristic cone, which, as was noted, has a complex shape, also requires that the 1D problem be considered. It is convenient to take advantage of the cone traces on appropriate coordinate planes. To illustrate the efficiency of our method more vividly, let us solve the model problem of Alfvén wave scattering by the plasma half-space after it has started moving.

Let a plane Alfvén wave

$$\begin{aligned} \mathbf{u}_0(\mathbf{r}, t) &= \mathbf{u}_0 \exp\lfloor ik_0^A z - i\omega_0^A t \rfloor, \\ \mathbf{b}_0(\mathbf{r}, t) &= \mathbf{b}_0 \exp\lfloor ik_0^A z - i\omega_0^A t \rfloor \end{aligned} \tag{8}$$

be normally incident on the plane interface ($z = 0$) between two media specified by the parameters $\mathbf{B}_i$, V_{Ai}, V_{Si}, and ρ_i ($i = 1, 2$). The inner medium ($i = 2$) starts uniformly moving with a velocity U_0 normally to its boundary at the time instant $t = 0$. Prior to the motion, the inhomogeneity occupied the space $z > 0$. Then, the law of motion will have the form $z(t) = U_0 t\Theta(t)$, where $\Theta(t)$ is the Heaviside function. The characteristic cone is defined by the expression

$$t - t' - |z - z'|/V_{\mathrm{A1}} = 0.$$

In such a statement of the diffraction problem, integral equations (7) for the Alfvén wave are scalar. Then, the integral expression for the Alfvén field both passed into the MHD inhomogeneity and reflected from it takes the form

$$\begin{aligned} u_x(\mathbf{r}, t) &= u_{0x}(\mathbf{r}, t) + \frac{1}{B_1}(V_{A1}^2 - V_{A2}^2)\frac{\partial^2}{\partial z \partial t} \\ &\times \int_{-\infty}^{\infty} dt' \int_{V(t')} b_x(\mathbf{r}', t') G^{\mathrm{A}}(\mathbf{r} - \mathbf{r}'', t - t') d\mathbf{r}' \\ &- V_{\mathrm{A1}}^2 \left(1 - \frac{B_2}{B_1}\right)\frac{\partial^2}{\partial t^2}\int_{-\infty}^{\infty} dt' \\ &\times \int_{V(t')} u_x(\mathbf{r}', t') G^{\mathrm{A}}(\mathbf{r} - \mathbf{r}', t - t') d\mathbf{r}', \end{aligned} \tag{9}$$

where

$$\begin{aligned} G^{\mathrm{A}}(\mathbf{r} - \mathbf{r}', t - t') &= -\frac{1}{V_{\mathrm{A1}}}\delta(x - x')\delta(y - y') \\ &\times \Theta\left(t - t' - \frac{|z - z'|}{V_{\mathrm{A1}}}\right) \end{aligned} \tag{10}$$

is the Alfvén Green function, $\delta(x)$ is the Dirac delta function, and $\Theta(t)$ is the Heaviside function.

The space integral in (9) is easily reduced to a one-dimensional integral for the inner problem,

$$\begin{aligned} &\int_{V(t')} u_x(\mathbf{r}', t') G^{\mathrm{A}}(\mathbf{r} - \mathbf{r}', t - t') d\mathbf{r}' \\ &= \int_{z(t')}^{z} u_x(z', t')\Theta\left(t - \frac{z}{V_{\mathrm{A1}}} - t' + \frac{z'}{V_{\mathrm{A1}}}\right) dz' \\ &+ \int_{z}^{\infty} u_x(z', t')\Theta\left(t + \frac{z}{V_{\mathrm{A1}}} - t' - \frac{z'}{V_{\mathrm{A1}}}\right) dz', \end{aligned} \tag{11}$$

and to a one-dimensional integral for the outer problem,

$$\begin{aligned} &\int_{V(t')} u_x(\mathbf{r}', t') G^{\mathrm{A}}(\mathbf{r} - \mathbf{r}', t - t') d\mathbf{r}' \\ &= \int_{z(t')}^{\infty} u_x(z', t')\Theta\left(t + \frac{z}{V_{\mathrm{A1}}} - t' - \frac{z'}{V_{\mathrm{A1}}}\right) dz'. \end{aligned} \tag{12}$$

To form the inner field, it is convenient to use space–time diagrams (Fig. 2). The field in the hatched regions satisfies an equation all terms of which are defined only in this region. Since motion is accomplished from left to right along the time axis, the evolution of the process may be traced by sequentially gaining information on the field in the previous time interval. Here, M is the world point, which bears information on the field at a given time instant and point of the space. Let us consider the inner field for $t > 0$ and $U_0 < V_{A1}$. To this end, we substitute relationship (11) into Eq. (9) to make the integral terms in (9) two-dimensional, that is, depending on the spatial, z', and time, t', coordinates.

Figure 2a shows the domain of integration for the two-dimensional integral in the case when the boundary of the inhomogeneity does not influence the field formation. Solving the initial problem for the unbounded medium, we find that the field in the range $V_{A1}t < z$ does not change either the frequency or the wavenumber. It changes only the amplitude of the Alfvén wave, which existed prior to the motion. In addition, a constant component due to the velocity discontinuity in the medium arises. The edge effects due to the moving boundary do not affect the field, because the disturbance propagates from the boundary with the velocity V_{A1} and its effect is confined in the domain $-V_{A1}t \le z \le V_{A1}t$.

On the domain $U_0t \le z \le V_{A1}t$, one should consider separately the cases of the "outgoing" and "incoming" boundary. The former case, where the boundary moves away from the incident field is shown in Fig. 2b. The latter case, where the boundary meets the incident wave, is shown in Fig. 2c. For either of them, the two-dimensional integral equation is reduced to the Volterra equation of the second kind with the Fredholm kernel. Such an equation is successfully solved by constructing a solution in the form of the Neumann convergent series [8]:

$$u_x(z, t) = F(z, t) + \sum_{n=1}^{\infty} (-\lambda^2 V_{A1})^n \iint_{(D)} K_n(z', t')F(z', t')dt'dz',$$

where $F(z, t)$ is the free term of the equation and the integrated kernels

$$K_1 = 1, \quad K_n = \iint_{(D)} \dots \iint_{(D)} (dD_1 \dots)dD_n, \quad n > 1$$

are polynomials in the variables $\tau^{\pm} = t \pm z/V_{A1}$.

Thus, from the results obtained it follows that the domains where the edge effects become significant arise on both sides of the inhomogeneity as it starts to move. The boundaries of these domains travel with the Alfvén velocity. When the half-space runs away from the incident wave, the field remains continuous at the boundaries of these domains; in the case of the counter motion, the field experiences a discontinuity presumably because of residual phenomena in the media. Similar qualitative effects are observed upon solving the nonstationary boundary-value problem of electromagnetic wave propagation in the electrodynamic statement

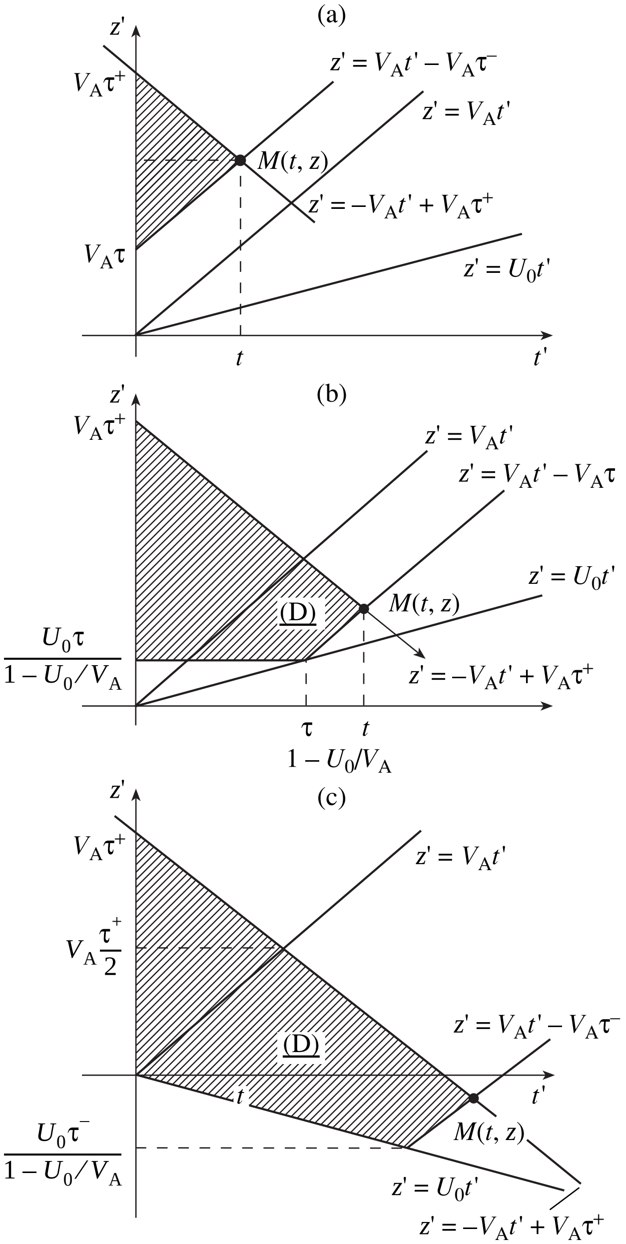


Fig. 2. Geometrical interpretation of space–time domains of integration to find the inner field after the beginning of motion: (a) $V_{A1}t < z$, (b) $U_0t \le z \le V_{A1}t$ ($U_0 > 0$), and (c) $U_0t \le z \le V_{A1}t$ ($U_0 < 0$).

[9]. This supports the validity of the unified physical approach to plasma problems.

The reflected field is easy to recover by quadrupole formulas (9) and (12) provided that the inner field is known. It should be noted that the stray field spectrum depends considerably on the initial condition. In the case of adiabatic switching at infinity, the transmitted field of the stray field spectrum has only the wave moving away from the boundary. When the initial time instant is finite, the field spectrum contains waves with different frequencies, which propagate both from and toward the boundary of the medium. This follows even from the zero-order term of the Neumann series.

The superficial analysis of the boundary-edge problem as applied to the Alfvén wave alone has shown that the method of integral equations is basically applicable to magnetohydrodynamically describing waves in the solar plasma. The value of rigorous mathematical solutions to simple model problems, such as the diffraction of the Alfvén wave, is that they may form the basis for constructing approximate solutions to more difficult boundary-edge problems where the propagation of a number of waves in the solar plasma is considered.

REFERENCES

1. E. R. Priest, *Solar Magnetohydrodynamics* (Reidel, Dordrecht, 1982; Mir, Moscow, 1985).
2. A. A. Aleksandrova and Yu. N. Aleksandrov, Zh. Tekh. Fiz. **67** (5), 6 (1997) [Tech. Phys. **42**, 460 (1997)].
3. A. A. Aleksandrova and Yu. N. Aleksandrov, Zh. Tekh. Fiz. **71** (7), 1 (2001) [Tech. Phys. **46**, 783 (2001)].
4. A. G. Neruch, I. V. Scherbatko, and M. Marciniak, *Electromagnetics of Modulated Media with Applications to Photonics* (Warsaw, 2001).
5. V. S. Vladimirov, *Equations of Mathematical Physics* (Nauka, Moscow, 1967; Marcel Dekker, New York, 1971).
6. A. A. Aleksandrova and N. A. Khizhnyak, *Boundary-Value Problems of Magnetic Hydrodynamics* (Test-Radio, Khar'kov, 1993).
7. A. A. Aleksandrova, Magn. Gidrodin., No. 2, 21 (1993).
8. P. P. Zabreiko, A. I. Koshelev, M. A. Krasnosel'skii, *et al.*, *Integral Equations: A Reference Text* (Nauka, Moscow, 1968; Noordhoff, Leyden, 1975).
9. A. G. Nerukh, Zh. Tekh. Fiz. **54**, 216 (1984) [Sov. Phys. Tech. Phys. **29**, 127 (1984)].

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1233–1237. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 13–17.
Original Russian Text Copyright © 2003 by Olemskoi, Yushchenko.

THEORETICAL AND MATHEMATICAL PHYSICS

Transition of a Loose Medium to the Fluid State: A Phenomenological Theory

A. I. Olemskoi and O. V. Yushchenko
Sumy State University, Sumy, 40007 Ukraine
e-mail: olemskoi@ssu.sumy.ua

Received September 10, 2002; in final form, March 17, 2003

Abstract—In terms of a self-consistent approach, it is shown that taking into account velocity and elastic stress fluctuations allows one to describe the transition of a loose medium to the fluid state in both continuous and stick-slip regimes. In the latter case, elastic stress fluctuations favor the manifestation of self-organized criticality. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

“Quicksand, I cried and leaned on the rifle, but it also started to cave in … Quicksands at the seacoast, where the surf loosens sand and hydrogen sulfide escapes from the bottom with hissing, is a common phenomenon … Once the sea had calmed down, sand consolidated to the point where even footprints became invisible.” This citation [1] gives a descriptive presentation of the dual nature of loose (granulated) media: under the action of swashing waves and a low pressure in an ascending stream of a gas, sand behaves as a viscous fluid; otherwise, its behavior resembles that of a solid. Extensive investigations over the past 10–15 years have shown that this duality is due to the fact that the rms fluctuation u of the particle velocity takes on a hydrodynamic character, causing a macroscopic degree of freedom, the effective (granulated) temperature $T \equiv mu^2$ (m is the mass of a particle), to appear [2–4].

The simplest case is a plane Couette flow, which is realized when the lower boundary moves with a uniform velocity U along the x axis. Here, nonuniformity developing in the perpendicular y direction obeys the following rules [4]: (i) near the moving boundary, the flow is confined within a thin layer, where velocity fluctuations u fall off more slowly than the mean velocity V; (ii) the spatial profile of the velocity $V(y)$ (divided by the maximal velocity U) is independent of V, pressure P, and flow regime (steady or stick-slip); and (iii) the shear stress $\sigma = \mu P$ is constant throughout the flow volume and is characterized by a friction factor $\mu = \mu(U)$, which drops with increasing U to a value typical of dry friction.

These rules could be accounted for in terms of the hydrodynamic theory [3, 4], where the system’s behavior is parametrized by temperature, $T(y, t)$, and velocity, $V(y, t)$, fields. The former field obeys the heat conduction equation

$$\dot{T} = -\varepsilon T + (\chi T')' + \sigma V', \tag{1}$$

which is supplemented by the flow equation $\sigma = \eta V'$. Here, ε is the parameter of dissipation due to inelastic collisions between particles, χ is the thermal diffusivity, η is the dynamic viscosity, and σ is the stress shear component. The dot means the total time derivative $d/dt \equiv \partial/\partial t + V\nabla$, and the primed quantities refer to derivatives with respect to the coordinate y. The velocity field is defined by the equation of motion

$$\dot{V} = \nu V'', \tag{2}$$

where the kinematic, ν, and dynamic, η, viscosity coefficients are related through the density of the medium ρ as $\eta = \nu\rho$.

In the case of a plane flow, $P' = 0$; that is, the pressure is constant throughout the volume. The set of Eqs. (1) and (2) is closed by the equation of state

$$d^3\left(1 - \frac{\rho}{\rho_c}\right) = \frac{T}{P}, \tag{3}$$

where d is the particle size and ρ_c is the critical density (the density at the glass transition point).

Along with ρ, the material constants ε, χ, and ν depend on the state parameters:

$$\varepsilon, \chi \sim \frac{P}{T^{1/2}}; \quad \nu \sim \left(\frac{P}{T}\right)^{\beta} T^{1/2}, \quad \beta \geq 1. \tag{4}$$

The hydrodynamic regime of velocity fluctuations separates two qualitatively different flow conditions: the classical flow with a finite temperature T, a low pressure P, and a low density ρ, and the solidlike state with negligible T, high P, and near-critical density $\rho \leq \rho_c$. In [4], such a difference is defined *a priori* by imparting necessary properties to the medium: in the classical case, β is taken to be equal to unity, while in the second case, it is much greater than unity $\beta \simeq 1.75$).

Thus, the hydrodynamic theory [4] explains the flow of a loose medium, which is observed in experiments;

at the same time, the transition from the solidlike state to the hydrodynamic regime remains uncovered.

In this work, we develop a phenomenological approach where this transition is viewed as the self-organization of an ensemble of particles subjected to an external load. The approach is based on a set of Lorentz equations that provides a simple field representation of a self-organizing medium [5].

SYNERGISTIC PICTURE OF THE TRANSITION TO THE FLIUD STATE

Taking into account the nonequilibrium character of the transition, one may assume that it can be described in terms of the synergistic representation [6], which generalizes the thermodynamic pattern of phase transformations. As follows from the consideration of transport flows [7], the problem is reduced to finding relaxation equations for the ordering parameter field, its conjugate, and a control parameter. When writing these equations, we will proceed from the fact that a description will be self-consistent if each degree of freedom is assigned a microscopic channel of dissipation (the motion of individual particles) and a macroscopic channel due to the cooperative motion (the flow of the medium as a whole).

As an ordering parameter, which separates the fluid state from the solidlike state, it is appropriate to take the rms velocity fluctuation $u \equiv \sqrt{T/m}$ instead of temperature T. Then, omitting nonlinear terms and introducing the hydrodynamic term $\dot{u} \sim \dot{V} = \nu V''$, we find from Eq. (1)

$$\dot{u} = -\frac{u}{t_u} + \chi u'' + a\nu V'', \quad a = \text{const} > 0. \tag{5}$$

The first term here stands for a microscopic channel of dissipation due to inelastic collisions, the rate of which (defined by the parameter ε in (1)) is inversely proportional to the relaxation time t_u. The remaining terms describe a macroscopic channel of dissipation due to the spatial variation of the velocity fluctuation u (thermal contribution) and velocity mean value V (hydrodynamic contribution). In terms of our self-consistent approach, hydrodynamic equation (2) should be supplemented by a linear term for a microscopic channel of dissipation due to the effect of shear stresses on the velocity fluctuation:

$$\dot{V} = -gu\sigma + \nu V'', \tag{6}$$

where g is a positive constant. Equation (6) describes the relaxation of elastic stresses σ to a value σ_e related with an external action:

$$\dot{\sigma} = \frac{\sigma_e - \sigma}{t_\sigma} - g_\sigma \nu u V''. \tag{7}$$

As above, the first term refers to a microscopic channel of dissipation due to the local particle redistribution with a characteristic time t_σ, while the second term includes the cooperative contribution due to the fluctuation redistribution of particles moving with an acceleration $\dot{V} \sim \nu V''$.

Synergistic equations (5)–(7) are conveniently analyzed in dimensionless variables that are obtained by reducing the time t, coordinate y, velocity fluctuation amplitude u, mean velocity V, and elastic stress σ to the following scales[1]:

$$t_u, \quad l \equiv \sqrt{\chi t_u}, \quad u_c \equiv (g g_\sigma t_\sigma)^{-1/2}, \quad V_c \equiv a^{-1}\frac{\chi}{\nu}u_c, \\ \sigma_c \equiv (a g t_u)^{-1}, \tag{8}$$

where the hydrodynamic coordinate and time are given by

$$\chi^2 \equiv \nu\tau = \frac{\eta^2}{\rho G}, \quad \tau \equiv \frac{\eta}{G} \tag{9}$$

(G is the characteristic value of the shear modulus).

Eventually, the behavior of a loose medium is described by the set of dimensionless equations

$$\dot{u} = -u + u'' + V'', \tag{10}$$

$$(\tau/t_u)(l/\lambda)^2\dot{V} = -u\sigma + V'', \tag{11}$$

$$(t_\sigma/t_u)\dot{\sigma} = (\sigma_e - \sigma) - uV''. \tag{12}$$

The solution of this set depends on the relationships between the times t_u, τ, and t_σ, as well as between the scales l and λ. The phase transition is dissipative when the inelastic character of collision is so weak that the relaxation time t_u far exceeds the hydrodynamic scale $(l/\lambda)^2\tau$ and microscopic time t_σ [6, 8]:

$$(l/\lambda)^2\tau \ll t_u, \quad t_\sigma \ll t_u. \tag{13}$$

Since the dimensionless velocities $\dot{u}$ and $\dot{V}$ and the rate $\dot{\sigma}$ are of the same order of magnitude, inequalities (13) make it possible to ignore the left-hand sides of (11) and (12). Then, from (11) and (12), we obtain

$$V'' = \frac{\sigma_e u}{1 + u^2}, \quad \sigma = \frac{\sigma_e}{1 + u^2}. \tag{14}$$

Thus, the spontaneous growth of the velocity fluctuations in the interval bounded above by $u = 1$ increases the curvature V'' of the mean velocity profile and causes the internal stresses σ to relax below the level σ_e ($\sigma < \sigma_e$), which is defined by external conditions. The substitution of the first equality in (14) into (10) yields the

---

[1] Under normal conditions, the kinematic viscosity ν does not exceed the thermal diffusivity χ and the range of the mean velocity V_c exceeds the amplitude u_c of its fluctuation [4].

Ginzburg–Landau–Khalatnikov equation

$$\dot{u} = u'' - \frac{\partial E}{\partial u}, \tag{15}$$

whose form is specified by the fluctuation energy

$$E = \frac{u^2}{2} - \frac{\sigma_e}{2}\ln(1+u^2), \tag{16}$$

measured in units of u_c^2. When the stress σ_e is small, the dependence $E(u)$ is a monotonically increasing function with a minimum at $u = 0$, which correlates with the solidlike state. As σ_e grows and exceeds the critical level σ_c, a minimum at

$$u_0 = \sqrt{\sigma_e - 1} \tag{17}$$

appears, which corresponds to the ordered state. In this case, the curvature of the velocity profile takes a finite value $V_0'' = u_0$ and internal stresses relax to the critical value $\sigma_c = 1$.

In the steady state ($\dot{u} = 0$), the first integral in Eq. (15) has the form

$$\begin{gathered} \frac{1}{2}(u')^2 = E + |E_0|, \\ E_0 \equiv E(u_0) \simeq -\frac{1}{4}(\sigma_e - 1)^2 < 0. \end{gathered} \tag{18}$$

Here, it is taken into account that the fulfillment of the conditions $u = u_0$ and $u' = 0$ in the ordered phase ($y = -\infty$) requires that the constant of integration equal the absolute value of the ordering energy E_0. This energy was estimated by expanding into a series and retaining terms quadratic in $\sigma_e - 1 \ll 1$.

Leaving terms of orders u^2 and u^4 in Eq. (18), we find the steady-state distribution of fluctuations in the form of a kink:

$$u = u_0 \tanh\left(\frac{y_0 - y}{\xi}\right), \quad \xi^2 \equiv \frac{2}{\sigma_e - 1}. \tag{19}$$

Here, ξ is the correlation length, which diverges at the critical value of the elastic stress. The constant of integration $y_0 \gg \xi$ defines the width of the transition domain where velocity fluctuations go to zero from the steady-state value given by (17). Substituting distribution (19), which falls within the correlation length ξ, into the second formula in (14), we see that the shear component in the transition region grows monotonically from the critical value $\sigma_c = 1$ to a value $\sigma_e > \sigma_c$ that depends on external conditions (Fig. 1a). Such a stress distribution leads to a critical increase in the friction factor $\mu \equiv \sigma/P$ from the hydrodynamic value $1/P \equiv (gat_uP)^{-1}$, which drops as the inelastic collision rate increases and the stress dependence of the fluctuations strengthens (see (6)), to the dry-friction value σ_e/P. Note that such behavior by no means contradicts the conclusion [4] that elastic stresses in a flow cell are constant. In fact, this conclusion is based on the macroscopic approximation in terms of which the interface is assumed to be indefinitely thin, while in our case, the elastic field varies over the correlation length $\xi \neq 0$.

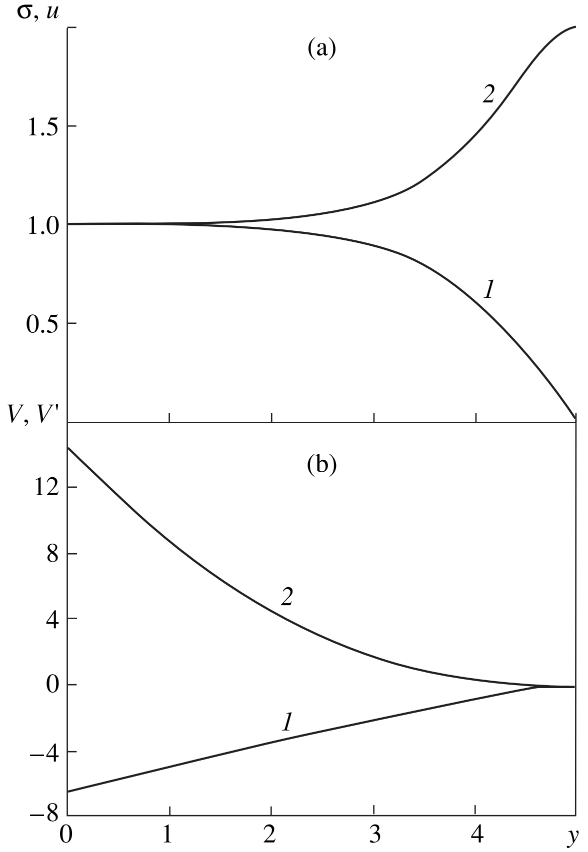


Fig. 1. (a) Spatial variation of the (*1*) rms velocity fluctuation u and (*2*) elastic stress σ and (b) the profiles of the (*1*) velocity gradient V' and (*2*) mean velocity V. $y_0 = 5$, $\sigma_e = 2$.

To avoid misunderstanding, we note that the formal use of equality (18) at the flow boundary $y = y_0$, where $u = 0$, results in a finite velocity fluctuation gradient:

$$u_0' = \frac{u_0}{\xi} \equiv \frac{1}{\sqrt{2}}(\sigma_e - 1). \tag{20}$$

Physically, the presence of such a gradient does not mean the appearance of a heat flux $J \propto T'$, since the temperature gradient $T' \equiv 2muu'$ at the flow boundary vanishes together with the velocity fluctuations ($u = 0$).

The first equality in (14) combined with an expression for $u'(y)$ that follows from (18) yields an equation for the flow mean velocity:

$$\sqrt{2}V' = -\ln\left[\frac{1+u^2}{(\sigma_e - 1) - u^2}\right]. \tag{21}$$

According to Fig. 1b, the gradient of this velocity increases monotonically from

$$V'(0) \simeq -\sqrt{2}\frac{y_0}{\xi} + \frac{1}{\sqrt{2}}\ln\frac{4(\sigma_e - 1)}{\sigma_e} \qquad (22)$$

to zero in the range $0 < y < y_0$. As a result, the velocity profile $V(y)$ splits into sections of fast ($0 < y < y_0 - \xi$) and slow ($y_0 - \xi < y < y_0$) fall. Since the fluctuations $u(y)$ vary only in the range $y_0 - \xi < y < y_0$, the mean velocity in the basic flow region ($0 < y < y_0 - \xi$) falls off faster than its fluctuations [4].

The above consideration demonstrates that the set of Lorentz equations (5)–(7) gives a self-consistent description of the transition of a loose medium from the solidlike to fluid state.

STICK-SLIP FLOW REGIME

Since Lorentz equations (5)–(7) are of hydrodynamic nature, the elastic stresses are represented by the value σ averaged over a physically small volume. Fluctuations that appear over distances on the order of the particle size must be taken into consideration by introducing a stochastic source $(\sqrt{I}/t_\sigma)\eta(t)$ into Eq. (7). This source is characterized by an intensity I and white noise $\eta(t)$: $\langle\eta(t)\rangle = 0$, $\langle\eta(t)\eta(t')\rangle = \delta(t - t')$. Then, in adiabatic approximation (13), the curvature $V''(t)$ of the velocity profile and the stress $\sigma(t)$ include stochastic additions:

$$V''(t) = V'' + \tilde{V}''\eta(t), \quad \sigma(t) = \sigma + \tilde{\sigma}\eta(t), \qquad (23)$$

$$\tilde{V}'' \equiv \frac{g}{\nu}\frac{\sqrt{I}u}{1 + u^2/u_c^2}, \quad \tilde{\sigma} \equiv \frac{\sqrt{I}}{1 + u^2/u_c^2}, \qquad (24)$$

where the first terms in (23) have the form of (14).

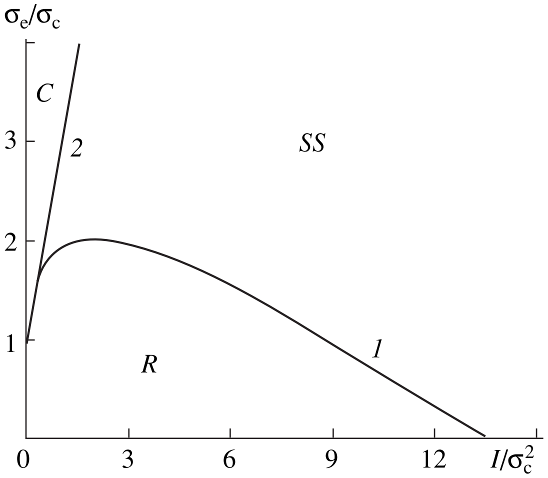


Fig. 2. Phase diagram of a loose medium subjected to elastic stresses with a mean value σ_e and a fluctuation intensity I. (*1*, *2*) Boundaries between the domains of continuous flow (*C*), rest (*R*), and stick-slip (*SS*) regime.

In view of (5), these equalities yield the space-dependent Langevin equation

$$t_u\dot{u} = l^2u'' + f(u) + \sqrt{I(u)}\eta(t), \qquad (25)$$

$$f(u) \equiv -u + \frac{\sigma_e}{\sigma_c}\frac{u}{1 + u^2/u_c^2}, \quad I(u) \equiv [a\nu t_u\tilde{V}''(u)]^2. \qquad (26)$$

The steady-state distribution of homogeneous solutions to (25) [8],

$$\mathcal{P}(u) \propto I^{-1}(u)\exp\left\{\int\frac{f(u)}{I(u)}du\right\}, \qquad (27)$$

has a maximum at the point given by the condition

$$x^3 - \frac{\sigma_e}{\sigma_c}x^2 - 2\frac{I}{\sigma_c^2}(x - 2) = 0, \quad x \equiv 1 + \frac{u^2}{u_c^2}. \qquad (28)$$

According to the phase diagram (Fig. 2), with σ_e exceeding the limit $\sigma_{c2} \equiv (1 + 2I)\sigma_c$, the most probable fluctuations meet the condition $u \neq 0$ and the flow regime is similar to that described above. As the stress decreases below σ_{c2}, $\sigma_e < \sigma_{c2}$, a maximum appears at the point $u = 0$, which corresponds to the solidlike state. Under the curve $\sigma_{c1} \equiv \sigma_e(I)$, which is given by

$$\frac{I^2}{\sigma_c^4} - \frac{I}{\sigma_c^2}\left[\frac{27}{2}\left(1 - \frac{1}{3}\frac{\sigma_e}{\sigma_c}\right) - \frac{1}{8}\left(\frac{\sigma_e}{\sigma_c}\right)^2\right] + \frac{1}{2}\left(\frac{\sigma_e}{\sigma_c}\right)^3 = 0, \qquad (29)$$

this maximum is single.

Thus, taking into account elastic stress fluctuations results in the appearance of the two-phase state at $\sigma_{c1}(I) < \sigma_e < \sigma_{c2}(I)$. In this range, the system may break into the stick-slip regime, randomly passing from the solidlike to fluid state. This regime may occur even in the absence of external stresses if the fluctuation intensity exceeds the value of $I_c = (27/2)\sigma_c^2$. Under such conditions, the system behaves as in the case of self-organized criticality [9].

Since the system is of stochastic character, the description of the stick-slip regime is reduced to finding the distribution $\mathcal{P}(\tau)$ of the durations τ of flow intervals alternating with rest intervals. This distribution is governed by the competition between negative feedback, which diminishes the elastic stress energy ζ, and a positive contribution increasing the entropy (complexity) s of the system. The duration τ plays the role of the ordering parameter, the entropy s represents the conjugate field, and the energy ζ serves as a control parameter. Eventually, the self-consistent behavior of the system is described by the generalization of Lorentz equations (5)–

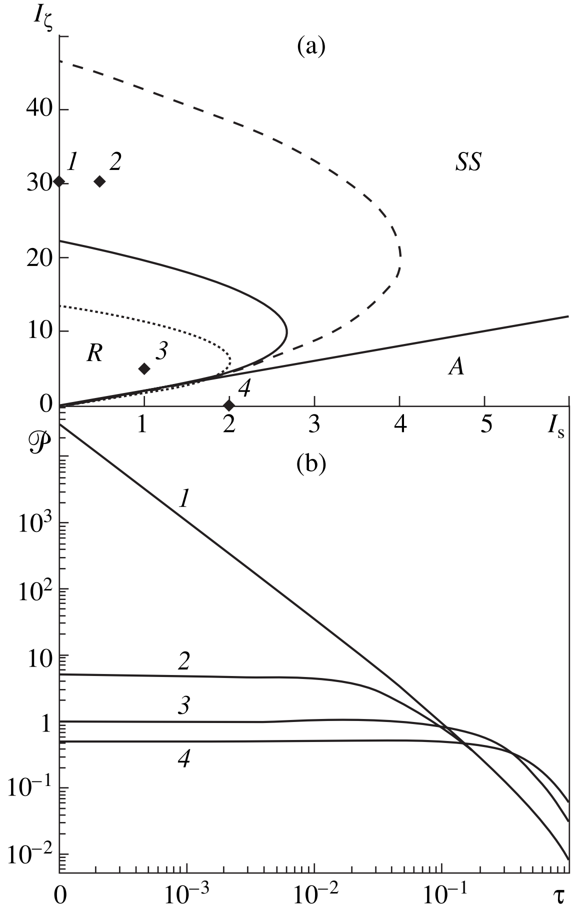


Fig. 3. (a) Phase diagram of the system according to the fluctuation intensities I_ζ and I_s, elastic stress energy, and entropy ($I_\tau = 0$, $\zeta_e = 0$). The dotted, continuous, and dashed lines correspond to the exponent a = 0.5, 0.75, and 1.0, respectively. Rhombi *1–4* correspond to the related curves in Fig. 3b. (b) Distributions of the flow interval durations for a = 0.75 and fluctuation intensities shown in Fig. 3a.

(7) for the stochastic case:

$$\begin{aligned} t_\tau \dot{\tau} &= -\tau + A_\tau s + \sqrt{I_\tau}\eta(t), \\ t_s \dot{s} &= -s + A_s \tau^a \zeta + \sqrt{I_s}\eta(t), \\ t_\zeta \dot{\zeta} &= (\zeta_e - \zeta) - A_\zeta \tau^a s + \sqrt{I_\zeta}\eta(t). \end{aligned} \tag{30}$$

Here, t_τ, t_s, and t_ζ are the relaxation times of the basic variables; A_τ, A_s, and A_ζ are coupling constants; I_τ, I_s, and I_ζ are the fluctuation intensities of the associated parameters; and ζ_e is the external action parameter. The distinctive feature of set (30) is its fractal feedback, the amount of which depends on the exponent a, $0 < a \le 1$.

In the adiabatic approximation $t_s, t_\zeta \ll t_\tau$, this set is reduced to spatially homogeneous Langevin equation (25), where the effective force and the intensity of the stochastic source are given by

$$f \equiv -\tau + \zeta_e \tau^a d_a(\tau), \quad I \equiv I_\tau + (I_s + I_\zeta \tau^{2a}) d_a^2(\tau). \tag{31}$$

Here, $d_a(\tau) \equiv (1 + \tau^{2a})^{-1}$ and the following scales are used:

$$\tau^{sc} \equiv (A_s A_\zeta)^{-\frac{1}{2a}}, \quad s^{sc} \equiv A_\tau^{-1}(A_s A_\zeta)^{-\frac{1}{2a}},$$

$$\zeta^{sc} \equiv A_\tau^{-1} A_s^{-\frac{1}{2}\left(1 + \frac{1}{a}\right)} A_\zeta^{\frac{1}{2}\left(1 - \frac{1}{a}\right)}, \quad I_\tau^{sc} \equiv (A_s A_\zeta)^{-\frac{1}{a}},$$

$$I_s^{sc} \equiv A_\tau^{-2}(A_s A_\zeta)^{-\frac{1}{a}}, \quad I_\zeta^{sc} \equiv A_\tau^{-2} A_s^{-\left(1 + \frac{1}{a}\right)} A_\zeta^{\left(1 - \frac{1}{a}\right)}.$$

The steady-state distribution $\mathcal{P}(\tau)$ of the flow interval durations is defined by expression (27), from which it follows that the behavior of the system under the self-organized criticality conditions can be found from the phase diagram shown in Fig. 3a. It is seen that a decrease in the feedback exponent a expands considerably the stick-slip flow range. The curves $\mathcal{P}(\tau)$ in Fig. 3b show that the power behavior, which is characteristic of the self-organized criticality conditions, takes place in the limit $I_\tau, I_s \ll I_\zeta$, where the distribution of the flow interval durations has the asymptotic $\mathcal{P} \sim \tau^{-2a}$ with $\tau \longrightarrow 0$.

Thus, our investigation demonstrates that the stick-slip flow regime of a loose medium under the self-organized criticality conditions is due to elastic stress fluctuations

ACKNOWLEDGMENTS

A.I.O. thanks the Ukrainian Research Center for their financial support (project no. 1976) and the Institute of Complex Systems (Dresden, Germany) for their hospitality.

REFERENCES

1. V. K. Arsen’ev, *Dersu Uzala* (Pravda, Moscow, 1983).
2. J. Rajchenbach, Adv. Phys. **49**, 229 (2000).
3. W. Losert, L. Bocquet, T. C. Lubensky, *et al.*, Phys. Rev. Lett. **85**, 1428 (2000).
4. L. Bocquet, W. Losert, D. Schalk, *et al.*, Phys. Rev. E **65**, 011307 (2002).
5. A. I. Olemskoi, Physica A **310**, 223 (2002).
6. A. I. Olemskoi and A. V. Khomenko, Zh. Éksp. Teor. Fiz. **110**, 2144 (1996) [JETP **83**, 1180 (1996)].
7. A. I. Olemskoi and A. V. Khomenko, Phys. Rev. E **63**, 036116 (2001).
8. A. I. Olemskoi, Usp. Fiz. Nauk **168**, 287 (1998) [Phys. Usp. **41**, 269 (1998)].
9. P. Bak, *How Nature Works: The Science of Self-Organized Criticality* (Oxford Univ. Press, Oxford, 1997).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1238–1242. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 18–22.
Original Russian Text Copyright © 2003 by Balashov, Bibikov.

ATOMS, SPECTRA, RADIATION

Inverse Problem in the Stopping Theory of Fast Multiply Charged Ions

V. V. Balashov and A. V. Bibikov
Skobeltsyn Research Institute of Nuclear Physics, Moscow State University,
Vorob’evy gory, Moscow, 119992 Russia
e-mail: balvse@anna19.npi.msu.su

Received January 10, 2003

Abstract—The method of coupled kinetic equations, which was put forward earlier to describe the passage of fast multiply charged ions through a material, is applied for the extraction of partial (resolved with respect to the charge state of an ion) stopping parameters from the experimental spectra of the ion energy loss at the exit from the target. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

The progressively expanding use of fast multiply charged ion beams in various areas of science (atomic and nucleus, physics, solid-state physics, materials science, biology, medicine, etc.) poses the problem of reliable determination of the stopping parameters for multiply charged ions passing through a material. Unlike protons or other elementary charged particles, multiply charged ions exchange electrons with it, when passing through the environment, thus stochastically changing their charge and, hence, stopping parameters. Taking into consideration the charge exchange effect makes the description of the ion energy spectra at the target exit difficult. This is especially true in the case of an ion beam passing through thin foils, when statistical equilibrium between different charge fractions of the beam has no time to establish.

Currently, the conventional method of analyzing the kinetics of multiply charged ion beams propagating in a material, including the determination of the ion stopping parameters, is Monte Carlo statistical simulation (see [1, 2] and Refs. therein). This approach is known to be tedious and machine-time-consuming. Another difficulty associated with this method is that the statistical simulation of the kinetics of such processes requires that a set of input parameters related to the effective cross section of ion–target interaction be predetermined. These are the rates of unit charge exchange events and the parameters of effective ion stopping that are related to each charge state of the ion. The theory of ion–atom collisions, while rapidly progressing on both the quantum and the classical basis, still cannot provide reliable data of this sort. The impression has formed (see, e. g., [3]) that the study of the process of multiply charged ion stopping (complex as it is) is just the way of finding basic stopping parameters. Here, we come to the concept of inverse problem: how to derive the parameters required from experimental data for ion beam passage and how to design experiments that fit the statement of the inverse problem to the greatest extent.

These questions were tackled in the fundamental works of Landau [4] and Winterbon [5]. In a great number of cases, the quantitative description of multiply charged ion stopping in a medium is possible without knowing details of ion–atom interaction. It suffices to know only several first moments of the energy loss distribution in an elementary act of inelastic ion–atom collision (in the diffusion approximation, these quantities are the effective stopping parameter $S_i(E)$ and the energy straggling parameter $\Omega_i^2(E)$ for each of the ion charge states considered). Thus, the inverse problem is reduced to the standard procedure of fitting to experimental data by varying a finite (though large) number of theoretical parameters. All one needs in this situation is an adequate stopping theory that would correctly describe the kinetics of ion beam passage through a material in the conditions of intense charge exchange.

We propose to fill this gap with the previously presented [6] method of coupled kinetic equations for partial (related to definite charge states) ion energy loss distributions. Based on this technique, we analyzed the evolution of such distributions in the case of nonequilibrium beam propagation [7], determined the effective charges of fast multiply charged ions in the target material [8], included charge exchange when considering the angular spread of an ion beam [9], and investigated the effect of the intermediate excitations of fast multiply charged ions on their stopping characteristics [10]. It was demonstrated that our approach offers advantages over the Monte Carlo method.

Turning back to the inverse problem, one has to state that experimental data on the partial energy distributions and other energy parameters of ion nonequilibrium passage are almost lacking; moreover, even those available from the literature usually do not carry information on direct measurements and error analysis, which is imperative for these results to be used in the inverse problem. Therefore, we will use pseudoexperimental (theoretically obtained) data that simulate the results of real experiments on multiply charged ion passage through a material.

THEORETICAL BACKGROUNDS

We proceed from the general Landau–Winterbon equations [4, 5] for the charge-state-resolved functions $F_i(\Delta, x)$ ($i = 1, \ldots, N$), which define the distribution of the energy loss $\Delta = E_0 - E$ for an ion at a distance x from the entrance to the target:

$$\frac{\partial F_i(\Delta, x)}{\partial x} = -\lambda_i^{\text{tot}} F_i(\Delta, x) + \int \lambda_{ii}(T) F_i(\Delta - T, x) dT + \int \sum_{j \neq i} \lambda_{ji}(T) F_j(\Delta - T, x) dT. \tag{1}$$

Here, λ_{ii} is the differential rate of collision energy loss for an ion with a charge i, $\lambda_{i \to j}(T) \equiv \lambda_{ij}(T)$ is the rate of charge exchange between two states, and $\lambda_i^{\text{tot}} = \int \sum_{j \neq i} \lambda_{ij}(T) dT$ is the total rate of transitions from a state i to other charge states. Input energy parameters of calculation are the effective stopping parameters $S_i = \int \lambda_{ii}(T) T dT$ and the straggling parameters $\Omega_i^2 = \int \lambda_{ii}(T) T^2 dT$ for an ion "frozen" in each of states i. They are the coefficients in the Taylor expansion of the distributions $F_i(\Delta - T, x)$ in the vicinity of Δ. At a certain distance x from the entrance to the target, the characteristic energy loss per collision, which is given by the distribution $\lambda_{ii}(T)$, becomes small compared with the mean energy loss $\langle \Delta \rangle$ over this distance (diffusion approximation) and all the moments $\int \lambda_{ii}(T) T^n dT$ with $n > 2$, starting from the skewness parameter $N_i = \int \lambda_{ii}(T) T^3 dT$, may be omitted from the expansion. In addition, if the energy loss per act of charge exchange is assumed to be constant, $\lambda_{ij}(T) = \lambda_{ij} \delta(T - \bar{\Delta}_{ij})(i \neq j)$, the system of equations (1) transforms into a simpler one, which in the integral form appears as [6]

$$F_i(\Delta, x) = e^{-\lambda_i x} \int g_i(\Delta - \Delta', x) F_i(\Delta', x = 0) d\Delta' + \sum_{j \neq i} \lambda_{ji} \int_0^x dx' \int e^{-\lambda_i (x - x')} g_i(\Delta - \bar{\Delta}_{ji} - \Delta', x - x') F_j(\Delta', x') d\Delta'. \tag{2}$$

This expression includes the Gaussian propagators

$$g_i(\Delta - \Delta', x - x') = \frac{\exp\{-[\Delta - \Delta' - S_i(x - x')]^2 / 2\Omega_i^2(x - x')\}}{\sqrt{2\pi\Omega_i^2(x - x')}}, \tag{3}$$

which are characteristic of the diffusion approximation and describe the stopping and energy spread of an ion beam should it stay in definite frozen-charge states $i = 1, \ldots, N$.

Kinetic equations (2) are solved numerically with the use of appropriate initial conditions at the entrance to the target $F_i(\Delta, x = 0)$. Solving (2), we obtain the energy loss distribution functions $F_i(\Delta, x)(i = 1, \ldots, N)$; the respective weights of the charge fractions of ions in the beam, $f_i(x) = \int F_i(\Delta, x) d\Delta$; the corresponding mean energy loss values $\langle \Delta \rangle_i(x) = \int F_i(\Delta, x) \Delta d\Delta / f_i(x)$; the partial straggling parameters; and the stopping power of the material for each of the ion charge states detected in the beam,

$$\left(-\frac{dE}{dx}\right)_i (x) = \frac{d}{dx}(\langle \Delta \rangle_i(x)). \tag{4}$$

Knowing these parameters, we determine the mean statistic charge of the ion beam,

$$q^{(\text{stat})}(x) = \sum_{i=1}^{N} q_i f_i(x), \tag{5}$$

and the so-called equivalent charge of the beam, which is expressed via its energy characteristics:

$$q^{(\text{equiv})}(x) \equiv \sqrt{\left\langle -\frac{dE}{dx} \right\rangle (x) \Big/ \left(-\frac{dE}{dx}\right)_{\text{p}}}, \tag{6}$$

where

$$\left\langle -\frac{dE}{dx} \right\rangle (x) = \sum_i \left(-\frac{dE}{dx}\right)_i (x) f_i(x)$$

is the stopping power averaged over the charge fractions in the beam and $(-(dE/dx))_{\text{p}}$ is the effective stopping power for protons.

For very thin films, the stopping power of a material should be calculated as the ratio of finite differences (see, e.g., [1])

$$\left(-\frac{\Delta E}{\Delta x}\right)_i (x) = \langle \Delta \rangle_i(x)/x$$

rather than by formula (4).

Because of this, we complement (6) by the modified equivalent charge [8], which is denoted by $Q^{(\text{equiv})}(x)$, where $\langle -(\Delta E/\Delta x) \rangle (x)$ replaces $\langle -(dE/dx) \rangle (x)$ in (6).

Table 1. Transition rates λ_{ij}, $[\mu g/cm^2]^{-1}$

$\lambda_{i \to j}$	O^{4+}	O^{5+}	O^{6+}	O^{7+}	O^{8+}
O^{4+}	0.	2.5389	0.	0.	0.
O^{5+}	0.0853	0.	0.8002	0.	0.
O^{6+}	0.	0.1338	0.	0.1281	0.
O^{7+}	0.	0.	0.1019	0.	0.0567
O^{8+}	0.	0.	0.	0.2354	0.

STOPPING AND CHARGE EXCHANGE FOR OXYGEN IONS O^{q+} PASSING THROUGH A THIN CARBON FILM

We consider the process that was studied experimentally and simulated with the Monte Carlo method by Datz *et al.* [1] for incident beam energies between 10 and 24 MeV. Let us take the energy 24 MeV and, following [1], construct system (2) based on the following set of charge states:

$$O^{q+} = O^{4+}, O^{5+}, O^{6+}, O^{7+}, O^{8+}. \qquad (7)$$

For the sake of simplicity, here we disregard intermediate excited states of ions in the stopping process (this point is considered in [10]). Using the charge exchange rate matrix λ_{ij} [1] (Table 1), we calculate the charge fractions distribution (7) as a function of the target thickness. Figure 1 shows that the equilibrium distribution, which is independent of the initial charge state of the beam, is achieved in the interval (20–25) $\mu g/cm^2$; therefore, a film thickness of 7.5 $\mu g/cm^2$ taken in [1] is adequate to investigate the process under nonequilibrium conditions. Note also that, at this thickness, the fraction of beryllium-like ions O^{4+} turns out to be negligible not only when the beam consists of bare (totally free of electrons) O^{8+} nuclei but also when O^{4+} ions themselves are the input charge state.

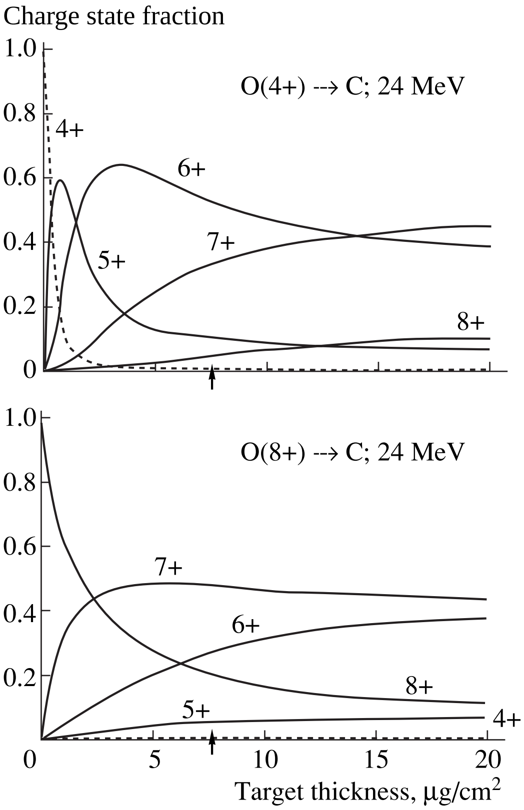


Fig. 1. Fractions of charge states O^{4+}, O^{5+}, O^{6+}, O^{7+}, and O^{8+} vs. the target thickness.

When choosing the stopping parameters for ions in frozen charge states, we will use the effective ion charges $q_i^{\text{stop}}(E)$ and $q_i^{\text{strag}}(E)$ calculated within the framework of the dielectric stopping theory, which is commonly used when the partial screening of an ion nucleus by the electron shell should be taken into consideration [11–14]. The corresponding values of the stopping power $S_i(E)$ are very close to those used in [1]. They are listed in Table 2 (second column), along with the straggling parameters $\Omega_i^2(E)$, which are taken according to Bohr [14].

(1) Examples of setting and solving the inverse problem. With the given set of parameters, we will calculate the energy loss spectrum for oxygen ions passing

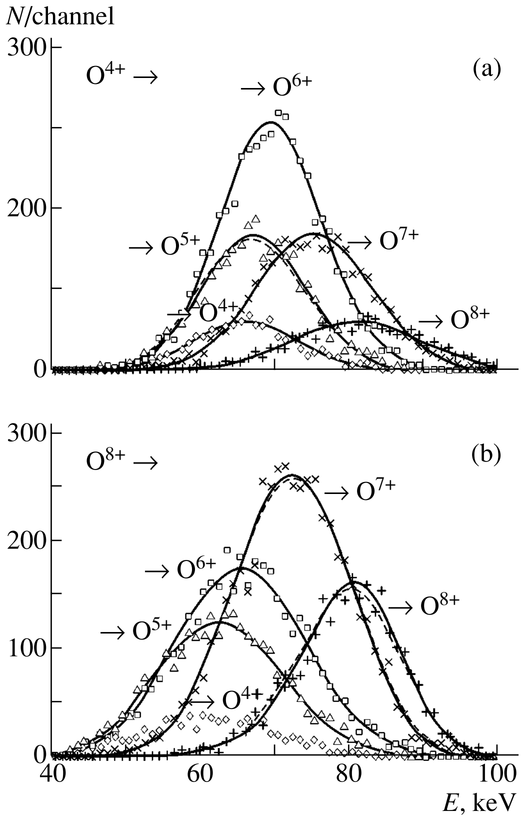


Fig. 2. Results of simulation by variant 1 ($O^{4+} \longrightarrow O^{q+}$ and $O^{8+} \longrightarrow O^{q+}$). Symbols, Monte Carlo simulation; the dashed and solid curves are the partial spectra calculated with the input parameters and the parameters derived from the analysis of the spectra, respectively.

through a 7.5-μg/cm^2-thick film in two variants. In either of them, the computation is performed twice: first, by the method of coupled equations (the dashed curves in Figs. 2, 3) to compare the partial energy loss spectra for different charge fractions and then by the Monte Carlo technique (with the same basic parameters) to simulate experimental results. The spectrum of ions that left the film in charge states (7) is used to plot histograms with a 1-keV/channel energy step in the range from 0 to 100 keV. The statistical spread of the resulting data gives an estimate of the data spread in the corresponding experiment.

Variant 1. Let an experiment give the complete set of the partial energy loss spectra $F_i(\Delta)$ for all the charge fractions $i = O^{4+}, \ldots, O^{8+}$ at the exit from the target with the states O^{4+} and O^{8+} chosen as the input ones (Figs. 2a, 2b). In order to detect a statistically reasonable number of events in each of the charge states at the exit from the target (Table 3), the total number of particles in the entering beam was substantially different for the input states O^{4+} (~10^6 particles) and O^{8+} (~10^4 particles).

These data are analyzed by the χ^2 minimization method with the system of equations (2) being solved at each step of the fitting procedure. The obtained values of the stopping parameters, along with the associated errors, are presented in Table 4 (third column), and the corresponding calculated spectra are shown by the solid curves in Figs. 2a and 2b.

Variant 2. Now we simulate another experiment that gives a set of the partial spectra $F_i(\Delta)$ for ions with the charge state at the entrance to $O^{4+}, \ldots, O^{7+}$ or O^{8+} and the exit from the target (Table 2; Fig. 3). The basic stopping parameters derived from such an "experiment" are listed in Table 4 (fourth column), and the corresponding calculated spectra are shown by the solid curves in Fig. 3.

(2) Discussion. Let us draw the main conclusions that follow from the two variants of simulation considered above.

(i) The discrepancies between the input and output stopping parameters, as well as the calculation errors, are the greatest for the charge state 4+, which is the lowest component in the charge fraction distribution when the effective target thickness is 7.5 μg/cm^2 (Fig. 1).

(ii) For the most intense fractions 6+ and 7+, both variants of simulation yield satisfactory results for the stopping parameters S_i and Ω_i^2: their values are close to the input ones, and the statistical errors are the least.

(iii) In both variants, the theoretical partial spectra plotted with the parameters extracted from the experiment (the solid curves in Figs. 2, 3) deviate from the curves constructed with the initial parameters (dashed curves) only slightly.

(iv) In view of the above observation, it is of interest to value the results in each of the variants considered: in variant 1, the final result is based on a sample of more than one million ion trajectories; in variant 2, the sample equals 550000. The difference is not large; so, in our opinion, neither of the two variants is preferred.

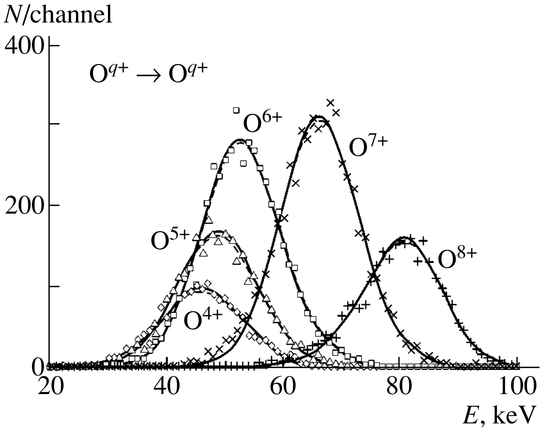


Fig. 3. The same as in Fig. 2 for variant 2 ($O^{q+} \longrightarrow O^{q+}$).

Our experience in calculating the partial spectra from a given set of basic parameters shows that the method of coupled kinetic equations may be conveniently applied for solving the inverse problem and is less time-consuming than the Monte Carlo method by several orders of magnitude. For example, with the use of an Athlon XP 1800+-based PC, the entire procedure for determining the optimal values of S_i and Ω_i^2 by the

Table 2. Number of trials (particles at the entrance to the target) and the number of detected events $O^{q+} \longrightarrow O^{q+}$ for simulation by variant 2

	Number of trials	Number of events detected
$O^{4+} \longrightarrow O^{4+}$	500 000	1702
$O^{5+} \longrightarrow O^{5+}$	30 000	2910
$O^{6+} \longrightarrow O^{6+}$	10 000	4797
$O^{7+} \longrightarrow O^{7+}$	10 000	5341
$O^{8+} \longrightarrow O^{8+}$	10 000	2693

Table 3. Number of detected events $O^{4+} \longrightarrow O^{q+}$ and $O^{8+} \longrightarrow O^{q+}$ for simulation by variant 1

Charge state at the exit	Input state O^{4+}	Input state O^{8+}
O^{4+}	1019	819
O^{5+}	2863	2666
O^{6+}	5297	3803
O^{7+}	3263	5139
O^{8+}	1303	2693

Table 4. Stopping parameters

Parameter	Used for simulation	Derived from final spectra	
		$\{O^{4+}, O^{8+}\} \longrightarrow O^{q+}$	$O^{q+} \longrightarrow O^{q+}$
$S(q = 4+)$, keV/(μg/cm^2)	4.02	4.450 ± 0.14	3.718 ± 0.23
$S(q = 5+)$, keV/(μg/cm^2)	5.52	5.603 ± 0.03	5.674 ± 0.04
$S(q = 6+)$, keV/(μg/cm^2)	7.10	7.060 ± 0.02	7.062 ± 0.02
$S(q = 7+)$, keV/(μg/cm^2)	8.83	8.790 ± 0.02	8.833 ± 0.02
$S(q = 8+)$, keV/(μg/cm^2)	10.9	10.928 ± 0.02	10.930 ± 0.02
$\Omega^2(q = 4+)$, keV^2/(μg/cm^2)	5.0	10.46 ± 1.5	3.56 ± 2.0
$\Omega^2(q = 5+)$, keV^2/(μg/cm^2)	5.0	3.35 ± 0.3	5.16 ± 0.4
$\Omega^2(q = 6+)$, keV^2/(μg/cm^2)	5.0	5.05 ± 0.2	4.88 ± 0.2
$\Omega^2(q = 7+)$, keV^2/(μg/cm^2)	5.0	4.71 ± 0.2	4.80 ± 0.2
$\Omega^2(q = 8+)$, keV^2/(μg/cm^2)	5.0	4.45 ± 0.2	4.66 ± 0.2

method of χ^2 minimization (Table 4) takes 20 min and 1 h in variants 1 and 2, respectively.

CONCLUSIONS

An advance in the experimental study of interaction between fast multiply charged ions and a material and, in particular, systematic measurements of the partial (charge-resolved) spectra of the energy loss under nonequilibrium stopping conditions is retarded by the lack of a theoretical basis that would provide a means to derive the basic stopping parameters from the results of such measurements. We took a step forward in this direction using the method of coupled kinetic equations, which has previously proved to be efficient for calculating the kinetics of fast multiply charged ions passing through a material. An algorithm was evolved, and a computer program was worked out, which demonstrates the potentialities for conveniently solving the inverse problem in the stopping theory under the conditions typical of recent experiments in this area.

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 00-02-17207).

REFERENCES

1. B. Rosner, S. Datz, W. Wu, *et al.*, Phys. Rev. A **57**, 2737 (1998).
2. A. Blazevic, H. G. Bohlen, and W. von Oertzen, Nucl. Instrum. Methods Phys. Res. B **190**, 64 (2002).
3. P. Sigmund, Nucl. Instrum. Methods Phys. Res. B **69**, 113 (1992).
4. L. Landau, J. Phys. (Moscow) **8**, 201 (1944).
5. K. B. Winterbon, Nucl. Instrum. Methods **144**, 311 (1977).
6. V. V. Balashov, A. V. Bibikov, and I. V. Bodrenko, Zh. Éksp. Teor. Fiz. **111**, 2226 (1997) [JETP **84**, 1215 (1997)].
7. V. V. Balashov and I. V. Bodrenko, Phys. Lett. A **232**, 231 (1997).
8. V. V. Balashov, A. V. Bibikov, and I. V. Bodrenko, Vestn. Mosk. Univ., Ser. 3: Fiz., Astron., No. 2, 28 (2002).
9. V. V. Balashov and I. V. Bodrenko, Phys. Scr. **T80**, 254 (1999).
10. V. V. Balashov, Nucl. Instrum. Methods (in press).
11. W. Brandt and M. Kitagawa, Phys. Rev. B **25**, 5631 (1982).
12. T. Kaneko, Phys. Rev. A **33**, 1602 (1986); Phys. Rev. A **41**, 4889 (1990).
13. L. L. Balashova and N. M. Kabachnik, Izv. Ross. Akad. Nauk, Ser. Fiz. **62**, 763 (1998).
14. N. Bohr, K. Dan. Vidensk. Selsk. Mat. Fys. Medd. **18**, 8 (1948).

Translated by A. Sidorova

Technical Physics, Vol. 48, No. 10, 2003, pp. 1243–1248. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 23–28.
Original Russian Text Copyright © 2003 by Gaibov, Grigor'ev.

GASES AND LIQUIDS

Acoustic Radiation from a Drop Due to Its Nonlinear Vibrations

A. R. Gaibov and A. I. Grigor'ev
Yaroslavl State University, Sovetskaya ul. 14, Yaroslavl, 150000 Russia
e-mail: shir@uniyar.ac.ru

Received February 20, 2003

Abstract—It is shown that the intensity of acoustic radiation from a vibrating drop depends mainly on the monopole and dipole components appearing only in the second order of smallness in vibration amplitude. The intensity of the quadrupole acoustic radiation generated by the vibration fundamental mode in the first order of smallness in amplitude turns out to be much weaker. This is associated with the fact that, if the acoustic wavelength is much larger than the drop characteristic size, their ratio becomes a governing small parameter, being lesser than the ratio of the drop vibration amplitude to the drop linear size. Analytical estimates of the amplitudes of monopole, dipole, and quadrupole components of the velocity field associated with the acoustic field of the drop. *© 2003 MAIK "Nauka/Interperiodica".*

(1) Interaction between a vibrating charged drop and an acoustic field is of interest for many scientific and applied problems, such as the physics of aerosols, the physics of thunderstorm electricity, and the electroacoustic levitation of drops in experiments aimed at producing superpure materials (see [1–8] and Ref. therein).

Nevertheless, a number of points concerning charged drop–acoustic field interaction remain poorly understood. For example, in most applications dealing with interaction between acoustic waves and liquid drop systems, drops are simulated as rigid sound-scattering objects without internal degrees of freedom. At the same time, it is known [3] that the frequency spectrum of the capillary vibrations of drops with sizes that are typical of rains and mists overlaps with the range of audible frequencies. This means that such drops may not only interact with an external acoustic field but also generate acoustic waves.

The investigation of acoustic radiation from a vibrating charged drop in approximations linear and nonlinear in vibration amplitude was carried out in [9–11].

(2) In [9], the analysis was based on the model of capillary vibrations of a charged drop immersed in a compressible medium and was performed in the approximation linear in amplitude of the drop initial deformation. It was found that acoustic waves generate all modes of drop surface vibrations starting from the fundamental one (that is, from $n = 2$) and also that the quadrupole component due to the fundamental mode ($n = 2$) has the highest intensity in the acoustic radiation spectrum.

The expression for the intensity of quadrupole acoustic radiation from a linearly vibrating charged drop has the form [9]

$$I \approx \frac{4\pi\rho_2 V C_2^2 R^2 \omega_2^2}{\left|k_2 R h_1^{(2)}(k_2R) - 3h_2^{(2)}(k_2R)\right|^2}; \quad k_2 \equiv \frac{\omega_2}{V}; \tag{1}$$

$$\omega_2^2 = \frac{4\sigma}{R^3}\left(1 - \frac{W}{4}\right)\left[\frac{\rho_1}{2} - \frac{\rho_2 h_2^{(2)}(k_2R)}{k_2 R h_1^{(2)}(k_2R) - 3h_2^{(2)}(k_2R)}\right]^{-1}; \tag{2}$$

$$W = \frac{Q^2}{4\pi\varepsilon\sigma R^3}.$$

Here, ρ_1 and ρ_2 are the densities of the drop and environment, V is the speed of sound in the environment, C_2 is the amplitude of the fundamental mode, ω_2 is the vibration frequency, R is the drop radius, $h_n^{(2)}(z)$ are spherical Hankel functions of the second kind, σ is the surface tension, Q is the drop charge, and ε is the permittivity of the environment.

The intensity of quadrupole acoustic radiation from the vibrating drop vs. drop radius and the amplitude of the initial deformation fundamental mode according to (1) is plotted in Fig. 1.

(3) In [10], the acoustic radiation of a charged incompressible liquid drop nonlinearly vibrating in a compressible environment was considered. The study was carried out in the approximation quadratic in amplitude of the initial deformation of the equilibrium spherical drop shape with the initially excited fundamental vibration mode. The acoustic radiation spectrum was found to have the monopole component, which arises from the time dependence of the ampli-

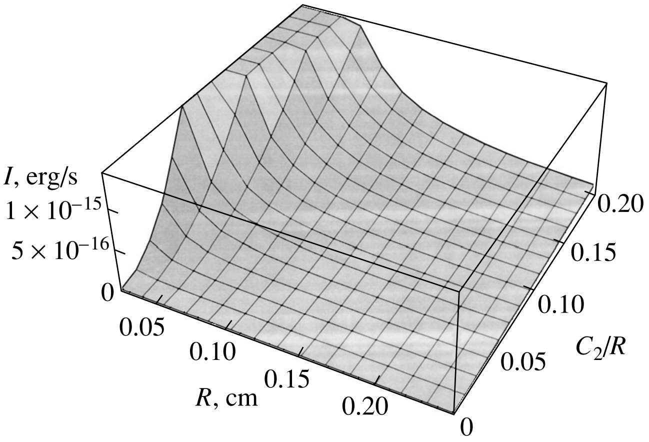


Fig. 1. Intensity of the quadrupole acoustic radiation from a water drop in air as a function of the drop radius and amplitude of the initial deformation of the fundamental mode at $W = 1$.

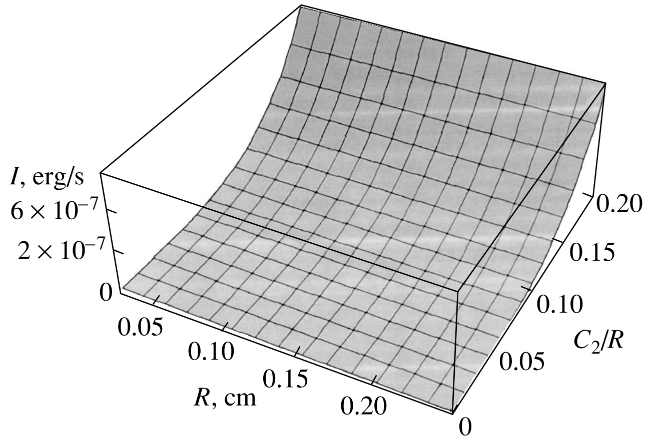


Fig. 2. Intensity of the monopole acoustic radiation from a nonlinearly vibrating water drop in air as a function of the drop radius and amplitude of the initial deformation of the fundamental mode at $W = 1$

tude of the drop vibration zeroth mode. This dependence is a consequence of the drop volume constancy. The time dependence of the zeroth mode is defined by the expression [10]

$$a_0^{(2)}(t) = -\frac{1}{10R}\delta_2^2\exp(-2\gamma t)[1+\cos(\Theta)] + \frac{1}{10R}[-2C_2^2 + \delta_2^2\cos(2\beta_2) + \delta_2^2],$$

where

$$\Theta \equiv (2\omega_2 t + 2\beta_2); \quad \delta_2 = C_2\sqrt{1+(\gamma/\omega_2)^2};$$

$$\beta_2 = \arctan\left(-\frac{\gamma}{\omega_2}\right);$$

C_2 is the initial disturbance amplitude; and ω_2 and γ are the vibration frequency and damping decrement of the fundamental mode of the drop vibrations, which are defined as the real and imaginary parts of the frequency given by expression (2).

Note that the vibration frequency of the zeroth mode is equal to the doubled frequency of the mode governing the initial deformation.

The expression for the intensity of monopole acoustic radiation from a nonlinearly vibrating charged drop that is related to the zeroth mode has the form

$$J = \frac{2\pi\rho_2 R^4\omega^4 a_0^2}{V(1+\omega^2R^2/V^2)}, \tag{3}$$

where a_0 is the zeroth mode amplitude and ω is the frequency of the same mode.

The intensity of monopole acoustic radiation from a nonlinearly vibrating drop vs. its radius and the amplitude of the fundamental-mode-induced initial deformation (expression (3)) is shown in Fig. 2.

(4) In [11], the acoustic radiation of a nonlinearly vibrating charged drop was considered in the case when several modes were initially excited. It was found that, when the initial spectrum has two modes with increasing numbers, the second-order approximation in deviation of the drop shape from the equilibrium (spherical) one yields the time dependence of the translational (first with $n = 1$) mode amplitude. This dependence transforms the drop into a source of dipole acoustic waves. The time dependence of the translational mode amplitude has the form [4]

$$a_1^{(2)}(t) = -\sum_{i\in N}\frac{9ih_{i-1}h_i}{2(2i-1)(2i+1)} \times \{\cos[(\omega_i+\omega_{i-1})t] + \cos[(\omega_i-\omega_{i-1})t]\}, \tag{4}$$

where N is a set of the numbers of initially excited vibrational modes and h_i are the coefficients determining the partial contribution of an ith vibration mode to the total initial excitation:

$$\sum_{i\in N} h_i = 1.$$

The intensity of dipole acoustic radiation is given by

$$I_s = \frac{3\pi\nu^2R^2}{2V^2}\left\{\delta^2 R\frac{9ih_{i-1}h_i}{(2i-1)(2i+1)}h_ih_{i-1}\right\}^2 \times \{(\omega_i+\omega_{i+1})^4 + (\omega_{i+1}-\omega_i)^4\}, \tag{5}$$

where ν is the kinematic viscosity and δ is the amplitude of the initial deformation of the equilibrium drop shape.

The dependence of the dipole acoustic radiation intensity on the drop radius and amplitude of the initial disturbance of the second and third modes (expression (5)) is shown in Fig. 3.

(5) Comparing the data presented in Figs. 1–3 shows that the dipole and monopole acoustic radiation intensities exceed the quadrupole radiation intensity, all other things being equal. This fact looks somewhat strange, because the zeroth and first modes, which cause the monopole and dipole radiations, appears only in the second order of smallness in deviation of the drop shape from the equilibrium one, whereas the second mode exists even in the first order of smallness. Therefore, it would be natural to expect that the monopole and dipole components of the acoustic radiation will be of lesser intensity compared with the quadrupole component. However, the actual distribution of the acoustic radiation intensity over the mode numbers [9–11] implies that the second-order monopole and quadrupole components of the radiation make a decisive contribution to the total intensity.

To elucidate reasons for this phenomenon, consider a model problem of acoustic radiation from an abstract spherical source that may excite all modes including the first one.

Let a sphere of radius R execute axisymmetric vibrations with a small amplitude $\xi(\Theta, t)$, where $|\xi|/R \ll 1$, in an external ideal compressible liquid. Such a system generates acoustic radiation. Our aim is to find the intensities of the total acoustic radiation components related to various modes of the sphere vibration.

Mathematically, the problem of potential motion of a compressible medium is stated as

$$\frac{1}{V^2}\frac{\partial^2\psi}{\partial t^2} - \Delta\psi = 0; \tag{6}$$

$$r = R: \quad \frac{\partial\xi}{\partial t} = \frac{\partial\psi}{\partial r}; \tag{7}$$

$$r \longrightarrow \infty: \quad \frac{\partial\psi}{\partial r} + ik\psi = o\left(\frac{1}{r}\right). \tag{8}$$

Here, $\psi(\mathbf{r}, t)$ is the velocity field potential in the environment and is the Laplacian. The time dependence of the velocity field potential in the environment is assumed to be periodic: $\psi \sim i\omega t$; wave equation (6) is then transformed into the Helmholtz equation

$$\Delta\psi + k^2\psi = 0; \quad k \equiv \omega/V. \tag{9}$$

(6) The velocity potential of the acoustic field radiated by a spherical source, which is a solution to Eq. (9) subject to radiation condition (8), can be represented in the form

$$\psi(\mathbf{r}, t) = \sum_{n=0}^{\infty} A_n h_n^{(2)}(kr) P_n(\mu) \exp(i\omega_n t); \qquad \mu \equiv \cos\Theta, \tag{10}$$

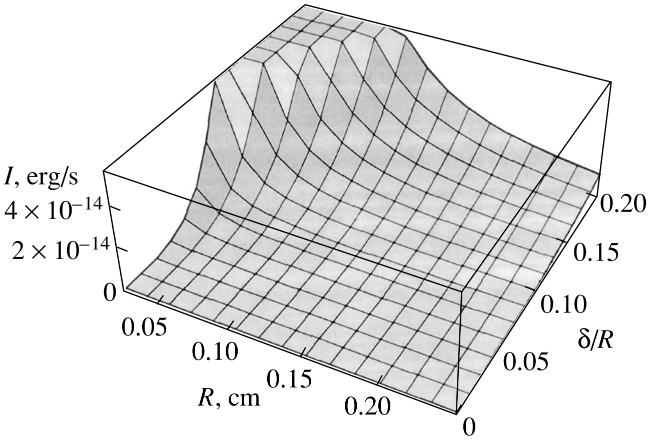


Fig. 3. Intensity of the dipole acoustic radiation from a nonlinearly vibrating water drop in air as a function of the drop radius and amplitude of the initial deformation of the fundamental mode at $W = 1$ and $h_2 = h_3 = 0.5$.

where $h_n^{(2)}(kr)$ are spherical Hankel functions of the second kind and $P_n(\mu)$ are Legendre polynomials.

The deviation of the vibrating drop surface from the spherical shape is described by the expression

$$\xi(\Theta, t) = \sum_{n=0}^{\infty} a_n \exp(i\omega_n t) P_n(\mu). \tag{11}$$

Substituting (10) and (11) into (7), we express the amplitude of the velocity field potential in the environment through the amplitude of the sphere surface vibration:

$$A_n = \frac{-i\omega_n a_n}{k[\partial h_n^{(2)}(kr)/\partial(kr)]_{r=R}}. \tag{12}$$

The amplitudes A_n of the velocity field potential for zeroth-, first-, and second-order acoustic waves are calculated by using (12) and expressing the spherical Hankel functions through trigonometric ones [12]:

$$h_0^{(2)}(z) = \frac{1}{2}[\sin(z) - i\cos(z)];$$

$$\frac{dh_0^{(2)}(z)}{dz} = \frac{\cos(z)}{z^2} - \frac{\sin(z)}{z^2} + i\left[\frac{\cos(z)}{z^2} + \frac{\sin(z)}{z}\right];$$

$$A_0(kR) = -i\omega_0 a_0 \left\{ k\left[\frac{\cos(kR)}{kr} - \frac{\sin(kR)}{(kr)^2}\right] + ik\left[\frac{\cos(kR)}{(kR)^2} + \frac{\sin(kR)}{kR}\right] \right\}^{-1};$$

$$h_1^{(2)}(z) = \frac{\sin(z)}{z^2} - \frac{\cos(z)}{z} + i\left[\frac{\cos(z)}{z^2} + \frac{\sin(z)}{z}\right];$$

$$\frac{dh_1^{(2)}(z)}{dz} = \left(\frac{1}{z} - \frac{2}{z^3}\right)\sin z + \frac{2\cos z}{z^2} + i\left[\left(\frac{1}{z} - \frac{2}{z^3}\right)\cos z - \frac{2\sin z}{z^2}\right];$$

$$A_1(kR) = -i\omega_1 a_1\left\{k\left[\left(\frac{1}{kR} - \frac{2}{(kR)^3}\right)\sin(kR) + \frac{2\cos(kR)}{(kR)^2}\right] + ik\left[\left(\frac{1}{kR} - \frac{2}{(kR)^3}\right)\cos(kR) - \frac{2\sin(kR)}{(kR)^2}\right]\right\}^{-1};$$

$$h_2^{(2)}(z) = \left(\frac{3}{z^3} - \frac{1}{z}\right)\sin z - \frac{3}{z^2}\cos z + i\left[\left(\frac{3}{z^3} - \frac{1}{z}\right)\cos z + \frac{3}{z^2}\cos z\right];$$

$$\frac{dh_2^{(2)}(z)}{dz} = \left(\frac{4}{z^2} - \frac{9}{z^4}\right)\sin z - \left(\frac{1}{z} - \frac{9}{z^3}\right)\cos z + i\left[\left(\frac{1}{z^2} - \frac{6}{z^3} - \frac{9}{z^4}\right)\cos z + \left(\frac{1}{z} - \frac{3}{z^2} - \frac{3}{z^3}\right)\sin z\right];$$

$$A_2(kR) = -i\omega_2 a_2\left\{k\left[\left(\frac{4}{(kR)^2} - \frac{9}{(kR)^4}\right)\sin(kR) - \left(\frac{1}{kR} - \frac{9}{(kR)^3}\right)\cos(kR)\right] + ik\left[\left(\frac{1}{(kR)^2} - \frac{6}{(kR)^3} - \frac{9}{(kR)^4}\right)\cos(kR) + \left(\frac{1}{(kR)} - \frac{3}{(kR)^2} - \frac{3}{(kR)^3}\right)\sin(kR)\right]\right\}^{-1}.$$

If the argument kr of the Hankel function entering into formula (10) is large (the so-called "wave zone" of acoustic radiation), this function can be represented as a series in negative powers of the argument according to the well-known relationships [12]:

$$h_n^{(2)}(z) = i^{n+1}z^{-1}\exp(-iz)\sum_{m=0}^{\infty}\frac{(n+m)!}{m!(n-m)!}(2iz)^{-m}. \quad (13)$$

In addition, at large values of the argument under the summation sign in expression (13), only the first term with $m = 0$ may be left:

$$z \gg 1: h_n^{(2)}(z) \approx i^{n+1}z^{-1}\exp(-iz).$$

Recall that such a representation is valid at large distances from the drop surface in the wave zone that is characterized by the conditions

$$r \gg R, \quad kR \ll 1.$$

In this zone, the geometry of the radiating system does not influence the shape of waves, which become spherical, and their amplitudes decrease with distance as $1/r$.

Thus, at distances much larger than the length of the acoustic wave radiated, the velocity field potential in the environment is represented by the superposition of divergent spherical waves and may be written in the form of the series

$$\psi(\Theta, t) \approx \sum_{n=0}^{\infty} A_n i^{n+1}\frac{1}{kr}\exp(i\omega t - ikr)P_n(\mu).$$

Since the velocity field of the environment is defined by the potential gradient $\mathbf{U} = \mathrm{Re}[\mathrm{grad}\psi]$, the velocity field in the environment in the approximation linear in r^{-1} has the form

$$\mathbf{U}(\mathbf{r}, t) \approx \mathrm{Re}\left[\sum_{n=0}^{\infty} -i^{n+2}A_n\frac{1}{r}\exp(i\omega t - ikr)P_n(\mu)\right]\mathbf{n}, \quad (14)$$

that is, contains only the radial component.

With the complex amplitude A_n represented in the form $A_n = \alpha_n\exp(i\beta_n)$, (14) can be recast as

$$\mathbf{U} = \left[\sum_{n=0}^{\infty} -\frac{\alpha_n}{kr}\cos(\omega t - kr + \kappa_n)P_n(\mu)\right]\mathbf{n};$$

$$\kappa_n = \frac{1}{2}\pi(n+2)\beta_n.$$

Now we determine the amplitude α_n of the velocity field for the zeroth, first, and second modes:

$$\alpha_n = \mathrm{mod}(A_n); \quad \alpha_0 = \sqrt{\frac{a_0^{(2)}k^4R^4\omega_0^2}{1 + k^2R^2}};$$

$$\alpha_1 = \sqrt{\frac{a_1^2k^6R^6\omega_1^2}{1 + k^4R^4}};$$

$$\alpha_2 = (a_2^2k^8R^8\omega_2^2)^{1/2}[81 + 54kR + 18k^2R^2 + 3k^3R^3 + k^4R^4 - 3k^5R^5 + k^6R^6 + 3kR(18 + 27kR - 5k^2R^2 - 6k^3R^3 + k^4R^4)\cos(2kR) - 3kR(18 - 15kR - 17k^2R^2 + 3k^3R^3 + k^4R^4)\sin(2kR)]^{-1/2}. \quad (15)$$

Let us compare the intensities of the acoustic waves with $n = 0$, 1, and 2, keeping in mind that the acoustic field intensity is proportional to the velocity field squared and assuming that $a_0 = a_1 = a_2$, $\omega_0 = \omega_1 = \omega_2$, and $kR \ll 1$:

$$\left(\frac{\alpha_1}{\alpha_2}\right)^2_{kR \ll 1} \approx \frac{81}{4(kR)^2}; \quad \left(\frac{\alpha_0}{\alpha_2}\right)^2_{kR \ll 1} \approx \frac{81}{4(kR)^4}. \qquad (16)$$

Since in these estimates the wavelength of the acoustic wave from the range of audibility is much larger than the drop radius, it follows that $kR \ll 1$. Eventually, from expressions (16), we may conclude that the monopole radiation is much more intense than the dipole and quadrupole ones and the dipole radiation is more intense than the quadrupole one, all other things being equal.

(7) The results of comparing the acoustic radiation intensities for the three first modes of the model spherical radiator should be complemented by two estimates concerning the amplitudes and frequencies of various modes prior to when they can be applied to a vibrating drop.

When making the estimates in the previous section, we assumed that the mode frequencies of the model radiator are equal to each other. Such an assumption is incorrect for the drop, and the frequencies of the linear oscillations of the fundamental mode of the drop in the environment are defined by expression (2). Thus, this expression gives us the frequency of the quadrupole acoustic radiation. The frequency of the nonlinear oscillations of the fundamental mode of the drop, which is responsible for the monopole radiation, is twice as high as the oscillation frequency of the mode governing the initial deformation (of the fundamental mode in our case); that is, $\omega_0 = 2\omega_2$.

Hence, according to (15), the ratio $(\alpha_0/\alpha_2)^2$ will be four times greater than that given by (16). Similarly, one can refine the value of the ratio α_1/α_2. It will also increase if one takes into account that, according to (5), the frequency of the translational mode in our situation depends on the sum of the frequencies of the fundamental and third modes of the drop oscillations and that the frequency increases as $n^{3/2}$ as the mode number increases [3]. It is easy to calculate that, with the dependence of the frequency on the mode number taken into consideration, the ratio $(\alpha_1/\alpha_2)^2$ increases approximately by one order of magnitude. Thus, if the actual dependence of the frequencies of the drop zeroth, translational, and fundamental modes on the mode numbers is included, estimates (16) strengthen.

Next, it should be noted that, when estimating the intensity of the monopole and quadrupole components of the acoustic radiation, we considered the initial excitation of the fundamental mode alone, whereas in estimating the intensity of the dipole radiation, we were forced to consider the excitation of the second and third modes, since only in this case the translational mode is generated via nonlinear interaction. Strictly speaking, to be correct, we had to evaluate the intensities of the monopole and quadrupole components at the initial excitation of the same two, second and third, modes. It is easy to check that the additional excitation of the third mode in no way influences the intensity of the quadrupole radiation, because in the approximation linear in drop vibration amplitude, where this component of the radiation appears, the modes of the drop vibration do not interact with each other and energy transfer from the third mode to the second one does not take place. At the same time, the third mode, when excited, changes the zeroth mode amplitude insignificantly up to a factor of ~ 1 [13]. Since the amplitude of the velocity field of the monopole acoustic radiation, which is related to the zeroth mode, for exceeds the amplitudes of the dipole and quadrupole radiations, the above ratio between the radiation intensities remains practically unchanged.

CONCLUSIONS

Analytical estimates of the acoustic radiation intensity from a vibrating drop are based on expansions of two types: in multipoles (in inverse powers of the dimensionless distance $1/kr$ to the point of observation in the so-called wave zone of radiation) and in drop vibration amplitudes (in powers of $|\xi|/R$). In natural liquid drop systems, for example, in rainy clouds, the characteristic values of the parameter $|\xi|/R$ are on the order of several tenths and equal (by order of magnitude) to $1/kr$ (although $1/kr$ may be much less at larger distances from the drop). The multipole expansion in the wave zone of the radiation yields the velocity field monopole component of the acoustic radiation on the order of $a_0(kR)^2/(kr)$; the dipole component, on the order of $a_1(kR)^3/(kr)$; and the quadrupole component, on the order of $a_2(kR)^4/(kr)$, where $kR \ll 1$, because the acoustic wavelength from the range of audibility exceeds the drop radius by many orders. In other words, at equal amplitudes a_i, the monopole component is the most intense. Next, the expansion in the drop vibration amplitude yields the peak value of the velocity field quadrupole component on the order of $a_0 \sim (|\xi|/R)$, while the monopole and dipole components are on the order of a_1 and $a_2 \sim (|\xi|/R)^2$, respectively. Eventually, the orders of smallness of the velocity field multipole components of the acoustic radiation are defined by the relationships

$$\alpha_0 \sim (|\xi|/R)^2(kR)^2/(kr), \quad \alpha_1 \sim (|\xi|/R)^2(kR)^3/(kr),$$

$$\alpha_2 \sim (|\xi|/R)(kR)^4/(kr).$$

From these expressions, it follows that the ratios of the amplitude coefficients of the acoustic radiation in the wave zone depend on the relationship between two small parameters $|\xi|/R$ and kR. As was mentioned above, $|\xi|/R \sim 0.1$, while kR for natural liquid drop systems that radiate sound in the range of audibility is less than unity by several orders (for example, for a coarse rainy drop with $R = 0.01$ cm radiating an acoustic wave with a frequency of 1 kHz, $kR = 3 \times 10^{-4}$). This fact explains the result obtained in [9–11]: the monopole and dipole components of acoustic radiation from a vibrating drop in the range of audibility, which are caused by the second-order vibration amplitude, turn out to be more intense than the quadrupole component, which is due to the first-order fundamental mode of the vibration. The value of the intensity ratio will also be the same in the ultrasonic range, where $|\xi|/R > kR$.

REFERENCES

1. H. L. Green and W. R. Lane, *Particulate Clouds: Dusts, Smokes, and Mists* (Spon, London, 1964, 2nd ed.; Khimiya, Leningrad, 1969).
2. L. G. Kachurin, *Physical Principles of Impact on Atmospheric Processes* (Gidrometeoizdat, Leningrad, 1990).
3. J. W. Strutt (Lord Rayleigh), *The Theory of Sound* (Macmillan, London, 1894; Gostekhizdat, Moscow, 1955), Vol. 2.
4. N. A. Fuks, *Mechanics of Aerosols* (Akad. Nauk SSSR, Moscow, 1955).
5. V. Sh. Shagapov, Izv. Akad. Nauk, Fiz. Atmos. Okeana **24**, 506 (1988).
6. P. V. R. Suryanarayana and Y. Bayazitoglu, Phys. Fluids A **3**, 967 (1991).
7. Won-Kyu Rhim, Sang Kun Chung, M. T. Hyson, *et al.*, IEEE Trans. Ind. Appl. **IA-23**, 975 (1987).
8. A. L. Yarin, G. Brenn, O. Kastner, *et al.*, J. Fluid Mech. **399**, 151 (1999).
9. A. I. Grigor’ev and A. R. Gaibov, Zh. Tekh. Fiz. **71** (11), 6 (2001) [Tech. Phys. **46**, 1351 (2001)].
10. A. R. Gaibov, S. O. Shiryaeva, A. I. Grigor’ev, and D. F. Belonozhko, Pis’ma Zh. Tekh. Fiz. **29** (4), 22 (2003) [Tech. Phys. Lett. **29**, 138 (2003)].
11. A. I. Grigor’ev, S. O. Shiryaeva, A. R. Gaibov, and D. F. Belonozhko, Pis’ma Zh. Tekh. Fiz. **27** (22), 7 (2001) [Tech. Phys. Lett. **27**, 934 (2001)].
12. *Handbook of Mathematical Functions*, Ed. by M. Abramowitz and I. A. Stegun (Dover, New York, 1971; Nauka, Moscow, 1979).
13. S. O. Shiryaeva, Zh. Tekh. Fiz. **72** (4), 15 (2002) [Tech. Phys. **47**, 389 (2002)].

Translated by N. Mende

Technical Physics, Vol. 48, No. 10, 2003, pp. 1249–1254. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 29–35.
Original Russian Text Copyright © 2003 by Okulov, Meledin, Naumov.

Experimental Investigation of a Swirling Flow in a Cubic Container

V. L. Okulov, V. G. Meledin, and I. V. Naumov
Institute of Thermal Physics, Siberian Division, Russian Academy of Sciences,
pr. Akademika Lavrent'eva 1, Novosibirsk, 630090 Russia
e-mail: okulov@itp.nsc.ru

Received January 8, 2003

Abstract—A flow induced by a rotating disk mounted at the top of a cubic container totally filled with a liquid is studied experimentally. The flow pattern is visualized for Reynolds numbers in the interval 1500–6000, and the velocity variation along the axis of the container, which coincides with the axis of revolution of the disk, is observed by means of Doppler laser anemometry (DLA). As Re grows, the velocity axial component starts fluctuating because of the vortex core precession. The breakdown of the vortex helical structure becomes pronounced at Re > 4000 without the formation of the return flow region (vortex breakdown bubble) at the axis. With the Reynolds number and the container height-to-disk radius ratio being the same, the axial flow patterns in standard cylindrical [1] and cubic containers differ radically. In the latter vessel at low Re, the steady flow regime and axisymmetric bubble breakdown of the vortex structure near the axis are absent. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

Variation of vorticity is known to cause significant structural changes in vortex flows, which reduces the efficiency of their applications. Among these changes are vortex breakdowns, which originate at the forward stagnation point near the axis behind which a limited area of return flow (a vortex breakdown bubble) forms. Two types of vortex breakdown occur most frequently: axisymmetric (bubble) and helical. Vortex breakdown is observed in vortex flows, in wing leading edge flow separations, as well as in intensely swirling flows in pipes and closed rotating-base cylinders [2].

A flow in a closed cylinder with a rotating lid, which generates a vortex structure coaxial with the axis of revolution of the lid, is the simplest model for studying the breakdown of an axial vortex. Structural changes in the flow depend, in this case, only on two parameters: the container height-to-disk radius ratio H/R and the Reynolds number $\mathrm{Re} = \Omega R^2/\nu$, where Ω is the angular velocity of the disk and ν is the kinematic viscosity of the fluid. Having visualized the flow pattern in a wide range of Re and H/R, Escudier [1] found regions with various steady and unsteady axisymmetric regimes of vortex breakdown. Figure 1 exemplifies the axisymmetric breakdown of a bubble-like vortex structure and shows the flow pattern in a cylindrical container.

The aim of this work is to find swirling flow regimes and visualize (for the first time) the breakdown of an axial vortex structure in a cubic container.

The pioneering investigation of flows in a container of cubic geometry was accomplished by numerical techniques [3]. The steady-state Navier–Stokes equations were solved for Re = 500, 1000, 1500, and 2000. In the absence of relevant experimental data, Chiang *et al*. [3] assumed that the flow regimes are steady-state and laminar like those in a cylindrical vessel [1]. However, tentative experimental data for the pulsation of a swirling flow in a closed rectangular container with $H/R = 2$ [4] suggested the dominant role of vibrations in the flow, which indicate that the flow is unsteady, including in the Re range studied in [3]. The precision DLA measurements for the temporal characteristics of the axial and tangential velocity components showed that the frequency of the first peak of these vibrations is proportional to (and roughly equal to 1/16 of) the frequency of an external excitation (disk rotation). Experimental studies of vortex breakdown in pipes also detected considerable flow velocity variations in time. Faler and Leibovich [5] obtained the velocity profiles and energy spectra in the vortex breakdown region. The particle motion in the breakdown region was found to be unsteady with regular low-frequency vibrations. The travel of the breakdown region made velocity field measurements in its neighborhood extremely difficult. The main vibration period in the experimental setup was about 2 Hz at a fixed flow rate. In general, the findings of Faler and Leibovich do not allow one to judge the reason for the nonstationarity of a swirling flow in a cubic container; therefore, the problem calls for a more careful examination.

The need for studying the vortex flow in square-section working chambers is dictated by their frequent use in practice. Facilities of rectangular geometry are simpler in design and cheaper [6, 7]. The latter factor becomes of particular importance for large vortex

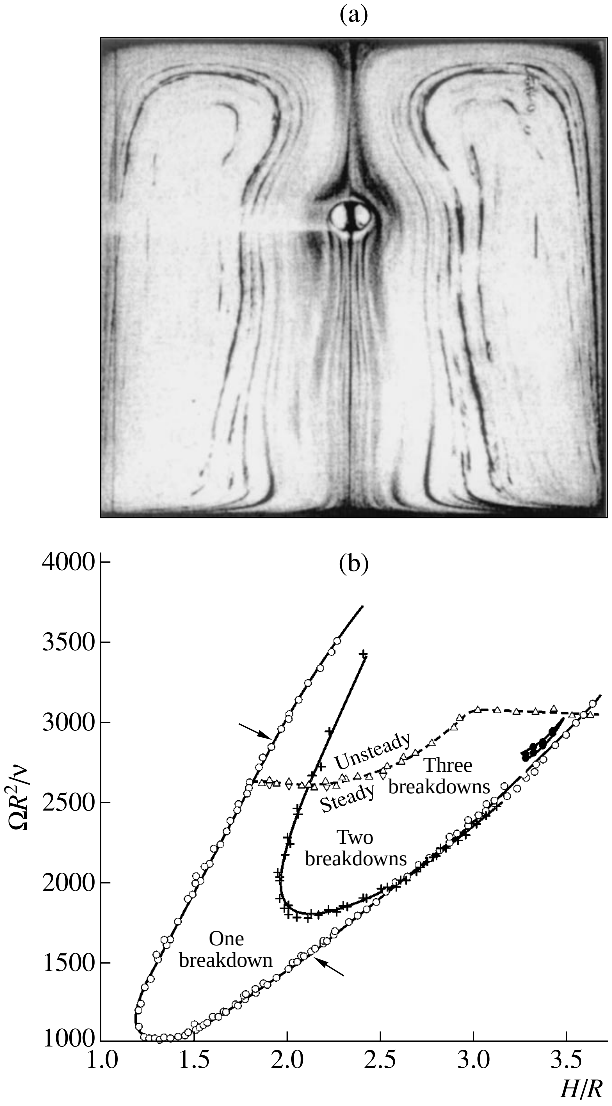


Fig. 1. (a) Flow pattern (Re = 1490) and (b) vortex evolution diagram in a cylindrical container [1].

chambers, e.g., cyclone tangential furnaces used in power engineering. Also, in the case of the rectangular geometry, special optical means for diagnosing vortex flows may be applied [7, 8].

It is reasonable to compare swirling flows in canonical (cylindrical) and noncanonical (in the form of a parallelepiped) vortex chambers. Alekseenko *et al.* [7] studied vortex flows in a rectangular chamber with transparent cylindrical insertions. In cylindrical and rectangular parts, the structure of concentrated unbroken axial vortices was found to be identical for several regimes. This fact was explained by the presence of smoothing corner vortices in a rectangular chamber, which produce an imaginary almost cylindrical surface inside of which the basic flow develops. However, the flow regimes and vortex breakdown conditions were not compared in that work and the statement that “the flows are almost identical to each other” was not supported in a wider range of flow parameters.

The vorticity and the shape of the vortex core to a great extent depend on its interaction with the circumferential boundary layer. In rectangular and cylindrical containers, this interaction is substantially different. In the former, the interaction is accomplished indirectly via smoothing corner vortices. Here, one may expect significant changes in the flow pattern, since the vortex breakdown is extremely sensitive to external effects. For these reasons, one more goal of this work is to compare the swirl flow patterns in canonical cylindrical [1] and noncanonical cubic experimental chambers.

To this end, the flow pattern in a cubic container was visualized for the first time with the technique used in [1]. The presence of vortex breakdown bubbles appearing along the container axis aligned with the axis of revolution of the disk were judged from the velocity axial component profiles taken with a Doppler laser anemometer.

EXPERIMENTAL

We used closed container *1* (Fig. 2) of square section with 120 mm on a side. Rotating disk *2* of radius R = 59.5 mm is mounted on the upper lid of the container. The height of the container can be varied from 0 to 140 mm. To provide cubic form, the height was set equal to 120 mm (H/R = 2).

The constancy of the angular velocity Ω of the disk is provided by a system of frequency–phase stabilization. The mean angular velocity of the disk was kept constant within 0.2% in the interval (3–100) rpm. A water–glycerol (40%) mixture was used as a working fluid. The kinematic viscosity of the fluid ν was 3.1×10^{-5} m^2/s at 25°C. Throughout the experiment, it was controlled with a viscosimeter. The temperature of the working fluid was controlled with a digital thermometer. The temperature was kept constant within 0.2°C. Since the viscosity of the water–glycerol fluid greatly depends on temperature, the viscosity/temperature ratio was carefully controlled so that the total error in determining Re be no more than ±10 for Re = 1000–6000.

The working section of the container was made of transparent optical-grade resin glass. Visualization was accomplished by introducing a dye at the container axis [1, 9]. A small amount of a water solution of fluorescein, which colors the working fluid, was added to the container from vessel *3* through 1-mm-diam. tube *4* and then through a 0.3-mm-diam. hole at the center of the container bottom (Fig. 2). The dye feed depends on the position (height) of the vessel relative to the container bottom. The height was selected in such a way that the dye filled the lower part of the container and then was entrained by the flow along the vortex axis. A 20-mW He–Ne laser was applied as a light source. Flow patterns in various regimes were recorded by video camera *5*, which was mounted perpendicularly to the side wall of the container. The images were examined visually and also were processed with PC *6*. Such a technique provides the reliable diagnostics of flow regimes in a cubic container.

The axial component of the flow velocity at the axis of revolution of the disk was measured by using a DLA optical sensor with adaptive time selection (ATS) [10]. ATS DLA sensor *7* is based on a two-frequency differential optical scheme with a carrier of 80 MHz and provides the acoustooptical switching of measuring channels. The laser radiation wavelength was 514 nm, and the focal length of the optical system was $F = 500$ mm. The probing light spot measured 0.5 × 0.05 mm. Polystyrene beads were used as a light-scattering medium.

A Doppler signal to be processed is sequentially applied to low-pass filter *8*, ADC *9*, and the PC. The synchronous digitization of the signal retains exact phase relationships between the cosinusoidal and sinusoidal components of a quadratic pair. Filtering and frequency determination were accomplished with a dedicated algorithm based on fast Fourier transformation [4]. The accuracy of velocity measurements were no worse than 1% in the velocity range 10^{-4}–10^{-1} m/s.

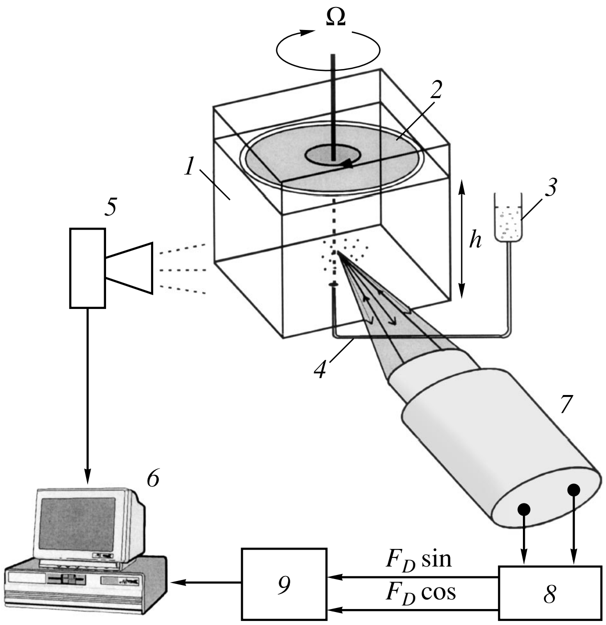


Fig. 2. Experimental setup used with a cubic container.

RESULTS AND DISCUSSION

A swirling flow was examined in the closed cubic container ($H/R = 2$) fully filled with the fluid in a wide range of Reynolds numbers.

Figure 3 shows the vortex core breakdown patterns for different Re in the cubic container. The patterns differ significantly from those in a cylindrical container. In all the regimes, the flow is unsteady. In the images, the stream filament colored executes rotary motion and precesses with an angular velocity that is lower than that of the disk.

Figure 4 demonstrates the variation of the velocity axial component along the axis of revolution of the disk (vertical axis of the container) for Re = 1500–4000. On the abscissa axis, the distance from the container bottom is plotted.

The velocity axial component profiles were recorded at random time instants relative to the phase oscillations of the vortex structure. At each point, the time of averaging was 10 s. From the data presented, it follows that fluctuations of the velocity axial component build up with Re. The maximal fluctuations are observed 50–110 mm away from the container bottom. However, unlike cylindrical containers, the zone of return flow (where V is negative) does not occur in the given interval of Re.

Figure 5 shows the images of vortex structure breakdown at Re = 5000, 5500, and 6000. In this case, fluorescein filled the bottom part of the container and then was entrained by the flow along the vortex axis, visualizing the core of the precessing vortex structure. With Re = 5000 and 5500, the stable picture of the vortex structure helical breakdown is observed; with Re = 6000, the reconstruction of the vortex structure takes place.

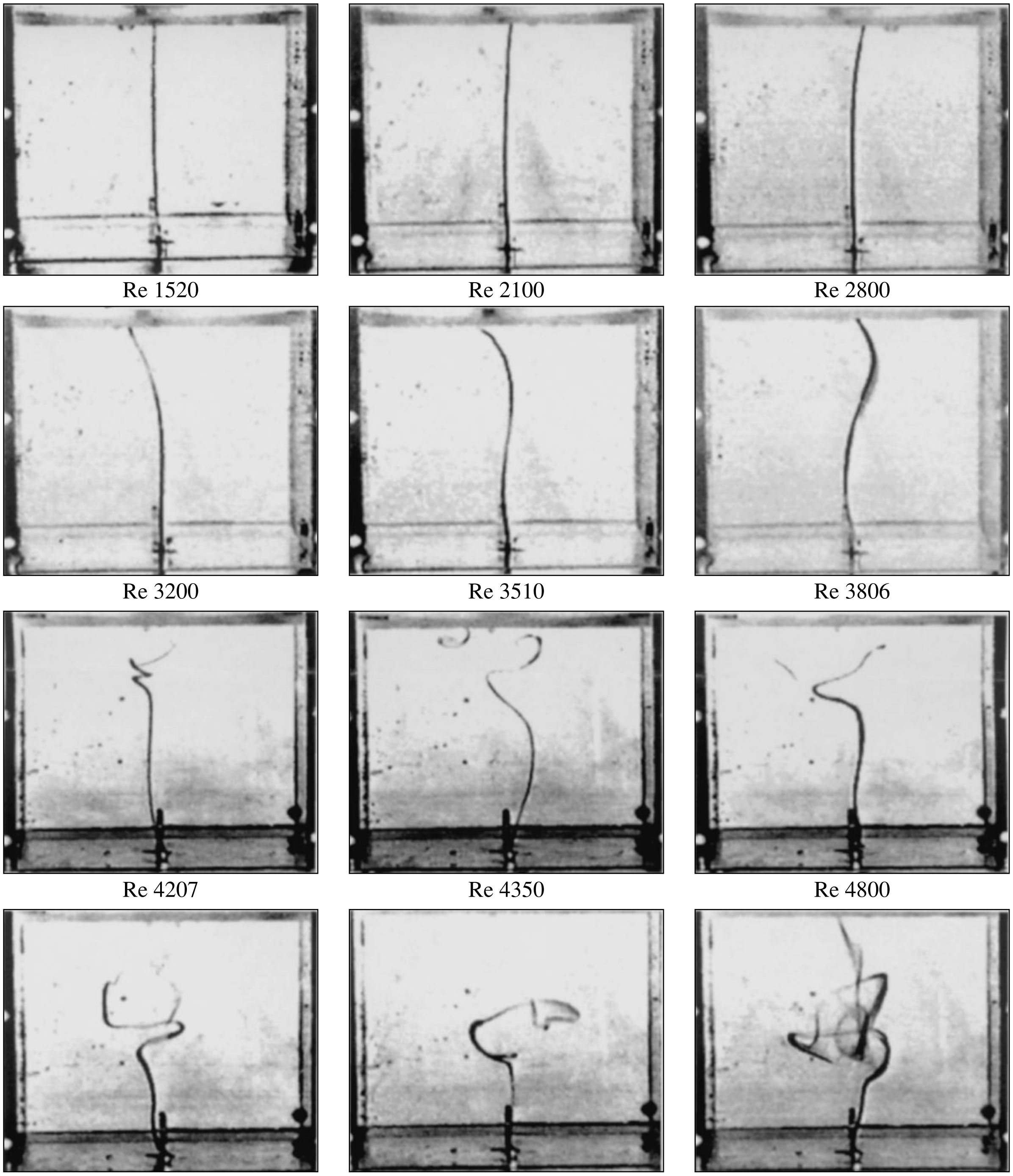
The evolution of the helical breakdown was studied by measuring the velocity axial component for Re = 5000 and 6000 (Fig. 6). The measurements were performed as follows.

(1) At each point, the axial component profile was averaged over four measurements. The time of averaging was 5 s. Averaging was made at random time instants:

$$\bar{V} = \frac{\sum_{i=1}^{n} V_i}{n}, \quad n = 4. \tag{1}$$

(2) In the maximal fluctuation zone (zone II in Fig. 6), the maximal and minimal values of the axial velocity component were measured at Re = 5000 and 6000. The time of averaging at each point was also 5 s.

Figure 6 shows the distribution of the axial velocity at Re = 5000 and 6000. In zone II (70–100 mm), the behavior of the axial component differs greatly from that for Re = 1500–4000 (Fig. 4). The fluctuation amplitude increases considerably: $V_{\max}/V_{\min} \geq 2$ versus $|V_{\max} - V_{\min}| \leq 30\%$ of the mean velocity in zones I and III. Comparing Figs. 6a and 6b, one may judge the extension of the intense velocity fluctuation zone and the growth of the fluctuation amplitudes with increasing Re. In addition, at Re = 6000, local short-lived regions of return flow (V is negative) arise. However, stable vortex breakdown bubbles, as in the case of bubble breakdown [1], do not form.



Re 1520 Re 2100 Re 2800

Re 3200 Re 3510 Re 3806

Re 4207 Re 4350 Re 4800

Re 5160 Re 5650 Re 6260

Fig. 3. Visualization of the vortex core breakdown in the closed cubic container (the dye is introduced at the geometrical center of the container).

Thus, it has been found that the flow regimes in a closed cubic container are unsteady in the Re range studied. At Re > 4000, only the helical breakdown of the vortex core is observed. The nonstationary zone of return flow appears at only Re = 6000; however, a stable vortex breakdown bubble does not form.

FLOW EVOLUTION IN CONTAINERS OF VARIOUS GEOMETRY

In the experiments described above, it was found that the swirling flow patterns in cylindrical and cubic containers differ substantially. In the latter, the bubble

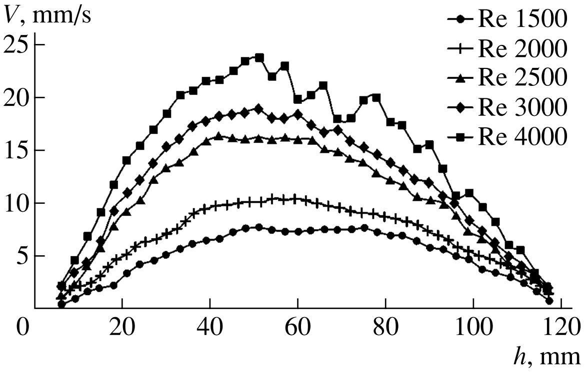


Fig. 4. Velocity distribution along the axis of revolution of the disk.

vortex breakdown and steady flow regimes at low Re are absent.

Visualization (Fig. 3) makes it possible to compare the vortex core structures in a cylindrical container [1]. The absence of the bubble vortex breakdown in the cubic container is clearly illustrated for Re = 1520–3200. In the cylindrical container, conversely, the steady bubble breakdown with one or two bubbles is distinctly seen in this Re range. It is noteworthy that, as Re increases roughly to 3200 in the cubic container, the colored flow core near the rotating disk is progressively deflected from the vertical axis and precesses slowly about it with a frequency that is proportional to, but much lower (roughly 16 times) than, the disk rotation frequency. The unsteady flow regime is distinctly seen.

For this Re range in the cylindrical container, the flow is steady and the bubble vortex breakdown is observed. The diagram (Fig. 1) shows the dependence of the number and position of breakdown bubbles in the cylindrical container as a function of Re and H/R. With $H/R = 2$, there is one breakdown bubble for Re varying between 1450 and 1800 and between 2200 and 3000. In the interval Re = 1800–2200, two breakdown bubbles appear. In this range of relatively low Re, the flow patterns in the central parts of the containers differ radically.

In the cylindrical container, the flow becomes unsteady only at Re > 2550, as follows from the diagram in Fig. 1. With Re = 2500, only one breakdown bubble is observed; as Re increases further, the breakdown region starts vibrating in the vertical direction but the flow retains axial symmetry. In the cubic container, the 3D helical structure appears at once. As Re grows, the deflection of the breakdown region from the vertical axis becomes more pronounced and the flow core takes the slightly conical helical shape and make a complete turn over the distance from the container bottom to its top (Fig. 3, Re = 3510 and 3806).

For Re more than 4000, the flow pattern undergoes new changes. The vortex filament exhibits a saw-toothed kink, which rotates and vertically vibrates. It may be assumed that the return flow zone appears and disappears in the flow, as in the case of the bubble breakdown of the vortex structure in the cylinder. Such a zone appears for a very short time, which is smaller than the swirling flow vibration period. Its appearance and disappearance distorts the helical shape of the trajectory of the dye introduced into the vortex core.

A region appears where the dye slows down in the axial direction and, going on rotating, makes a tight turn (the saw-toothed region for Re = 4207 in Fig. 3). Visual observations suggest that, as Re grows, the existence time of the return flow region increases and the region itself extends down to the container bottom. In this situation, the dye outlines this region in Fig. 3 more distinctly.

DLA measurements also reliably detect the return flow region (Fig. 6b). As Re grows further, the existence time of the vortex breakdown bubble becomes comparable to the vortex core precession period and the bubble in the flow persists and vibrates about the vertical axis. Thus, at high Re, the complicated unsteady flow executing two types of vibration is established: the precession of the vortex helical structure about the vertical axis of the container and the vibration of the return flow region along this axis. As the Reynolds number increases, the amplitude of both types of vibration grows and visualization at Re > 6000 makes no sense.

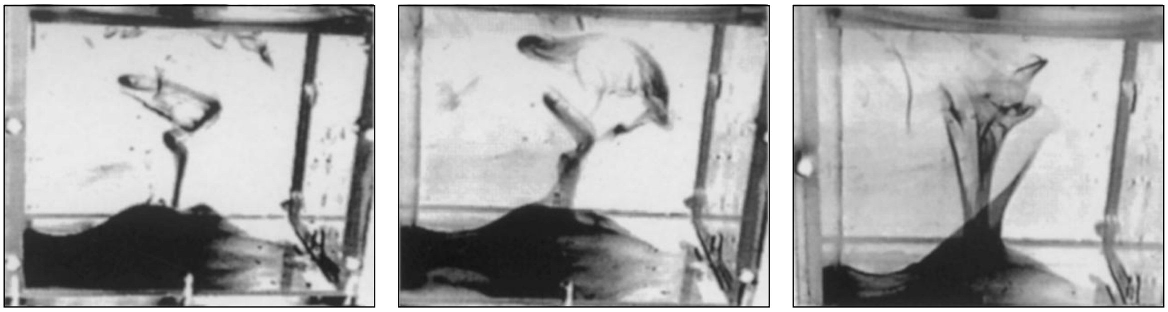

Re 5000 Re 5500 Re 6000

Fig. 5. Visualization of the vortex core breakdown in the closed cubic container (the dye covers the bottom of the container).

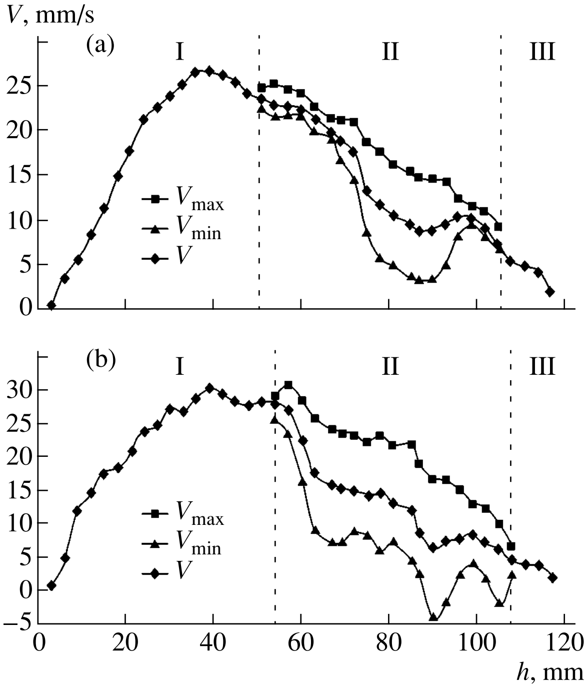


Fig. 6. Velocity distribution along the axis of revolution of the disk in the closed cubic container ($H/R = 2$). The velocity values are averaged over four measurements. Re = 5000 (a) and 6000 (b).

Summarizing the results obtained, we may conclude that, as Re grows, the flow pattern in cylindrical and cubic containers undergoes essentially different space–time changes. While the flow in a cylindrical container is axisymmetric up to Re ≈ 4000, the flow in a cubic container is asymmetric and unsteady even at very low Re (Re ≈ 1500). The changes of the flow regimes in a cubic container are as follows.

(1) Up to Re ≈ 3200, the flow exhibits a slight precession of the vortex core and return flow regions (vortex breakdown bubbles) do not appear. Near the disk rotating at the top of the container, the core is somewhat deflected from the vertical axis.

(2) With Re exceeding 3200, the precession of the vortex core smoothly passes into the precession of the vortex helical structure, which occupies the entire container. Vortex breakdown bubbles do not form either.

(3) With Re > 4200, the 3D unsteady flow prevents the formation of a stable breakdown bubble; however, along with the precession of the vortex helical core, clear-cut axial vibrations appear.

(4) As Re grows further, these two types of vibration become comparable in intensity, breakdown bubbles are detected by DLA measurements, and the structure becomes so complicated that flow visualization seems to be impossible (at Re > 6000).

It should be noted that the flow structure variation scenarios in a cubic container (see above) and a cylindrical container (see the diagram in Fig. 1b for $H/R = 2$) radically differ.

CONCLUSIONS

The swirling flow pattern evolution in a closed cubic container was experimentally studied at Re varying from 1500 to 6000. The swirl was generated by the rotating disk mounted on the top of the container. The velocity distribution along the container axis, which is aligned with the axis of revolution of the disk, was measured. Throughout the Re range, only unsteady flows were observed. This should be taken into account in mathematical simulation (that is, the model considered in [3] is invalid). Based on the experimental data, we showed that the amplitude and frequency of vortex core vibrations grows with increasing Re. In a closed cube, the distinct helical breakdown of the vortex core was observed with Re > 4000. However, unlike the flow in a cylinder, here a stable return-flow region does not arise. While in a cylinder the axisymmetric steady flow is known to change to axisymmetric unsteady and then to asymmetric unsteady flow, in a cube, the asymmetric unsteady flow is established almost at once.

ACKNOWLEDGMENTS

This work was partially supported by INTAS (grant no. 00-00232) and the Russian Foundation for Basic Research (grant no. 01-01-00899).

REFERENCES

1. P. Escudier, Exp. Fluids **2**, 189 (1984).
2. S. Leibovich, Annu. Rev. Fluid Mech. **4**, 185 (1978).
3. T. R. Chiang, W. H. Sheu, and S. F. Tsai, Comput. Fluids **28**, 41 (1999).
4. Yu. N. Dubnishchev, V. G. Meledin, and I. V. Naumov, Avtometriya, No. 5, 30 (2000).
5. J. H. Faler and S. Leibovich, J. Fluid Mech. **86**, 312 (1978).
6. S. V. Alekseenko and V. L. Okulov, Teplofiz. Aéromekh., No. 2, 101 (1996).
7. S. V. Alekseenko, P. A. Kuibin, V. L. Okulov, *et al.*, J. Fluid Mech. **382**, 195 (1999).
8. V. A. Arbuzov, Yu. N. Dubnishchev, A. V. Lebedev, *et al.*, Pis'ma Zh. Tekh. Fiz. **23** (23), 84 (1997) [Tech. Phys. Lett. **23**, 938 (1997)].
9. A. Spohn, M. Mory, and E. J. Hopfinger, J. Fluid Mech. **370**, 73 (1998).
10. P. Ya. Belousov, Yu. N. Dubnishchev, V. G. Meledin, *et al.*, Optica Applicata **20** (3), 187 (1990).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1255–1257. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 36–39.

GASES AND LIQUIDS

New Equation for the Polarization Fourier Components of the Small-Scale Velocity of an Incompressible Fluid in Anisotropic Turbulence

A. M. Balonishnikov
St. Petersburg State University of Economics and Engineering, St. Petersburg, 191002 Russia
e-mail: balonalex@yahoo.co.uk, balonale@engec.ru

Received December 24, 2002; in final form, February 25, 2003

Abstract—A new equation for the small-scale polarization Fourier components of the incompressible fluid velocity in the case of anisotropic turbulence is suggested. The principal invariant of the strain rate tensor for the large-scale velocity is found. This invariant is of most significance for the subgrid simulation of fully developed turbulence. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

Most researchers (among them the author) believe that fully developed turbulence in an incompressible fluid may be described by the Navier–Stokes equations (see, e.g., [1, 2]). The basic difficulty in this case is that modern computers cannot solve these equations when a Reynolds number is large. Such a situation is widely met in practice, since a large number of degrees of freedom are usually excited. Yet, reasonable results have been obtained with semiempirical models. This is an indirect indication that excited degrees of freedom are not all of significance if we are interested in only the large-scale long-term evolution of fully developed turbulence. It may be assumed that if one subdivides in the standard way the instantaneous velocity $\mathbf{v}(\mathbf{x}, t)$ and pressure $p(\mathbf{x}, t)$ into large-scale and small-scale components, the energetically significant small-scale components will accommodate the large-scale ones. One way of closing turbulence consists in approximately solving equations for the small-scale velocity and pressure components by expressing them through the large-scale ones. Such an approach is employed, for example, in the rapid distortion turbulence (RDT) theory (see, e.g., [3]). Below, we will derive an approximate equation for the small-scale velocity that is simpler than that used in [3].

EQUATION FOR SMALL-SCALE POLARIZATION FOURIER COMPONENTS OF VELOCITY IN THE CASE OF ANISOTROPIC TURBULENCE

Let us subdivide the pressure $p(\mathbf{x}, t)$ and velocity $\mathbf{v}(\mathbf{x}, t)$ into large-scale, $P(\mathbf{x}, t)$ and $\mathbf{V}(\mathbf{x}, t)$, and small-scale, $p'(\mathbf{x}, t)$ and $\mathbf{u}(\mathbf{x}, t)$, components:

$$p = P + p'; \quad \mathbf{v} = \mathbf{V} + \mathbf{u}. \tag{1}$$

We assume that $\langle p' \rangle = 0$ and $\langle \mathbf{u} \rangle = 0$, where $\langle \ldots \rangle$ means averaging over a cubic cell of edge L. Then, for the small-scale components, we obtain the equations [2]

$$\begin{gathered} \partial_t u_i + \partial_k (V_i u_k + V_k u_i) + \rho^{-1} \partial_i p' \\ = \partial_k (\nu \partial_k u_i - u_i u_k + \langle u_i u_k \rangle), \\ k = 1, 2, 3; \quad i = 1, 2, 3, \end{gathered} \tag{2}$$

where ν is the molecular kinematic viscosity and ρ is the density of the fluid, and the incompressibility equation

$$\partial_i u_i = 0 \tag{3}$$

(hereafter, summation goes over repeating indices).

As is known, the large-scale velocity also satisfies the incompressibility equation:

$$\partial_i V_i = 0. \tag{4}$$

In view of Eqs. (3) and (4), Eq. (2) may be written as

$$\begin{gathered} \partial_t u_i + u_k \partial_k V_i + V_k \partial_k u_i + \rho^{-1} \partial_i p' \\ = \nu \partial_k \partial_k u_i - u_k \partial_k u_i + \partial_k \langle u_i u_k \rangle. \end{gathered} \tag{5}$$

For a given cell, it is assumed that $\langle u_i u_k \rangle = \text{const}$, $V_k = \text{const}$, and $\partial_i V_k = \text{const}$. This approximation seems to be first used in the model of random forces [4] (note that the RDT theory uses the Proudman–Batchelor linear approximation of large-scale or medium-scale velocity [3, 5]). Since the macrovariables are assumed to be constant, the last term on the right of (5) vanishes and the third term on the left of (5) vanishes in the coordinate system moving with a velocity $\mathbf{V}$. The dynamics of small-scale turbulence will be described in this coordinate system. Eventually, Eq. (5) takes the form

$$\partial_t u_i + u_j \partial_j V_i + \rho^{-1} \partial_i p' - \nu \partial_j \partial_j u_i + u_j \partial_j u_i = 0. \tag{6}$$

Let us apply Fourier transformation to the small-scale pressure and velocity components

$$\begin{aligned} p(\mathbf{x}, t) &= \Sigma_k p(\mathbf{k}, t)\exp(i\mathbf{k}\cdot\mathbf{x}), \\ u_j(\mathbf{x}, t) &= \Sigma_k u_j(\mathbf{k}, t)\exp(i\mathbf{k}\cdot\mathbf{x}), \end{aligned} \tag{7}$$

where $\mathbf{k} = (2\pi/L)(n_1, n_2, n_3)$; $n_1, n_2, n_3 = \pm 1, \pm 2, \ldots$; and L is the edge length of the cubic cell separating large-scale and small-scale motions.

As a result, Eq. (6) for the Fourier components takes the form

$$\begin{aligned} &\partial_t u_l(\mathbf{k}, t) + u_j(\mathbf{k}, t)\partial_j V_l + ik_l\rho^{-1}p'(\mathbf{k}, t) \\ &+ \nu k^2 u_l(\mathbf{k}, t) + ik_n\Sigma_{\mathbf{q}} u_n(\mathbf{q}, t)u_l(\mathbf{k}-\mathbf{q}, t) = 0, \end{aligned} \tag{8}$$

and Eq. (3) is recast as

$$k_l u_l(\mathbf{k}, t) = 0. \tag{9}$$

Multiplying both sides of (8) by k_l, performing summation over l, and using Eq. (9), we come to

$$\begin{aligned} \rho^{-1}p' &= iu_j\partial_j V_q k_q k^{-2} \\ &- k_m k_l k^{-2}\Sigma_j u_l(\mathbf{j}, t)u_m(\mathbf{k}-\mathbf{j}, t). \end{aligned} \tag{10}$$

Substituting expression (10) into Eq. (8), we obtain an equation for the small-scale Fourier components of the velocity that does not involve small-scale pressure components:

$$\begin{aligned} &\partial_t u_l + u_j P_{lq}\partial_j V_q + \nu k^2 u_l + i[k_n\Sigma_{\mathbf{p}} u_n(\mathbf{p}, t)u_l(\mathbf{k}-\mathbf{p}, t) \\ &- k_l k_m k_j k^{-2}\Sigma_{\mathbf{p}} u_j(\mathbf{p}, t)u_m(\mathbf{k}-\mathbf{p}, t)] = 0, \end{aligned} \tag{11}$$

where $P_{lq} = \delta_{lq} - k_l k_q k^{-2}$.

The nonlinear terms in the brackets can be written in a more elegant form (see, e.g. [6]):

$$\begin{aligned} &\partial_l u_l + u_j P_{lq}\partial_j V_q + \nu k^2 u_l \\ &= -\frac{i}{2}P_{lmn}(\mathbf{k})\Sigma u_m(\mathbf{p}, t)u_n(\mathbf{k}-\mathbf{p}, t), \end{aligned} \tag{12}$$

where $P_{lmn} = k_m P_{ln} + k_n P_{lm}$.

Since the three components of the small-scale velocity are related through incompressibility condition (9), it is convenient to pass to two independent components of the small-scale velocity directed along vectors $\boldsymbol{\varepsilon}^1$ and $\boldsymbol{\varepsilon}^2$ that are orthogonal to each other and to vector $\mathbf{k}$, as was done in the case of isotropic turbulence [7] without large-scale velocity gradients ∂V:

$$\begin{aligned} &(\partial_t + \nu k^2)u^\gamma(\mathbf{k}, t) = a^{\gamma\mu}u^\mu \\ &+ \Sigma_{\mathbf{p}, \mathbf{q}, \mathbf{p}+\mathbf{q}=\mathbf{k}}\Phi^{\gamma\alpha\beta}(\mathbf{k}, \mathbf{p}, \mathbf{q})u^\alpha(\mathbf{p}, t)u^\beta(\mathbf{q}, t). \end{aligned} \tag{13}$$

Here, the superscripts and subscripts run through the values 1, 2 and 1, 2, 3, respectively; $a^{\gamma\mu} = -\varepsilon_j^\gamma\varepsilon_m^\mu\partial_m V_j$; and $\Phi^{\gamma\alpha\beta}(\mathbf{k}, \mathbf{p}, \mathbf{q}) = -ik_m\varepsilon_j^\gamma(\mathbf{k})\varepsilon_j^\alpha(\mathbf{p})\varepsilon_m^\beta(\mathbf{q})$. The vectors appearing in these formulas have Cartesian components expressed through the Euler angles [7]: $\mathbf{k} = (k\cos\theta\cos\eta, k\sin\theta\cos\eta, k\sin\eta)$, where $\cos\theta = k_1/k''$, $\sin\theta = k_2/k''$, $\cos\eta = k''/k$, and $k'' = \sqrt{k_1^2 + k_2^2}$.

When selecting the unit vector $\mathbf{e} = (\cos\theta, \sin\theta, 0)$, we introduce the unit vectors $\varepsilon^1(\mathbf{k}) = \mathbf{e}\times\mathbf{k}/|\mathbf{e}\times\mathbf{k}|$ ($\boldsymbol{\varepsilon}^1 = (\sin\theta, -\cos\theta, 0)$) and $\boldsymbol{\varepsilon}^2 = \mathbf{k}\times\varepsilon^1(\mathbf{k})/|\mathbf{k}\times\boldsymbol{\varepsilon}^1(\mathbf{k})|$ ($\boldsymbol{\varepsilon}^2 = (\cos\theta\sin\eta, \sin\theta\sin\eta, -\cos\eta)$). It is easy to check that $\boldsymbol{\varepsilon}^1\times\boldsymbol{\varepsilon}^2 = \mathbf{k}/k$. The relationship between the normal (subscripts) and polarized (superscripts) Fourier components of the small-scale velocity is [7]

$$u_i(\mathbf{k}, t) = \varepsilon_i^\mu(\mathbf{k})u^\mu(\mathbf{k}, t), \tag{14}$$

with

$$\mathbf{k}\times\boldsymbol{\varepsilon}^\mu = 0, \tag{15}$$

$$\boldsymbol{\varepsilon}^\mu(\mathbf{k})\times\varepsilon^\lambda(\mathbf{k}) = \delta_{\mu\lambda}, \tag{16}$$

$$\varepsilon_i^\mu(\mathbf{k})\varepsilon_j^\mu(\mathbf{k}) = P_{ij}(\mathbf{k}). \tag{17}$$

Making the change of variables $\mathbf{u} = B\mathbf{v}$, where the matrix B is composed of the eigenvectors of the matrix A ($B = (\mathbf{b}_1, \mathbf{b}_2)$ and $\mathbf{A}\mathbf{b}_{1,2} = \lambda_{1,2}\mathbf{b}_{1,2}$), one can bring matrix A to diagonal form, i.e., to matrix J with eigennumbers λ_1 and λ_2 as diagonal elements (in the general case, matrix J is a Jordan matrix). Equation (13) then takes the form

$$\begin{aligned} &(\partial_t + \nu k^2)v^\lambda = J^{\lambda\mu}v^\mu \\ &+ (B^{-1})^{\lambda\gamma}\Sigma_{\mathbf{p}+\mathbf{q}=\mathbf{k}}\Phi^{\gamma\alpha\beta}(B\mathbf{v})^\alpha(\mathbf{p})(B\mathbf{v})^\beta(\mathbf{q}), \end{aligned} \tag{18}$$

where

$$\lambda_{1,2} = -\frac{P}{2} \pm \sqrt{\frac{P^2}{4} - Q}, \tag{19}$$

$$P = \varepsilon_J^1\varepsilon_m^1\partial_m V_j + \varepsilon_j^2\varepsilon_m^2\partial_m V_j, \tag{20}$$

$$Q = (\varepsilon_j^1\varepsilon_m^1\partial_m V_j)(\varepsilon_l^2\varepsilon_p^2\partial_p V_l) - (\varepsilon_j^1\varepsilon_m^2\partial_m V_j)(\varepsilon_j^2\varepsilon_j^1\partial_j V_l). \tag{21}$$

If the variable $\boldsymbol{\partial}\mathbf{V}$ is divided into symmetric and asymmetric parts, $\partial_m V_j = S_{jm} + \boldsymbol{\Omega}_{jm}$, where

$$S_{ij} = \frac{1}{2}(\partial_i V_j + \partial_j V_i), \quad \Omega_{ij} = \frac{1}{2}(\partial_i V_j - \partial_j V_i),$$

one can introduce the unit vector $\mathbf{n}$ with the components $n_j = k_i/k$. Then, using incompressibility condition (4) and the properties of vectors ε^λ, the expressions for P and Q simplify to

$$P = -n_i n_j S_{ij}, \tag{22}$$

$$Q = (\varepsilon_j^1\varepsilon_m^1 S_{mj})(\varepsilon_l^2\varepsilon_p^2 S_{lp}) - (\varepsilon_j^1\varepsilon_m^2 S_{mj})^2 + \frac{1}{4}(\mathbf{n}\cdot\boldsymbol{\Omega})^2, \tag{23}$$

where vorticity vector $\boldsymbol{\Omega}$ is defined in the usual manner as $\boldsymbol{\Omega} = \boldsymbol{\nabla}\times\mathbf{V}$ or $\Omega_i = 2\varepsilon_{ijk}\Omega_{kj}$, where ε_{ijk} is the components of the alternating tensor.

By eliminating auxiliary vectors $\boldsymbol{\varepsilon}^{\lambda}$, one can show that Q is directly dependent on the orientation of vector $\mathbf{k}$ (or $\mathbf{n}$) with respect to tensor $\partial\mathbf{V}$. Expression (23) can be recast as follows:

$$Q = (\varepsilon_m^1\varepsilon_l^2 - \varepsilon_l^1\varepsilon_m^2)\varepsilon_j^1\varepsilon_p^2 S_{mj}S_{lp} + \frac{1}{4}(\mathbf{n}\cdot\mathbf{\Omega})^2. \quad (24)$$

Defining the antisymmetric tensor as $\tau_{ml} = \varepsilon_m^1\varepsilon_l^2 - \varepsilon_l^1\varepsilon_m^2$ and separating out terms other than zero (i.e., at $m \neq l$) from the right of (24), we obtain

$$Q = (\tau_{12}S_{1j}S_{2p} + \tau_{13}S_{1j}S_{3p} + \tau_{23}S_{2j}S_{3p})\tau_{jp} + \frac{1}{4}(\mathbf{n}\cdot\mathbf{\Omega})^2. \quad (25)$$

It is easy to check that $\tau_{12} = n_3$, $\tau_{23} = n_1$, and $\tau_{13} = -n_2$. Therefore,

$$\begin{aligned} Q = {} & n_3^2(S_{11}S_{22} - S_{12}S_{21}) + n_2^2(S_{11}S_{33} - S_{13}S_{31}) \\ & + n_1^2(S_{22}S_{33} - S_{23}S_{32}) + 2n_1n_2(S_{23}S_{13} - S_{12}S_{33}) \\ & + 2n_1n_3(S_{12}S_{23} - S_{22}S_{13}) \\ & + 2n_2n_3(S_{12}S_{13} - S_{11}S_{23}) + \frac{1}{4}(\mathbf{n}\cdot\mathbf{\Omega})^2. \end{aligned}$$

In invariant form, we have

$$Q = -\frac{1}{2}\mathrm{Tr}[(\mathbf{n}\times\mathbf{S})^2] + \frac{1}{4}(\mathbf{n}\cdot\mathbf{\Omega})^2, \quad (26)$$

where the trace of the tensor is $\mathrm{Tr}[(\mathbf{n}\times\mathbf{S})^2] = \varepsilon_{ijk}n_jS_{kp}\varepsilon_{pab}n_aS_{bi}$.

Eventually, we arrive at

$$\lambda_{1,2} = \frac{1}{2}\mathrm{Tr}(\mathbf{CS}) \pm\sqrt{\frac{1}{4}[\mathrm{Tr}(\mathbf{CS})]^2 + \frac{1}{2}\mathrm{Tr}[(\mathbf{n}\times\mathbf{S})^2] - \frac{1}{4}(\mathbf{n}\cdot\mathbf{\Omega})^2}, \quad (27)$$

where $C_{ij} = n_in_j$ and $\mathrm{Tr}(\mathbf{CS}) = C_{ij}S_{ji}$.

Thus, we derived Eq. (18), which approximates the dynamics of small-scale polarization modes for the case when the large-scale dynamic is anisotropic. The linear part of the equation depends significantly on λ_1 and λ_2.

ANALYSIS OF THE RESULTS AND CONCLUSIONS

Equation (18), primarily its linear part, is simpler than the equation used in the RDT theory (see, e.g., [3, 5]), which simplifies its analysis. According to the Haken subordination principle [9] (the special case of the central manifold theorem [10]), the dynamics of nonlinear systems is governed first of all by the most unstable modes (in our case, these are modes corresponding to the directions maximizing λ_1). To this point, the subgrid simulation of turbulence has used the Smagorinsky parametrization, where the coefficient of subgrid turbulent viscosity ν_t depends solely on the invariant of the strain rate tensor $S = \sqrt{2S_{ij}S_{ji}}$; that is, $\nu_t = c_1L^2S$, where c_1 is an empirical constant. However, the use of the vorticity vector magnitude $|\Omega|$ in the expression for ν_t, $\nu_t = c^2L^2|\Omega|$ [11], where c_2 is also an empirical constant, sometimes gives more adequate results. To parametrize the coefficient ν_t in terms of our approach, it is appropriate to take the real part $(\Re(\lambda_1))_{\max}$ of λ_1:

$$\nu_t = c_3L^2(\Re(\lambda_1))_{\max}, \quad (28)$$

where the maximum is taken over all directions of the vector $\mathbf{k}$.

The quantity $(\Re(\lambda_1))_{\max}$ may be termed the principal invariant of strain rate tensor $\partial\mathbf{V}$, since this quantity to a great extent specifies the energy transfer from the large-scale motion to the small-scale motion. Exceptions are regions with high vorticity of the large-scale motion and small values of the components of tensor $\mathbf{S}$ (for these components, the value of $(\Re(\lambda_1))_{\max}$ is small in magnitude). In these regions, we deal with two complex conjugates λ_1 and λ_2. These regions call for special analysis. The consideration of the small-scale motion dynamics will be continued in the next paper.

REFERENCES

1. U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov* (Cambridge University, Cambridge, 1995).
2. W. Frost and T. H. Moulden, *Handbook of Turbulence* (Plenum, New York, 1977; Mir, Moscow, 1980), Vol. 1, Chap. 2.
3. S. Nazarenko, N.-K. R. Kevlahan, and B. Dubrulle, Physica D **139**, 158 (2000).
4. G. E. Skvortsov and L. V. Timokhov, Vestn. Leningr. Gos. Univ., No. 13, 106 (1980).
5. G. K. Batchelor and I. Proudman, Q. J. Mech. Appl. Math. **7**, 83 (1954).
6. W. D. McComb, Rep. Prog. Phys. **58**, 1117 (1995).
7. J. Lee, J. Math. Phys. **16**, 1359 (1975).
8. A. D. Bryuno, *Local Method in Nonlinear Analysis of Differential Equations* (Nauka, Moscow, 1979).
9. H. Haken, *Synergetics: An Introduction* (Springer-Verlag, Berlin, 1977; Mir, Moscow, 1980).
10. A. Davey, R. C. DiPrima, and J. T. Stuart, J. Fluid Mech. **31**, 17 (1968).
11. R. Peyret and T. D. Taylor, *Computational Methods for Fluid Flow* (Springer-Verlag, New York, 1983; Gidrometeoizdat, Leningrad, 1986).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1258–1263. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 40–45.
Original Russian Text Copyright © 2003 by Mukhomorov.

GASES AND LIQUIDS

On the Existence of Electron Periodic Structures in Metal–Ammonia Systems

V. K. Mukhomorov
Agrophysical Research Institute, St. Petersburg, 195220 Russia
e-mail: ivl@agrophys.spb.su

Received January 29, 2003

Abstract—For metal–ammonia systems, the critical temperature and electron concentration at which the spatially uniform temperature and electron distributions may become spatially periodic are found. A dispersion relation for the excitation of the uniformly distributed electron system is derived, and conditions for sustained excitation are found. Tests for instability correlate with conditions where the periodical structures appear. © *2003 MAIK “Nauka/Interperiodica”.*

The existence of polarons and their bound states, bipolarons, in metal–ammonia systems has been confirmed by many studies of magnetic [1, 2], optical [3, 4], and transport [3] properties of related media, as well as by investigating metal–insulator phase transitions [5]. Atoms of dissolved alkali metals are known to dissociate in ammonia so that valence electrons and metal cations are spatially separated. The complete independence of the dynamic properties of metal–ammonia systems on the nature of dissolved metals suggests that the electron states are unrelated to electron–ion interaction. This assumption is supported by experimental data for the conductivity and heat of solution of electrons, as well as by the fact that the optical spectrum remains unchanged when divalent metals are substituted for monovalent ones. The absence of electron–cation interaction is also indicated by the extremely narrow spin resonance line with a g factor of 2.0012 ± 0.0002 [6], which almost coincides with the g factor of a free electron (2.0023).

The electron states in metal–ammonia systems were characterized in terms of the theory of adiabatically and strongly coupled continual polarons [3, 5]. A polar medium is viewed as a dielectric continuum described by macroscopic static and high-frequency permittivities, with the inertial orientation of dipole molecules assumed to be elastic. The strong coupling model, combined with the effective mass method, is self-consistent and macroscopically applicable if $\hbar\omega_0 \ll \hbar\omega_1 \ll \hbar\omega_2$, where $\hbar\omega_1 \approx 0.85$ eV is the energy of the first allowed (in the dipole approximation) polaron optical transition, $\hbar\omega_2 \approx 6$ eV is the excitation energy of electrons of the parent material, and $\hbar\omega_0 \approx 0.025$ eV is the limit energy of long-wave optical phonons.

The possibility of bound two-electron dipole states forming in ammonia has been indicated in [5, 7, 8]. The study of the magnetic properties of metal–ammonia systems has shown that paramagnetism due to the spin properties of electrons weakens with growing concentration of a dissolved alkali metal and when the concentration becomes on the order of 10^{20} cm^{-3}, the system passes into the diamagnetic state. The occurrence of bound electron states also changes other parameters of the electron subsystem: the equivalent conductivity, the Knight shift, optical parameters, etc.

Detailed analysis of the bipolaron structure [9, 10] has demonstrated that the polaron interaction potential is attracting if the dielectric parameters of the medium and the electron–phonon coupling constant meet certain requirements. Using the kinetic equation of self-consistent field for the particle distribution function, Vlasov [11] showed that the space-periodic distribution of interacting particles may be established (under certain conditions) in multiparticle systems where the pair potential has an attracting component. In this work, we find conditions under which the spatially uniform distribution of polarons becomes unstable and may pass into periodic one.

A most effective technique for studying multiparticle systems is the method of equilibrium distribution functions $F_N(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$ that satisfy a chain of Bogoliubov equations [12]. The distribution functions define the probability that a group of N particles occupies positions with radius vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ in the unit volume. In this method, the functions F_N are successively introduced and a set of coupled equations for these functions is found. This set consists of N one-particle and $N(N-1)/2$ two-particle functions. Of the latter, one can construct a set of equations that involves three-particle functions, etc. In this way, a chain of sets of equations can be constructed.

To study the probability of a uniform and isotropic polaron system passing into a periodical structure, we will take the first equation from the chain of coupled integro-differential Bogoliubov equations [12] for asymptotic ($N \longrightarrow \infty$, $V \longrightarrow \infty$) equilibrium distribution functions of polarons,

$$\begin{gathered}\frac{\partial F_1^{(e)}(\mathbf{r}_1)}{\partial r_1^{\alpha}} + \frac{1}{\vartheta}\frac{\partial\varphi(\mathbf{r}_1)}{\partial r_1^{\alpha}}F_1^{(e)}(\mathbf{r}_1)\\ + \frac{1}{\vartheta}\int\frac{\partial}{\partial r_1^{\alpha}}\left(\frac{N}{V}\Phi(|\mathbf{r}_1-\mathbf{r}_2|)\right)F_2^{(e)}(\mathbf{r}_1,\mathbf{r}_2)d\mathbf{r}_2 = 0,\end{gathered}\tag{1a}$$

and metal ions,

$$\begin{gathered}\frac{\partial F_1^{(i)}(\mathbf{r}_1)}{\partial r_1^{\alpha}} + \frac{1}{\vartheta}\frac{\partial\varphi(\mathbf{r}_1)}{\partial r_1^{\alpha}}F_1^{(i)}(\mathbf{r}_1) = 0,\\ \Delta\varphi = -4\pi e(\rho^{(e)}-\rho^{(i)}),\end{gathered}\tag{1b}$$

where

$$\vartheta = k_{\mathrm{B}}T; \quad \alpha = x, y, z;$$

$$\lim_{V\to\infty}\frac{1}{V}\int F_1^{(\mathrm{e,\,i})}(\mathbf{r}_1)d\mathbf{r}_1 = 1;$$

$$\lim_{V\to\infty}\frac{1}{V}\int F_2^{(\mathrm{e,\,i})}(\mathbf{r}_1,\mathbf{r}_2)d\mathbf{r}_2 = F_1^{(\mathrm{e,\,i})}(\mathbf{r}_1).$$

Here,

$$\rho^{(\mathrm{e,\,i})}(\mathbf{r}) = e\frac{N}{V}F_1^{(\mathrm{e,\,i})}(\mathbf{r})$$

is the charge density (superscripts *e* and *i* refer to electrons and ions, respectively), $F_1^{(\mathrm{e,\,i})}(\mathbf{r}_1)$ is the one-particle distribution function, $F_2^{(\mathrm{e,\,i})}(\mathbf{r}_1, \mathbf{r}_2)$ is the binary distribution function, N/V is the polaron concentration, φ is the electrostatic potential (for a system of charged polarons to be stable, its charge must be neutralized by the uniformly distributed positive charge of metal cations; $\rho^{(i)} = \mathrm{const}$), and $\Phi(|\mathbf{r}_1 - \mathbf{r}_2|)$ is the pair potential energy of interaction between polarons at the points $\mathbf{r}_1$ and $\mathbf{r}_2$. The interaction of each pair of particles is central and depends on only the distance between their centers of inertia. Equation (1a) is not closed. To decouple the coupled equations of the chain, we make a simplifying assumption that the binary function may be written in multiplicative form:

$$F_2^{(e)}(\mathbf{r}_1,\mathbf{r}_2) = F_1^{(e)}(\mathbf{r}_1)F_1^{(e)}(\mathbf{r}_2).\tag{2}$$

Multiplicative approximation (2) applies if the particle spacing $|\mathbf{r}_1 - \mathbf{r}_2| \gg R_{\mathrm{p}} = 10a_0^*$ far exceeds the polaron radius. This means that the one-particle functions do not overlap and each of the particles moves in the field generated by all surrounding particles.

With approximation (2), Eq. (1) can be reduced to the N equilibrium equations of the self-consistent field for the one-particle distribution functions:

$$\begin{gathered}\frac{\partial F_1^{(e)}(\mathbf{r}_1)}{\partial r_1^{\alpha}} + \frac{1}{\vartheta}\frac{\partial\varphi(\mathbf{r}_1)}{\partial r_1^{\alpha}}F_1^{(e)}(\mathbf{r}_1)\\ + \frac{1}{\vartheta}\frac{\partial}{\partial r_1^{\alpha}}\left(\int\frac{N}{V}\Phi(|\mathbf{r}_1-\mathbf{r}_2|)F_1^{(e)}(\mathbf{r}_2)d\mathbf{r}_2\right)F_1^{(e)}(\mathbf{r}_1) = 0.\end{gathered}\tag{3}$$

Integrating (3) yields the closed nonlinear Hammerstein integral equation

$$\begin{gathered}\ln[\lambda F_1^{(e)}(\mathbf{r}_1)]\\ + \frac{1}{\vartheta}\left(\frac{N}{V}\int\Phi(|\mathbf{r}_1-\mathbf{r}_2|)F_1^{(e)}(\mathbf{r}_2)d\mathbf{r}_2 + \varphi(\mathbf{r}_1)\right) = 0,\end{gathered}\tag{4}$$

where λ is the constant of integration.

Nonlinear integral equation (4) includes the temperature and concentration of the particles as parameters and represents a one-particle distribution function that takes into account the interaction of a given particle with the remaining ensemble.

A trivial solution to integral equation (4) is that corresponding to the spatially uniform particle distribution ($F_1^{(e)} = \mathrm{const}$). At sufficiently high temperatures T, a solution to (4) that corresponds to the spatially uniform distribution is unique. However, from the general theory of nonlinear integral equations, it follows that a nonlinear integral equation like (4) has other solutions if the parameter T entering into this equation is below its critical value. Such solutions belong to the class of ramified solutions. The method to search for ramified solutions to Hammerstein integral equations was described at length in [13]. This method gives an opportunity to find solution branches based on any exact solution with equation parameters continuously varying.

Let us seek nontrivial solutions to (4) in the vicinity of its trivial solution, which corresponds to the spatially nonuniform particle distribution, and find critical parameters associated with the existence of the solutions. Nontrivial solutions to integral equation (4) will be found by the method of successive approximations. Following [14], we expand F_1, λ, and ϑ into series in the small parameter $(\vartheta - \vartheta_0)^{1/2}$ to linearize Eq. (4):

$$\begin{gathered}F_1^{(e)}\mathbf{r} = \chi_0 + (\vartheta-\vartheta_0)^{1/2}\chi_1(\mathbf{r}) + (\vartheta-\vartheta_0)\chi_2(\mathbf{r}) + \ldots,\\ \lambda = \lambda_0 + (\vartheta-\vartheta_0)^{1/2}\lambda_1 + (\vartheta-\vartheta_0)\lambda_2 + \ldots,\\ \frac{1}{\vartheta} = \frac{1}{\vartheta_0} + \frac{\vartheta-\vartheta_0}{\vartheta_0^2} + \ldots,\\ \varphi(\mathbf{r}) = \varphi_0 + (\vartheta-\vartheta_0)\varphi_1(\mathbf{r}) + \ldots, \quad \vartheta_0 = k_{\mathrm{B}}T_0,\\ \lim_{V\to\infty}\frac{1}{V}\int\chi_0(\mathbf{r})d\mathbf{r} = 1, \quad \lim_{V\to\infty}\frac{1}{V}\int\chi_1(\mathbf{r})d\mathbf{r} = 0.\end{gathered}\tag{5}$$

Here, we consider only small deviations from the spatially uniform distribution: $\chi_0 \gg \chi_1$. The temperature $T_0 = \vartheta_0/k_B$ is the ramification point.

Now we substitute series (5) into integral equation (4). Collecting terms with the same powers of $(\vartheta - \vartheta_0)^{1/2}$, we arrive at the set of integral equations

$$\ln(\lambda_0\chi_0) = -\frac{1}{\vartheta_0}\left(\int\frac{N}{V}\Phi(|\mathbf{r}_1 - \mathbf{r}_2|)d\mathbf{r}_2 + \varphi_0\right), \qquad (6)$$

$$\begin{gathered}\lambda_1\chi_0 + \lambda_0\chi_1(\mathbf{r}_1) \\ = -\frac{\lambda_0}{\vartheta_0}\left(\int\frac{N}{V}\Phi(|\mathbf{r}_1 - \mathbf{r}_2|)\chi_1(\mathbf{r}_2)d\mathbf{r}_2 + \varphi_0\chi_1(\mathbf{r}_1)\right)\ldots.\end{gathered} \qquad (7)$$

Equation (6) corresponds to the spatially uniform particle distribution under the condition of temperature equilibrium and has the exact solution χ_0 = const. Since the system as a whole is electroneutral, the potential $\varphi_0 = 0$. In view of this condition and also the constancy of $F_1^{(i)}$, Eq. (1b) identically goes to zero. Thus, the zeroth-order approximation is characterized by the neutrality of the system and spatially uniform polaron distribution.

Nontrivial solutions to Eq. (7) appear as solutions to the homogeneous linear integral equation

$$\chi_1(\mathbf{r}_1) = -\frac{N}{\vartheta_0 V}\int\Phi(|\mathbf{r}_1 - \mathbf{r}_2|)\chi_1(\mathbf{r}_2)d\mathbf{r}_2. \qquad (8)$$

In general, Eq. (8) has a continuous spectrum of eigenvalues. Therefore, we additionally require that solutions to (8) be periodic and seek them in the form

$$\chi_1(\mathbf{r}) = C\exp(i\mathbf{r}\mathbf{k}). \qquad (9)$$

Substituting (9) into (8), we find the condition where the spatially uniform distribution gives way to periodic solutions:

$$-\frac{N}{\vartheta_0 V}\int\exp(i\mathbf{r}_1\mathbf{k})\Phi(|\mathbf{r} - \mathbf{r}_1|)d\mathbf{r}_1 = 1. \qquad (10)$$

Expression (10) can be written in the form

$$-\frac{N}{\vartheta_0 V}\sigma(k) = 1, \quad \sigma(k) = \frac{4\pi}{k}\int_0^\infty\Phi(r)r\sin(kr)dr. \qquad (11)$$

It is obvious that solutions to (8) exist if the inequality $\sigma(k) < 0$ is met; in other words, the interaction potential must have an attracting component. Hence, ramified solutions other than those corresponding to the spatially uniform distribution arise if the requirement $N|\sigma(k)|/V\vartheta > 1$ is imposed on the temperature, concentration, and interaction forces and, in addition, the condition $d\sigma(k)/dk = 0$ is met.

The interaction potential for two adiabatically coupled continual polarons as a function of the distance between their centers of inertia and macroscopic dielectric properties of a polar medium can analytically be represented as [2, 9]

$$\begin{gathered}\Phi(r) = 2\alpha^2\hbar\omega_0\{(\varepsilon^*/\varepsilon_\infty - 1)a_0^*(1 \\ -\exp(-\delta r))\exp(-\kappa r)/r\} \\ + 2\alpha^2\hbar\omega_0\{[d + cr^2 + (a + b(r - r_0)^2) \\ \times(1 - \exp(-\gamma r))]\exp(-\delta r)\}.\end{gathered} \qquad (12)$$

The parameters of potential (12) for polarons in ammonia are as follows: $a = -5.175 \times 10^{-7}$, $b = 6.885 \times 10^{-3}a_0^{*-2}$, $c = -6.857 \times 10^{-3}a_0^{*-2}$, $d = 0.001$, $\gamma = 0.245a_0^{*-1}$, $\delta = 0.2a_0^{*-1}$, $a_0^* = (\varepsilon^* m/m^*)a_0$ is the effective Bohr radius, $\kappa = (4\pi e^2N/\varepsilon_s k_< T)^{1/2}$ is the inverse Debye screening length for long-range Coulomb interaction, $\alpha = (1/2\varepsilon^*)(\varepsilon^2/\hbar\omega_0)(2m^*\omega_0/\hbar)^{1/2}$ is the dimensionless constant of electron–phonon coupling, $m^* = 1.73m$ is the isotropic effective mass of an electron at the conduction band bottom (it was found by comparing the theoretical maximum of the polaron optical absorption band with its experimental value [9]), N is the equilibrium number of polarons, and $\hbar\omega_0$ is the longitudinal long-wave optical phonon energy. The first term in potential (12) describes screened long-range Coulomb interactions; the second one, short-range attractive forces due to phonon exchange. The limiting frequency ω_0 is related to the elastic displacements of molecular dipoles from their equilibrium positions and may be found from the half-width $W_{1/2}$ of the polaron optical absorption spectrum. The absorption spectrum represents a wide structureless band with a single peak in the IR range [3]. At low temperatures, $W_{1/2} = 2(A_0^1\hbar\omega_0\ln2)^{1/2}$. The energy A_0^1 of polar medium rearrangement refers to the most intense $s \longrightarrow p$ photoinduced transition [15]. From the experimental value of $W_{1/2}$ = 0.46 eV [16], we estimate the limiting frequency as $\omega_0 = 5.5 \times 10^{13}$ s^{-1}, which falls into the interval (5.1–6.8) × 10^{18} s^{-1} of long-wave librational vibrations of ammonia [17]. This value of ω_0 will be used below. The effective dielectric constant is $\varepsilon^{*-1} = \varepsilon_\infty^{-1} - \varepsilon_s^{-1}$, where $\varepsilon_s = 22.7$ is the static permittivity of ammonia and $\varepsilon_\infty^{1/2} = 1.325$ is the refractive index on the plateau between the IR absorption by librational vibrations of ammonia dipole molecules and the fundamental absorption by electrons (in the dispersion curve).

Figure 1 shows the pair polaron interaction potential for polaron equilibrium concentrations $N = 10^{18}$, 10^{19}, and 10^{20} cm^{-3}. Potential (12) is physically realistic. From the translation-invariant theory of adiabatically coupled bipolarons [9, 10], it follows that the polaron interaction potential has a fairly deep well, a “softly” repulsing kernel, and the Coulomb asymptotics at $r \longrightarrow \infty$. Also, it is finite at $r = 0$. The attracting component of the pair potential is associated with exchange by longitudinal optical phonons and also by the effects of quantum-mechanical exchange, which is taken into account by the Fermi statistics.

Figure 2 plots the curve $\sigma(k)$, where k is the wavenumber. The minimum in this curve corresponds to the stable state [11] of the polaron system with an optimal value of the wavenumber $k_{\text{opt}} = 0.091/a_0^*$, which is found from the minimum condition for $\sigma(k)$. Since the shape of the polaron interaction potential depends on the temperature and polaron concentration through κ, one can, using relationships (10) and (11), find a critical temperature (for a given concentration) below which a periodic component appears in the distribution function $F_1^{(e)}(r)$ against the background of the spatially uniform polaron distribution:

$$T \le T_{\text{cr}}^{(1)} = -N\sigma(k)/Vk_{\text{B}}. \qquad (13)$$

At temperatures below the critical temperature $T_{\text{cr}}^{(1)}$, the polaron system exhibits periodic structures with a node spacing (period) $d_{\text{opt}} = 2\pi/k_{\text{opt}} = 69.05a_0^* > R_{\text{p}}$, which exceeds the mean particle spacing. The optimal period d_{opt} obtained validates the neglect of pair correlations and use of approximation (2). Figure 3 plots the concentration dependence of the critical temperature. As the polaron concentration grows, so does the critical temperature of the onset of the periodic component. Under the phase transition line, Eq. (8) admits periodic solutions. Above it ($T > T_{\text{cr}}^{(1)}$), uniform polaron distributions are only possible.

The value of k_{opt} remains unchanged throughout the concentration and critical temperature ranges. Thus, at $T \le T_{\text{cr}}^{(1)}$ and a given particle concentration, condition (13) may be considered as the demarcation line between uniform and periodic solutions. At the same time, from (13) at $k \longrightarrow 0$, one can find the instability condition for the spatially uniform particle distribution

$$T_{\text{cr}}^{(2)} = -\frac{4\pi N}{Vk_{\text{B}}}\int_0^\infty \Phi(r)r^2dr = -\frac{N\sigma(0)}{Vk_{\text{B}}}, \qquad (14)$$

that is, the uniform polaron distribution becomes unstable subject to $\sigma(0) < 0$, which imposes one more condition on the particle interaction force. However, this condition is necessary but not sufficient. From Fig. 2 it is seen that, although k_{opt} exists and $\sigma(k_{\text{opt}}) < 0$ for the top curve ($N = 10^{18}$ cm^{-3}) at 10 K, periodic solutions are absent, as follows from Fig. 3 (they may appear only at temperatures below 10 K).

It is therefore of interest to analyze the stability condition for the uniform distribution by using the time dependence of the particle distribution function with the pair interaction potential given by (12). To do this, we write a nonstationary equation [11] for the kinetic one-particle distribution function $F_1^{(e)}(\mathbf{r}, \mathbf{v}, t)$ that differs from the associated spatially uniform distribution only slightly:

$$\frac{\partial F_1^{(e)}}{\partial t} + \mathbf{v}\nabla_{\mathbf{r}}F_1^{(e)} - \frac{1}{M}\nabla_{\mathbf{v}}F_1^{(e)}\nabla_{\mathbf{r}} \times \left(\int \Phi(|\mathbf{r}-\mathbf{r}'|)F_1^{(e)}(\mathbf{r}', \mathbf{v}', t)d\mathbf{v}'d\mathbf{r}' + \varphi(\mathbf{r})\right) = 0. \qquad (15)$$

Here, α is the electron–phonon coupling constant,

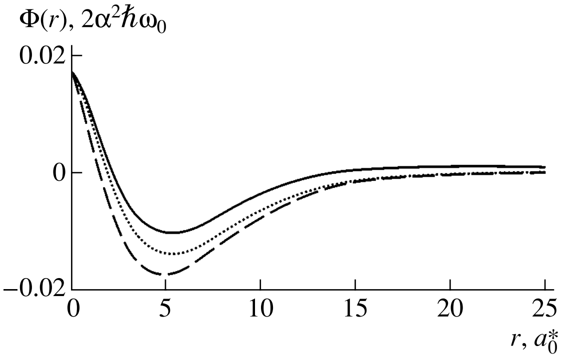


Fig. 1. Pair polaron interaction potential for $N = 10^{18}$ (continuous line), $N = 10^{19}$ (dotted line), and $N = 10^{20}$ cm^{-3} (dashed line). $T = 10$ K.

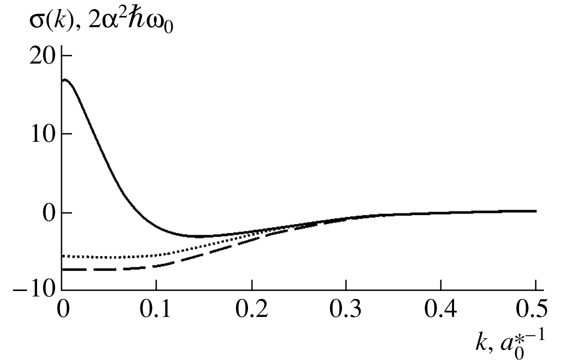


Fig. 2. Fourier transform of the pair polaron interaction potential for $N = 10^{18}$ (continuous line), $N = 10^{19}$ (dotted line), and $N = 10^{20}$ cm^{-3} (dashed line). $T = 10$ K.

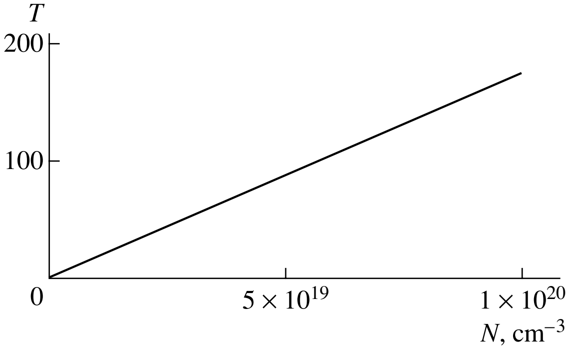


Fig. 3. Concentration dependence of the critical temperature at which the polaron distribution acquires a periodic component.

M is the effective translational mass of a polaron, and $\mathbf{v}$ is the translational velocity of a polaron. For $\alpha \gg 1$, $M = 0.023\alpha^4 m^* \gg m^*$.

Let us determine conditions where spontaneous oscillations in a system of uniformly distributed polarons are sustained and trace a correlation between the conditions for sustained oscillations with a test for appearance of periodic solutions to (8). If the amplitude of a perturbation is small, the distribution function is considered as a perturbation of the stationary spatially uniform distribution $F_0(\mathbf{v})$, which depends only on the velocity. In this case, Eq. (15) can be linearized and brought to the form

$$\frac{\partial f}{\partial t} + \mathbf{v}\nabla_{\mathbf{r}} f - \frac{1}{M}\nabla_{\mathbf{v}} F_0 \nabla_{\mathbf{r}}\int \Phi(|\mathbf{r}-\mathbf{r}'|)\, f(\mathbf{r}', \mathbf{v}', t)\, d\mathbf{v}' d\mathbf{r}' - \frac{1}{M}\nabla_{\mathbf{v}} f(\mathbf{r}, \mathbf{v}, t)\nabla_{\mathbf{r}}\int \Phi(|\mathbf{r}-\mathbf{r}'|) F_0\, d\mathbf{v}' d\mathbf{r}' = 0, \tag{16}$$

where it is assumed that the polaron charge is neutralized, on average, by the charge of positively charged alkaline ions and $\varphi(\mathbf{r}) = \varphi_0$. Since (16) is a linear equation and does not contain coordinates in explicit form, its solution will be sought in the form

$$\exp(i\mathbf{kr} + (i\omega + \gamma)t),$$

where ω is the frequency of sustained acoustic waves in the polaron system and γ is the damping decrement.

Substituting this exponential into Eq. (16) allows one to rewrite it in operator form in the Fourier representation:

$$\left(-iZ - ik\mathbf{r} + \frac{i\mathbf{k}\sigma(k)}{M}\frac{\partial F_0}{\partial \mathbf{v}}\int d\mathbf{v}\right) f_k(\mathbf{v}) = 0. \tag{17}$$

This expression takes into account that the last term in (16) vanishes. Using the results obtained in [18], we find a dispersion relation that relates the frequency ω and the wavenumber k:

$$\frac{k_B T V}{N\sigma(k)} = -1 + 2Z\exp(-Z^2)\int_0^Z \exp(s^2)\,ds, \tag{18}$$

where

$$Z = \frac{1}{k}(\omega + i\gamma)\sqrt{\frac{M}{2k_B T}}$$

and

$$F_0(v) = \frac{N}{V}\left(\frac{M}{2\pi k_B T}\right)^{3/2}\exp\left(-\frac{Mv^2}{2\pi k_B T}\right)$$

is the equilibrium Maxwell distribution.

Solutions to Eq. (18) give the eigenfrequencies of oscillations in the polaron system. When deriving (18), we used the assumption that the velocity and space distributions of the particles are statistically independent, $F(\mathbf{r}, \mathbf{v}) = F(\mathbf{r})F(\mathbf{v})$, and that the distribution function depends on velocity only through the kinetic energy.

Using Eq. (18), one may find the instability condition for the spatially uniform distribution upon propagation of the perturbation. If purely imaginary values of the decrement γ appear in solutions to (18), oscillations in the spatially uniform distribution will build up, causing a nonuniform structure to form. Dispersion relation (18) shows that, at a given polaron concentration, purely imaginary decrements γ appear only if $T \le T_{\text{cr}}^{(1)}$. Figure 4 shows the temperature variation of γ for $N = 10^{20}$ cm^{-3}. As the temperature decreases continuously, γ stepwise takes a purely imaginary value at $T = T_{\text{cr}}^{(1)} = 172$ K and reaches a maximum near 75 K. It is at these values of the parameters that periodic components appear in a solution to Eq. (8). As the frequency ω rises, the purely imaginary value of γ also increases. At the same time, at $T > T_{\text{cr}}^{(1)}$, the damping decrement takes (also stepwise) a real value.

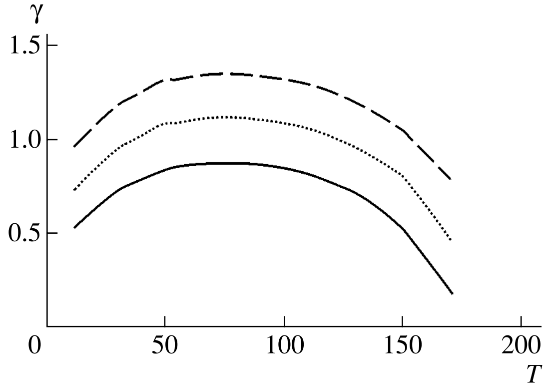


Fig. 4. Temperature dependence of the purely imaginary value of the decrement γ at $\omega = 0$ (continuous line), $\omega = 0.25$ (dotted line), and $\omega = 0.5$ (dashed line).

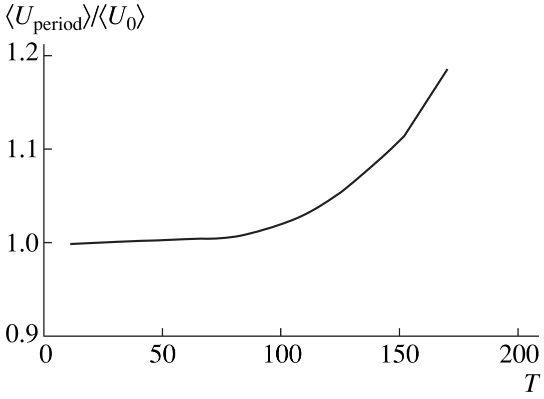


Fig. 5. Temperature dependence of the ratio between the mean potential energies of the periodic and uniform polaron distributions. $N = 10^{20}$ cm^{-3}.

In this case, the perturbation of the spatially uniform distribution decays with time. Thus, the stability boundaries for a polaron system that are obtained from nonlinear integral equation (8) and dispersion relation (18) coincide. For $T > N|\sigma(\kappa_{\text{opt}})|/Vk_B$, the uniform polaron distribution with $\sigma(\kappa_{\text{opt}}) < 0$ is stable against perturbations, while at temperatures and concentrations such that $T < N|\sigma(\kappa_{\text{opt}})|/Vk_B$, the uniform distribution breaks.

Setting the interaction potential $\Phi(|\mathbf{r} - \mathbf{r}'|)$, one may see which type of distribution is stable at parameters given. Let us compare the internal energy of the polaron periodic distribution,

$$\begin{aligned} W_1 &= 3Nk_BT + \frac{1}{2}\int\Phi(|\mathbf{r}-\mathbf{r}'|)\rho_1(\mathbf{r})\rho_1(\mathbf{r}')d\mathbf{r}d\mathbf{r}' \\ &= 3Nk_BT + \langle U_{\text{period}}\rangle, \end{aligned} \tag{19}$$

with that of the uniform distribution,

$$W_0 = 3Nk_BT + \frac{2\pi N^2}{V}\int_0^\infty \Phi(r)r^2dr = 3Nk_BT + \langle U_0\rangle. \tag{20}$$

Here, it is taken into account that the particle concentration is related to the one-particle distribution function as $\rho_1(\mathbf{r}) = NF_1(\mathbf{r})/V$ and to the uniform distribution function as $\rho = N/V$. Figure 5 shows the temperature dependence of the ratio between the mean energies $\langle U_{\text{period}}\rangle/\langle U_0\rangle$, which are involved in (19) and (20). The mean potential energies are negative. Throughout the range of permissible temperatures, the internal temperature of the periodic state of interacting polarons is lower than the internal energy of the uniform state of polarons.

REFERENCES

1. E. Huster, Ann. Phys. (Leipzig) **33**, 477 (1938).
2. V. K. Mukhomorov, Zh. Tekh. Fiz. **67** (8), 1 (1997) [Tech. Phys. **42**, 855 (1997)].
3. J. Thompson, *Electrons in Liquid Ammonia* (Clarendon, Oxford, 1974; Mir, Moscow, 1979).
4. V. K. Mukhomorov, Phys. Status Solidi **231**, 462 (2002); Opt. Spektrosk. **79**, 188 (1995) [Opt. Spectrosc. **79**, 170 (1995)].
5. N. F. Mott, *Metal–Insulator Transformations* (Taylor&Fransis, London, 1974; Nauka, Moscow, 1979).
6. A. Demortier and G. Lepoutre, C. R. Hebd. Seances Acad. Sci. **268**, 453 (1969).
7. R. Catteral and M. C. R. Symons, J. Chem. Soc. A 13 (1966).
8. S. Schettler and G. Lepoutre, J. Phys. Chem. **79**, 2823 (1975).
9. V. K. Mukhomorov, Opt. Spektrosk. **55**, 246 (1983) [Opt. Spectrosc. **55**, 145 (1983)].
10. V. K. Mukhomorov, J. Phys. C **13**, 3633 (2001); Fiz. Tverd. Tela (St. Petersburg) **44**, 232 (2002) [Phys. Solid State **44**, 241 (2002)].
11. A. A. Vlasov, *Theory of Many Particles* (GITTL, Moscow, 1950).
12. N. N. Bogolyubov, *Selected Works on Statistical Physics* (Izd. Mosk. Gos. Univ., Moscow, 1979).
13. N. Nazarov, *Hammerstein Nonlinear Integral Equations* (1941).
14. S. V. Tyablikov, Zh. Éksp. Teor. Fiz. **17**, 386 (1947).
15. V. K. Mukhomorov, Opt. Spektrosk. **68**, 947 (1990) [Opt. Spectrosc. **68**, 553 (1990)].
16. E. Hart and M. Anbar, *The Hydrated Electron* (Wiley, New York, 1970; Atomizdat, Moscow, 1973).
17. A. Anderson and S. H. Walmsley, Rev. Mod. Phys. **9**, 1 (1965).
18. N. N. Zryakov and A. E. Mozol’kov, Vestn. Mosk. Univ., Ser. 3: Fiz., Astron., No. 2, 116 (1970).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1264–1269. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 46–50.
Original Russian Text Copyright © 2003 by Burmakinskiĭ, Rogov.

GAS DISCHARGES, PLASMA

Calculation of the Erosion Profile of a Cathode for Magnetron Ion-Sputtering Systems

I. Yu. Burmakinskiĭ and A. V. Rogov

Russian Research Centre Kurchatov Institute,
pl. Kurchatova 1, Moscow, 123182 Russia
e-mail: alex-rogov@yandex.ru

Received October 2, 2002

Abstract—A method is proposed for calculating the cathode erosion profile in planar magnetron ion-sputtering systems. The method uses integral characteristics of the magnetic and electric fields and does not require detailed information on the spatial characteristics of the discharge plasma. The calculated results are compared with experimental cathode profiles for rectangular and circular planar ion-sputtering systems. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

One of the important characteristics of magnetron ion-sputtering systems (MISSs), which are widely used to deposit thin films, is the utilization ratio of the target material [1, 2]. This parameter is defined as $K_m = (m_0 - m_k)/m_0$, where m_0 is the cathode mass before sputtering and m_k is the final cathode mass, which corresponds to the maximum cathode erosion in a given MISS. This parameter can easily be measured by weighing the cathode. However, the experimental optimization of the parameter K_m for newly created MISSs is a very laborious process, because it requires carrying out a series of long-run experiments with various combinations of the relative positions of magnets and with cathodes of various thicknesses and shapes until their maximum erosion. In this paper, we describe a method for calculating the profile of the erosion region. The method has been successfully applied to circular and rectangular planar MISSs. It allows one to estimate the efficiency of a magnetron ion-sputtering system with respect to material consumption as a function of the magnetic system parameters and the integral MISS characteristics, such as the total current I_d and the discharge voltage U_d.

MODEL

The erosion region on the cathode surface in an MISS is formed due to the sputtering of the cathode by ions (e.g., argon ions) arriving from the discharge region. Preliminary experiments showed that the discharge voltage and the anode position only slightly affect the shape of the erosion region and the position of the erosion maximum. Therefore, we can assume in calculations that the main factor determining the parameters of the erosion region is the magnetic configuration of the sputtering system, i.e., the relative positions and sizes of magnets, as well as the relationship among the their magnetic inductions.

As a rule, the shape of the erosion region on the cathode surface is similar to the shape of the discharge [2]. In view of this fact, we will assume to a first approximation that the ions move from the discharge region toward the cathode along nearly vertical trajectories. Accordingly, we will assume that the electric field lines are also directed vertically in this region.

The ion bombardment results not only in the cathode sputtering, but also in the secondary electron emission from the cathode surface. The secondary emission coefficient depends on the energy and sort of ions and the material, structure, and state of the bombarded surface. For typical targets, sputtering gases, and ion energies, this coefficient does not exceed 0.1 electrons per ion. Although the secondary electrons are emitted from almost the entire cathode area exposed to the ion flow, they arrive at the discharge region primarily from the cathode region F (Fig. 1). In the cathode region where the magnetic field lines are nearly parallel to the cathode surface (the gap between the magnets, region S in Fig. 1), electrons drift in the crossed electric and magnetic fields; i.e., they move across the magnetic field $\mathbf{B}$, describing a cycloid in the (xz) plane. The return point of the cycloid lies practically at the cathode surface, and the height of the cycloid is on the order of two Larmor radii R_L:

$$R_L \propto \frac{E_z}{B_y}\frac{m_e}{eB} \approx \frac{E_z}{B_y}\frac{1}{\omega_{Ce}}, \qquad (1)$$

where m_e and e are the mass and charge of an electron; E_z and B_y are the z and y components of the electric and magnetic fields, respectively; B is the magnetic induction; and ω_{Ce} is the electron cyclotron frequency near the cathode.

We assume that, in region S, the following conditions are satisfied:

$$|\partial E_z/\partial z|R_L \ll |E_z|, \quad |\partial B_y/\partial z|R_L \ll |B_y|. \tag{2}$$

Since an electron possesses some thermal velocity in the y direction, it also drifts along a magnetic field line toward the central magnet or toward the periphery (Fig. 1, region F). The electrons that have reached region F (or originated in this region) can move along the magnetic field lines away from the cathode and gain the energy W_e that is sufficient for ionization. The motion of the Larmor centers of these electrons is a combination of the motion along a magnetic field line with a subsequent reflection at the return point near the cathode and the drift motion in the crossed electric and magnetic fields. The calculations show that the centrifugal, gradient, diamagnetic, and other kinds of drifts [1] are negligible. At the return point, an electron can pass with a 50% probability onto a higher magnetic line (thus increasing the height of its drift trajectory) or a lower magnetic line. Shifting from one magnetic line to another at the return points, an electron after some reflections falls into region D (Fig. 1) and, then, escapes from the system. The electron moving along a magnetic field line that passes through the cathode region D can be considered free. The frequency with which such an electron oscillates along the magnetic field line (bounce frequency) tends to zero, and the height of its drift trajectory is comparable with the system size.

THEORETICAL ANALYSIS

Let us consider the generation of ions in the discharge region at an arbitrary point $A(y, z)$ (in the MISS symmetry plane) with coordinates y and z (Fig. 1). The number of ionization events in unit volume per unit time in the vicinity of this point can be written as

$$N_i(y, z) \approx n_e(y, z)\nu_i(y, z), \tag{3}$$

where $n_e(y, z)$ is the density of ionizing electrons at the point A and $\nu_i(y, z)$ is the ionization frequency.

Assuming the steady ion flow in the discharge region to be aligned with the z axis and integrating expression (3) with respect to z, we find a quantity proportional to the ion flux density $J_i(y)$ at the cathode at a point with the coordinate y:

$$J_i(y) \propto \int_{z_{\text{cat}}}^{z^*} n_e(y, z)\nu_i(y, z)dz. \tag{4}$$

Here, the integral is taken from z_{cat} to a certain coordinate z^*, which is chosen such that the entire discharge region lies within the integration limits.[1]

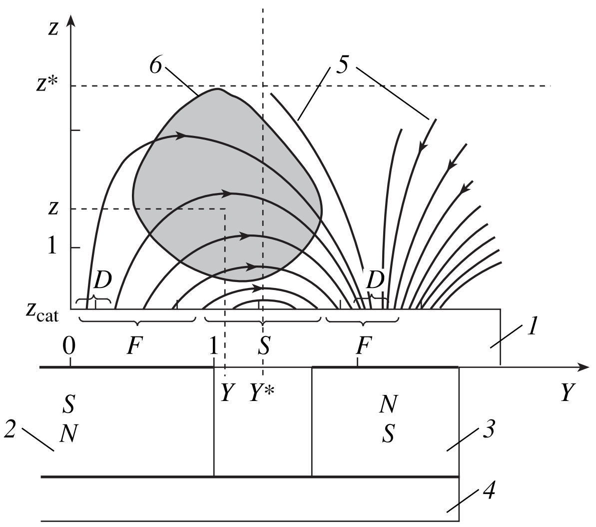


Fig. 1. Arrangement of the cathode and the magnetic system and the configuration of the magnetic field in a MISS: (*1*) cathode, (*2*) central magnet, (*3*) peripheral magnet, (*4*) magnetic core, (*5*) magnetic field lines, and (*6*) discharge region. The lengths are normalized to the width of the peripheral magnet.

The erosion depth $h(y)$ at a point with a coordinate y is proportional to the product of the ion flux density $J_i(y)$ by the energy-averaged material sputtering coefficient $\langle S_y(y)\rangle$:

$$h(y) \propto J_i(y)\langle S_y(y)\rangle \propto \langle S_y(y)\rangle \int_{z_{\text{cat}}}^{z^*} n_e(y, z)\nu_i(y, z)dz. \tag{5}$$

Here, the energy-averaged sputtering coefficient [3] is defined as

$$\langle S_y(y)\rangle = \frac{\dfrac{e}{M_i}\displaystyle\int_0^{\infty} f_i(\varepsilon, y)S_y(\varepsilon)d\varepsilon}{J_i(y)}, \tag{6}$$

where e is the ion charge equal in magnitude to the electron charge (the ions are assumed to be singly charged); M_i is the ion mass; and $f_i(\varepsilon, y)$ is the ion energy distribution function at the cathode, such that the integral over energy gives the total current density,

$$J_i(y) = \frac{e}{M_i}\int_0^{\infty} f_i(\varepsilon, y)d\varepsilon. \tag{7}$$

Assuming that the current density of the sputtering ions depends strongly on y, whereas the energy-averaged sputtering coefficient depends only slightly on it, we will assume this coefficient to be constant. Then,

---

[1] As will be shown below, the choice of the upper limit of integration z^* is rather arbitrary; we can assume this value to be constant and independent of y.

expression (5) can be written in the form

$$h(y) \propto \int_{z_{\text{cat}}}^{z^*} n_e(y,z)\nu_i(y,z)dz. \tag{8}$$

Applying the mean-value theorem to the latter expression, we obtain

$$h(y) \propto \{\nu_i(y)\} \int_{z_{\text{cat}}}^{z^*} n_e(y,z)dz, \tag{9}$$

where $\{\nu_i(y)\}$ is the ionization frequency averaged over the z coordinate.

The integral on the right-hand side of formula (9) is the effective ionization length in terms of the electron density in section y. The plasma density rapidly decreases with distance from the cathode [1, 2]. Therefore, when integrating in formula (9), we can set the upper limit to be equal to infinity and can assume that the integral value is independent of y:

$$\int_{z_{\text{cat}}}^{z^*} n_e(y,z)dz \longrightarrow \int_{z_{\text{cat}}}^{\infty} n_e(y,z)dz = [nl]_e. \tag{10}$$

To calculate the erosion profile, the absolute value of the constant $[nl]_e$ is of minor importance, because, eventually, this value will only enter in the normalization factor. Hence, we have

$$h(y) \propto \{\nu_i(y)\}. \tag{11}$$

The ionization frequency is a function of the total kinetic energy W_e (velocity V_e) of electrons and the working-gas density. The working-gas density may be assumed to be constant over the entire discharge region; hence, the neutral density does not affect the shape of the erosion profile. Assuming the mean value of the function to be close to the function of the mean value of the argument [4], we rewrite expression (11) in the form

$$h(y) \propto \{\sigma_i(W_e(y))V_e(y)\} \approx F(\{V_e(y)\}), \tag{12}$$

where $\sigma_i(y)$ is the ionization cross section of the working gas, $\{V_e(y)\}$ is the average electron velocity in section y, and $F(\{V(y)\})$ is a function of the average electron velocity.

To calculate the average electron velocity, we represent it in the form

$$\{V_e\}^2 = \{V_\parallel\}^2 + \{V_\perp\}^2, \tag{13}$$

where $V_\parallel$ is the longitudinal (with respect to the magnetic field) velocity component, which is determined by the length of the accelerating gap under the discharge, and

$$V_\perp = \frac{\mathbf{E}\times\mathbf{B}}{B^2} \approx \frac{E_z B_y}{B^2} \tag{14}$$

is the transverse velocity component.

We define the average transverse component as

$$\{V_\perp(y)\} \approx \frac{\int_{z_{\text{cat}}}^{z^*} V_\perp(y,z)B^2(y,z)dz}{\int_{z_{\text{cat}}}^{z^*} B^2(y,z)dz} \approx \frac{\int_{z_{\text{cat}}}^{z^*} E_z(y,z)B_y(y,z)dz}{\int_{z_{\text{cat}}}^{z^*} B^2(y,z)dz} \approx [E_z]\frac{\int_{z_{\text{cat}}}^{z^*} B_y(y,z)dz}{\int_{z_{\text{cat}}}^{z^*} B^2(y,z)dz}. \tag{15}$$

The upper integration limit in this expression can be set equal to infinity; then we obtain

$$\{V_\perp(y)\} \approx [E_z]\frac{\int_{z_{\text{cat}}}^{\infty} B_y(y,z)dz}{\int_{z_{\text{cat}}}^{\infty} B^2(y,z)dz}. \tag{16}$$

Here, the average electric field $[E_z]$ is defined as

$$[E_z] = \frac{\int_{z_{\text{cat}}}^{\infty} E_z(y,z)B_y(y,z)dz}{\int_{z_{\text{cat}}}^{\infty} B_y(y,z)dz} = \frac{(-\varphi(y,z)B_y(y,z))_{z_{\text{cat}}}^{\infty} + \int_{z_{\text{cat}}}^{\infty} \varphi(y,z)\frac{\partial B_y(y,z)}{\partial z}dz}{\int_{z_{\text{cat}}}^{\infty} B_y(y,z)dz}, \tag{17}$$

where $\varphi(y, z)$ is the potential at the point with coordinates y and z.

We find the average electric field in the gap between the magnets assuming that it depends only slightly on y. In this gap, the magnetic field component parallel to the cathode ($B_y(z)$) can be approximated to a high accuracy by the dependence

$$B_y(z) = B_{\text{cat}}\exp\{-z/\lambda_\beta\}, \tag{18}$$

where λ_β is a constant, which depends on the relative positions and sizes of the magnets and on the ratio between their magnetizations.

Since the following conditions are satisfied:

$$\begin{aligned} \varphi(y, z_{\mathrm{cat}}) &= U_d; \quad \varphi(y, \infty) \longrightarrow 0, \\ B_y(y, z_{\mathrm{cat}}) &= B_{\mathrm{cat}}; \quad B_y(y, \infty) \longrightarrow 0 \end{aligned} \tag{19}$$

(where U_d is the total discharge voltage and B_{cat} is the magnetic induction on the cathode surface), formula (17) takes the form

$$[E_z] = \frac{U_d B_{\mathrm{cat}} + \langle\varphi\rangle \int_{B_{\mathrm{cat}}}^{0} dB_y}{\int_{z_{\mathrm{cat}}}^{0} B_{\mathrm{cat}} \exp\{-z/\lambda_\beta\} dz} \approx \frac{U_d B_{\mathrm{cat}} - \langle\varphi\rangle B_{\mathrm{cat}}}{\lambda_\beta B_{\mathrm{cat}}}. \tag{20}$$

Assuming in this expression that $\langle\varphi\rangle \approx U_d/2$ [5, 6], we find the average electric field

$$[E_z] \approx \frac{U_d}{2\lambda_\beta}. \tag{21}$$

Hence, the transverse velocity component is equal to

$$\{V_\perp(y)\} \approx \frac{U_d}{2\lambda_\beta} \frac{\int_{z_{\mathrm{cat}}}^{\infty} B_y(y, z) dz}{\int_{z_{\mathrm{cat}}}^{\infty} B^2(y, z) dz}. \tag{22}$$

We represent the total energy as a sum of the energies related to the motion along and across the magnetic field (the longitudinal and transverse energies, respectively)

$$W_e = W_\parallel + W_\perp. \tag{23}$$

The transverse energy varies periodically with time, because the transverse velocity component oscillates harmonically from 0 at the lower point of the cycloid to $2\{V_\perp(y)\}$ at the upper point. Therefore, we should average it over the oscillation period,

$$[W_\perp(y)] \approx \{V_\perp(y)\}^2 m_e. \tag{24}$$

We assume that the longitudinal electron energy is the same throughout the entire discharge region and is close to or slightly higher than the ionization energy of the working gas

$$W_\parallel(y, z) = W_\parallel^0 \geq eJ_p, \tag{25}$$

where J_p is in electronvolts.

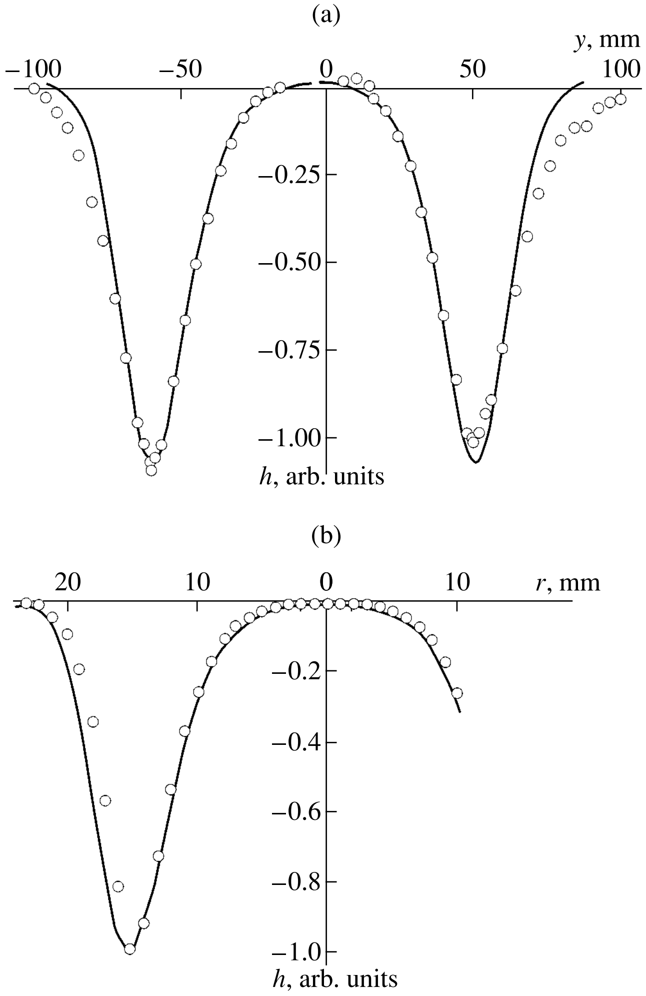


Fig. 2. Comparison of the calculated (curves) and measured (symbols) erosion profiles for a rectangular (a) and a circular (b) magnetic system.

Then, using formulas (22),[2] (24), and (25), we find the average electron energy in section y,

$$[W_e(y)] = W_\parallel^0 + [W_\perp(y)]. \tag{26}$$

To calculate the dependence of the ionization cross section on the electron energy $[W_e(y)]$, we can use the following formula [7]:

$$\sigma_i(W_e) = 2.66\pi a_0^2 \frac{J_n}{J_p} g \frac{(W_e/J_p - 1)}{(W_e/J_p)^2} \ln(W_e/J_p), \tag{27}$$

where energy is in electronvolts, J_p is the ionization energy (15.8 eV for argon), J_n is a constant taken to be 13.6 eV, πa_0^2 is a constant equal to 8.8×10^{-21} m^2, and g is the number of electrons at the upper level ($g = 8$ for argon).

---

[2] The magnetic fields $B_y(y, z)$ and $B(y, z)$, entering formula (22), are to be calculated for a given magnetron magnetic system.

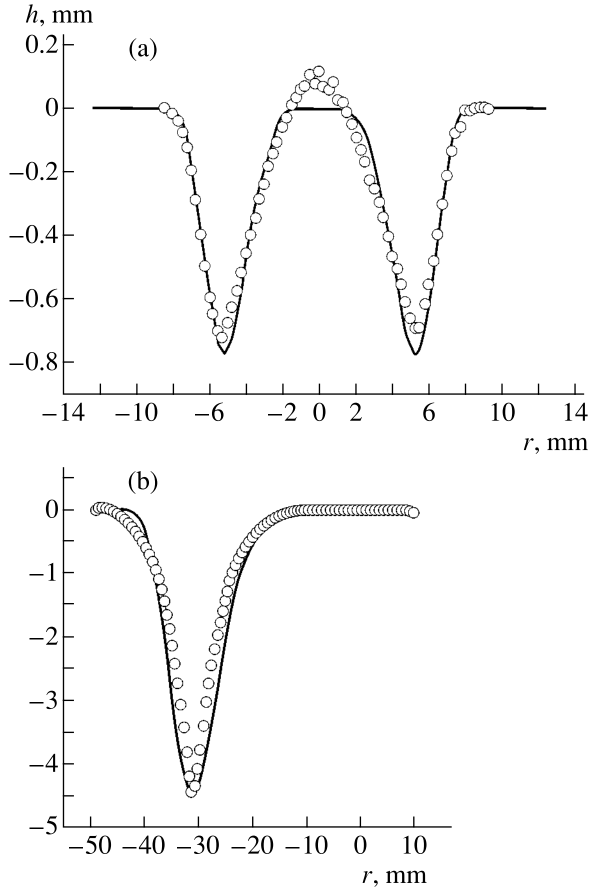


Fig. 3. Comparison of the calculated (curves) and measured (symbols) erosion profiles for two circular magnetrons with cathode diameters of (a) 20 and (b) 100 mm.

The electron velocity expressed through the total kinetic energy (26) is

$$V_e(y) \approx \sqrt{\frac{2[W_e(y)]}{m_e}}. \tag{28}$$

Using formulas (26)–(28) and introducing the normalization coefficient h_{norm} (which allows for the cathode thickness) in expression (12), we find the cathode erosion depth h as a function of y:

$$h(y) \approx h_{\text{norm}}\sigma_i([W_e(y)])V_e(y). \tag{29}$$

When calculating the erosion profile of a circular magnetron, it is sufficient to make the formal substitution $y \longrightarrow r$ in order to find the erosion profile as a function of radius. In numerical calculations with known magnetic fields, the upper integration limit z_{cal} is chosen from the condition

$$\frac{B(z_{\text{cal}})}{B_{\text{cat}}} \ll 1 \tag{30}$$

or

$$\int_{z_{\text{cat}}}^{z_{\text{cal}}} B^2(y,z)dz \approx \int_{z_{\text{cat}}}^{\infty} B^2(y,z)dz,$$
$$\int_{z_{\text{cat}}}^{z_{\text{cal}}} |B_y(y,z)|dz \approx \int_{z_{\text{cat}}}^{\infty} |B_y(y,z)|(dz). \tag{31}$$

The constant λ_β is determined by calculating the magnetic field for a given magnetic configuration.

COMPARISON OF EXPERIMENT AND CALCULATIONS

The cathode erosion profiles were measured in a series of experiments carried out in the Module, Mavr-2, and Mavr-3 MISS devices. The results obtained were compared with the profiles calculated by the method described above. The Module device operated with a rectangular magnetic system, and, in the Mavr-2 and Mavr-3 devices, circular systems were used. In the latter two devices, experiments were carried out with three similar magnetic systems of different size, the cathode diameters being 21, 65, and 110 mm. The measurements and calculations for the rectangular cathode of the Module device were carried out for one of the MISS symmetry planes. A comparison of the results obtained is presented in Figs. 2 and 3.

The experimental and calculated positions of the erosion maximums are found to practically coincide to each other. However, at the periphery, experimental and calculated profiles are somewhat different. This difference may be explained by the defects of the calculation model (in particular, by the assumption of a nearly constant sputtering coefficient over the cathode surface), the assembly defects of the real magnetic system (which are most pronounced in the region above the magnet edge), etc.

CONCLUSIONS

In this paper, we have proposed a rather simple method for calculating the cathode erosion profile in planar MISSs. The method has been successfully used to design the cathode units for MISS devices. The method proposed does not require detailed information on the distributions of the electric potential and plasma density in the discharge and the position of the upper boundary of the discharge. It is only necessary to calculate the magnetic configuration in a given MISS and know the integral characteristics of the discharge, in particular, the discharge voltage U_{d}. The method proposed allows one to choose the optimal magnetic configuration and, taking into account the design features, to find the optimal thickness and shape of the cathode. The implementation of the method requires a computer program that allows one to vary the parameters of the system when calculating the magnetic fields in a given MISS.

The method proposed also makes it possible to find the utilization ratio of the cathode material K_m without performing laborious measurements and to optimize the system with respect to this parameter.

REFERENCES

1. B. S. Danilin and V. K. Syrchin, *Magnetron Sputtering Systems* (Radio i Svyaz’, Moscow, 1982).
2. B. S. Danilin, *Application of Low-Temperature Plasma for Depositing Thin Films* (Énergoatomizdat, Moscow, 1989).
3. M. Mayer, R. Behrish, C. Gowers, P. Andrew, and A. T. Peacock, *Diagnostics for Experimental Thermonuclear Fusion Reactors 2*, Ed. by P. E. Stott *et al.* (Plenum, New York, 1998).
4. L. D. Landau and E. M. Lifshitz, *Mechanics* (Nauka, Moscow, 1982; Pergamon, New York, 1988).
5. V. K. Kalashnikov and V. Kim, Fiz. Plazmy **17**, 1003 (1991) [Sov. J. Plasma Phys. **17**, 584 (1991)].
6. L. Pekker and S. I. Krasheninnikov, Phys. Plasmas **7**, 382 (2000).
7. Yu. P. Raizer, *Gas Discharge Physics* (Nauka, Moscow, 1987; Springer-Verlag, Berlin, 1991).

Translated by E. Satunina

Technical Physics, Vol. 48, No. 10, 2003, pp. 1270–1279. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 51–60.
Original Russian Text Copyright © 2003 by Olevanov, Mankelevich, Rakhimova.

GAS DISCHARGES, PLASMA

Coagulation Rate of Dust Grains in a Low-Temperature Plasma

M. A. Olevanov, Yu. A. Mankelevich, and T. V. Rakhimova

Skobeltsyn Institute of Nuclear Physics, Moscow State University,
Vorob'evy gory, Moscow, 119992 Russia
e-mail: Michael.Olevanov@algodign.com

Received March 19, 2003

Abstract—The mechanisms for the interaction between dust grains in a low-temperature plasma are analyzed theoretically with the aim of describing the grain coagulation process. It is shown that the experimentally observed coagulation process cannot be described by taking into account only electrostatic interaction between the grains. A theoretical model is proposed that describes the interaction between dust grains by accounting for the redistribution of the ion fluxes over the grain surfaces under the action of the electrostatic field of a neighboring grain. The model is employed to analytically calculate the rate constant for the dust grain coagulation. The theory developed is used to explain for the first time the nature of the experimentally observed coagulation threshold and to estimate the critical grain size above which the grains stop growing by the deposition from a gaseous phase and start to coagulate. The applicability of the model proposed to a quantitative description of the coagulation dynamics is discussed. *© 2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

In recent years, rapid technological advances have stimulated increased attention to problems associated with various nanoscale processes and objects. Such objects are of considerable interest because of their importance for both the production of new materials with special properties providing further miniaturization of electronic devices and increasing their productivity and for the study of regular features of the fundamental mechanisms governing the processes that occur on spatial scales intermediate between micro- and macroscales.

Dusty plasma is an example of a system in which nanoparticles are actively produced. It has been known for a long time that, in most devices used in the semiconductor industry, submicron and micron dust grains are produced as a by-product. The formation of dust grains was observed in devices for plasma etching, chemical deposition from a gaseous phase, and deposition by sputtering. Although the processes by which the dust particles grow have been investigated both experimentally [1–7] and theoretically [8–10] for many years, a complete understanding of the driving mechanisms for these processes is still lacking.

Based on the available experimental results, the process of formation and growth of dust particles can be described as follows. The process usually occurs in three steps: the phase of the initial growth of the grains, the coagulation phase, and the saturation phase. During the first phase, the particles grow uniformly to sizes of about 10 nm. The grains thus formed are nearly spherical in shape and have nearly the same radii. During the coagulation phase, the particle growth rate rapidly increases because of an abrupt increase in the probability of collisions between them. The shape of the grains formed in this phase, as well as the topology of their surfaces, can vary over a broad range—from spherical [1, 2] to fractal [3, 4]. In the latter case, the grains are produced by the connected chains of the original crystallites. In the saturation phase, the coagulation process has essentially come to an end and the particles continue to grow very slowly by the material deposition from the gaseous phase. On the whole, the cluster growth time can vary from fractions of a second to several hours, depending on the particular experimental conditions; however, the average growth rate of dust grains is substantially higher than the rate of film deposition from a gaseous phase. This can be illustrated by the data from [4], in which the dust grains were observed to grow at a rate of 100 to 800 nm/s, while the characteristic rate of film deposition is as low as 0.064–0.120 nm/s [5].

Along with the growth of dust grains and the formation of clusters, the self-organization processes that occur in dust–plasma systems are of great interest. The existence of dust crystals, which were first observed in laboratory plasmas [11–13], is an established fact, although the possibility of their formation was discussed theoretically over a fairly long period of time [14, 15]. Dust crystals are formed of dust grains whose sizes may be as large as tens of microns, depending on the particular experimental conditions. The lattice constant of such crystals is usually much longer than the Debye screening length and can reach hundreds of microns. Very often, in both industrial and experimental devices, not only dust crystalline structures were found to form, but also dust–plasma droplets were revealed and gas–liquid phase transitions were observed [16–18]. An understanding of the dynamics of the above processes is of fundamental importance for both plasma physics and the physics of condensed media.

One of the central problems in the physics of dusty plasma is the nature of interaction between the dust grains. Obviously, there are two distinctly different types of the phenomenon under consideration. The phenomena of the first type occur in experiments with dust–plasma crystals and droplets. The dust grains in these structures repel each other, and the structures themselves are confined in electrostatic traps. In such experiments, the dust density is relatively low ($n_d \lesssim 10^7$ cm^{-3}) in comparison with the ion density ($n_i \sim 10^9$ cm^{-3}), the characteristic grain sizes being $a \gtrsim 1$ μm, and there are practically no collisions between the grains. On the other hand, in many experiments, the dust particles are observed to grow rapidly by coagulation. In such systems, the dust density is relatively high ($n_d \gtrsim n_i \sim 10^9$ cm^{-3}) and the particles grow by coagulation to sizes of $a \sim 10$–20 nm. The above examples show that the nature of the interaction between the grains can be radically different, depending on the conditions created in the system. In this paper, we will try to analyze the main reasons why dust–plasma systems exhibit such a variable behavior.

MECHANISMS FOR THE INTERACTION BETWEEN DUST GRAINS IN A PLASMA

The main factor causing a strong interaction between the dust grains in a plasma is the accumulation of an electric charge by the grain. Depending on the experimental conditions, the grains can acquire a charge of one to 10^3–10^4 electron charges. Since the electrons are far more mobile than the ions, the charge acquired by the grains is usually negative. However, some effects like secondary electron emission and photoemission can make dust particles to be positively charged. Let us consider how this factor influences the interaction between the grains. To begin, we assume that the grains interact only through their electrostatic fields. Having made this assumption, we then calculate the rate constant for the dust grain coagulation.

The problem of finding the cross section for collisions between the grains interacting through their given spherically symmetric potentials is solved exactly. The result is

$$\sigma_{dd} = \pi R_{\min}^2\left(1 - \frac{2U_{\text{e.s}}(R_{\min})}{\mu v^2}\right), \qquad (1)$$

where $\mu = m_1 m_2/(m_1 + m_2)$ is the reduced mass, $U_{\text{e.s}}(r)$ is the electrostatic interaction energy, $R_{\min}$ is the shortest distance between the interacting grains, and v is the relative velocity.

For spherical grains of radii a_1 and a_2, we have $R_{\min} = a_1 + a_2$. We also introduce the notation $\sigma_T = \pi(a_1 + a_2)^2$ for the cross section for the interaction between uncharged grains.

Now, we calculate the rate constant for the coagulation of dust grains of different radii:

$$k_{a_1a_2} = \langle \sigma_{a_1a_2}(v)v \rangle. \qquad (2)$$

Here, the averaging is performed over the relative velocities of the grains. We also assume that the colliding grains stick together with a unit probability. Of course, this assumption is actually somewhat idealized. Moreover, the coagulation can be accompanied by a reverse process—a partial destruction of the cluster by a collision with another grain. In our analysis, however, we neglect this possibility and assume that the probability of the destruction of clusters is sufficiently low. We also assume that the grains obey a Maxwellian velocity distribution with a temperature T_0 equal to the temperature of the neutral gas in the chamber. We will consider two cases: the interaction between likely charged grains and the interaction between oppositely charged grains.

In the case of oppositely charged grains, they attract each other, so that we have

$$k_{a_1a_2} = \left(\frac{\mu}{2\pi T_0}\right)^{3/2}\sigma_T \times \int_0^{\infty} v\left(1 - \frac{2U_{\text{e.s}}}{\mu v^2}\right)\exp\left(-\frac{\mu v^2}{2T_0}\right)4\pi v^2 dv. \qquad (3)$$

Integrating over velocities yields

$$k_{a_1a_2} = 2\bar{v}\sigma_T\left(1 - \frac{U_{\text{e.s}}}{T_0}\right), \qquad (4)$$

where $\bar{v}$ is the mean relative velocity of the dust grains and the electrostatic interaction energy $U_{\text{e.s}}$ is negative.

In the case of likely charged grains, they repel each other, so that we can write

$$k_{a_1a_2} = \left(\frac{\mu}{2\pi T_0}\right)^{3/2}\sigma_T \times \int_{v_{\min}}^{\infty} v\left(1 - \frac{2U_{\text{e.s}}}{\mu v^2}\right)\exp\left(-\frac{\mu v^2}{2T_0}\right)4\pi v^2 dv, \qquad (5)$$

where $v_{\min} = \sqrt{2U_{\text{e.s}}/\mu}$ is the minimum relative velocity of the grains at which they can collide.

As in the previous case, we integrate over velocities to obtain

$$k_{a_1a_2} = 2\bar{v}\sigma_T\exp\left(-\frac{U_{\text{e.s}}}{T_0}\right). \qquad (6)$$

Let us analyze the results obtained. We can see that taking into account the electrostatic interaction produces an additional factor in the expression for the coagulation rate constant. In the case of oppositely charged grains, which attract each other, the coagulation rate constant increases linearly with the absolute value of $U_{\rm e.s}$, while, for likely charged grains, this rate constant decreases exponentially. As was noted above, the most typical situation in a dusty plasma is that in which all dust grains are charged negatively. Consequently, the electrostatic interaction should reduce the coagulation rate constant in comparison with that corresponding to a purely thermal motion of the uncharged grains. Note also that the expressions obtained are independent of the functional form of the long-range interaction potential of the grains, in particular, the screening of the electrostatic interaction potential in a plasma.

Now, we estimate how important the difference is between the rate constants for the coagulation of charged and uncharged grains. The electrostatic interaction energy $U_{\rm e.s}$ is given by the expression $U_{\rm e.s}(R_{\rm min}) = Z_{d1}Z_{d2}e^2/(a_1 + a_2)$, where Z_{d1} and Z_{d2} are the charges of the grains. In this expression, we also took into account the fact that, for the shortest distance between the interacting grains, the screening effect of the plasma can be neglected. For estimates, we choose the parameter values corresponding to conditions typical of experiments on coagulation, namely, $n_d \sim n_i \sim 10^9$ cm^{-3} and $T_0 \simeq 0.025$ eV, the grain radius being $a = 10$ nm. The charges of the grains can be estimated from the quasineutrality condition

$$\bar{Z}_d n_d + n_e = Z_i n_i, \qquad (7)$$

where $\bar{Z}_d$ is the mean charge acquired by the grains.

Under the conditions adopted, the electron density is, as a rule, much lower than the ion density, in which case we have $\bar{Z}_d \simeq n_i/n_d$. Substituting all the above values into expression (6) for the coagulation rate constant, we obtain $k/k_0 \sim 3 \times 10^{-3}$, where $k_0 = 2\bar{v}\sigma_T$ is the rate constant for the coagulation of uncharged grains. Therefore, even a slight charge on the dust grains should suppress the coagulation process. However, practical experience shows that this conclusion contradicts the experimentally observed coagulation process: the growth of charged grains is not suppressed but rather occurs at a much faster rate in comparison with that for uncharged grains.

In the literature, several models have been proposed that solve this problem. The so-called ballistic model assumes that dust grains coagulate into clusters due to collisions between charged and uncharged grains [8] or collisions between slightly charged grains [9, 10]. An argument supporting this assumption is based on the fact that, in the initial coagulation stage, the dust density is usually higher than the electron and ion densities in the plasma; hence, according to the quasineutrality condition, most of the grains are uncharged. The ballistic model is capable of explaining such a distinctive feature of the process as the termination of coagulation at a certain time, followed by the saturation phase. The explanation is that, as the grains grow to a certain size, the dust density decreases and becomes lower than the ion density; as a result, almost all of the grains turn out to be charged; as a result, according to the ballistic model, the coagulation process comes to an end. However, this model fails to explain the existence of a certain initial critical grain size below which coagulation does not occur. In addition, it was often pointed out [4, 6] that the experimentally observed high growth rates of the clusters cannot be explained in terms of the collision frequency corresponding to the cross section for the interaction between thermal dust grains. Hence, we arrive at the conclusion that there should exist an additional attractive force between the grains.

An anomalously high coagulation rate can be explained by assuming that some of the grains acquire a positive charge and collisions occur predominantly between the grains carrying charges of opposite sign. The authors of the corresponding model [19, 20] pointed out the following circumstance: since the mean grain charge in the initial coagulation stage is small and fluctuates about a zero value and since the bombardment of the grains by high-energy electrons can give rise to secondary electron emission from the grain surfaces, any of the grains indeed carries a positive charge for a time. However, this hypothesis is unlikely to be productive because it assumes that, in the plasma volume, there should be a sufficient number of high-energy electrons (with energies $\gtrsim$100–300 eV) which are capable of ensuring high secondary-emission yields from the grain surfaces. Nevertheless, practical experience shows that dust clusters grow actively even when such electrons are present in negligible amounts.

It should also be emphasized that it is more correct to consider the evolution of the charge of an individual grain rather than the evolution of the positively charged dust in the plasma volume. The characteristic time of fluctuations of the grain charge is usually much shorter than the time between its collisions with other grains. The grain charging time is determined primarily by the rate at which its surface is bombarded by the electrons; consequently, for estimates, we can write $\tau_e/\tau_d \sim n_d v_d/n_e v_e$, where n_e and n_d are the densities of the electrons and dust grains, respectively, and v_e and v_d are their thermal velocities. At room temperature, the velocity of 10-nm silicon grains is about $v_d \sim 10^2$ cm/s. The velocity of electrons with a temperature of $T_e \sim 3$ eV (which is typical of many experiments) is about $v_e \sim 10^8$ cm/s. Under these conditions, the grain-to-electron velocity ratio is about $v_d/v_e \sim 10^{-5}$–10^{-6}. In the overwhelming majority of cases, the grain-to-electron density ratio n_d/n_e is no larger than 10^2–10^3. As a result, we obtain $\tau_e/\tau_d \ll 1$. This indicates that binary collisions between the grains should be considered in terms of the mean (rather than instantaneous) grain charge.

Hence, we are forced to conclude that the coagulation phenomena cannot be described satisfactorily in terms of only the electrostatic interaction between dust grain. Therefore, the coagulation model should be constructed by taking into account a wider range of processes.

Along with the electrostatic interaction between the charged dust grains, the plasma can give rise to other mechanisms by which the grains interact with each other. In the literature, the interaction models that take into account the fluxes of ions and neutral gas particles to the grain surfaces have been studied for a long time [21]. In the most widely accepted model, effective attraction between the dust grains is assumed to originate from the screening of the plasma fluxes by a neighboring grain. This interaction mechanism was described in detail in [22, 23]. Here, we restrict ourselves to a brief description of the main aspects of the model.

In this model, the dust grains are treated as spherical macroparticles immersed in a plasma. Since the grain surfaces are assumed to be absorbing, each of the grains produces an isotropic plasma flow in which the ions or neutral gas particles move toward its center. If a neighboring grain occurs within the flow created by a test grain, it absorbs the ions or neutral gas particles in the corresponding conical part of the flow to the test grain and, thereby, partially screens it. As a result, some amount of momentum is transferred from the plasma to the test grain. The energy of the interaction between the grains through the screening of the plasma fluxes can be written in the form

$$U(r) = -\frac{3}{4}\pi n T\frac{a_1^2 a_2^2}{r}, \tag{8}$$

where n and T are the density and temperature of the absorbed particles.

Let us consider the main features of the interaction mechanism provided by this model. First, we can readily see that, because of the screening of the fluxes, the grains are subject to an effective attraction with a Coulomb-like potential. Another distinguishing feature of this interaction mechanism is that the action and reaction forces between each pair of grains are equal and opposite. Formally, this feature follows immediately from expression (8); actually, however, it is a consequence of one of the assumptions underlying the model, specifically, the assumption that the plasma flux absorbed by each individual grain is isotropic. We can easily determine the momentum transferred from the plasma to the dust subsystem for the cases when we treat one or two particles. In the first case, the momentum transferred is equal to zero because of the spherical symmetry of the plasma flow to the grain. In the second case, the neighboring grain screens (absorbs) some part of the plasma flow to the test grain. In this case, however, the total momentum transferred to a dust system consisting of two grains does not change because it is independent of which of the grains the momentum has been transferred to. As a result, the total momentum transferred is again zero, which yields the equality of the action and reaction forces. However, the analogy with the Coulomb interaction cannot be pursued further. As was shown by Ignatov [22], the screening mechanism, by its essence, results in the fundamentally non-pairwise nature of the interaction between the grains and thereby cannot be described using a field approach.

According to the above considerations, an important point in the model in question is that the forces should be calculated under the assumption that the plasma flows to the dust grains are isotropic and are directed toward the grain centers. In the case of dust grains carrying electrostatic charges, this assumption generally fails to hold for ions and, as was pointed out in [22], is only partly justified for grains whose sizes are much larger than the Debye screening length, $a \gg \lambda_d$. For dusty plasmas, this situation is very exotic because, usually, $\lambda_p \gtrsim 10$ μm and because the grains having sizes of tens of microns are regarded as being too large to be involved in the coagulation process.

For nanometer dust grains, the approach described above should be somewhat modified by taking into account the fact that the trajectories of charged plasma particles are curved by the electric fields of the grains. Consequently, the additional effect of the plasma fluxes is associated not so much with their screening by a neighboring grain but more with the redistribution of the absorbed ion flux over the grain surface under the action of the electric fields of the nearest neighbors. In this case, under the conditions of dynamic equilibrium between the ionization and ion loss processes in the plasma, the mean number of the ions absorbed by each of the grains remains unchanged.

The interaction model that accounts for the redistribution of plasma fluxes over the grain surfaces was proposed in our earlier paper [24]. The main idea of the model is that, when a test grain is subjected to the electric field of the neighboring grain, it becomes polarized and redistribute the ion fluxes over its surface (see Fig. 1). The resulting asymmetry of the fluxes gives rise to an additional force exerted on the grain in the direction of the electric field. The main distinguishing feature of the model proposed in [24] is that the force exerted by the plasma fluxes on the dust grains turns out to be proportional to the electric field strength, as is the case with the Coulomb repulsive force. The expression derived in that paper for the force exerted by the ion fluxes on the grains makes it possible to calculate the pairwise interactions by parametrically taking into account the effect of the plasma. Under certain conditions in the system, this effect can become dominant and give rise to effective attraction between dust grains.

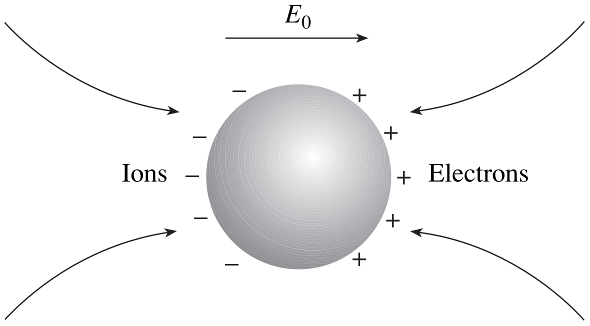


Fig. 1. Redistribution of the plasma fluxes over the surface of a grain in an external electric field.

The model was used to calculate the forces in the approximation in which the electric field produced by the neighboring grains was assumed to be uniform. Hence, the expressions obtained are applicable when the sizes of the grains are much smaller than the mean distance between them, $a \ll n_d^{-1/3}$. In the next section, we will calculate the rate constant for coagulation of the dust grains with allowance for the above-described additional effect of the plasma fluxes.

RATE CONSTANT FOR THE COAGULATION OF DUST GRAINS

In [24], we obtained an expression for the force exerted by an electric field on a dust grain whose surface is affected by an ion flux. In deriving this expression, we assumed for simplicity that all the grains in the system are of the same size. However, this assumption fails to hold for growing dust grains, in which case, in the expression for the interaction force, it is necessary to take into account the distribution of the grains over their sizes. Generalizing the expression derived in [24] to dust grains of different radii yields

$$F_d = Z_d e E_0 \left(\frac{Z_i \tilde{n}_i}{Z_d n_d} \frac{a^2}{\overline{a^2}} \frac{\frac{3}{2} + \frac{n_0 \sigma_0}{n_d \pi \overline{a^2}} \Phi(\xi)}{1 + \frac{n_0 \sigma_0}{n_d \pi \overline{a^2}}} - 1 \right), \qquad (9)$$

where Z_d is the grain charge, E_0 is the electric field strength, Z_i is the ion charge, $\tilde{n}_i$ is the ion density in the vicinity of the grain surface, n_d is the dust density, n_0 is the density of neutral gas particles, δ_0 is the cross section for the interaction of ions with neutral gas particles, a is the grain radius, $\overline{a^2}$ is the mean square radius of the grains, and

$$\Phi(\xi) = \xi^2 (1 - 2\xi^2 + 2\sqrt{\pi}\xi^3 e^{\xi^2} erfc(\xi)), \qquad (10)$$

$$\xi^2 = \frac{Z_i e |\varphi_0|}{T_i} \frac{1}{1 + \frac{n_0 \sigma_0}{n_d \pi \overline{a^2}}}. \qquad (11)$$

In expression (11), T_i is the ion temperature and φ_0 is the electric potential at the grain surface. The potential at the surface of a spherical grain is related to its radius by

$$e|\varphi_0| = \frac{Z_d e^2}{a}. \qquad (12)$$

We consider two dust grains of radii a_1 and a_2 that interact with one another through the forces $\mathbf{f}_{12}$ and $\mathbf{f}_{21}$, where $\mathbf{f}_{12}$ is the force exerted by the second grain on the first grain and $\mathbf{f}_{21}$ is the force exerted by the first grain on the second grain. We describe the force acting on each of the grains by expression (9) in which the electric field is assumed to be produced by the other grain. As a result, we are dealing with a system of two grains the interaction between which is pairwise and is characterized by a spherically symmetric potential. The effect of the plasma in this system can be described parametrically. The task now is to solve the problem of scattering of the grains by one another.

It should be stressed, however, that, although the problem of the interaction of two dust grains in a plasma has much in common with the conventional problem about the pairwise interaction between two particles, there is one essential difference. Specifically, when ion fluxes to the surfaces of the grains are taken into account, Newton's third law for the interaction forces fails to hold; in other words, for grains of different sizes, we have $\mathbf{f}_{12} \neq -\mathbf{f}_{21}$. This conclusion is not surprising because, in the presence of plasma, the system of two grains is open. As a result, in some situations, one grain can "move away" from the other grain (because its size is such that the electrostatic repulsive force exceeds the force exerted by the ion flux), while the other grain "catches up" with the first one. As a consequence, the plasma can transfer a portion of its energy to the grains, thereby heating the dust subsystem. Dust heating effect was actually observed experimentally [25].

We introduce the radius vector $\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$, connecting the first and second grains, where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the position vectors of the first and second grains in the rest frame. The time evolution of the radius vector so defined is described by the equation

$$\ddot{\mathbf{r}} = \frac{\mathbf{f}_{21}}{m_2} - \frac{\mathbf{f}_{12}}{m_1}. \qquad (13)$$

Substituting the expressions for the forces into

Eq. (13), we obtain

$$\ddot{\mathbf{r}} = -\frac{1}{\mu}\left\{\frac{Z_i}{n_d\overline{a^2}}\left(\frac{\tilde{n}_{i1}a_1^2}{Z_{d1}}\frac{m_2}{m_1+m_2}\frac{\dfrac{3}{2}+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}\Phi(\xi_1)}{1+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}} + \frac{\tilde{n}_{i2}a_2^2}{Z_{d2}}\frac{m_1}{m_1+m_2}\frac{\dfrac{3}{2}+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}\Phi(\xi_2)}{1+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}}\right)-1\right\}\mathbf{f}_{\text{e.s}}, \tag{14}$$

where the numerical subscripts denote the parameters of the first and second grains and $\mathbf{f}_{\text{e.s}}$ is the electrostatic repulsive force between two grains in the absence of plasma.

We thus see that the original problem is equivalent to the problem of the motion of a particle in a spherically symmetric field with the interaction potential energy $U(r)$, which is equal to

$$U(r) = -\chi U_{\text{e.s}}(r), \tag{15}$$

where $U_{\text{e.s}}(r)$ is the energy of the electrostatic repulsion between the grains and the parameter χ has the form

$$\chi = \frac{Z_i}{n_d\overline{a^2}}\left(\frac{\tilde{n}_{i1}a_1^2}{Z_{d1}}\frac{m_2}{m_1+m_2}\frac{\dfrac{3}{2}+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}\Phi(\xi_1)}{1+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}} + \frac{\tilde{n}_{i2}a_2^2}{Z_{d2}}\frac{m_1}{m_1+m_2}\frac{\dfrac{3}{2}+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}\Phi(\xi_2)}{1+\dfrac{n_0\sigma_0}{n_d\pi\overline{a^2}}}\right)-1. \tag{16}$$

Now, we reduce expression (16) to a simpler form. For spherical grains, we have m_1, $m_2 \sim a_1^3$, a_2^3. We will also assume that, on the average, the potential at the surfaces of all grains is the same, $\varphi_0 = \text{const}$. This assumption is based on the fact that the grains are immersed in a conducting medium, in which induced currents rapidly reduce the potential differences between them to zero.

The ion densities $\tilde{n}_{i1}$ and $\tilde{n}_{i2}$ in the vicinities of the grains are determined by the grain radii and the potentials at the grain surfaces and may vary considerably depending on the particular plasma conditions. The main factor that influences the spatial distribution of the plasma ions near the grains is that the ions are absorbed by the grain surfaces. This circumstance is important primarily for large grains, because, for $a \gtrsim \lambda_i$ (where λ_i is the ion mean free path), the lifetime of the ions in the vicinity of an absorbing grain is much shorter than that far from the grain, and, accordingly, the ion density near the absorbing grain is much lower. Nevertheless, under the conditions typical of experiments on the growth of dust grains through the deposition from a gaseous phase, the opposite case prevails, namely, that in which the density can be correctly estimated by the Boltzmann exponent:

$$\tilde{n}_i \simeq n_i \exp\left(\frac{Z_i e|\varphi_0|}{T_i}\right), \tag{17}$$

where n_i is the volume-averaged ion density.

However, if we consider conditions such that the sizes of the dust grains are much smaller than the characteristic ion mean free path in the plasma, $a \ll \lambda_i$, then we can see that, most of the time between their collisions with neutral gas particles, the ions move along finite trajectories in the fields of the grains. Consequently, we cannot assume that the ion velocity distribution is equilibrium because, in the vicinities of the grains, the mean kinetic energy of the ions is much higher than that in the surrounding plasma. In the model developed here, this circumstance is taken into account by correcting the ion temperature parameter under the assumption that, on the average, the kinetic energy of the ions increases proportionally to the energy acquired by them in the fields of the grains,

$$T_i \simeq \zeta Z_i e|\varphi_0|, \tag{18}$$

where $\zeta \lesssim 1$ is a dimensionless parameter.

Of course, we were forced to make this assumption (which is, however, justified by the results of our earlier calculations [26] based on the methods of molecular dynamics) because, in the parameter range under consideration, an exact expression for the ion distribution function is unlikely to be obtained analytically.

Now, we consider the expression for ξ. The required estimates will be obtained for conditions typical of experimental devices [3]. We set the neutral gas density and dust density to $n_0 \simeq 3.25 \times 10^{15}$ cm^{-3} (which corresponds to the pressure $P = 13$ Pa) and $n_d \simeq 1.5 \times 10^9$ cm^{-3}, respectively. We also estimate the cross section for the interaction of ions with neutral gas atoms by the value $\sigma_0 \simeq 0.5 \times 10^{-14}$ cm^2, in which case the condition $n_0\sigma_0/n_d\pi\overline{a^2} \gg 1$ is well satisfied up to mean square grain radii of $a \sim 0.8$ μm. For these parameter values, we arrive at the approximate expression

$$\xi^2 \simeq \frac{1}{\zeta}\frac{n_d\pi\overline{a^2}}{n_0\sigma_0}. \tag{19}$$

This indicates that, for the grain sizes in the range 10–100 nm, which is of interest to us, and for dust densities of about $n_d \sim 10^9$ cm^{-3}, the inequality $\xi^2 \ll 1$ is, as a rule, satisfied. We also take into account the fact that, for small ζ values, the function $\Phi(\xi)$ behaves asymptotically as

$$\Phi(\xi) \simeq \xi^2. \tag{20}$$

Then, substituting expressions (17) and (20) into the formula for χ, using expressions (18) and (19), and employing the condition $n_0\sigma_0/n_d\pi\overline{a}^2 \gg 1$, we obtain

$$\chi = \frac{Z_i e\pi n_i}{n_0\sigma_0|\varphi_0|}\frac{a_1a_2(a_1^2+a_2^2)}{a_1^3+a_2^3}\frac{\exp(1/\zeta)}{\zeta} - 1. \tag{21}$$

Now, we determine how the account of the ion flux and electrostatic interaction changes the expression for the coagulation rate constant. Expression (15) shows that, in comparison with the case of purely electrostatic interaction, the problem contains an additional factor that parametrically takes into account the contribution of the plasma to the pairwise interaction. It turns out that this factor, which has been denoted above by χ, plays a governing role in the calculation of the rate constant for the coagulation of dust grains, because it, in fact, determines the nature of the effective interaction between two grains. Specifically, for $\chi > 0$, the grains attract one another, while, for $\chi < 0$, the grains repel. In terms of this factor, expressions (4) and (6), which were obtained above for the coagulation rate constant, can be rewritten as

$$k_{a_1a_2} = 2\bar{v}\sigma_T\left(1 + \frac{U_{\text{e.s}}}{T_0}\chi\right), \quad \chi \ge 0, \tag{22}$$

$$k_{a_1a_2} = 2\bar{v}\sigma_T\exp\left(-\frac{U_{\text{e.s}}}{T_0}|\chi|\right), \quad \chi < 0, \tag{23}$$

where the energy $U_{\text{e.s}}$ corresponds to electrostatic repulsion and has the form

$$U_{\text{e.s}}(a_1 + a_2) = \varphi_0^2\frac{a_1a_2}{(a_1+a_2)}. \tag{24}$$

Now, we consider how the parameter χ depends on the radii of two interacting dust grains. To be specific, we denote by a_1 the radius of the smaller grain, in which case the radius of the larger grain is naturally denoted by a_2. We also denote by α the ratio between the radii, $\alpha = a_2/a_1 \ge 1$. In this notation, the part of expression (21) that includes the grain radii takes the form

$$\frac{a_1a_2(a_1^2+a_2^2)}{a_1^3+a_2^3} = a_1 f(\alpha) = a_1\frac{\alpha(1+\alpha^2)}{1+\alpha^3}. \tag{25}$$

The function $f(\alpha)$ in expression (25) lies in the range $1 \le f(\alpha) \lesssim 1.1$. We thus see that the character (sign) of the interaction between two dust grains of different radii is almost completely determined by the radius of the smaller grain.

Now, we consider the stage preceding coagulation, namely, the growth phase of dust grains, and determine how the coagulation rate constant depends on the radii of the growing grains. The potential at the grain surfaces can be evaluated from expression (12) and quasineutrality condition (7). To do this, we assume that the distribution of grains over their radii is sufficiently narrow. This characteristic feature of the growing dust particles is typical for the initial stage of dust production in a plasma, when the particles grow by the deposition from a gaseous phase. In this case, we can assume that all the grains carry almost the same charges equal to $\bar{Z}_d$. We also introduce the notation $P = n_e/n_i$.

Then, we carry out the necessary manipulations to arrive at the relationship

$$\chi\frac{U_{\text{e.s}}}{T_0} = \left(\frac{Z_in_i(1-P)}{n_d}\right)^2\frac{e^2/2a}{T_0}\left(\frac{a}{a_*} - 1\right), \tag{26}$$

where we have also introduced the notation

$$a_* \simeq \sqrt{\frac{n_0\sigma_0(1-P)}{\pi n_d}}\sqrt{\zeta}\exp\left(-\frac{1}{2\zeta}\right). \tag{27}$$

When the dust function in the volume is high, i.e., when the dust density is comparable with the ion density (as is almost always the case for the variety of the problems under consideration), we have $P \ll 1$, so that this parameter can be neglected in further calculations.

Relationship (26) is the main relationship determining the rate of grain coagulation. If the radius of the dust particles is smaller than a certain critical radius (which we denote by a_*), then the parameter χ is negative and, according to expression (23), the coagulation rate constant decreases exponentially as the ratio $U_{\text{e.s}}/T_0$ increases. However, as the radii of the growing particles exceed the critical radius, the coagulation rate increases very rapidly because the energy of the electrostatic interaction between the grains is much higher than their thermal energy. Figure 2 shows the rate constant $k_{a_1a_2}$ for the coagulation of dust grains in units of the rate constant associated with their thermal collisions. One of the axes is the grain radius, and the other is the ion density. The dependence shown in Fig. 2 was obtained under the assumption that the distribution of the grains over their sizes is monodisperse, which is characteristic of the initial stage of the coagulation process. The calculations were carried out for the above values of the dust density and the density of neutral gas particles. The parameter ζ was set equal to 0.13. For the critical grain radius, we obtained the estimate $a_* \simeq 4.8$ nm. Figure 2 shows that, as the radii of the grains increase above this critical value, the cross section for their interaction rapidly increases. In the next section, we will discuss the results obtained.

DISCUSSION

Let us summarize the main results obtained in this study and make some remarks on their possible applications.

In the previous sections, we showed that the experimentally observed pattern of the coagulation process in a dusty plasma cannot be described adequately by taking into account only the electrostatic interaction between the grains and that the coagulation problem can be solved by accounting for additional forces exerted on the grains by the ion fluxes absorbed by the grain surfaces. When these forces are taken into account, the coagulation rate constant depends not only on the properties of the grains but also on a wide variety of additional factors responsible for the interaction of plasma ions with neutral gas particles and with the dust subsystem. We thus can conclude that, during the coagulation process, the conditions favorable for the growth of the dust particles in a plasma can also change dynamically. This change is associated not only with the accumulation of an electric charge by the dust but also with other effects. As a consequence, the dust component can form in a variety of ways, depending on particular experimental conditions, so that the coagulation process may also proceed in vastly different manners.

Let us consider in more detail the dependence of the coagulation rate constant on the main parameters of the system. As may be seen from Fig. 2, the probability for small grains to coagulate is fairly low. However, after the grains have grown to a certain critical radius, the coagulation rate rapidly increases and may become 200 times higher than that for uncharged grains. Hence, we can conclude that a transition from the phase of the initial growth to the coagulation phase is definitely threshold in nature, which agrees completely with many experimental observations of the phenomenon in question. Note also that, up to the present, there were no theoretical models that took into account this distinguishing feature.

Another distinctive feature of the model proposed here is accounting for the effect of the electrostatic field when describing the coagulation process. Up to now, it was generally believed that the role of the electrostatic interaction is to stop the coagulation process as the grains acquire like electrostatic charges. In the literature, this stage of the process is called the saturation phase. In our model, however, the electrostatic interaction plays an opposite role. The reason is that the forces exerted on the dust grains by ion fluxes are proportional to the electrostatic field strength, as is the case with electrostatic repulsive forces. From Fig. 2, we can distinctly see that the coagulation rate constant rapidly increases as the ion density grows (or, equivalently, as the grains acquire the charge). This conclusion is confirmed experimentally by the fact that the rate of dust grain coagulation is observed as increasing with increasing applied power, which, in turn, determines the ionization rate in the discharge plasma [7]. As for the saturation phase, our model gives a radically different explanation for its onset. However, before proceeding to a more detailed description of the onset of the saturation phase, we need to consider some features of the growth of clusters in a dusty plasma.

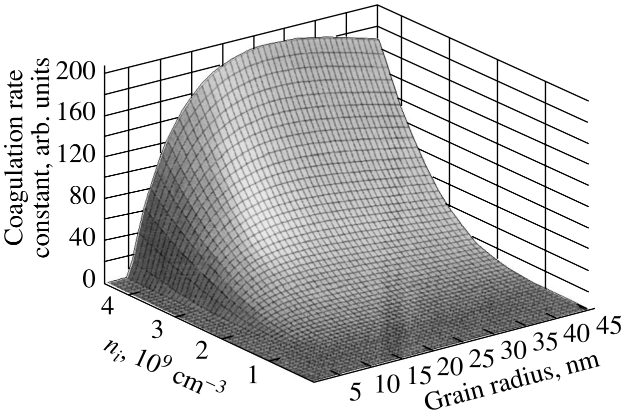


Fig. 2. Rate constant for the coagulation of spherical grains as a function of the grain radius a and ion density n_i. The coagulation rate constant is expressed in units of the rate constant associated with thermal collisions, $k_T = 2\bar{v}\sigma_T$, where $\bar{v}$ is the mean relative velocity of the grains and $\sigma_T = 2\pi a^2$ is the cross section for thermal collisions.

Practical experience shows that the structures of clusters observed in dusty plasmas may be very different. Nascent clusters may possess fairly regular structures resembling that of a compact hard head cabbage or may develop into branched fractal structures. When determining the way in which dust clusters form, it is very important to analyze the properties of the crystallite material that serves as a starting material for the formation of dust structures. It turns out that the higher the conductivity of the material, the larger the degree to which the growing structures are fractal. On the whole, it is this tendency that was observed in a number of experimental papers [16, 17] aimed at investigating the growth of dust structures from different substances (Fig. 3).

The formulas that have been derived above for the coagulation rate constant describe the interaction of two spherical conducting dust grains of arbitrary radii. However, these formulas are inapplicable for calculating the interaction between the clusters formed through the sticking together of dust grains and thus need to be further generalized. Here, in order to understand the main mechanisms for the coagulation-related growth of clusters and its main features, we restrict ourselves to considering merely a qualitative pattern of the coagulation phenomenon.

In order to model the interaction between branched clusters, it is necessary to take into account the following basic feature of their formation: each such cluster consists of a certain number of fractal chains. If a clus-

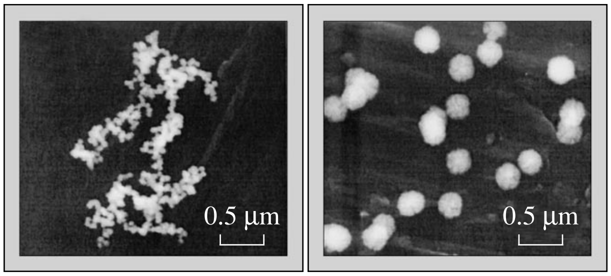


Fig. 3. Dust structures grown in a plasma: a cluster composed of aluminum grains (left) and clusters composed of carbon grains (right). In experiments, the clusters were grown from the products of electrode erosion [16].

ter can be regarded as being conducting, then the charge acquired by it will concentrate at the ends of the chains; as a result, the ion fluxes will exert directed forces only on very limited sites on the cluster surface. Since, in spherical geometry, the drag force exerted by the ion fluxes on a particle is proportional to the particle surface area, we can see that the effective attraction between fractal structures is weaker than that between spherical particles of equivalent size.

For large fractal structures, the number of chains increases far more gradually than the number of macroparticles forming the chains. Consequently, the efficiency of attraction between large clusters decreases sharply as their sizes increase. In this case, the coagulation rate constant is determined completely by electrostatic repulsion. We thus arrive at the conclusion that fractal clusters can only grow by gaining the smallest macroparticles, i.e., the macroparticles that serve as a starting material in the coagulation phase. Taking into account the fact that the force by which macroparticles are attracted to a cluster is proportional to the electric field strength and assuming that the charges in the cluster concentrate at the ends of the chains, we see that the ends will be the most probable places for gaining new particles. It is this result that explains the typical pattern of the formation of such fractal structures, specifically, the fact that, in a growing cluster, the number of chains remains nearly constant while the chains themselves increase rapidly in length.

Turning to the question about the causes for the onset of the saturation phase of the coagulation process, note again that the main source of material for the growing clusters is provided by the smallest crystallites, i.e., by the dust grains that have grown in the initial phase through the deposition from a gaseous phase rather than by the grains that have been produced as a result of coagulation of the smaller grains. Recall also that the coagulation phase can occur only when the original crystallites grow to sizes larger than a certain critical size, in which case the probability of collisions between them increases manyfold. However, throughout the entire stage of the subsequent cluster formation, the sizes of the original crystallites remain close to the critical size. In turn, the critical size is highly sensitive to the global parameters of the system and may become larger or smaller, depending on how the values of these parameters change. It is this circumstance that plays a critical role in the coagulation process and severely restricts the possibility of a continuous growth of the dust particles. The reason is that the higher the electric potential at the surfaces of the dust particles, the larger the size to which they should grow in order to evolve into the coagulation phase. As a result, the parameters of the bulk of the crystallites involved in the formation of clusters occur in the range in which the probability of collisions between them is low. This conclusion also agrees with estimate (27), which was obtained above for the critical grain radius and implies that this radius should increase as the dust density decreases. The above mechanism by which the coagulation process comes to an end may have a number of consequences. In particular, based on the description of this mechanism, we can conclude that, some time after the beginning of coagulation, all clusters fall into two groups, distinguished by their sizes. The first group includes the original crystallites that will not take part in the coagulation process, while the second group consists of large dust agglomerates, which will grow progressively by gaining small clusters of several original crystallites until this source of building material is also exhausted. Such a behavior of the distribution function of the clusters over their sizes is indeed typical of many experiments [1–4].

Hence, we can conclude that the model proposed here for describing the interaction of dust grains in a plasma provides a qualitative explanation of many characteristic features of the coagulation process and can serve as a basis for constructing a quantitative theory of the coagulation phenomenon. The conclusions obtained in this study are confirmed by the results of numerical calculations of the dynamics of the coagulation process and by their direct comparison with the experimental data. At present, these numerical results are being prepared for publication.

ACKNOWLEDGMENTS

We are grateful to A.M. Dykhne, who actively participated in the discussion of the questions concerning the problems addressed here and whose comments on the subject were exceptionally useful for our study. This study was supported in part by the Russian Foundation for Basic Research, project nos. 01-02-17726 and 00-15-96554.

REFERENCES

1. Y. Hayashi and K. Tachibana, Jpn. J. Appl. Phys. **33**, 4208 (1994).
2. A. Garscadden, Pure Appl. Chem. **66**, 1319 (1994).
3. Y. Watanabe, M. Shiratani, H. Kawasaki, *et al.*, J. Vac. Sci. Technol. A **14**, 540 (1996).

4. H. Kawasaki, J. Kida, K. Sakamoto, *et al.*, J. Appl. Phys. **83**, 5665 (1998).
5. T. Fukuzava, S. Kushima, Y. Matsuoka, *et al.*, J. Appl. Phys. **86**, 3543 (1999).
6. A. Bouchoule, L. Boufendi, J. Hermann, *et al.*, Pure Appl. Chem. **68**, 1121 (1996).
7. M. Shiratani, S. Maeda, K. Koga, and Y. Watanbe, Jpn. J. Appl. Phys. **39**, 287 (2000).
8. V. A. Schweigert and I. V. Schweigert, J. Phys. D **29**, 655 (1996).
9. F. Y. Huang, H. H. Hwang, and M. J. Kushner, J. Vac. Sci. Technol. A **14**, 562 (1996).
10. F. Y. Huang and M. J. Kushner, J. Appl. Phys. **81**, 5960 (1997).
11. N. Thomas, G. E. Morfill, V. Demmel, *et al.*, Phys. Rev. Lett. **73**, 652 (1994).
12. J. H. Chu and I. Lin, Phys. Rev. Lett. **72**, 4009 (1994).
13. A. Melzer, T. Trottenberg, and A. Piel, Phys. Lett. A **191**, 301 (1994).
14. E. Wigner, Trans. Faraday Soc. **34**, 678 (1938).
15. H. Ikezi, Phys. Fluids **29**, 1764 (1986).
16. D. Samsonov and J. Goree, J. Vac. Sci. Technol. A **17**, 2835 (1999).
17. D. Samsonov and J. Goree, Phys. Rev. E **59**, 1047 (1999).
18. G. E. Morfill, H. M. Thomas, U. Konopka, *et al.*, Phys. Rev. Lett. **83**, 1598 (1999).
19. K. Watanabe, K. Nishimura, and T. Sato, *Advances in Dusty Plasmas* (World Sci., Singapore, 1998).
20. M. Horanyi and C. K. Goerts, Astrophys. J. **361**, 155 (1990).
21. V. N. Tsytovich, Usp. Fiz. Nauk **167**, 57 (1997) [Phys. Usp. **40**, 53 (1997)].
22. A. M. Ignatov, Fiz. Plazmy **22**, 648 (1996) [Plasma Phys. Rep. **22**, 585 (1996)].
23. A. M. Ignatov, Usp. Fiz. Nauk **171**, 213 (2001) [Phys. Usp. **44**, 199 (2001)].
24. Yu. A. Mankelevich, M. A. Olevanov, and T. V. Rakhimova, Zh. Éksp. Teor. Fiz. **121**, 1288 (2002) [JETP **94**, 1106 (2002)].
25. R. A. Quinn and J. Goree, Phys. Plasmas **7**, 3904 (2000).
26. M. A. Olevanov, Yu. A. Mankelevich, and T. V. Rakhimova, Zh. Éksp. Teor. Fiz. **123**, 503 (2003) [JETP **96**, 444 (2003)].

Translated by G. Shepekina

Technical Physics, Vol. 48, No. 10, 2003, pp. 1280–1288. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 61–69.
Original Russian Text Copyright © 2003 by Yakovenkova, Karkina, Rabovskaya.

SOLIDS

Superdislocation Core Structure in Type I and Type II Pyramidal Planes of Ti_3Al Intermetallic: Glissile Dislocations and Dislocation Barriers

L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya
Institute of Metal Physics, Ural Division, Russian Academy of Sciences,
ul. S. Kovalevskoĭ 18, Yekaterinburg, 620219 Russia
e-mail: yakovenkova@imp.uran.ru

Received December 17, 2002; in final form, February 20, 2003

Abstract—The energies of surface defects in the basal and prismatic planes, as well as in the planes of type I and type II pyramids, are calculated by using *N*-particle interaction potentials for the Ti_3Al intermetallic with the $D0_{19}$ superlattice. The core structure of 2**c** + **a** edge and screw glissile and sessile (barrier-forming) dislocations in pyramidal planes of type I, $\{20\bar{2}1\}$, and type II, $\{\bar{1}\bar{1}21\}$, in Ti_3Al is analyzed. © *2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

Titanium aluminides TiAl, Ti_3Al, Ti_3(Al, Nb), etc. offer a unique set of strength properties at elevated temperatures, good corrosion resistance, and high fatigue properties, which makes them promising for space technology, automotive industry, and other applications. The Ti_3Al intermetallic with the $D0_{19}$ ordered hexagonal superlattice (α_2 phase) is the basic constituent of several single- and two-phase alloys (including those with a lamellar structure). Room-temperature mechanical tests show [1] that polycrystalline Ti_3Al has a high yield stress (500–800 MPa) and an elongation at fracture of 0.3–5.0%. Thus, a disadvantage of Ti_3Al and other titanium aluminides is low-temperature brittleness. As follows from electron microscopy data, the deformation of the α_2 polycrystalline phase occurs largely through the motion of **a** dislocations with the Burgers vector $1/3\langle 11\bar{2}0\rangle$ in the (0001) basal planes and $\{\bar{1}100\}$ prismatic planes.

Recent investigations of single-crystal Ti_3Al [2–5] revealed the strong orientation dependence of the yield stress σ_y and the presence of a complex dislocation ensemble including superdislocations with a variety of Burgers vectors and slip planes. According to experimental conditions (temperature, strain rate, etc.), **a**, **c** + **a**, and 2**c** + **a** superdislocations may glide in the basal, prismatic, or pyramidal planes. It was found [2–5] that the yield strength at room temperature equals 1000–1300 MPa (the ductility is $\varepsilon < 10\%$) for deformation axes close to the [0001] *c* direction. For other directions of deformation, $\sigma_y = 40$–250 MPa (the ductility is $\varepsilon \geq 20\%$). It was shown that, in the former case, deformation is due to the glide of 2**c** + **a** superdislocations with the Burgers vector $1/3\langle 11\bar{2}6\rangle$ in the type I, $\{20\bar{2}1\}$, and type II, $\{11\bar{2}1\}$, pyramidal planes. In the latter case, where the yield stress is much lower, deformation is accomplished through the glide of **a** superdislocations in the basal and prismatic planes, as in Ti_3Al polycrystals.

These two cases also differ in the temperature dependence of the yield stress. For the basal and prismatic slip planes, σ_y has a normal slowly decreasing temperature dependence. In the temperature range studied, the basal-slip ductility is somewhat lower than in the case of prismatic slip. For the pyramidal planes, the temperature dependence of the yield strength exhibits peaks at $T \approx 500°C$ [4] and ≈850°C [2, 3]. In the range from room temperature to the temperature of the peak, anomalous rise in the yield stress is observed, which is typical of many intermetallics with the $L1_2$, $L1_0$, and $D0_{19}$ superlattices.

From electron microscopy data [6], it follows that **a** superdislocations in the basal and prismatic planes dissociate into two superpartial dislocations $1/6\langle 2\bar{1}\bar{1}0\rangle$ with an antiphase boundary (APB) in between. However, in experiments made *in situ*, **a** superdislocations in basal planes dissociated with the formation of a superlattice intrinsic stacking fault (SISF). It was also noted [7] that **a**/2 superpartial dislocations may dissociate into two partial ones; however, the Burgers vectors of the partial dislocations were not identified.

Weak-beam electron microscopy studies performed in [2] showed that 2**c** + **a** superdislocations in Ti_3Al dissociate into two superpartial dislocations in the $(2\bar{1}\bar{1}\bar{1})$ plane, which are spaced at ≤16 nm. In Ti_3(Al, Nb), this

1063-7842/03/4810-1280$24.00 © 2003 MAIK “Nauka/Interperiodica”

spacing is about 50 nm. The related APB energies in type II pyramidal planes are $\xi_{(2\bar{1}\bar{1}\bar{1})} \approx 240$ and 74 mJ/m^2, respectively. Unfortunately, detailed experimental data for the fine structure of the core of **c** + **a**/2 superdislocations in $\{20\bar{2}1\}$ and $\{\bar{2}111\}$ planes are lacking. Therefore, a mechanism of dislocation transformations that are responsible for the deformation behavior of single-crystal Ti_3Al in pyramidal slip planes remains unclear. Computer-aided simulation of dislocation motion in various slip planes makes it possible to study the dislocation core structure and find lowest energy dislocation configurations. We performed such simulation [8–11] for **a** superdislocations in the basal and prismatic planes by the molecular dynamics method.

In this work, we determine the core structure and energy characteristics of a $1/6\langle 2\bar{1}\bar{1}6\rangle$ superpartial dislocation in the pyramidal planes of types I and II by the molecular dynamics method using *N*-particle interaction potentials found for Ti_3Al with the embedded atom technique [12, 13]. We analyze γ surfaces and isenergic contoured maps [12–14], from which one can infer the presence of stable surface defects (APBs and stacking faults) and reveal possible dissociations of a 2**c** + **a** superdislocation in pyramidal planes. Here, we study the core structure of screw and edge 2**c** + **a** superdislocations.

RESULTS OF COMPUTER-AIDED SIMULATION

The computer simulation of the dislocation core in various structures demonstrates that final results may depend on the initial configuration. If the Burgers vector of a superpartial dislocation is large, the dissociation of this dislocation into two or more partial dislocations is energetically favorable (the greater the Burgers vector and the lower the stacking fault (SF) energy, the higher the energy gain upon dissociation).

The choice of the initial configuration depends on the shape of the γ surface for a given plane. To construct this surface, one should cut a model crystallite along a crystallographic plane selected and displace the upper part relative to the lower one by a vector **f** that is parallel to the cut plane. The energy difference between the parts of the crystallite with and without a planar defect per unit surface area is the energy corresponding to the vector **f** (the energy of a generalized stacking fault). If the γ surface has a deep local minimum corresponding to a low-energy stacking fault, the Burgers vector coincides, as a rule, with the displacement vector of this defect. If local minima on the γ surface are absent, it is necessary to consider several initial configurations that differ by the number and Burgers vector of partial dislocations, as well as by the position of their axes, in order to find the core structure that has a minimal energy after relaxation.

The core structure of **c** + **a**/2 screw (the axis $[\bar{1}\bar{1}26]$) and edge (the axes $[\bar{1}\bar{1}20]$ and $[\bar{1}100]$) superpartial dislocations with the Burgers vector $1/6[\bar{1}\bar{1}26]$ in the type I and type II pyramidal planes were studied. The choice of such orientations was dictated by electron microscopy data for the dislocation structure of strained single-crystal Ti_3Al. In [3], 2**c** + **a** edge dislocation bands were observed in strained Ti_3Al single crystals of [0001] **c** orientation at temperatures ranging between from the room value to ≈700°C. In [15–17], both edge and screw 2**c** + **a** superdislocations in $\{11\bar{2}1\}$ and $\{20\bar{2}1\}$ pyramidal planes were observed when the deformation axis was deflected from the exact **c** orientation through angles within ≈30°. The feature of the $1/6[\bar{1}\bar{1}26]$ orientation of the superpartial dislocations taken as model objects in this work is that the axis $[\bar{1}\bar{1}26]$ of the screw dislocation belongs to the type II pyramidal plane $(11\bar{2}1)$, type I pyramidal plane $(20\bar{2}1)$, and the prismatic plane $(\bar{1}100)$; for the edge dislocation, the axis $[11\bar{2}1]$ belongs to the type I pyramidal plane $(20\bar{2}1)$ and basal plane (0001) and the axis $[\bar{1}100]$ belongs to the type II pyramidal plane $(11\bar{2}1)$ and basal plane (0001) (Fig. 1). Thus, as the initial con-

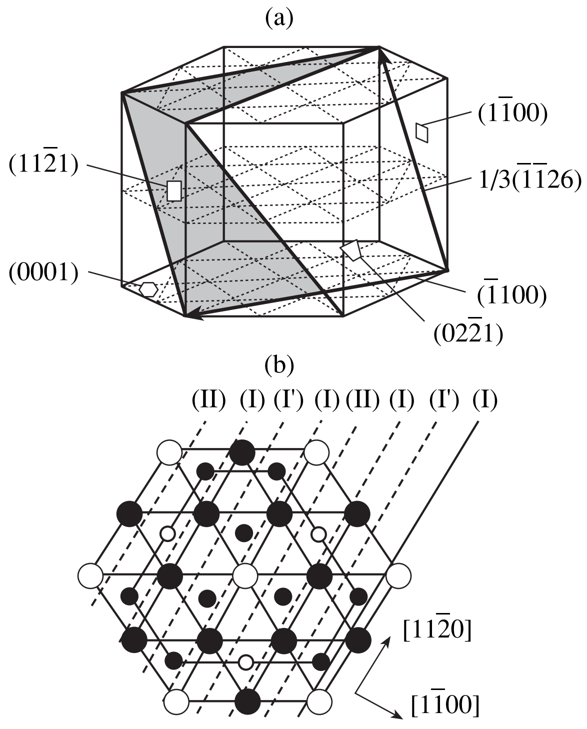


Fig. 1. (a) Unit cell and (b) its projection onto the basal plane for the $D0_{19}$ superlattice. Large atoms (circles) lie in the basal plane; small ones are the projections of atoms located at a distance of **c**/2 onto the basal plane.

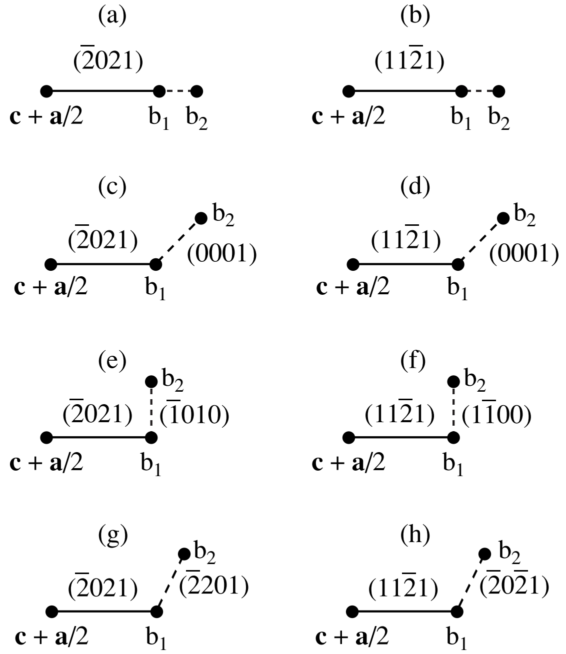


Fig. 2. Dissociation of the (2**c** + **a**) superdislocation with the Burgers vector 1/3[$\bar{1}\bar{1}26$] in pyramidal planes of type I, ($20\bar{2}1$), and type II, ($11\bar{2}1$). (a, b) Glissile configurations and (c–h) dislocation barriers formed when superpartials are split simultaneously in (c, d) pyramidal and basal planes, (e, f) pyramidal and prismatic planes, and (g, h) planes of different pyramids. — APB; - - - SF.

figuration of the superpartial dislocation upon the simulation, one must consider its dissociation in the initial pyramidal plane (which contains the APB) and also in the intersecting pyramidal, prismatic, and basal planes. Figure 2 schematically shows the dissociation of the 2**c** + **a** superdislocation in the type I and type II pyramidal planes. In all the configurations, the superdislocation is split into two superpartial **c** + **a**/2 dislocations with the Burgers vector 1/6[$\bar{1}\bar{1}26$], which bound the APB.

Upon the simulation, one of the superpartial dislocations was placed at the center of the crystallite, while the other (marked by **c** + **a**/2) was at a distance calculated in terms of the elasticity theory for the interatomic interaction potential chosen and had an effect on the crystallite through its displacement fields. In the configuration shown in Figs. 2a and 2b, the superpartial dislocation inside the crystallite is split into two partials with the Burgers vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ lying in the same pyramidal plane. The magnitudes of $\mathbf{b}_1$ and $\mathbf{b}_2$ depend on the shape of the γ surfaces in the ($20\bar{2}1$) and ($11\bar{2}1$) pyramidal planes. The dissociation of the edge superpartial dislocation simultaneously in the pyramidal and basal planes is shown in Figs. 2c and 2d; the dissociation of the screw superpartial dislocation in the pyramidal and prismatic planes, in Figs. 2e and 2f; and the dissociation of the screw dislocation in the intersecting type I and type II pyramidal planes, in Figs. 2g and 2h. The magnitudes of the Burgers vectors $\mathbf{b}_1$ and $\mathbf{b}_2$ depend on the shape of the γ surface in the pyramidal, prismatic, and basal planes. For the configuration in Figs. 2a and 2b, all the partial dislocations and planar defect bands belong to one pyramidal plane; therefore, such configurations are gliding. For the configurations in Figs. 2c–2h, the partial dislocations and planar defect bands belong to two intersecting planes; therefore, such dislocations cannot glide readily in the pyramidal planes and pile up, forming barriers. Which of the configurations is energetically favorable depends on the energy of interaction between the partial dislocations and planar defects incorporated into the superdislocations. The most favorable configuration is found during the relaxation procedure in computer-aided simulation.

1. γ Surfaces in the Basal, Prismatic, and Pyramidal Planes: The Choice of the Initial Configuration

Figures 3a and 3b depict the isenergic lines on the γ surfaces in the type I and type II pyramidal planes [12–14]. In the type I pyramidal plane ($20\bar{2}1$), there exist two different sections, for either of which the dependence of the surface defect energy on the displacement vector in this plane has been obtained. For the section that corresponds to closely spaced {$20\bar{2}1$} planes, the surface defect energies are much higher than for the other. This means that the elastic energy of the superpartial dislocation in this section is also high; therefore, we will consider the dissociation of the dislocation in only the section corresponding to widely spaced {$20\bar{2}1$} planes. The surface defect energy vs. displacement vector in this section is plotted in Fig. 3a. Here, the vector **OA** corresponds to the APB with the vector displacement **a**/2; the vector **OB**, to the APB with the displacement **c** + **a**/2. The respective APB energies are $\xi_{(20\bar{2}1)}(\mathbf{a}/2) = 78$ mJ/m^2 and $\xi_{(20\bar{2}1)}(\mathbf{c} + \mathbf{a}/2) = 205$ mJ/m^2. Along with the minima corresponding to the APBs, the γ surface has one more minimum (the vector **OC** in Fig. 3a) that is associated with the generalized stacking fault with the displacement vector $\mathbf{b}_1 = \mathbf{OC} = (1 - \alpha)/12[1\bar{2}10] + \beta/4[\bar{1}014]$, where $\alpha = 0$ and $\beta = 0.475$. The hatched region near point **C** in Fig. 3a is the domain of displacements on the γ surface where the stacking fault energy is low. This domain covers the Burgers vectors of partial dislocations into which the **c** + **a**/2 superpartial dislocation may dissociate. Point **E**, which belongs to the hatched region, determines the symmetric dissociation of the **c** + **a**/2 superpartial dislocation, based on which we calculated the activation energy for the formation of dislocation barriers [18]. Thus, as the initial configuration of the superpartial dislocation in the type I pyramidal plane, we take the dissociation by the reaction

$$1/6[\bar{1}\bar{1}26] = \mathbf{b}_1 + \text{SF} + \mathbf{b}_2, \qquad (1)$$

where

$$\mathbf{b}_1 = \frac{1-\alpha}{12}[1\bar{2}10] + \frac{\beta}{4}[\bar{1}014],$$

$$\mathbf{b}_2 = \frac{\alpha}{12}[1\bar{2}10] + \frac{1-\beta}{4}[\bar{1}014].$$

The coefficients α and β take the values $\alpha = 0$, $\beta = 0.475$ and $\alpha = 0.5$, $\beta = 0.5$.

In Fig. 3b, the only minimum of the γ surface in the type II pyramidal plane $(11\bar{2}1)$ corresponds to the APB with the displacement vector **c** + **a**/2 (the vector **AB**). The APB energy is $\xi_{(11\bar{2}1)} = 228$ mJ/m^2. In this plane, only the metastable dissociation of the **c** + **a**/2 dislocation into partial ones is possible. The lowest SF energies are obtained for displacement vectors lying along the $[\bar{1}\bar{1}26]$ direction from point **A** to point **C** (Fig. 3b). This suggests that the dissociation into partial dislocations with a Burgers vector parallel to the initial **c** + **a**/2 superpartial dislocation is the most plausible; however, the type of dissociation cannot be established with certainty. The number *n* of partial dislocations varied from two to ten. For $n = 10$, the dislocation spacing was on the order of the dislocation core radius. Varying *n* from two to ten, we, in essence, effect the transition from the discrete to continuous distribution of partial dislocations. Thus, as the initial configuration of the superpartial dislocation in the type II pyramidal plane, we consider three versions of dissociation by the reactions

$$1/6[\bar{1}\bar{1}26] = 1/6(0.5[\bar{1}\bar{1}26] + \text{SF} + 0.5[\bar{1}\bar{1}26]), \quad (2)$$

$$1/6[\bar{1}\bar{1}26] = 1/6(0.4[\bar{1}\bar{1}26] + \text{SF} + 0.6[\bar{1}\bar{1}26]), \quad (3)$$

$$1/6[\bar{1}\bar{1}26] = 1/6(0.1[\bar{1}\bar{1}26] + \text{SF} + \ldots + 0.1[\bar{1}\bar{1}26]). \quad (4)$$

Reaction (2) describes the dissociation into partial dislocations with equal Burgers vectors: $\mathbf{b}_1 = \mathbf{b}_2 = 1/12[\bar{1}\bar{1}26]$ (Fig. 2a); reaction (3), the dissociation into partial dislocations with $\mathbf{b}_1 = \mathbf{AC}$ and $\mathbf{b}_2 = \mathbf{CB}$ (Fig. 3b), where vector **AC** corresponds to the minimal energy of the generalized SF along the $[\bar{1}\bar{1}26]$ direction; and reaction (4), the continuous distribution of partial dislocations.

The γ surface in the basal plane is similar to those found earlier [16] and is characterized by three local minima corresponding to the APB, complex SF (CSF), and SISFs that are associated with the displacement vectors $1/6[\bar{2}110]$, $1/6[0\bar{1}10]$, and $1/3[\bar{1}010]$. The energies of these defects are, respectively, 161, 197, and 78 mJ/m^2. The three local minima on the γ surface define three possible dissociation scenarios for the superpartial dislocation:

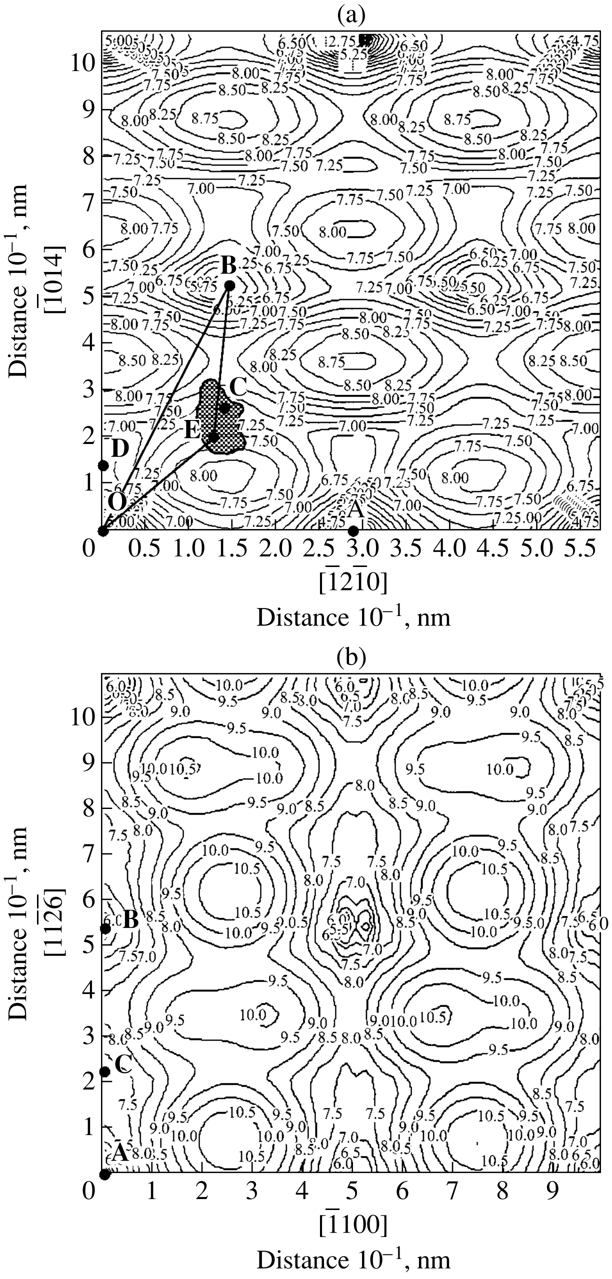


Fig. 3. Isenergic lines on the γ surfaces in Ti_3Al for pyramidal planes of (a) type I, $(20\bar{2}1)$, and (b) type II, $(11\bar{2}1)$.

$$1/6[\bar{1}\bar{1}26] = \mathbf{b}_1 + \mathbf{b}_2 = [0001] + \text{APB} + 1/6[\bar{1}\bar{1}20], \quad (5)$$

$$1/6[\bar{1}\bar{1}26] = \mathbf{b}_1 + \mathbf{b}_2 = 1/6[\bar{1}016] + \text{CSF} + 1/6[0\bar{1}10], \quad (6)$$

$$1/6[\bar{1}\bar{1}26] = \mathbf{b}_1 + \mathbf{b}_2 = 1/6[\bar{1}\bar{1}06] + \mathrm{SISF} + 1/3[\bar{1}010]. \quad (7)$$

Reaction (7) is energetically unfavorable, since the partial dislocations with $\mathbf{b}_1$ and $\mathbf{b}_2$ are attracted together; reaction (5) is energetically neutral; and reaction (6) is energetically favorable. Therefore, for the basal-plane dissociation, $\mathbf{b}_2 = 1/6[\bar{1}\bar{1}20]$ or $\mathbf{b}_2 = 1/6[0\bar{1}10]$ in the initial configuration.

Superlattice $D0_{19}$ may be obtained by translating four prismatic planes of type II, I, I', and I, which differ by arrangement of atoms of two sorts and are variously spaced, in the direction $\langle 01\bar{1}0 \rangle$ (Fig. 1b). It was shown [12] that the APB energy in the type II prismatic plane (section II in Fig. 1b) is 318 mJ/m^2, whereas in the type I prismatic planes (sections I and I' in Fig. 1b), this energy equals 6 mJ/m^2. For the three sections I, I', and I in Fig. 1b, the APB energy is the same but the shapes of the γ surface and, accordingly, the surface defect energies for sections I and I' differ. Earlier [10, 11], we considered various ways of dissociation of the **a** superdislocation with the Burgers vector $1/3\langle \bar{1}\bar{1}20 \rangle$ in the prismatic planes of types I, I', and II. For all of the ways studied, it was found that the post-relaxation energy values in type I' prismatic planes turn out to be much lower than in the prismatic planes of type I and II. As the initial configuration in type I' planes, two ways of dissociation of the screw (or edge) superpartial dislocation with the Burgers vector $1/6\langle \bar{1}\bar{1}20 \rangle$ were considered. One variant is the dissociation into n screw (edge) partial dislocations with the same Burgers vectors $\mathbf{b}_n = (1/6\langle \bar{1}\bar{1}20 \rangle)/n$. The second one is the dissociation into two groups of partial dislocations. For $n = 2$, the related reactions of dissociation of the $1/6\langle 2\bar{1}\bar{1}0 \rangle$ superpartial dislocation have the from

$$1/6[\bar{1}\bar{1}20] = 1/12[\bar{1}\bar{1}20] + \mathrm{SF} + 1/12[\bar{1}\bar{1}20], \quad (8)$$

$$1/6[\bar{1}\bar{1}20] = 1/12[\bar{1}\bar{1}2x] + \mathrm{SF} + 1/12[\bar{1}\bar{1}2x], \quad (9)$$

where $x = 1.038$.

The analysis of the core structure of edge and screw superpartial dislocations in type I' prismatic planes indicates that the core is planar; that is, the region of maximal displacements is localized near a type I' plane that contains the APB. It appeared that, for the initial configuration with the same Burgers vectors (reaction (8)), the energy is somewhat lower than for the other initial configuration (reaction (9)). For each number n of partial dislocations, the distribution of displacements characterizing the superpartial dislocation core is nonuniform along type I' prismatic planes and may be viewed as the dissociation into two partial dislocations. Thus, for a dislocation barrier (Fig. 2d), the Burgers vector $\mathbf{b}_2$ of the partial dislocation in type I' planes was set equal to $\mathbf{b}_2 = 1/12[\bar{1}\bar{1}20]$.

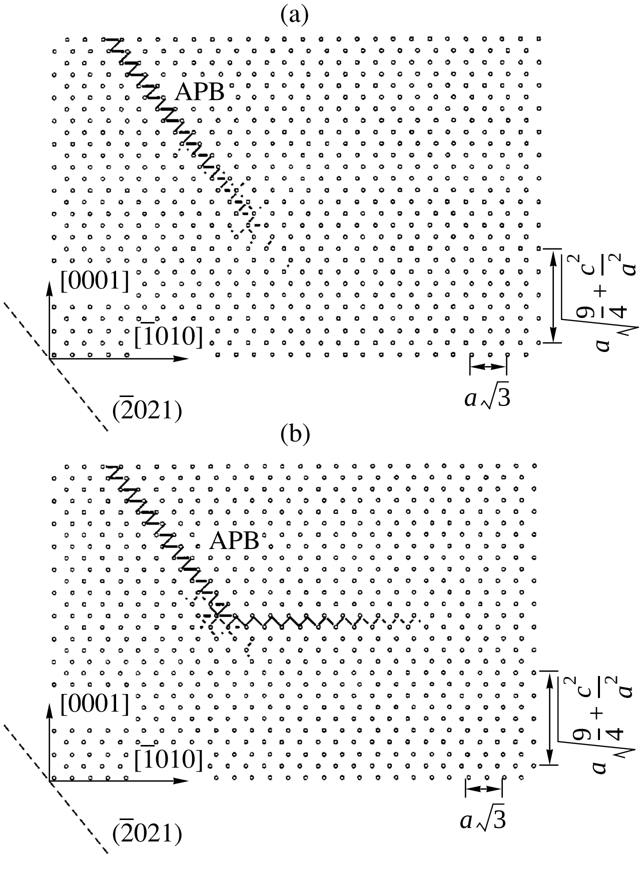


Fig. 4. Core structure of the edge superpartial with the Burgers vector $1/6[2\bar{1}\bar{1}6]$, which dissociates (a) in the $(\bar{2}021)$ pyramidal plane of type I (glissile configuration) and (b) in the $(\bar{2}021)$ pyramidal plane of type I and basal plane with the formation of a CSF (dislocation barrier). The displacement edge components are shown.

2. Glissile and Sessile Configurations of an Edge Superdislocation in Pyramidal Planes of Type I and II

As the initial configuration of a glissile superpartial dislocation in type I planes, we took its dissociation according to reaction (1). For this dislocation, the dissociation scheme in the planes considered is similar to that shown in Fig. 2a. Figure 4a shows the core structure of the superpartial dislocation with the Burgers vector $1/6[2\bar{1}\bar{1}6]$ and axis $[1\bar{2}10]$ that is split into two partial dislocations according to (1) with $\alpha = 0.5$ and $\beta = 0.5$.[1] When imaging the core structure, we used the method of differential displacements [20]. The circles

---

[1] For the dislocation with the axis $[1\bar{2}10]$, the angle between the axis and Burgers vector equals 85.08°. Therefore, this dislocation will be considered as an edge dislocation.

are the projections of atoms of the crystallite studied onto the plane $(1\bar{2}10)$, which is perpendicular to the axis of the edge dislocation considered. We computed atomic positions providing the lowest energy of the crystallite with a dislocation. Displacements not exceeding $0.05a$ (a is the Ti_3Al lattice parameter) are shown by the arrows between two neighboring atoms. From the spatial distribution of the arrows, one can judge the dislocation core structure. The core is planar, since most of both the edge and screw components of the displacement lie near the plane $(\bar{2}021)$, which contains an APB. For $\alpha = 0$ and $\beta = 0.475$, the post-relaxation core is similar to that shown in Fig. 4a; however, the energy of this configuration turned out to be considerably higher.

For the dislocation with the Burgers vector $1/6[2\bar{1}\bar{1}6]$ and axis $[1\bar{2}10]$ in type I pyramidal planes, we considered two types of dislocation barriers that appear when the superpartial dislocation overdissociates in the basal plane with the formation of an APB or a CSF (reactions (5) and (6)). The Shockley partial dislocation with the Burgers vector $1/6[10\bar{1}0]$ and the superpartial dislocation with the Burgers vector $1/6[2\bar{1}\bar{1}0]$ bound the CSF and APB in the basal plane. The core structure of the dislocation barrier forming by reaction (6) is shown in Fig. 4b. Both the edge and screw components of the displacements are distributed along one basal plane. For the dislocation located on the line of intersection of type I planes, the displacements are distributed along the pyramidal, prismatic, and basal planes.

For type II pyramidal planes, we studied three ways of dissociation of the edge superpartial dislocation with the Burgers vector $1/6[\bar{1}\bar{1}26]$ and axis $[\bar{1}100]$. Its initial configuration is defined by reactions (2)–(4). As follows from the results of simulation, the initial configuration has a minor effect on the post-relaxation energy and core structure of both edge and screw dislocations in this case.

The core structure of the glissile edge dislocation with the Burgers vector $1/6[\bar{1}\bar{1}26]$ in the type II planes is, in general, planar, since the displacements concentrate mainly near the plane $(11\bar{2}1)$, which contains the APB (the displacements are distributed as in Fig. 4). For the edge dislocation with the Burgers vector $1/6[\bar{1}\bar{1}26]$ and axis $[1\bar{1}00]$ in type II planes, also two kinds of dislocation barriers with the formation of an APB and a CSF were considered. For a barrier with an APB band in the basal plane, the superpartial dislocation dissociates into two partials as before, although the split in the basal plane somewhat increases, reaching $4a$, and the displacements are localized mainly in those planes where the initial configuration was split. The energy of the dislocation barrier with the CSF band in the basal plane is slightly below than for the case considered above, which reflects the fact that reaction (6) is more favorable than reaction (5). After relaxation, the dissociation into two partials in the basal plane persists but the spatial distribution becomes more complicated. For the dislocation with the Burgers vector $\mathbf{b}_2 = 1/6[0\bar{1}10]$, which bounds the CSF in the basal plane, the core is planar: both the edge and screw components are distributed in the basal plane. The core of the dislocation with the Burgers vector $\mathbf{b}_1 = 1/6[\bar{1}016]$ is nonplanar: the screw component of the Burgers vector $\mathbf{b}_1$ is distributed over several parallel pyramidal and basal planes.

Thus, as follows from the simulation of edge dislocations in type I and type II pyramidal planes, the energy of a glissile dislocation configuration is higher than that of a dislocation barrier.

3. Glissile and Sessile Configurations of the Screw Superdislocation in Type I and Type II Pyramidal Planes

For the glissile screw configuration in type I pyramidal planes, we considered three ways of dissociation of the superpartial dislocation in its initial configuration (reactions (2)–(4)). For type II planes, one version according to reaction (1) was studied. The analysis of the post-relaxation displacement distribution near the core of the screw superpartial dislocation with the Burgers vector $1/6[2\bar{1}\bar{1}6]$, which is split into two partials in the type I plane according to (1) with $\alpha = \beta = 0.5$, shows that the screw component of the partial displacements is distributed in the pyramidal plane $(\bar{2}021)$, as well as in the prismatic plane $(01\bar{1}0)$ and pyramidal plane $(\bar{2}201)$, which have the line of intersection parallel to the $[2\bar{1}\bar{1}6]$ axis of the superpartial. The edge component of the displacements is distributed in four neighboring pyramidal planes $\{\bar{2}201\}$; therefore, the core of this configuration may generally be viewed as nonplanar. As for the edge dislocation, the core of the screw superpartial with $\alpha = 0$ and $\beta = 0.475$ has a higher energy.

Figure 5 demonstrates the core of the glissile screw dislocation with the Burgers vector $1/6[\bar{1}\bar{1}26]$ in the type II pyramidal plane after the relaxation of its initial configuration by reaction (3). The distribution of the screw component is shown on the plane that is normal to the screw dislocation axis $1/6[\bar{1}\bar{1}26]$. Although the displacements are the greatest near the pyramidal plane, the configuration is generally nonplanar (in Fig. 5, the normal to the pyramidal plane $(11\bar{2}1)$ coincides with the direction $[11\bar{2}2]$). The displacement distribution is nonuniform along the pyramidal plane, and

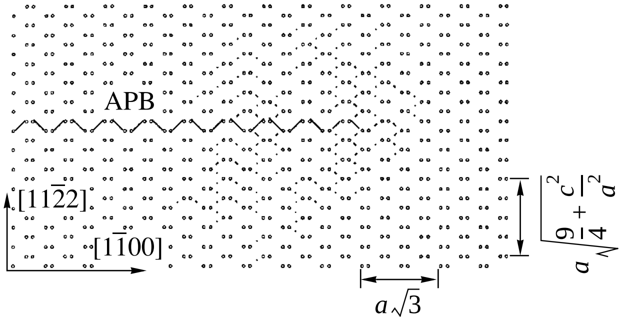


Fig. 5. Core structure of the screw superpartial with the Burgers vector 1/6[$\bar{1}\bar{1}26$], which dissociates in the $(11\bar{2}1)$ pyramidal plane of type II.

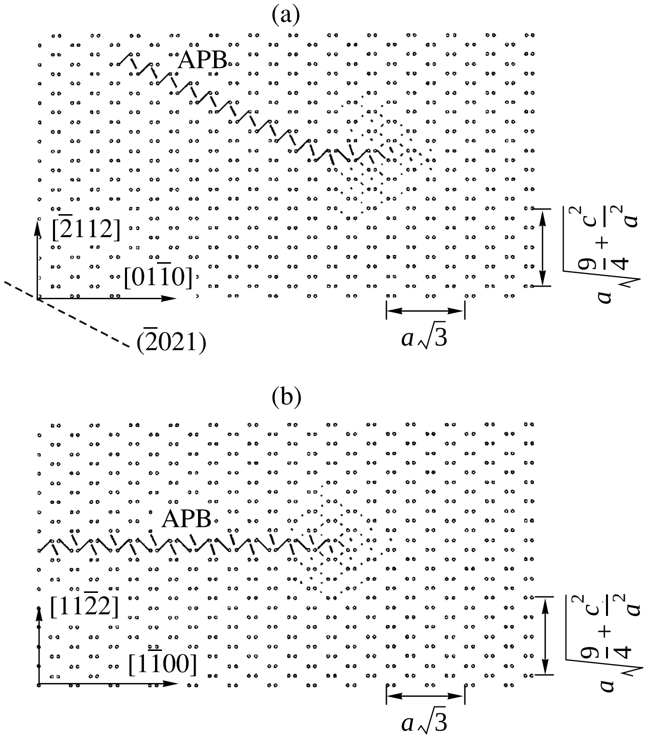


Fig. 6. (a) Core structure of the screw superpartial with the Burgers vector 1/6[$2\bar{1}\bar{1}6$] (dislocation barrier), which (a) dissociates and (b) does not dissociate in the $(\bar{2}021)$ pyramidal plane of type I and in the $(\bar{2}111)$ pyramidal plane of type II.

the situation may be considered as the dissociation into two partials. The displacements around either of the partials are located not only near the pyramidal planes but cover a range of (6–7)a in the plane $(1\bar{1}00)$. Such a dissociation occurs when the energy of the SF in the initial plane (in our case, in the pyramidal plane $(11\bar{2}1)$) is high and the SF overdissociation in the plane that passes through the dislocation line and has a lower SF energy (the prismatic plane $(1\bar{1}00)$ in our case) becomes energetically favorable.

In the type I pyramidal plane $(\bar{2}021)$, three kinds of dislocation barriers for the screw superpartial with the Burgers vector 1/6[$2\bar{1}\bar{1}6$] were considered. In the first case, the superpartial is split in the type I pyramidal plane and in the type I' prismatic plane $(01\bar{1}0)$; in the second case, in the type I pyramidal plane and in the type II pyramidal plane $(\bar{2}111)$; and in the third case, in the type I pyramidal planes $(\bar{2}021)$ and $(\bar{2}201)$. For the screw dislocation in type II pyramidal planes, two kinds of dislocation barriers were studied. In the first case, the superpartial dislocation is split in the type II pyramidal plane and in type I' prismatic plane; in the second case, in the type I and type II prismatic planes (Figs. 2f–2h). All the planes listed intersect along the [$2\bar{1}\bar{1}6$] direction, which coincides with the Burgers vector and axis of the dislocation.

In the former case (overdissociation in the type I or type II pyramidal plane and in the type I' prismatic plane), the energy of the dislocation barrier turns out to be higher than the energy of the glissile configuration and the energy of dislocations in the two other kinds of dislocation barriers; therefore, the formation of such barriers is unfavorable. Figure 6a shows the core structure of the screw superdislocation that is split in type I and type II pyramidal planes. As a whole, the core structure of these configurations is nonplanar: the displacements are distributed in pyramidal planes of both types and in prismatic planes simultaneously. The core of the dislocation barrier that is formed by extending the superpartial dislocation in two intersecting type I pyramidal planes (Fig. 2g) is the same. In the final configuration, the displacements are basically distributed largely in those type I pyramidal plane where the initial dissociations of the superpartial dislocation were preset. Figure 6b shows the core structure of the screw superdislocation that arises from the nondissociated superpartial dislocation with the Burgers vector 1/6[$\bar{1}\bar{1}26$]. The resulting configuration may be characterized as nonplanar, since the displacements are simultaneously distributed in pyramidal and prismatic planes. Estimates show that the energies of all the barriers are close to each other (for the configuration in Fig. 6b, the energy is slightly lower) and that the core of the **c** + **a**/2 screw superpartial is nonplanar.

CONCLUSIONS

In this work, γ surfaces in the pyramidal, prismatic, and basal planes are constructed for Ti_3Al by using N-particle potentials of interatomic interaction. These surfaces exhibit local minima corresponding to APBs. The APB energies, $\xi_{(0001)}$ = 161 mJ/m^2, $\xi_{(\bar{1}100)}$ = 6 mJ/m^2,

$\xi_{(11\bar{2}1)}$ = 228 mJ/m^2, and $\xi_{(20\bar{2}1)}$ = 225 mJ/m^2, are close to the experimental values [10].

Local minima are also observed on the γ surface corresponding to the superlattice intrinsic and complex stacking faults. The energies of the defects, γ_{SISF} = 197 mJ/m^2 and γ_{CSF} = 78 mJ/m^2, also agree with experimental data. It was shown that, in prismatic planes and type I and type II pyramidal planes, deep local minima associated with stable SFs are absent.

When simulating the core structure of superdislocations in type I and type II pyramidal planes, we assumed that they are split into two superpartials with the Burgers vectors 1/6$[\bar{1}\bar{1}26]$ with an APB in between. Such a configuration of superdislocations is related to deep minima on the γ surfaces in type I and type II pyramidal planes (Figs. 3a, 3b) with the displacement vector 1/6$[\bar{1}\bar{1}26]$. For the superpartial dislocation in its initial configuration, several versions of dissociation in type II pyramidal planes and only one way of dissociation in type I pyramidal planes were studied, since there are no other deep minima on the γ surface in pyramidal planes of type I. The analysis of the post-relaxation core structure of the edge glissile superpartial showed that the core as a whole is planar (Fig. 4a) and the displacement distribution may be viewed as the dissociation of the dislocation with the Burgers vector 1/6$[\bar{1}\bar{1}26]$ into two partials, since the displacements around the partials are localized largely near the pyramidal plane. For the screw dislocation with the axis $[\bar{1}\bar{1}26]$, the core of the superpartials is nonplanar (Figs. 5, 6), since the displacements around either of the partials are distributed not only near the pyramidal plane containing the APB but also in the prismatic planes and other intersecting pyramidal planes.

To date, the computer-aided simulation of the dislocation core structure in pyramidal planes for the D0$_{19}$ superlattice has not been performed; dislocations have been simulated in only basal and prismatic planes [21]. Let us compare our results with the simulation of the dislocation core structure in hcp metals [20–22] that was carried out for the slip system 1/3$\langle\bar{1}\bar{1}23\rangle\{11\bar{2}2\}$ using Lenard–Jones pair potentials. The slip system 1/3$\langle\bar{1}\bar{1}23\rangle\{11\bar{2}2\}$ for hcp crystals is equivalent to the system 1/6$\langle\bar{1}\bar{1}26\rangle\{11\bar{2}1\}$ for the ordered superlattice D0$_{19}$. The simulation of the screw dislocation core showed that the final result depends on the initial configuration for both the hcp lattice and D0$_{19}$ superlattice. For the glissile configuration, the results for the hcp lattice and D0$_{19}$ superlattice are similar. For hcp crystals, only one barrier for a dislocation split simultaneously in pyramidal planes of type I and type II. For the D0$_{19}$ superlattice, we considered several kinds of barriers for a screw superpartial dislocation (Figs. 2c–2h). One of them (Fig. 2h) is equivalent to that considered for the hcp lattice; in the others, the superpartial is split simultaneously in pyramidal and prismatic planes. The energies of the barriers considered are close to each other.

For an edge dislocation, the simulations of the hcp lattice and D0$_{19}$ superlattice gave different results. Based on experimental data for a number of hcp metals, Minonishi *et al.* [23, 24] simulated an edge dislocation, assuming that its initial configuration results from the dissociation into a two-layer twin in a $\{11\bar{2}2\}$ plane and a three-layer SF in a $\{11\bar{2}1\}$ plane. Experimental data for Ti$_3$Al and other crystals with the superlattice D0$_{19}$ suggest that twinning is observed at only very high deformation temperatures. At moderate temperatures close to the operating temperatures of these materials, twinning is absent and slip is accomplished through the motion of superdislocations.

We compared the glissile configurations of the superpartial dislocation with dislocation barriers. For the edge dislocation with the Burgers vector 1/6$[\bar{1}\bar{1}26]$, the displacements around its core are distributed mainly in pyramidal planes of type I and type II. After relaxation, the energy of the superpartial glissile configuration turns out to be higher than that of the dislocation barrier. For screw superdislocations with the axis $[\bar{1}\bar{1}26]$, three kinds of dissociation barriers were considered. The barriers form when the superdislocations overdissociate from initial pyramidal planes of type I or type II into other planes (prismatic planes of type I' or other pyramidal planes of type I and type II) that intersect in the initial one along the direction parallel to the Burgers vector of the dislocations. Also, the nondissociated configuration of the superpartial in pyramidal planes was considered. The energies of the dissociated configurations of the screw superpartial (glissile configurations and dislocation barriers) were found to be close to each other.

Thus, by applying computer-aided simulation of the core structure for edge and screw superpartials in pyramidal planes of type I and type II, we established that, for the edge dislocation, the glissile configuration has a higher energy than the dislocation barrier; for the screw dislocation, the energies of both configurations are nearly the same. A lower energy of the dislocation barrier is a prerequisite for the thermally activated transformation of a glissile configuration to a dislocation barrier. This, in turn, may explain thermal hardening observed experimentally in Ti$_3$Al. Mechanisms of thermally activated superdislocation reconfiguring and the deformation behavior of Ti$_3$Al are the subject of our next report [25].

ACKNOWLEDGMENTS

M.Ya.R. thanks the Young Scientist Promotion Commission at the Russian Academy of Sciences (grant no. 69, 1999).

REFERENCES

1. J. P. A. Lafvander, S. A. Couret, and H. A. Fraser, Philos. Mag. A **59**, 280 (1989).
2. Y. Minonishi and M. H. Yoo, Philos. Mag. Lett. **61**, 203 (1990).
3. Y. Minonishi, Philos. Mag. A **63**, 1085 (1991).
4. Y. Umakoshi, T. Nakano, and T. Takenaka, Acta Metall. Mater. **41**, 1149 (1993).
5. M. Legros, A. Couret, and D. Caillard, Philos. Mag. A **73**, 61 (1996).
6. M. Legros, A. Couret, and D. Caillard, Philos. Mag. A **73**, 81 (1996).
7. Y. Minonishi, Philos. Mag. Lett. **62**, 153 (1990).
8. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Fiz. Met. Metalloved. **93** (3), 1 (2002).
9. L. E. Karkina and L. I. Yakovenkova, Izv. Akad. Nauk, Ser. Fiz. **65**, 807 (2001).
10. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Izv. Akad. Nauk, Ser. Fiz. **66**, 891 (2002).
11. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Zh. Tekh. Fiz. **73** (1), 60 (2003) [Tech. Phys. **48**, 56 (2003)].
12. L. I. Yakovenkova, V. V. Kirsanov, L. E. Karkina, *et al.*, Fiz. Met. Metalloved. **89** (3), 31 (2000).
13. L. I. Yakovenkova, V. V. Kirsanov, L. E. Karkina, *et al.*, J. Soc. Photo-Opt. Instrum. Eng. **3987**, 159 (1999).
14. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, in *Proceedings of the Conference "Structures and Properties of Metals and Alloys"* (Ural. Otd. Ross. Akad. Nauk, Yekaterinburg, 1999), pp. 237–242.
15. E. V. Panova, L. E. Karkina, and E. P. Romanov, Fiz. Met. Metalloved. **75** (4), 166 (1993).
16. E. V. Panova, E. P. Romanov, and L. E. Karkina, Fiz. Met. Metalloved. **80** (3), 166 (1995).
17. E. V. Panova, L. E. Karkina, B. A. Grinberg, and E. P. Romanov, Fiz. Met. Metalloved. **85** (1), 111 (1998).
18. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Fiz. Met. Metalloved. **87** (3), 20 (1999).
19. V. Vitek, Cryst. Lattice Defects **5**, 1 (1974).
20. J. Cserti, M. Khantha, V. Vitek, and D. P. Pope, Mater. Sci. Eng., A **152**, 95 (1992).
21. S. Ando, T. Gotan, and H. Tonda, Metall. Mater. Trans. A **33**, 823 (2002).
22. V. Vitek and M. Igarashi, Philos. Mag. A **63**, 1059 (1991).
23. Y. Minonishi, S. Ishioka, M. Koiwa, and S. Morizumi, Philos. Mag. A **45**, 835 (1982).
24. Y. Minonishi, S. Ishioka, M. Koiwa, and S. Morizumi, Philos. Mag. A **46**, 761 (1982).
25. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Zh. Tekh. Fiz. **73** (10), 70 (2003) [Tech. Phys. **48** (10), 1289 (2003)].

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1289–1295. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 70–76.
Original Russian Text Copyright © 2003 by Yakovenkova, Karkina, Rabovskaya.

SOLIDS

Superdislocation Core Structure in Pyramidal Slip Planes and Temperature Anomalies of Ti_3Al Intermetallic Deformation Behavior

L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya
Institute of Metal Physics, Ural Division, Russian Academy of Sciences,
ul. S. Kovalevskoĭ 18, Yekaterinburg, 620219 Russia
e-mail: yakovenkova@imp.uran.ru

Received December 17, 2002; in final form, February 20, 2003

Abstract—For Ti_3Al intermetallic with the $D0_{19}$ superlattice, the computer-aided simulation of the superdislocation core structure in the basal, prismatic, and type I and type II pyramidal planes is performed. A model of thermally activated transformation of 2**c** + **a** glissile edge superdislocations into dislocation barriers in the pyramidal planes is used to account for the anomalous temperature dependence of the yield stress of single-crystal Ti_3Al compressed in the [0001] direction. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

In [1], we studied the core structure of glissile and sessile (barrier-forming) 2**c** + **a** superdislocations in type I, $(\bar{2}021)$, and type II, $(11\bar{2}1)$, pyramidal planes. The core structure of **a** superdislocations in basal and prismatic planes was visualized in [2–5]. The computer-aided simulation of the cores of various dislocations was carried out by the molecular dynamics methods using N-particle atomic interaction potentials for Ti_3Al. The need for computer simulation stems from the fact that a large body of experimental data for the strong orientation dependence of the yield stress σ_y, complex slip geometry, and temperature dependence of strain characteristics for single-crystalline Ti_3Al have been obtained to date. In particular, the normal run of the $\sigma_y(T)$ curve for the basal and prismatic planes and the anomalous run of this curve (with a peak) for pyramidal slip were observed. Yet, available experimental data cannot provide a detailed information on the fine core structure for **a** and 2**c** + **a** superdislocations in the basal, prismatic, and pyramidal planes and cannot shed light on the mechanism of dislocation transformations that are responsible for the deformation behavior of Ti_3Al.

The presence of glissile and sessile edge and screw dislocations in type I and type II pyramidal planes have been confirmed by electron microscopy data [6–9]. It was shown that, during deformation, sessile 2**c** + **a** superdislocations in these planes are observed in the temperature range where the yield stress $\sigma_y(T)$ exhibits an anomalous rise. As in other superlattices, such as $L1_2$ and $L1_0$ [10–12], the anomalous temperature dependence of the Ti_3Al strain characteristics seems to be associated with the thermally activated transformation of glissile dislocations, which are responsible for plastic deformation, into dislocation barriers. The theory of plasticity for ordered alloys, where several types of transformations in a dislocation ensemble [13] are considered, suggests the following prerequisites for a peak in the $\sigma_y(T)$ curve. First, the sessile configuration must be energetically more favorable than the glissile one and, second, the activation energy of the glissile–sessile transformation must be lower than the energy of barrier breakdown.

The aim of this work is to analyze the computer simulation results for the core structure of superdislocations forming glissile and sessile configurations in basal, prismatic, and type I and type II pyramidal planes. A correlation between the superdislocation core structure and anomalous temperature dependence of the Ti_3Al strain characteristics is considered in view of energy relationships between these dislocation configurations. Experimentally observed features of plastic deformation in Ti_3Al oriented for basal, prismatic, and pyramidal slips are discussed.

DISLOCATION ENSEMBLE AND STRAIN BEHAVIOR OF Ti_3Al

The experimentally observed high orientation dependence of the yield stress σ_y for single-crystal Ti_3Al reflects a variety of the Burgers vectors and slip planes of superdislocations. The analysis of the slip geometry indicates that **a** and 2**c** + **a** superdislocations may glide in the basal, prismatic, or pyramidal planes according to the experimental conditions. The computer-aided simulation of the superdislocation core structure in these planes has shown that each of the slip systems, which are characterized by the Burgers vector and slip plane, may form both glissile configurations

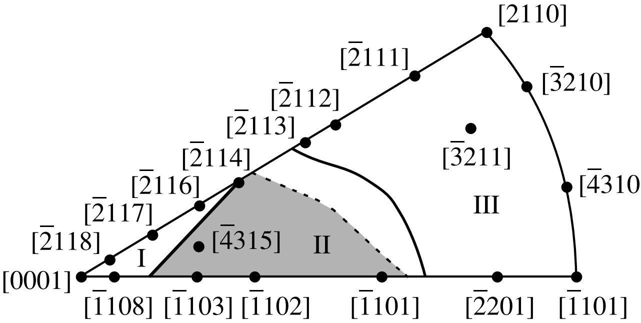


Fig. 1. Stereographic triangle $[0001]$–$[\bar{1}100]$–$[\bar{2}110]$ for the hcp lattice. In regions I–III, pyramidal, basal, and prismatic slip, respectively, prevails.

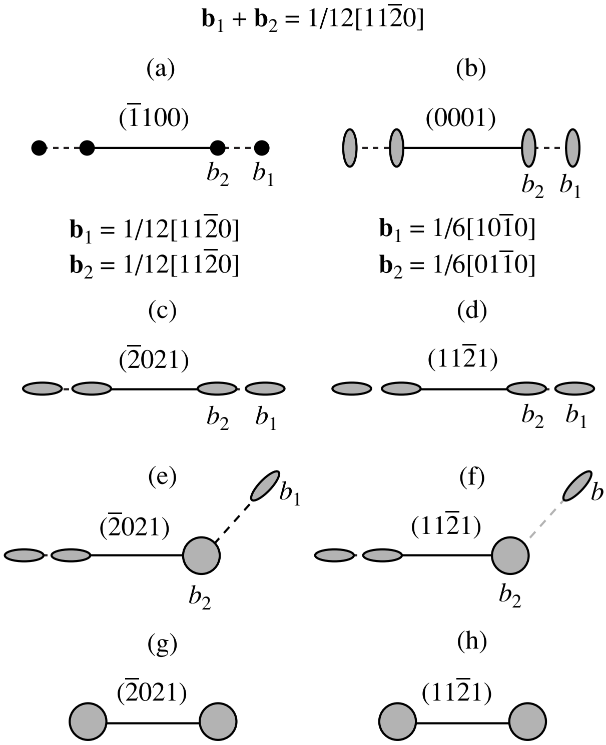


Fig. 2. Dissociation of an **a** superdislocation in the (a) type I' prismatic $(\bar{1}100)$ and (b) basal (0001) planes and the dissociation of a $2\mathbf{c} + \mathbf{a}$ superdislocation in the (c, e, g) type I $(\bar{2}021)$ and (d, f, h) type II $(11\bar{2}1)$ pyramidal planes. (a–d) Glissile configurations, (e, f) dislocation barriers that form when an edge superpartial dissociates in pyramidal and basal planes simultaneously, and (g, h) nonplanar core of a screw superpartial in pyramidal planes. — APB; - - - SF.

and dislocation barriers. The complex dislocation ensemble may be divided into several groups of dislocations depending on the deformation direction in single-crystal Ti_3Al. Figure 1 shows the standard stereographic triangle $[0001]$–$[\bar{1}100]$–$[\bar{2}110]$, where basal-, prismatic-, and pyramidal-slip regions are marked. In region I (bounded by the solid line) within 20°–25° from the [0001] direction, the dislocations in the single crystal glide largely in the type I and type II pyramidal planes. For the single crystals with the deformation axis within this region, the Schmid factors f ($f = (\mathbf{b} \cdot \mathbf{t})(\mathbf{b} \cdot \mathbf{n})$), where **b**, **n**, and **t** are the unit vector along the Burgers vector, unit normal to the slip plane, and unit normal to the deformation axis **t**) in pyramidal planes, $f_{\{2021\}\langle 11\bar{2}6\rangle}$ and $f_{\{\bar{1}\bar{1}21\}\langle 11\bar{2}6\rangle}$, are close to 0.5, while in the prismatic and basal planes, they approach zero. In region II in Fig. 1, which is also bounded by solid lines, the deformation axis orientation favors basal slip (the Schmid factor $f_{\{0001\}\langle\bar{1}\bar{1}20\rangle}$ is maximal). Finally, in region III, slip is primarily prismatic (the Schmid factor $f_{\{\bar{1}100\}\langle\bar{1}\bar{1}20\rangle}$ is maximal).

To begin with, we will consider the dislocation ensemble in the prismatic and basal planes, where the yield stress is low and the deforming stress shows a normal temperature dependence (i.e., decreases with increasing temperature) [14, 15]. Next, we will analyze the dislocation ensemble in the type I and type II pyramidal planes, for which the $\sigma_y(T)$ curve has a peak, the yield stress is much higher, and the stress at fracture is low [6, 16, 17].

1. Prismatic and Basal Slip

Experimental data [15] show that the room-temperature critical resultant shear stress[1] is $\sigma_y^{\{1\bar{1}00\}} = 50$–150 MPa in the prismatic planes and $\sigma_y^{\{0001\}} = 180$–260 MPa in the basal planes; that is, the yield stress in the basal planes is roughly twice as high as that in the prismatic planes. One can easily explain this fact by comparing experimental data with our simulated results for the core structure of **a** superdislocations in the basal and prismatic planes (Figs. 2a, 2b). In [18–20], we studied the shape of the γ surface and the core structure of the superdislocation with the Burgers vector $1/3[2\bar{1}\bar{1}0]$ in prismatic planes of types I, I', and II (for details, see [1]). It turned out that the post-relaxation energies of the antiphase boundary (APB), stacking faults (SFs), and dissociated **a** superdislocation in the type I' prismatic plane are lower than in the type I and type II prismatic planes. The core of both edge and screw superpartials in the type I' prismatic plane is planar; that is, the displacements concentrate largely near the type I' prismatic plane that contains the APB (Fig. 2a). This is the reason why the Peierls force, which triggers the motion of the superdislocation, providing its high mobility in type I' prismatic planes, is low and, accordingly, the experimentally found yield stress in these planes of Ti_3Al is also low. When the axis of deformation of single-crystal Ti_3Al is within region

---

[1] The critical resultant shear stress is the component of the shear stress σ_{bn} acting on a dislocation with a Burgers vector **b** and a normal **n** to the slip plane.

III of the standard stereographic triangle (Fig. 1), the Schmid factor is maximal in the prismatic plane; therefore, the yield stress is bound to be the lowest.

The simulation of **a** superdislocations in basal planes that dissociate with the formation of an APB demonstrated that the core structure of the 30° and edge dislocations remains planar. For the 60° and screw superdislocations, as well as for Shockley partials forming the superpartial with the Burgers vector $1/6[2\bar{1}\bar{1}0]$, considerable relative displacements in the prismatic plane of cross slip are observed (Fig. 2b). As a whole, the core region of the **a** superdislocation is localized near the basal plane and the Peierls force is relatively low; therefore, the yields stress in the basal plane is not too high but is still higher than in the prismatic plane. Also note that the nonplanar character of Shockley partials shows up in the jerky movement of superdislocations in the basal plane, which was observed in *in situ* experiments [21].

When the axis of deformation falls into region II of the stereographic triangle, the Schmid factor is the highest in the basal plane. With $f_{\{\bar{1}100\}\langle\bar{2}110\rangle} < 0.5 f_{(0001)\langle\bar{2}110\rangle}$ (the unhatched part of region II), the reduced shear stress is higher than the yield stress in the basal plane but lower than the shear stress in the prismatic plane. Thus, in this range of compression axis orientations, deformation will be accomplished through the dislocation motion in the basal plane alone. In the hatched part of region II of the stereographic triangle, $0.5 f_{(0001)\langle\bar{1}\bar{1}20\rangle} < f_{\{\bar{1}100\}\langle\bar{1}\bar{1}20\rangle} < f_{(0001)\langle\bar{1}\bar{1}20\rangle}$; therefore, basal and prismatic slips will proceed simultaneously. This situation may not considerably increase the yield stress compared with the other orientations in region II. However, the interaction between these two slip systems may produce indestructible dislocation barriers and reduce the ductility of Ti_3Al single crystals with the above orientations.

For example, for the $[\bar{4}315]$ orientation (Table 1, Fig. 1), the Schmid factors in the primary prismatic and basal planes meet the inequality $f_{\{\bar{1}010\}\langle 1\bar{2}10\rangle} > 0.5 f_{(0001)\langle\bar{2}110\rangle}$. Therefore, one may expect that deformation will proceed in these slip systems. It should be noted that secondary slip systems in the basal and prismatic planes (Table 1) have much lower Schmid factors and do not take part in deformation. Attracting dislocations lying in the primary prismatic and basal slip systems may react to form barriers at the line of intersection of these planes.

Table 1. Maximal Schmid factors *f* in basal and prismatic slip systems for the orientation $[\bar{4}315]$

No.	Slip system	*f*
1	$[\bar{2}110]$ (0001)	0.44
2	$[1\bar{2}10]$ $(\bar{1}010)$	0.35
3	$[1\bar{2}10]$ (0001)	0.33
4	$[\bar{2}110]$ $(0\bar{1}10)$	0.19

As follows from the simulated results [18–20], **a** superdislocations in the prismatic and basal planes dissociate into two superpartials with an APB in between:

$$1/3[\bar{1}2\bar{1}0]_{(\bar{1}010)} = 1/6[\bar{1}2\bar{1}0]_{(\bar{1}010)} + APB_{(\bar{1}010)} + 1/6[\bar{1}2\bar{1}0]_{(\bar{1}010)}, \tag{1}$$

$$1/3[\bar{2}110]_{(0001)} = 1/6[\bar{2}110]_{(0001)} + APB_{(0001)} + 1/6[\bar{2}110]_{(0001)}. \tag{2}$$

In both the basal and prismatic planes, either of the superpartials is split into two partials. In the basal plane, the partial dislocation bounds a complex stacking fault (CSF). The displacement vector that is equal to the Burgers vector corresponds to a local minimum on the γ surface in the basal plane. The computer simulation determined the Burgers vectors of partials in the type I' prismatic plane, which were found to be associated with a metastable SF. In the basal plane, the dissociation of the superpartials has the form

$$1/6[\bar{2}110]_{(0001)} = 1/6[\bar{1}100]_{(0001)} + CSF_{(0001)} + 1/6[\bar{1}010]_{(0001)}; \tag{3}$$

in the prismatic plane,

$$1/6[1\bar{2}10]_{(\bar{1}010)} = 1/12[1\bar{2}10]_{(\bar{1}010)} + SF_{(\bar{1}010)} + 1/12[1\bar{2}10]_{(\bar{1}010)}. \tag{4}$$

Figure 3 illustrates the stages of formation of dislocation barriers on superdislocations interacting in the (0001) basal plane and in the $(\bar{1}010)$ prismatic plane. The axes of barrier-forming dislocations coincide with the line $[1\bar{2}10]$, which is the line of intersection of these planes. Thus, the dislocations involved in the reaction are the screw **a** superdislocation in the prismatic plane and the 60° **a** superdislocation in the basal plane. The analysis shows that, for all the orientations in the hatched part of region II in Fig. 1, dislocation barriers forming as a result of interaction between the **a** superdislocations in the primary prismatic and basal planes are equivalent to those represented in Figs. 3c and 3g. Two types of barriers form according to the sequence of dislocation reactions between the partials that constitute the **a** superdislocations in the basal and

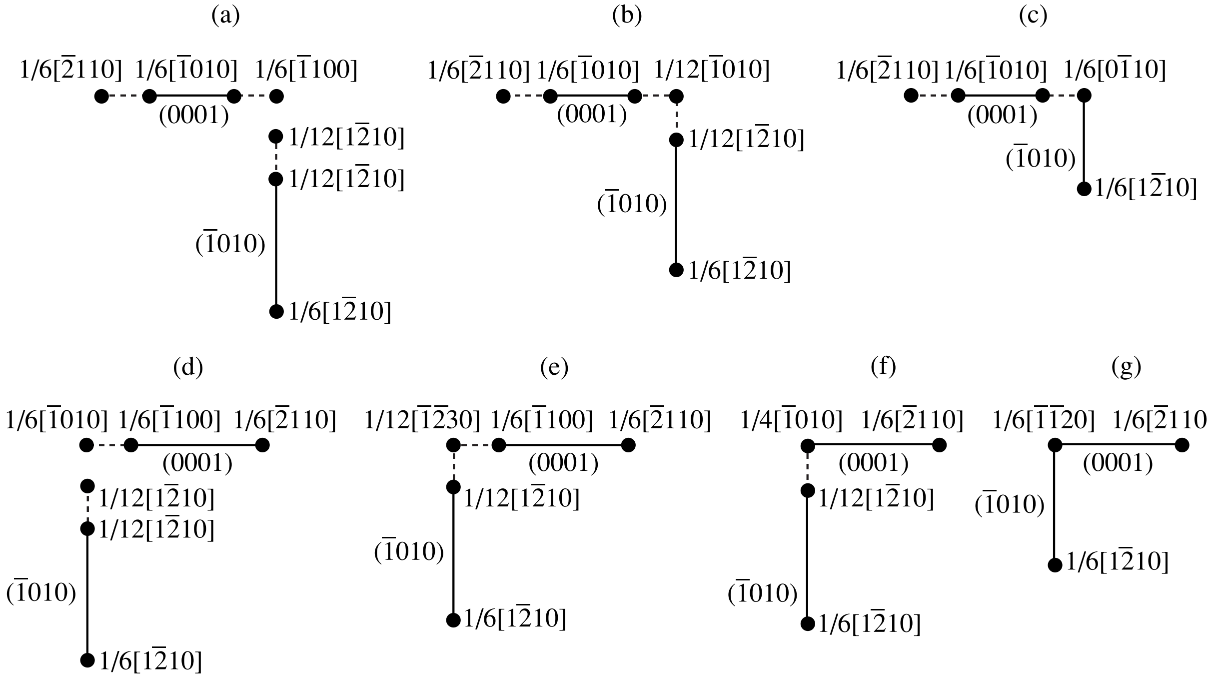


Fig. 3. Formation of barriers when superdislocations react in basal and prismatic planes. — APB; - - - SF.

prismatic planes. In view of (3) and (4), the partials with the Burgers vector 1/12$[1\bar{2}10]$, which lie in the prismatic plane, may react either with the dislocation with the Burgers vector 1/6$[\bar{1}100]$ or with the dislocation with the Burgers vector 1/6$[\bar{1}010]$, which lie in the basal plane. In the former case, the barrier formation stages are shown in Figs. 3a–3c; in the latter case, in Figs. 3d–3g. The partials with the Burgers vectors 1/6$[\bar{1}100]$ and 1/12$[1\bar{2}10]$ are attracted together, producing a stair-rod dislocation with the Burgers vector 1/12$[\bar{1}010]$ (Fig. 3b). This stair-rod dislocation is repelled from the second basal partial with the Burgers vector 1/6$[\bar{1}010]$ but may react with the prismatic partial with the Burgers vector 1/12$[1\bar{2}10]$ to form the configuration shown in Fig. 3c. The stress field of the prismatic superpartial with the Burgers vector 1/6$[1\bar{2}10]$ causes the recombination of the dislocations with the Burgers vectors 1/12$[\bar{1}010]$ and 1/12$[1\bar{2}10]$.

Similarly, the stress field of the dislocation with the Burgers vector 1/6$[1\bar{2}10]$ generates a stair-rod dislocation with the Burgers vector 1/12$[1\bar{2}30]$ (Figs. 3d, 3e) at the initial stage of formation of the second barrier. The resulting stair-rod dislocation is repelled from the prismatic partial with the Burgers vector 1/12$[1\bar{2}10]$ but is attracted to the basal partial with the Burgers vector 1/6$[\bar{1}100]$. Their recombination results in the formation of the stair-rod dislocation with the Burgers vector 1/4$[\bar{1}010]$ (Fig. 3f). At the last step, the dislocations with the Burgers vectors 1/4$[\bar{1}010]$ and 1/12$[1\bar{2}10]$ recombine in the stress field of the superpartials with the Burgers vectors 1/6$[\bar{2}110]$ and 1/6$[1\bar{2}10]$, which lie in the basal and prismatic planes, respectively, and produce a dislocation barrier (Fig. 3g).

The configurations shown in Figs. 3c and 3g are indestructible dislocation barriers and may considerably reduce the ductility of the single crystals with deformation axis orientations within the hatched part of region II in the stereographic triangle (Fig. 1) because these barriers are crack nuclei.

2. Pyramidal Slip

Experimental data suggest that the yield stress anomaly in Ti_3Al is related to the motion of 2**c** + **a** dislocations in type I and type II pyramidal planes. After deformation at 600 and 700°C, Minonishi and Yoo [16] and Minonishi [17] observed slip bands of sessile edge and near-edge 2**c** + **a** dislocations in the range of anomalous rise in σ_y. Extended straight-line screw 2**c** + **a** superdislocations after room-temperature deformation and extended straight-line edge superdislocations after deformation at 400°C were found in [7–9].

Thus, from experiments, it follows that sessile screw 2**c** + **a** superdislocations are observed in the range where the yield stress depends on temperature only slightly (before the anomalous rise in $\sigma(T)$). In the range of anomalous rise, edge and near-edge sessile 2**c** + **a** superdislocations appear in type I and type II pyramidal planes. It is well known that the anomalous behavior of $\sigma_y(T)$ in ordered alloy with the $L1_2$ superlattice [12, 13] is due to the thermally activated transformation of glissile superdislocations into Kear–Wilsdorff dislocation barriers. We believe that the anomalous behavior of $\sigma_y(T)$ in Ti_3Al is also related to thermally activated dislocation transformations.

Let us consider the simulated results for the core structure of 2**c** + **a** superdislocations in type I and type II pyramidal planes in order to explain the anomalous temperature dependence of the yield stress in Ti_3Al. Figures 2c and 2d show the core structure of glissile superdislocations in type I and type II pyramidal planes. According to [1], the core of edge superdislocations in type I and type II pyramidal planes is planar; that is, the displacements are localized largely near type I and type II pyramidal planes containing APBs. Figures 2e and 2f depict the structure of barriers made up of edge superdislocations. It is seen that one of the superpartials is split in the basal plane, while the other and the APB lie in the pyramidal plane. The thermally activated transformation of superdislocations implies the presence of glissile and sessile dislocation configurations. Also, the barrier energy must be lower than the glissile dislocation energy. Our calculations show that the latter condition is met for 2**c** + **a** superdislocations in type I and type II pyramidal planes.

Table 2 gives energy gains per unit dislocation length when edge glissile and sessile (barrier-forming) 2**c** + **a** superdislocations dissociate in type I and type II pyramidal planes.[2] The lowest values of the barrier energy are listed. In type I and type II pyramidal planes, dislocation barriers are produced when the superpartial dislocation overdissociates into the basal plane with the formation of a CSF (Figs. 2e, 2f).

Table 2. Energy gain upon splitting glissile and barrier-forming **c** + **a**/2 superdislocations

Slip system, dislocation axis	Glissile configuration, eV/nm	Dislocation barrier, eV/nm
$[\bar{1}\,\bar{1}\,26](11\bar{2}\,1)$, $[\bar{1}\,010]$	27.14	30.15
$[2\bar{1}\,\bar{1}\,6](\bar{2}\,021)$, $[1\bar{2}\,10]$	46.48	46.86

In [24], we calculated the activation energy of barrier formation in type I and type II pyramidal planes using approximations made in [22, 23]. Prerequisites for the formation of the barriers are the recombination of the **c** + **a**/2 superpartial in the pyramidal planes and the sag of the partial with the Burgers vector $1/6\langle\bar{1}100\rangle$ toward the basal plane. Figure 4 shows the dependence of the activation energy of barrier formation on the applied shear stress. The activation energies in the type I and type II pyramidal planes are close to each other. Thus, if the **c** + **a**/2 superpartial may be pinned in one of the pyramidal planes, it may also be pinned in the other. It is reasonable to compare the computed values of the activation energy of barrier formation with those extracted from the temperature dependence of the yield stress for Ti_3Al [8, 9, 16, 17], which can be represented by the empirical formula $\sigma \sim \exp(-W/kT)$ (k is the Boltzmann constant). From available experimental curves $\sigma_y(T)$ ($\sigma_y \approx 1.2$ GPa), one may estimate the activation energy in the temperature interval 300–650°C as $W = 0.07$–0.15 eV. For barriers made up of edge dislocations, the curves in Fig. 4 give the activation energy values, which are very close to those found experimentally.

The analysis of the slip geometry shows that, with the deformation axes located within region I of the stereographic triangle in Fig. 1, the deformation of single-crystal Ti_3Al involves both type I and type II pyramidal planes. An important result of the computer-aided simulation is that thermally activated dislocations move in both slip systems. Otherwise, the anomaly in the curve $\sigma_y(T)$ would disappear, which comes into conflict with the experiment. The anomaly is not observed for those orientations making the glide of **a** superdislocations in prismatic and basal planes possible (regions II and III of the stereographic triangle). The presence of two pin-

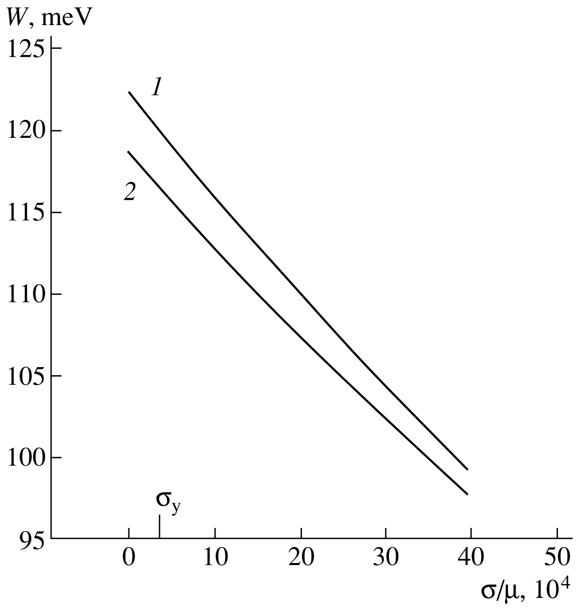


Fig. 4. Activation energy of formation of barriers from edge superdislocations vs. applied stress (μ is the shear modulus). The curves refer to barriers in the (*1*) type I pyramidal plane $(\bar{2}021)$ and (*2*) type II pyramidal plane $(11\bar{2}1)$.

---

[2] The energy gain is defined as the difference between the superpartial dislocation energy after relaxation and the energy of the unsplit dislocation.

ning mechanisms for superdislocations sheared by various slip systems reflects the complexity of the dislocation ensemble in Ti_3Al and similar intermetallics.

As follows from the electron microscopy examination of the dislocation structure [7, 8], the deformation of single-crystal Ti_3Al at room temperature produces extended screw 2**c** + **a** superdislocations in the type I and type II pyramidal planes. The pinning of edge superdislocations takes place at higher temperatures ($T > 300°C$) in the range of anomalous rise of $\sigma_y(T)$. For pyramidal slip in the temperature range from room temperature to $T \approx 300°C$, the yield stress is roughly five times as high as the yield stress for prismatic and basal slips. Our calculations show that the core structure of the screw **c** + **a**/2 superpartial in the type I and type II pyramidal planes is nonplanar both in the initial glissile configuration and in the barrier configuration (Figs. 2g, 2h). The displacements here are distributed in pyramidal planes of both types and in prismatic planes simultaneously. Thus, the computer-aided simulation suggests that screw superpartials in pyramidal planes are subjected to spontaneous pinning. The nonplanar structure of the screw dislocation core has also been found for fcc lattices [25], bcc lattices [26], and hcp lattices [27], as well as ordered alloys with the $L1_2$ [25] and $D0_{19}$ [28] superlattices. This feature of the screw superpartial core may be an explanation for a high yield stress of Ti_3Al single crystals in pyramidal planes when the axis of deformation falls into regions II and III.

When the deformation axis of the single crystals lies in region I of the stereographic triangle (Fig. 1), usually several, rather than one, pyramidal slip systems are involved. If the deformation axis coincides exactly with the direction [0001], the Schmid factors are maximal for 24 slip systems $\langle 11\bar{2}6\rangle\{\bar{2}201\}$. If the deformation axis is offset from the direction [0001], the number of slip systems decreases while remaining relatively large. For example, for the orientation $[\bar{2}118]$, which also belongs to region I of the stereographic triangle, two primary slip systems, $[1\bar{2}16](\bar{2}201)$ and $[11\bar{2}6](\bar{2}021)$, have the Schmid factors $f = 0.49$ and two secondary slip systems, $[2\bar{1}\bar{1}6](\bar{2}201)$ and $[2\bar{1}\bar{1}6](\bar{2}111)$, have nearly the same Schmid factors $f = 0.46$. Thus, for the deformation axis in the direction $[\bar{2}118]$, deformation may proceed in four slip systems. The multiplicity of pyramidal slip systems causes (as in the case of interaction between basal and prismatic slip systems) indestructible dislocation barriers to arise upon the interaction of dislocations from different slip systems. These barriers significantly raise the yield stress at pyramidal slip and may act as crack nuclei. Furthermore, because of a low mobility of both edge and screw superdislocations in pyramidal planes, originating cracks open up in a brittle manner without forming a plastic zone. This explains the low ductility of the material in experiments where the axis of deformation lies in region I [7–9, 16].

CONCLUSIONS

Based on the simulated results for the core structure of **a** and 2**c** + **a** superdislocations in prismatic, pyramidal, and basal planes of Ti_3Al, one can draw the following conclusions.

(1) The core of both edge and screw **a**/2 superpartials in type I prismatic planes is nonplanar, which provides a low "prismatic" yield stress. The fact that a "basal" yield stress is higher is related to the nonplanar core of the Shockley partial as a component of both screw and 60° **a**/2 superpartials.

(2) The energy of barriers made up of edge **c** + **a**/2 dislocations dissociated in pyramidal and basal planes simultaneously is lower than the energy of the dislocations in the glissile configuration.

(3) Screw **c** + **a**/2 superpartials in type I and type II pyramidal planes have a nonplanar core. In this case, the displacements are distributed in one prismatic and several pyramidal planes.

(4) A dislocation theory of deformation of single-crystal Ti_3Al at pyramidal, basal, and prismatic slips is developed based on the computer-aided simulation of the superdislocation core structure.

(5) The anomalous temperature dependence of the yield stress, which is observed in single-crystal Ti_3Al compressed in the direction [0001], is explained by the thermal model of transformation of glissile edge 2**c** + **a** superdislocations into a dislocation barrier.

ACKNOWLEDGMENTS

M.Ya.R. thanks the Young Scientist Promotion Commission at the Russian Academy of Sciences (grant no. 69, 1999).

REFERENCES

1. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Zh. Tekh. Fiz. **73** (1), 60 (2003) [Tech. Phys. **48**, 56 (2003)].
2. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Fiz. Met. Metalloved. **93** (3), 1 (2002).
3. L. E. Karkina and L. I. Yakovenkova, Izv. Akad. Nauk, Ser. Fiz. **65**, 807 (2001).
4. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Izv. Akad. Nauk, Ser. Fiz. **66**, 891 (2002).
5. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Zh. Tekh. Fiz. **73** (10), 61 (2003) [Tech. Phys. **48** (10), 1280 (2003)].
6. Y. Minonishi, Philos. Mag. A **63**, 1085 (1991).
7. E. V. Panova, L. E. Karkina, and E. P. Romanov, Fiz. Met. Metalloved. **75** (4), 166 (1993).
8. E. V. Panova, E. P. Romanov, and L. E. Karkina, Fiz. Met. Metalloved. **80** (3), 166 (1995).

9. E. V. Panova, L. E. Karkina, B. A. Grinberg, *et al*, Fiz. Met. Metalloved. **85** (1), 111 (1998).
10. S. M. Copley and B. H. Kear, Trans. Metall. Soc. AIME **239**, 977 (1967).
11. T. Kawabata, Y. Takezono, T. Kanai, *et al.*, Acta Metall. **36**, 963 (1988).
12. B. A. Grinberg, M. A. Ivanov, Yu. N. Gornostyrev, *et al.*, Fiz. Met. Metalloved. **46**, 813 (1978).
13. B. A. Grinberg and V. I. Syutkina, *New Methods of Alloy Hardening* (Metallurgiya, Moscow, 1985).
14. H. Inui, Y. Toda, Y. Shirai, *et al.*, Philos. Mag. A **269**, 1161 (1994).
15. M. Legros, A. Couret, and D. Caillard, Philos. Mag. A **73**, 61 (1996).
16. Y. Minonishi and M. H. Yoo, Philos. Mag. Lett. **61**, 203 (1990).
17. Y. Minonishi, Philos. Mag. A **63**, 1085 (1991).
18. L. I. Yakovenkova, V. V. Kirsanov, L. E. Karkina, *et al.*, Fiz. Met. Metalloved. **89** (3), 31 (2000).
19. L. I. Yakovenkova, V. V. Kirsanov, L. E. Karkina, *et al.*, J. Soc. Photo-Opt. Instrum. Eng. **3987**, 159 (1999).
20. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, in *Proceedings of the Conference "Structures and Properties of Metals and Alloys*" (Ural. Otd. Ross. Akad. Nauk, Yekaterinburg, 1999), pp. 237–242.
21. M. Legros, A. Couret, and D. Caillard, Philos. Mag. A **73**, 81 (1996).
22. B. A. Greenberg and L. I. Yakovenkova, Phys. Status Solidi A **18**, K129 (1973).
23. B. A. Grinberg and L. I. Yakovenkova, Fiz. Met. Metalloved. **36**, 1166 (1973).
24. L. I. Yakovenkova, L. E. Karkina, and M. Ya. Rabovskaya, Fiz. Met. Metalloved. **87** (3), 44 (1999).
25. L. E. Karkina, B. A. Grinberg, and L. I. Yakovenkova, Fiz. Met. Metalloved. **61**, 695 (1986).
26. V. Vitek, Cryst. Lattice Defects **5**, 1 (1974).
27. V. Vitek and M. Igarashi, Philos. Mag. A **63**, 1059 (1991).
28. J. Cserti, M. Khantha, V. Vitek, and D. P. Pope, Mater. Sci. Eng., A **152**, 95 (1992).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1296–1303. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 77–84.
Original Russian Text Copyright © 2003 by Romanovskiĭ.

SOLIDS

Magnetic Relaxation in Superconductors with Varying Creep Behavior

V. R. Romanovskiĭ

Russian Research Centre Kurchatov Institute, pl. Kurchatova 1, Moscow, 123182 Russia

Received October 1, 2002; in final form, February 12, 2003

Abstract—The relaxation of an electromagnetic field inside low- and high-temperature superconductors in the case when the magnetic flux occupies the cross section incompletely is studied theoretically in the self-similar approximation. On the macroscopic level, creep is described by phenomenological equations for exponential and power *I*–*V* characteristics. A relation between the nonlinearity of the *I*–*V* curves and the magnetic relaxation process is found. The range of low electric fields is shown to be of considerable importance for the relaxation in high-temperature superconductors. In this range, the equations for exponential and power *I*–*V* curves may differ substantially from those at high fields. Complete analytical solutions are contrasted with a numerical solution to the problem. © *2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

By studying relaxation processes in superconductors, one can shed light on pinning mechanisms and transients in superconducting devices operating in rapidly varying magnetic fields. Therefore, magnetic relaxation has been the subject of much investigation (see, e.g., [1–7] and Refs. cited therein). Yet, in spite of extensive theoretical efforts, the results obtained up to now are incomplete. In particular, comparative analysis of vortex macrodiffusion in view of various creep models has not been performed. In the diffusion approximation, creep models are described by equations for *I*–*V* characteristics. In the simplest case, these equations have the form [8, 9]

$$E = E_C(J/J_C)^n, \tag{1}$$

$$E = E_C\exp[(J - J_C)/J_\delta]. \tag{2}$$

Here, J_C is the current density determined at an electric field strength E_C, n is an exponent defining the steepness of the curve, and J_δ is the rise rate constant of the curve.

In this work, we compare in a macroscopic approximation the magnetic relaxation processes in superconductors for which *I*–*V* equations (1) and (2) are valid. The rules formulated in this work follow from a set of solutions to the problems of electromagnetic field relaxation inside a cooled superconductor. These solutions agree with experimental measurements of the magnetic moment in the case when an external magnetic field $B(x, t) = B_0 + B_i$ (B_0 is the large constant component and B_i is the small variable component, $B_0 \gg B_i$) takes a constant value $B_{0,\,i}$ at $t = t_i$. Without loss in generality, it is assumed that the external magnetic field is applied parallel to the boundary surface of a semi-infinite superconductor. Therefore, according to the Maxwell equations, the relaxation of an electric field induced by an external perturbation in the Cartesian coordinate system is described by the equations [6]

$$\frac{\partial^2 E}{\partial x^2} = \mu_0 \begin{cases} \dfrac{J_C}{nE}\left(\dfrac{E}{E_C}\right)^{1/n}\dfrac{\partial E}{\partial t} & \text{for } I\text{–}V \text{ curve } (1), \quad (3) \\[2ex] \dfrac{J_\delta}{E}\dfrac{\partial E}{\partial t} & \text{for } I\text{–}V \text{ curve } (2), \quad (4) \end{cases}$$

with the boundary and initial conditions

$$\partial E/\partial x(0, t) = 0, \quad E(\infty, t) = 0, \qquad E(x, t_i) = E_0(x). \tag{5}$$

Here, $E_0(x)$ is the distribution of the induced electric field at the instant the external magnetic field ceases to rise. The second boundary condition assumes that the image current occupies completely the cross section of the superconductor irrespective of the form of the *I*–*V* curve. This assumption forms the basis for the numerical analysis of the electrodynamic states of superconductors with creep (see, e.g., [7]).

Below, we will also use the condition of conservation of the magnetic flux introduced:

$$\mu_0\int_0^\infty J(x, t)dx = B_{0,\,i}, \quad t \ge t_i. \tag{6}$$

It will be shown that this condition is necessary for the correct characterization of relaxation under creep.

As in [3], relaxation processes taking place under creep are studied with self-similar approximations.

RELAXATION OF THE ELECTROMAGNETIC FIELD INSIDE A SUPERCONDUCTOR WITH A POWER *I–V* CURVE

Let us introduce the dimensionless variables

$$e = E/E_C, \quad X = x/L_x, \quad \tau = t/t_x,$$

$$L_x = B_{0,i}/(\mu_0 J_C), \quad t_x = B_{0,i}^2/(\mu_0 J_C E_C).$$

A solution to the problem given by (3), (5), and (6) will be sought in the form

$$e = (\tau+\tau_0)^q W(Z), \quad X = (\tau+\tau_0)^p Z,$$

$$q = -n/(n+1), \quad p = 1/(n+1),$$

where τ_0 is a constant to be determined.

Passing from differentiation with respect to the independent variables to differentiation with respect to *Z*, we obtain

$$\begin{gathered}(n+1)\frac{d^2W}{dZ^2} + \frac{Z}{n}W^{(1-n)/n}\frac{dW}{dZ} + W^{1/n} = 0,\\ dW/dZ(0) = 0, \quad W(\infty) = 0.\end{gathered} \tag{7}$$

Problem (7) has an analytical solution in the form

$$W(Z) = \left[\frac{n-1}{2n(n+1)}(Z_0^2 - Z^2)\right]^{\frac{n}{n-1}}, \tag{8}$$

where Z_0 is an unknown constant.

An interesting feature of the invariant $W(Z)$ is noteworthy. According to (8), the allowable values of $W(Z)$ lie within the segment $0 \le Z \le Z_0$ with $W(Z_0) = 0$. Hence, condition (6) relative to the self-similar variables should be written as

$$\int_0^{Z_0} W^{1/n} dy = 1. \tag{9}$$

Substituting (8) into (9) yields

$$Z_0 = \left[\frac{2n(n+1)}{n-1}\right]^{1/(n+1)} \left(\frac{1}{\Psi_1}\right)^{\frac{n-1}{n+1}},$$

$$\Psi_1 = \int_0^1 (1-y^2)^{\frac{1}{n-1}} dy.$$

The existence of the finite value Z_0 at which the function $W(Z)$ vanishes defines its specific behavior in the vicinity of $Z = Z_0$, since, according to (8), $d^kW/dZ^k = 0$ ($k = 1, 2, 3, \ldots$) at $Z = Z_0$.

Thus, an internal electromagnetic field in a superconductor with a power *I–V* curve decays within a finite region. In dimensional form, the decay is described by the expressions

$$\begin{gathered}E(x,t) = E_a(t)(1 - x^2/x_0^2)^{\frac{n}{n-1}},\\ B(x,t) = B_{0,i}\left[1 - \frac{1}{x_0\Psi_1}\int_0^x \left(1 - \frac{y^2}{x_0^2}\right)^{\frac{1}{n-1}} dy\right],\end{gathered} \tag{10}$$

where

$$\begin{gathered}E_a(t) = E_C\left(\frac{t_n}{t+t_0}\right)^{\frac{n}{n+1}}, \quad t_n = \frac{n-1}{2n(n+1)}\frac{t_x}{\Psi_1^2},\\ x_0(t) = \frac{B_{0,i}}{\mu_0 J_C \Psi_1}\left[\frac{E_C}{E_a(t)}\right]^{\frac{1}{n}}.\end{gathered} \tag{11}$$

This solution is determined up to the constant t_0. We will find it by setting the electric field strength on the superconductor surface, $E_{0,i} = E_0(0)$, at a time instant t_i. As a result, we have

$$t_0 = t_e - t_i, \quad t_e = t_n(E_C/E_{0,i})^{(n+1)/n},$$

where t_e is the characteristic time of magnetic flux relaxation.

Figure 1 shows the distribution of the magnetic field induction inside the domain of magnetization versus the exponent of the *I–V* curve. The region of the smooth transition from the disturbed to undisturbed state is seen to be negligible. Moreover, the condition $n > 10$ adequately describes the states where the image current relaxation is characterized by the almost linear distribution of the magnetic field induction inside the superconductor. This observation substantiates the use of simplified approaches to analyzing relaxation in superconductors (for example, that suggested in [5]).

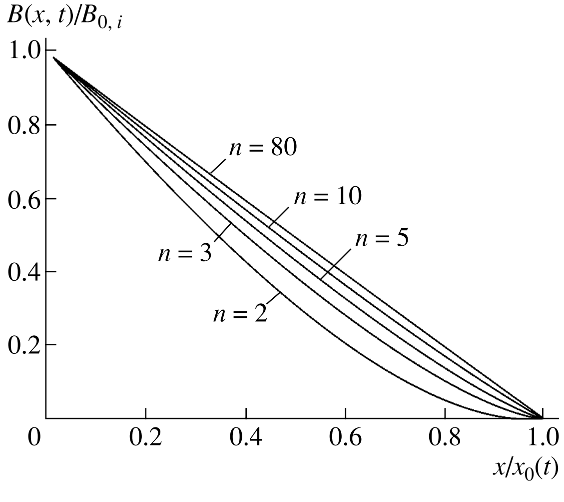


Fig. 1. Self-similar distribution of the magnetic field in the domain of magnetization as a function of the exponent *n* of *I–V* characteristic (1).

Thus, the relaxation in a superconductor with the power *I–V* curve generates the particular state. It features a finite extent of the domain of magnetization with the mobile boundary, at which the values of induced electric and magnetic fields match smoothly with their unperturbed values:

$$\partial^k E/\partial x^k = 0, \quad \partial^k B/\partial x^k = 0, \quad k = 1, 2, 3, \ldots.$$

The image current penetrates into the material with a finite rate

$$\frac{dx_0}{dt} = \frac{B_{0,i}}{(n+1)\mu_0 J_C \Psi_1}\frac{E_a(t)}{E_C}.$$

This solution allows one to judge whether creep-related processes in superconductors with the logarithmic current dependence of the potential barrier are described correctly [3, 4]. As was indicated [5], the estimator [3] for the magnetic moment of a superconducting plate in the case of incomplete image current penetration is in error. Moreover, as follows from the above solution, the induced and unperturbed values of the electric and magnetic fields match smoothly at the mobile boundary of the domain of magnetization. It is easy to check that this matching condition must also be met in the problem stated in [3]. However, the solution found in [3] does not satisfy this condition. The mistake made in [3] is that unknown constants appearing in the solution are found not by (9) but with the additionally introduced condition

$$J(x,t)_{|x=0,\, t=0|} = J_C.$$

Because of this, the self-similar solution found in [3, 4] fails not only to estimate the magnetic moment of the plate but also to calculate the magnetic field during the relaxation process.

Using the above self-similar solution, we will find the time variation of the magnetic moment of a superconducting plate with a half-width *a* for states preceding the complete penetration of the image current into the plate. According to (10) and (11), we have

$$-\mu_0 M(t)/B_{0,i} = 1 - \varphi_n x_0(t)/a$$

or

$$M(t) = M(t_i) + M_1\left(\frac{t - t_i + t_e}{t_n}\right)^{\frac{1}{n+1}} \times \left[1 - \left(\frac{t_e}{t - t_i + t_e}\right)^{\frac{1}{n+1}}\right]\varphi_n, \quad t \geq t_i.$$

Here, $M(t_i)$ is the magnetic moment of the plate at $t = t_i$, $M_1 = B_{0,i}^2/(\mu_0^2 a J_C \Psi_1)$,

$$\varphi_n = 1 - \frac{1}{\psi_1}\int_0^1 \psi(\eta)d\eta, \text{ and } \psi(\eta) = \int_0^\eta (1 - y^2)^{\frac{1}{n-1}}dy.$$

From these expressions, it follows that the relaxation of the magnetic moment of a superconductor in the case when the image current partially occupies its cross section depends on the dynamics of the mobile boundary of the magnetized domain. The fact that the magnetization of a superconductor is directly related to the penetration depth of the magnetic flux is of particular importance, since such a conclusion, which is drawn for the first time, allows one, based on relevant experimental data, to directly analyze the finiteness of creep-related processes.

Differentiating $M(t)$ with respect to time yields

$$\frac{dM}{dt} = \frac{M_1\varphi_n}{(n+1)(t - t_i + t_e)}\left(\frac{t - t_i + t_e}{t_n}\right)^{\frac{1}{n+1}}.$$

From this formula, we can estimate typical values of the magnetic relaxation rate. With $t - t_i \ll t_e$, the magnetic moment relaxes virtually with a constant rate:

$$\frac{dM}{dt} \approx \frac{M_1\varphi_n}{(n+1)t_e}\left(\frac{t_e}{t_n}\right)^{\frac{1}{n+1}}.$$

If $t - t_i \gg t_e$, the magnetic moment relaxes with a monotonically decreasing rate:

$$\frac{dM}{dt} \approx \frac{M_1\varphi_n}{(n+1)(t - t_i)}\left(\frac{t - t_i}{t_n}\right)^{\frac{1}{n+1}}, \quad t - t_i \gg t_e$$

or

$$\frac{dM}{d\ln t} \approx \frac{M_1\varphi_n}{n+1}\frac{t}{t - t_i}\left(\frac{t - t_i}{t_n}\right)^{\frac{1}{n+1}},$$

that is, for long times, the relaxation is described by a logarithmic function up to a factor $(t/t_n)^{1/(n+1)}$, as was expected.

This complete analytical solution to the problem of magnetic flux relaxation inside a superconductor with a power *I–V* characteristic gives a clear qualitative description of attendant effects. First of all, it shows that magnetic relaxation depends primarily on the behavior (decay) of the electric field on the surface. Specifically, all other things being equal, an increase in the exponent *n* cuts the decay time of $E_a(t)$. Accordingly, the penetration rate of the image current and the relaxation rate of the magnetic moment of the superconductor decrease, since $dM/dt \sim dx_0/dt \sim E_C/(n+1)$. In the limit $n \longrightarrow \infty$, we have $E_a \longrightarrow 0$, $dx_0/dt \longrightarrow 0$, $dM/dt \longrightarrow 0$, and $x_0 \longrightarrow B_{0,i}/\mu_0 J_C$. These relationships are easy to explain. They immediately follow from an increase in the differential resistance of the superconductor: $\rho_d \longrightarrow \infty$ for $n \longrightarrow \infty$. As is known, this is also the reason for the radical difference in the relaxation times in low- and high-temperature superconductors.

The results obtained with the self-similar approximation were verified by comparing with numerical calculations. Figure 2 shows the time variation of the electric field strength on the surface of the Nb–Ti low-temperature superconductor for various t_i. The associated parameters were assumed to be $E_C = 10^{-4}$ V/cm, $J_C = 4 \times 10^9$ A/m^2, and $n = 80$. The continuous curves depict a numerical solution to the problem given by (3), (5), and (6). The distribution of the electric field $E_0(x)$ at $t = t_i$ was set based on the image current penetration calculated by numerically integrating Eqs. (3) subject to

$$E(x, 0) = 0, \quad \partial E/\partial x(0, t) = -dB/dt, \quad E(x_0, t) = 0 \tag{12}$$

for the case when the magnetic induction of the surface rises with a rate $dB/dt = 1$ T/s.

The curves depicted, together with similar calculations made for high-temperature superconductors, demonstrate that the analytical solution suggested makes it possible to describe adequately the entire relaxation process in a superconductor with a power *I–V* characteristic and not only its asymptotics at long times.

RELAXATION OF THE ELECTROMAGNETIC FIELD INSIDE A SUPERCONDUCTOR WITH AN EXPONENTIAL *I–V* CURVE

Let a superconductor have the *I–V* characteristic in the form (2). Stated in the dimensionless variables

$$e = E/E_C, \quad X = x/L_x, \quad \tau = t/t_x,$$

$$L_x = B_{0,i}/(\mu_0 J_\delta), \quad t_x = B_{0,i}^2/(\mu_0 J_\delta E_C),$$

problem (4)–(6) is reduced to integrating the equation

$$\partial e/\partial\tau = e\partial^2 e/\partial X^2 \tag{13}$$

with the boundary condition

$$\partial e/\partial X(0, \tau) = 0 \tag{14}$$

and the condition of conservation of the induced current with a density $j = 1 + \delta \ln e$ ($j = J/J_C$, $\delta = J_\delta/J_C$):

$$\int_0^{X_0} j dX = \delta. \tag{15}$$

The latter condition assumes that there exists a finite-size domain of magnetization and that $e(X_{0,\tau}) = 0$ at the mobile boundary.

Problem (13)–(15) falls outside the scope of the class of self-similar problems. Therefore, its solution will be described in greater detail than before.

The desired distribution of the electric field will be sought in the form

$$e = T(\tau)W(Z), \quad Z = X/X_0(\tau).$$

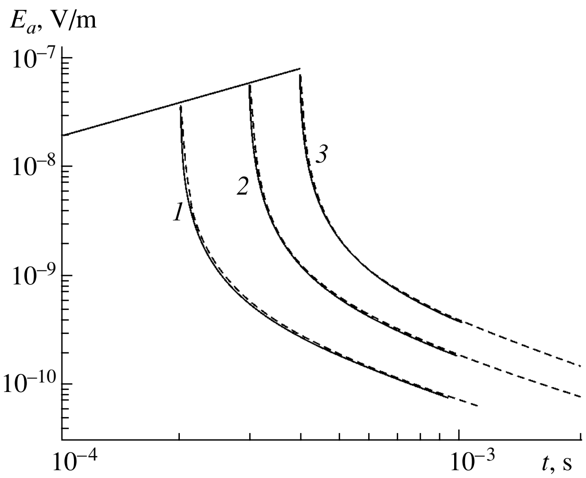


Fig. 2. Time variation of the electric field strength on the surface of Nb–Ti superconductor with the power characteristic. The continuous curves depict the numerical solution to the problem of introduction and relaxation of image current; dashed curves show the self-similar solution. t_i = (*1*) 2×10^{-4}, (*2*) 3×10^{-4}, and (*3*) 4×10^{-4} s.

Then, Eq. (13) may be recast as

$$\frac{dT}{d\tau}W - \frac{dX_0}{d\tau}\frac{Z}{X_0}T\frac{dW}{dZ} = \frac{T^2}{X_0^2}W\frac{d^2W}{dZ^2}. \tag{16}$$

Since the functions $T(\tau)$ and $X_0(\tau)$ are assumed to be arbitrary, we require that they satisfy the equation

$$\frac{dX_0}{d\tau}\frac{T}{X_0} = -\alpha\frac{dT}{d\tau},$$

where α is a constant to be determined.

This equation provides the relationship between the functions $T(\tau)$ and $X_0(\tau)$,

$$X_0(\tau) = \gamma T^{-\alpha}(\tau), \tag{17}$$

which can be determined up to an arbitrary factor γ.

In view of (17), the variables in (16) can be separated; that is,

$$\frac{T^2}{X_0^2}\frac{dT}{d\tau} = \frac{W\frac{d^2W}{dZ^2}}{W + \alpha Z\frac{dW}{dZ}} = -\beta = \text{const.} \tag{18}$$

From (17) and (18), it is easy to find (up to the constant τ_0) that

$$T(\tau) = T_1(\tau)/\beta, \quad X_0(\tau) = \gamma_1(\tau + \tau_0)^{\frac{\alpha}{1+2\alpha}},$$

where

$$T_1(\tau) = \gamma_1^2/\left[(1 + 2\alpha)(\tau + \tau_0)^{\frac{1}{1+2\alpha}}\right],$$

$$\gamma_1 = [\gamma\beta^{\alpha}(1+2\alpha)^{\alpha}]^{\frac{1}{1+2\alpha}}.$$

In terms of the new variable $W_1 = W/\beta$, the desired solution takes the form

$$e = T_1 W_1 = \gamma_1^2 W_1 \Big/ \left[(1+2\alpha)(\tau+\tau_0)^{\frac{1}{1+2\alpha}}\right],$$

which does not depend on β. The problem is then reduced to integrating the equation

$$W_1\frac{d^2W_1}{dZ^2} + \alpha Z\frac{dW_1}{dZ} + W_1 = 0 \qquad (19)$$

with the boundary conditions

$$dW_1/dZ(0) = 0, \quad W_1(1) = 0.$$

Let us study the variation of $W_1(Z)$. According to (19) and in view of the boundary condition at $Z = 1$, the first and all higher order derivatives of the invariant $W_1(Z)$ vanish at the right-hand boundary. In addition, Eq. (19) yields two integral equalities

$$1 + \alpha\int_0^1 \frac{Z}{W_1}\frac{dW_1}{dZ}dZ = 0,$$

$$\int_0^1 \left(\frac{dW_1}{dZ}\right)^2 dZ + (\alpha-1)\int_0^1 W_1 dZ = 0.$$

Since physically $W_1 \geq 0$ and $dW_1/dZ \leq 0$, these equalities are valid only if $0 < \alpha < 1$.

Thus, after the integration of Eq. (19), the desired distribution of the electromagnetic field inside a superconductor with an exponential *I–V* characteristic is determined up to the three constants α, γ, and τ_0. Formally, α and γ can be found from condition (15). However, it is impossible to directly write the solution in the self-similar approximation from this condition. In physical terms, this means the violation the self-similarity conditions in the variation of the sought solutions throughout the relaxation. At the same time, as the process develops, the self-similar state may be asymptotically approached. This allows the determination of the constants α and γ. To this end, we simplify *I–V* characteristic (2) by expanding it into the exponential series. Using the linear term of this expansion, we will seek α and γ for the *I–V* curve in the form $J \approx J_C(E/E_0)^{\delta}\xi$, where $E_0 = E_C\exp(-1/\delta)$ and $\xi < 1$ is a correction factor for this linear approximation. To define this factor, we take advantage of the condition $E = E_C$ at $J = J_C$. Then, $\xi = (E_0/E_C)^{\delta}$. Within this approximation, condition (15) transforms into

$$\delta = \gamma_1^{1+2\delta}\frac{(\tau+\tau_0)^{\frac{\alpha-\delta}{1+2\alpha}}}{(1+2\delta)^{\delta}}\int_0^1 W_1^{\delta}(Z)dZ.$$

Taking into account that the right of this expression must not depend on time, we may write

$$\alpha = \delta, \quad \gamma_1 = \left[\delta(1+2\delta)^{\delta}\Big/\int_0^1 W_1^{\delta}(Z)dZ\right]^{\frac{1}{1+2\delta}}. \qquad (20)$$

Since $\delta \ll 1$, γ_1 can be approximated by the expression

$$\gamma_1 \approx [\delta(1+2\delta)^{\delta}]^{\frac{1}{1+2\delta}}. \qquad (21)$$

To determine $W_1(Z)$, we transform differential equation (19) into an equivalent integral equation. Straightforward mathematics gives

$$W_1(Z) = \frac{1-Z^2}{2} - \alpha\int_1^Z dx\int_0^x \frac{y}{W_1}\frac{dW_1}{dy}dy.$$

Since α is small, the desired invariant $W_1(Z)$ can be approximated with a good accuracy by the zero-order solution of the above integral equation:

$$W_1(Z) \approx (1-Z^2)/2.$$

Therefore, in terms of the approximations used, the decay of an electromagnetic field in a superconductor with an exponential *I–V* characteristic is described by the expressions

$$\begin{aligned} E(x,t) &= E_a(t)(1-x^2/x_0^2), \\ B(x,t) &= B_{0,i} - \mu_0 J_C x - \mu_0 J_\delta x \ln[E_a(t)/E_C] \\ &\quad - \mu_0 J_\delta x[(1+x/x_0)\ln(1+x/x_0) \\ &\quad - (1-x/x_0)\ln(1-x/x_0) - 2x/x_0], \end{aligned} \qquad (22)$$

where

$$E_a(t) = \frac{E_C\gamma_1^2}{2(1+2\delta)}\left(\frac{t_x}{t+t_0}\right)^{\frac{1}{1+2\delta}}.$$

In this case, according to the self-similar approximation, the coordinate of the mobile boundary varies as

$$x_{a,0}(t) = \frac{B_{0,i}\gamma_1}{\mu_0 J_\delta}\left(\frac{t+t_0}{t_x}\right)^{\frac{\delta}{1+2\delta}}. \qquad (23)$$

At the same time, according to (15), this coordinate grows by the law

$$x_0(t) = \frac{B_{0,i}}{\mu_0 J_C[1+\delta\ln[E_a(t)/E_C] + 2\delta(\ln 2 - 1)]}. \qquad (24)$$

The constant t_0 is defined by appropriately setting the electric field strength $E_{0,i}$ on the superconductor surface at $t = t_i$. Then,

$$t_0 = t_e - t_i, \quad t_e = t_x\left[\frac{\gamma_1^2}{2(1+2\delta)}\frac{E_C}{E_{0,i}}\right]^{1+2\delta}.$$

In view of the approximate solution obtained, the magnetic moment of a superconducting plate with a half-width a relaxes by the law

$$-\frac{\mu_0 M(t)}{B_{0,i}} = 1 - \frac{x_0(t)}{a} + \frac{\mu_0 J_C x_0^2(t)}{2aB_{0,i}}$$
$$+ \delta\frac{\mu_0 J_C x_0^2(t)}{4aB_{0,i}}\left[\ln\frac{E_a(t)}{E_C} + \frac{3}{4}\ln 2 - \frac{10}{9}\right].$$

Thus, as in the case of superconductors with a power I–V characteristic, the relaxation of the magnetic moment of a superconductor with the exponential characteristic directly depends on the coordinate of the mobile boundary of the magnetization domain. In its turn, the position of the mobile boundary is influenced by the electric field on the surface. However, unlike superconductors with the power characteristic, the relaxation in superconductors with the exponential characteristic exhibits three stages. At $t - t_i \ll t_e$, the initial stage of relaxation takes place. Then, the relaxation passes to the so-called quasi-self-similar regime. The duration of this stage depends primarily on the superconducting properties: in low-temperature superconductors, it is much longer than in high-temperature ones. At both stages, the coordinate of the mobile boundary varies almost as in the self-similar regime. Therefore, one can use the self-similar approximation and determine the characteristic rates of the relaxation at the three stages. For the initial stage ($t - t_i \ll t_e$), the rate is practically constant and can be estimated by the formula

$$\frac{dM}{dt} \sim \frac{dx_{a,0}}{dt} \approx \frac{B_{0,i}\gamma_1}{\mu_0 J_\delta}\frac{\delta}{1+2\delta}\frac{1}{t_e}\left(\frac{t_e}{t_x}\right)^{\frac{\delta}{1+2\delta}}.$$

The relaxation rate in the quasi-self-similar state drops with time. With $t - t_i \gg t_e$, the decay of the magnetic moment approaches the logarithmic law

$$\frac{dM}{dt} \sim \frac{dx_{a,0}}{dt} \approx \frac{B_{0,i}\gamma_1}{\mu_0 J_\delta}\frac{\delta}{1+2\delta}\frac{1}{t-t_i}\left(\frac{t-t_i}{t_x}\right)^{\frac{\delta}{1+2\delta}}.$$

At the third stage, the relaxation process falls outside the scope of the self-similar approach: the magnetic flux penetrates with a rate that is higher than that following from the self-similar approximation.

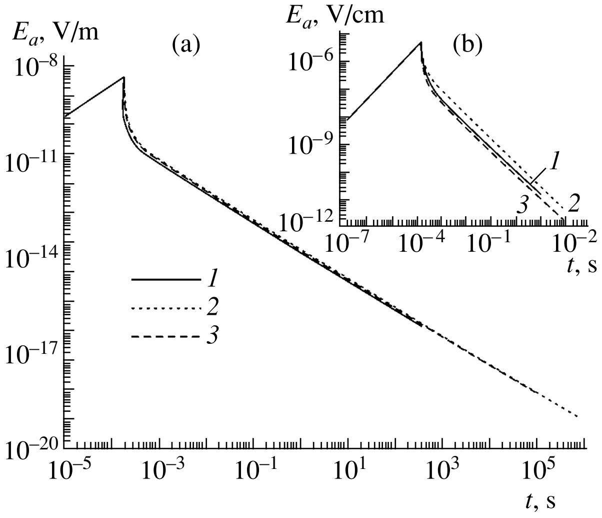


Fig. 3. Time variation of the electric field strength on the surface of (a) Nb–Ti and (b) YBCO superconductors with the exponential characteristic. (*1*) Numerical solution to the problem of introduction and relaxation of image current; (*2*) self-similar solution described by formulas (20), (22), and (23); and (*3*) solution described by formulas (21), (22), and (24).

To verify the analytical solutions obtained, we compare them with the computational results for the relaxation of the electric field strength on the surface of the Nb–Ti low-temperature superconductor (Fig. 3a). The parameters of the I–V characteristic were taken to be as follows: $E_C = 10^{-4}$ V/m, $J_C = 4 \times 10^9$ A/m^2, and $J_\delta = 4 \times 10^7$ A/m^2. The external magnetic field increased at a rate of $dB/dt = 1$ T/s and ceased to grow at $t = 2 \times 10^{-4}$ s. The continuous curves depict the numerical solution to the problem of introduction and relaxation of the image current; the dashed and dotted curves show the analytical solutions that use the value of $E_{0,i}$ determined by solving the problem given by (4) and (12). Figure 3b compares the numerical and analytical results obtained for the YBCO high-temperature superconductor with the parameters

$$E_C = 10^{-8}\ \text{V/cm}, \quad J_C = 1.15 \times 10^5\ \text{A/cm}^2,$$
$$J_\delta = 8.686 \times 10^3\ \text{A/cm}^2. \tag{25}$$

It was assumed that, at $t < 2 \times 10^{-4}$ s, the superconductor is in an external magnetic field that grows with a rate $dB/dt = 7.5 \times 10^{-3}$ T/s and ceases to grow at $t = 2 \times 10^{-4}$ s.

THE EFFECT OF THE SHAPE OF THE *I*–*V* CHARACTERISTIC ON THE RELAXATION PROCESS

The solutions obtained show that the relaxation in hard superconductors with I–V characteristics of both types (see (1) and (2)) proceeds in a similar way. The spatial profile of the electric field (which decreases with time) is described by a parabolic law, while the distribution of the magnetic field is almost linear. However, the equivalence in the description of the electrodynamic states of superconductors with power and exponential characteristics depends on the exponent n. In general, the difference in the behavior of the systems grows with decreasing n. If $n < 10$, creep greatly affects the distribution of the magnetic field throughout the domain of magnetization of a superconductor with the power characteristic and this distribution deviates markedly from linearity.

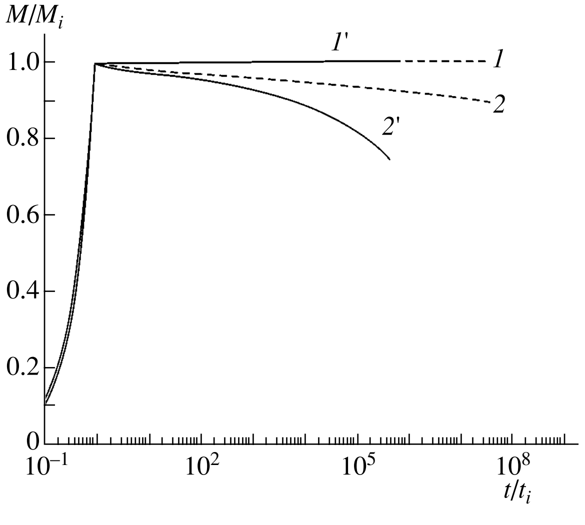


Fig. 4. Variation of the magnetic moment of superconducting plates with the different characteristics upon the introduction and relaxation of image current. (*1*) Nb–Ti with characteristic (1): $E_C = 0.6522 \times 10^{-2}$ V/m, $J_C = 3.5 \times 10^9$ A/m^2, and $n = 87.5$; (*1*') Nb–Ti with characteristic (2): $E_C = 0.6522 \times 10^{-2}$ V/m, $J_C = 3.5 \times 10^9$ A/m^2, and $J_\delta = 4 \times 10^7$ A/m^2; (*2*) YBCO with characteristic (1): $E_C = 1.778 \times 10^{-2}$ V/m, $J_C = 2 \times 10^9$ A/m^2, and $n = 23$; and (*2*') YBCO with characteristic (2) and the parameters given by (25).

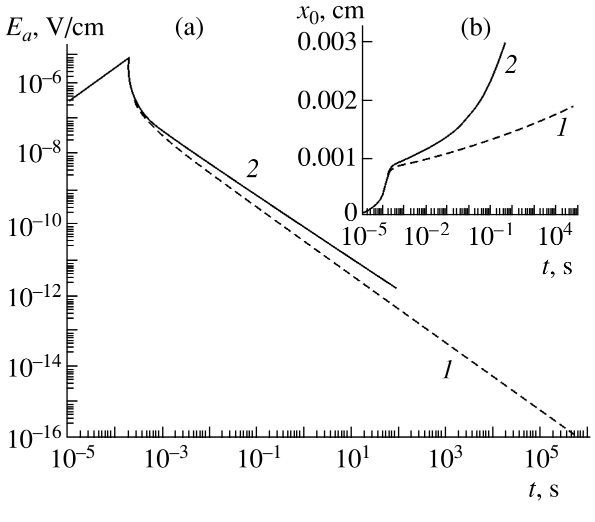


Fig. 5. Effect of the *I–V* characteristic of high-temperature superconductors on the (a) electric field strength on the surface and (b) dynamics of the mobile boundary of magnetization upon the introduction and relaxation of image current. (*1*) Superconductor with characteristic (1): $E_C = 1.778 \times 10^{-2}$ V/m, $J_C = 2 \times 10^9$ A/m^2, and $n = 23$ and (*2*) superconductor with characteristic (2) and the parameters given by (25).

In addition, the qualitative difference between the relaxation processes is also related to the variation of the electric field on the surface of the superconductor, since the surface state directly influences the travel of the mobile boundary of magnetization. The relaxation dynamics also depends on the type of the *I–V* characteristic. In the case of the power characteristic, the relaxation is a two-stage process, these stages being approximated well by the self-similar solution. In the case of the exponential characteristic, the process has three stages. The first two are qualitatively similar to those observed in superconductors with the power characteristic. However, the third stage of the relaxation in superconductors with the exponential characteristic cannot be described in terms of self-similar states. At this stage, the velocity of the mobile boundary of magnetization exceeds the rate of its penetration into a superconductor with the power characteristic.

To illustrate the aforesaid, Fig. 4 compares the magnetization curves for a superconducting plate with a half-width *a*. The calculation was performed for Nb–Ti ($a = 10^{-5}$ m) and YBCO ($a = 5 \times 10^{-6}$ m) superconductors with power and exponential characteristics. The parameters of the superconductors were selected so that the power characteristic touched the exponential characteristic at a given point (J_C, E_C). This takes place with $n = J_C/J_\delta$. Such a selection of the parameters of the characteristics makes it possible to contrast the calculated values of $E(J)$. For the superconductors under study, these are given in the caption to Fig. 4. The current values of the magnetic moments were normalized to those reached at the instant the magnetic field is switched off ($t_i = 2 \times 10^{-4}$ s). For $t \le t_i$, the magnetic moment of the plate was found by numerically evaluating the penetration of the image current induced by a varying magnetic field into a superconductor. At $t > t_i$, the magnetic moment was found in the self-similar approximation. The rate of rise of the external magnetic field was set equal to $dB/dt = 1$ T/s for Nb–Ti and 7.5×10^{-3} T/s for YBCO. The formation of the electrodynamic states in YBCO is shown in Fig. 5.

Thus, power and exponential *I–V* characteristics basically describe the relaxation dynamics of electrodynamic states in a superconductor with creep in a different manner. The smaller is the exponent of the power characteristic, the greater are quantitative and qualitative discrepancies. The reason is that the curves variously run in the range of low electric fields. As follows from (1) and (2), at the same value of the current density, the differential resistance of a superconductor with the exponential characteristic is always lower than that of a superconductor with the power characteristic when $E < E_C$. Since the rate of electric field decay grows with the differential resistance, the field on the surface of the former will fall less rapidly than on the surface of the latter. Because of this, when the same current is introduced into superconductors with power or exponential *I–V* characteristics, the diffusion of the mobile boundary of magnetization proceeds at a higher rate in the latter, all other things being equal. Since the above-mentioned difference in differential resistances is greater in high-temperature superconductors, their behavior is more sensitive (on both the qualitative and quantitative level) to the shape of the *I*–*V* characteristic (Figs. 4, 5).

ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research (project no. 01-02-16252a).

REFERENCES

1. M. R. Beasely, R. Labusch, and W. W. Webb, Phys. Rev. **181**, 682 (1969).
2. C. J. van der Beek, G. J. Nieuwenhuys, and P. H. Res, Physica C **185–189**, 2241 (1991).
3. V. M. Vinokur, M. V. Feigel’man, and V. B. Geshkenbein, Phys. Rev. Lett. **67**, 915 (1991).
4. G. Blatter, M. V. Feigel’man, V. B. Geshkenbein, *et al.*, Rev. Mod. Phys. **66**, 1125 (1994).
5. H. G. Schnack and R. Griessen, Phys. Rev. Lett. **68**, 2706 (1992).
6. A. Gurevich and H. Küpfer, Phys. Rev. B **48**, 6477 (1993).
7. E. H. Brandt, Phys. Rev. B **54**, 4246 (1996).
8. M. N. Wilson, *Superconducting Magnets* (Oxford University Press, London, 1983; Mir, Moscow, 1985).
9. A. V. Gurevich, R. G. Mints, and A. L. Rakhmanov, *Physics of Composite Superconductors* (Nauka, Moscow, 1987).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1304–1306. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 85–87.
Original Russian Text Copyright © 2003 by Levin, Tatarintsev, Kostsova, Kostsov.

SOLID-STATE ELECTRONICS

Activation of a Semiconductor Surface by a Pulsed Magnetic Field

M. N. Levin, A. V. Tatarintsev, O. A. Kostsova, and A. M. Kostsov
Voronezh State University, Universitetskaya pl. 1, Voronezh, 394693 Russia
e-mail: levin@lev.vsu.ru

Received December 24, 2002

Abstract—The possibility of semiconductor surface activation, which shows up as a long-term increase in the adsorption capacity in response to a short exposure to a pulsed magnetic field, is demonstrated for the first time. Magnetic-field-induced surface activation is studied on silicon, germanium, and gallium arsenide crystals. The effect revealed extends the capabilities of thin-film growth on the semiconductor surface. © *2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

It has been reported that the exposure of semiconductor crystals to weak pulsed magnetic fields (PMFs) causes long-term nonmonotonic changes in their structure and structure-dependent properties. The near-surface layers of semiconductors, especially silicon, are highly sensitive to such a treatment. The following PMF-induced effects have been observed: the long-term nonmonotonic relaxation of the silicon conductivity at the Si–SiO_2 interface [1], the nonmonotonic variation of the lattice parameter near the silicon surface [2], the gettering of surface defects and the recrystallization of amorphized surface layers [3], the enrichment of surface layers by oxygen being released from the bulk [4], and the change in the sorption properties of the silicon surface [5]. However, systematic research on the kinetics of these processes in different semiconductor materials has not been carried out and a correlation between them has not been established.

The aim of this study is to find ways of semiconductor surface activation to make advanced technologies feasible.

EXPERIMENTAL

The object of study was Czochralski-grown plane-parallel polished (111) Si wafers doped by phosphorous to $\approx 5 \times 10^{15}$ cm^{-3}, (100) Ge wafers doped by antimony to $\approx 2 \times 10^{16}$ cm^{-3}, and (110) GaAs wafers doped by tellurium to $\approx 10^{17}$ cm^{-3}.

The samples were exposed to a PMF with an amplitude $B_0 = 0.4$ T, pulse duration $\tau = 1–4 \times 10^{-5}$ s, and a pulse repetition rate $f = 50$ Hz. The treatment time was varied from 30 to 60 s. The samples were treated and stored at room temperature. The surface topology, microstructure, and sorption properties of the samples were examined.

The surface topology was studied by scanning tunnel microscopy (STM), which images the sample surface and allows the surface parameters to be measured with the atomic resolution. The samples were freed of surface oxide by chemical treatment, rinsed in deionized water, and dried. Surface imaging was carried out with the SKAN-8 scanning tunnel microscope at the Advanced Technology Center (Moscow State University) in the constant tunnel current mode.

Figure 1 shows the variation of the surface topology for the silicon crystal under the PMF action. At the first stage of aging (Fig. 1b), the surface becomes less regular compared with the original surface (Fig. 1a): it becomes uneven, patterned, and rough. In ≈200 h after the PMF action, these changes are the most pronounced. With an increase in the aging time, the surface smoothes out (Fig. 1c) and in ≈400 h becomes even smoother than the initial one. In the caption to Fig. 1, we give the average surface roughness parameter *Rz* (the ten-point height [6]) measured over two mutually perpendicular cross sections 10 nm from each other.

The modifications of the surface topology and cross-sectional relief due to the PMF treatment are accompanied by the variation of the sorption properties of the semiconductor surfaces. The adsorption capacity of the crystal surface exposured to the PMF was studied by microwave spectroscopy, which is highly sensitive to the presence and state of water molecules in the studied medium [7]. We recorded the reflection spectra in the frequency range from 26 to 37 GHz.

Sorption experiments were carried out as follows. The sample cut from the central part of the wafer was placed in a measuring cell of a wave duct. The sample size was matched to the internal size of the duct. The cell with the sample was urged by dry air for 30 min, and then the reflection spectra were measured (Fig. 2, curves *1*). Air was dried by passing through a column filled with silica gel. Repeat purging by dry air did not

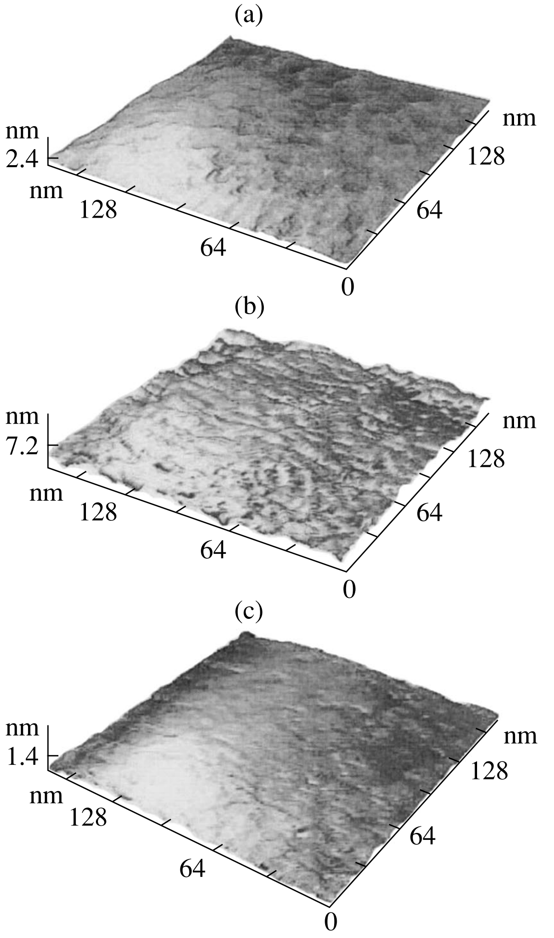


Fig. 1. STM images of the Cz–Si crystal surface: (a) as-grown (the roughness parameter *Rz* is 1.34 nm), (b) 200, and (c) 400 h after the exposure to the PMF (*Rz* is 5.81 and 0.92 nm, respectively).

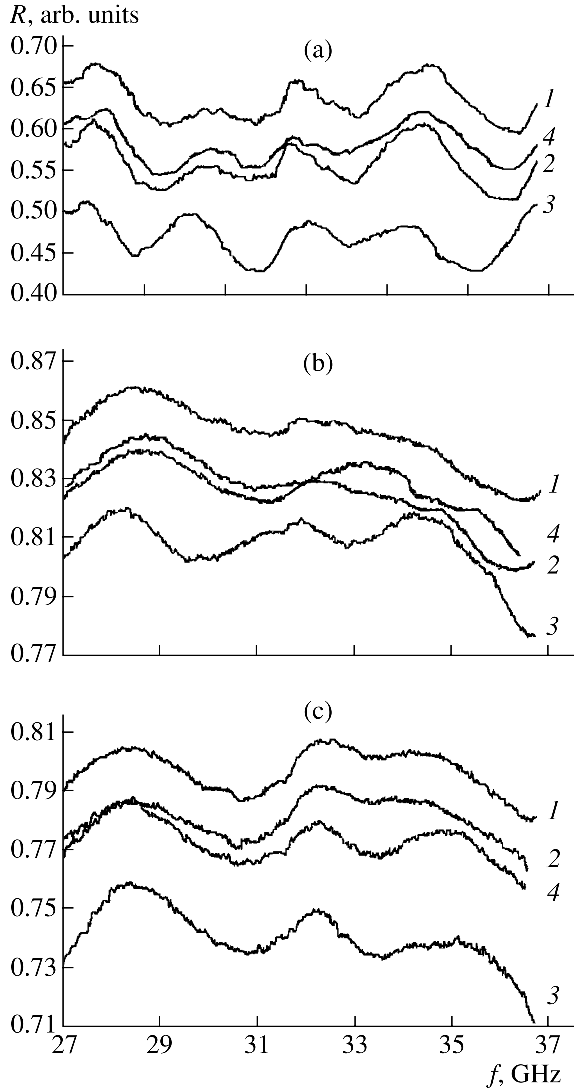


Fig. 2. Microwave radiation reflection spectra for (a) Si, (b) Ge, (c) and GaAs. (*1*, *2*) Before exposure to the PMF (purging by dry and humid air, respectively); (*3*) 200, 150, and 120 h after the exposure for Si, Ge, and GaAs, respectively (purging by humid air); and (*4*) 350 h after the exposure (purging by humid air).

change the spectra. Next, the cell was purged by air of 100% humidity (Fig. 2, curves *2*). To reach 100% humidity, air was passed through a bubbler where it was saturated by water vapor. The no-condensation conditions of purging the measuring cell were provided. The humidity was measured by a hygrometer.

Purging by humid air decreased the coefficient of microwave radiation reflection, which appeared as a shift of the spectra (almost without distortion) along the intensity scale. The shift depended on the wet purging time and reached a maximum at a purging time of ≈5 min, when the moisture adsorption capacity of the surface became the highest.

The reversibility of moisture adsorption–desorption processes on the semiconductor surface made it possible to trace the time variation of the sorption properties by multiply repeating the cycle with the same sample.

The spectra of the as-prepared samples did not change with time. The short PMF treatment resulted in the long-term variation of the reflection spectra (Fig. 2; curves *3*, *4*), indicating a change in the sorption capacity of the surface.

Common to the PMF-induced changes in the microwave radiation reflection spectra for all the crystals is the nonmonotonic variation of the spectral shift ΔR_S (Fig. 3). First, ΔR_S increases markedly, reaching a maximum in several hundreds of hours, decreases, and takes the steady-state value in ≈350 h after the PMF exposure.

At the same time, there are differences in the behavior of samples exposed to the PMF: first, silicon and germanium have a latent period (≈50 h), over which the spectra do not change, whereas GaAs does not have such a period; second, the time over which ΔR_S reaches

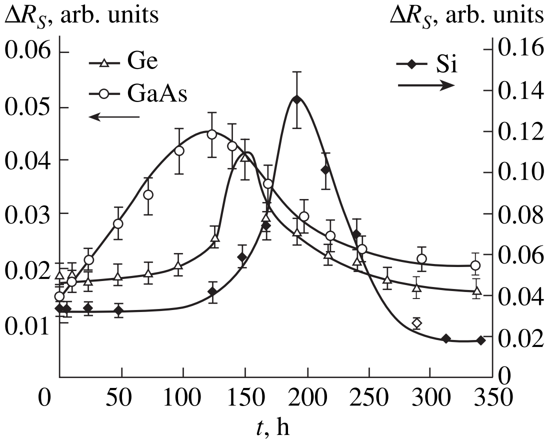


Fig. 3. Time variation of the microwave reflection coefficient ΔR_S for the Si, Ge, GaAs crystals after exposure to the PMF.

a maximum is ≈200 h for silicon, ≈150 h for germanium, and ≈120 h for gallium arsenide; third, the scales of ΔR_S variation are different.

Our experimental data indicate that the PMF action results in the nonmonotonic variation of the chemical activity of the crystal surfaces. The increase in the adsorption capacity is changed to the steady reduction of the surface capability to moisture adsorption either below (silicon and germanium) or to (gallium arsenide) the initial value. The modification of the sorption properties lasts several hundreds of hours at room temperature.

It is noteworthy that, as a result of the PMF action, the sorption capacity of silicon, germanium, and gallium arsenide increases more than four, two, and three times, respectively.

The reduction of the silicon surface hydrophobicity under the PMF action was used to form thin films of metal oxides and sulfides by thermally spraying aqueous solutions of thiocarbamide complexes [5].

Thus, STM and microwave spectroscopy studies of the semiconductor crystals exposed to a PMF revealed the nonmonotonic long-term variation of surface physicochemical properties. The surface adsorption capacity first increases and then declines to the initial value (for gallium arsenide) or to a level below the initial value (for silicon and germanium crystals). The increase in the sorption capacity of the semiconductor crystal surface goes in parallel with the modification of its topology. This effect is observed at room temperature and shows up most vividly several hundreds of hours after the treatment.

The PMF-induced increase in the chemical activity of the semiconductor surface may be applied to improve different process steps, e.g., to reduce the temperature of oxidation and increase the rate of chemical reactions on the semiconductor surface, in thin-film technology.

REFERENCES

1. A. L. Danilyuk and A. I. Nareĭko, Poverkhnost, No. 9, 27 (1996).
2. N. V. Kukushkin, S. N. Postnikov, Yu. A. Terman, *et al.*, Zh. Tekh. Fiz. **55**, 2083 (1985) [Sov. Phys. Tech. Phys. **30**, 1227 (1985)].
3. M. N. Levin and B. A. Zon, Zh. Éksp. Teor. Fiz. **111**, 1373 (1997) [JETP **84**, 760 (1997)].
4. M. N. Levin and B. A. Zon, Mater. Res. Soc. Symp. Proc. **583**, 278 (2000).
5. M. N. Levin, V. N. Semenov, and A. V. Naumov, Pis’ma Zh. Tekh. Fiz. **27** (7), 35 (2001) [Tech. Phys. Lett. **27**, 279 (2001)].
6. A. I. Yakushev, L. N. Vorontsov, and N. M. Fedotov, *Interchangeability, Standardization, and Technical Measurements* (Mashinostroenie, Moscow, 1986).
7. A. M. Kostsov, Izv. Ross. Akad. Nauk, Ser. Fiz. **64**, 1712 (2000).

Translated by M. Lebedev

Technical Physics, Vol. 48, No. 10, 2003, pp. 1307–1310. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 88–92.
Original Russian Text Copyright © 2003 by Zaporojets, Yakovenko.

OPTICS, QUANTUM ELECTRONICS

Design of an Electromagnetic System for Wide-Band Optical Modulators

D. V. Zaporojets and N. A. Yakovenko
Kuban State University, ul. Karla Libknekhta 9, Krasnodar, 350040 Russia
e-mail: den zap@mail.ru

Received October 7, 2002

Abstract—The parameters of a slot-type electromagnetic system that can be used in ultrahigh-frequency wide-band optical modulators are studied. Analytical expressions for the bandwidth of the device are obtained, and this parameter in prototypes is estimated. Experiments on optical radiation modulation by microwave signals show that these electromagnetic systems are promising for volume optical modulators. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

External modulators of optical radiation play an important role in extending the capabilities of communication by improving the data rate, functionality, and level of miniaturization. Usually, ultrahigh-frequency wide-band modulators with a low control voltage are built on microelectronic guiding systems in which the propagation velocities of optical and microwave radiations are synchronized [1]. The fabrication technology of the device is complex. It is made of a large lithium niobate single crystal cut along appropriate crystal axes. A controllable half-wavelength phase shift of optical radiation is achieved along the crystal, which must as long as several centimeters.

We consider the microwave electrodynamic system of an external ultrahigh-frequency volume modulator where a lithium niobate single crystal is placed in a section of a below-cutoff waveguide. The system is intended for a high control voltage across the lithium niobate crystal to provide a required half-wavelength phase shift of optical radiation. The crystal length determines the operating range of modulation frequencies and is chosen according to an allowable phase mismatch between microwave oscillations at the entrance to the crystal, where an optical ray enters into the device, and at the end, where the ray leaves the crystal. The transverse size of the crystal specifies the amplitude of microwave oscillations excited in the crystal: the smaller the transverse size, the lower the voltage at which a required half-wavelength phase shift is achieved. The minimum transverse size of the crystal depends primarily on the power of the optical radiation that carries information. The power of the optical radiation is limited by depolarization, which occurs in nonlinear crystals [2]. Next, the transverse size of the crystal depends on the crystal growing process and on the technology of fabricating an electromagnetic system for which the crystal is intended. Also, the transverse crystal size cannot be too small; otherwise the electromagnetic radiation applied to the crystal may produce electrical breakdown in air. It should be noted that the breakdown voltage in the crystal is much higher than in air. This allows designers to find new ways of separating the crystal from air, for example, by covering it with appropriate insulating materials outside the metallic electrodes.

In this paper, we study the parameters of an electromagnetic system that is based on below-cutoff waveguides and is a part of an optical modulator. Below-cutoff transmission lines were used in beyond-cutoff attenuators and in dielectric resonator filters placed into evanescent waveguides [3–5]. Inserted into a waveguide section, low-loss dielectric materials (alumina, teflon, or single crystals, including those with dielectric anisotropy, such as quartz or lithium niobate) make it possible to create high-Q resonators ($Q >$ 1000). Such resonators may be used not only in hybrid microwave integrated circuits as filters but also in microwave optical modulators [6]. An advantage of these modulators is that the microwave power necessary to produce a nearly 100% modulation index is low. However, their bandwidth is narrow. This drawback may be eliminated by designing wide-band electromagnetic systems, for example, short sections of below-cutoff waveguides (diaphragms) placed into a regular waveguide.

AMPLITUDE–FREQUENCY RESPONSE OF A SLOT-TYPE ELECTROMAGNETIC SYSTEM

Short sections of dielectric-filled transmission lines may be analyzed by the method of equivalent wave impedances [3]. Consider a section of an $a \times b$ rectangular waveguide of length L that is filled with a dielectric material of permittivity ε and butt joint to a section of an $a_0 \times b_0$ rectangular waveguide. We assume that a_0

and b_0 are smaller than a and b, respectively, and also that the length of the diaphragm and the width b of the slot are small. The latter assumption follows from the smallness of the cross section of the crystal. To simplify calculations, we put $L = b$.

To obtain the amplitude–frequency response of this electromagnetic system, we will calculate its power transmission coefficient [3]

$$k = 1 - |\Gamma|^2. \tag{1}$$

Here, Γ is the electric field reflection coefficient at the junction of the waveguides of different cross sections:

$$\Gamma = \frac{Z_{\text{in}} - Z_0}{Z_{\text{in}} + Z_0}, \tag{2}$$

where Z_{in} is the input impedance at the junction and Z_0 is the equivalent wave impedance of the $a_0 \times b_0$ rectangular waveguide. In turn,

$$Z_{\text{in}} = Z_1 \frac{Z_0 + Z_1 \tanh(2\pi L/\lambda_{\text{w}})}{Z_1 + Z_0 \tanh(2\pi L/\lambda_{\text{w}})}, \tag{3}$$

where Z_1 is the equivalent wave impedance of the $a \times b$ rectangular waveguide filled with a dielectric of permittivity ε and λ_{w} is the wavelength in this waveguide.

For the H_{01} wave, the parameters Z_0, Z_1, and λ_{w} have the form

$$Z_0 = \frac{kb_0}{\sqrt{4a_0^2 - \lambda^2}}, \tag{4}$$

where k is a constant factor depending on the definition of the equivalent parameters [3] (it commonly equals $480\pi\Omega$) and λ is the wavelength of the electromagnetic radiation generated by a microwave source;

$$Z_1 = \frac{kb}{\sqrt{4a^2\varepsilon - \lambda^2}}, \tag{5}$$

$$\lambda_{\text{w}} = \frac{2a\lambda}{\sqrt{4a^2\varepsilon - \lambda^2}}. \tag{6}$$

Since L is small ($L \ll \lambda_{\text{w}}$), expression (3) can be represented as

$$Z_{\text{in}} = Z_1 \frac{Z_0 + jZ_1 2\pi L/\lambda_{\text{w}}}{Z_0 + jZ_0 2\pi L/\lambda_{\text{w}}}. \tag{7}$$

Since $(L/\lambda_{\text{w}})^2 \ll 1$, this expression can be reduced to

$$Z_{\text{in}} = Z_0 + (j2\pi L/\lambda_{\text{w}})(Z_1 - Z_0^2/Z_1). \tag{8}$$

The substitution of (4), (5), and (6) into (8) yields

$$Z_{\text{in}} = \frac{kb_0}{\sqrt{4a_0^2 - \lambda^2}} + \frac{jkb_0^2 \pi L(4a^2\varepsilon - \lambda^2)}{ba\lambda(4a_0^2 - \lambda^2)}. \tag{9}$$

Substituting (9) into (2) and then (2) into (1), we obtain

$$k = 1 - \frac{\pi^2 L^2 b_0^2 (4a^2\varepsilon - \lambda^2)^2}{4a^2 b^2 \lambda^2 (4a^2 - \lambda^2) + \pi^2 L^2 b_0^2 (4a^2\varepsilon - \lambda^2)^2}. \tag{10}$$

To simplify the analysis of expression (10), we set $L = b$ and $a_0 = a\sqrt{\varepsilon}$. Formula (10) gives

$$k = 1 - \frac{\pi^2 b_0^2 (4a_0^2 - \lambda^2)}{4a_0^2\lambda^2/\varepsilon + \pi^2 b_0^2 (4a_0^2 - \lambda^2)}. \tag{11}$$

As can be seen from (11), at $\lambda = \lambda_0 = 2a_0$, $k = 1$, while the condition $k = 0.5$ is met at $\lambda = \lambda_1$, which is determined from the relationship

$$4a_0^2\lambda_1^2/\varepsilon = \pi^2 b_0^2 (4a_0^2 - \lambda^2). \tag{12}$$

Thus, we have

$$\lambda_1 = \frac{2a_0\pi b_0}{\sqrt{4a_0^2/\varepsilon + \pi^2 b_0^2}}. \tag{13}$$

The half-power bandwidth is given by

$$\Delta\lambda = \lambda_0 - \lambda_1 = 2a_0\left(1 - \frac{\pi b_0}{\sqrt{4a_0^2/\varepsilon + \pi^2 b_0^2}}\right). \tag{14}$$

As follows from formula (14), the waveguide height b_0 is the most important parameter among those that define the bandwidth of the electromagnetic system: the smaller b_0, the greater $\Delta\lambda$. When $b_0^2 \ll a_0^2$, formula (14) gives

$$\Delta\lambda = 2a_0(1 - \pi b_0\sqrt{\varepsilon}/2a_0). \tag{15}$$

Expression (15) shows that the crystal permittivity ε significantly affects the bandwidth of the device: the greater ε, the narrower the bandwidth of the electromagnetic system.

Let us estimate the bandwidth of the electromagnetic system based on standard $a_0 \times b_0 = 28.5 \times 12.5$-mm waveguides butt joint to an $a \times b = 5.4 \times 0.5$-mm waveguide. The value $a = 5.4$ mm follows from the relationship $a = a_0/\sqrt{\varepsilon}$, where $\varepsilon = 28$ is the permittivity of a Z-cut lithium niobate single crystal when the electric field is perpendicular to the cut plane. The value $b = 0.5$ mm is chosen because it is much smaller than $\lambda_0 = 2a_0 = 57$ mm. Also, crystals with such cross-sectional dimensions are easy to grow. Formula (14) yields $\Delta\lambda = 2.03$ mm. Since the transmission coefficient k equals unity when the source frequency is 5263 MHz, we find that the half-bandwidth is $\Delta f = 195$ MHz. At $b_0 = 5$ mm, $\Delta\lambda = 9.99$ mm and $\Delta f = 1120$ MHz; at $b_0 = 1$ mm, $\Delta\lambda = 41.0$ mm and $\Delta f = 13\,500$ MHz. Thus, the bandwidth of the electromagnetic system increases with decreasing b_0.

It should be noted that, in the frequency range considered, the electromagnetic system is a resonance system and, as follows from expression (5), Z_1 approaches infinity near the cutoff frequency. This allows us to expect that the efficiency of the optical modulator will be high. One should keep in mind that the value of Z_0 of the $a_0 \times b_0$ waveguide sections increases without bound in this frequency range; therefore, the effect of evanescent propagation is undoubtful in this case [4].

To design a highly efficient electromagnetic system, the resonance excitation conditions must be provided even for an electromagnetic system with small-size single crystals. However, the bandwidth of such a resonator must be wide enough. This condition is met by optimizing the dimensions of the waveguide diaphragms with crystals chosen according to the calculations presented above. The bandwidth is controlled with tunable coupling elements (evanescent waveguide sections similar to those described in [4]) inserted into the electromagnetic system.

EXPERIMENTAL RESULTS

We studied the electromagnetic system without and with a dielectric insert.

In the former case, the parameters of the electromagnetic system according to its dimensions were examined. As was noted above, the length of the diaphragm determines the resonance frequency of the electromagnetic system. The height of the diaphragm, with its thickness being fixed, specifies the bandwidth. The material of the conductor of which the diaphragm is made defines the unloaded Q factor Q_0 of the resonator.

To characterize the transfer of the microwave energy to the crystal placed in the electromagnetic system, one must know the basic parameters of resonators: the unloaded Q factor (Q_0), which is associated with losses due to a finite conductance of the walls and the presence of the dielectric insert, and the coupling Q factor (Q_{coup}), which is related to radiation losses in a resonator coupled with a transmission line. With Q_0 and Q_{coup} known, the loaded Q factor Q_1, which is measured in experiment, can easily be evaluated.

The relationships used to calculate Q_0, Q_{coup}, and Q_1 in experiments are given below [3, 7]. The first parameter is calculated as

$$Q_0 = f_0/(2\Delta f). \tag{16}$$

Here, f_0 is the resonance frequency of the microwave resonator and $2\Delta f$ is the resonator bandwidth found at a level of the transmission coefficient k_1:

$$k_1 = \frac{k_0}{(\sqrt{k_0} - 1)^2 + 1}, \tag{17}$$

where k_0 is the resonator's power transmission coefficient at resonance when coupling apertures are identical.

The parameter Q_{coup} is given by

$$Q_{\text{coup}} = Q_0/\beta, \tag{18}$$

where β is the factor of coupling between the resonator and the transmission line.

The coupling coefficient β is defined as

$$\beta = \frac{\sqrt{k_0}}{2(1 - \sqrt{k_0})}. \tag{19}$$

The loaded Q factor can be calculated as

$$Q_1 = Q_{\text{coup}} Q_0/(2Q_0 + Q_{\text{coup}}). \tag{20}$$

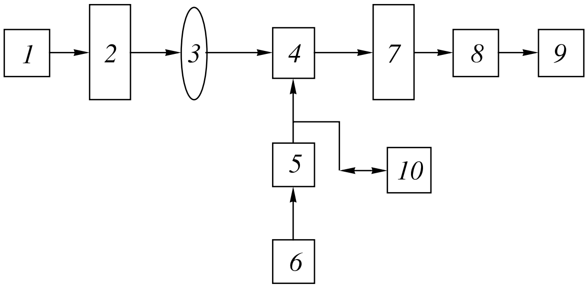
The bandwidth of the loaded resonator, $2\Delta f_1$, is determined as

$$2\Delta f_1 = f_0/Q_1. \tag{21}$$

The table lists the parameters of typical 17-mm-long diaphragms versus their height h at different thick-

Aperture parameters

Material	d, mm	h, mm	Q_0	Q_{coup}	Q_1	$2\Delta f_1$, GHz
Copper	0.5	0.3	110	13	6.1	1.5
Copper	0.5	0.4	200	16	7.7	1.2
Brass	0.3	0.1	110	21	9.7	0.93
Brass	0.3	0.2	190	18	7.7	1.2
Brass	0.3	0.3	190	14	6.6	1.4
Brass	0.3	0.4	260	12	5.9	1.5
Brass	0.3	0.5	320	10	5.6	1.6
Aluminum	0.1	0.1	100	20	8.7	1.0
Aluminum	0.1	0.2	300	13	6.4	1.4
Steel	0.02	0.1	18	15	5.3	1.7
Steel	0.02	0.2	22	14	5.3	1.7
Steel	0.02	0.4	42	10	4.6	2.0



Block diagram of the setup: (*1*) laser, (*2*) polarizer, (*3*) lens, (*4*) sample in diaphragm, (*5*) GKCh-42 signal source, (*6*) G3-56/1 audio-frequency generator, (*7*) analyzer, (*8*) photodetector, (*9*) selective amplifier, and (*10*) Ya2R-19 standing-wave indicator.

nesses *d* of the conductor. The resonance frequency is close to 9 GHz.

It is seen that the bandwidth of all but two diaphragms is greater than 1 GHz (in one case, it is even 2 GHz) and Q_0 and Q_{coup} are such that the coupling between the electromagnetic system and the resonator is always greater than unity.

At the second stage, we studied the diaphragms partially filled with lithium niobate single crystals. With a 2.0 × 0.5 × 0.5-mm crystal inserted into *a* and 17-mm-long diaphragm, the results were as follows: the frequency is 5.8 GHz, $Q_0 = 60$, $Q_{\text{coup}} = 36$, and $Q_l = 13.3$. For a 2.0 × 0.3 × 0.3-mm crystal, the frequency is 5.9 GHz, $Q_0 = 40$, $Q_{\text{coup}} = 38$, and $Q_l = 12.6$. The bandwidth is seen to narrow to 0.45 and 0.43 GHz, respectively.

Thus, these resonators are promising for ultrafast optical modulators with a bandwidth as high as several hundreds of megahertz.

In further experiments, we studied the modulation of optical radiation by microwave signals. A prototype of the microwave optical modulator was built around a lithium niobate crystal and used the transverse electrooptical effect. The crystal measuring 2.0 × 0.5 × 0.5 mm was inserted into a 0.5-mm-thick 17-mm-long copper diaphragm with a 0.5-mm-wide slot. Two 17 × 8 × 1-mm silver-plated brass evanescent diaphragms were used as coupling elements. The device was fixed between two standard waveguide-to-coaxial (28.5 × 12.5 mm) adapters. The microwave radiation was generated by the signal source of an R2-42 standard standing-wave indicator. The block diagram of the device is shown in the figure. As an optical source, we used a 0.63-μm LGN-208A laser, which emitted circularly polarized radiation. The microwave modulator was placed between crossed polarizer and analyzer. A short-focus lens focused the optical radiation on the central part of the crystal. The signal was recorded by a low-frequency photodetector at 6 kHz. Therefore, the signal from the microwave source was also modulated at this frequency by an external audio-frequency generator. The signal from the photodetector versus microwave modulator frequency was measured. At 6025 MHz, we managed to pick up a 30-μV signal with a noise level of less than 1 μV. The 3-dB bandwidth of the modulator was $2\Delta f = f_1 - f_2 = 6155 - 5790 = 365$ MHz. The frequency dependence was nonmonotonic: three resonant peaks were observed within $2\Delta f$. This is because the microwave radiation power varied from 50 to 150 mW in this frequency range.

The tuning of the microwave system to resonance provides the signal as high as 85 μV. However, the bandwidth reduced to 120 MHz.

By varying the angle between the polarizer and analyzer, one may suppress the unmodulated signal down to 500 μV and achieve a depth of modulation of about 15%.

CONCLUSIONS

(1) A technique for analyzing a slot-type electromagnetic system intended for an ultrafast external wide-band volume optical modulator is proposed.

(2) The method of equivalent wave impedances is used to calculate the transmission coefficient and bandwidth of microwave waveguides with short dielectric inserts. It is found that slot-type inserts with small slot heights provide a wide bandwidth of about 1–2 GHz. A dielectric insert narrows the bandwidth; however, with a small slot height, a sufficiently wide band of about 1 GHz can be achieved.

(3) The transfer of the microwave energy to the crystal is studied by considering the slot systems as cavity resonators. This theoretical method is verified with slots that have various dimensions and are made of different materials.

(4) The testing of optical modulator prototypes will make simple and cheap devices feasible.

REFERENCES

1. O. Mitomi, K. Noguchi, and H. Miyazawa, IEE Proc.: Optoelectron. **145**, 360 (1998).
2. E. R. Mustel’ and V. N. Parygin, *Methods of Light Modulation and Scanning* (Nauka, Moscow, 1970).
3. A. D. Grigor’ev, *Electrodynamics and Microwave Technology* (Vysshaya Shkola, Moscow, 1990).
4. S. A. Vyzulin, N. I. Vyrodov, V. V. Zaporojets, *et al.*, Radiotekhnika, No. 6, 9 (1991).
5. G. V. Zaporojets, V. V. Zaporojets, and N. A. Yakovenko, *Resonant Frequencies of Dielectric Filters Built on Evanescent Lines: Problems of Physicomathematical Simulation* (Krasnodar, 1997), pp. 77–82.
6. V. P. Gladkiy, V. V. Zaporojets, and N. A. Yakovenko, in *Proceedings of the 3rd All-Russia Research Conference on Problems of Solid-State Electronics, Taganrog, 1996*, p. 86.
7. V. V. Zaporojets and N. A. Yakovenko, *Wave Processes and Quantum Radiophysics: Laboratory Course* (Kubansk. Gos. Univ., Krasnodar, 2001).

Translated by A. Khzmalyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1311–1314. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 93–96.
Original Russian Text Copyright © 2003 by Golyshev.

OPTICS, QUANTUM ELECTRONICS

Analytical Model of an Ytterbium–Erbium Fiber Saturating Amplifier

V. Yu. Golyshev
Institute of Radio Engineering and Electronics (Fryazino Branch), Russian Academy of Sciences,
pl. Vvedenskogo 1, Fryazino, Moscow oblast, 141120 Russia
e-mail: vyug@mail.ru

Received December 15, 2002

Abstract—The amplification of an optical signal in an erbium- and ytterbium-doped fiber saturation-mode amplifier is considered. Analytical distributions of the inverse population, pump power, and signal intensity along an active light guide are derived. Spectral dependences of the saturation power and amplification threshold are obtained for different mode sizes of the radiation and ytterbium-to-erbium active ion ratios. The optimal active-fiber length is calculated. *© 2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

Erbium- and ytterbium-doped fiber amplifiers operating in the 1.5 μm range offer a number of advantages over conventional erbium amplifiers. The concentration of erbium atoms in the core of an Yb–Er light guide may reach 10^{25} m^{-3} without noticeably reducing the energy efficiency of the amplifier because of weak up-conversion. Furthermore, ytterbium transmits a part of the energy to erbium via the radiationless transfer of excitation. Ytterbium ions have a wide absorption band between 850 and 1100 nm (the absorption band of erbium ions near 980 nm is much narrower). Yb–Er amplifiers are pumped by multimode semiconductor (InGaAs) laser diodes with an output exceeding 1 W. Ytterbium–erbium active optical fibers pumped by several high-power laser diodes are widely used in booster amplifiers and fiber lasers with an output of 5 W or more in the 1535–1565 nm range.

Early theoretical works [1–5] considered amplification by active fibers doped by similar rare-earth ions such as Yb^{3+}, Nd^{3+}, Er^{3+}, Tm^{3+}, or Pr^{3+}. In active media consisting of dissimilar rare-earth ions, excitation may be transmitted between the ions nonradiatively. An example of such a medium is a quartz fiber doped by ytterbium and erbium where the former serves as a donor and the latter, as an acceptor. The numerical simulation and optimization of an Yb–Er fiber amplifier were made in [6]. However, analytical data for the optical performance of the amplifier are lacking. In this work, we derive analytical relationships for the optical characteristics of an Yb–Er saturating fiber amplifier in which self-saturation due to enhanced spontaneous self-luminescence is absent.

RATE EQUATIONS

In an Yb–Er system, the concentrations of active ions are usually taken such that each Er atom is surrounded by 15 to 50 Yb atoms. Such screening markedly diminishes the up-conversion effect for Er ions compared with the "all-erbium" medium [7, 8]. Therefore, we ignore up-conversion mechanisms associated with the excitation of a metastable Er level to higher energy levels. The energy level diagram for Yb^{3+} and Er^{3+} ions is shown in Fig. 1. Ytterbium in the ground state $^2F_{7/2}$ absorbs a pump radiation. An Yb^{3+} ion excited to the level $^2F_{5/2}$ transmits the energy to a neighboring Er^{3+} ion, which passes from the state $^4I_{15/2}$ into the state $^4I_{11/2}$. Excited erbium quickly and nonradiatively relaxes to the longer lived metastable $^4I_{13/2}$ level. A lasing transition at a wavelength of about 1550 nm occurs between the levels $^4I_{13/2}$ and $^4I_{15/2}$. Let the numbers 1, 2, and 3 be assigned to the erbium levels $^4I_{15/2}$, $^4I_{13/2}$, and $^4I_{11/2}$, respectively, and 4 and 5, to the ytterbium levels $^2F_{7/2}$ and $^2F_{5/2}$. The related populations will be designated by n_1, n_2, n_3, n_4, and n_5. The lifetimes of

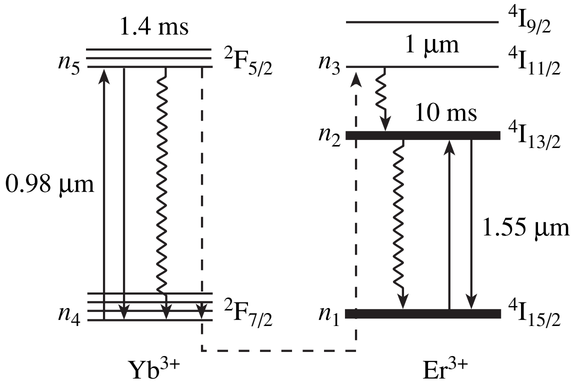


Fig. 1. Energy level diagram for Yb^{3+} and Er^{3+} ions.

1063-7842/03/4810-1311$24.00 © 2003 MAIK "Nauka/Interperiodica"

levels 2, 3, and 5 are $\tau_2 = 10^{-2}$ s, $\tau_3 = 10^{-6}$ s, and $\tau_5 = 1.4 \times 10^{-3}$ s. We introduce three spatial cylindrical coordinates: z, coordinate along the fiber; r, magnitude of the radius vector across the fiber; and φ, angle of rotation of the radius vector. In our model of active medium, all equations for pump power, fundamental mode power in the fiber, and energy level population do not depend on the angle φ; accordingly, the associated functions are also φ independent. Rate equations for the populations are given by (see, e.g., [6])

$$\begin{aligned}\frac{\partial}{\partial t}n_2 &= -\frac{1}{\tau_2}n_2 + \frac{1}{\tau_3}n_3 + w_{12}n_1 - w_{21}n_2,\\ \frac{\partial}{\partial t}n_3 &= -\frac{1}{\tau_3}n_3 + c_{\rm cr}n_1n_5 - c_{\rm cr}n_3n_4,\\ \frac{\partial}{\partial t}n_5 &= -\frac{1}{\tau_5}n_5 + w_{45}n_4 - w_{54}n_5 + c_{\rm cr}n_3n_4 - c_{\rm cr}n_1n_5,\end{aligned} \tag{1}$$

where n_i and w_{ij} ($i, j = 1$–5) are functions of only z and r.

In (1), $c_{\rm cr}$ is the coefficient of excitation transfer between Yb and Er ions[1]; w_{ij} is the rate of induced transitions from level i to level j:

$$w_{12}(r) = \frac{\psi_{\rm s}(r)}{hc\pi\omega_{\rm s}^2}\int d\lambda\sigma_{12}(\lambda)P_{\rm s}(\lambda)\lambda,$$

$$w_{21}(r) = \frac{\psi_{\rm s}(r)}{hc\pi\omega_{\rm s}^2}\int d\lambda\lambda\sigma_{12}(\lambda)P_{\rm s}(\lambda),$$

$$w_{45}(r) = \frac{\psi_{\rm p}(r)}{hc\pi\omega_{\rm p}^2}\int d\lambda\lambda\sigma_{45}(\lambda)P_{\rm p}(\lambda),$$

h is the Planck constant; c is the speed of light;

$$\omega_{\rm s,\,p} = \sqrt{2\int_0^{+\infty} drr\psi_{\rm s,\,p}(r)}$$

are the mode sizes of the signal and pump in the fiber; $\psi_{\rm s,\,p}(r)$ are the signal and pump distribution functions over the cross section ($\psi_{\rm s,\,p}(0) = 1$); σ_{ij} are the cross sections of the induced transition (via absorption or luminescence) from level i to level j for the signal wavelength $\lambda_{\rm s}$ and pump radiation wavelength $\lambda_{\rm p}$; and $P_{\rm s,\,p}(\lambda)$ is the power spectral density.

The distributions of the pump and signal along the fiber are given by the differential equations

$$\begin{aligned}\frac{d}{dz}P_{\rm p}(z) &= -\frac{2}{\omega_{\rm p}^2}P_{\rm p}(z)\\ &\times\int_0^{+\infty} drr(\sigma_{45}n_4(r) - \sigma_{54}n_5(r) + \alpha_{\rm p})\psi_{\rm p}(r),\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{d}{dz}P_{\rm s}(z) &= -\frac{2}{\omega_{\rm s}^2}\int_0^{+\infty} drr(\sigma_{21}n_2(r)[P_{\rm s}(z) + 2P_0]\\ &- \sigma_{12}n_1(r)P_{\rm s}(z) - \alpha_{\rm s}P_{\rm s}(z))\psi_{\rm s}(r),\end{aligned} \tag{3}$$

where $\alpha_{\rm s,\,p}$ are the coefficients of passive losses due to scattering by inhomogeneities and impurity centers (their typical values are $\alpha_{\rm s} \approx 0.1$ dB/m and $\alpha_{\rm p} \approx 0.2$ dB/m). The spontaneous luminescence power $P_0 = h\nu\delta\nu$ depends on one noise photon in a frequency band $\delta\nu$, and factor two before P_0 takes into account two orthogonal states of polarization.

SOLUTION AND ANALYSIS

Using the relationships $a = n_1 + n_2 + n_3$ and $d = n_4 + n_5$ (a and d are the concentrations of erbium and ytterbium atoms, respectively), one easily obtains a steady-state solution to system (1):

$$\begin{aligned}n_2 &= \frac{a}{2}(1 + \nu_1 + \nu_2 + \nu_3\\ &- \sqrt{(\nu_1 + \nu_2 - 1)^2 + \nu_3^2 + 2\nu_3(\nu_1 + 1 - \nu_2)}),\end{aligned} \tag{4}$$

$$n_3 = d\frac{\tau_3w_{45}(a - n_2)}{(d\tau_3 + c_{\rm cr}^{-1})(w_{45} + w_{54} + \tau_5^{-1}) + a - n_2}, \tag{5}$$

$$n_5 = d\frac{c_{\rm cr}^{-1}w_{45} + n_3}{c_{\rm cr}^{-1}(w_{45} + w_{54} + \tau_5^{-1}) + a - n_2}, \tag{6}$$

where

$$\nu_1 = \frac{d}{a}\frac{\tau_2w_{45}(1 - \tau_3w_{12})}{1 + \tau_2(w_{12} + w_{21})}, \tag{7}$$

$$\nu_2 = \frac{\tau_2w_{12}}{1 + \tau_2(w_{12} + w_{21})}, \tag{8}$$

$$\nu_3 = \frac{1}{a}\left(d\tau_3 + \frac{1}{c_{\rm cr}}\right)\left(\frac{1}{\tau_5} + w_{45} + w_{54}\right). \tag{9}$$

Let us introduce a number of simplifications. In phosphate fibers, the transfer rate $c_{\rm cr}d$ exceeds 10^7 s^{-1} [9].[2]

---

[1] In thermal equilibrium, the rates of excitation transfer from Yb to Er, $c_{\rm cr}^{+}$, and from Er to Yb, $c_{\rm cr}^{-}$, are related as $c_{\rm cr}^{-}/c_{\rm cr}^{+} = \exp\{-(E_5 - E_3)/k_{\rm B}T\}$, where $E_5 - E_3$ is the transfer energy loss [8]. Since the energy levels of Er and Yb are close to each other, we put $c_{\rm cr}^{-} = c_{\rm cr}^{+} \equiv c_{\rm cr}$.

[2] According to [9], if the interaction between higher multipoles is taken into account and the minimal spacing between Er and Yb atoms is assumed to be less than 4 Å, the excitation transfer rate exceeds 10^7 s^{-1}. Such spacings in optical fibers are possible if rare-earth atoms form clusters [7].

Therefore, one may consider the limiting case $c_{cr} \longrightarrow +\infty$ (see formulas (4)–(6), (9)). We also put $\tau_3 = 0$, since the radiationless transition of excitation from the third to second level of Er proceeds much more rapidly than spontaneous and induced lasing transitions. With this approximation, we arrive at two steady-state solutions to Eqs. (1). If $dw_{45} \le a(\tau_2^{-1} + w_{21})$, we have

$$n_2 = \frac{dw_{45} + aw_{12}}{\tau_2^{-1} + w_{12} + w_{21}}, \tag{10}$$

$$n_3 = n_5 = 0. \tag{11}$$

The other solution (at $dw_{45} > a(\tau_2^{-1} + w_{21})$) corresponds to the complete inversion of the active medium; that is,

$$n_2 = a. \tag{12}$$

To provide a high output power, a booster amplifier usually operates in the high saturation mode. In this case, the steady-state solution to the rate equations is given by formulas (10) and (11).

Let us introduce the saturation power

$$P_{\text{sat}} = hc\pi\omega_s^2\{\lambda_s\tau_2(\sigma_{12} + \sigma_{21})\}^{-1}$$

and the threshold power

$$P_{\text{th}} = \frac{a\sigma_{12}}{d\sigma_{45}} hc\pi\omega_p^2\{\lambda_p\tau_2(\sigma_{12} + \sigma_{21})\}^{-1}.$$

The plots of $P_{\text{sat}}(\lambda)$ and $P_{\text{th}}(\lambda)$ for a phosphate fiber are depicted in Figs. 2 and 3. To simplify algebraic expressions, we normalize the optical powers of the signal and pump: $s = P_s/P_{\text{sat}}$ and $p = P_p/P_{\text{th}}$. Then, the rates of induced transitions will have the form

$$w_{12}\tau_2 = \sigma_{12}(\sigma_{12} + \sigma_{21})^{-1}s\psi_s,$$

$$w_{21}\tau_2 = \sigma_{21}(\sigma_{12} + \sigma_{21})^{-1}s\psi_s,$$

$$w_{45}\tau_2 = \frac{a}{d}\sigma_{12}(\sigma_{12} + \sigma_{21})^{-1}p\psi_p.$$

Substituting formulas (10) and (11) into (2) and (3) and neglecting the spontaneous luminescence power P_0 yields

$$\frac{d}{dz}p(z) = -\frac{2}{\omega_p^2}p(z)\int_0^{+\infty} drr(d\sigma_{45}\Theta(r_0 - r) + \alpha_p)\psi_p(r), \tag{13}$$

$$\frac{d}{dz}s(z) = \frac{2}{\omega_s^2}s(z)\int_0^{+\infty} drr \times\left(a\sigma_{12}\Theta(r_0 - r)\frac{p\psi_p(r) - 1}{s\psi_s(r) + 1} - \alpha_s\right)\psi_s(r), \tag{14}$$

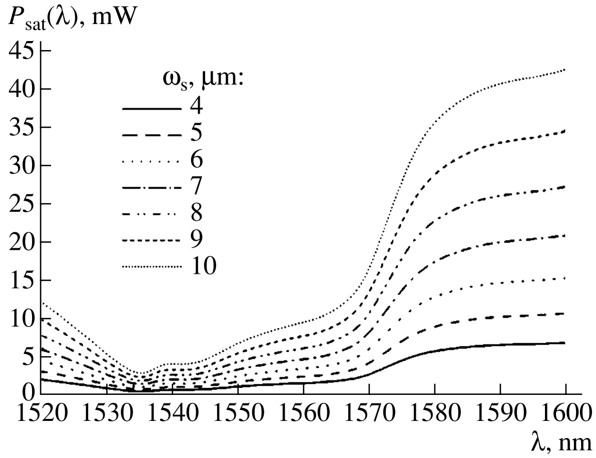


Fig. 2. Spectral dependences $P_{\text{sat}}(\lambda)$ for different mode sizes of the signal in the fiber.

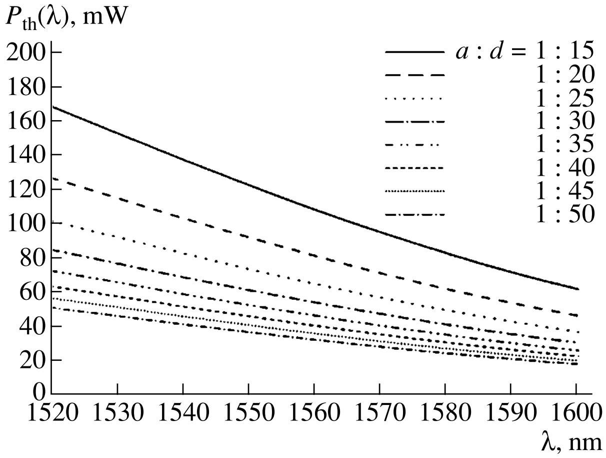


Fig. 3. Spectral dependences $P_{\text{th}}(\lambda)$ for different concentrations of Er^{3+} and Yb^{3+} ions. $\omega_p = 110$ μm, $\sigma_{45} = 1.7 \times 10^{-25}$ m^2.

where

$$\Theta(x) = \begin{cases} 1, & x \ge 0 \\ 0, & x < 0 \end{cases}$$

is the unit step function and r_0 is the fiber core radius.

We assume that the fundamental mode LP_{01} of the signal radiation has the Gaussian distribution over the cross section of the fiber; that is, $\psi_s(r) = \exp(-r^2/\omega_s^2)$. For multimode pumping, we may put $\psi_p(r) = \Theta(\omega_p - r)$, where ω_p is the outer radius of the fiber cladding. In (14), the integral represents the integral overlap factor, which depends on the signal and pump powers:

$$\Gamma(s, p) = \frac{2}{\omega_s^2}\int_0^{+\infty} drr\Theta(r_0 - r)\psi_s(r)\frac{p\psi_p(r) - 1}{s\psi_s(r) + 1}.$$

This factor includes the overlap of the signal and pump modes with the active fiber core and inverse population variation across the fiber. It differs from the overlap factor $\Gamma_s = 1 - \exp(-r_0^2/\omega_s^2)$ [1], which is valid for the unsaturated operating regime and reflects the partial overlap of the signal with the activated core. Integrating over r in formulas (13) and (14), we obtain

$$\frac{d}{dz}p(z) = -\left\{\left(\frac{r_0}{\omega_p}\right)^2 d\sigma_{45} + \alpha_p\right\}p(z), \qquad (15)$$

$$\frac{d}{dz}s(z) = a\sigma_{12}(p(z)-1) \times \ln\left[\frac{1+s(z)}{1+s(z)\exp(-r_0^2/\omega_s^2)}\right] - \alpha_s s(z). \qquad (16)$$

Equation (15) has the solution

$$p(z) = p_+\exp[-\gamma z] + p_-\exp[-\gamma(L-z)], \qquad (17)$$

where p_+ is the pump power applied to the entrance into the fiber ($z = 0$) in the direction of signal propagation, p_- is the pump power applied to the exit from the fiber ($z = L$) in the direction opposite to the signal propagation, and $\gamma = r_0^2\omega_p^{-2}d\sigma_{45} + \alpha_p$ is the pump absorption coefficient.

Equation (16) is readily integrable over z if $s \gg 1$:

$$s(z) = s(0) + a\sigma_{12}\left(\frac{r_0}{\omega_s}\right)^2\exp[-\alpha_s z] \times \int_0^z dx(p(x)-1)\exp[\alpha_s x]. \qquad (18)$$

Substituting expression (17) into (18) yields the signal power at the output of the amplifier:

$$s(L) = s(0) + a\sigma_{12}\left(\frac{r_0}{\omega_s}\right)^2\left\{\frac{p_+}{\gamma-\alpha_s}(\exp[-\alpha_s L] - \exp[-\gamma L]) + \frac{p_-}{\gamma+\alpha_s}(1-\exp[-(\alpha_s+\gamma)L]) - \frac{1}{\alpha_s}(1-\exp[-\alpha_s L])\right\}. \qquad (19)$$

From (19), one can find the optimal length of the fiber, i.e., the length providing the maximal output at a given pump power. For $\alpha_s = 0$, we have

$$L_{\text{opt}} = \frac{\ln(p_+ + p_-)}{\gamma}. \qquad (20)$$

The corresponding output power is

$$s(L_{\text{opt}}) = s(0) + a\sigma_{12}\left(\frac{r_0}{\omega_s}\right)^2 \times \frac{1}{\gamma}\{p_+ + p_- - 1 - \ln(p_+ + p_-)\}. \qquad (21)$$

Let us evaluate the output power of a single-stage amplifier pumped from the exit. With $\lambda_s = 1550$ nm, $P_s(0) = 20$ mW, $a = 1.336 \times 10^{25}$ m^{-3}, $a/d = 1/25$, $P_p = 2$ W, $\sigma_{12} = 2.394 \times 10^{-25}$ m^2, $\sigma_{21} = 3.6 \times 10^{-25}$ m^2, $\sigma_{45} = 1.7 \times 10^{-25}$ m^2, $r_0 = \omega_s = 7$ µm, $\sigma_p = 110$ µm, and $L = 5$ m, formula (19) yields the output of the amplifier $P_s(L) = 815$ mW.

CONCLUSIONS

Our model is based on the following assumptions: (i) the signal exceeds the power of enhanced spontaneous luminescence, $P_s(0) \gg GP_0$ (G is the gain, $P_0 \sim 1$ µW within the 30-nm-wide band), and (ii) the amplifier operates in the saturation regime with the incomplete inversion of the active medium: $P_s(0) \gg P_{\text{sat}}$ and

$$P_s(L) \geq P_{\text{sat}}\exp\left(-\frac{r_0^2}{\omega_s^2}\right)\left[\frac{\sigma_{12}}{\sigma_{21}}\left(\frac{P_p}{P_{\text{th}}}-1\right)-1\right].$$

The analytical expression obtained for the output power of the amplifier makes it possible to optimize the waveguide characteristics of an active fiber and calculate its optimal length, which depends on both the absorption coefficient and total pump power applied to the fiber. We used the integral overlap factor that is a function of the signal and pump powers, adequately describes the active erbium medium in the saturation mode, and differs from the overlap factor Γ_s for the unsaturated regime.

REFERENCES

1. E. Desurvire, *Erbium-Doped Fiber Amplifiers Principles and Applications* (Wiley, New York, 1994).
2. C. Barnard, P. Myslinski, J. Chrostowski, and M. Kavehrad, IEEE J. Quantum Electron. **30**, 1817 (1994).
3. F. Sanchez, B. Mezianei, T. Chartier, *et al.*, Appl. Opt. **34**, 7674 (1995).
4. A. Hardy and R. Oron, IEEE J. Quantum Electron. **33**, 307 (1997).
5. I. Kelson and A. Hardy, IEEE J. Quantum Electron. **34**, 1570 (1998).
6. M. Karasek, IEEE J. Quantum Electron. **33**, 1699 (1997).
7. M. Fedrighi and F. Di Pasquale, IEEE Photonics Technol. Lett. **7**, 303 (1995).
8. A. M. Burshteĭn, Avtometriya, No. 5, 65 (1978).
9. V. P. Gapontsev and N. S. Platonov, Mater. Sci. Forum **51**, 165 (1989).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1315–1318. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 97–100.
Original Russian Text Copyright © 2003 by Ostrovskiĭ, Nadtochiĭ, Korotchenkov, Nikandrova.

ACOUSTIC, ACOUSTOELECTRONICS

Q Factor of Lithium Niobate Piezoelectric Transducers at High Excitation Levels

I. V. Ostrovskiĭ, A. B. Nadtochiĭ, O. A. Korotchenkov, and M. V. Nikandrova
Shevchenko National State University, Kiev, 03127 Ukraine
e-mail: nadt@gala.net

Received July 15, 2002; in final form, March 17, 2003

Abstract—The variations of the electroacoustic parameters (Q factor, electromechanical coupling coefficient, and capacitance) of lithium niobate piezoelectric transducers with increasing high-frequency excitation voltage are studied experimentally. The relative acoustic strain is found to reach a maximum of about 10^{-4} in the frequency range from 2 to 3 MHz. The Q factor of the transducers may increase by 100% in the range of acoustic strains studied. This increase is accompanied with acoustic emission. The reason for this effect is the block structure of the lithium niobate crystal. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

Lithium niobate ($LiNbO_3$) piezoelectric, one of the basic materials of modern acoustoelectronics and optoelectronics [1–3], is widely used in sources of ultrasonic surface and bulk waves, where the high-frequency voltage that excites ultrasonic waves is often high. However, the variation of the transducer parameters due to its vibration has been studied insufficiently. It has been found that the Q factor of piezoelectric transducers decreases with increasing vibration amplitude [4]. Therefore, the study of such effects in $LiNbO_3$ plates is of current interest. In this paper, we experimentally investigate the variations of the electroacoustic parameters of lithium niobate piezoelectric transducers (Q factor, electromechanical coupling coefficient, and capacitance) with growing amplitude of the high-frequency electric voltage that excites the transducer.

MEASURING TECHNIQUE

The parameters that define the operation of a piezoelectric transducer are its capacitance C, Q factor, and electromechanical coupling coefficient K. Traditionally, these parameters are determined by [5] (i) measuring the capacitance C at a frequency much lower than the transducer resonance frequency, (ii) calculating the Q factor from the resonance width of the frequency dependence of the admittance at a level of 0.707, and (iii) calculating the electromechanical coupling coefficient squared, K^2, from the measured resonance, f_r, and antiresonance, f_a, frequencies of the admittance:

$$K^2 = \frac{\pi f_r}{2 f_a}\cot\left(\frac{\pi f_r}{2 f_a}\right). \tag{1}$$

It should be noted that, at sufficiently high vibration amplitudes, the parameters of piezoelectric transducers start depending on the electric voltage applied. In this case, the accuracy with which the parameters are measured is limited by the fact that the loading conditions of the transducer under which C, Q, and K^2 are measured may differ. In this paper, we propose a technique that allows one to find all three parameters at a given amplitude of the applied voltage in one measurement cycle. This is achieved by using a composite two-frequency exciting signal given by

$$U = U_r\sin(2\pi f_r) + U_a\sin(2\pi f_a + \Delta\varphi), \tag{2}$$

where U_r is the voltage amplitude at the resonance frequency f_r; U_a is the voltage amplitude at the antiresonance frequency f_a; and $\Delta\varphi$ is a constant phase shift, which is insignificant in our measurements.

With the use of such a signal, the transducer impedance can be measured at its resonance and antiresonance frequencies simultaneously.

Let us find relationships between the impedance of a piezoelectric transducer at the resonance and antiresonance frequencies and its parameters. The wave propagation coefficient γ is defined as

$$\gamma = \alpha + j\frac{\omega}{V}, \tag{3}$$

where $\omega = 2\pi f$ is the angular frequency, V is the velocity of sound; j is the imaginary unit; and α is the ultrasound attenuation factor, which is related to the Q factor of the transducer as

$$Q = \frac{\pi}{\alpha\lambda}. \tag{4}$$

Here, λ is the acoustic wave wavelength, which may be taken as twice the transducer thickness in this case.

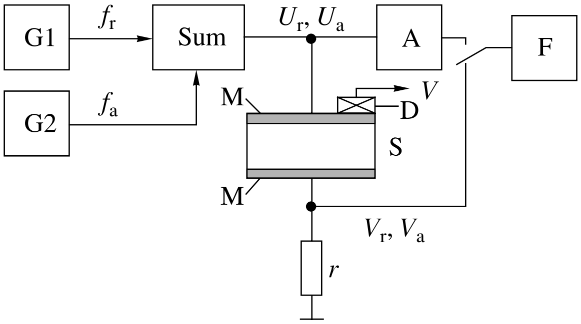


Fig. 1. Block diagram of the setup for measuring the transducer parameters: G1 and G2, signal sources; Sum, power combiner; A, attenuator; S, sample ($LiNbO_3$ plate); M, metal layer; D, acoustic detector; *r*, 1- to 10-Ω resistance; and F, spectrum analyzer.

The transducer admittance can be represented as [6]

$$Y = \frac{j\omega C}{1 - \frac{K^2 \tanh(\gamma d)}{\gamma d}}, \qquad (5)$$

where *C* is the transducer capacitance and *d* is its half-thickness.

Expanding admittance *Y* (5) into the series in 1/*Q* near the antiresonance frequency and leaving only the linear term yields an expression that relates the transducer resistance R_a at its antiresonance frequency f_a to the Q factor:

$$R_a = \frac{1}{2\pi f_a C} \frac{8}{\pi^2} K^2 Q. \qquad (6)$$

It can be seen that the transducer resistance at the antiresonance frequency is directly proportional to its Q factor. Next, we expand 1/*Y* given by (5) into the series in 1/*Q* near the resonance frequency, leave the linear term alone, and take into account that K^2 is small to obtain a relationship between the transducer resistance R_r at its resonance frequency f_r to its Q factor:

$$R_r = \frac{1}{2\pi f_a C K^2 Q} \frac{\pi^2}{8}. \qquad (7)$$

From (6) and (7), we find the expressions for the transducer capacitance,

$$C = \frac{1}{2\pi f_a \sqrt{R_r R_a}}, \qquad (8)$$

and its Q factor,

$$Q = \left(\frac{\pi}{2}\right)^2 \frac{1}{2K^2} \sqrt{\frac{R_a}{R_r}}. \qquad (9)$$

When the length and width of a plated piezoelectric resonator are much smaller than its thickness, the strain ε induced in the transducer by the high-frequency voltage *U* applied can be calculated in terms of the standard one-dimensional model [7]. The calculations yield [6]

$$\varepsilon = U \sqrt{\frac{K^2 C}{m v^2}} \frac{\cosh(\gamma z)/\cosh(\gamma d)}{1 - \frac{K^2 \tanh(\gamma d)}{\gamma d}}, \qquad (10)$$

where *m* is the transducer weight and *z* is the coordinate along its thickness.

The strain is seen to be maximal at the center of the transducer (*z* = 0) at the resonance frequency. Using the expansion in 1/*Q* and leaving its linear term, we obtain from (10), (3), and (4) the expression for the strain

$$\varepsilon \approx U \sqrt{\frac{K^2 C}{m v^2}} \frac{4}{\pi} Q. \qquad (11)$$

The block diagram of the setup for measuring the parameters of the resonator excited by an amplitude-modulated signal is shown in Fig. 1. Let V_r and V_a be the voltages at the resonance, f_r, and antiresonance, f_a, frequencies across the resistor of value *r*. Then,

$$R_r = \left(\frac{V_r}{U_r} - 1\right) r, \quad R_a = \left(\frac{V_a}{U_a} - 1\right) r. \qquad (12)$$

The voltages U_r, U_a, V_r, and V_a, as well as the frequencies f_r and f_a, were measured directly by the spectrum analyzer *F*. The resonator parameters were calculated from formulas (1), (8), and (9), and the strain was evaluated from (11).

To detect internal mechanical processes (e.g., the motion of dislocations and grain boundaries) that accompany the high-amplitude vibration of the piezoelectric resonator, we applied the acoustic emission method. Acoustic noise attendant on these processes was recorded using an AF-15 instrument and filters that suppress spurious signals at the excitation frequency. We also used standard acoustic emission detectors of the AF-15. When the dimensions of the sample were much smaller than those of the detector, we employed industrial PZT-based piezoelectric transducers (D in Fig. 1). The sensitivity of the acoustic emission detector is maximal near its antiresonance frequency. Under the assumption that the acoustic signal is due to longitudinal waves, the amplitude of acoustic emission pulses can be estimated from the expression [6]

$$A = \frac{V}{4h_{33}}, \qquad (13)$$

where *V* is the voltage generated by the detector under the action of an acoustic wave of amplitude *A* and h_{33} is the piezoelectric coefficient of the detector.

The coefficient h_{33} is related to other parameters of the detector as

$$h_{33} = \sqrt{\frac{K_d^2 4 m_d f_a^2}{C_d}}, \qquad (14)$$

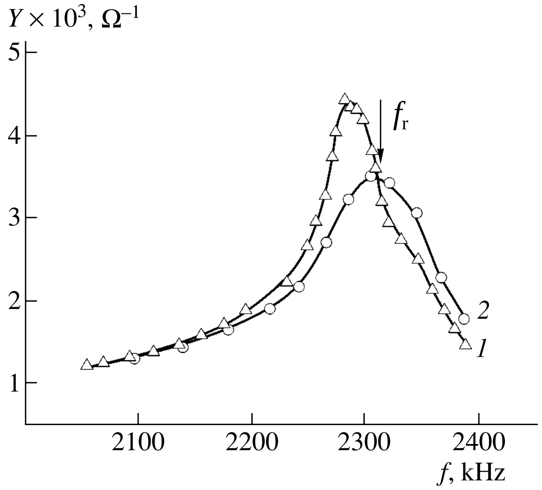


Fig. 2. Admittance Y of the LNY2 sample versus frequency f at U = (*1*) 1 and (*2*) 30 V. The resonance vibration frequency f_r for curve *1* is indicated by the arrow.

where m_d, K_d, and C_d are the weight, electromechanical coupling coefficient, and capacitance of the detector, and f_a is the antiresonance frequency.

We studied eight X- and Y-cut $LiNbO_3$ plates. Below are the results for three Y-cut samples marked LNY1, LNY2, and LNY4. The length and width of the samples were varied from 6 to 11 mm, and their thickness was 1 mm. The dislocation structure of the $LiNbO_3$ crystals was revealed by etching in an $HNO_3 : HF = 2 : 1$ selective etchant for several tens of minutes at 110°C (the boiling temperature of the etchant). The etched surface of the plate was examined under an MIM metallographic microscope.

RESULTS AND DISCUSSION

In some of the X- and Y-cut samples, we discovered a significant increase in the Q factor with increasing acoustic strain. This effect is illustrated in Fig. 2 as the evolution of the frequency dependence of the admittance, $Y(f)$, with growing amplitude of the excitation voltage U. The slight decrease in the resonance frequency f_r (and in the antiresonance frequency f_a not shown in Fig. 2) is associated with the extension of the sample heated by ultrasound. At the same time, the resonance curve narrows substantially, which shows that the Q factor increases. It is important that this effect cannot be reproduced by merely heating the sample when curve *1* in Fig. 2 is taken at a low excitation level. Therefore, the increase in the Q factor (curve *2* in Fig. 2) is not a purely thermal effect, as also demonstrated by the insignificant increase in Q of $LiNbO_3$ crystals from another group (see below).

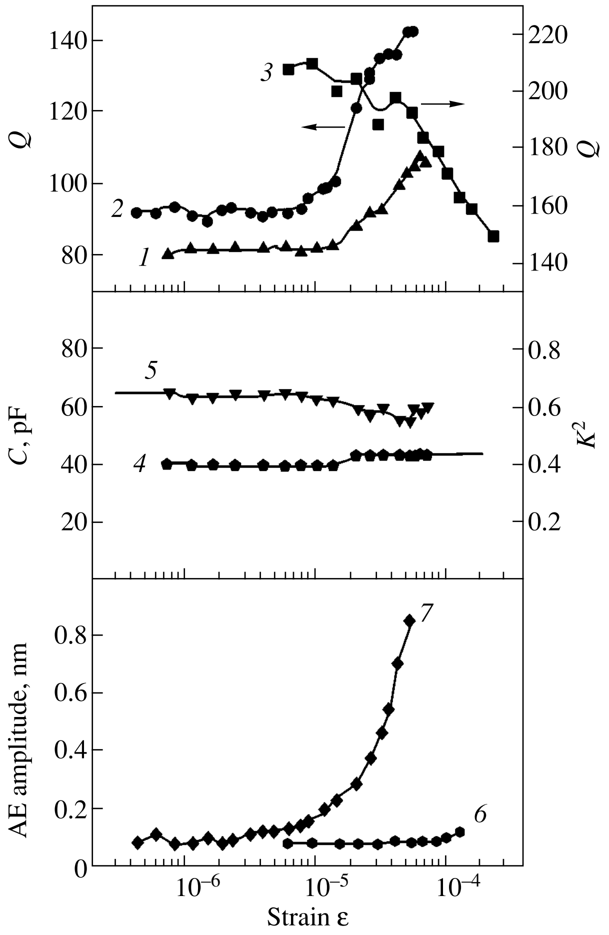


Fig. 3. (*1*) Q factor, (*4*) electromechanical coupling coefficient squared K^2, and (*5*) capacitance C versus strain for the LNY2 sample. (*2*, *3*) Strain dependences of Q for the (*2*) LNY4 and (*3*) LNY1 samples. Curve *1* is shifted 50 units upwards. (*6*, *7*) Acoustic emission versus strain for the LNY1 and LNY4 samples, respectively.

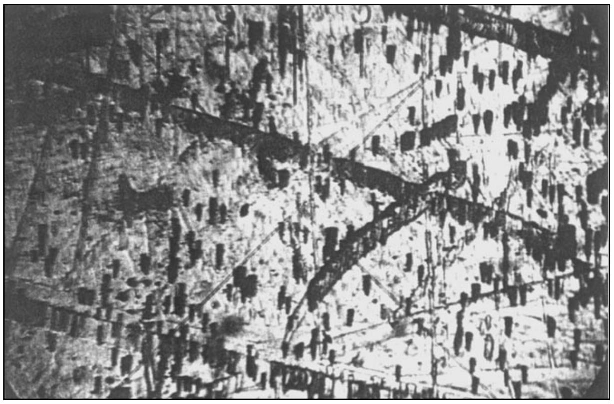


Fig. 4. Micrograph of the etched surface for the LNY4 sample. The size of the region shown is 480×320 μm.

The parameters Q, C, and K^2 calculated as a function of the acoustic strain from data similar to those shown in Fig. 2 are plotted in Fig. 3 (curves *1*, *2*, *4*, and *5*). It can be seen that the capacitance C and the electromechanical coupling coefficient squared, K^2, remain almost unchanged (curves *4*, *5*), while the Q factor starts growing from strains of about 10^{-5} and increases by almost 100% at $\varepsilon \approx 10^{-4}$ (curves *1*, *2*). With the excitation level increasing further, *Q* goes down (not shown in curves *1* and *2*).

Note that the above effect depended on which group of $LiNbO_3$ crystals was used as the resonator. In a number of cases, the increase in the Q factor was not observed (curve *3* in Fig. 3). The study of etch patterns on the selectively etched surfaces made it possible to find the nature of this effect. The etched surface of one of the crystals that exhibit the increase in *Q* is shown in Fig. 4. Etch pits (dark spots elongated in the vertical direction) appear at the sites where individual dislocations cross the surface. Such pits of variable density were observed in all the crystals studied irrespective of the behavior of *Q*. At the same time, the surfaces of the samples that exhibited the effect of increasing *Q* had pit chains characteristic of grain boundaries (three inclined pit chains in Fig. 4). X-ray examination revealed a subblock structure in such $LiNbO_3$ crystals.

Thus, we may assume that the decrease in the Q factor with increasing excitation level (curve *3* in Fig. 3) is of dislocation nature similar to that described earlier [4]. The growth of *Q* observed may be associated with the presence of grain boundaries in the $LiNbO_3$ crystals, as also corroborated by the strain dependence of the acoustic emission amplitude (curves *6*, *7* in Fig. 3). In the sample where the density of dislocations is low, the emission is very weak and appears only if $\varepsilon \geq 10^{-4}$ (curve *6*). In this case, the emission may also be of dislocation nature [8]. By contrast, in the other samples, the behavior of the emission amplitude as the excitation level increases correlates with the growth of the Q factor, as demonstrated by curve *7* and curve *2* (Fig. 3), which are plotted for the same sample. Thus, the acoustic emission signal in this case is apparently generated by the motion of grain boundaries and the increase in the Q factor corresponds to the orientation of subblocks in the acoustic field. The latter effect reduces the acoustic losses (the factor alpha in formula (4)) and, consequently, increases *Q*.

CONCLUSIONS

It is shown that the Q factor of $LiNbO_3$ piezoelectric transducers may rise significantly (by about 100%) with increasing excitation level. The other parameters of the transducer (electromechanical coupling coefficient and capacitance) vary insignificantly. The increase in *Q* is related to the block structure of $LiNbO_3$ crystals. This effect must be taken into account when designing and manufacturing high-intensity $LiNbO_3$-based ultrasound sources.

REFERENCES

1. Yu. S. Kuz'minov, *Electrooptical and Nonlinear Optical Lithium Niobate Crystal* (Nauka, Moscow, 1987).
2. C. Campbell, *Surface Acoustic Wave Devices and Their Signal Processing Applications* (Academic, Boston, 1989).
3. C. Campbell, *Surface Acoustic Wave Devices for Mobile and Wireless Communications* (Academic, Boston, 1998).
4. O. A. Korotchenkov, A. T. Marchenko, and I. V. Ostrovskii, Zh. Tekh. Fiz. **62** (8), 187 (1992) [Sov. Phys. Tech. Phys. **37**, 895 (1992)].
5. J. Zelenka, *Piezoelectric Resonators and Their Applications* (Elsevier, Amsterdam, 1986; Mir, Moscow, 1990).
6. A. B. Nadtochii, Candidate's Dissertation (Kiev, 2000).
7. I. Ya. Kotsarenko, S. V. Koshevaya, and I. V. Ostrovskii, Radiotekh. Élektron. (Moscow) **18**, 1208 (1973).
8. V. A. Kalitenko, O. A. Korotchenkov, I. Ya. Kucherov, *et al.*, Ukr. Fiz. Zh. **30**, 1358 (1985).

Translated by A. Khzmalyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1319–1322. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 101–105.
Original Russian Text Copyright © 2003 by Dvoesherstov, Cherednick, Chirimanov, Petrov.

Parameter Optimization of Surface Acoustic Waves Via Multilayer Structures

M. Yu. Dvoesherstov, V. I. Cherednick, A. P. Chirimanov, and S. G. Petrov
Lobachevski State University, pr. Gagarina 23, Nizhni Novgorod, 603950 Russia
e-mail: Dvoesh@rf.unn.ru

Received February 4, 2003

Abstract—A method and algorithm for numerical analysis and optimization of the basic parameters of electroacoustic surface waves propagating in multilayer piezoelectric structures are described. Combinations of layer materials and piezoelectric substrates for which an electroacoustic surface wave has optimal parameters of propagation (low dispersion, high electromechanical coupling coefficient, high thermal stability, low diffraction losses, etc.) are found. © *2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

Basic parameters of electroacoustic surface waves (EASWs) are [1] wave velocity V, electromechanical coupling coefficient K^2, power flux angle *pfa*, anisotropy parameter β, temperature coefficient of delay *tcd*, velocity dispersion γ in the case of layer systems, propagation losses δ for pseudosurface electroacoustic waves [2], etc.

The parameters of EASWs propagating in multilayer structures depend on material properties of the substrate and individual layers. When the total thickness of the layers is comparable to or greater than the wavelength λ, the EASW structure is rather complex and, in addition, contains higher order modes. If $H \ll \lambda$, the parameters and structure of propagating EASWs are governed largely by the properties of a semi-infinite piezoelectric substrate. Layers applied on its surface generally modify the EASW parameters only slightly. Nevertheless, even if $H \ll \lambda$, the wave parameters may be considerably improved by using film–piezoelectric substrate structure [3, 4]. For example, if the piezoelectric substrate is of poor thermal stability in terms of EASW propagation, the thermal stability and simultaneously the electromechanical coupling coefficient may be improved by applying an appropriate layer of certain thickness. In the case of pseudosurface acoustic waves (PSAWs), an applied film of certain thickness may sometimes greatly decrease the propagation losses δ [5].

However, unlike bare piezoelectric substrates, the layer systems suffer from EASW velocity dispersion. With even one layer of finite thickness h applied on the substrate surface, the EASW velocity starts depending on the frequency. One may try to eliminate this drawback with multilayer structures. Let two layers be applied on a piezoelectric surface. If one layer increases the wave velocity and the other, conversely, decreases, the combination of both (with appropriate thicknesses) will help to eliminate dispersion and simultaneously improve the thermal stability of the wave.

Thus, varying the material and thickness of layers applied on a piezoelectric substrate of certain orientation, one can solve the problem of parameter optimization for EASWs propagating in multilayer structures by using numerical experiment. This work is devoted to the solution of this problem. A general method of numerical analysis and EASW parameter optimization by using multilayer piezoelectric structures is suggested.

STATEMENT OF THE PROBLEM

To determine the EASW basic parameters in multilayer structures, it is necessary to solve piezoacoustic equations [6, 7] that describe the EASW propagation in piezoelectric crystals and a set of equations of wave propagation in a separate layer [7]. For an arbitrary number of layers applied on a piezosubstrate of arbitrary crystal symmetry, the EASW parameters can be evaluated only numerically. When solving these equations, one may invoke the Farnell–Jones [7] or Adler [2] approach. To find the EASW phase velocity and the other parameters listed above, it is also necessary to state appropriate mechanical and electrical boundary conditions at the interfaces and outer boundaries of the layers and substrate. In the case of multilayer systems, three components of mechanical displacements u_j, three normal components of the mechanical stress tensor T_{3j}, the electrical potential φ, and the normal component of the electric field induction D_3 must satisfy the continuity conditions at the interfaces [7]. Clearly, as the number of layers applied increases, so does the number of mechanical and electrical boundary conditions. Furthermore, electrical boundary conditions vary

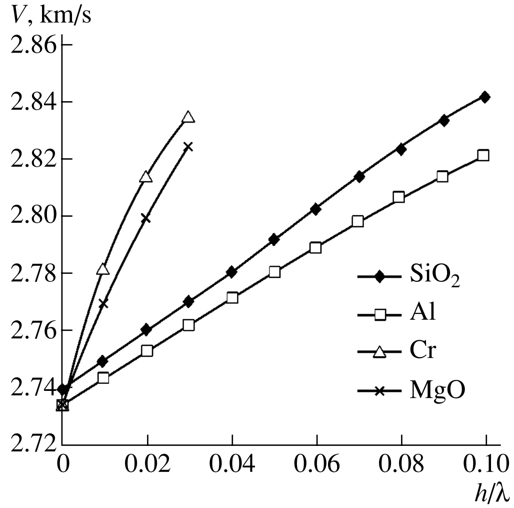


Fig. 1. SAW velocity vs. h/λ in (0, 140°, 24°) LGS covered by isotropic SiO_2, Al, Cr, and MgO films.

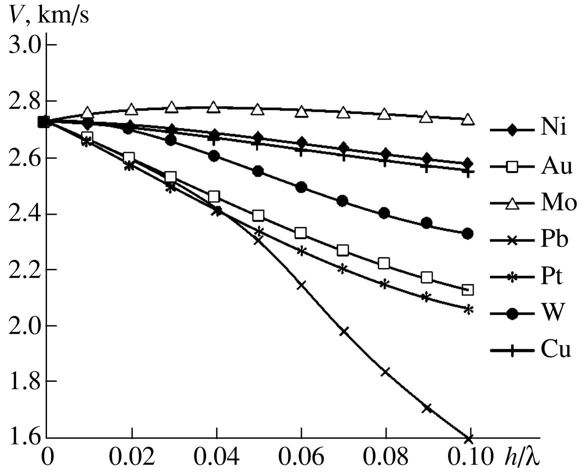


Fig. 2. The same as in Fig. 1 for Al, Au, Mo, Pb, Pt, W, and Cu films.

according to the material of the layers (metal, insulator, or piezoelectric).

Generally, the phase velocity of an EASW in a multilayer medium can be found with any method of searching a global extremum of a boundary objective function Φ [8]. In the Farnell–Jones approach, Φ is taken to be the boundary condition determinant [7]; in the Adler approach, it is the effective permittivity function [2]. Upon searching solutions for surface acoustic waves (SAWs), a global extremum of the boundary objective function Φ is found by scanning the wave's phase velocity V, which serves as a parameter. Upon searching for solutions for PSAWs, a global extremum of Φ is found by simultaneously scanning the phase velocity V and propagation losses δ. The phase velocity V found depends on the material properties of each of the layers and the piezoelectric substrate, as well as on the crystal orientation and wave propagation direction, which are usually described by the Eulerian angles ϕ, Θ, and Ψ [7]. Then, one can calculate all the other parameters of the wave (electromechanical coupling coefficient, temperature coefficient of velocity (delay), anisotropy parameter, and so on) [6].

SEARCH FOR MULTILAYER CONFIGURATIONS OPTIMAL FOR EASW PROPAGATION

Consider the mechanism of decreasing the dispersion of EASWs propagating in multilayer structures in greater detail. If one layer is applied on a piezoelectric substrate, the velocity of the acoustic wave changes, depending on the layer thickness, compared to the bare surface and becomes a function of frequency (wavelength λ). Using the above methods, we calculated the phase velocities of an SAW propagating in langasite (LGS) of orientation (0, 140°, 24°) with various isotropic films covering its surface (Figs. 1, 2). From Fig. 1, it follows that applied films of fused quartz (SiO_2), aluminum, chromium, and magnesia increase the SAW velocity with increasing h/λ (the film thickness h is normalized to the wavelength λ). Nickel, gold, molybdenum, lead, platinum, tungsten, and copper films, conversely, decrease the velocity. The material constants for these films were taken from [9, 10].

If two films variously changing the EASW phase velocity are applied on a piezosubstrate, their effects may compensate each other if the thicknesses of the films are appropriately selected. In this case, the dispersion may be considerably reduced and the wave velocity becomes virtually frequency (wavelength) independent. From Figs. 1 and 2, one can see that the EASW phase velocity vs. h/λ ($h \ll \lambda$) curves run in the opposite manner when the SiO_2 and nickel films are applied on (0, 140°, 24°) LGS piezocrystal. In both cases, the run is linear up to $h/\lambda \approx 0.1$. Therefore, with both films applied on the LGS surface, the EASW velocity will be virtually dispersionless.

Let us find a dispersionless solution for an EASW propagating in a multilayer system. When the films are thin ($h_i \ll \lambda$), the phase velocity varies with h/λ linearly in most cases (Figs. 1, 2). Therefore, in a first approximation, one may introduce a factor γ that characterizes the dispersion properties of a layer system. Then, the relative change in the EASW phase velocity can be written as

$$\frac{(V_0 - V(h))}{V_0} = \gamma\frac{h}{\lambda} = \frac{\gamma}{V_0}\omega. \tag{1}$$

Here, V_0 is the phase velocity in the absence of the layer, $V(h)$ is the velocity in the presence of the layer, and ω is the frequency. Generally, the law of EASW velocity variation with layer thickness may be nonlinear; therefore, the factor γ may vary with layer thick-

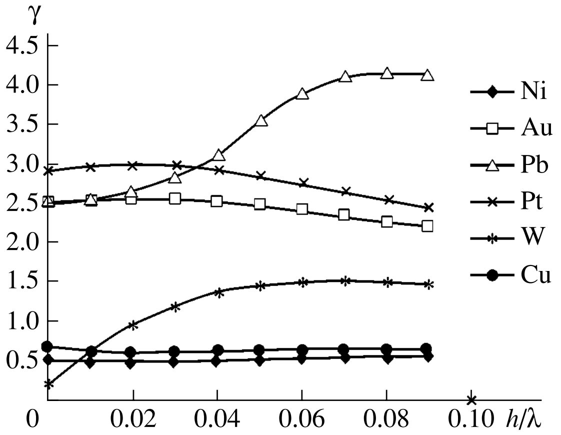


Fig. 3. Factor γ vs. h/λ for SAWs propagating in (0, 140°, 24°) LGS covered by isotropic Ni, Au, Pb, Pt, W, and Cu films.

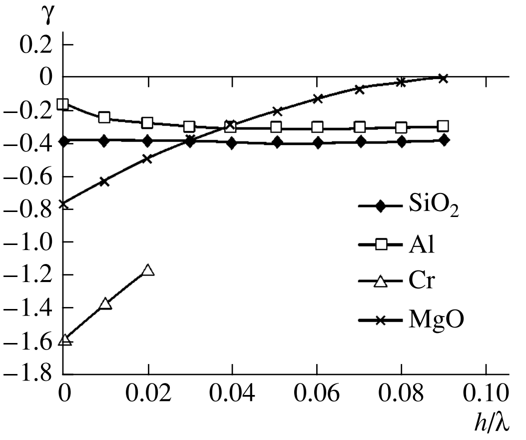


Fig. 4. The same as in Fig. 3 for SiO_2, Al, Cr, and Mo films.

ness. Its value depends on the material of the layer (of thickness h) and piezosubstrate orientation.

Figures 3 and 4 show the values of γ calculated for (0, 140°, 24°) LGS piezocrystal covered by films of various materials. It is seen that γ may be both positive (Fig. 3) and negative (Fig. 4).

If many layers are applied so that the condition

$$G = \sum_{i=1}^{N}\gamma_i h_i = 0 \tag{2}$$

is met (N is the number of layers and γ_i and h_i are the dispersion factor and thickness of an ith layer, respectively), one may again argue that the EASW velocity is dispersion-free.

Figure 5 shows the values of the parameter $G = (\gamma_1 h_1 + \gamma_2 h_2)$ calculated for the case when an SAW propagates in the SiO_2/Ni/(0, 140°, 24°) LGS system with various h_1/λ and h_2/λ (h_1 and h_2 are the thicknesses of the lower, Ni, and upper, SiO_2, films, respectively). A family of curves corresponding to the SAW dispersionless velocity is seen. For example, if the Ni film is $h_1 = 0.04\lambda$ thick, the SAW velocity dispersion is absent ($G = 0$) for a SiO_2 film thickness $h_2 = 0.05\lambda$ (the curve $h_2/\lambda = 0.05$ in Fig. 5). The SAW phase velocities calculated for each of the curves were as follows (in km/s): 2.7345 for the bare (0, 140°, 24°) LGS surface, 2.6844 for the Ni ($h_1/\lambda = 0.04$)/LGS system, 2.7908 for the SiO_2/Ni/LGS system, and 2.7305 for the SiO_2/Ni/LGS system. With the nickel film $h_1 = 0.015\lambda$ thick, the dispersion will be absent when the SiO_2 thickness is $h_2 = 0.02\lambda$ (the curve $h_2/\lambda = 0.02$ in Fig. 5), etc.

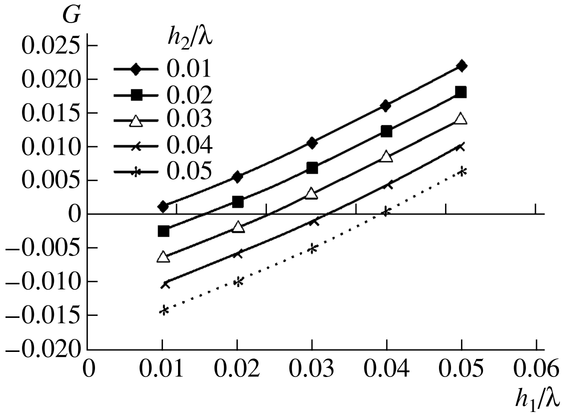


Fig. 5. Parameter G vs. h_1/λ at different h_2/λ for SAWs propagating in the SiO_2/Ni/(0, 140°, 24°) LGS system.

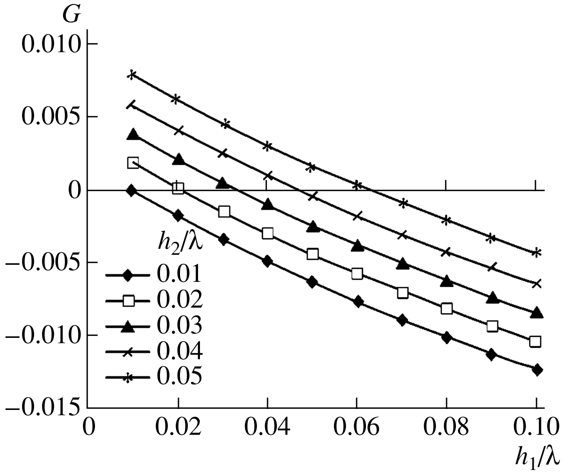


Fig. 6. The same as in Fig. 5 for PSAWs propagating in the Al/Cr/(0, –48°, 0) $LiTaO_3$ system.

One can also eliminate the velocity dispersion when PSAWs propagate in multilayer systems. Figure 6 shows the values of $G = (\gamma_1 h_1 + \gamma_2 h_2)$ calculated for the Al/Cr/(0, –48°, 0) $LiTaO_3$ system. Here, again, the PSAW velocity has no dispersion at certain relationships between the Al and Cr film thicknesses. For example, $G = 0$ if the chromium thickness is $h_1 = 0.03\lambda$ and the aluminum thickness is $h_2 = 0.04\lambda$, etc. The PSAW parameters calculated for each of the cases were as follows (V in km/s, δ in dB/λ: $V_m = 4.11485$, $\delta_m = 0.012$, $V_b = 4.226$, $\delta_b = 0.00075$, $K^2 = 5.6\%$, $pfa = 0$, $\beta = -2.7$, and $tcd = 63.2 \times 10^{-6}$ /°C for bare (subscript b) and metallized (m) (0, –48°, 0) $LiTaO_3$, respectively; $V_m = 4.1366$ and $\delta_m = 0.04$ for the one-layer Cr($h_1/\lambda = 0.03$)/$LiTaO_3$ system; $V_m = 4.0829$ and $\delta_m = 0.01$ for the one-layer Al($h_2/\lambda = 0.04$)/$LiTaO_3$ system; and $V_m = 4.1099$, $\delta_m = 0.04$, $pfa = 0$, $\beta = -2.56$, and $tcd = 63.7 \times 10^{-6}$ /°C for the two-layer Al/Cr/$LiTaO_3$ system. The material constants for $LiTaO_3$ were taken from [11].

Thus, using the method described above and carrying out a numerical experiment, one may select a piezosubstrate of certain orientation, layer materials, and their thicknesses such that the EASW velocity is dispersionless.

It should also be noted that, when selecting the layer materials and thicknesses that minimize the EASW velocity dispersion, one can simultaneously find their combinations optimizing other EASW parameters. To this end, it is necessary to write the objective function Φ as a linear combination of EASW parameters with individual weighting coefficients a_i, which specify the contribution of each of the terms to the objective function [12]:

$$\Phi(\phi, \Theta, \Psi, h_i/\lambda, \gamma) = \min\{a_1 pfa + a_2 tcd + a_3(\beta + 1) + a_4 G + \ldots\}. \tag{3}$$

By applying the procedure of searching a global extremum of the objective function Φ of many variables [12], one can, by exhaustively searching the objective function variables ϕ, Θ, Ψ, h_i/λ, and γ with a minimal step in a cycle, optimize the EASW parameters in a multilayer system by means of numerical experiment.

Using such an approach, we constructed an Al/Ni/(0, 90°, 0) *YX* piezoquartz multilayer system where a propagating SAW is temperature compensated ($tcd \approx 0$) and dispersion free ($G \approx 0$). Note that the Ni film applied on the *YX* cut of piezoquartz reduces markedly the SAW velocity in this system (the phase velocity is $V = 2.86$ km/s at $h_1 = 0.05\,\lambda$) compared with the bare surface of piezoquartz ($V_b = 3.16$ km/s, $V_m = 3.157$ km/s, $tcd = -22 \times 10^{-6}$ /°C. On the contrary, the Al film on the *YX* cut of piezoquartz somewhat increases the SAW velocity ($V = 3.1602$ km/s at $h_1 = 0.05\lambda$). In constructed two-layer Al($h_2 = 0.05\lambda$)/Ni($h_1 = 0.09\lambda$)/*YX* piezoquartz (the phase velocity $V = 2.706$ km/s, $tcd = 0.9 \times 10^{-6}$/°C, $G = 0$) and Al ($h_2 = 0.05\lambda$/Ni($h_1 = 0.005\lambda$)/*YX* piezoquartz (the phase velocity $V = 3.13$ km/s, $tcd = -0.2 \times 10^{-6}$ /°C, $G = 0.001$) systems, a propagating SAW has no velocity dispersion and is temperature compensated.

CONCLUSIONS

We suggest a numerical method of EASW parameter analysis and optimization in multilayer structures. Particular combinations of films that optimize the characteristics of propagating electroacoustic waves are found. The efficiency of the approach suggested is demonstrated with lithium tantalate, piezoquartz, and langasite (new promising piezoelectric crystal) substrates.

REFERENCES

1. R. Schmidt and F. Voltmer, IEEE Trans. Microwave Theory Tech. **17**, 920 (1969).
2. E. L. Adler, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **41**, 876 (1994).
3. R. Raghavan, J. Vetelino, and A. Jhunjunwala, *Proceedings of the IEEE Ultrason. Symp., 1979*, p. 606.
4. M. Yu. Dvoesherstov, V. I. Cherednick, A. P. Chirimanov, *et al.*, Akust. Zh. **47**, 485 (2001) [Acoust. Phys. **47**, 415 (2001)].
5. P. Wallner, W. Ruile, and R. Weigel, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **47**, 1235 (2000).
6. *Acoustic Surface Waves*, Ed. by A. A. Oliner (Springer-Verlag, New York, 1978; Mir, Moscow, 1981).
7. *Surface Wave Filters: Design, Construction, and Use*, Ed. by H. Mathews (Wiley, New York, 1977; Radio i Svyaz', Moscow, 1981).
8. M. Yu. Dvoesherstov, V. I. Cherednick, and A. P. Chirimanov, Izv. Vyssh. Uchebn. Zaved. Radiofiz. **43**, 801 (2000).
9. S. Ballandras, E. Gavignet, E. Bigler, *et al.*, Appl. Phys. Lett. **71**, 1625 (1997).
10. Yu. Pisarevsky, P. Senushencov, P. Popov, *et al.*, in *Proceedings of the IEEE Frequency Control Symposium, 1995*, p. 653.
11. G. Kovacs, M. Anhorn, H. Engan, *et al.*, *Proceedings of the IEEE Ultrason. Symp., 1990*, p. 435.
12. M. Yu. Dvoesherstov, S. G. Petrov, V. I. Cherednick, *et al.*, Zh. Tekh. Fiz. **72** (8), 103 (2002) [Tech. Phys. **47**, 1032 (2002)].

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1323–1329. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 106–112.
Original Russian Text Copyright © 2003 by Cherednick, Dvoesherstov.

ACOUSTIC, ACOUSTOELECTRONICS

Numerical Calculation of the Parameters of Surface and Pseudosurface Acoustic Waves in Multilayer Structures

V. I. Cherednick and M. Yu. Dvoesherstov
Lobachevski State University, Nizhni Novgorod, pr. Gagarina 23, 603950 Russia
e-mail:017Cherednik@ichem.unn.runnet.ru, Dvoesh@rf.unn.ru

Received February 4, 2003

Abstract—A method for numerically calculating the parameters of surface and pseudosurface electroacoustic waves propagating in multilayer piezoelectrics is suggested. The feasibility of wave parameter optimization by using various materials of the layers and piezoelectric substrate is demonstrated with particular examples. *© 2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The application of various layers on a piezoelectric substrate is a way of improving the parameters of propagating surface acoustic waves (SAWs). For example, a metal film of certain thickness may thermally suppress the SAW propagation when the orientation of the substrate provides a high piezoelectric coupling. Using the overlayer, one can vary the wave propagation velocity and, hence, the operating frequency of a piezoelectric device. The effect of the environment (gas or liquid) on the properties of the wave and overlayer is used in related sensors. Finally, the layer may protect the piezoelectric surface against undesired external impacts. One more advantage of multilayer compositions is the reduction of a velocity dispersion, which is observed in single-layer structures. Therefore, analysis and optimization of the SAW parameters in multilayer structures seems to be topical. Various aspects of this problem are discussed elsewhere [1–4].

In this paper, we formulate boundary conditions for insulating, metallic, and piezoelectric insulating overlayers and for a piezoelectric substrate of any crystal symmetry. Also, a general method for numerically calculating the parameters of SAWs propagating in multilayer piezoelectric crystal structures is reported.

STATEMENT OF THE PROBLEM AND BOUNDARY CONDITIONS

The parameters of SAWs propagating in the multilayer structures depend on the properties of the substrate and each of the layers. It is necessary to solve the set of piezoacoustic equations [5]

$$\rho\frac{\partial^2 u_j}{\partial t^2} - c_{ijkl}\frac{\partial^2 u_k}{\partial X_i \partial X_l} = e_{kij}\frac{\partial^2 \varphi}{\partial X_k \partial X_i},$$
$$\varepsilon_{ik}\frac{\partial^2 \varphi}{\partial X_i \partial X_k} = e_{ikl}\frac{\partial^2 u_k}{\partial X_i \partial X_l}; \quad i, j, k, l = 1, 2, 3, \tag{1}$$

where X_i are coordinates; c_{ijkl}, e_{ijk}, and ε_{ij} are the tensors of the piezoelectric, elastic, and dielectric constants; ρ is the density of the medium; u_i are mechanical displacements; φ is the electrical potential; and t is time. The set of Eqs. (1) must be solved for each of the system's component (Fig. 1). A solution to set (1) is sought in the form

$$u_j = \alpha_j \exp(ikb_3X_3)\exp[ik(b_1X_1 - Vt)],$$
$$\varphi = \alpha_4 \exp(ikb_3X_3)\exp[ik(b_1X_1 - Vt)]; \tag{2}$$
$$j = 1, 2, 3.$$

Here, α_j are the mechanical displacement amplitudes, α_4 is the electric potential; k is the wavenumber; V is the wave velocity; and b_1 and b_3 are the coefficients that relate the wave amplitude to the coordinates X_1 and X_3, respectively. In the general case, the coefficient b_1 may be represented as $b_1 = 1 + i\delta$, where δ is a real positive quantity that has the meaning of the decay of the wave along the propagation direction. For surface acoustic waves, the decay $\delta = 0$; for pseudosurface acoustic waves, $\delta > 0$ [6]. In the general case, substituting (2) into (1) yields a set of Christoffel equations, from which one, knowing the wave velocity V and the decay coefficient δ, can find eight complex roots $b_3^{(n)}$ (n = 1–

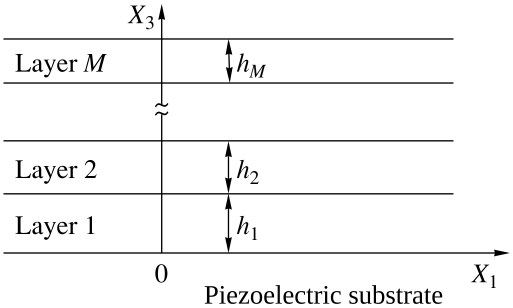


Fig. 1. Multilayer structure.

1063-7842/03/4810-1323$24.00 © 2003 MAIK "Nauka/Interperiodica"

8) and the associated set of complex amplitudes $\alpha_1^{(n)}$, $\alpha_2^{(n)}$, $\alpha_3^{(n)}$, and $\alpha_4^{(n)}$. For each of the structure constituents, a general solution is represented as a linear combination of partial waves in this constituent:

$$u_j^{(m)} = \sum_{n = N_{m-1}+1}^{N_m} C_n \alpha_j^{(n)} \exp[ik(b_i^{(n)} X_i - Vt)],$$
$$\varphi^{(m)} = \sum_{n = N_{m-1}+1}^{N_m} C_n \alpha_4^{(n)} \exp[ik(b_i^{(n)} X_i - Vt)]; \tag{3}$$
$$i, j = 1, 2, 3.$$

Here, $b_1^{(n)} = 1 + i\delta$, $b_2^{(n)} = 0$, $b_3^{(n)}$ is a set of the roots of the polynomial equation (specific for each of the media), $N_m = n_0 + n_1 + \ldots + n_m$, n_m is the number of partial modes in a medium of number m ($m = 0$ refers to the substrate; $m = 1$, to the first layer, etc.; $N_{0-1} = n_{0-1} = 0$), and C_n are unknown amplitude coefficients.

When analyzing surface or pseudosurface acoustic waves, one must properly select the roots $b_3^{(n)}$ for the substrate. For the layers, the rule of root selection is straightforward: for each of them, a solution must be constructed as a linear combination of the partial modes. The unknown coefficients C_n of the linear combination in (3) are found from boundary conditions at the interfaces and at the outer boundary of the top layer. Unfortunately, general boundary conditions that are applicable to any combination of the substrate and layer materials cannot be stated in principle.

1. PIEZOELECTRIC LAYERS ON A PIEZOELECTRIC SUBSTRATE

In this case, three components of mechanical displacements u_j, three normal components of the mechanical stress tensor T_{3j}, the electrical potential φ, and the normal component of the electric field induction D_3 must satisfy the continuity conditions at the interfaces. The conditions at the boundary between mth and (m + 1)th media are as follows:

$$\sum_{n = N_{m-1}+1}^{N_m} C_n \alpha_j^{(n)} \exp[ikb_3^{(n)} X_3^{(m)}]$$
$$= \sum_{n = N_m+1}^{N_{m+1}} C_n \alpha_j^{(n)} \exp[ikb_3^{(n)} X_3^{(m)}], \tag{4a}$$

$$\sum_{n = N_{m-1}+1}^{N_m} C_n (c_{3jkl} \alpha_k^{(n)} b_l^{(n)} + e_{k3j} \alpha_4^{(n)} b_k^{(n)}) \exp[ikb_3^{(n)} X_3^{(m)}]$$
$$= \sum_{n = N_m+1}^{N_{m+1}} C_n (c_{3jkl} \alpha_k^{(n)} b_l^{(n)} + e_{k3j} \alpha_4^{(n)} b_k^{(n)}) \exp[ikb_3^{(n)} X_3^{(m)}], \tag{4b}$$

$$\sum_{n = N_{m-1}+1}^{N_m} C_n \alpha_4^{(n)} \exp[ik(b_3^{(n)} X_3^{(m)})]$$
$$= \sum_{n = N_m+1}^{N_{m+1}} C_n \alpha_4^{(n)} \exp[ik(b_3^{(n)} X_3^{(m)})], \tag{4c}$$

$$\sum_{n = N_{m-1}+1}^{N_m} C_n (e_{3jk} \alpha_j^{(n)} b_k^{(n)} - \varepsilon_{3j} \alpha_4^{(n)} b_j^{(n)}) \exp[ikb_3^{(n)} X_3^{(m)}]$$
$$= \sum_{n = N_m+1}^{N_{m+1}} C_n (e_{3jk} \alpha_j^{(n)} b_k^{(n)} - \varepsilon_{3j} \alpha_4^{(n)} b_j^{(n)}) \exp[ikb_3^{(n)} X_3^{(m)}]. \tag{4d}$$

Here, $m = 0, 1, 2, \ldots, M - 1$, where M is the number of layers; $X_3^{(m)} = h_1 + h_2 + \ldots + h_m$; and $X_3^{(0)} = 0$. Equations (4a)–(4d) are the continuity conditions for mechanical displacements, stresses, potential, and electric field induction.

If any surface $X_3 = X_3^{(m)}$ is covered by a metal layer of infinitesimal thickness and short-circuited, Eqs. (4c) and (4d) change. The right of (4c) and the left of (4d) vanish, and the right of (4d) is replaced by the right of (4c).

The potential $\varphi^{(\mathrm{f})}$ in free space must meet the Laplace equation and decrease down to zero with $X_3 \longrightarrow \infty$. These conditions are satisfied if $\varphi^{(\mathrm{f})}$ is taken in the form

$$\varphi^{(\mathrm{f})} = \varphi^{(M)} e^{-kb_1(X_3 - X_3^{(M)})}, \quad X_3 \geq X_3^{(M)}. \tag{5}$$

Here, $\varphi^{(M)}$ is the potential at the outer boundary of the top layer ($X_3 = X_3^{(M)}$). Eventually, for the top surface uncovered, we obtain the electrical boundary condition

$$i \sum_{n = N_{M-1}+1}^{N_M} C_n (e_{3jk} \alpha_j^{(n)} b_k^{(n)} - \varepsilon_{3j} \alpha_4^{(n)} b_j^{(n)}) \exp[ikb_3^{(n)} X_3^{(M)}]$$
$$= b_1 \varepsilon_0 \sum_{n = N_{M-1}+1}^{N_M} C_n \alpha_4^{(n)} \exp[ikb_3^{(n)} X_3^{(M)}]. \tag{6}$$

For the short-circuited top boundary, the electrical boundary condition may be derived from Eq. (4c) with $m = M$ and zero on the right.

When all the constituents of the system are piezoelectrics, the complete set of boundary conditions contains $n_0 + n_1 + n_2 + \ldots + n_M$ equations for the same number of the unknowns C_n. For such a system, $n_0 = 4$ and $n_1 = n_2 = \ldots = n_M = 8$ in the general case.

2. METALLIC LAYERS ON A PIEZOELECTRIC SUBSTRATE

The situation where the first layer is metallic and the others are either metallic or insulating in any combination may also be assigned to this case. Here, the mechanical boundary conditions at the interfaces remain the same as above and the electrical boundary conditions simplify to

$$\sum_{n=1}^{n_0} C_n \alpha_4^{(n)} = 0. \tag{7}$$

3. ISOTROPIC INSULATING LAYERS ON A PIEZOELECTRIC SUBSTRATE

Let a piezoelectric substrate support M isotropic insulating layers with permittivities ε_m. Then, electric boundary conditions become awkward, since any of the interfaces may by either open or short-circuited. In the general case, the electrical potential inside an mth layer depends on X_3 as follows:

$$\begin{gathered}\varphi^{(m)}(X_3) = A_m e^{-kb_1(X_3 - X_3^{(m-1)})} + B_m e^{kb_1(X_3 - X_3^{(m-1)})}, \\ X_3^{(m-1)} \le X_3 \le X_3^{(m)}.\end{gathered} \tag{8}$$

Having defined all the coefficients A_m and B_m through the interfacial potentials, one may then, using the continuity conditions for the potential and normal component of the electric induction vector at each of the interfaces, eliminate the interfacial potentials and obtain the X_3 dependence of the potential $\varphi^{(1)}$ ($\varphi^{(1)}$ is the potential of the first layer) that involves only $\varphi^{(0)}(X_3 = 0)$, which is the potential on the substrate surface. The potential $\varphi^{(1)}$ appears in the expression for the normal component of the electric field induction in the first layer:

$$D_3^{(1)} = -\varepsilon_1 \varepsilon_0 \frac{d\varphi^{(1)}}{dX_3}. \tag{9}$$

The induction calculated by (9) for $X_3 = 0$ should now be set equal to the induction calculated for the substrate at the same X_3. This yields a single electrical boundary condition for isotropic insulating layers on a piezoelectric substrate. Its shape will depend considerably on the number of layers and on the electrical state of the interfaces (open or short-circuited).

3.1. All interfaces are open. If there is a single layer, electrical boundary conditions have the form

$$\begin{aligned} i\sum_{n=1}^{n_0} C_n(e_{3jk}\alpha_j^{(n)} b_k^{(n)} - \varepsilon_{3j}\alpha_4^{(n)} b_j^{(n)}) \\ = \frac{b_1\varepsilon_1\varepsilon_0}{\sinh(kb_1h_1)} S_1 \sum_{n=1}^{n_0} C_n \alpha_4^{(n)}, \end{aligned} \tag{10a}$$

where

$$\begin{aligned} S_1 = \cosh(kb_1h_1) \\ - \frac{\varepsilon_1}{\varepsilon_1\cosh(kb_1h_1) + R_2\sinh(kb_1h_1)}. \end{aligned} \tag{10b}$$

In (10b),

$$R_2 = \frac{\varepsilon_2}{\sinh(kb_1h_2)} S_2 \tag{11}$$

is the recursive coefficient that makes it possible to derive a formula for two layers from the formula for one layer; that is, for two layers, the electrical boundary condition takes the form

$$\begin{aligned} i\sum_{n=1}^{n_0} C_n(e_{3jk}\alpha_j^{(n)} b_k^{(n)} - \varepsilon_{3j}\alpha_4^{(n)} b_j^{(n)}) = \frac{b_1\varepsilon_1\varepsilon_0}{\sinh(kb_1h_1)} \\ \times \left[\cosh(kb_1h_1) - \frac{\varepsilon_1}{\varepsilon_1\cosh(kb_1h_1) + \dfrac{\varepsilon_2\sinh(kb_1h_1)}{\sinh(kb_1h_2)} S_2}\right] \\ \times \sum_{n=1}^{n_0} C_n \alpha_4^{(n)}, \end{aligned} \tag{12a}$$

where

$$\begin{aligned} S_2 = \cosh(kb_1h_2) \\ - \frac{\varepsilon_2}{\varepsilon_2\cosh(kb_1h_2) + R_3\sinh(kb_1h_2)}. \end{aligned} \tag{12b}$$

The recursive coefficient

$$R_3 = \frac{\varepsilon_3}{\sinh(kb_1h_3)} S_3 \tag{13}$$

allows one to obtain a formula for three layers from the formula for two layers; that is, for three layers, we have

$$i\sum_{n=1}^{n_0} C_n\left(e_{3jk}\alpha_j^{(n)}b_k^{(n)} - \varepsilon_{3j}\alpha_4^{(n)}b_j^{(n)}\right) = \frac{b_1\varepsilon_1\varepsilon_0}{\sinh(kb_1h_1)}$$

$$\times\left\{\cosh(kb_1h_1) - \frac{\varepsilon_1}{\varepsilon_1\cosh(kb_1h_1) + \dfrac{\varepsilon_2\sinh(kb_1h_1)}{\sinh(kb_1h_2)}\left[\cosh(kb_1h_2) - \dfrac{\varepsilon_2}{\varepsilon_2\cosh(kb_1h_2) + \dfrac{\varepsilon_3\sinh(kb_1h_2)}{\sinh(kb_1h_3)}S_3}\right]}\right\}\sum_{n=1}^{n_0} C_n\alpha_4^{(n)}, \qquad (14)$$

$$S_3 = \cosh(kb_1h_3) - \frac{\varepsilon_3}{\varepsilon_3\cosh(kb_1h_3) + R_4\sinh(kb_1h_3)}. \qquad (14b)$$

An electrical boundary condition for (m + 1) isotropic insulating layers with open boundaries can be derived from a formula for m layers with a recursive coefficient R_{m+1}. If the number of layers is M, $R_{M+1} = 1$. To derive an electrical boundary condition for M layers, it is necessary to sequentially derive expressions for one, two, etc. layers until a formula for M layers is derived.

3.2. Short-circuited interface. If one of the interfaces, $X_3 = X_3^{(m)}$, is covered by a metallic layer of infinitesimal thickness and short-circuited, the electrical states of all interfaces further from the substrate are of no significance, since the electric field beyond the short-circuited interface ($X_3 > X_3^{(m)}$) is zero. The result will be the same if m layers counted from the substrate are insulating, the (m + 1)th layer is metallic, and all other more distant layers are both insulating and metallic in any combination.

If the substrate surface is short-circuited ($X_3 = 0$), the electrical boundary condition is defined by expression (7). If any other interface is short-circuited, $X_3 = X_3^{(m)}$, the electrical boundary condition can be easily obtained as follows. Using the recursive coefficients R_i ($i = 1, 2, 3, \ldots, m$), one obtains an expression for m insulating open-boundary layers as was done above. Then, the first term $\cosh(kb_1h_m)$ in the expression for S_m (i.e., for the layer preceding the short-circuited interface) is left, while the second term with R_{m+1} in the denominator is set equal to zero. The resulting expression corresponds to the zero potential on the surface $X_3 = X_3^{(m)}$. For example, if the outer boundary of the second layer is short-circuited, $X_3 = X_3^{(2)}$ (that is, $\varphi^{(2)} = 0$), we find that the electrical boundary condition coincides with (12a) but (12b) should be replaced by $S_2 = \cosh(kb_1h_2)$. The total number of isotropic layers $M > m$ in this case, and the electrical states of all more distant surfaces are of no significance for electrical boundary conditions.

SOLUTION OF EQUATIONS FOR BOUNDARY CONDITIONS

Having stated all the boundary conditions, we come to a set of linear homogeneous equations for the unknown coefficients C_n:

$$\begin{aligned} a_{11}C_1 + a_{12}C_2 + \ldots + a_{1N}C_N &= 0, \\ a_{21}C_1 + a_{22}C_2 + \ldots + a_{2N}C_N &= 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{N1}C_1 + a_{N2}C_2 + \ldots + a_{NN}C_N &= 0. \end{aligned} \qquad (15)$$

The number of equations and the number of unknowns (N) equal to the total number of partial modes in all the system's components. Set (15) may have a nontrivial solution only if the complex determinant made up of the coefficients of this set equals zero. To find the zero of the boundary condition determinant, we applied the global search method [6] to find an extremum of a function of two variables (velocity V and coefficient δ). As an objective function, we used the absolute value of the boundary condition determinant squared ($|d|^2$).

The velocity V and coefficient δ can also be determined by the Adler technique [7], which is based on effective permittivity analysis. Let all the layers are piezoelectric. Then, the electric boundary condition at the outer boundary of the top layer (of number M) is stated either as the continuity of the normal component $D_3^{(M)}$ of the induction vector for the open surface or as the continuity (vanishing) of the potential $\varphi^{(M)}$ for the short-circuited surface. The induction and potential may be combined into one quantity as [7]

$$\varepsilon_{\text{eff}} = \frac{D_3^{(M)}}{kb_1\varphi^{(M)}}. \qquad (16)$$

The wavenumber k here is added to match the dimensions of the left- and right-hand sides. The coefficient b_1 appears if pseudosurface waves are considered. In such a representation, the continuity condition for the induction or potential is replaced by the equivalent continuity condition for the effective permittivity. For an open surface, the effective permittivity is ε_0; for a short-circuited one, it is infinitely large (the permittivity of a metal). Having substituted potential (3) and induction (9) for the Mth layer into (16), we arrive at

$$\frac{i\sum\limits_{n=N_{M-1}+1}^{N_M} C_n(e_{3jk}\alpha_j^{(n)}b_k^{(n)} - \varepsilon_{3j}\alpha_4^{(n)}b_j^{(n)})\exp[ikb_3^{(n)}X_3^{(M)}]}{b_1\varepsilon_0\sum\limits_{n=N_{M-1}+1}^{N_M} C_n\alpha_4^{(n)}\exp[ikb_3^{(n)}X_3^{(M)}]} \qquad (17)$$

$$= \begin{cases}\varepsilon_0\\ \infty.\end{cases}$$

The upper equality in (17) is totally equivalent to Eq. (6). The lower one is equivalent to Eq. (4c) if $m = M$ and the right-hand side is set equal to zero. The absolute value of (17) squared is used as an objective function in searching for the velocity V and the decay coefficient δ. For an open surface, the solution corresponds to the zero (minimum) of this function; for a short-circuited surface, to the pole. With piezoelectric layers, the use of the effective permittivity is valid not only for the outer surface of the top layer but also for any interface. In the latter case, when considering an open surface, one must take equal values of the effective permittivity on both sides of the surface. For a short-circuited surface, the potential of this surface should be set equal to zero. Then, for the construction of the effective permittivity, it is necessary to use the appropriate equation from set (15). An objective function for isotropic insulating layers can be constructed in the same way. In this case, we have only one equation of electrical boundary conditions, which is the continuity equation for the normal component of the induction on the substrate surface. This equation is found in view of the continuity of both electrical parameters (potential and field) on all other surfaces (for the open substrate surface) or in view of the zero potential on the short-circuited substrate surface.

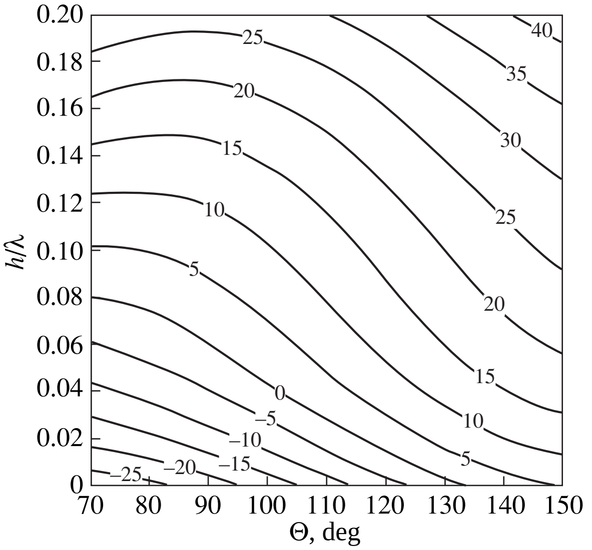


Fig. 2. Dependence of the *tcd* on Θ and reduced thickness *h*/λ of the aluminum layer on piezoquartz ($\varphi = 0$, $\Theta = 70°$–$150°$, $\psi = 0$). The figures by the curves are *tcd* in 10^{-6} 1/°C.

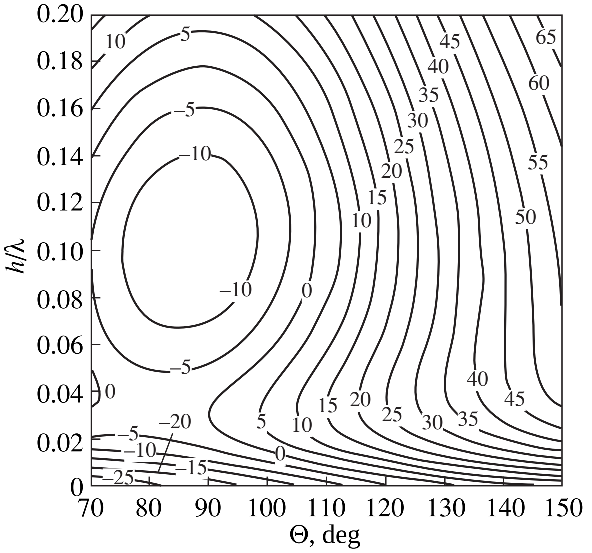


Fig. 3. The same as in Fig. 2 for gold on piezoquartz.

COMPUTATIONAL RESULTS

Based on the above approach, we implemented an algorithm for computing the basic parameters of electroacoustic surface waves propagating in multilayer structures. Figures 2 and 3 demonstrate the effect of individual layers of various metals on the temperature dependences of piezoquartz with the second Eulerian angle [5] Θ ranging from 70° to 150° and the first and third Eulerian angles being zero. These ranges of Eulerian angles correspond to piezoquartz orientations (*YX*, *AT*–*X*, *ST*–*X*, et al.) that are most widely used in related devices. Specifically, the two-dimensional dependences of the temperature coefficient of delay (*tcd*) on Θ and on the normalized (to the wavelength λ) Al and Au layer thickness *h* are shown. The material constants for Al and Au were taken from [8]. From Figs. 2 and 3 it follows that negative values of *tcd* can be compen-

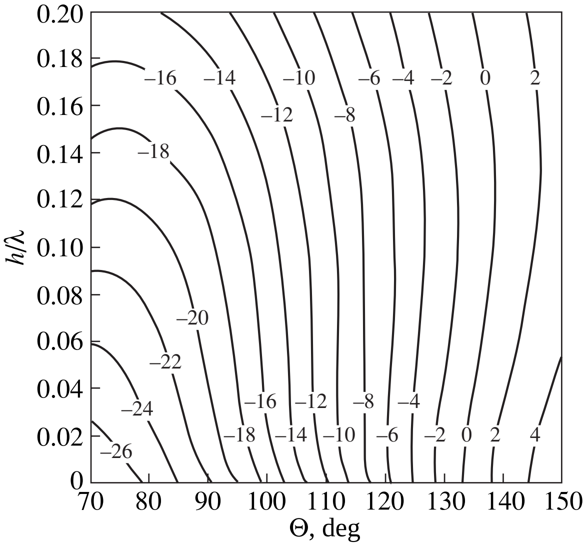


Fig. 4. The same as in Fig. 2 for an isotropic fused quartz layer on piezoquartz.

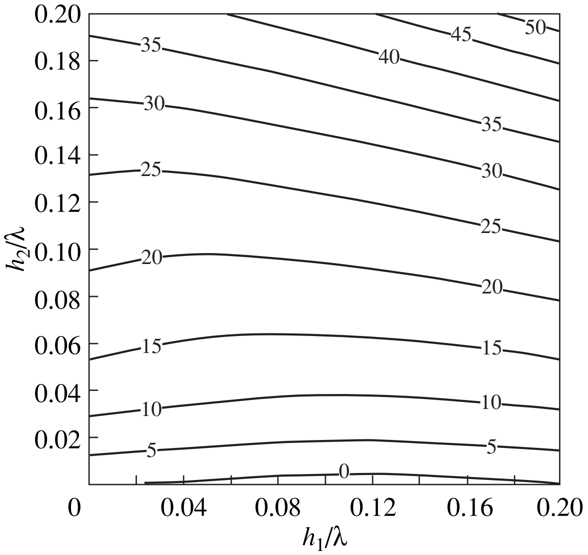


Fig. 5. Dependence of the *tcd* on the reduced thicknesses of isotropic fused quartz, h_1/λ, and aluminum layer, h_2/λ, on *ST–X* quartz (0, 132.75°, 0). The figures by the curves are *tcd* in 10^{-6} 1/°C.

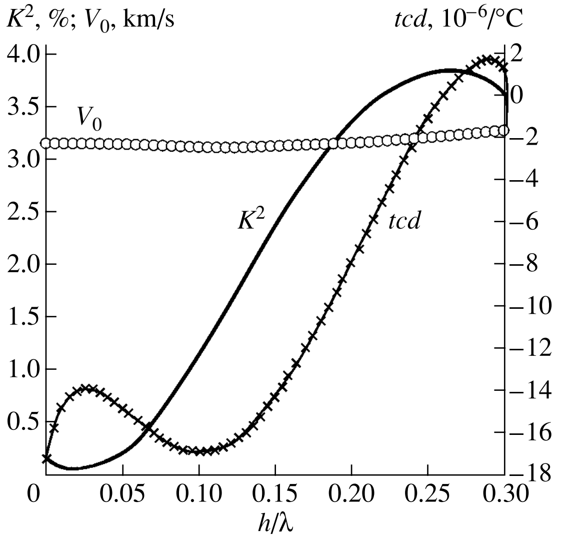


Fig. 6. *tcd*, K^2, and V_0 vs. the thickness h/λ of $LiNbO_3$ (0, 38°, 0) on the quartz substrate (0, 100°, 0).

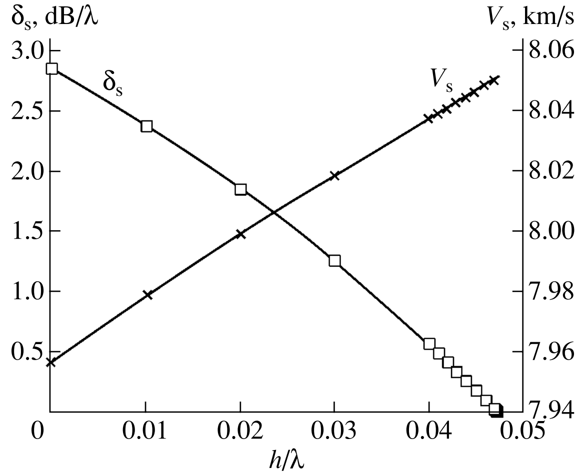


Fig. 7. Propagation losses δ_s and velocity V_s of the second-order pseudosurface wave vs. the thickness h/λ of the Al layer on the $LiNbO_3$ crystal surface (0, –49°, 0).

sated for by a metal layer of certain thickness. For example, the orientation *YX* (Θ = 90°) can be stabilized by applying an Al layer about 0.06λ thick on the quartz surface. Figure 4 shows the effect of an isotropic insulating fused quartz layer on the temperature behavior of the same crystal orientations. The effect is seen to be insignificant: the related isolines are arranged largely normally to the Θ axis. This means that fused quartz is promising as a protective layer against undesired environmental (mechanical and chemical) impacts that almost does not change the temperature properties of the system (in particular, the thermal stability of the orientation *ST–X*, Θ = 132.75°, remains practically the same).

A thin metal layer applied on an isotropic insulator provides protection against electrical fields. The combined effect of an isotropic fused quartz layer covered by an aluminum layer is illustrated in Fig. 5, which shows the two-dimensional dependence of the *tcd* on the thicknesses of the fused quartz layer and aluminum overlayer in the crystal–fused quartz–aluminum system. The effect of the Al layer is significant; therefore, its thickness must be small so as not to break the thermal stability of the given orientation. Figure 6 plots the wave velocity V_0, temperature coefficient of delay *tcd*, and electromechanical coupling coefficient K^2 (an interdigital transducer on the layer surface) vs. the thickness of lithium niobate $LiNbO_3$ (0, 38°, 0) applied on the open-surface quartz substrate (0, 100°, 0). It is seen that K^2 reaches a value of 3.85% and *tcd* goes to zero near $h/\lambda = 0.26$. Thus, a $LiNbO_3$ layer on quartz may provide a combination of high K^2, which is typical of lithium niobate, and a high thermal stability, which is inherent in quartz. Finally, let us consider the influence of a metal layer on the properties of pseudosurface waves propagating in $LiNbO_3$ (0, –49°, 0). For this orientation, a second-order pseudosurface wave has the following parameters: $V_s = 7.9576$ km/s, $V_0 = 8.3144$ km/s, $\delta_s = 2.865$ dB/λ, $\delta_0 = 0.531$ dB/λ, and $K^2 = 8.58\%$. The subscripts 0 and s refer to the open and short-circuited surfaces, respectively. An aluminum layer of finite thickness decreases losses. The dependences of δ_s and V_s on the Al layer thickness are shown in Fig. 7. As h/λ grows from zero to 0.047, the losses decrease from 2.865 dB/λ to about 10^{-3} dB/λ. It appears that the metal layer changes the wave propagation conditions so that the angle at which the wave goes inward to the crystal diminishes.

CONCLUSIONS

A general method for numerically calculating the parameters of surface and pseudosurface acoustic waves propagating in multilayer structures is suggested. The structures involve piezoelectric, insulating, and metal layers on a semi-infinite substrate of any crystal orientation. The effect of layers of different materials on electroacoustic surface wave propagation is studied. In a number of special cases, the thermal stability, electromechanical coupling coefficient, and propagation losses for pseudosurface waves may be improved. Various coatings in various combinations may improve considerably the characteristics of acoustoelectron devices.

REFERENCES

1. R. S. Raghavan, J. T. Vetelino, and A. Jhunjunwala, Proc.-IEEE Ultrason. Symp. 606 (1979).
2. S. Camou, Th. Pastureaud, S. Ballandras, *et al.*, Electron. Lett. **37**, 1053 (2001).
3. T. Eng Leong, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **49**, 929 (2002).
4. P. Wallner, W. Ruile, and R. Weigel, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **47**, 1235 (2000).
5. J. Campbell and W. Jones, IEEE Trans. Sonics Ultrason. **SU-15**, 209 (1968).
6. M. Yu. Dvoesherstov, V. I. Cherednik, A. P. Chirimanov, *et al.*, Proc. SPIE **3900**, 283 (1999).
7. E. L. Adler, IEEE Trans. Ultrason. Ferroelectr. Freq. Control **41**, 876 (1994).
8. S. Ballandras, E. Gavignet, E. Bigler, *et al.*, Appl. Phys. Lett. **71**, 1625 (1977).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1330–1332. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 113–115.
Original Russian Text Copyright © 2003 by Martirosyan.

ELECTRON AND ION BEAMS, ACCELERATORS

Effect of Fringing Fields in Storage Rings

Yu. L. Martirosyan
CANDLE Center of Synchrotron Radiation, Yerevan, 375040 Armenia
e-mail: martirosyan@asls.candle.am

Received December 27, 2002

Abstract—Electron beams in storage rings of third-generation synchrotron radiation sources feature a low emittance and extended lifetime. The provision of such characteristics requires a detailed study of higher order effects related to magnetic fields and cooperative effects associated with beam density. Fringing fields, being an unavoidable attribute of magnets of any type, may significantly affect the beam dynamics, since they appear in equations of particle motion of first and higher orders. A simple technique for evaluating the effect of fringing fields on the beam dynamics is suggested. Numerical results obtained with this technique for the quadrupoles of the CANDLE storage ring [1] are reported. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

In this work, a new approach to studying the effect of fringing fields on the beam dynamics is suggested. In designing large accelerator rings, the fringing fields of dipole and quadrupole magnets are usually ignored; i.e., the so-called hard-edge model is used. The matrix description of the beam dynamics provide vast information on the particle motion and makes it possible to estimate basic dynamic parameters. While mathematically correct, such an approach fails from the physical point of view, since magnetic fields do not satisfy the Maxwell equations in this case. More detailed investigations show that fringing fields may cause aberrations of different orders, have an adverse effect on the dynamic aperture, shift the betatron oscillation frequency, etc. This accounts for recently sharpened interest in exploring fringing fields in accelerators (see [2] and references cited therein).

Upon studying the beam dynamics in accelerators, fringing fields in real magnetic systems are taken into account with various simulation methods. In [3], the magnet was split into small segments and matrix analysis was applied to each of them. In a more rigorous analysis of fringing fields [4], they were simulated by the Enge function and calculated in terms of differential algebra. Finally, in [5], the study was based on the Lie algebra. Each of the approaches has its own domain of applicability, advantages, and disadvantages. Here, we apply the canonical–variational method of analysis, which involves the direct integration of canonical path equations and the solution of a related set of variational equations.

COMPUTATIONAL METHOD

The components of the magnetic field of a quadrupole that are obtained from the scalar potential (from a solution to the Laplace equation in the natural coordinate system) can be represented in the form

$$B_x(x, y, s) = G(s)y - \frac{1}{12}G''(s)(3x^2y + y^3) + \ldots;$$

$$B_y(x, y, s) = G(s)x - \frac{1}{12}G''(s)(x^3 + 3y^2x) + \ldots; \quad (1)$$

$$B_s(x, y, s) = G'(s)xy - \frac{1}{12}G'''(s)(x^3y + xy^3) + \ldots,$$

where the primes mean differentiation with respect to the longitudinal coordinate *s*.

A magnetic field gradient along the magnet axis is described by the function [6]

$$G(s) = \frac{\mu_0 I}{R^2}\left[\frac{9}{8}f_1(t) - f_3(t) + \frac{3}{8}f_5(t)\right]_{s-L}^{s+L}, \quad (2)$$

where the fitting functions are given by

$$f_k(t) = \left[\frac{t}{\sqrt{R^2 + t^2}}\right]^k. \quad (3)$$

Here, *R* is the radius of a circle inscribed into a quadrupole magnet, *I* is the current in the magnet winding, and *L* is the magnet half-length.

By properly choosing these parameters, one can fairly exactly describe the field gradient of a real magnet with (2).

In view of the Hamiltonian of the problem in the curvilinear coordinate system (*x*, *y*, *s*) related to the ideally flat orbit of an equilibrium particle, the path equation for the particle can be represented in the matrix form

$$Z' = F(Z, s), \quad (4)$$

1063-7842/03/4810-1330$24.00 © 2003 MAIK “Nauka/Interperiodica”

where $Z^t = [x, x', y, y']$, $F^t = [F_1, F_2, F_3, F_4]$, and s is an independent variable.

In the absence of solenoids in the magnetic systems, the generalized velocities x' and y' play the role of canonical momenta. The canonicity of set of Eqs. (4) provides the simplecticity of their solutions. This makes it possible to judge the stability of the particle motion during many turns by integrating over one turn of the particle in the ring. Below, we will follow the method used in [7].

Set (4) is intimately related to a set of 16 linear variational equations

$$B'(s) = A(s)B(s) \tag{5}$$

with the unit boundary conditions $B(0) = I$, where I is the unit matrix.

The matrix of coefficients in (5) is found from (4) as

$$A(s) = \left.\frac{\partial F_i}{\partial Z_j}\right| x = 0; \quad y = 0,$$

$$i, j = 1, \ldots, 4.$$

For one turn, the transition matrix $B_0 = B(s = L_0)$ of set (5) coincides with the Jacobian matrix for set (4). This fact can be used to check the simplicticity of a solution to (4). Solving the problem for the eigenvalues of the matrix B_0, one can calculate phase advances, the fractional part of betatron oscillations, and other dynamic parameters. The accuracy of calculation can be checked by comparing the results obtained in the hard-edge approximation with those of the matrix analysis.

NUMERICAL RESULTS

The magnetic system of the CANDLE (Center for Advancement of Natural Discoveries using Light Emission) storage ring allows for three types of quadrupole magnets. Their basic characteristics are listed in the table.

Figure 1 shows the distributions of the normalized gradients of these magnets where fringing fields are simulated by a characteristic function like (3). Since the quadrupole magnets of the CANDLE storage ring are still under development, below we will use the distribution of the quadrupole lens gradients that is defined by function (3). After magnetic measurements are complete, the calculations will be applied to real magnets.

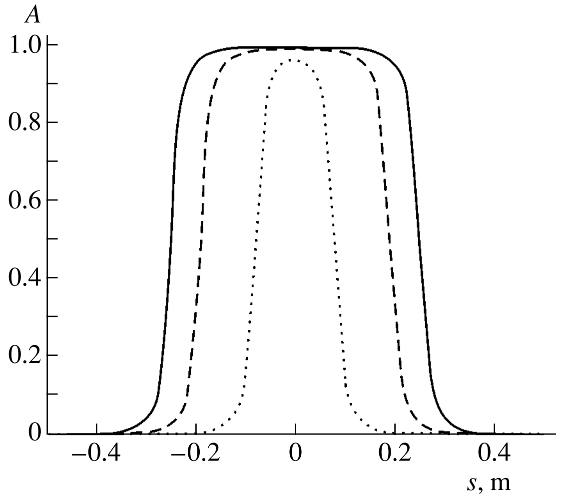


Fig. 1. Longitudinal distributions of the normalized quadrupole gradients: continuous line, quadrupole focusing central; dashed line, quadrupole focusing; and dotted line, quadrupole defocusing. —— *qfc*; – – *qf*; - - - *qd*.

The working point of the betatron oscillation frequency in the CANDLE, which was obtained based on the data listed in the table, corresponds to (13.22, 4.26). With the fringing-field effect taken into consideration by using the canonical–variational method, the working point shifts to (13.08, 4.11). The tracking of particles in the ring showed that the position of the working point is recovered to (13.22, 4.26) if the gradient of the central quadrupole is increased by 0.39%; those of the horizontally focusing outer quadrupoles, by 0.67%; and that of the 0.16-m-long vertically focusing quadrupole, by 1.48%. The relatively large increase in the gradient of the quadrupole that focuses in the vertical direction may be explained by two reasons. First, during the particle tracking, the gradient of the dipole magnet with combined functions remained constant and the vertical correction of the working point was accomplished by the quadrupole magnet. Second, a short length of the magnet (or, in other words, its large aspect ratio) causes the extension of the fringing field region outside the magnet and decreases the peak value of the gradient (Fig. 1).

The horizontal and vertical chromaticities calculated in view of fringing fields were estimated as –18.43 and –14.11, respectively. For the new quadrupole lens gradients that correspond to the recovered working point, the respective chromaticities are –18.87 and –14.72, respectively. These values are slightly lower than –18.91 and –14.86, which were obtained without considering fringing fields.

To estimate the effect of fringing fields on the beam dynamics in the ring, we compared the phase portraits of horizontal oscillations with and without considering the fields prior to increasing the gradients of the quadrupoles (Fig. 2).

Tracking was accomplished for one thousand particles per turn. This is equivalent to one thousand turns of

Basic parameters of quadrupoles NC CANDLE

Type	L, mm	K, m^{-2}	G, T/m	R, mm	$I(A)$
QF	380	1.6497	16.5	35	82
QFC	500	1.703	17.0	35	84
QD	160	1.2896	12.9	35	64

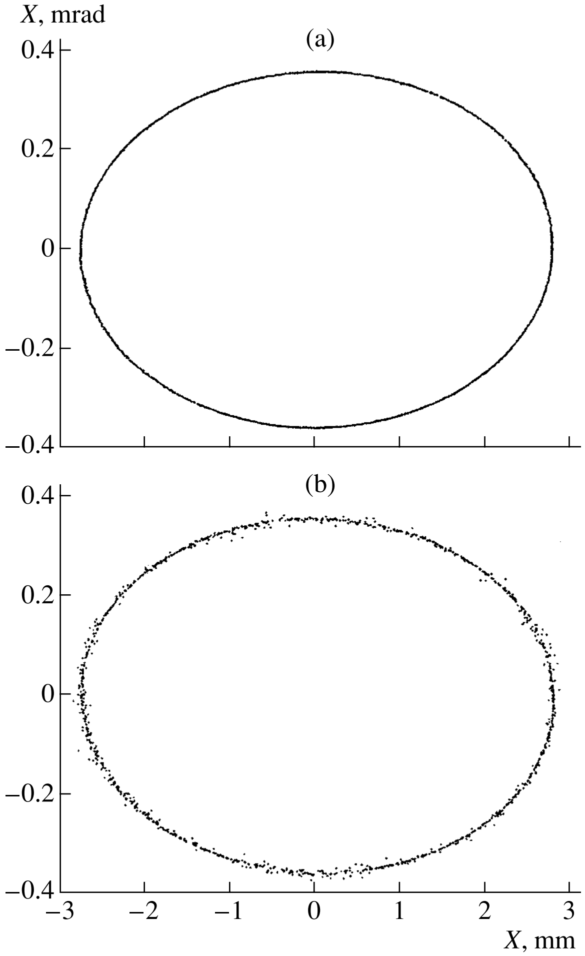


Fig. 2. Phase portraits of horizontal oscillations (a) without and (b) with considering fringing fields at an amplitude of 10σ and zero energy spread.

one particle. Bearing in mind that the damping time of betatron oscillations in a storage ring is several milliseconds, we may assume that one thousand turns would suffice to judge the stability of the motion over a longer time period. The diffused form of the phase portrait when the fringing fields of the quadrupoles are taken into account reflects the fact that the number of horizontal oscillations approaches an integer resonance.

CONCLUSIONS

Our method of evaluating the effect of fringe fields, which is based on the direct integration of canonical path equations combined with the associated variational equations (canonical–variational method), makes it possible to find a shift of the working point, as well as the dynamic aperture, chromaticity, and other characteristic parameters of the beam.

The numerical results obtained for the CANDLE storage ring indicate that an increase in the quadrupole magnet gradients by roughly 1% will recover the working point desired.

ACKNOWLEDGMENTS

The author thanks V. Tsakanov and M. Ivanyan for the valuable discussions.

REFERENCES

1. *CANDLE—3 GeV Synchrotron Light Source: Design Report* (Yerevan, 2002).
2. M. Berz, B. Erdelyi, and K. Makino, Phys. Rev. ST Accel. Beams **3**, 124001 (2000).
3. G. Singh and A. Streun, SLS-TME-TA-2001-0178.
4. K. Makino and M. Berz, Nucl. Instrum. Methods Phys. Res. A **427**, 338 (1999).
5. A. Dragt, in *Proceedings of Particle Acceleration Conference, New York, 1999,* p. 2761.
6. M. Bassetti and C. Biscari, Frascati Phys. Ser. **10**, 247 (1998).
7. A. Dragt, AIP Conf. Proc. **87**, 147 (1982).

Translated by V. Isaakyan

Technical Physics, Vol. 48, No. 10, 2003, pp. 1333–1337. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 116–121.
Original Russian Text Copyright © 2003 by Golbraikh, Rapis, Moiseev.

EXPERIMENTAL INSTRUMENTS AND TECHNIQUES

On the Crack Pattern Formation in a Freely Drying Protein Film

E. Golbraikh*, E. G. Rapis, and S. S. Moiseev*****
** Center for MHD Studies, Ben-Gurion University, Beer-Sheva, 84105 Israel*
e-mail: golbref@bgumail.bgu.ac.il
*** Sour Medical Center, University of Tel-Aviv, Tel-Aviv, 64239 Israel*
**** Space Research Institute, Russian Academy of Sciences, Profsoyuznaya ul. 84/32, Moscow, 117810 Russia*

Received January 28, 2003

Abstract—As shown for the first time by Rapis [7, 8], cracks appearing on the surface of an aqueous protein solution upon its dehydration under normal conditions usually form the same pattern. In this work, a model of processes responsible for the formation of this pattern is proposed. *© 2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

In recent decades, the drying of various liquid substances, including colloidal suspensions, under normal conditions (i.e., at normal temperature and pressure) has been extensively investigated [1–6]. Such studies of drops or films of substances differing in composition, with their dimensions varying from several microns to several centimeters, are necessary for solving a number of technological problems. Drops of biological fluids, however, have been studied inadequately in this respect. The data presented in this paper concern the behavior of aqueous solutions of various proteins, as well as blood, upon natural drying. In other words, we describe the behavior of this open system under natural conditions.

As was indicated in [7], the drying of protein solutions, including the complex of blood proteins, is little understood. Namely, it remains unclear why the crack pattern appearing on the surface of a drying protein film acquires a quasi-cellular structure and why a helical pattern of defects (fractures) is then formed within each cell.

Such a behavior of protein films upon drying under natural conditions has not been explained to date. Here, we propose a model accounting for these observations.

EXPERIMENTS ON DEHYDRATION OF PROTEIN SOLUTIONS

Experiments were performed with a variety of proteins, including albumin, globulin, hemoglobin, cytochrome, lysozyme, crystallin, etc. A protein solution (water–protein system) was placed on solid substrates made of wettable materials (glass, plastic, etc.). A total of 8000 samples of protein films were studied [8]. It is noteworthy that the general parameters of dehydration depended mainly on the protein concentration in the solution, while the qualitative characteristics of cracks were independent of the protein type: even the protein complex of the normal blood serum produced phenomenologically the same crack pattern. However, this pattern changed significantly in the case of disturbances in the protein structure (denaturation) or other forms of pathology.

These experiments with the protein solutions (including the normal blood serum) allowed us to distinguish the following stages of their dehydration.

(1) Dehydration begins at the periphery of a drop (film), and the vitrification front moves toward its center. An example of a protein film at this stage is shown in Fig. 1. The concentric rings observed may be associated with the formation of the stagnation point (see [2]) and its displacement in the course of drying or with the generation of concentration standing autowaves [9].

(2) As the water evaporates, a daisy-like pattern begins to form (Figs. 2–4). The daisy consists of cracks

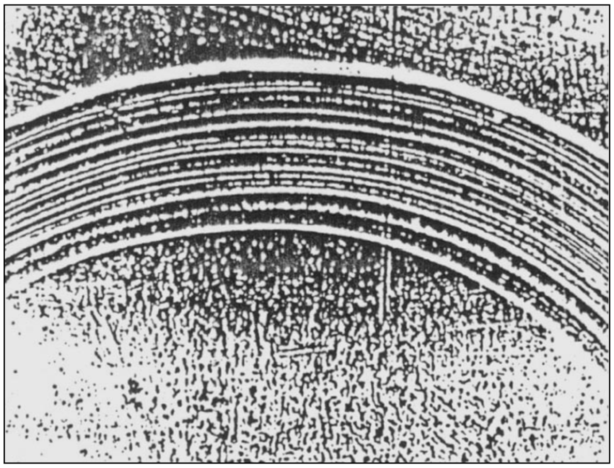


Fig. 1. Initial stage of the natural dehydration of the film formed by an aqueous protein solution.

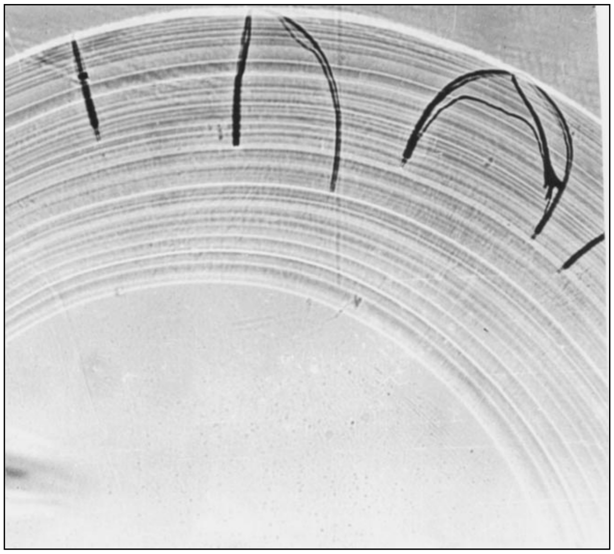


Fig. 2. Initial stage of daisy formation.

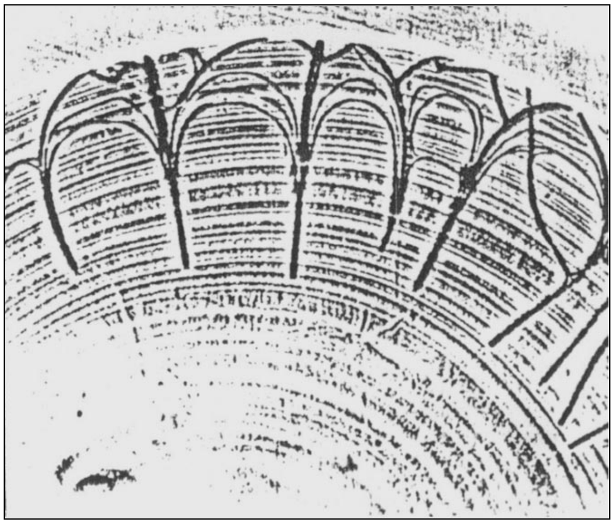


Fig. 3. Subsequent stage of daisy formation.

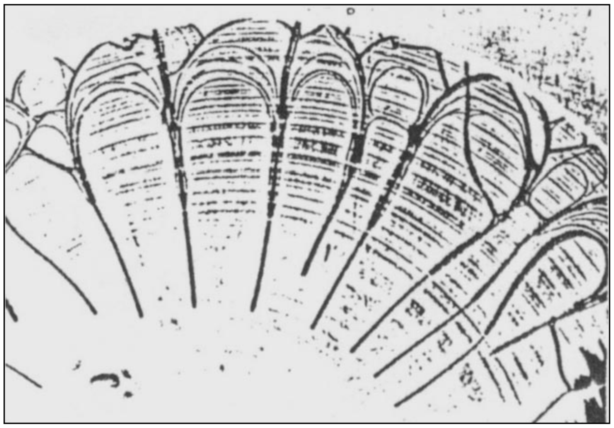


Fig. 4. Daisy pattern (final stage) and the initial stage of cell formation.

that originate at the periphery of the drop and extend toward its center, crossing the annular structures.

(3) At the next stage, the crack pattern acquires a quasi-cellular configuration (Fig. 5) and this process also goes from the periphery to the center.

(4) Next, helical fractures appear in each of the cells. There may be both double (countertwisted, Fig. 6) and single helices (Fig. 7).

Let us consider each of the stages separately.

(1) The initial stage of the dehydration of the aqueous protein solution takes approximately 10–20 min, depending on the concentration, mass, and temperature [7, 8]. In our experiments, the protein concentration N ranged from 0.001 to 20%. The formation of the concentric circles at this stage is virtually independent of N, and the sol behavior can be considered in terms of the general approach to dehydration of such drops. As was shown in [2], the field of hydrodynamic flow forms in the drop, which is generally a shear flow with the stagnation point (see [2], Fig. 12). Once the stagnation point has formed, stagnation zones appear in its vicinity, where sol particles are accumulated. As the drop dries up, the stagnation point moves toward its center, producing concentric circular areas with the increasing concentrations of these particles.

On the other hand, such a distribution of the sol particle density may be due to the generation of standing autowaves [9] as a result of interaction between the direct flow (from the center to the periphery) and the flow reflected from the drop edge or the surface flow (from the stagnation point). Apparently, both processes could be responsible for the formation of the concentric pattern shown in Fig. 1.

Asymmetric particles, such as protein molecules or protein clusters, present in sol solutions are eventually aligned with the shear flow [10]. Therefore, the gel resulting from the dehydration of a protein solution is bound to possess the properties of a liquid crystal because of a director appearing in it. Since protein molecules and clusters lack the center of symmetry, this liquid crystal is a cholesteric or a C-type smectic [11].

(2) As the protein film is dehydrated, it cracks and the cracks eventually form the daisy pattern shown in Figs. 2–4. Such behavior of colloidal suspensions (which our protein solutions are) upon dehydration is described elsewhere [12]. The authors of [12] observed daisy-like cracking at an ionic strength I below the critical value $I = 0.18$ mol/l at which the characteristic drying time is much shorter than the drop gelation time. Since the conditions of protein film dehydration in our work are similar to those in [12], we may assume that the daisy in both cases forms in the same way: sol particles concentrate near the three-phase line, where the highest stresses are generated. These stresses increase with time (i.e., proportionally to the degree of dehydration), eventually leading to the formation of the characteristic crack pattern. Note that, in protein solutions,

this process develops much more slowly than in the sols studied in [12].

So far, we have considered the initial stages of protein film dehydration and compared our data with the results of other experiments where the properties of suspension drops at these stages have already been observed and analyzed. However, the behavior of protein films at subsequent stages of dehydration is described only in our publications [7, 8].

(3) **Quasi-cellular structure.** Experimental data show that, after the daisy has formed (this process takes approximately 25 min), secondary fracturing begins in its petals and the pattern of cracks eventually acquires the quasi-cellular structure (Fig. 5). Like at the previous stages of dehydration, this process goes from the periphery to the center.

The formation of the quasi-cellular structure can be explained as follows (note that, according to the experimental data, the daisy remains unchanged, so that each its petal may be studied separately). As was noted in [12], the highest stresses in an evaporating suspension drop arise at the three-phase line (point). During the evaporation, the lower part of the drop begins to shrink and simultaneously the film strongly adheres to the substrate [8]. This causes stresses to arise at the petal boundaries, leading to the formation of the quasi-cellular cracking.

(4) Helical fracturing within quasi-cells appeared in all the experiments on drying films from the protein solutions (with an initial protein concentration exceeding 5%). The only difference is in the topological density of these structures: depending on the protein concentration, the number of helical “nuclei” per cell may vary considerably (Figs. 6, 7).

The formation of helical structures in biological media is an important problem. It should be noted, however, that the scope of this problem is not limited to biological media (see, e.g., [13]). Pseudoscalars that emerge when describing systems with violated mirror symmetry play a certain role in the generation of helical macrostructures, and this role is now becoming clearer. As follows from numerous sources (see [13] and Refs. therein), media characterized by helicity or chirality (one of the key features of nonlinear dynamics) are unstable: when subjected to weak external disturbances, they pass to a new state with the formation of large-scale structures. It may well be that the dehydration of protein solutions is also governed by the helicity parameter, which is responsible for the helical cracking. As was noted above, protein clusters in the presence of a fluid flow from the center to the periphery become aligned with this flow. The medium acquires liquid crystal properties and apparently retains them as the solvent (water) evaporates, although it is known that large-scale helical structures may be initiated by small helical or chiral nuclei (see [14] and Refs. therein).

It is also very important that, in the case described, neither protein molecules nor their conglomerates possess mirror symmetry. All protein molecules are of either right-handed or left-handed helices in this case; however, their conglomerates may be of either type (both right- and left-handed but, naturally, with violated symmetry). Thus, we distinguish patterns formed on the microscale, where the symmetry of protein molecules is of crucial importance, and on the macroscale, where the symmetry of protein conglomerates becomes of significance. Therefore, the type of protein used in the experiments is of minor concern.

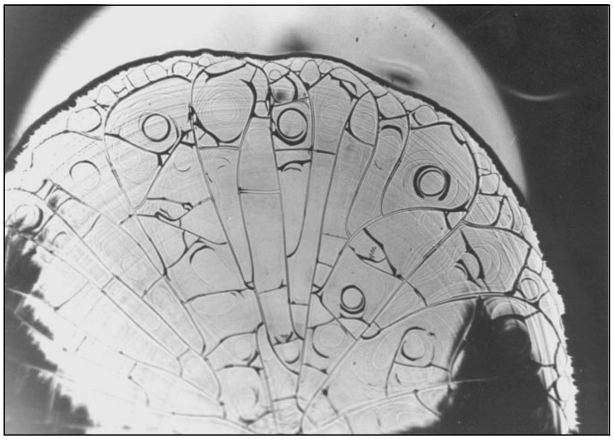


Fig. 5. Quasi-cellular pattern with incipient helical nuclei.

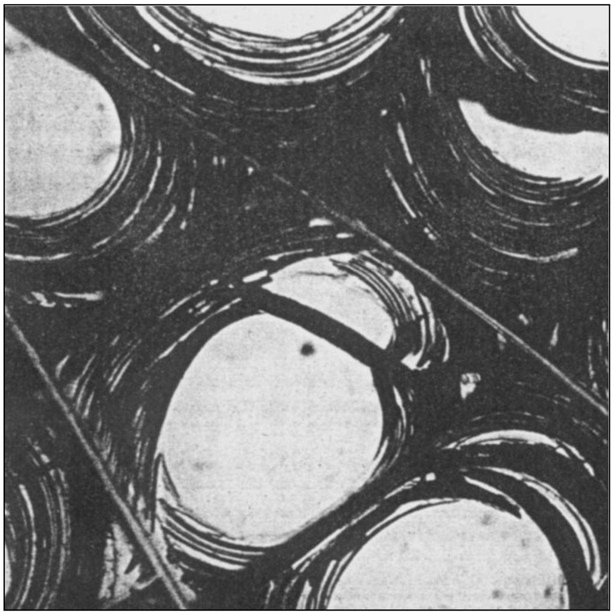


Fig. 6. Typical pattern of helical nucleus formation.

MODEL OF SELF-ORGANIZATION

Let us consider a model of self-organization in protein solutions. Since protein conglomerates have no mirror symmetry, small disturbances, which always

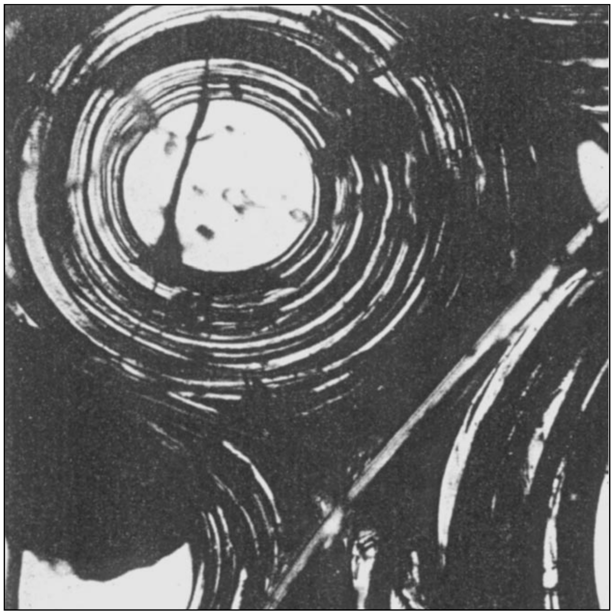


Fig. 7. Another type of helical nuclei.

occur in the system, make them unstable and this instability leads to the formation of helical structures in the areas with the quasi-liquid-crystal properties. The free energy of deformation in the case of a cholesteric liquid crystal can be described as

$$\Phi = 0.5\chi_1(\mathrm{div}\mathbf{n})^2 + 0.5\chi_2(\mathbf{n}\cdot\mathrm{curl}\mathbf{n} + q)^2 + 0.5\chi_3[\mathbf{n}\times\mathrm{curl}\mathbf{n}]^2, \tag{1}$$

where χ_i are Franck's moduli, **n** is a director (a unit vector with a certain direction at a given point), and q is the helicity parameter (in units of inverse length). In equilibrium ($\Phi = 0$), the director components have the form [15]

$$n_x = \cos(qz), \quad n_y = \sin(qz), \quad n_z = 0, \tag{2}$$

i.e., they form a helical structure with the axis aligned with z.

However, this is an equilibrium structure. Below, we consider its transition to the helical state by analyzing a nematic that lacks the symmetry plane for some reasons. The equation of motion of the director has the form [15]

$$dn_i/dt = N_i. \tag{3}$$

Let the nematic be at rest; that is, its velocity V equals zero or affects the motion of the director only slightly. The quantity $\mathbf{N} = \mathbf{h}/\gamma$ describes the director relaxation in the molecular field, and γ is a coefficient that has the dimension of viscosity. The molecular field **h** has the form

$$\mathbf{h} = \mathbf{H} - \mathbf{n}(\mathbf{H}\cdot\mathbf{n}), \tag{4}$$

where

$$H_i = \partial[\partial\Phi/\partial(\partial n_i/\partial x_k)]/\partial x_k - \partial\Phi/\partial n_i. \tag{5}$$

In equilibrium, $h = 0$. For a cholesteric, H is determined as follows:

$$\begin{aligned}\mathbf{H} &= \nabla(\chi_1\mathrm{div}\mathbf{n})\\ &- \{\chi_2(\mathbf{n}\cdot\mathrm{curl}\mathbf{n} + q)\mathrm{curl}\mathbf{n}\\ &+ \mathrm{curl}[\chi_2(\mathbf{n}\cdot\mathrm{curl}\mathbf{n} + q)\mathbf{n}]\}\\ &- \{\chi_3(\mathbf{n}\times\mathrm{curl}\mathbf{n})\times\mathrm{curl}\mathbf{n} + \mathrm{curl}[\chi_3\mathbf{n}\times(\mathbf{n}\times\mathrm{curl}\mathbf{n})]\}.\end{aligned} \tag{6}$$

Consider small fluctuations of the director: $\mathbf{n} = \mathbf{n}_0 + \mathbf{n}'$, where $|\mathbf{n}'| \ll 1$ (consequently, $\mathbf{n}_0\cdot\mathbf{n}' = 0$) and $\mathbf{n}_0$ is the undisturbed director, which remains constant throughout the medium. In the linear approximation, $\mathbf{h}' = \mathbf{H}' - \mathbf{n}_0(\mathbf{H}'\cdot\mathbf{n}_0)$, where, as follows from Eq. (6),

$$\begin{aligned}\mathbf{H}' &= \nabla(\chi_1\mathrm{div}\mathbf{n}')\\ &- \{\chi_2 q\mathrm{curl}\mathbf{n}' + \mathrm{curl}[\chi_2(\mathbf{n}_0\cdot\mathrm{curl}\mathbf{n}')\mathbf{n}_0 + q\mathbf{n}']\}\\ &+ \mathrm{curl}\{\chi_3\mathbf{n}_0\times(\mathbf{n}_0\times\mathrm{curl}\mathbf{n}')\}.\end{aligned} \tag{7}$$

For the fluctuation part of Eq. (3),

$$\partial n_i'/\partial t = h_i'/\gamma. \tag{8}$$

Assume that $\chi_i = \mathrm{const}$. Then, if $\mathbf{n}'$ corresponds to a plane wave $\exp(i\omega t - ikr)$, the formula for $\mathbf{h}'$ takes the form

$$\begin{aligned}\mathbf{h}' &= -\chi_1\{\mathbf{k} - \mathbf{n}_0(\mathbf{n}_0\cdot\mathbf{k})\}(\mathbf{k}\cdot\mathbf{n}')\\ &- \chi_2\{\beta(\beta\mathbf{n}') + iq(\mathbf{k}\times\mathbf{n}')\} - \chi_3(\mathbf{n}_0\cdot\mathbf{k})^2\mathbf{n}',\end{aligned} \tag{9}$$

where $\beta = (\mathbf{n}_0\cdot\mathbf{k})$ (cf. [15]; problem 3 in Sect. 42).

Scalarly multiplying Eq. (8) by **k** and taking into account Eq. (9), we obtain the dispersion relation

$$\begin{aligned}&\{i\omega + (\chi_1/\gamma)[k^2 - (\mathbf{n}_0\cdot\mathbf{k})^2]\\ &+ (\chi_3/\gamma)(\mathbf{n}_0\cdot\mathbf{k})^2\}(\mathbf{k}\cdot\mathbf{n}') = 0.\end{aligned} \tag{10}$$

Here, two variants are possible.

(1) $(\mathbf{k}\cdot\mathbf{n}') = 0$; hence, $\mathbf{k}\perp\mathbf{n}'$. Then, we find from Eqs. (8) and (9) that

$$\begin{aligned}i\omega\mathbf{n}' &= -(\chi_2/\gamma)\{\beta(\beta\mathbf{n}') + iq(\mathbf{k}\times\mathbf{n}')\}\\ &- (\chi_3/\gamma)(\mathbf{n}_0\cdot\mathbf{k})^2\mathbf{n}'.\end{aligned} \tag{11}$$

Note that β is parallel or antiparallel to $\mathbf{n}'$. Therefore, if $\mathbf{n}' = n'\mathbf{e}$ (where **e** is the unit vector),

$$\begin{aligned}&\{i\omega + (k^2/\gamma)(\chi_2\sin^2\theta + \chi_3\cos^2\theta)\}\mathbf{e}\\ &= -(i\chi_3/\gamma)q(\mathbf{k}\times\mathbf{e}),\end{aligned} \tag{12}$$

where θ is the angle between **k** and $\mathbf{n}_0$.

Let us choose the coordinate system where the wavevector **k** is aligned with $\mathbf{e}_3$; in this case, the vector **e**

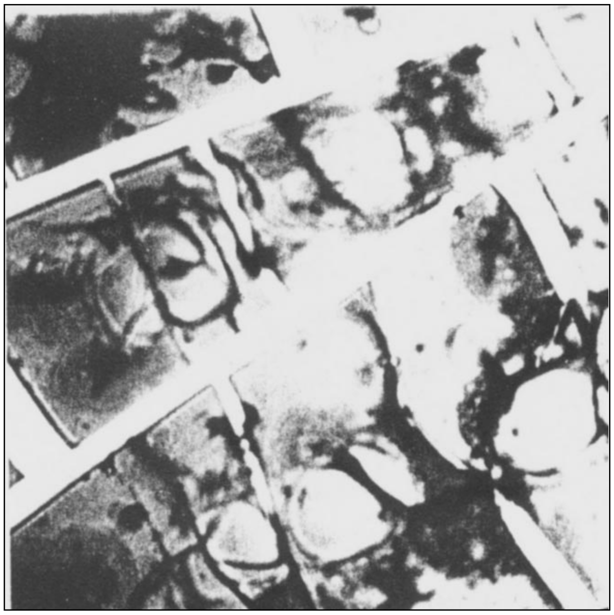


Fig. 8. Film formed by the drying blood serum of a patient with cancer.

has two components, which meet the relationships

$$\{i\omega + (k^2/\gamma)(\chi_2 \sin^2\theta + \chi_3 \cos^2\theta)\}e_1 = (i\chi_2/\gamma)qke_2, \tag{13}$$
$$\{i\omega + (k^2/\gamma)(\chi_2 \sin^2\theta + \chi_3 \cos^2\theta)\}e_2 = -(i\chi_2/\gamma)qke_1.$$

It follows from Eqs. (13) that

$$i\omega = \pm(q\chi_2 k/\gamma) - k^2(\chi_2 \sin^2\theta + \chi_3 \cos^2\theta)/\gamma, \tag{14}$$

hence, $i\omega$ is an increment only if

$$k < (q\chi_2)/(\chi_2 \sin^2\theta + \chi_3 \cos^2\theta). \tag{15}$$

It is apparent that, depending on the orientation of **k** and $\mathbf{n}_0$, the boundary of the instability zone will vary from $k < q\chi_2/\chi_3$ to $k < q$ at $\theta = 0$ and $\pi/2$.

(2) If $(\mathbf{k} \cdot \mathbf{n}') \neq 0$, we have

$$i\omega + (\chi_1/\gamma)[k^2 - (\mathbf{k} \cdot \mathbf{n}_0)^2] + (\chi_3/\gamma)(\mathbf{k} \cdot \mathbf{n}_0)^2 = 0. \tag{16}$$

As the second and third terms of this equation are positive, these modes decay. Thus, the appearance of helicity makes the system unstable against the generation of large-scale structures in the director field. Note that the extremum of the instability increment corresponds to the modes with $\theta = 0$ and $\pi/2$. With these values, the dependence of $i\omega$ on the angle θ has a maximum at $\chi_3 < \chi_2$ and a minimum at $\chi_3 > \chi_2$. This fact may be important when films of blood serum from sick persons are analyzed. Helical structures in such films (Fig. 8) are formed very rarely, if at all. This can be explained by different ratios of Franck's coefficients χ_2 and χ_3, which may vary considerably in such films.

CONCLUSIONS

The model proposed in this paper describes the natural dehydration of films of aqueous protein solutions. To date, this process has been poorly studied. While its initial stages resemble the drying of sols and gels, the formation of the quasi-cellular structure with helical nuclei at its final stage has remained unclear. Our simple model separates out processes responsible for the cracking pattern in a drying protein film. The point in the model is the development of instability in the liquid-crystal phase due to the helicity factor. The pattern of cracks observed in the experiments depends apparently on the ratio between the Franck's parameters.

REFERENCES

1. D. Avnir and M. L. Kagan, Chaos **5**, 589 (1995).
2. R. D. Deegan, O. Bakajin, T. F. Dupont, *et al.*, Phys. Rev. E **62**, 756 (2000).
3. F. Parisse and C. Allain, J. Phys. II **6**, 1111 (1996).
4. Y. Sezen, Int. Commun. Heat Mass Transfer **27**, 559 (2000).
5. J. J. Hegseth, N. Rashidnia, and A. Chai, Phys. Rev. E **54**, 1640 (1996).
6. R. D. Deegan, Phys. Rev. E **61**, 475 (2000).
7. E. G. Rapis and G. Yu. Gasanova, Zh. Tekh. Fiz. **61**, 406 (1991) [Sov. Phys. Tech. Phys. **36**, 406 (1991)].
8. E. G. Rapis, Zh. Tekh. Fiz. **70** (1), 122 (2000) [Tech. Phys. **45**, 121 (2000)].
9. V. A. Vasil'ev, Yu. M. Romanovskiĭ, and V. G. Yakhno, Usp. Fiz. Nauk **128**, 625 (1979) [Sov. Phys. Usp. **22**, 615 (1979)].
10. C. G. Caro, T. J. Pedley, R. C. Schroter, and W. A. Seed, *The Mechanics of the Circulation* (Oxford Univ. Press, Oxford, 1978; Mir, Moscow, 1991).
11. S. Chandrasekhar, *Liquid Crystals* (Cambridge Univ. Press, Cambridge, 1992, 2nd ed.; Mir, Moscow, 1980).
12. L. Pauchard, F. Parisse, and C. Allain, Phys. Rev. E **59**, 3737 (1999).
13. H. Branover, A. Eidelman, E. Golbraikh, and S. S. Moiseev, *Turbulence and Structures* (Academic, New York, 1999).
14. V. A. Kizel', *Physical Causes for the Asymmetry of Gel-Like Systems* (Nauka, Moscow, 1985).
15. L. D. Landau and E. M. Lifshitz, *Course of Theoretical Physics*, Vol. 7: *Theory of Elasticity* (Nauka, Moscow, 1987; Pergamon, New York, 1986).

Translated by N. Gorgolyuk

Technical Physics, Vol. 48, No. 10, 2003, pp. 1338–1345. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 122–129.
Original Russian Text Copyright © 2003 by L. Veresov, O. Veresov.

EXPERIMENTAL INSTRUMENTS AND TECHNIQUES

Ion Source with a Cold Magnetron Cathode and Magnetic Plasma Compression

L. P. Veresov and O. L. Veresov
Sukhumi Physicotechnical Institute, Sukhumi, 384914 Abkhaziya, Georgia
e-mail: ol_veres@mail.ru

Received February 10, 2003

Abstract—The operating principles, design, and characteristics of an ion source with a cold magnetron cathode and magnetic plasma compression are described. The source is intended for the injector of a linear proton accelerator. *© 2003 MAIK "Nauka/Interperiodica".*

INTRODUCTION

The ion source known as a duoplasmatron was designed by Ardenne in 1949 [1]. Up to the present, the duoplasmatron is one of the most widely used types of ion sources. The reason for this is the following two factors: the outstanding characteristics of this source (the unsurpassed plasma density, low working-gas consumption, and high efficiency) and the possibility of further modifying it. The latter factor allowed one to substantially improve the parameters of the source over the past decades. A specific feature of the duoplasmatron is a two-stage plasma contraction: the geometric contraction by means of narrowing the intermediate electrode and the magnetic contraction by applying a nonuniform magnetic field near the anode. All this results in a substantial contraction of the discharge and the production of a dense plasma. This feature provides the higher ion emissivity and the lower working-gas consumption of the duoplasmatron in comparison with other ion sources.

A substantial modification in the duoplasmatron was the employment of an expander [2] to extract an ion beam from a large-area plasma surface, which allowed one to substantially increase the beam current. The next step in increasing the beam current was the introduction of electron oscillations in the duoplasmatron in order to achieve an even denser plasma and a higher gas efficiency, which made it possible to enlarge the emission aperture [3]. The use of a three-electrode ion-optical acceleration–deceleration system in such a duoplasmatron allowed one to form hydrogen ion beams with currents about one ampere [4]. The applying of permanent magnets (instead of electromagnets) made it possible to reduce the energy consumption and mass of the duoplasmatron [5].

Further, the duoplasmatron was modified to prolong the time of its continuous operation, which is primarily determined by the service life of the heater cathode. To do this, attempts were made to replace the hot cathode with a cold one. The use of different types of hollow cathodes made it possible to substantially prolong the service time of the device [6, 7]. On the other hand, those studies revealed serious disadvantages of a hollow-cathode duaplasmotron, such as the low gas efficiency and the high breakdown potential. The use of a multipin cathode [8] did not solve the problem of the high breakdown potential, but raised the new problem of the high impurity content in the generated ion beam, because the ionization of the cathode-erosion products attained 90%. Many of these problems were overcome in a duoplasmatron with a cold magnetron cathode [9]. The studies on such a duoplasmatron have led the authors to a new version of an ion source with a cold magnetron cathode and magnetic plasma compression. This type of ion source is intended for the injector of a linear proton accelerator [10].

PHYSICAL PROCESSES IN THE SOURCE

The ion source with a cold magnetron cathode and magnetic plasma compression consists of two main units: a plasma generator (Fig. 1) and a system for extracting and forming an ion beam. The plasma generator can be conventionally divided into three regions: the magnetron (cold-cathode) region, the main-discharge (magnetic-compression) region, and the plasma-expansion region (the expander).

Glow discharge in secondary-emission cold cathodes is characterized by gamma-processes occurring on the emitting metal surface when it is bombarded (predominantly) by positive ions. This type of discharge is self-controlled and self-sustained [11]. At a fixed gas flow rate and fixed input power, the discharge current and voltage are self-adjusted due to variations in the emissivity of the plasma surface: when there is a deficit of electrons maintaining the discharge, the discharge current decreases and the discharge voltage increases, and vice versa when there is an excess of electrons maintaining the discharge. This effect allows one to produce both high-voltage (a few kilovolts) and low-voltage (a few tens of volts) glow discharges. To create an ion source with a low energy consumption, it is necessary to minimize the discharge voltage. This can be achieved in two ways: by increasing the number of particles bombarding the emitting surface or by enlarging the emitting surface itself. Both these methods were realized in a cylindrical reversed magnetron cathode [12]. By introducing diaphragms into the cavity of the reversed magnetron cathode and, thus, dividing it into several cells, it became possible to enlarge the emitting surface, increase the gas efficiency, and achieve a more stable cathode operation [9]. The tapered end of the magnetron cathode in the form of a truncated cone transforms the annular beam into a cylindrical one. This enables one to solve two problems: to further increase the gas efficiency and to equalize and increase the density of the plasma flow entering the main-discharge region from the cold-cathode region.

In the main-discharge region, the plasma is compressed into a thin column by a strong magnetic field, so that it can penetrate through a millimeter aperture into the expander. It should be noted that, near the intermediate contracting electrode of the duoplasmatron, a bubble-shaped electric double layer arises, which is a source of fast electrons. Being accelerated in the field of the double layer, the flow of these electrons substantially increases the plasma density [13]. In sources with a cold magnetron cathode, such a bubble does not form. There, the double layer arises near the negative electrode of the magnetron and accelerates the electrons produced by secondary electron emission from the cathode. For this reason, in this type of ion source, the intermediate contracting electrode was excluded as useless. It is known [14] that, in plasma ion sources employing low-pressure discharges, charged particles recombine primarily at the walls of the gas-discharge chamber. The reduction in the area of the chamber walls due to the exclusion of the intermediate electrode led to an increase in the plasma density in the source.

After passing through the magnetic-compression region, the plasma enters the expander, where it expands. The plasma expander [2] is applied because an intense high-quality ion beam cannot be extracted from a high-density plasma. As the plasma expands, its density decreases. In the course of expansion, a meniscus plasma boundary is produced, which forms the extracted beam. The expander is a part of the plasma generator and, at the same time, a part of the ion-optical system (IOS) for extracting and forming the beam. In the expander, the particle flow leaving the plasma generator is matched to the ion flow forming the beam. The ion beams are best formed with a concave plasma surface, and they are worst formed with a convex plasma surface. There is a variety of expanders with different sizes and designs. Among those are expanders with conical or cylindrical geometry, with grids stabilizing the plasma boundary, with flaps cutting the plasma coat, etc.

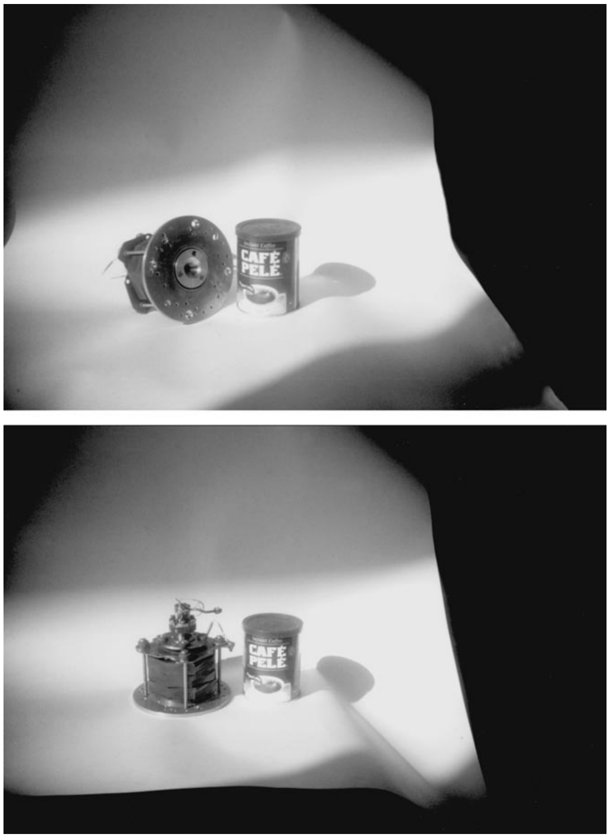


Fig. 1. External view of the plasma generator.

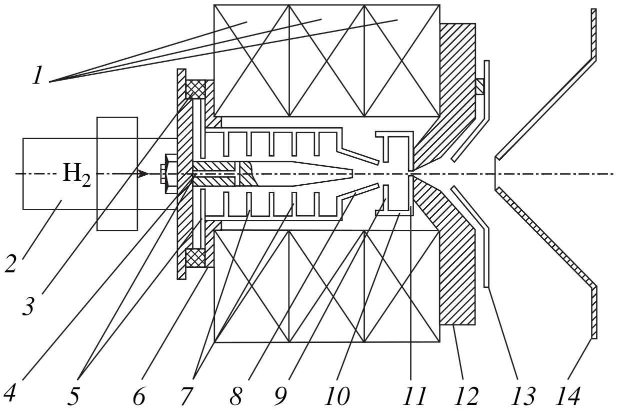


Fig. 2. Design of the ion source: (*1*) annular ferrite–barium magnets, (*2*) electromagnetic valve for gas puffing, (*3*) annular ceramic insulator, (*4*) magnetron anode, (*5*) casing of the magnetron cathode, (*6*) annular centering support, (*7*) cathode diaphragms, (*8*) conical insert of the magnetron cathode, (*9*) anode diaphragm, (*10*) protecting shield, (*11*) collimating diaphragm, (*12*) emission electrode, (*13*) conical insert of the expander, and (*14*) extraction electrode.

In the ion source described in this paper, we used a conical expander with a smoothly varied angle. A conical insert (Fig. 2) introduced in the expander served the following purposes: (i) to form and stabilize the accelerating electric field and the concave plasma boundary; (ii) to cut the plasma coat; and (iii) to reduce the area of the plasma ion emitter, thus improving the beam emittance. From the peripheral region of the conical insert, the excess gas was removed in the radial direction, which resulted in a substantial reduction in breakdown events in the high-voltage gap. A magnetic field of 150 G in the expander hampered electron recombination at its walls; as a result, a negative space charge neutralizing the positive space charge of the beam was produced in the high-voltage gap. Such an expander design allowed us to form and extract 300-mA neutralized proton beams with the help of a two-electrode IOS. The same IOS allowed us to extract partially neutralized proton beams with a current above 1 A.

In the immediate future, a three-electrode acceleration–deceleration IOS and a double-focusing electrostatic system will be applied in this ion source. The three-electrode IOS (expander–extracting electrode–decelerating electrode) creates a potential barrier for the beam-plasma electrons drifting toward the source. The barrier produces a virtual cathode for electrons neutralizing the positive space charge of the ion beam. This makes it possible to increase the intensity of a neutralized ion beam. We note that, with a two-electrode IOS, the return flow of the beam-plasma electrons was accelerated in the high-voltage gap and melted down the collimating copper diaphragm of the expander.

The double-focusing lens consists of two matched lenses with different focal lengths. An advantage of this lens is that it operates more efficiently that a single lens. At the same time, this lens can also be supplied from one power source.

The concept discussed above was implemented in an ion source with a cold magnetron cathode and magnetic plasma compression, whose design will be described below.

DESIGN OF THE SOURCE

Exclusion of the intermediate electrode from the duoplasmatron allowed us to significantly simplify the design of the new ion source (see Fig. 2). The magnetic field is produced by three 16-mm-thick annular ferrite–barium magnets. Two of these magnets have the same polarity and produce a longitudinal field in the magnetron-cathode region. The third annular magnet has the opposite polarity and produces a strong nonuniform magnetic field, which causes the discharge to contract near the emission aperture. As a cold cathode, we use a reversed multicell cylindrical magnetron [9]. Its design is the same as in the last version of the magnetron cathode of the duoplasmatron. The cathode unit is a stainless-steel casing, into which the cathode components (nonmagnetic washers and tubes) are inserted. The washers act as diaphragms, and tubes serve as crossbars. They divide the casing cavity into six cells. The casing is ended with an insert, which is a truncated cone with an aperture 4 mm in diameter in its smaller base. The casing is inserted into a centering annular support made of a nonmagnetic material. The magnetron anode is attached to the annular support through an annular ceramic insulator. The anode is a cylindrical metal rode tapered at the end. The working gas is supplied through it by using an electromagnetic valve [15]. The cathode unit is in the longitudinal magnetic field. The third annular magnet, whose polarity is opposite to that of the two first annular magnets, and the iron emission cathode form the magnetic compression region. Such a design allows us to produce a strong nonuniform magnetic field in a narrow contraction gap. The destination of this field is the same as in a duoplasmatron. Figure 3 shows the longitudinal profile of the magnetic field produced with the help of the annular magnets described above and the iron emission cathode. The emission electrode serves the following three functions: (i) it forms the magnetic field of a required configuration, (ii) it acts as a main anode of the source, and (iii) it produces the conical cavity of the plasma expander. The emission electrode is an iron disk of complicated shape; it is attached to the third magnet. Its central conical part is buried in the cavity of this magnet by 6 mm. The central region of the opposite side of the disk forms the conical cavity of the expander. The diameter of an aperture in the cone vertex of the emission electrode is 2 mm. The cone vertex is adjacent to a 1-mm-thick circular copper insert (a collimating diaphragm) with a 1-mm-diameter collimating aperture. Another function of the collimating diaphragm is to remove heat toward the emission electrode. This diaphragm is joint to a metal shield protecting the magnets from the plasma particles and discharge radiation. In the conical cavity of the expander with a smoothly broadening angle, a conical insert is placed. Between the cavity and the insert, there is a clearance to remove a portion of the gas from the expander. The conical insert is attached via three supports to the emission electrode and is at the electrode potential. The insert follows the shape of the expander cavity and, then, smoothly transforms into a plane ring. The diameter of the insert aperture is 0.8 mm, and the length of the insert (from its end to the collimating hole in the copper insert) is 9 mm. The extracting electrode is attached through three ceramic insulators to the emission electrode of the plasma generator. The diameter of the aperture in the extraction electrode is 9 mm. The length of the accelerating gap was varied from 8 to 11 mm.

DIAGNOSTIC FACILITY

The parameters of the ion beam generated by the source were measured using a specific diagnostic facility.

To estimate the beam emittance, we used an angular gauge of the following design. A 15-mm-diameter metal tube with a base length of 350 mm was inserted in a sectioned Faraday cup (FC) [16] with the entrance

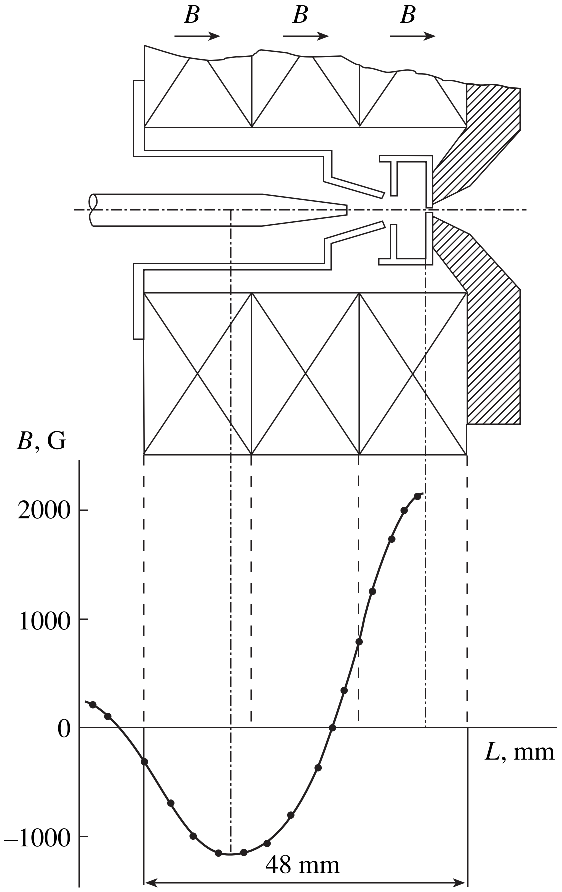


Fig. 3. Longitudinal profile of the magnetic field in the source.

aperture 60 mm in diameter. On the source side, the tube was tapered and had a 0.1-mm slit, which cut out a narrow jet from the beam. On the opposite end of the tube, we placed a grid made of 0.2-mm wires. The interwire distance was 0.2 mm, so that the grid step was 0.4 mm. The grid was specially shaped to facilitate visual counting (Fig. 4). The grid was tight against a quartz glass doped with some lanthanide [17]. The glass served as an optical detector of the accelerated ions, whose action manifested itself by bright strips on the glass. From the number of these strips and the known base length, we determined the angular divergence of the ion jet. Behind the doped quartz glass, an optical lens was placed for convenience of visual observation and counting of the bright strips.

The beam current was measured with the same sectioned FC.

The current density profile was measured by a small FC with the entrance aperture 1.1 mm in diameter.

The beam portrait was observed and measured with the help of a glass doped with some lanthanide and covered with a grid for removing the accumulated electric charge.

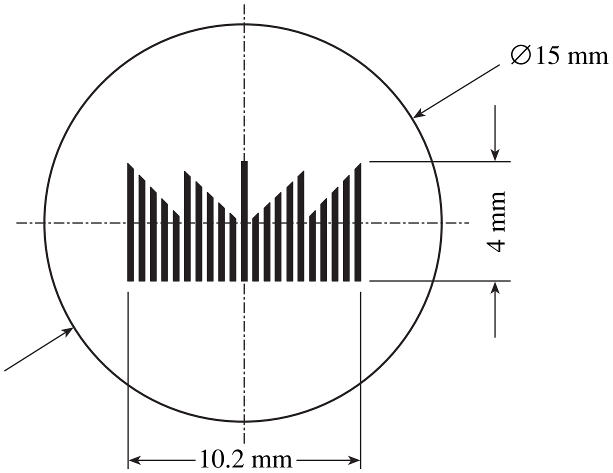


Fig. 4. Measuring rule at the end of the angular gauge: the number of slits is 21, the slit width is 0.2 mm, the grid-wire thickness is 0.2 mm, and the meander step is 0.4 mm.

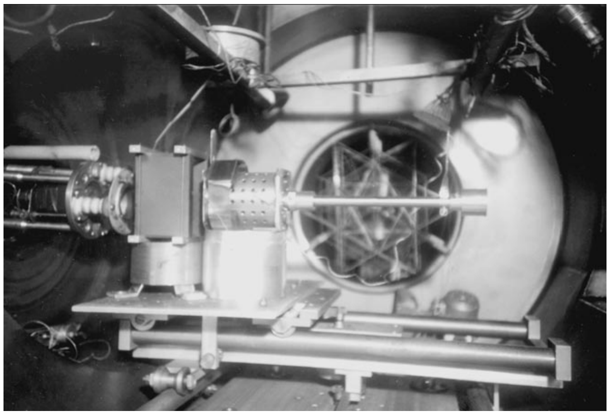


Fig. 5. Arrangement of the source and the angular gauge inside the vacuum chamber.

The mass composition of the beam was measured by a Dempster mass spectrometer [18]. A magnetic field of up to 6 kG was produced by an electromagnet.

All diagnostic devices were mounted on a mobile platform in a 1-m^3 vacuum chamber. The platform could be moved by electric motors in two orthogonal directions (Fig. 5).

EXPERIMENTAL RESULTS

We investigated the performance of the ion source with a cold magnetron cathode and magnetic plasma compression in the pulsed operation mode. The pulse duration varied from 1 to 100 μs, and the repetition rate varied from 1 to 10 Hz.

As can be seen from the source design described above, the magnetron cathode of this source differs from the magnetron cathode described in [9]. The cathode length was doubled, which allowed us to increase

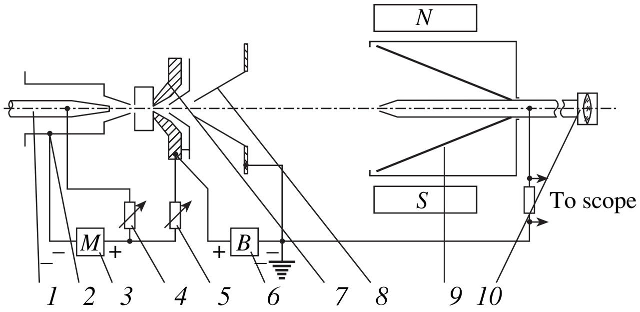


Fig. 6. Power supply diagram of the source and the measuring circuit: (*1*) magnetron anode, (*2*) magnetron cathode, (*3*) modulator of the discharge power supply, (*4*) ballast resistor of the auxiliary discharge, (*5*) ballast resistor of the main discharge, (*6*) high-voltage rectifier, (*7*) source anode (emission electrode), (*8*) extraction electrode, (*9*) sectioned cone of the FC with a shield, and (*10*) angular gauge with an optical lens.

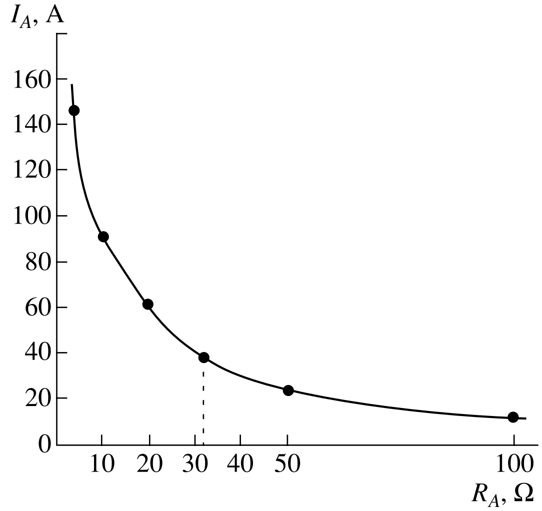


Fig. 7. Main-discharge current vs. ballast resistance in the main-discharge circuit at a fixed ballast resistance in the auxiliary-discharge circuit.

the number of cells from three to six. The longitudinal magnetic field was produced by two magnets instead of one. The cathode diameter was increased from 19 to 22 mm. All this made it possible to considerably enlarge the emitting area of the cathode and strengthen the longitudinal magnetic field. As a result, the discharge voltage was lowered from 400–500 to 250–300 V and the working-gas flow rate was reduced from 10–12 to 7–9 torr cm^3/pulse. The conical extension at the end of the cathode unit, which serves as a diaphragm between the magnetron region and the main-discharge region, allowed us to further decrease the gas flow rate (from 7–9 to 5–7 torr cm^3/pulse). Gas measurements were carried out using the technique described in [9].

Figure 6 shows the electric circuit of the ion source. The negative terminal of the discharge modulator was connected to the magnetron cathode, and its positive terminal was connected through two parallel-connected variable ballast resistors to the magnetron anode and the emission electrode. The discharge excited by the modulator operated in different regimes in different regions of the plasma generator. The ion-source emissivity depended exclusively on the parameters of the contracted discharge, whose characteristics, in turn, depended on the parameters of the discharge in the magnetron cathode. For this reason, the contracted discharge was called the main discharge and the magnetron discharge was called the auxiliary discharge. Variations in the ratio between the ballast resistances resulted in variations in the ratio between the main and auxiliary discharge currents. As was mentioned above, both the amplitude and duration of the ion-beam current extracted from the source were governed by the main discharge. Varying the currents in the circuits of the auxiliary and main discharges with the help of the resistors, we could affect the source operation and the ion-beam parameters. Thus, we revealed the following feature: when the current in the magnetron circuit was higher than the current in the main discharge $I_m > I_{\text{ch}}$, we observed quiescent discharges with ion-beam currents up to 100 mA. For beam currents above 100 mA, the optimum condition was $I_m < I_{\text{ch}}$. For this reason, the operating conditions were chosen such that the resistance in the auxiliary-discharge circuit was fixed and equal to $R_M = 32\ \Omega$, whereas the resistance in the main-discharge circuit varied over a wide range (Fig. 7). The required value of the beam current in the injector under construction is 200–300 mA, which corresponds the range of ballast resistances in the main-discharge circuit from 5 to 20 Ω.

At the minimum flow rate of the working gas (H_2), the auxiliary discharge in the magnetron was excited at a potential of 400–600 V. As the working-gas flow rate or the input power was increased, the main discharge was excited; as a result, the distribution of the potentials over the electrodes (i.e., the voltage and current distributions along the discharge) changed. After the main discharge was excited, the plasma filled the expander and the ion-beam current began to be extracted (Fig. 8). When intense ion beams were extracted, we usually observed steplike initial distributions of the voltage and current along the discharge. These steps are associated with transitions from one form of a glow discharge to another. The ion beam was extracted and formed when the main discharge operated in anyone of its possible forms, each corresponding to a “shelf” in the current and voltage waveforms (Fig. 8). The source operation can easily be adapted to the required form of the discharge by changing the ballast resistances, regulating the working-gas rate, or changing the power deposited in the discharge with the help of the modulator. In this case, the waveforms of the discharge and beam currents acquired a rectangular (i.e., single-shelf) shape. The auxiliary-discharge current could reach 100 A; however, under the optimum operating conditions of the plasma generator, its value was in the range 5–30 A. The main-discharge current could reach 200 A; however, for ion beams with currents of 200–300 mA, its value was in the range 20–60 A. For low-current (below 50 mA) beams, the voltage across the main discharge was maintained in the range 250–300 V. As the ion-beam current increased, the voltage across the main discharge decreased and the discharge current increased. The increase in the discharge current was accompanied by the change in the glow color from pale rose to red-orange. At currents of the main-discharge above 100 A, the voltage across it could decrease to 50 V. This corresponds to an anomalous glow discharge rather than an arc discharge. The transition of the magnetron discharge into an arc was always accompanied by bright blue bursts, which occurred very rarely in a multicell magnetron cathode.

From the extensive literature devoted to duoplasmatrons, it is known that, as the discharge current increases, the percentage of H^+ and H_3^+ ions increases, whereas the percentage of H_2^+ ions decreases. The ion percentage also depends on the distance from the intermediate electrode to the anode [19]. In the ion source under study, this tendency was also observed, with the only difference that the ion percentage depended on the distance between the emission electrode and the end of the magnetron cathode (i.e., on the length of the main-discharge region). With a 1-mm gap, the proton content in the beam reached 40–45%. With a gap length of 6–8 mm, the proton content in the beam reached 75–85%, the other discharge parameters being the same. When the distance between the anode and the magnetron cathode was less than 3 mm, the plasma freely flowed out from the magnetron region into the main-discharge region of the plasma generator. When the distance between these electrodes was longer than 3 mm, the plasma flow was impeded and the plasma generator operated unstably. This fact provides indirect evidence that the plasma bubble does not form in the cone of the magnetron cathode. To provide the stable operation of the ion source, a diaphragm with an inner diameter of 4.5 mm was placed at a distance of 1 mm from the cone end of the magnetron cathode. The diaphragm was at the anode potential. Due to this diaphragm, the electric field in the gap between the magnetron cathode and the anode diaphragm was sufficient to maintain a stable plasma flow from one region to another. This provided stable ion-source operation at any distance between the magnetron cathode and the emission electrode. Under certain operating conditions, relaxation oscillations of the current and voltage of the main discharge arose. The frequency of these oscillations was several tens of kilohertz, and their modulation depth attained 10%. These oscillations gave rise to the beam-current oscillations with the same frequency. These oscillations were suppressed by changing the operating conditions of the main discharge. It should be noted that, when the distance between the magnetron cathode and the emission electrode was longer than 3 mm, the main discharge was often unstable and the discharge current exhibited relaxation oscillations. This can be attributed to the change in the strength and distribution of the electric field. Applying the above diaphragm led to the stabilization of the main discharge.

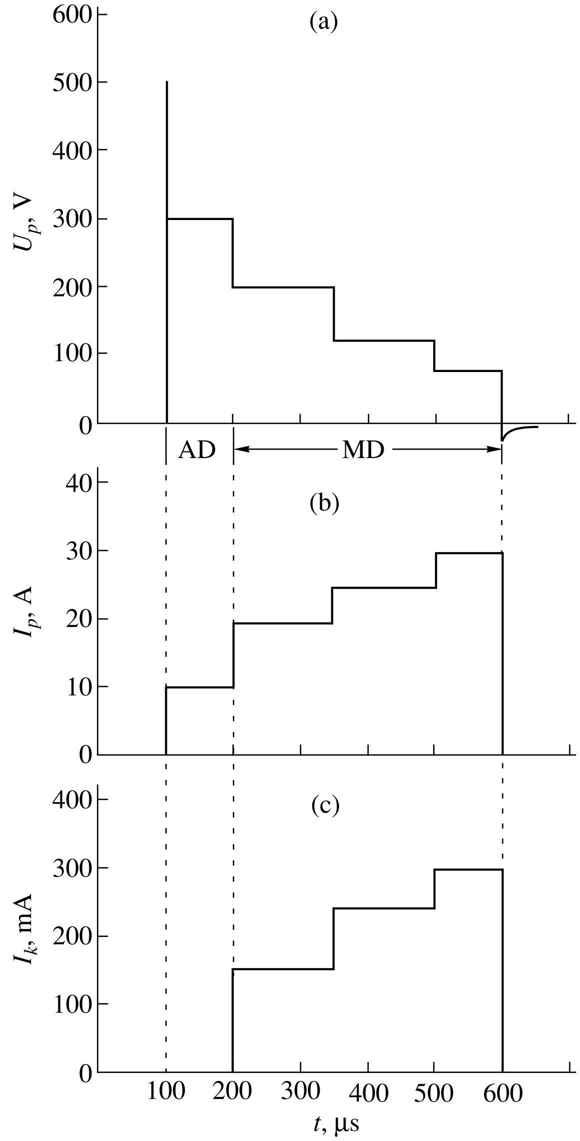


Fig. 8. Waveforms of the (a) discharge voltage, (b) discharge current, and (c) ion-beam current; AD and MD stand for the auxiliary and main discharge, respectively.

The conical expander that was used previously in the duoplasmatron operated unsatisfactorily. The improvement of various components of the duoplasmatron, including the expander, resulted in the creation of a new type of ion source. The design of the modified conical expander and the physical principles underlying its operation were described above. Here, we only

Table

Main-discharge current, A	Beam current measured by an FC, mA	Beam diameter, mm	Number of bright strips in the angular gauge	Length of two opposite sides, mm	Divergence angle of an ion jet, rad	Normalized ε, m rad
18	120	20	3	1.4	4×10^{-3}	2×10^{-7}
22	150	22	4	1.8	5.1×10^{-3}	2.85×10^{-7}
40	200	31	8	3.4	9.7×10^{-3}	7.5×10^{-7}
60	300	43	11	4.6	13.1×10^{-3}	1.4×10^{-6}

point out one of its advantages over the expander employed previously in the duoplasmatron. With the previous expander, it was impossible to apply an extraction voltage higher than 45 kV to the high-voltage gap because of frequent breakdowns. In contrast, with the modified expander, we could apply the extraction voltage as high as 70 kV to the high-voltage gap. The two-electrode system for extracting and forming the beam with the previous expander produced highly diverging beams, whose diameter at a distance of 15 cm from the source reached 180 mm. The two-electrode system for extracting and forming the beam with the modified expander allows one to produce weakly diverging beams, whose diameter at a distance of 15 cm from the source does not exceed 45 mm. This fact evidences that a concave meniscus does form in this type of expander. The parameters of the ion beam were measured in four different operating modes of a source with a two-electrode IOS at an extraction voltage of 45 kV. The aim of these measurements was to estimate the main characteristic of the source, namely, the beam emittance. The beam emittance was estimated roughly; in fact, it was somewhat overestimated, because the beam cross-sectional area was assumed to be equal to the area of a rectangle circumscribing the ellipse, instead of the area of the ellipse itself. In the center of the beam, in one transverse phase plane, a small jet was cut out from the beam with the help of a slit. Then, the jet divergence and the beam diameter were measured. The beam emittance was calculated by multiplying the jet divergence by the beam diameter and dividing the product by 4. In this manner, we estimated the emittance of the total current of the axially symmetric beam. The normalized emittance is given in the table.

We also conducted an experiment intended to estimate the maximum source emissivity. At a main-discharge current of 100 A, the beam current measured by an FC was about 2 A; however, the beam had a rather poor quality and exhibited strong current oscillations, which were accompanied by strong breakdowns in the source. This result is still to be verified.

When the source operated in a repetitive mode with an off-duty factor of 1000, no cooling was needed. However, when this factor was decreased to 100, it was necessary to cool the source.

At magnetron-discharge currents lower than 500 mA, the plasma generator could also operate in a continuous mode without cooling. The generator was powered from a BP-100 power source of a magnetic electric-discharge pump. We expect that, at higher input powers, but with a proper cooling, the ion source with a cold magnetron cathode and magnetic plasma compression will be able to generate continuous intense ion beams.

CONCLUSIONS

The ion source with a cold magnetron cathode and magnetic plasma compression is more advantageous than a duoplasmatron with the same cathode, because it has a simpler design, the higher gas efficiency, the higher emissivity, and the higher quality of the extracted ion beam.

The study of this ion source is still in progress. We plan to test the source with a new three-electrode acceleration–deceleration IOS and a double-focusing lens. Our estimates show that, in this case, the quality of the beams extracted from the source will improve and it will be possible to extract more intense beams with satisfactory parameters. In order to extract more intense beams, it will also be necessary to optimize the sizes of the expander and the conical insert. The optimization of magnetic fields in the plasma generator can lead to a more stable operation of the source in the regime of extracting intense ion beams.

ACKNOWLEDGMENTS

We thank S.V. Grigorenko for his attention to the manuscript and valuable remarks.

REFERENCES

1. M. Ardenne, *Tabellen der Elektronenphysik. Ionenphysik und Ubermikroskopie* (Berlin, 1956).
2. M. D. Gabovich, *Ion Plasma Sources* (Naukova Dumka, Kiev, 1964).
3. R. A. Demirkhanov, Yu. V. Kursanov, and V. M. Blagoveshchenskiĭ, Prib. Tekh. Éksp., No. 1, 30 (1964).

4. O. F. Poroshin and Zh. Zh. Kutan, *Production and Investigation of Intense Hydrogen-Ion Beams* (Fiz. Tekh. Inst. GK Inst. At. Energ. SSSR, 1963), p. 102.
5. L. J. Christiansen and E. J. Zaharis, Rev. Sci. Instrum. **37**, 1571 (1966).
6. V. T. Barchenko, S. N. Zagranichnyĭ, and A. Yu. Sokolovskiĭ, Izv. Leningr. Élektrotekh. Inst. im. V. I. Ul'yanova **303**, 42 (1982).
7. V. V. Nizhegorodtsev, in *Proceedings of the 5th All-Union Workshop on Charged-Particle Accelerators, Protvino, 1978*, Vol. 1, p. 368.
8. J. P. Delbarre, J. Faure, R. Noel, and R. Vienet, C. R. Acad. Sci. B (France) **262**, 1494 (1966).
9. L. P. Veresov, O. L. Veresov, M. I. Dzkuya, *et al.*, Zh. Tekh. Fiz. **71** (10), 50 (2001) [Tech. Phys. **46**, 1256 (2001)].
10. L. V. Arinin, G. V. Kuznetsov, A. I. Markoliya, *et al.*, in *Proceedings of the 9th All-Russia Meeting on Industrial and Medical Applications of Charged-Particle Accelerators, St. Petersburg, 1998*, p. 122.
11. Yu. P. Raizer, *Gas Discharge Physics* (Nauka, Moscow, 1987; Springer-Verlag, Berlin, 1991).
12. J. A. Thornton, J. Vac. Sci. Technol. **15**, 171 (1978).
13. R. A. Demirhanov, H. Frohlich, U. V. Kursanov, and T. I. Gutkin, Tech. Rep. Brookhaven Nat. Lab. BNL-767 (C-36), 224 (1962).
14. M. D. Gabovich, *Physics and Technology of Ion Plasma Sources* (Atomizdat, Moscow, 1972).
15. L. P. Veresov, P. A. Litvinov, Yu. P. Neustroev, and É. S. Nikolaev, USSR Inventor's Certificate No. 4,364,894 (1989).
16. L. P. Veresov, O. L. Veresov, and L. P. Skripal', Zh. Tekh. Fiz. **67** (9), 135 (1997) [Tech. Phys. **42**, 1104 (1997)].
17. E. A. Meleta, L. P. Veresov, V. N. Kizhin, *et al.*, Fiz. Khim. Stekla **17** (3), 77 (1991).
18. L. A. Artsimovich and S. Yu. Luk'yanov, *Charged-Particle Motion in Electric and Magnetic Fields* (Nauka, Moscow, 1972).
19. E. M. Kellog and K. E. Eklund, Rev. Sci. Instrum. **33**, 1338 (1962).

Translated by N. Larionova

Technical Physics, Vol. 48, No. 10, 2003, pp. 1346–1348. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 130–133.
Original Russian Text Copyright © 2003 by Tukmakov, Aksenov.

BRIEF COMMUNICATIONS

Object Identification from the Function of Number of Acoustic Response States

A. L. Tukmakov and I. B. Aksenov

Institute of Mechanics and Mechanical Engineering, Kazan Scientific Center,
Russian Academy of Sciences, Kazan, 420111 Tatarstan, Russia
e-mail: tukmakov@mail.knc.ru

Received November 10, 2002; in final form, January 10, 2003

Abstract—A method of acoustic response analysis to identify an echo signal source is suggested. It is based on the construction of the echo signal image in the phase space and on contrasting the dynamic process with a discrete set of states. A test for identity is worked out by comparing the discrete states of the echo signal from an object to be identified and the states typical of the echo signal from a known object. *© 2003 MAIK "Nauka/Interperiodica".*

The composition of an acoustic echo signal is usually analyzed by the methods of Fourier and wavelet expansions. In this case, a test of identity of object, i.e., sources of response, is found by comparing the coefficients of the echo signal expansion for given parameters of external excitation. In this paper, we suggest a method of object identification from an echo signal that is based on the analysis of a discrete set of dynamic system states [1, 2] and work out a test for identity.

We use the method to work out a test of identity of or difference between objects that are the sources of an acoustic echo signal. Let there be necessary to find the qualitative difference between two blades of the intermediate stage of a high-pressure compressor (Fig. 1a). The parts have standard dimensions but different geometries. An acoustic signal is generated when a striker hits the blade. The striker represents an inclined trough, in which a steel ball rolls down (Fig. 1b). The blade is mounted in a suspension, which reproducibly returns the blade to the same position relative to the striker and has a low damping factor.

The collision of the ball with the fixed blade generates acoustic vibrations, which are perceived by a capacitance microphone connected to the sound card of a computer. The power spectra of the acoustic signals from the first and second blades had the same structure but separated from each other along the frequency axis (Fig. 2).

THE FUNCTION OF NUMBER OF DYNAMIC SYSTEM STATES

To investigate an acoustic signal carrying information about an object, the function of number of dynamic system states is applied [1, 2]. Let us consider the image of a dynamic process $u(t)$ in the phase space using two dimensions for the sake of simplicity. The values of the function $u(t_i)$ at discrete time moments t_i are plotted along the longitudinal axis, and the difference $u(t_i) - u(t_{i-1})$, along the transverse axis [1, 2]. In

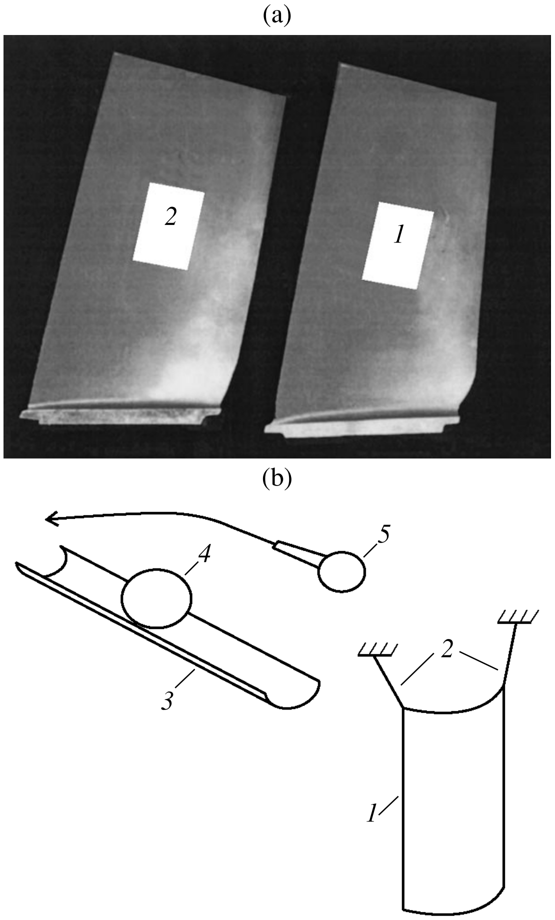


Fig. 1. (a) Compressor blades and (b) scheme of the experiment: (*1*) blade, (*2*) suspension, (*3*) trough, (*4*) steel ball; and (*5*) microphone.

the general case, the dimension of the phase space of embeddings is determined by the Hausdorff dimension of the attractor.

Let us introduce a set of states of the dynamic process considered and cover the attractor projection onto the phase plane by a grid with J and K meshes in the X and Y directions, respectively (Fig. 3). The mesh sizes are ΔX and ΔY [1, 2]. The attractor point in a mesh (j, k) is assigned to a system state with the number $(k-1)J+j$. Thus, a state of the dynamic system is described by a range ΔX of the function values and, up to a factor, by a range of rate of function variation ΔY. A state of the system in the multidimensional phase space is specified by the attractor point falling into a certain cell of the hypercube. To render our analysis efficient, the cells must be out of the regions of saturation and poor statistic [3].

We construct the function of number of dynamic system states as follows: at a current time instant, the function increases by unity if the state being analyzed is outside some base set. Depending on the depth of analysis, by a state we mean either a state $N(t_i)$ at a current instant of time t_i or a set of states $(N(t_i), N(t_{i-1}), N(t_{i-2}), \dots, N(t_{i-n}))$ that includes a number of preceding instants of time. The function of number of states thus defined responds to a change in the amplitude and frequency of a series realization in values in the base set. The sensitivity of the function may be varied in a wide range by appropriately selecting the dimension of a phase curve being analyzed (a phase curve in the space of embeddings or its projection) and the mesh sizes in the grid covering the area occupied by the phase curve. The sensitivity of the function also depends on the depth of the analysis, which may vary in time.

FORMATION OF THE BASE SET

A base set, with reference to which a new current state is defined, is formed during the analysis of the series realization of the signal and is complete if all (or almost all) its possible states are exhausted within a finite interval of time. When investigating an acoustic response to an external excitation, we will form a base set from states occurring in a number of blade-striker collisions. Consider how the rate of increase of the number of new states depends on the number of collisions. Let the function of number N of states increase by unity at a time instant t_i if a new state was previously absent. Otherwise, the function N remains unchanged. Simultaneously with the function, a base set, to where the current new state is added if it causes N to increase, is constructed.

ANALYSIS RESULTS

The function of number of states and the signal from the microphone in a series of collisions with blade 1 (the first eleven strikes) are shown in Fig. 4a. The rate of increase of the number N of new states decreases with every strike; i.e., new states are added to the base set, training the system. The relation between the number of new states and the serial number of a strike was obtained by the least squares method as applied to a power function. It turns out that the rate of system training is described by the Zipf law $\Delta N \approx A/n^{(1+\alpha)}$ [4], where $A = 573$, $\alpha = 0.38$, and n is the strike serial number (Fig. 4b). In this case, the Zipf alphabet is formed by a set of system states in the discrete phase space. Each of the states has its own repetition rate, which specifies the state rank in the base set [4].

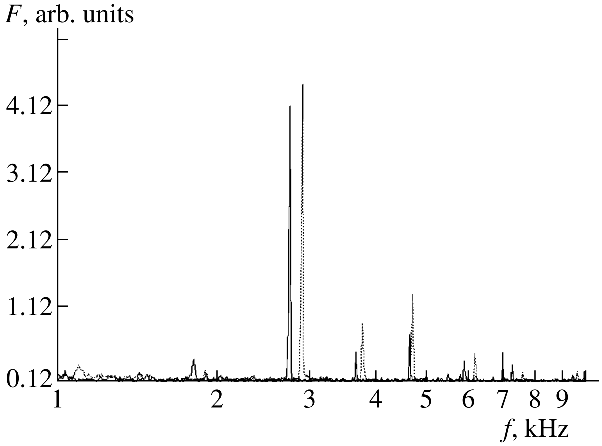


Fig. 2. Vibration power spectra for blades 1 and 2. The spectral lines of blade 2 (solid lines) are shifted to the left from those of blade 1 (dashed lines).

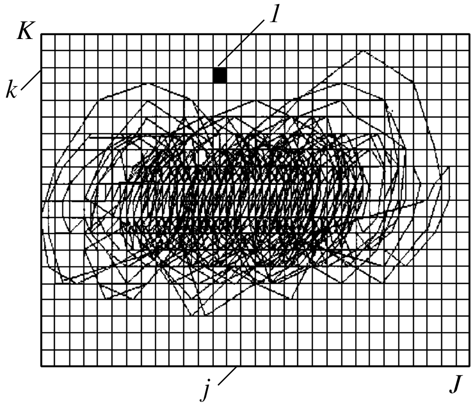


Fig. 3. States space of the dynamic system. (*1*) State $(k-1)J+j$.

The autoregressive scheme of increase of the state number, $\Delta N = A/n^{(1+\alpha)} + \varepsilon(n)$, where $\varepsilon(n)$ is the noise term, allows one to estimate the rate of system training [5]. If the increment of the number of states that are new for the basic set differs considerably from the value predicted by the autoregressive scheme, one can conclude that properties of the object have changed or its attributes do not fit the base set. For example, in Fig. 4a, the last strike was on blade 2 with the previously established parameters of excitation and fixing; as a result, the increment of the number of new states is much

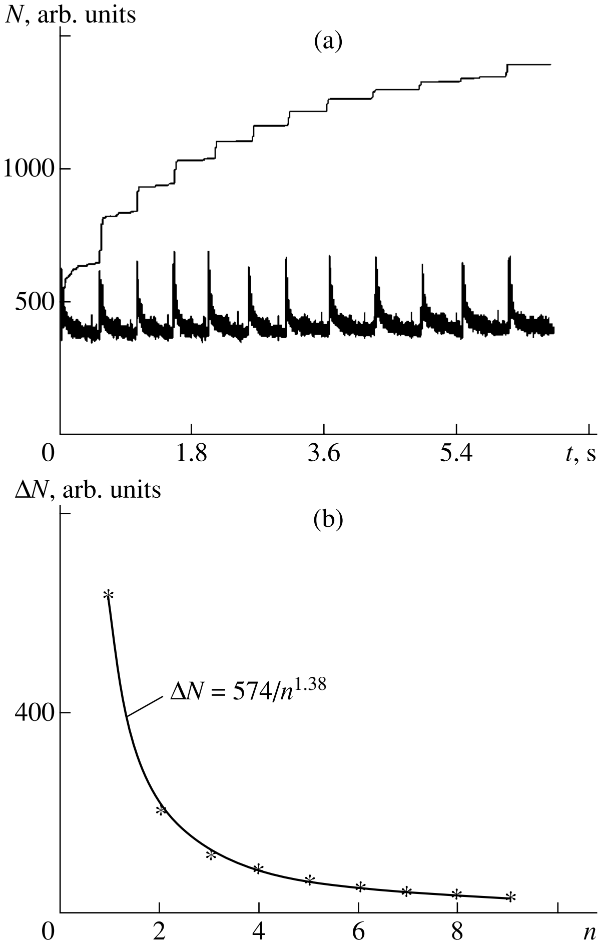


Fig. 4. (a) Function of number of states and a series of strikes on blade 1; (b) number of new states as a function of the strike number.

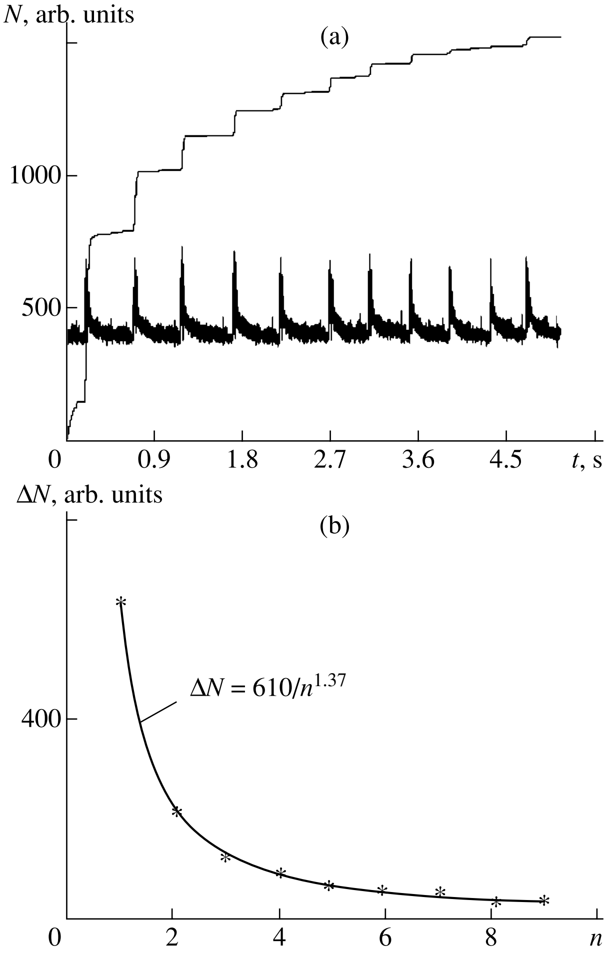


Fig. 5. (a) Function of number of states and a series of strikes on blade 2; (b) number of new states as a function of the strike number.

greater than predicted by the autoregressive scheme for blade 1 (Fig. 4b).

The results of system training aimed at recognizing the attributes of blade 2 are shown in Fig. 5a. The rate of training is given by the relationship $\Delta N \approx A/n^{(1+\alpha)}$, where $A = 610$ and $\alpha = 0.37$ (Fig. 5b). After the system training has been complete (this means that the number of new strike-induced states decreased), blade 2 was replaced by blade 1. This led to a sharp increase in the number ΔN of new states, which is inconsistent with the autoregressive scheme (Fig. 5a).

Thus, a function defined on a discrete set of states in the phase space of a dynamic system may be used to construct a system for recognizing objects with different geometric or structural features. In this case, the test for identity of or difference between objects is an increment of the number of acoustic response states that are new for the base set. The resolution of this function depends on the dimension of the space where the analysis is carried out; the depth of analysis, which may vary with time; and mesh size of the grid covering the area occupied by the phase curve. With the state-defining parameters properly adjusted, this function reveals the number of states that are new for the base set of states and also time instants at which the number of states increases. This allows one to find qualitative differences in the properties of objects responding to external excitation.

REFERENCES

1. A. L. Tukmakov, Pis'ma Zh. Tekh. Fiz. **28** (6), 18 (2002) [Tech. Phys. Lett. **28**, 224 (2002)].
2. A. L. Tukmakov, Zh. Tekh. Fiz. **72** (7), 137 (2002) [Tech. Phys. **47**, 929 (2002)].
3. H. G. Shuster, *Deterministic Chaos: An Introduction* (Wiley, New York, 1995; Mir, Moscow, 1998).
4. M. Shreder, *Fractals, Chaos, and Power Laws: Miniatures from Infinite Paradise* (Nauchn. Issledov. Tsentr "Regul. Khaot. Din.," Izhevsk, 2001).
5. *Analysis of Autoregressions*, Ed. by Yu. P. Lukashin (Statistika, Moscow, 1978).

Translated by M. Astrov

Technical Physics, Vol. 48, No. 10, 2003, pp. 1349–1353. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 134–138.
Original Russian Text Copyright © 2003 by Migal, Rom, Chugaj.

BRIEF COMMUNICATIONS

Modification of the Defect Structure and Electrophysical Properties in Zinc Selenide Crystals under the Action of an Alternating Electric Field

V. P. Migal*, M. A. Rom, and O. N. Chugaj***

** Zhukowsky National Aerospace University (KhAI), ul. Chkalova 17, Kharkov, 61070 Ukraine*
*** Research Concern Institute of Single Crystals, National Academy of Sciences of Ukraine,*
pr. Lenina 60, Kharkov, 61070 Ukraine
e-mail: kaf505@xai.edu.ua

Received February 11, 2003

Abstract—Modifications in the structure and electrophysical properties of ZnSe crystals subjected to a high alternating electric field of industrial frequency are investigated. It is established that such an action changes their defect structure and dielectric parameters. The latter exhibit different dependences on temperature and photoexcitation wavelength. The modifications are caused by ionic conduction, which is associated with local anomalies in the electric and elastic subsystems of the crystals. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

Investigation of external actions on the properties of solids may help in the understanding of the attendant processes, elucidation of the mechanisms behind phenomena observed, and control of the physical properties. Despite extensive research on the electrophysical properties of different materials, the effect of strong electric fields on the structure and properties of real crystals still remains little studied. In any case, we are not familiar with relevant works concerning II–VI compounds. This paper is an attempt to fill the gap in this field.

EXPERIMENTAL

The samples used were 5- to 6-mm-thick plates cleaved from bulk ZnSe crystals along the (110) planes. The ZnSe crystals were grown from melt in the argon atmosphere under a high pressure. The surface area of the samples was 500–700 mm^2. To these surfaces, an alternating electric field of industrial frequency was applied for 8 h. The intensity of the field reached 4×10^4 V/cm.

The permittivity ε' and loss tangent $\tan \delta$ were measured by R571 and R589 standard ac bridges in the frequency range of 10^2–10^4 Hz.

The crystal perfection was investigated by the X-ray diffraction method. Diffraction curves were taken using Cu$K\alpha_1$ radiation. We used both the symmetric (220) and (440) reflections, which do not broaden in the presence of stacking faults, and {111} reflections, which are sensitive to stacking faults, on the condition that $h + k + l = 3N \pm 1$, where $N = 0, 1, 2, \ldots$ [1].

If a ZnSe crystal represents a polysynthetic twin, its (110) stereographic projection also contains traces of two extra (111) planes. The rotation of the inclined sample about the normal to the sample surface enables us to take all four (111) diffraction reflections. From their relative intensity, one can judge the concentration of blocks of a certain orientation, whereas the total width of a diffraction curve, $\beta = I^R/I^m$ (I^R is the integral intensity and I^m is the maximum intensity), is a measure of the misorientation of mosaic blocks and concentration of stacking faults. The diffraction geometry used in [2] (the rotation of a crystal about the diffraction vector in the case of asymmetric Bragg reflection) makes it

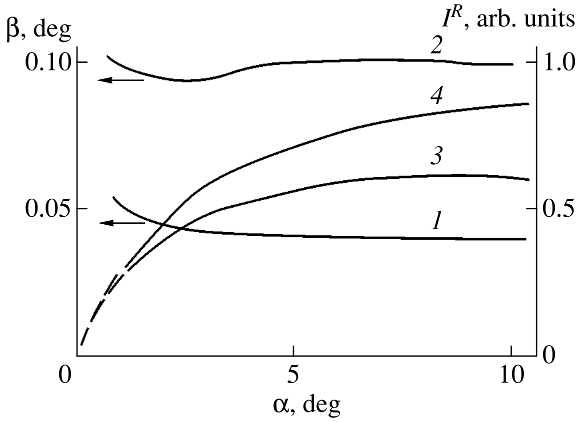


Fig. 1. β and I^R parameters (without corrections for the diffraction spread of the primary beam) of diffraction curves as functions of the glancing angle α of the primary beam. Curves *1*, *3* and *2*, *4* are taken from the crystal in the initial state and after the action of the alternating electric field, respectively. (110) ZnSe, (111) reflection, Cu$K\alpha_1$ radiation.

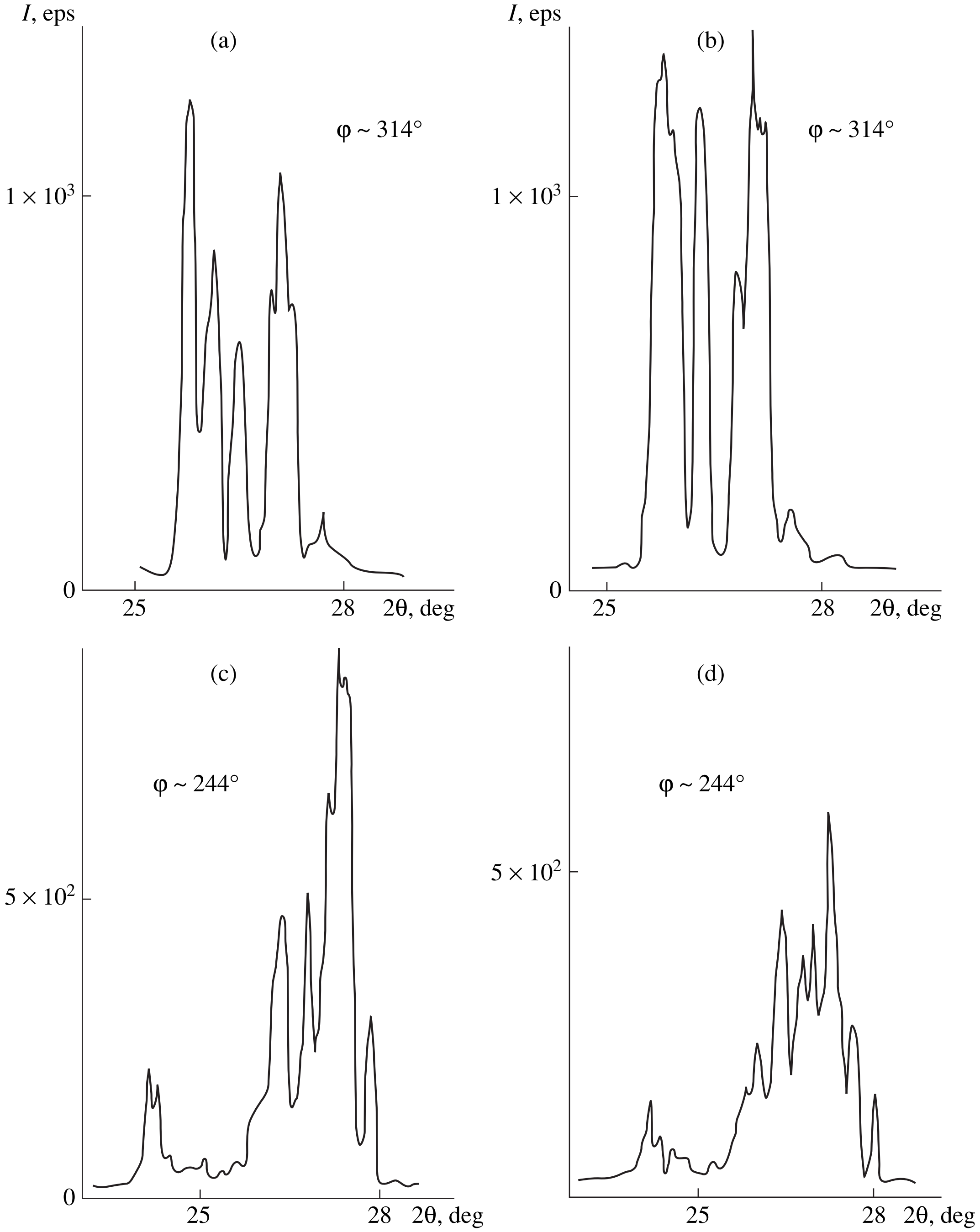


Fig. 2. Diffraction pattern from twins in the (110) ZnSe crystal ((111) reflections). The crystal is fixed, $\alpha \approx 1°$, and the detector scans the crystal within 2θ. Panels (a, c) and (b, d) refer to twin interlayers in the initial state of the crystal and after the action of the field, respectively. Angles φ are measured on the scale of a GP-14 attachment (Cu$K\alpha_1$ radiation).

possible to vary the thickness of a layer being analyzed. At glancing angles of the X-ray beam ($\alpha \approx 1°$), a considerable part of the sample surface is irradiated and one can obtain an analog of the twin structure topogram by scanning a narrow-slit detector.

RESULTS

The X-ray examination showed that the application of the alternating electric field adversely affects the crystal perfection of the ZnSe samples, as demonstrated by an increase in I^R and β for all the (111) reflections. In a number of cases, the value of I^R increases by 60% and β by 160% compared to their initial values. The dependences $I^R(\alpha)$ and $\beta(\alpha)$ which characterize the distributions of these quantities across the layer being analyzed are shown in Fig. 1. However, for other (111) reflections, I^R and β vary nonmonotonically. This suggests that, both near the surface (1–5 μm) and at a depth of 20–30 μm, which is accessible for Cu$K\alpha$ radiation, the competing processes of generation and annihilation of defects take place. 2θ scanning made it possible to detect new stacking faults and also the migration of twin interlayers. The migration shows up in a shift of the peaks and in the redistribution of the reflection intensity from separate blocks forming the polysynthetic twin. This is shown in Fig. 2.

The transformation of the defect structure in the crystals that were subjected to the alternating electric field is accompanied by the modification of their electrophysical properties: the low-frequency values of ε' and tan δ increase by 50 and 120%, respectively. The run of the temperature dependences of these parameters also changes (Fig. 3), but their weak frequency dependences in the frequency range mentioned above remains unchanged.

The alternating electric field also changes the photoelectric properties of the ZnSe crystals in the entire spectral range of photosensitivity. As a rule, one may observe a shift not only of the intrinsic maximum (λ_{m1} = 0.470–0.485 μm) but also of the impurity maximum (λ_{m2} = 0.535–0.575 μm) in the spectral dependences of the permittivity and dielectric loss factor ε'' = ε' × tanδ (Figs. 4 and 5, respectively). However, the features of the photodielectric effect in the sample are more evident in the dependences ε'(λ) and ε''(λ) represented in the form of a diagram on the complex plane [3]. Curvilinear or linear sections on the diagram correspond to the relaxation oscillator groups specified by the configuration of internal electric fields and distribution of nonequilibrium carriers over the crystal volume. Our experiments indicate that the application of the alternating electric field changes the sequence of the curvilinear and linear sections in this diagram, as well as the wavelengths at their edges (Fig. 6).

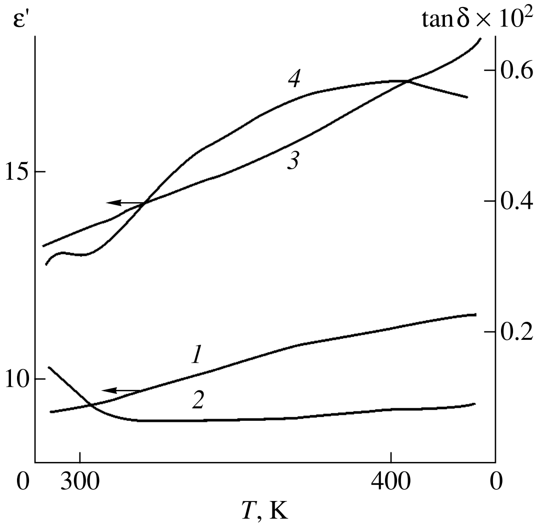


Fig. 3. Temperature dependences of ε' and tan δ in the initial state of the crystal and after the action of the field (curves *1*, *2* and *3*, *4*, respectively). f = 1 kHz.

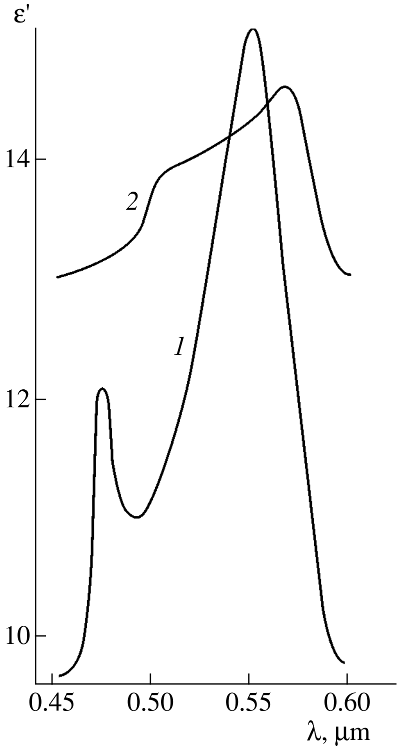


Fig. 4. Spectral dependences of ε' (*1*) in the initial state of the crystal and (*2*) after the action of the field. f = 1 kHz.

DISCUSSION

Initially, all the samples subjected to the alternating field were characterized by a high concentration of twin boundaries, dislocations, and stacking faults, which broaden considerably both the symmetric, (220), and asymmetric, (111), reflections. The defects formed during the growth and subsequent cooling of the ingots. Here, the basic source of the defects seems to be thermoelastic stresses, the nonuniform distribution of which on cooling generates a nonuniform field of residual mechanical stresses. For a number of reasons (the piezoelectric effect among them), large-scale fluctuations of the electric potential (internal electric field) also arose in the crystal. Note that these fluctuations are not screened by free carriers at the test temperature because of the large energy gap (E_g = 2.6 eV).

The dissipation of the alternating electric field energy near the dielectric strength limit causes a number of interrelated processes associated with charge and mass transfer (ionic current). These processes increase the general disorder in the system, including the number of defects in the crystals. Since structural defects have a decisive effect on these processes, the quantitative consequences of the action of the alternating electric field (residual changes in the elastic and electric subsystems) are bound to be different in different samples. However, common to all the samples is the signif-

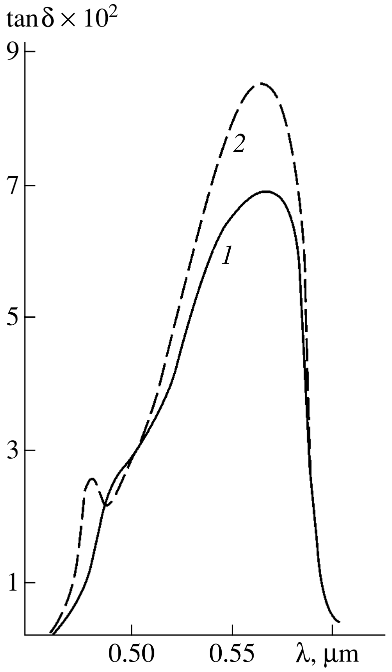


Fig. 5. Spectral dependences of tan δ (*1*) in the initial state of the sample and (*2*) after the action of the field. $f = 1$ kHz.

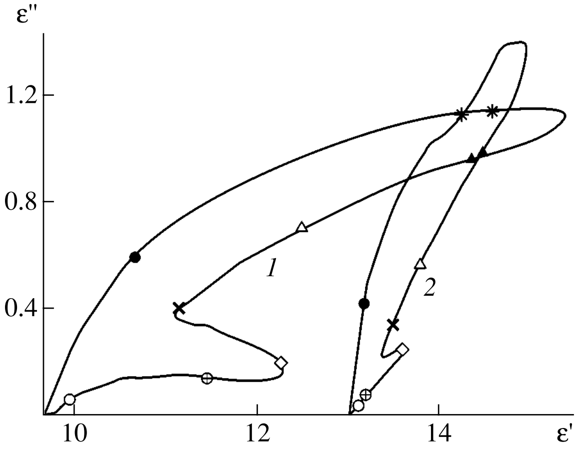


Fig. 6. Spectral dependences of ε' and ε'' on the complex plane. Curves (*1*) and (*2*) were taken before and after applying the alternating electric field, respectively. The light wavelength is (○) 465, (⊕) 470, (◇) 475, (×) 490, (△) 510, (▲) 530, (∗) 570, and (●) 590 nm.

icant effect of local anomalies in both subsystems on the processes occurring in them.

For example, in the case of the migration of interstitial ions, the hopping frequency of an ion from one equilibrium position to another is given by the expression [4]

$$\omega = \nu \exp(-\Delta g/kT),$$

where ν is the hopping frequency of an ion, Δg is the Gibbs free energy of activation, k is the Boltzmann constant, and T is absolute temperature.

The elastic and electric fields near an inhomogeneity, influencing Δg, may rise the hopping frequency of ions in a certain direction. As a result, the ions migrate (perhaps, in several hops) to the position where the value of Δg is much greater in any direction. In essence, the case in point is the trapping of an ion by a defect. As a result, the defect changes its configuration. Note that other physical processes associated with the ion migration (be it diffusion due to elastic stresses [5] or an electric field [6] or energy transfer by the relay-race mechanism [7]) are also responsible for the transformation of the defect structure. The motion of twin boundaries also becomes possible if we give up the assumption that the interface between blocks is atomically smooth. Indeed, a section oblique to the (111) plane that is normal to the [111] twinning axis showed curved (i.e., steplike) traces of twin boundaries on the surface.

As was mentioned above, because of different defect structures, as well as different internal electric and elastic field configurations, in the ZnSe crystals, the modifications of their physical properties under the action of the alternating electric field also differ. Common to all the samples is an increase in the dielectric loss factor and, correspondingly, in the rate of the alternating electric field energy dissipation. This fact testifies to a relation between the transformation of the defect structure and the modification of the electrophysical properties, since it appears quite natural that, as the structural disorder grows, so do the dielectric losses in the crystal. The increase in the electrical permittivity, the stronger temperature dependence of the dielectric loss factor, and the changed spectral dependences of these parameters, reflect a new state of the defect structure and internal fields after the external action. In this state, the density of levels in the energy gap rises and the interaction energy of traps changes. One may suggest with caution that this new state is related to a different phonon spectrum.

CONCLUSIONS

The application of a strong alternating electric field to ZnSe crystals results in the transformation of the defect structure and, as a consequence, modifies many of their physical properties. However, the participation of a variety of defects in the transformation makes it difficult to separate out the dominating mechanism in this process. An important problem to be tacked is the instant the transformation of the defect structure is complete. In a number of experiments, such an action caused the formation of a conducting surface film, which, screening the field in the bulk of the crystal, prevented further transformations.

REFERENCES

1. Ya. D. Vishnyakov, *Stacking Faults in Crystal Structure* (Metallurgiya, Moscow, 1970).
2. M. A. Rom and I. N. Chukanova, Funct. Mater. **6**, 915 (1999).
3. Yu. A. Zagoruĭko, V. K. Komar’, V. P. Migal’, *et al.*, Fiz. Tekh. Poluprovodn. (St. Petersburg) **29**, 1065 (1995) [Semiconductors **29**, 552 (1995)].
4. A. Lidiard, *Ionic Conductivity* (Springer, Berlin, 1957; Inostrannaya Literatura, Moscow, 1962).
5. V. V. Kirsanov, S. B. Kislitsyn, and E. M. Kislitsyna, Zh. Tekh. Fiz. **58**, 1440 (1988) [Sov. Phys. Tech. Phys. **33**, 860 (1988)].
6. V. A. Kashcheev and P. P. Poluéktov, Zh. Tekh. Fiz. **61** (8), 51 (1991) [Sov. Phys. Tech. Phys. **36**, 875 (1991)].
7. A. C. Damask and G. J. Dienes, *Point Defects in Metals* (Gordon and Breach, New York, 1963; Mir, Moscow, 1966).

Translated by Yu. Vishnyakov

Technical Physics, Vol. 48, No. 10, 2003, pp. 1354–1357. Translated from Zhurnal Tekhnicheskoĭ Fiziki, Vol. 73, No. 10, 2003, pp. 139–142.
Original Russian Text Copyright © 2003 by Nosov, Derkachenko.

BRIEF COMMUNICATIONS

Aftereffect in Microhardness Testing of Corundum

Yu. G. Nosov and L. I. Derkachenko
Ioffe Physicotechnical Institute, Russian Academy of Sciences,
Politekhnicheskaya ul. 26, St. Petersburg, 194021 Russia
e-mail: p.antonov@mail.ioffe.ru

Received February 13, 2003

Abstract—Cleavage near microhardness indentations on the surface of corundum is found to appear long (20–30 days) after unloading. *© 2003 MAIK “Nauka/Interperiodica”.*

INTRODUCTION

Microhardness testing is widely applied to study the properties of corundum. Indentations, their variation, and the behavior of the material around them have been analyzed to investigate deformation mechanisms upon loading [1–3], deformation anisotropy [4], the temperature dependence of deformation [5], and acoustic emission [6]. It has been repeatedly noted that indentations produced by a diamond pyramid on the surface of corundum may contain cracks [2, 6–8]. Cracks appear at many indentations when the load is 80–100 g; at loads of 150–200 g, each indentation usually has several cracks. The basic types of cracks detected are shown in Fig. 1. If an internal crack running at an angle to the indenter axis reaches the surface, cleavage is observed near the edge of the indentation. Vertical cracks decrease the strength of the material, whereas inclined cracks cause cleavages; material separation; and, thus, erosion of the surface.

In the works cited above, either the behavior of the material was studied during loading or the indentation was examined immediately after unloading. In this work, we investigate the formation of cracks and cleavages in corundum with emphasis on cleavage arising near the indentation long after unloading.

EXPERIMENTAL

The material under test was 8 × 1.4-mm single-crystal corundum ribbons grown from melt by the Stepanov method. The as-grown (1000) face of the crystal and the (1000) surface prepared by grinding and polishing to a depth of 0.2 mm were indented. A PMT-3 microhardness meter with a Vickers pyramid was used to indent 15 × 8 × 1-mm corundum plates at a load of 200 g. Microindentation was performed in air at room temperature and normal humidity. Indentation, holding under load, and unloading were carried out by the standard method for 30–35 s. The length of the indentation diagonal was ≈14 μm at an indentation depth of 1.8–2.0 μm. Vertical cracks and sometimes cleavage appear near the indentations immediately after unloading. Visual observation and photos showed that new cleavages arise continuously within 20–30 days after unloading.

Figure 2 illustrates the cleavage evolution in time near six indentations after unloading. Early after indentation, cleavage is detected near indentation *6* alone. After 0.5 h, the only change in the pattern is the increase in the cleavage area. Within 16 h, cleavage near indentation *2* appears and the size of the cleavage are near indentation *6* increases. In 17 days, cleavage near indentation *5* appears and the shape of the cleavage area near indentation *6* changes further.

From Fig. 2, we may draw the conclusion that the crack formation after unloading is to a great extent a probabilistic process. Under the same load, cleavages appear only near some of the indentations. From the

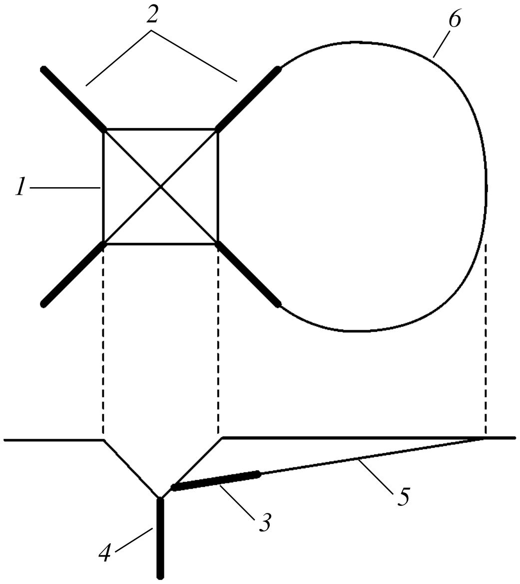


Fig. 1. Schematic of cracks forming near an indentation. *1*, indentation; *2*, radial cracks running normally to the sample surface; *3*, lateral inclined crack; *4*, median crack; *5*, cleavage; *6*, apparent contour of cleavage on the crystal surface.

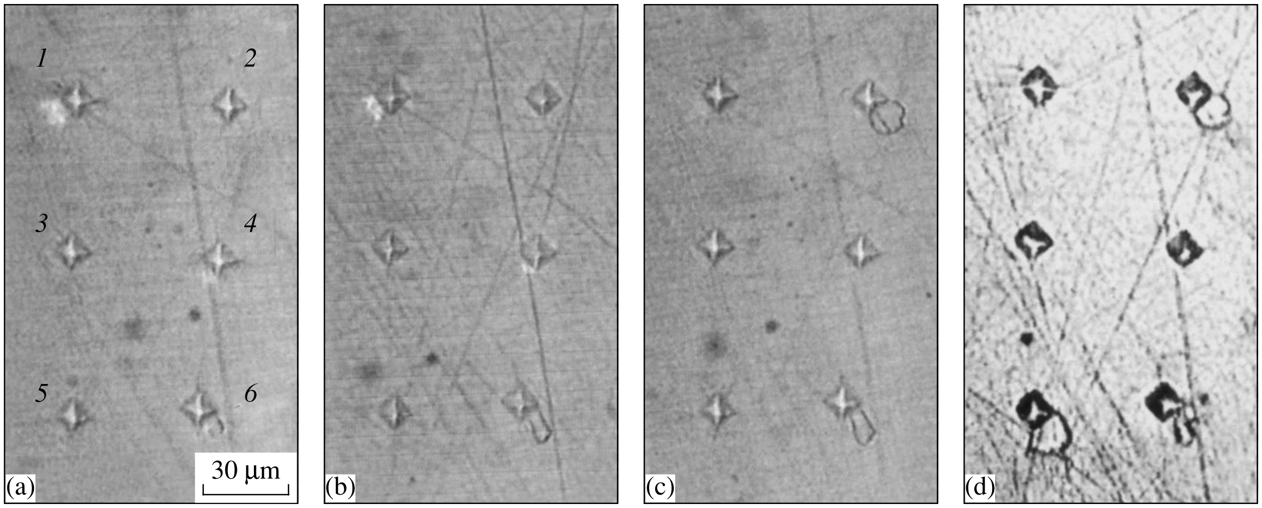


Fig. 2. Cleavage evolution on the surface of the corundum single crystal near the edges of indentations after unloading: (a) immediately after unloading, the initial stage of cleavage near indentation *6*; (b) 0.5 h after unloading, the size of cleavage near indentation *6* increases; (c) 16 h after unloading, cleavage near indentation *2*; and (d) 17 days after unloading, cleavage near indentation *5* and the change of the cleavage contour near indentation *6*.

appearance of the initial indentation, one cannot predict whether it will induce cleavage or not. For example, bright regions, which are likely to be caused by stresses or by incipient cracks, are visible near indentations *1* and *4* immediately after unloading; however, these indentations do not produce cleavages with cracks reaching the surface. On average, only 15–20% of indentations induce cleavage after unloading.

Analysis of a large number of indentations allowed us to construct the sequence of events involved in cleavage after unloading (Figs. 3a–3d). At the initial stage, radial cracks originating most often at the corners of the indentation are observed. Then, the crack paths bend, and the cracks move toward each other. Finally, the cracks meet each other to form a closed cleavage contour. Later on, the region bounded by this contour may separate from the crystal. Note that usually three cracks are involved in cleavage; one inclined crack and two vertical ones, and the development of only the latter two may be observed on the surface.

After unloading, usually one cleavage forms near one indentation; however, two (Fig. 3f) and even three cleavages may appear. Separating corundum particles have the shape of a plate with a maximum thickness (0.5–2.0 μm) in the region adjacent to the indentation and a minimum thickness (tending to zero) away from the indentation. The length of this plate is different for different indentations and varies from 0.1–0.2 to 2–2.5 times the indentation diagonal; i.e., it is at most 30–35 μm. A large cleavage is shown in Fig. 3e. Here, one comment is necessary. Before cleavage, this indentation was identical to its neighbors and had a standard size of ≈14 μm. After cleavage, its apparent diagonal grew to ≈18 μm. It is likely that the residual elastic field near the indentation sharply changes upon cleavage, which leads to a change (increase) in the indentation size.

The fracture (cleavage) surface was never mirror-smooth. It had steps and breaks of random orientation over a large area. The fracture surface near the thin edge was often wavy, with “waves” propagating from the indentation (Fig. 4).

Most of our observations of cleavage near indentations produced by the diamond pyramid were made on Stepanov-grown corundum crystals. Some of the experiments were carried out on the crystals annealed in a vacuum at 1800°C. Also, we examined corundum crystals grown from melt by the method elaborated in the State Optical Institute (St. Petersburg). In all the cases, the cleavage pattern long after unloading was the same.

DISCUSSION

Upon indentation, the energy of external forces is partially spent to alter the material shape (indentation formation) and to rupture the material (crack formation). The remaining part evolves as heat and is stored as the energy of the residual elastic field. With corundum, it was shown that the energy stored may do work in the region adjacent to an indentation long after unloading.

Since cleavage studied in this work can be considered as material fracture, two remarks are needed.

(1) Any fracture is known to follow plastic deformation [9], which is also true for such a brittle material as corundum [5, 10]. However, a high binding energy in Al_2O_3 (5–7 eV [11]) and a high activation energy of dislocation motion (≈5 eV [12]) prevent the restructuring

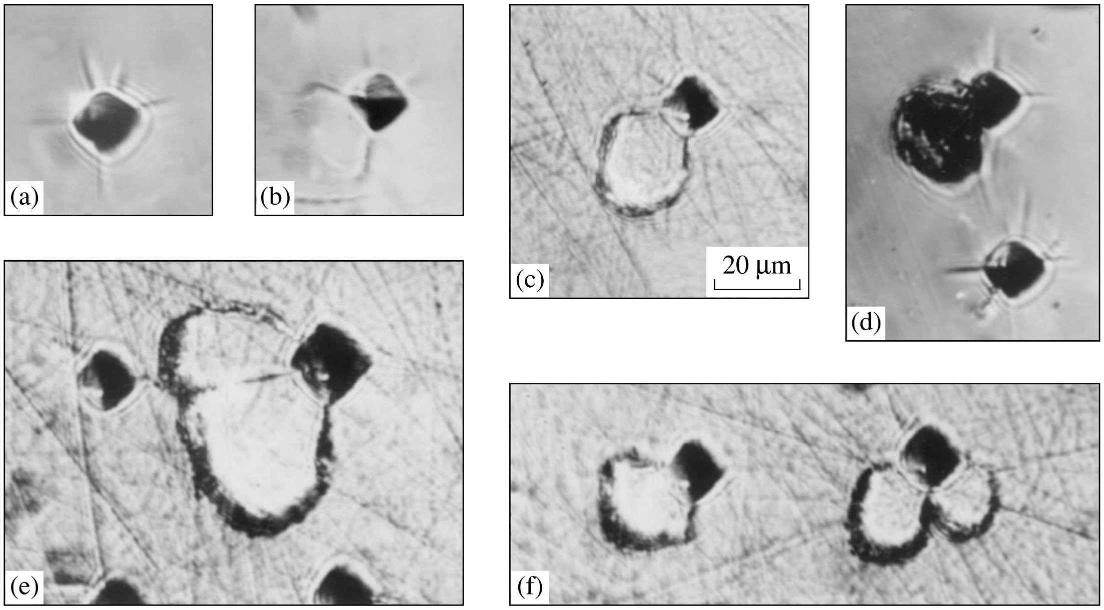


Fig. 3. Various cleavages formed near the indentations on the surface of the corundum single crystal after unloading. (a–d) The stages of cleavage: (a) an indentation with many straight cracks originating at the indentation corners, no cleavage; (b) two cracks bend toward each other; (c) the two cracks meet to form a closed cleavage contour; and (d) two indentations, the material at the upper indentation separated from the crystal. (e) An example of a large cleavage (≈35 μm in size) and (f) indentations with one and two cleavages.

of dislocation cores and dislocation motion at room temperature.

(2) Water adsorbed on the crystal surface significantly looses bonds in corundum. It was noted [10, 13–17] that the hydrolytic mechanism of fracture is mostly responsible for cracking in corundum upon loading at near-room temperatures. This mechanism consists in breaking Al–O bonds on the surface under the action of atmospheric moisture. For example, the strength of corundum in a vacuum at 10°C is higher than that in humid air by 32% [10]. Water vapor affects the fracture kinetics. Chen and Knapp [13] established that the average service time of corundum ceramics in dry argon exceeds that in air of 50% humidity by four orders of magnitude at the same stresses. In experiments on cyclic loading of single-crystal synthetic corundum, the fracture activation energy was found to be 1.3 eV [14], which led them to assume that water adsorbed upon fracture is of crucial importance. The anomalous microcreep of corundum at room temperature [15] is related to moisture present on the surface of the samples.

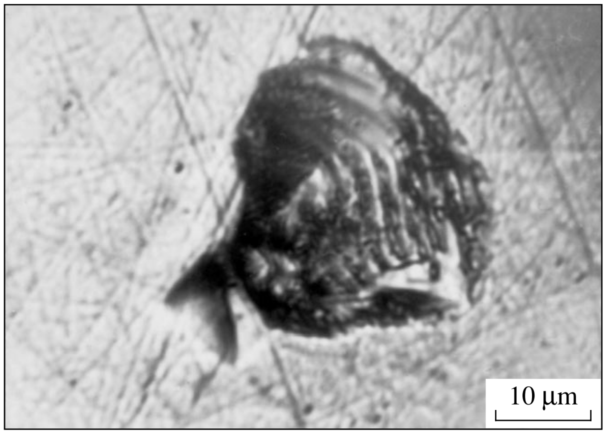


Fig. 4. Indentation with a wavy cleavage surface.

In view of the aforesaid, we may believe that cracking and cleavage after unloading is due to residual tensile elastic stresses at the crack tips. The plastic deformation of the material is facilitated by the hydrolytic mechanism of fracture in this case.

From our results, it follows that when analyzing the microindentation of materials that are considered as brittle, one should take into account the kinetics and aftereffect of the process.

To conclude, one additional remark regarding the abrasive treatment of corundum is necessary. The polishing of corundum diamond pastes or powders produces scratches, which cause cracks and cleavages on the surface. It is natural to suppose that, as for indentation, cleavage with the separation of corundum particles near the scratches is also a long-term process; that is, the final microrelief of the surface is established within a certain time after the mechanical treatment. For scratches ≈2 μm in depth, this time is about one month.

ACKNOWLEDGMENTS

We thank V.M. Krymov for his attention and assistance.

REFERENCES

1. B. Rother, J. Mater. Sci. **30**, 5394 (1995).
2. M. S. Bobji and S. K. Biswas, Philos. Mag. A **73**, 399 (1996).
3. E. H. Yoffe, Philos. Mag. A **46**, 617 (1982).
4. M. Kaji, M. E. Stevenson, and R. C. Bradt, J. Am. Ceram. Soc. **85**, 415 (2002).
5. W. Kollenberg, J. Mater. Sci. **23**, 3321 (1988).
6. E. Breval, V. Srikanth, and E. C. Subbarao, J. Am. Ceram. Soc. **78**, 2541 (1995).
7. N. Yu. Ikornikova, *Microhardness* (Moscow, 1951), pp. 211–225.
8. N. Yu. Ikornikova, Tr. Inst. Kristallogr. Akad. Nauk SSSR, No. 8, 293 (1953).
9. A. V. Stepanov, *Fundamentals of Crystal Hardness: Technical Applications* (Nauka, Moscow, 1974).
10. R. J. Stokes, in *Fracture*, Ed. by H. Leibowitz (Academic, New York, 1971; Mir, Moscow, 1975), Vol. 7, Chap. 1, p. 129.
11. V. I. Vedeneev, L. V. Gurvich, V. N. Kondrat'ev, V. A. Medvedev, and E. L. Frankevich, *The Energy of Chemical Bond Dissociation* (Izd. Akad. Nauk SSSR, Moscow, 1962).
12. *Ruby and Sapphire*, Ed. by L. M. Belyaev (Nauka, Moscow, 1974).
13. C. P. Chen and W. J. Knapp, J. Am. Ceram. Soc. **60**, 87 (1977).
14. V. V. Shpeĭzman, V. B. Vovnoboĭ, and A. I. Kozachuk, Fiz. Khim. Obrab. Mater., No. 3, 113 (1982).
15. J. H. Westbrook and P. J. Jorgensen, Trans. AIME **233**, 425 (1965).
16. V. A. Bershteĭn, *Mechanohydrolytic Processes and Hardness of Solids* (Nauka, Leningrad, 1987).
17. V. A. Ivantsov, V. I. Nikolaev, R. S. Chudnova, *et al.*, Izv. Ross. Akad. Nauk, Ser. Fiz. **58** (9), 63 (1994).

Translated by K. Shakhlevich