Physica D 238 (2009) 540–548

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Modulation instability: The beginning

V.E. Zakharov [a,b], L.A. Ostrovsky [c,d,*]

[a] Department of Mathematics, University of Arizona, Tucson, AZ, USA
[b] Lebedev Institute of Physics, Russian Acad. Sci., Russia
[c] Zel Technologies/NOAA Earth Science Research Laboratory, USA
[d] Institute of Applied Physics, Russian Acad. Sci., Russia

ARTICLE INFO

Article history:
Received 20 March 2008
Received in revised form
26 November 2008
Accepted 4 December 2008
Available online 24 December 2008
Communicated by A.C. Newell

PACS:
41.20.Jb
42.65.SF
42.65.TG
47.10.Df
47.35.Bb

Keywords:
Instability
BFL criterion
Nonlinear Schrodinger equation
Envelope waves
Active systems

ABSTRACT

We discuss the early history of an important field of "sturm and drang" in modern theory of nonlinear waves. It is demonstrated how scientific demand resulted in independent and almost simultaneous publications by many different authors on modulation instability, a phenomenon resulting in a variety of nonlinear processes such as envelope solitons, envelope shocks, freak waves, etc. Examples from water wave hydrodynamics, electrodynamics, nonlinear optics, and convection theory are given.

1. Introduction

To give the reader an idea of how widespread is the notion of modulation instability (MI), we can recommend to do a simple Internet search. There are between one and two million entries on "Modulation instability" and even more for "Self-modulation" in, e.g., Yahoo. Even if these references are not all equally relevant, the numbers are still impressive. We believe that most of the researchers in the area of nonlinear waves would agree that the MI is one of the most ubiquitous types of instabilities in nature. Thus, it seems useful to briefly outline the beginnings of the research in this area: it is remarkable that different groups of physicists in different countries have started research in this area almost simultaneously, albeit independently, an indicator that the idea was emerging when the time was indeed ripe.

In its simplistic version, the effect of modulation instability is the result of interaction between a strong carrier harmonic wave at a frequency ω, and small sidebands $\omega \pm \Omega$. This is the particular case of four-wave interaction (two quanta at ω create the quantas at $\omega + \Omega$ and $\omega - \Omega$). Growth of the sidebands can be treated in terms of amplification of weak modulation imposed on a harmonic wave (Fig. 1).

At the same time, in modern nonlinear physics, MI (or self-modulation) is considered as a basic process that classifies the qualitative behavior of modulated waves ("envelope waves") and may initialize the formation of stable entities such as envelope solitons. It was observed in numerous physical situations including water waves, plasma waves, laser beams, and electromagnetic transmission lines. In theoretical models, the phenomenon was considered for even broader range of phenomena, from biological moleculae to galactics.

As mentioned, the development of the theory of MI started almost simultaneously and occurred in parallel in hydrodynamics and electrodynamics/nonlinear optics. Thus, it should be stressed from the very beginning that our goal is not to set priorities, but on the contrary, to show how the same or similar ideas may arise independently when they are in demand.

* Corresponding author at: Zel Technologies/NOAA Earth Science Research Laboratory, USA.
E-mail address: Lev.A.Ostrovsky@noaa.gov (L.A. Ostrovsky).

0167-2789/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physd.2008.12.002

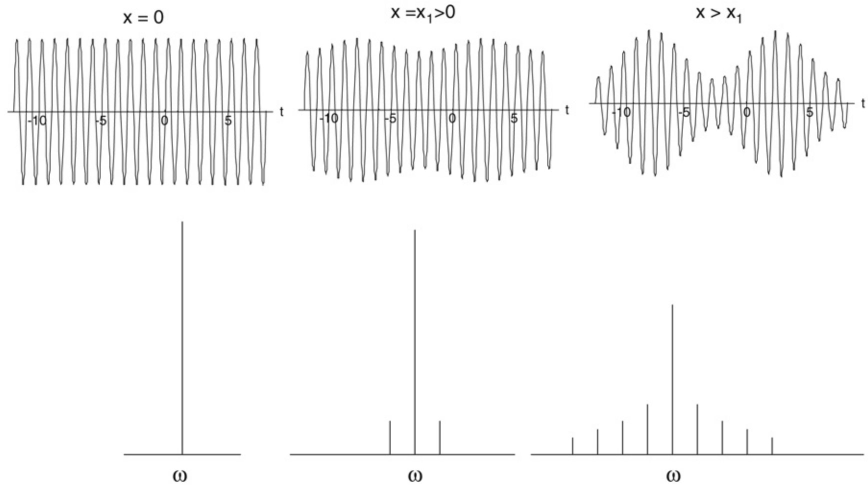

Fig. 1. Top: evolution of a nonlinear wave train in the course of MI. Bottom: the corresponding evolution of wave spectrum.

The mathematical models used in all works considered below are largely similar and universal. Most of the authors understood well this universality and often explicitly stated it or even used the universal approach from the very beginning. Still, our narrative here will follow, whenever possible, this parallel development that seems instructive and characteristic of the early progress. We limit ourselves by relatively few works, mostly from the 1960s and early 1970s, and only briefly mention the later, intensive development. Whenever possible we preserve notation and logics of the original papers.

It might be interesting to note that the research in this area had been started by the Western and Soviet scientists in the 1960s almost independently, and often implied different physical applications. Most of the early Western work has been related to classical hydrodynamics: water waves, convection etc. On the other hand, Soviet works on MI of about the same period were largely based on the then recent progress in electromagnetics, including nonlinear optics (lasers, self-focusing, nonlinear radiowaves etc.), and plasma physics (there were exceptions, however, even at that time; e.g., the Russian paper [1] cited below is concerned with water waves). Both authors of this paper began their scientific careers in the Soviet Union, with rather limited international contacts. We subsequently acquired an access to the papers by Whitham, then by Lighthill, and somewhat later by Benjamin and Feir. A publication in Russian of the materials of a discussion on nonlinear dispersive waves organized by Lighthill and published in English in 1967 (see Ref. [2]), especially enhanced our knowledge of the Western contributions to the area.

Naturally, here we have cited only published works and may have missed some interesting historical details, in particular relating to priorities. Indeed, A. Newell, in a private communication, has told us that when his advisor, D. Benney, first asked him, in spring of 1965, to look into the possibility of such an instability, it was because Benney had heard from Benjamin that he had had trouble reproducing the Stokes wave experimentally and believed it was unstable. Moreover, Newell also recalls Whitham saying at a much later date that he was initially puzzled by the fact that his modulation equations could be both hyperbolic (expected) and elliptic (unexpected) and it was only after he heard of Benjamin's result that "the penny dropped". It therefore may well have been that Benjamin was the first to derive, in the context of water waves, the criterion for the modulational instability. However, because the approach is quite general, and because it was from the Lighthill paper of 1965 [3] and from the volume [2] edited by Lighthill that we first learned of the result, we have decided to tell the story using Lighthill's version. In Section 3, we carry out the calculation for water waves, the path that Benjamin and Feir and, later, one of us followed.

2. Benjamin–Feir–Lighthill criterion

In 1965, Whitham [4] suggested the averaged variational principle for quasi-periodic waves based on a period-averaged Lagrangian, $\mathcal{L}(\omega, k, a)$ which depends on the wave phase θ (actually on its derivatives, local frequency $\omega = -\partial\theta/\partial t$ and wave number $k = \partial\theta/\partial x$), amplitude a, and possibly other slowly varying parameters. Using θ and a as canonical variables, one obtains equations describing slowly varying wave characteristics having in a 1-D case the form

$$\frac{\partial\mathcal{L}}{\partial a} = 0, \qquad \frac{\partial\mathcal{L}_\omega}{\partial t} - \frac{\partial\mathcal{L}_k}{\partial x} = 0, \qquad \frac{\partial k}{\partial t} + \frac{\partial\omega}{\partial x} = 0. \tag{1}$$

Lighthill [3] further developed Whitham's theory, considering a specific case of small nonlinearity where the averaged Lagrangian can be reduced to

$$\mathcal{L} = G(\omega, k)a^2 + B(\omega, k)a^4. \tag{2}$$

Variation of this over a gives

$$G(\omega, k) + 2B(\omega, k)a^2 = 0, \tag{3}$$

or, after resolving with respect to ω,

$$\omega = \omega_0(k) + \omega_1(k)a^2. \tag{4}$$

The latter expression can be considered as a nonlinear dispersion equation in which $\omega_0(k)$ follows from $G(\omega, k) = 0$ corresponding to the linear approximation, and ω_1 is due to nonlinearity. Note that in the linear case when $B = 0$, from Eq. (3) we have $\mathcal{L} = 0$. This relation has a simple mechanical interpretation: average values of the kinetic and potential energy densities are equal in a linear traveling wave.

The rest of the two equations (1) gives

$$\frac{\partial a^2}{\partial t} + \frac{\partial}{\partial x}(v_{gr}a^2) = 0, \tag{5}$$

and

$$\frac{\partial k}{\partial t} + v_{gr}\frac{\partial k}{\partial x} + \frac{\partial}{\partial x}(\omega_1 a^2) = 0, \tag{6}$$

where $v_{gr} = -G_k/G_\omega = \mathrm{d}\omega_0/\mathrm{d}k$ is the linear group velocity. Characteristic velocities for this system are

$$C_\pm = v_{gr}(k) \pm \sqrt{v_{gr'}\omega_1 a^2} + O(a^2). \tag{7}$$

Note that nonlinearity in Eq. (5) leading to the terms of order a^2 in Eq. (7) is neglected; these terms become important if dispersion is small (see Section 3.3 below). Hence, the above equations are hyperbolic (C is real) if

$$\beta = \omega_1 \mathrm{d}v_{gr}/\mathrm{d}k > 0 \tag{8}$$

and elliptic ($C_\pm$ are complex) if $\beta < 0$. We shall refer to this condition as *Benjamin–Feir–Lighthill (BFL) criterion*. As it is easy to deduce from these expressions, if Eqs. (5) and (6) are linearized around the harmonic wave with constant a, ω, and k, and perturbations are sought in the form of $\exp \mathrm{i}(Kx - \Omega t)$, the result is

$$\Omega = C_\pm K. \tag{9}$$

Hence, in the hyperbolic case the harmonic wave is stable, and, according to (7), can propagate with two slightly different "group velocities", whereas in the elliptic case it is unstable with respect to small modulation.

Lighthill then considered a weakly nonlinear Stokes wave on deep water when the nonlinear dispersion equation (4) reads

$$\omega = \omega_0(k)\left(1 + \frac{1}{2}k^2a^2\right), \quad \omega_0 = \sqrt{gk}. \tag{10}$$

In this case, from (8) it follows that

$$\beta = -\frac{\omega_0''^2}{8} < 0. \tag{11}$$

Hence, this is an elliptic case. Although Lighthill did not explicitly discuss wave stability in that paper, it is clear that, according to (7) and (9), a weakly nonlinear Stokes wave is unstable with an increment

$$\gamma = \mathrm{Im}(C)K = \frac{\omega_0(k)}{2} a v_{gr}(k)\Omega. \tag{12}$$

In this approximation, the increment increases monotonically with modulation frequency. As described below, a limitation of this result was established shortly thereafter in both water wave theory and in electrodynamics.

Modulation instability can also be explained as follows. Suppose that at some moment a local "bump" of intensity in a propagating wave appears. If, for example, $\omega_1 > 0$ in (4), the derivative ω_x is positive before the bump and negative after it. According to the wave phase conservation expressed in the last equation (1), that means that $k_t < 0$ before maximum and $k_t < 0$ after it. Suppose now that $\omega_{kk} = \partial v_{gr}/\partial k < 0$. Thus, the group velocity (more exactly, its linear part) tends to increase behind the peak and decrease in front of it. This means that the wave groups neighboring the amplitude maximum tend to compress the bump; due to the energy conservation, the amplitude increases cumulatively. The same reasoning shows that an initial trough in a harmonic wave would deepen. This is the case of modulation instability. In case of $\omega_{kk} > 0$, the effect is the opposite: the initial bump tends to be smeared; this is the case of neutral stability.

From the spectral viewpoint, a simple interpretation of the BFL criterion is as follows: small sidebands interact with the strong carrier wave; for their effective interaction, the simultaneous fulfillment of resonance (synchronism) conditions is needed:

$$\begin{aligned} &\omega_1 + \omega_2 = 2\omega_c, \qquad k_1 + k_2 = 2k_c, \quad \text{or} \\ &\omega_{1,2} = \omega_c \pm \Omega, \qquad k_{1,2} = k_c \pm K \end{aligned} \tag{13}$$

(the latter equalities are for slow modulation when Ω and K are small). Here subscripts c, 1, 2 correspond to the carrier wave and the sideband waves, respectively. In the linear case these conditions are not met because of dispersion. In the nonlinear case, the velocities of these waves differ due to two factors: dispersion (due to their frequency differences) and nonlinearity (they propagate on the background of the nonlinear carrier wave). In cases when these detunings are of the same sign, no synchronism occurs and the sidebands do not increase (hyperbolic case). If, however, these detunings differ in signs, they can compensate each other, and the waves interact in a synchronous, resonant manner, which results in their amplification. It is reflected in the BFL criterion; indeed, β is a product of the parameters of dispersion and nonlinearity.

3. Higher-order dispersion. Nonlinear Schrödinger equation

3.1. Water waves. Benjamin–Feir instability

Whitham's equations (1) can be considered non-dispersive with respect to the "complex envelope" $a(x, t)$. As a result, the BFL criterion (8) does not depend on modulation frequency (provided it is small as compared to the carrier frequency). However, a more detailed account of the effects of dispersion imposes an additional limitation on modulation instability.

For water waves it was demonstrated by Benjamin and Feir [5,6], who discovered modulation instability for nonlinear Stokes waves on the water surface. Such a discovery came as a surprise. Indeed, for decades the existence of stationary nonlinear (Stokes) waves on deep water was the subject of an involved mathematical proof (e.g., [7,8]). Suddenly it was determined that although such solutions do exist mathematically, they are unstable! Benjamin and Feir [5] experimentally demonstrated and theoretically explained this fact. Feir's experiments were performed in a water channel with a wave maker producing a wave with a length of 2.2 m; some results are shown in Figs. 2 and 3. Their theoretical model for such instability used a spectral approach, starting from the equations and boundary conditions for the one-dimensional potential, $\varphi(x, z, t)$, and the surface displacement, $z = \eta(x, z, t)$, in the form (for deep water)

$$\begin{aligned} &\varphi_{xx} + \varphi_{zz} = 0, \\ &\eta_t + \eta_x[\varphi_x]_{z=\eta} - [\varphi_z]_{z=\eta} = 0, \\ &g\eta + \left[\varphi_t + \frac{1}{2}\left(\varphi_x^2 + \varphi_z^2\right)\right]_{z=\eta} = 0, \end{aligned} \tag{14}$$

where g is gravity acceleration; $z = 0$ corresponds to a non-perturbed surface. A known solution of these equations is a progressive (Stokes) water wave, in which only the basic (first) and the second harmonics are retained:

$$\begin{aligned} &\eta = H \approx a\left(\cos\zeta + \frac{1}{2}ka\cos 2\zeta\right), \\ &\varphi = \Phi \approx \omega k^{-1} a e^{kz} \sin\zeta, \\ &\omega^2 \approx gk(1 + k^2a^2). \end{aligned} \tag{15}$$

Here $\zeta = kx - \omega t$, and a is its amplitude. Then small perturbations are added to this solution, each being represented as a sum of spectral components at frequencies $\omega \pm \Omega$, where Ω is a modulation frequency and $\Omega \ll \omega$. In other words, the wave is now represented in the form $\eta = H + \eta_1 + \eta_2$, $\varphi = \Phi + \theta_1 + \theta_2$. The sideband waves η_1 and η_2 are supposed to have amplitudes $\varepsilon_{1,2}$ and phases

$$\zeta_{1,2} = k(1 \pm \kappa)x - \omega(1 \pm \delta)t - \gamma_{1,2}, \tag{16}$$

where κ and $\delta = \Omega/\omega$ are small fractions satisfying the relation $\delta\omega = c_g \kappa k$, and $c_g = g/(2\omega)$ is the linear group velocity at the main frequency. The parameters $\gamma_{1,2}$ are corrections that arise due to dispersion (a difference of group velocities at the main wave and the side components) and to nonlinearity. If $\theta = \gamma_1 + \gamma_2$, the four-wave resonance mentioned in the Introduction occurs when $2\zeta = \zeta_1 + \zeta_2 + const$. As a result, the perturbations can increase, which is equivalent to the MI.

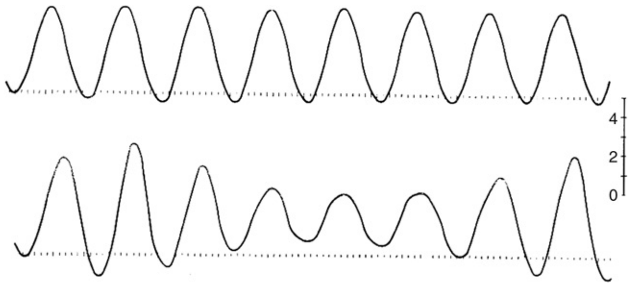

Fig. 2. Evolution of a wave train with the main frequency of 0.85 Hz in a water tank. Upper record is taken at a distance of 60 m from the wave maker; lower record is for 120 m. Time marks are at each 0.1 s. Vertical bar is in inches. From [6].

Fig. 3. Photographs of progressive wave trains illustrating the wave breaking due to the instability. Upper photograph is made near the wave maker; lower at 60 m from it. The main wave length is 2.2 m. From [6].

After substitution of the perturbed η and φ with slowly varying $\varepsilon_{1,2}(t)$ and $\theta(t)$ into Eq. (14) and keeping only resonance terms, the following equations follow after some transformations:

$$\frac{\mathrm{d}\varepsilon_\pm}{\mathrm{d}t} = \frac{1}{2}\left(\omega k^2 a^2 \sin\theta\right)\varepsilon_\mp, \tag{17}$$
$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \omega k^2 a^2 \left(1 + \frac{\varepsilon_1^2 + \varepsilon_2^2}{2\varepsilon_1\varepsilon_2}\cos\theta\right) - \Omega^2/\omega.$$

For a Stokes wave this yields instability with growth rate

$$\gamma = \frac{1}{2}\delta\left(2k^2a^2 - \delta^2\right)^{1/2}. \tag{18}$$

From here it is evident that the instability exists in a limited range of modulation frequencies,

$$\Omega < \Omega_s = \omega k a\sqrt{2}. \tag{19}$$

The maximum of growth rate (increment) is achieved at $\Omega = \Omega_s/\sqrt{2} = \omega k a$ (Fig. 4).

At small Ω formula (18) reduces to (12) following from Lighthill's consideration, but in general it represents a more specific condition for instability.

The works by Whitham, Lighthill, Benjamin, and Feir stimulated a lively discussion organized by M. J. Lighthill [2]. Within that event, the contributions by the above authors have expanded their previous studies. For example, Benjamin and Feir have shown that for water of finite depth h, the instability takes place at $kh > 1.363$, i.e., as expected, waves on shallow water are

Fig. 4. Dependence of the growth rate of the side-band amplitudes on frequency. From [5].

modulationally stable. Lighthill suggested a phenomenological averaged Lagrangian for Stokes waves. Whitham showed an equivalence between the spectral method used by Benjamin and Feir and his own modulational approach as regards to the MI. The range of relevant problems has been broadened by other authors. In particular, Phillips, Hasselmann, and Longuet-Higgins and Gill have studied resonance interactions of waves beyond the limits of the MI.

3.2. Hamiltonian approach for water waves

Zakharov [1] has shown that the equations of type (14) for weakly nonlinear waves on the surface of deep fluid can be reduced to a Hamiltonian form

$$\frac{\partial\eta}{\partial t} = \frac{\delta E}{\delta\varphi_s}, \qquad \frac{\partial\varphi_s}{\partial t} = -\frac{\delta E}{\delta\eta}, \tag{20}$$

where φ_s is the potential at the surface, $z = \eta$, and E is energy (Hamiltonian). Then the dynamic equations are expressed in terms of Fourier components $a(k)$ that can be considered as new complex canonical variables:

$$\eta(\mathbf{k}) = \sqrt{\frac{|\mathbf{k}|}{2\omega(\mathbf{k})}}\left[a(\mathbf{k}) + a^*(-\mathbf{k})\right], \tag{21}$$
$$\varphi_s(\mathbf{k}) = -\mathrm{i}\sqrt{\frac{\omega(\mathbf{k})}{2|\mathbf{k}|}}\left[a(\mathbf{k}) - a^*(-\mathbf{k})\right].$$

The resulting Hamiltonian equation is

$$\frac{\partial a(\mathbf{k})}{\partial t} = -\mathrm{i}\frac{\delta E}{\delta a^*(\mathbf{k})}. \tag{22}$$

The energy E is then represented as a series in powers of $a(k)$ and $a^*(k)$ up to the quartic terms, integrated over all ranges of wave vectors. For weakly nonlinear waves, the complex amplitudes can be presented in the form $a(k) \approx A(k,t)\exp[-\mathrm{i}\omega(\mathbf{k})t]$, where A is a slowly varying function.

For a wave packet with a narrow spectrum, the nonlinear Schrödinger equation (NSE) follows from here for the wave envelope; in one-dimensional case it has the form

$$\frac{\partial\varphi_s}{\partial t} - \frac{\mathrm{i}\lambda}{2}\frac{\partial^2\varphi_s}{\partial\xi^2} = -w|\varphi_s|^2\varphi_s, \tag{23}$$

where $\xi = x - v_{gr}t$, $v_{gr} = \mathrm{d}\omega/\mathrm{d}k$, $\lambda = \mathrm{d}^2\omega/\mathrm{d}k^2$. This equation has an obvious solution in the form of a constant-amplitude harmonic wave, the phase velocity of which depends on amplitude. Namely, at a given $k = k_0$, the frequency is $\omega = w_0 b_0^2$, where b_0 is proportional to the wave amplitude. Adding a perturbation so that