Steven B. Larkin
Editor

Lasers and Electro-Optics Research at the Cutting Edge

NOVA

LASERS AND ELECTRO-OPTICS RESEARCH AT THE CUTTING EDGE

LASERS AND ELECTRO-OPTICS RESEARCH AT THE CUTTING EDGE

STEVEN B. LARKIN
EDITOR

Nova Science Publishers, Inc.
New York

Library of Congress Cataloging-in-Publication Data

ISBN 978-1-60692-529-4

Published by Nova Science Publishers, Inc. ✣New York

CONTENTS

PREFACE

It is expected that ongoing advances in optics will revolutionize the 21st century as they began doing in the last quarter of the 20th. Such fields as communications, materials science, computing and medicine are leaping forward based on developments in optics. This new book presents leading edge research on optics and lasers from researchers spanning the globe.

The existence of resonant enhancement and near-field collimation of light waves by subwavelength apertures in metal films [for example, see T.W. Ebbesen et al., Nature (London) 391, 667 (1998) and H.J. Lezec et al., Science, 297, 820 (2002)] leads to the basic question: Can a light wave be enhanced and simultaneously localized in space and time by a subwavelength slit? To address this question, the spatial distribution of the electromagnetic field of an ultrashort (femtosecond) wave-packet scattered by a subwavelength (nanometer-size) slit was analyzed by using the conventional approach based on the Neerhoff and Mur solution of Maxwell's equations. The results show that a light wave can be resonantly enhanced by orders of magnitude and simultaneously localized in the near-field diffraction zone at the nm and fs scales. Chapter 1 includes many illustrations to facilitate an understanding of the natural spatial and temporal broadening of light beams and the physical mechanisms that are contributing to the resonantly enhanced scattering and localization of fs pulses by subwavelength nm-size metal slits. The results are discussed in the context of possible applications in the near-field scanning optical microscopy (NSOM).

As explained in chapter 2, accurate knowledge of absorption coefficient of a sample is a prerequisite for measuring the third order optical nonlinearity of materials, which can be a serious limitation for unknown samples. The authors introduce a method, which measures both the absorption coefficient and the third order optical nonlinearity of materials with high sensitivity in a single experimental arrangement. They use a dual-beam pump-probe experiment and conventional single-beam z-scan under different conditions to achieve this goal. They also demonstrate a counterintuitive coupling of the non-interacting probe-beam with the pump-beam in pump-probe z-scan experiment.

With the development of stable, compact and reliable pulsed laser sources the field of characterising materials, and in particular interfaces through their nonlinear optical response has bloomed. Second harmonic generation due to symmetry breaking at interfaces has provided a new spectroscopic tool of great utility. For example, it has been used to observe diffusion processes in semiconductors and to monitor the surface concentration of glucose oxidase in water. The surface response of various stained cellular features has been used as a source of illumination in microscopy, providing results similar to fluorescence microscopy but with less absorption and resulting damage.

The nonlinear optical response of various materials provides a very sensitive technique for measuring and monitoring some phase changes such as crystallisation from solution, micellisation of surfactants and polymorphic transitions of explosives. The work has focused on characterising and determining concentrations of pharmaceutically interesting bulk compounds, dispersions and emulsions. This work has potential application for in-line monitoring and quality control of pharmaceutical manufacturing.

In chapter 3 the authors present an extensive review of various spectroscopic techniques that make use of the nonlinear optical response of one or more media. They also present the results of our own work in this field.

Since the discovery of laser, optical radars or lidars have been successfully applied to the monitoring of the three main environments of our planet: lithosphere, hydrosphere and atmosphere. The transmission of light depends on the medium: while in soil it does not propagate, in water and air it can typically travel for meters and kilometers, respectively. Analogously, three theoretical frameworks can be established: lidar equations for hard, dense and transparent targets. Such different behaviors drive the range of lidar applications in the three above mentioned environments. In the lithosphere, laser radars have been applied to three-dimensional scans of underground cavities. In the hydrosphere, lidar fluorosensors have been very effective in the bio-optical characterization of the first layers of sea waters. Such instruments are usually aboard planes and ships and can help filling the gap between in situ measurements and satellite imagery. The medium where lidars are unbeaten, at least for some purposes, is air. Atmospheric applications of laser radars range from troposphere (e.g. pollution monitoring in urban areas and wind speed measurements) to stratosphere (e.g. polar stratospheric cloud detection and ozone hole assessment), and even mesosphere (e.g. profiling of K and Na). The purpose of Chapter 4 is twofold: from one hand, the interested reader is introduced in lidar science and technology, to the other one, the researcher familiar with laser remote sensing is faced with some current investigations. The first aim is achieved by introducing the lidar principles, for hard, dense and transparent targets, and by illustrating selected case studies, taken from the experience of the author, in order to exemplify some relevant applications of such principles. The second one is pursued by describing in more detail the most recent results obtained by the author in the field of environmental monitoring by laser radar.

In Chapter 5, thermal effects of diode-pumped solid-state (DPSS) lasers were theoretically analyzed and modeled for Nd doped lasing materials. The models determined the thermal lens and separated the end effect from the thermal effects. The knowledge of thermal lensing in the laser cavity is critical to scale the laser output power with the diffraction-limit laser beam output and to optimize the design of DPSS lasers. To validate the model, a new modified Twyman-Green interferometer is proposed to measure the thermal effects directly and accurately. Good agreement with the analytical model was obtained. Based on these pre-determined thermal lens effects, a power-scaling model was setup to guide the development of DPSS lasers. Within the fracture limits of the lasing crystals under high-power end-pumping, together with the design aims of the TEM_{00} mode output and the conversion efficiency, the model optimized the mode-to-pump ratio, doping concentration and dimensions of the lasing crystals. It also predicted the output power of DPSS lasers. The model showed a good agreement with the experiment later on. Based on the model, a practical diode-pumped Nd:YVO_4 laser system with a simple linear cavity was demonstrated. The laser produced an output power of 9.8W in the TEM_{00} mode under a pump power of 28W

where the rod surface acting as one of end mirrors in the linear plano-concave resonator. In conclusion, the models are quite useful for broadly guiding the design and development of most DPSS lasers, including rare earth ions doped YAG, YVO_4, and YLF, etc.

As presented in Chapter 6, optical catastrophe occurring in far field of waves of drop-refracted laser beam implies a wealth of information about drop spreading. When a parallel laser beam passes through a sessile drop on a slide glass, the rays are refracted by the drop and produce a new wave field in space after the drop. Very interesting optical image patterns occur on a screen far from the drop. Stable spreading-and-evaporations of sessile drops on an isothermal and isotropic slide-glass surface give nearly perfect circular far-field sectional images of the drop-refracted laser beam on the distant screen with a bright-thick ring and a set of fringes. The image configurations are time dependent and sensitive to the drop properties. In some cases of evaporating drops, rapidly varying optical image patterns occur on the screen. Classical geometric optics cannot correctly interpret these optical image patterns. An analysis based on catastrophe optics reveals and interprets the formation of these optical image patterns. The circular caustic line manifested as the bright-thick ring on the screen implies that the caustic is the lowest hierarchy of optical catastrophes, called fold caustic, which is produced from the initial wavefront with a local third-order surface given by the drop refraction. The set of fringes next to the fold caustic is the caustic diffraction instead of an aperture effect misunderstood by many investigators. The caustics and their diffraction fringes on the far-field sectional images of the drop-refracted laser beam can sensitively detect the interface instability of sessile drops, measure real-time local contact angles, estimate the drop-foot heights, visualize thermo-capillary convection flows, and investigate the spreading characteristics of evaporating liquid drops, such as drop profiles, effects of evaporation and thermo-capillary convection flows on the spreading, etc. The successful applications of catastrophe optics to the study of drop spreading convincingly show that catastrophe optics is not only able to qualitatively explain optical phenomena in nature but also able to quantitatively measure physical parameters of the objects of study, such as spreading drops.

The new results obtained in the area of repetitive and single longitudinal mode operation of TEA CO_2 lasers have been presented in Chapter 7. Helium-free operation of a conventional TEA CO_2 laser excited by a novel pulser circuit and its advantages in repetitive and single longitudinal mode lasing has been described. Much higher detunability of the single longitudinal mode achieved by making use of a traveling wave cavity in conjunction with a saturable absorber is discussed. A latch-proof pulser driven by an indigenously developed rotating dielectric spark gap switch makes the repetitive operation possible.

Chapter 8 discusses broadband infrared luminescence covering the optical telecommunication wavelength region of O, E, S, C and L bands (1200~1600nm), with FWHM (Full Width at Half Maximum) of 200~400nm and fluorescent lifetime of 200~700μs, which was observed from bismuth-doped SiO_2-Al_2O_3, GeO_2-Al_2O_3, GeO_2-B_2O_3, GeO_2-Ga_2O_3, GeO_2-Ta_2O_5, SiO_2-B_2O_3, SiO_2-Ta_2O_5, Al_2O_3-B_2O_3, Al_2O_3-P_2O_5 and ZnO-Al_2O_3-SiO_2 glass systems at room temperature in the case of 808nm excitation. In the absorption spectra of these glasses, a strong absorption peak near ~500nm was always detected in all samples, while absorption peaks at ~700, ~800 and ~1000nm were only in some of hosts such as GeO_2-M_2O_x glasses where M stood for B, Al, Ga and Ta. As usual, the emission and excitation peaks for Bi^{3+}-doped luminescent materials are located in ultraviolet or visible wavelength region, and the radiative lifetime seldom exceeds 10μs. Thus, the

luminescent properties of these Bi-doped glasses are distinguishably different from the Bi^{3+} doped luminescent materials previously reported in literatures. Compared with the experimental results on BiO molecule in gas phase reported by Fink *et al* and Shestakov *et al*, we tentatively assign the broadband infrared luminescence to the BiO molecule dissolved in these bismuth doped vitreous matrices. The observed absorptions at ~500nm, 700nm, 800nm and 1000nm could then be attributed to the transitions of $X_1 \rightarrow$(H, I), $X_1 \rightarrow A_2$, $X_1 \rightarrow A_1$ and $X_1 \rightarrow X_2$ respectively, and the emission peak at 1300nm could be to $X_2 \rightarrow X_1$. These glasses may have potential applications in widely tunable laser and super-broadband optical amplifier for the optical communications.

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN 1-60021-194-1
Editor: Steven B. Larkin, pp. 1-42

Chapter 1

RESONANT ENHANCEMENT AND NEAR-FIELD LOCALIZATION OF FS PULSES BY SUBWAVELENGTH NM-SIZE METAL SLITS

S. V. Kukhlevsky
1Institute of Physics, University of Pcs,
Ifjsg u. 6, Pcs 7624, Hungary

Abstract

The existence of resonant enhancement and near-field collimation of light waves by subwavelength apertures in metal films [for example, see T.W. Ebbesen et al., Nature (London) 391, 667 (1998) and H.J. Lezec et al., Science, 297, 820 (2002)] leads to the basic question: Can a light wave be enhanced and simultaneously localized in space and time by a subwavelength slit? To address this question, the spatial distribution of the electromagnetic field of an ultrashort (femtosecond) wave-packet scattered by a subwavelength (nanometer-size) slit was analyzed by using the conventional approach based on the Neerhoff and Mur solution of Maxwell's equations. The results show that a light wave can be resonantly enhanced by orders of magnitude and simultaneously localized in the near-field diffraction zone at the nm and fs scales. The chapter includes many illustrations to facilitate an understanding of the natural spatial and temporal broadening of light beams and the physical mechanisms that are contributing to the resonantly enhanced scattering and localization of fs pulses by subwavelength nm-size metal slits. The results are discussed in the context of possible applications in the near-field scanning optical microscopy (NSOM).

1. Introduction

In the last decade nanostructured optical elements based on scattering of light waves by subwavelength-size metal objects, such as particles and screen holes, have been investigated, intensively. The most impressive features of the optical elements are resonant enhancement and spatial localization of optical fields by the excitation of electron waves in the metal (for example, see the studies [1–76]). In the last few years, a great number of studies have been devoted to the nanostructures in metal films, namely a single aperture, a grating of apertures and an aperture surrounded by grooves. The study of collimation and

resonant enhanced transmission of waves in close proximity to a single subwavelength aperture acting as a microscope probe [1–5] was connected with developing near-field scanning microwave and optical microscopes with subwavelength resolution [6–9]. The resonant transmission of light by a grating with subwavelength apertures or a subwavelength aperture surrounded by grooves is an important effect for nanophotonics [10–76]. The transmission, on the resonance, can be orders of magnitude greater than out of the resonance. It was understood that the enhancement effect has a two-fold origin: First, the field increases due to a pure geometrical reason, the coupling of incident plane waves with waveguide mode resonances located in the slit, and further enhancement arises due to excitation of coupled surface plasmon polaritons localized on both surfaces of the slit (grating) [16, 38, 40]. A dominant mechanism responsible for the extraordinary transmission is the resonant excitation of the waveguide mode in the slit giving a Fabry-Perot like behavior [38]. In addition to the extraordinary transmission, a series of parallel grooves surrounding a nanometer-size slit can produce a micrometer-size beam that spreads to an angle of only few degrees [15]. The light collimation, in this case, is achieved by the excitation of coupled surface plasmon polaritons in the grooves [40]. At appropriate conditions, a single subwavelength slit flanked by a finite array of grooves can act as a "lens" focusing a light wave [54]. It should be noted, in this connection, that the diffractive spreading of a beam can be reduced also by using a structured aperture or an effective nanolens formed by self-similar linear chain of metal nanospheres [55, 56].

New aspects of the problem of resonantly enhanced transmission and collimation of light are revealed when the nanostructures are illuminated by an ultra-short (fs) light pulse [57–62]. For instance, in the study [57], the unique possibility of concentrating the energy of an ultrafast excitation of an "engineered" or a random nanosystems in a small part of the whole system by means of phase modulation of the exciting fs-pulse was predicted. The study [58] theoretically demonstrated the feasibility of nm-scale localization and distortion-free transmission of fs visible pulses by a single metal slit, and further suggested the feasibility of simultaneous super resolution in space and time of the near-field scanning optical microscopy (NSOM). The quasi-diffraction-free optics based on transmission of pulses by a subwavelength nano-slit has been suggested to extend the operation principle of a 2-D NSOM to the "not-too-distant" field regime (up to ~ 0.5 wavelength) [59]. Some interesting effects namely the pulse delay and long leaving resonant excitations of electromagnetic fields in the resonant-transition gratings were recently described in the studies [41, 60].

The great interest to resonant enhanced transmission, spatial localization (collimation) of continuous waves and light pulses by subwavelength metal apertures leads to the basic question: Can a light pulse be enhanced and simultaneously localized in space and time by a subwavelength slit? If the field enhancement can be achieved together with nm-scale spatial and fs-scale temporal localizations, this could greatly increase a potential of the nanoslit systems in high-resolution applications, especially in near-field scanning microscopy and spectroscopy. In the present study we test whether the resonant enhancement could only be obtained at the expense of the spatial and temporal broadening of a light wavepacket. To address this question, the spatial distribution of the energy flux of an ultrashort (fs) pulse diffracted by a subwavelength (nano-size) slit in a thick metal film of perfect conductivity will be analyzed by using the conventional approach based on the Neerhoff and Mur solution of Maxwell's equations. The chapter is organized as follows. In short, we first

will discuss the natural spatial and temporal broadening of light beams in free-space (Section 2.). The theoretical development of Neerhoff and Mur will be described in Section 3. and the model will then be used to calculate the spatial distribution of the electromagnetic field of the continuous wave (Section 4.) and pulse (Section 5.) scattered by the subwavelength slit under various regimes of the near-field diffraction. We will show that a light can be enhanced by orders of magnitude and simultaneously localized in the near-field diffraction zone at the nm and fs scales. The results presented in Sections 4. and 5. are discussed in the context of possible applications in the near-field scanning optical microscopy (NSOM). In Section 6., we summarize results and present conclusions.

2. Natural Spatial and Temporal Broadening of Light Beams

In Section 2., we consider several general properties of the optical beams, which are important for understanding the central topic of this chapter, the resonant enhancement and near-field localization of fs pulses by subwavelength nm-size metal slits. The material of the section is reprinted from Optics Communications, Vol. 209, S.V. Kukhlevsky and G. Nyitray, Correlation between spatial and temporal uncertainties of a wave-packet, 377-382, Copyright (2002), with permission from Elsevier. The section discuses the physical mechanisms that are contributing to the natural spatial and temporal broadening of continuous waves and light pulses in free space.

The natural spatial and temporal broadening of light beams in free space are closely connected with the spatial and temporal uncertainties of a wave-packet. The spatial ($\Delta\mathbf{x} \sim 1/\Delta\mathbf{k}$) and temporal ($\Delta t \sim 1/\Delta\omega$) uncertainties of a wave-packet are the universal laws of physics, and ones of the best understood. The detailed description of the phenomena can be found in optical textbooks. The uncertainty relations affect all classical and quantum-mechanical fields without exception. For instance, given the de Broglie postulate $\mathbf{p}$=$\hbar\mathbf{k}$, the spatial uncertainty in quantum mechanics is determined by the Heisenberg relation $\Delta\mathbf{x} \sim \hbar/\Delta\mathbf{p}$. Because of the uncertainty it is impossible to define at a given time both the position $\Delta\mathbf{x}$ of a wave packet and its direction $\Delta\mathbf{p}/p$ to an arbitrary degree of accuracy. The temporal uncertainty ($\Delta t \sim 1/\Delta\omega$), which in quantum mechanics is often called the fourth Heisenberg uncertainty relation ($\Delta t \sim \hbar/\Delta E$), shows that for a wave packet, the smaller the frequency uncertainty $\Delta\omega$, the more long the packet duration Δt. The mathematical principle behind the spatial and temporal uncertainties of wave packets involves a well-known relation between the widths of two functions, which are Fourier transforms of each other. Owing to the relation, the greater the energy uncertainty ΔE of a quantum system, the more rapid the time evolution Δt. Another example is diffraction phenomenon: the limitation of the phase space ($\Delta\mathbf{k}$) of a wave by the $x_{ap.}$ and $y_{ap.}$ boundaries of an aperture causes the spatial broadening of the wave in the $\{x\}\{y\}$ domain. According to the space-frequency uncertainty, the beam angle divergence $\Delta\mathbf{k}/k$ increases with decreasing the region of spatial localization $\Delta\mathbf{x}$. Similarly, the decrease of the spectral region $\Delta\mathbf{k}$ causes the increase of the value $\Delta\mathbf{x}$. According to the quantum-mechanics operator formalism, the spatial and temporal uncertainties arise because the conjugate variables must satisfy the Heisenberg postulate.

The spatial $\Delta\mathbf{x}$ and temporal Δt uncertainties of a wave packet are usually considered to be independent each other. In the frames of such consideration the unlimited small

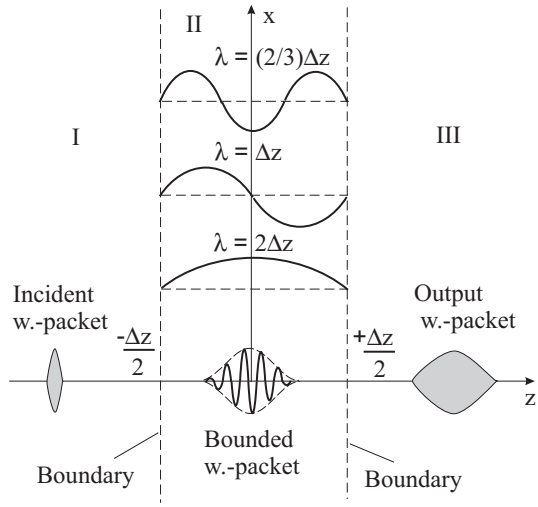


Figure 1. Propagation of a wave-packet through the bounded space.

temporal and spatial uncertainties ($\Delta t \to 0, \Delta \mathbf{x} \to 0$) of a wave packet can be achieved simultaneously. This can be realized, for instance, by using an ultra-short wave-packet ($\Delta t \to 0, \Delta \omega \to \infty$) passing through a small aperture ($\Delta \mathbf{x} \to 0, \Delta \mathbf{k} \to \infty$). The value of the central wavelength λ_0 of the wave packet can be arbitrary. The following simple consideration shows that the spatial ($\Delta \mathbf{x} \sim 1/\Delta \mathbf{k}$) and temporal ($\Delta t \sim 1/\Delta \omega$) uncertainties of a wave-packet correlate with each other, the smaller uncertainty $\Delta \mathbf{x}$, the bigger uncertainty Δt. The value of the uncertainty Δt is determined by the width $\Delta \mathbf{x}$ and the value of the central wavelength λ_0 of the wave packet.

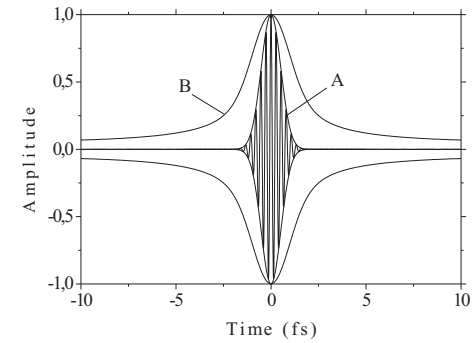
Let us investigate propagation of a wave packet of light through the bounded space. For the sake of simplicity, we consider an one-dimensional scalar packet passing through the space confined by two boundaries, as shown in Fig. 1. At a point $P(x, z)$ of the region I, an incident wave-packet $\psi(P, t)$ of the duration Δt_I can be presented in the form of the Fourier time expansion:

$$\psi(P, t) = \int_0^\infty \psi(P, \omega) \exp(-i\omega t) d\omega, \tag{1}$$

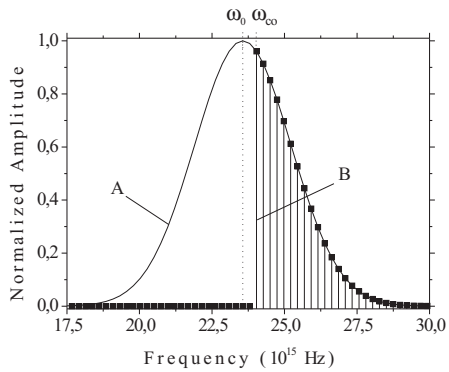
where $\omega = ck_z$ is the light frequency. The propagation of the Fourier components $\psi(P, \omega(k_z))$ in the regions I, II and III is governed by the Helmholtz equation

$$[\nabla_z^2 + k_z^2]\psi(P, k_z) = 0, \tag{2}$$

with boundary conditions imposed. In the non-bounded regions I and III, the Fourier spectra is continuous with the value $\omega(k_z) \in [0, \infty]$. In the confined space II, the values of momenta $k_z = \pi n/\Delta z$ are discrete ($n = 1, 2, ...$) and a nonzero lower limit $k_z^{min} = \pi/\Delta z$ exists due to the boundaries (see, Fig. 1). Compared with the regions I and III, the spectral width in the bounded region II is smaller, $\omega \in [\omega(k_z^{min}), \infty]$. According to the Fourier analysis, the decrease of the spectral width $\Delta\omega$ of the wave packet in the region II (the spectral interval changes from $[0, \infty]$ to $[\omega(k_z^{min}), \infty]$ causes increasing of the packet duration $\Delta t_{II,III}$ in



(a)



(b)

Figure 2. (a) The amplitudes of the original (A) and modified (B) wave-packets. (b) The Fourier spectra of the original (A) and modified (B) wave-packets. Here, $\psi(P,t) = exp[-2ln(2)(t/\tau)2]exp(i\omega_0 t)$; τ = 1 fs is the wave-packet duration; ω_0 is the central frequency of the wave packet. The figure demonstrates the broadening of the fs pulse by cutting off the wave-packet spectra at the frequency ω_{co}.

the regions II and III. As an example, Fig. 2. demonstrates the increasing of the width of a femtosecond wave-packet by cutting off the wave-packet spectra at the frequency $\omega_{co} = \omega(k_z^{min})$. The cut-off frequency ω_{co} depends on the boundary conditions. The approximate value is given by $\omega(k_z^{min}) = c\pi/\Delta z$. The cut-off frequency increases with decreasing of the width Δz of the packet. Thus, the temporal uncertainty $\Delta t_{II,III}$ is limited by the spatial uncertainty Δz. Such behavior indicates the correlation between the spatial $\Delta z \sim 1/\Delta k_z$ and the temporal $\Delta t \sim 1/\Delta\omega$ uncertainties, the smaller uncertainty Δz, the bigger uncertainty Δt. In the case of a wave-packet confined in three directions, the nonzero lower limits of the momenta $\mathbf{k}$ at x, y and z directions are given respectively by $k_x^{min} = \pi/\Delta x$, $k_y^{min} = \pi/\Delta y$ and $k_z^{min} = \pi/\Delta z$, and the cut-off frequency is given by

$$\omega(\mathbf{k}^{min}) = c((\pi/\Delta x)^2 + (\pi/\Delta y)^2 + (\pi/\Delta z)^2)^{1/2}. \tag{3}$$

It can be easily demonstrated, using Eq. 3 and the relation $\Delta t \sim 1/\Delta\omega$, that the space-time uncertainty of a wavepacket is given by

$$\Delta t \sim 1/(\omega_0 + \Delta\omega/2 - c((\pi/\Delta x)^2 + (\pi/\Delta y)^2 + (\pi/\Delta z)^2)^{1/2}), \tag{4}$$

where Δt is the duration of the spatially localized wave-packet, and ω_0 and $\Delta\omega$ are the central frequency and the spectral width of the free-space (unbounded) packet. The temporal uncertainty Δt of a wave-packet increases with decreasing the spatial uncertainties Δx, Δy and Δz. The increase of the temporal uncertainty Δt of the wave packet causing by decreasing its spatial uncertainty $\Delta\mathbf{x}$ can be reduced by increasing the central frequency ω_0 of the packet. In the limit of $\omega_0 \to \infty$, the light wave-packets are described by geometrical optics. We believe that the space-time uncertainty relation (4) can be considered as a general property of a wave-packet, because it is based on the basic properties of the Fourier transformation.

The spatial, temporal and space-time uncertainty relations for a light wave are in agreement with the results of exact analytical and numerical solutions of Maxwell's equations for both the under-wavelength ($\Lambda > \lambda$) and subwavelength ($\Lambda < \lambda$) domains. Here, Λ is the characteristic transversal dimension of the wave. For an example, in the case of a plane wave $E(z,t) \sim E(t)\exp(ik_z z)$, the pulse $E(t)$ of the duration Δt corresponds to the field $E(\omega)$ that is localized in the Fourier spectral region $\Delta\omega$. The regions of localization Δt and $\Delta\omega$ satisfy the temporal uncertainty relation ($\Delta t \sim 1/\Delta\omega$). In the under-wavelength domain, the angle divergence $\Delta\mathbf{k}/k$ determined by the spatial uncertainty $\Delta\mathbf{x} \sim 1/\Delta\mathbf{k}$ is in agreement with the diffractive broadening of a beam predicted by the Fresnel-Kirchhoff diffraction theory [63]. Because of the spatial uncertainty, light spreads out in all directions after passing a metal slit smaller than its wavelength as shown in Fig. 3 (see, for example Refs. [1–9]).

The harder one tries to decrease the beam transverse dimension by narrowing the slit, the more it broadens out. Similarly, the beam width dramatically increases with increasing the distance from the slit. The spatial uncertainty imposes a fundamental limit on the transverse dimension of a beam at a given distance from a subwavelength aperture and consequently limits the resolution capabilities and makes harder the position requirements of subwavelength-beam optical devices, such as near-field scanning optical microscopes (NSOM) and spectroscopes (see, Refs. [1–9]). One of the most important consequences

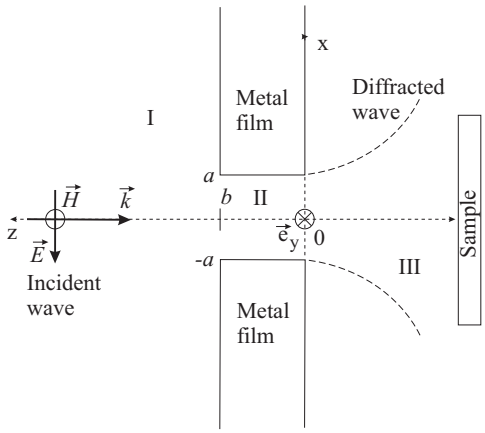


Figure 3. Diffraction of a continuous wave by a subwavelength nm-sized slit (waveguide) in a thick metal film and its implication to near-field optical microscopy. The slit width and length are $2a$ and b, respectively.

of the space-time uncertainty is that the duration Δt of a light pulse localized in the space increases with decreasing the region of spatial localization $\Delta \mathbf{x}$. In the under-wavelength domain, for an example, such a behavior is well known for the Gaussian beams [63, 64].

Although, the exact values of Δt and $\Delta \mathbf{x}$ of a light pulse scattered (localized) by a subwavelength aperture can be computed for the particular experimental conditions by solving the Maxwell equations, the correlation between the spatial and temporal uncertainties of the light wave-packet can be demonstrated without the computer simulation. Indeed, according to the subwavelength diffraction theory [1–5], the Fourier components $\psi(P, \omega)$ (continuous wave) having the frequency lower than the cut-off frequency ($\omega < \omega_{co}(z, \Delta x)$) does not contribute energy to the components $\psi'(P', \omega)$ of the diffracted wave-packet at a given distance z from the aperture. Thus, the effective duration $\Delta t(z = 0)$ of the wave-packet at the distance $z = 0$ increases due to the decrease of the effective spectral band width from $[0, \infty]$ to $[\omega_{co}, \infty]$. The effective increase of the duration $\Delta t(z = 0)$ implies an increase of the pulse duration $\Delta t'(z > 0)$ at a distance z, in other words, an increase in the temporal uncertainty associated with the incident wave-packet. The value of ω_{co} increases with decreasing of the width Δx and increasing of the distance z from the aperture [1–5]. Thus, the temporal uncertainty $\Delta t'$ is limited by the spatial uncertainty $\Delta x' = \Delta x'(z, \omega, \Delta x)$. The transverse dimension $\Delta x'$ of the diffracted wave-packet increases with increasing of the distance z and the value $1/\Delta x$. Such behavior indicates the correlation between the spatial $\Delta \mathbf{x} \sim 1/\Delta \mathbf{k}$ and temporal $\Delta t \sim 1/\Delta \omega$ uncertainties of the light wave-packet, which could be observed in the subwavelength diffraction experiment. It can be mentioned that the "cut-off frequency" phenomenon, in fact, is the Faraday screening of electromagnetic waves. The phenomenon is well-known also in theory of short waveguides, which can be considered as the limit of thick aperture. It should be noted that the time-space uncertainty correlation should be not limited to the case of diffraction on an aperture. Owing to Babinett's principle, the similar behavior is expected for the diffraction on a small particle (Rayleigh-Mie scattering). Thus, the correlation should affect the fs-scale formation

of highly enhanced fields localized in nm-size regions of random clusters, composites and rough surfaces [23]. The phenomenon probably will be important also for the future experiments on the scattering of attosecond and zeptosecond light pulses by electrons, neutrons and protons [22].

The spatial and space-time uncertainty relations are in agreement with the results of an exact numerical model presented in Section 3., which demonstrate respectively the spatial (Section 4.) and spatio-temporal (Section 5.) broadenings of a continuous wave and a pulse scattered by a subwavelength nm-size metal slit.

3. Model

An adequate description of the scattering of light through a subwavelength nano-sized slit in a metal film requires solution of Maxwell's equations with complicated boundary conditions. The light-slit interaction problem even for a continuous wave can be solved only by extended two-dimensional (x, z) numerical computations. The three-dimensional (x, z, t) character of the pulse-slit interaction makes the numerical analysis even more difficult. Let us consider the near-field diffraction of an ultrashort pulse (wave-packet) by a subwavelength slit in a thick metal screen of perfect conductivity by using the conventional approach based on the Neerhoff and Mur solution of Maxwell's equations. Before presenting a treatment of the problem for a wave-packet, we briefly describe the Neerhoff and Mur formulation [3, 5] for a continuous wave (a Fourier ω-component of a wavepacket). The transmission of a plane continuous wave through a slit (waveguide) of width $2a$ in a screen of thickness b is considered. The perfectly conducting nonmagnetic screen placed in vacuum is illuminated by a normally incident plane wave under TM polarization (magnetic-field vector parallel to the slit), as shown in Fig. 3 The magnetic field of the wave is assumed to be time harmonic and constant in the y direction:

$$\vec{H}(x, y, z, t) = U(x, z)\exp(-i\omega t)\vec{e}_y. \tag{5}$$

The electric field of the wave (5) is found from the scalar field $U(x, z)$ using Maxwell's equations:

$$E_x(x, z, t) = -\frac{ic}{\omega\epsilon_1}\partial_z U(x, z)\exp(-i\omega t), \tag{6}$$

$$E_y(x, z, t) = 0. \tag{7}$$

$$E_z(x, z, t) = \frac{ic}{\omega\epsilon_1}\partial_x U(x, z)\exp(-i\omega t). \tag{8}$$

Notice that the restrictions in Eq. 5 reduce the diffraction problem to one involving a single scalar field $U(x, z)$ in two dimensions. The field is represented by $U_j(x, z)$ (j=1,2,3 in each of the three regions I, II and III). The field satisfies the Helmholtz equation:

$$(\nabla^2 + k_j^2)U_j = 0, \tag{9}$$

where $j = 1, 2, 3$. In region I, the field $U_1(x, z)$ is decomposed into three components:

$$U_1(x, z) = U^i(x, z) + U^r(x, z) + U^d(x, z), \tag{10}$$

each of which satisfies the Helmholtz equation. U^i represents the incident field, which is assumed to be a plane wave of unit amplitude:

$$U^i(x, z) = \exp(-ik_1 z). \tag{11}$$

U^r denotes the field that would be reflected if there were no slit in the screen and thus satisfies

$$U^r(x, z) = U^i(x, 2b - z). \tag{12}$$

U^d describes the diffracted field in region I due to the presence of the slit.

With the above set of equations and standard boundary conditions for a perfectly conducting screen, a unique solution exists for the diffraction problem. To find the field, the 2-dimensional Green's theorem is applied with one function given by $U(x, z)$ and the other by a conventional Green's function:

$$(\nabla^2 + k_j^2)G_j = -\delta(x - x', z - z'), \tag{13}$$

where (x, z) refers to a field point of interest; x', z' are integration variables, $j = 1, 2, 3$. Since U_j satisfies the Helmholtz equation, Green's theorem reduces to

$$U(x, z) = \int_{Boundary} (G\partial_n U - U\partial_n G)dS. \tag{14}$$

The unknown fields $U^d(x, z)$, $U_3(x, z)$ and $U_2(x, z)$ are found using the reduced Green's theorem and boundary conditions on G

$$U^d(x, z) = -\frac{\epsilon_1}{\epsilon_2} \int_{-a}^{a} G_1(x, z; x', b) DU_b(x')dx' \tag{15}$$

for $b < z < \infty$,

$$U_3(x, z) = \frac{\epsilon_3}{\epsilon_2} \int_{-a}^{a} G_3(x, z; x', 0) DU_0(x')dx' \tag{16}$$

for $-\infty < z < 0$,

$$\begin{aligned} U_2(x, z) = &- \int_{-a}^{a} [G_2(x, z; x', 0) DU_0(x') - U_0(x')\partial_{z'} G_2(x, z; x', z')|_{z\to 0^+}]dx' \\ &+ \int_{-a}^{a} [G_2(x, z; x', b) DU_b(x') - U_b(x')\partial_{z'} G_2(x, z; x', z')|_{z\to b^-}]dx' \end{aligned} \tag{17}$$

for $|x| < a$ and $0 < z < b$. The boundary fields in Eqs. 15-17 are defined by

$$U_0(x) \equiv U_2(x, z)|_{z\to 0^+}, \tag{18}$$

$$DU_0(x) \equiv \partial_z U_2(x, z)|_{z\to 0^+}, \tag{19}$$

$$U_b(x) \equiv U_2(x, z)|_{z\to b^-}, \tag{20}$$

$$DU_b(x) \equiv \partial_z U_2(x, z)|_{z\to b^-}. \tag{21}$$

In regions I and III the two Green's functions in Eqs. 15 and 16 are given by

$$G_1(x,z;x',z') = \frac{i}{4}[H_0^{(1)}(k_1 R) + H_0^{(1)}(k_1 R')], \tag{22}$$

$$G_3(x,z;x',z') = \frac{i}{4}[H_0^{(1)}(k_3 R) + H_0^{(1)}(k_3 R'')], \tag{23}$$

with

$$R = [(x-x')^2 + (z-z')^2]^{1/2}, \tag{24}$$

$$R' = [(x-x')^2 + (z+z'-2b)^2]^{1/2}, \tag{25}$$

$$R'' = [(x-x')^2 + (z+z')^2]^{1/2}, \tag{26}$$

where $H_0^{(1)}$ is the Hankel function. In region II, the Green's function in Eq. 17 is given by

$$G_2(x,z;x',z') = \frac{i}{4a\gamma_0}\exp(i\gamma_0|z-z'|) + \frac{i}{2a}\sum_{m=1}^{\infty}\gamma_m^{-1} \times\cos[m\pi(x+a)/2a]\cos[m\pi(x'+a)/2a] \times\exp(i\gamma_m|z-z'|), \tag{27}$$

where $\gamma_m = [k_2^2 - (m\pi/2a)^2]^{1/2}$. The field can be found at any point once the four unknown functions in Eqs. 18-21 have been determined. The functions are completely determined by a set of four integral equations:

$$2U_b^i(x) - U_b(x) = \frac{\epsilon_1}{\epsilon_2}\int_{-a}^{a} G_1(x,b;x',b)DU_b(x')dx', \tag{28}$$

$$U_0(x) = \frac{\epsilon_3}{\epsilon_2}\int_{-a}^{a} G_3(x,0;x',0)DU_0(x')dx', \tag{29}$$

$$\frac{1}{2}U_b(x) = -\int_{-a}^{a}[G_2(x,b;x',0)DU_0(x') - U_0(x')\partial_{z'}G_2(x,b;x',z')|_{z'\to 0^+}]dx' + \int_{-a}^{a}[G_2(x,b;x',b)DU_b(x')]dx', \tag{30}$$

$$\frac{1}{2}U_0(x) = \int_{-a}^{a}[G_2(x,0;x',b)DU_b(x') - U_b(x')\partial_{z'}G_2(x,0;x',z')|_{z'\to b^-}]dx' - \int_{-a}^{a}[G_2(x,0;x',0)DU_0(x')]dx', \tag{31}$$

where $|x| < a$, and

$$U_b^i(x) = \exp(-ik_1 b). \tag{32}$$

The coupled integral equations 28-31 for the four boundary functions are solved numerically. The magnetic $\vec{H}(x,z,t)$ and electric $\vec{E}(x,z,t)$ fields of the diffracted wave in

region I, II and III are found by using Eqs. 15-17. The magnetic $\vec{H}(x,z,t)$ fields in regions I, II, and III are given by

$$H^1(x,z) = \exp(-ik_1 z) + \exp(-ik_1(2b-z)) - \frac{ia}{N}\frac{\epsilon_1}{\epsilon_2}\sum_{j=1}^{N} H_0^1(k_1\sqrt{(x-x_j)^2+(z-b)^2})(DU_b)_j, \quad (33)$$

$$\begin{aligned} H^2(x,z) = &-\frac{i}{2N\sqrt{k_2^2}}\mathrm{e}^{i\sqrt{k_2^2}|z|}\sum_{j=1}^{N}(DU_0)_j + \frac{i}{2N\sqrt{k_2^2}} \\ &\times\mathrm{e}^{i\sqrt{k_2^2}|z-b|}\sum_{j=1}^{N}(DU_b)_j - \frac{1}{2N}\mathrm{e}^{i\sqrt{k_2^2}|z|}\sum_{j=1}^{N}(U_0)_j + \frac{1}{2N} \\ &\times\mathrm{e}^{i\sqrt{k_2^2}|z-b|}\sum_{j=1}^{N}(U_b)_j - \frac{i}{N}\sum_{m=1}^{\infty}\frac{1}{\gamma_1}\cos\frac{m\pi(x+a)}{2a}\mathrm{e}^{i\gamma_1|z|} \\ &\times\sum_{j=1}^{N}\cos\frac{m\pi(x_j+a)}{2a}(DU_0)_j - \frac{1}{N}\sum_{m=1}^{\infty}\cos\frac{m\pi(x+a)}{2a} \\ &\times\mathrm{e}^{i\gamma_1|z|}\sum_{j=1}^{N}\cos\frac{m\pi(x_j+a)}{2a}(U_0)_j + \frac{i}{N}\sum_{m=1}^{\infty}\frac{1}{\gamma_1}\mathrm{e}^{i\gamma_1|z-b|} \\ &\times\cos\frac{m\pi(x+a)}{2a}\sum_{j=1}^{N}\cos\frac{m\pi(x_j+a)}{2a}(DU_b)_j \\ &+\frac{1}{N}\sum_{m=1}^{\infty}\cos(m\pi\frac{x+a}{2a})\mathrm{e}^{i\gamma_1|z-b|} \\ &\times\sum_{j=1}^{N}\cos\frac{m\pi(x_j+a)}{2a}(U_b)_j, \end{aligned} \quad (34)$$

$$H^3(x,z) = i\epsilon_3\sum_{j=1}^{N}\frac{a}{N\epsilon_2}(D\vec{U}_0)_j \times H_0^{(1)}\left[k_3\sqrt{(x-x_j)^2+z^2}\right], \quad (35)$$

where $x_j = 2a(j-1/2)/N - a$, $j = 1,2,\ldots,N$; $N > 2a/z$; $H_0^{(1)}(X)$ is the Hankel function; $\vec{H}^i = H^i \cdot \vec{e}_y$, $i = 1,2,3$. The coefficients $(D\vec{U}_0)_j$ are found by solving numerically the four coupled integral equations (28-31). The electric field of the wave is found from the scalar field $U(x,z)$ using Eqs. 6-8. For instance, the electrical field in the region III is given by

$$E_x^3(x,z,t) = -\frac{a}{N}\frac{(\epsilon_3)^{1/2}}{\epsilon_2}\sum_{j=1}^{N}\frac{z}{((x-x_j)^2+z^2)^{1/2}}H_1^{(1)}[k_3((x-x_j)^2+z^2)^{1/2}] \times(D\vec{U}_0)_j\exp(-i\omega t), \quad (36)$$

$$E_y^3(x,z,t)=0, \tag{37}$$

$$E_z^3(x,z,t)=\frac{a}{N}\frac{(\epsilon_3)^{1/2}}{\epsilon_2}\sum_{j=1}^{N}\frac{x-x_j}{((x-x_j)^2+z^2)^{1/2}}H_1^{(1)}[k_3((x-x_j)^2+z^2)^{1/2}] \times(D\vec{U_0})_j\exp(-i\omega t). \tag{38}$$

For more details on the model and the numerical solution of the Neerhoff and Mur coupled integral equations, see the references [3, 5].

Let us now consider the scattering of an ultra-short pulse (wave packet). The magnetic field of the incident pulse is assumed to be Gaussian-shaped in time and both polarized and constant in the y direction:

$$\vec{H}(x,y,z,t)=U(x,z)\exp[-2\ln(2)(t/\tau)^2]\exp(-i\omega_0 t)\vec{e}_y, \tag{39}$$

where τ is the pulse duration and $\omega_0=2\pi c/\lambda_0$ is the central frequency. The pulse can be composed in the wave-packet form of a Fourier time expansion (for example, see ref. [58, 59]):

$$\vec{H}(x,y,z,t)=\int_{-\infty}^{\infty}\vec{H}(x,z,\omega)\exp(-i\omega t)d\omega. \tag{40}$$

The electric and magnetic fields of the diffracted pulse are found by using the expressions (5-38) for each Fourier's ω-component of the wave-packet (40). This algorithm is implemented numerically by using the discrete fast Fourier transform (FFT) instead of the integral (40). The spectral interval $[\omega_{min},\omega_{max}]$ and the sampling points ω_i are optimized by matching the FFT result to the original function (39).

The above presented model deals with the incident waves having TM polarization. This polarization is considered for the following reasons. According to the theory of waveguides, the vector wave equations for this polarization can be reduced to one scalar equation describing the magnetic field H of TM modes. The electric component E of these modes is found using the field H and Maxwell's equations. The reduction simplifies the diffraction problem to one involving a single scalar field in only two dimensions. The TM scalar equation for the component H is decoupled from the similar scalar equation describing the field E of TE (transverse electric) modes. Hence, the formalism works analogously for TE polarization exchanging the E and H fields.

4. Non-resonant Localization of Continuous Waves and Femtosecond Pulses

In Section 4., we consider the physics of non-resonant transmission and diffraction of continuous waves and femtosecond pulses by a subwavelength nm-size metal slit. The physics is important for understanding the resonant near-field localization of light considered in Section 5.. In the subsection (4.1.)), we investigate the transverse broadening of a continuous wave (a Fourier's ω-component of a wave-packet) and its connection with the spatial

uncertainty $\Delta\mathbf{x} \sim 1/\Delta\mathbf{k}$. The limitations on the transverse dimension and duration of a wave-packet, at a given distance from a subwavelength aperture, imposed by the spatio-temporal uncertainty (Eq. 4) are demonstrated in the subsection (4.2.). The limitations are considered in the context of the spatio-temporal resolution capabilities of subwavelength-beam optical devices, such as near-field scanning microscopes and spectroscopes. The material used in Section 4. is reprinted from Physics Letters A, Vol. 319, S.V. Kukhlevsky, M. Mechler, L. Csapo, K. Janssens 439447, Copyright (2003), with permission from Elsevier, and from Journal of Optics A: Pure and Applied Optics, Vol. 5, 256-262, Copyright (2003), with permission from Institute of Physics Publishing.

4.1. Non-resonant Near-Field Localization of Continuous Waves

The scattering of light waves by a subwavelength slit is a process that depends on the wavelength $\lambda = 2\pi c/\omega$ (see, Eqs. 5-38). Owing to the dispersion, the amplitude of a time-harmonic continuous plane wave (a FFT ω-component of a wave-packet) does change under propagation through the slit. At the appropriate conditions, this leads to attenuation or amplification of the wave intensity. The dispersion of a time-harmonic continuous wave is usually described by the normalized transmission coefficient $T_{cw}(\lambda)$ for the energy flux ($\vec{S} = (c/8\pi)Re(\vec{E}\times\vec{H}^*)$, in CGS units). The coefficient is calculated by integrating the normalized energy flux S_z/S_z^i over the slit value [4, 5]:

$$T_{cw} = -\frac{\sqrt{\epsilon_1}}{4a\cos\theta}\int_{-a}^{a}\lim_{z\to 0^-}[(E_x H_y^* + E_x^* H_y)]dx, \tag{41}$$

where S_z^i and S_z are the energy fluxes (intensities) of the incident and transmitted waves, respectively.

The first objective of our analysis is to check the consistency of the results by comparing the transmission coefficient $T_{cw} = T_{cw}(\lambda, a, b)$ calculated in the studies [4, 5] with those obtained by our computations. We have computed the coefficient for different slit widths $2a$ and a variety of film thicknesses b. As an example, Fig. 4 shows the transmission coefficient T_{cw} as a function of screen thickness b for the different slit widths $2\,a$, at the fixed wavelength λ=500 nm.

We notice that the transmission resonances of $b = \lambda/2$ periodicity predicted in Refs. [4, 5] are reproduced. The resonance positions and the peak heights ($T_{cw}\approx\lambda/2\pi a$) at the resonances are also in agreement with the results [4, 5]. The dispersion function $T_{cw} = T_{cw}(\lambda)$ computed for a continues wave passing through the 50-nm slit (see, Fig. 5) also demonstrates the possibility of the resonant enhancement of the wave intensity.

Analysis of Figs. 4 and 5 shows that the non-resonant transmission of a continuous wave can be achieved by using a slit in a thin metal film $b << \lambda$.

Figure 6 shows the spatial distribution of energy flux S_z of a continuous wave localized by a thin slit in the near-field diffraction zone under the non-resonant regime of light transmission.

The spatial broadening of the continuous wave is in agreement with the spatial uncertainty relation discussed in Section 2.. Notice, in this connection, that in the near-field scanning optical microscopy, a subwavelength aperture illuminated by a continuous wave is used as a near-field light source providing the ultra-high optical resolution in space (for example, see Refs. [6–9]). The spatial broadening of the subwavelength continuous beam

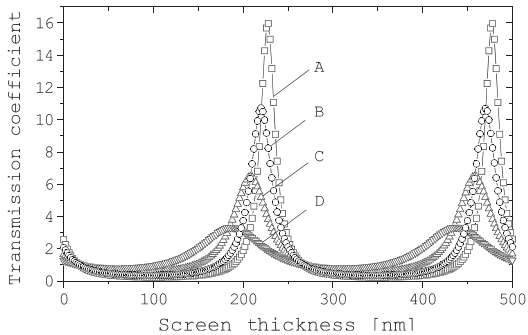


Figure 4. The transmission coefficient T_{cw} for a time-harmonic continuous wave with $\lambda = 500$ nm as a function of screen thickness b for the different aperture widths $2a$: A - 10 nm, B - 15 nm, C - 25 nm and D - 50 nm.

(Fig. 6) indicates a limitation of the optical resolution achievable with such a beam at a given distance from the slit.

4.2. Non-resonant Near-Field Localization of FS Pulses

The amplitude of a FFT ω-component of the wave-packet transmitted through the aperture depends on the wavelength (see, Sec. 4.1.). Due to the dispersion $T_{cw} = T_{cw}(\omega)$, the Fourier spectra of the wave-packet changes under the propagation through the aperture leading to modification of the temporal and spatial characteristics of the wave-packet. From a theoretical point of view, the most important questions for the non-resonant transmission are the degree of collimating, the duration of an ultra-short pulse past the aperture and the rate of spatial and temporal broadening of the pulse farther from the aperture.

To address the aforementioned questions, we have computed the near-field amplitude distributions of sub-fs and fs pulses passed non-resonantly through the nm-size slit in the thin metal film. It is desirable to present the computed data in the form of energy flux $\vec{S} = (c/8\pi)Re(\vec{E}\times\vec{H}^*)$, since it is the intensity distribution that determines the spatial resolution of the near-field scanning optical microscopy. The temporal resolution of NSOM, however, is better described by the field amplitudes $\vec{E}$ and $\vec{H}$. Indeed, directly after interaction of the pulse with a sample (the interaction Hamiltonian $V_{p-s}(x, y, z, t) = V_{p-s}(\vec{P}, \vec{E}, \vec{H})$, where $\vec{P} = \vec{P}(x, y, z, t)$ is the sample polarization operator), the pulse temporal spreading may progressively appear, but the physical time information contained in the sample response is not distorted or lost. Hence, the distributions $\vec{E}(x, y, z, t)$ and $\vec{H}(x, y, z, t)$ of the diffracted pulse are the quantities that determine the temporal resolution of NSOM. It should be noted, in this connection, that the NSOM using a continuous wave provides nm-scale spatial resolution [6–9]. Nowadays, the high-harmonic generation produces near 0.5-fs (500 as) wave-packets with $\sim$50-nm central wavelengths [65,66]. A method of generation of sub-as pulses has been also suggested [22]. The perspective is the achievement of nm-scale spatial resolution together with sub-fs or fs temporal resolution.

We have analyzed the distributions of the energy flux $\vec{S}$=$(S_x, 0, S_z)$, the electric field

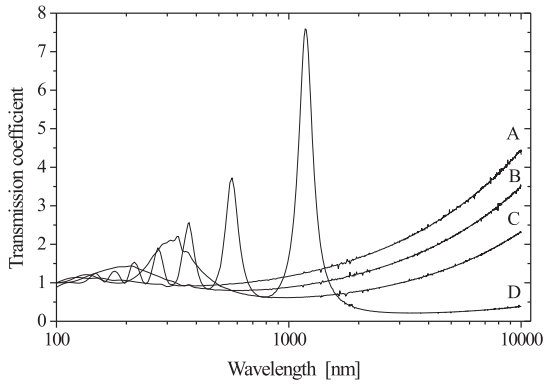


Figure 5. The transmission coefficient T_{cw} for a time-harmonic continuous wave passing a 50-nm slit as a function of the wavelength λ for the different screen thickness b: A - 25 nm, B - 50 nm, C - 100 nm and D - 500nm.

$\vec{E}$=$(E_x, 0, E_z)$, and the magnetic field $\vec{H}$=$(0, H_y, 0)$ past the slit and further from the film. The distributions have been computed for the experimental conditions that are relevant to the spatial and temporal resolution of NSOM. We first studied the scattering of the sub-fs pulses at the different values of the duration τ, central wavelength λ_0, width $2a$ and thickness b. As an example, the distributions S_z, S_x (see, Fig. 7), $|\vec{E_x}|$, $|\vec{E_z}|$ (see, Fig. 8) and $|\vec{H_y}|$ (see, Fig. 9) are shown at the three distances z from the screen: at the edge of the film $(z = -0.1a)$, in the near-field $(z = -a)$ and far-field $(z = -10a)$ zones of the diffraction.

The figures show the results for the following parameters: τ = 0.5 fs, λ_0 = 500 nm, $2a = 50$ nm, $b = 25$ nm and $U(x, z = 0) = 1$. The coordinate x and time t are in meter and second units, respectively. Notice, that the value $2a = 50$ nm is the minimum aperture size, which is generally accepted for the practical near-field ($z \approx -a$) scanning microscopy in the visible spectral region ($\lambda \approx 500$nm) [5].

Analysis of the distribution $S_z(x, z)$ (see, Fig. 7a) shows that the flux is collimated to exactly the aperture width at the edge of the film ($z = -0.1a$). The spatial resolution of the NSOM, in which the sample lies in a plane parallel to the screen containing the aperture (conventional geometry of NSOM), is determined by the longitudinal energy flux S_z. Hence, the ultra-short pulse passing through a subwavelength aperture can be used to provide a nm-size image, super-resolution in space. The distribution $S_z(x, z)$ is characterized by formation of maxima at the rims of the aperture (Fig. 7a). It is interesting that a similar effect exists in the case of the NSOM using a continuous wave [67]. Analysis of Figs. 7b,c indicates practical difficulties inherent in the NSOM using sub-fs pulses because of the spatial divergence of the wave-packet further from the screen. At the distance $z = -a$ the spatial resolution is approximately equal to the geometrical projection of the aperture (Fig. 7b). This demonstrates the practical possibility of the nanometer ($2a = 50$ nm) spatial resolution at $z = -a$. Notice that the value $2a$ was considered as the practical limit of the

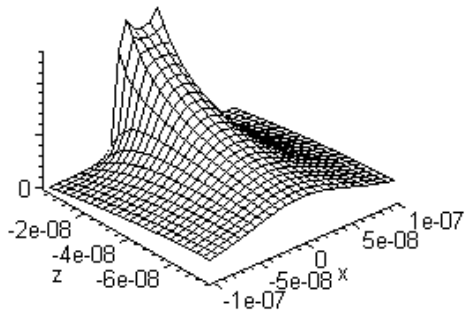


Figure 6. The energy flux distributions S_z (a, b, c) of a continuous wave diffracted by the slit of width $2a$ = 50 nm in the perfectly conducting screen of thickness b = 25 nm. The wavelength is λ=500 nm, the flux is in arbitrary units. The values x and z are in meters.

spatial resolution also in the case of the NSOM using a continuous wave [5]. The results for the near-field diffraction (Figs. 7a,b) are in agreement with the conclusion [68, 69] that the spatial resolution of NSOM using ultra-short pulses is defined by the aperture size and independent of the wavelength. It should be noted, in this connection, that the width Δx of the ω-Fourier component of the wave packet decreases slowly with increasing the frequency ω. This should lead to some increasing of the spatial resolution of the NSOM with decreasing the wavelength. Indeed, the highest spatial resolution is achieved for the spectral component having the highest frequency ω_{HG}. The value $\omega_{HG} = \omega_0 + \Delta\omega/2$ depends on the central frequency ω_0 and spectral width $\Delta\omega$ of the wave packet. The width $\Delta\omega \approx 1/\tau$ increases with decreasing the duration τ. Therefore, strictly speaking, the spatial resolution of the NSOM even in the near-field diffraction zone is the wavelength dependent. In agreement with the spatio-temporal uncertainty (Eq. 4), the spatial resolution increases with increasing the pulse duration τ and the central frequency ω_0 of the wave-packet. This effect is stronger in the far-field $(z = -10a)$ zone, where the pulse experiences very strong transverse broadening and shape deformation (Fig. 7c).

The energy flux S_x in the transverse direction $\vec{e}_x$ does not affect the spatial resolution of the conventional NSOM. Nevertheless, in some configurations of NSOM, the sample lies in a plane which is not parallel to the screen containing the aperture. In such a case, the NSOM is sensitive also to the transverse flux S_x. In order to better understand the spatial resolution capabilities of the NSOM that uses the flux S_x of a sub-fs pulse for illumination of the sample, we computed the distribution $S_x(x, z)$ (Fig. 7d,e,f). Figure 7d shows that the flux $S_x(x, z)$ is not collimated to the aperture width even at the edge of the screen ($z = -0.1a$). We also notice that at the distances $z = -0.1a, -a$ the transverse dimensions of the flux $S_x(x, z)$ (Fig. 7d,e) are much bigger of the dimensions of the flux $S_z(x, z)$ (Fig. 7a,b). The results presented in Figs. 7d,e,f indicate that the use of the flux $S_x(x, z)$ instead of $S_z(x, z)$ for the illumination of the sample decreases the spatial resolution capabilities of the NSOM at the edge of the screen $(z = -0.1a)$, in the near-field $(z = -a)$ and far-field $(z = -10a)$ zones of the diffraction.

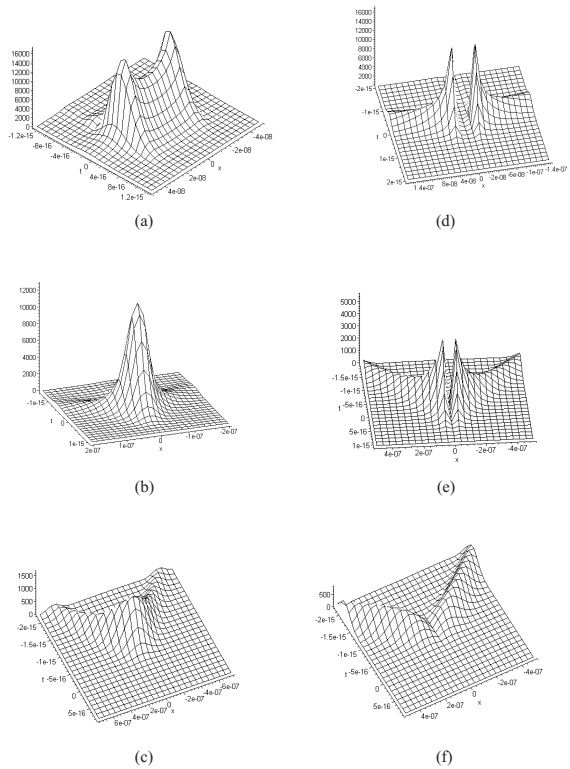


Figure 7. The energy flux distributions S_z (a, b, c) and S_x (d, e, f) of the 0.5-fs pulse diffracted by the slit of width $2a$ = 50 nm in the perfectly conducting film of thickness b = 25 nm, at the three distances z from the screen: (a, d) - $(z = -0.1a)$, (b, e) - $(z = -a)$ and (c,d) - $(z = -10a)$. Here, Δt is the temporal delay of the pulse with respect to the point $t = 0$.

The distributions of the electric $\vec{E}$=$(E_x, 0, E_z)$ and magnetic $\vec{H}$=$(0, H_y, 0)$ fields, which determine the temporal resolution of the NSOM using a sub-femtosecond (0.5 fs) incident pulse, are shown respectively in Figs. 8 and 9. The amplitude distribution $|\vec{E_x}|$ for the NSOM using a 0.75-fs pulse is shown in Fig. 10 for the comparison.

In agreement with the temporal uncertainty $\Delta t \sim 1/\Delta\omega$, the Fourier spectrum of the transmitted pulse experiences considerable narrowing leading to the increase of the pulse duration. In Figs. 8a,b,d,e, 9a,b and 10a,b we notice that the duration of the field components E_x, E_z and H_y of the diffracted pulses at the film edge (z=-0.1a) and in the near-field $(z = -a)$ zone of the diffraction are in a few times longer compared to the duration of the incident pulse. This phenomenon imposes a fundamental limit on the simultaneous spatial and temporal resolutions of NSOM in the visible spectral region ($\lambda_0 \sim 500$ nm). It could be noted that the decrease of the temporal resolution of NSOM caused by the increase of the pulse duration can be reduced by decreasing the central wavelength (short-wavelength spectral region) of the incident wave-packet [70]. The results presented in Figs. 8 and 9 also

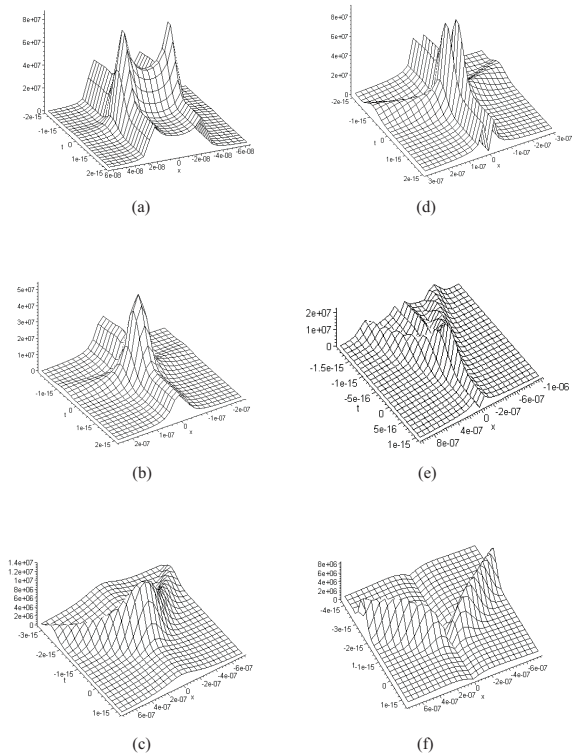


Figure 8. The electric field distributions $|\vec{E_x}|$ (a, b, c) and $|\vec{E_z}|$ (d, e, f) of the 0.5-fs pulse diffracted by the slit of width $2a$ = 50 nm in the perfectly conducting film of thickness b = 25 nm, at the three distances z from the screen: (a, d) - $(z = -0.1a)$, (b, e) - $(z = -a)$ and (c,d) - $(z = -10a)$. Here, Δt is the temporal delay of the pulse with respect to the point $t = 0$.

indicate that the use of the electric field component $E_x(x, z, t)$ for the illumination of the sample provides us with higher temporal resolution compared to that achieved by the use of the field components $E_z(x, z, t)$ and $H_y(x, z, t)$ at the edge of the screen $(z = -0.1a)$, in the near-field $(z = -a)$ and far-field $(z = -10a)$ zones of the diffraction.

The numerical analysis has been performed also for fs pulses ($\tau \geq 1$ fs). The filed distributions of the diffracted wave-packet was calculated for different values of the incident-pulse duration τ, central wavelength $\lambda_0 = 2\pi c/\omega_0$, slit width $2a$ and screen sickness $2b$. As an example, Fig. 11 shows the distribution $|\vec{E_x}|$ of the 2-fs wave-packet diffracted by the slit. The field was computed for the following experimental conditions: $2a = 50$ nm, $b = 25$ nm, $\lambda_0 = 500$ nm, τ = 2 fs. The distribution $|\vec{E_x}|$ is shown at the three distances z from the screen: at the edge of the screen $(z = -0.1a)$, in the near-field $(z = -a)$ and far-field $(z = -10a)$ zones of the diffraction. Analysis of Fig. 11a shows that the pulse is collimated to exactly the aperture width $2a = 0.1\lambda_0$ at the edge $(z = -0.1a)$ of the screen. Hence, the basic concept of NSOM remains valid also for the pulse duration $\tau \geq 2$ fs: the fs

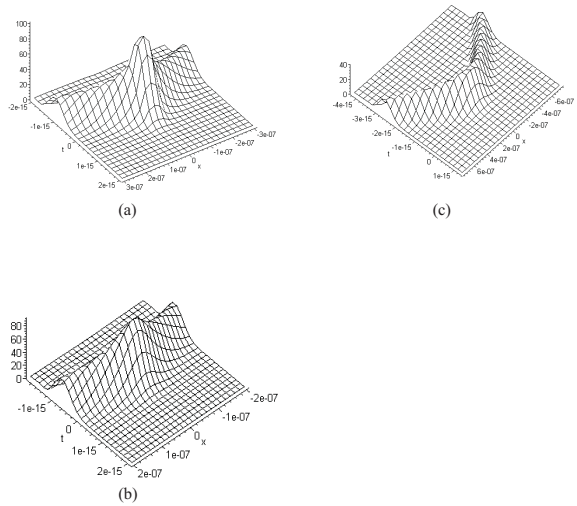


Figure 9. The magnetic field distribution $|\vec{H}_y|$ of the 0.5-fs pulse diffracted by the slit of width $2a$ = 50 nm in the perfectly conducting film of thickness b = 25 nm, at the three distances z from the screen: (a, d) - $(z = -0.1a)$, (b, e) - $(z = -a)$ and (c,d) - $(z = -10a)$. Here, Δt is the temporal delay of the pulse with respect to the point $t = 0$.

wave-packets passing through a nm-size aperture can be used to provide a nm-scale image, super-resolution in space. The aperture provides a subwavelength image ($2a << \lambda_0$) in respect to the central wavelength of the wave-packet. The duration τ of the diffracted pulse at the screen edge is practically the same as the incident pulse ($\tau \approx$ 2fs). Thus, the temporal resolution associated with the 2-fs duration of the incident pulse is practically unchanged past the aperture. This demonstrates the principal feasibility of nm-scale resolution in space together with fs-scale resolution in time. The possibilities and limitations of the practical NSOM using fs pulses are demonstrated by the results presented in Figs. 11b,c. The spatial resolution of the practical ($z \approx -a$) NSOM performing with 2-fs pulses is approximately equal to the geometrical projection of the aperture at the distance of half the aperture width (Fig. 11b). At the distance $z \approx -a$, the duration of the diffracted pulse is practically the same as the incident pulse. This demonstrates the practical feasibility of the 50-nm resolution in space together with the 2-fs resolution in time. Notice, that the 50-nm spatial resolution is the practical limit also for the NSOM based on the use of a continuous wave [5].

Analysis of Figs. 8, 9, 10 and 11 show that the spatial resolution of the practical NSOM ($z \approx -a$) decreases with shortening the incident pulse. The spatial resolution of the NSOM performing with 2-fs pulses (Fig. 11b) is approximately two times higher than the resolution achieving with 0.75-fs pulses (see, Fig. 10b). Notice, in this connection, that the spatial resolution of the NSOM using ps pulses is defined only by the aperture size [68]. At the distance $z \approx -a$, the temporal broadening of sub-fs pulses (Fig. 11b) imposes a limit on the temporal resolution. In the far-field region $(z = -10a)$, the 2-fs and sub-fs pulses experience the additional temporal and spatial broadenings and the distortion of the pulse

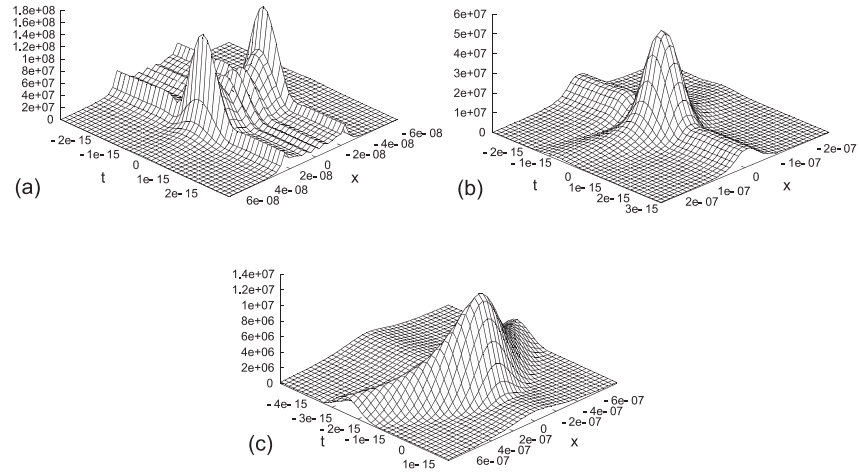


Figure 10. The amplitude distribution $|\vec{E_x}|$ at the three distances z from the screen: (a) - $(z = -0.1a)$, (b) - $(z = -a)$ and (c) - $(z = -10a)$, in the arbitrary units. Here, the incident-pulse duration $\tau = 0.75$ fs, the aperture width $2a = 50$ nm and thickness $b = 25$ nm, the packet central wavelength $\lambda_0 = 500$ nm and the amplitude $U(x, z = b) = 1$. The time t and coordinate x are in the second and meter units, respectively.

shape (Figs. 7c,f, 8c,f, 9, 10c and 11c). This leads to the further decrease of both the spatial and temporal resolutions associated with the pulse width and duration.

The profile and duration of the fs and sub-fs pulses are modified by the slit transmission function $T_{cw} = T_{cw}(\lambda)$. At the appropriate conditions, however, the wave-packet propagates through the slit practically without spatial and temporal broadening (Figs. 11a,b. Figure 5 helps us to understand the conditions of the broadening-free propagation. According to the data presented in Fig. 5, the transmission coefficient $T_{cw}(\lambda)$ of the aperture in the thin ($b < 50$nm) screen is approximately constant in the spectral band $\Delta\lambda \approx [300nm, 1000nm]$. The width of the Fourier spectra of the wave-packet having the fs-scale ($\tau > 2$ fs) duration and the central wavelength ($\lambda_0 \sim 500$ nm) does not exceed the dispersion-free interval $\Delta\lambda$. In addition, the diffraction of the fs-scale pulses is small immediately ($z < -a$) past the screen. Therefore, in the near-field zone the slit in the thin metal film provides the broadening-free diffraction of the fs-scale pulses. Thin slit transmits the light without increasing its intensity. The increase of the film thickness and/or decreasing the input pulse duration leads to the distortion of the diffracted pulses and decreasing the resolution of the NSOM-pulse system.

The above-presented analysis indicated the limitation of the non-resonant spatio-temporal localization of an ultra-short pulse imposed by the spatio-temporal uncertainty of a wave-packet. The pulse can be localized in the near-field diffraction zone at the nm and fs scales by the appropriate adjusting of the slit and pulse parameters.

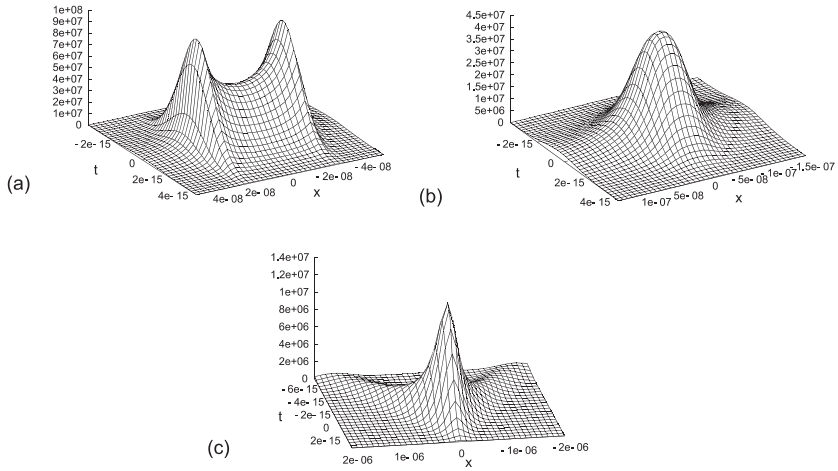


Figure 11. The amplitude distribution $|\vec{E_x}|$ at the three distances z from the screen: (a) - $(z = -0.1a)$, (b) - $(z = -a)$ and (c) - $(z = -10a)$, in the arbitrary units. Here, the incident-pulse duration $\tau = 2$fs, the aperture width $2a = 50$nm and thickness $b = 25$nm, the packet central wavelength $\lambda_0 = 500$nm and the amplitude $U(x, z = b) = 1$. The time t and coordinate x are in the second and meter units, respectively.

4.3. Limitations of the Model

The numerical analysis of the non-resonant scattering presented in Sections 4.1. and 4.2. shows that the shape deformation and spatio-temporal broadening of the pulse are the main effects that limit the spatial and temporal capabilities of the near-field optics, in which a fs pulse localized by the subwavelength slit is used to provide nm spatial resolution together with fs temporal resolution. The accuracy of the above-presented analysis must be considered before the results are used to apply to a particular experimental situation in NSOM.

The slit model can be directly used for a 2-D NSOM-pulse system. The advantages and limitations of the two-dimensional-aperture model of a NSOM based on the non-resonance transmission of a continuous wave and a light pulse by a slit have been discussed in detail in the studies [5, 67] and [68, 69], respectively. We consider here only the most important questions associated with the limitations of our model. The simultaneous super resolutions in space and time of the slit-pulse system is a consequence of the assumption of the screen perfect conductivity. In the reality, the metal films which can be used to form a thin screen have finite conductivity. As a general criterion, the perfect conductivity assumption should remain valid as far as the slit width and the screen thickness exceed the extinction length for the Fourier components of the wavepacket within the metal. The light intensity decays in the screen at a rate of $I_s = I_0 \exp(-b/\delta)$, where $\delta = \delta(\lambda)$ is the extinction length in the screen. The aluminum has the largest opacity δ<11 nm in the spectral region $[100nm, 1000nm]$ [71]. The extinction length increases from 11 to 220 nm with decreasing the wavelength from 100 to 50 nm. Hence, the perfect conductivity is a very good approximation in a situation involving a relatively thick ($b = 25 - 50$nm) aluminum screen

and a 2-fs wave-packet with the central wavelength $\lambda_0 = 500$ nm and the spectral band $\Delta\lambda \approx [300nm, 1000nm]$. However, in the case of the more thin $b < 25$nm screens, the shorter ($\tau <$2 fs) pulses and the shorter central wavelengths ($\lambda_0 <$ 500 nm), the metal films are not completely opaque. This would decrease the spatial and temporal resolutions of a 2-D NSOM-pulse system due to passage of the light in the region away from the slit. Notice, that the model is also limited to the situation where the magnetic field of the incident pulse is aligned along the slit. For other polarizations, the transmitted field distribution may be different, but the degree of collimating should be about the same.

We have studied the slit model, while in reality a circular aperture is usually used in NSOM ([6–9]). The transmission properties of a round aperture are different from the slit ones. The slit in a thick screen acts as a waveguide which can support a propagation mode, whereas a subwavelength circular aperture in a thick screen can support only evanescent modes. Furthermore, there is an additional reduction in transmission due to reflections at the screen, which is different for the slit and round apertures. Despite these differences, the slit model was chosen here for a number of reasons. The problem of a subwavelength diffraction of a continuous wave by a circular aperture is usually solved using a series expansion for the fields within the aperture, Fourier optics or the method of moments (see, for example the studies [2, 72–74] and reference therein). These methods yield results of finite accuracy, even when sophisticated numerical methods are used. The use of ultrashort pulse in a NSOM makes the numerical solution of the circular aperture diffraction problem even less accurate. In contrast, the slit model provides accuracy of the results limited only by the speed and memory of the computer used. Nevertheless, the results of circular aperture diffraction [73] can be used to obtain an approximate estimate of the minimum aperture size which can be used in a practical NSOM-pulse system. The intensity decay of the least attenuated mode (TE_{11} waveguide mode) within a subwavelength round aperture is given by $I_w = I_0 \exp(-3.682b/r)$, where r is the aperture radius. When the attenuation due to the waveguide effect exceeds the attenuation in the metal ($I_s(\lambda) > I_w$), the contrast between the aperture and surrounding screen becomes insufficient for the super resolution. Hence the minimum aperture diameter is given by $2r \approx 7.364\delta$. In the case of a 2-fs wave-packet with the central wavelength $\lambda_0 = 500$ nm and the spectral band $\Delta\lambda \approx [300nm, 1000nm]$ passing through an aluminum aperture, this yields a diameter of about 50 nm, which explains our choice of a 50-nm slit for NSOM imaging. The width of the Fourier spectra of the 2-fs wavepacket does not exceed the dispersion-free spectral interval of subwavelength diffraction ($I_w \approx$const for $\lambda > 2r$). The diffraction is relatively small immediately ($z < -r$) past the screen. Therefore, in the near-field zone ($z < -r$), the 50-nm circular aperture should provide the diffraction of the 2-fs pulse free of the temporal broadening. These result is in a good agreement with the 50-nm resolution in space and the 2-fs resolution in time derived here for a practical 2-D NSOM-pulse system. It can be noted that due to the dispersion-free property ($I_w \approx$const for $\lambda > 2r$) of a circular subwavelength aperture in a perfect-conductivity screen the minimum screen thickness b does not depend on the pulse spectra, in contrast to the wavelength-dependent diffraction by a subwavelength slit. In the case of the thin ($b < 25$ nm) screens, however, the metal films are not completely opaque [71]. This would decrease the spatial and temporal resolutions of a NSOM-pulse system using the thin ($b < 25$ nm) screens due to passage of the light in the region away from the circular aperture. It should be also noted that in the case of very short ($\tau <$1 fs)

pulses, the wavepacket components with $\lambda << 2r$ propagate through a circular aperture practically without attenuation, while the amplitude of the long-wavelength ($\lambda >> 2r$) components exponentially decreases. Due to this effect, the Fourier spectra narrows that leads to the increase of the transmitted pulse duration. The similar effect exists also in the case of the diffraction by a slit in the thin ($b < 25$ nm) perfect-conductivity screen. In this case, however, the Fourier spectra narrows due to the amplification of the long-wavelength ($\lambda >> 2r$) components of a sub-fs wavepacket (see, curve (A) of Fig. 3). It can be mentioned that for the case $\lambda << 2a$, the the super-resolution imaging was theoretically studied in Refs. [70, 75].

The presence of a microscopic sample (a molecule, for example) placed in strong interaction with a NSOM aperture modifies the boundary conditions. In the case of the strong interaction, which takes place in the region $z < -0.1a$, the application of the standard boundary conditions practically impossible. Probably, in this region the problem can be overcome by combination of a microscopic and macroscopic descriptions of NSOM [76], where the response function accounting for the modification of the quantum mechanical behavior of the molecule is derived self-consistently through the solution of Dyson's equation.

5. Resonant Enhancement and Localization of Continuous Waves and FS Pulses

In Section 5., we study the resonant scattering of continuous waves and fs pulses by a subwavelength nm-size metal slit. The subsection (5.1.) considers interplay between the resonant enhancement and spatial localization of a continuous wave. The section (5.2.)) describes correlation between the resonant enhancement and the spatio-temporal localization of a wave-packet, which imposes a limit on the transverse dimension and duration of the resonantly enhanced fs pulse and consequently limits the spatio-temporal resolution capabilities of subwavelength-beam optical devices using the resonance effect.

5.1. Resonant Enhancement and Near-Field Localization of Continuous Waves

In the last decades, a number of studies have been devoted to the resonantly enhanced scattering of light by nano-apertures in metal films, namely a single aperture, a grating of apertures and an aperture surrounded by grooves. Since the recent paper of Ebbesen and colleagues [14] on the resonantly enhanced transmission of light observed for a 2D array of subwavelength holes in metal films, the resonant phenomenon is intensively discussed in the literature [10–76]. Such a kind of light scattering is usually called a Wood's anomaly. In the early researches, Hessel and Oliner showed that the resonances come from coupling between inhomogeneous diffraction orders and eigenmodes of the grating [11]. Neviere and co-workers discovered two other possible origins of the resonances [12, 13]. One appears when the surface plasmons of a metallic grating are excited. The other occurs when a metallic grating is covered by a dielectric layer, and corresponds to guided modes resonances in the dielectric film. The role of resonant Wood's anomalies and Fano's profiles [10] in the

resonant transmission were explained in the study [36].

The phenomena involved in propagation through hole arrays are different from those connected with slit arrays. In a slit waveguide there is always a propagating mode inside the channel, while in a hole waveguide all modes are evanescent for hole diameters smaller than approximately a wavelength. In the case of slit apertures in a thick metal film, the transmission exhibits enhancement due to a pure geometrical reason, the resonant excitation of propagating modes inside the slit waveguide [16, 17, 24, 37, 38]. At the resonant wavelengths, the transmitted field increases via the strong coupling of an incident wave with the waveguide modes giving a Fabry-Perot-like behavior [17, 18, 37, 59]. In the case of films, whose thickness are too small to support the intra-cavity resonance, the extraordinary transmission can be caused by another mechanism, the generation of resonant surface plasmon polaritons and coupling of them into radiation [14,16,19,38,40]. Both physical mechanisms play important roles in the extraordinary transmission through arrays of two-side-open slits (transmission gratings) and the resonant reflection by arrays of one-side-open slits (reflection gratings). A model of trapped (waveguide) modes has been recently used to show that an array of two-side-open slits can operate like a reflection grating totally reflecting TE-polarized light [41]. The surface plasmons and Rayleigh anomalies were involved in explanation of reflective properties of such a kind of gratings [29].

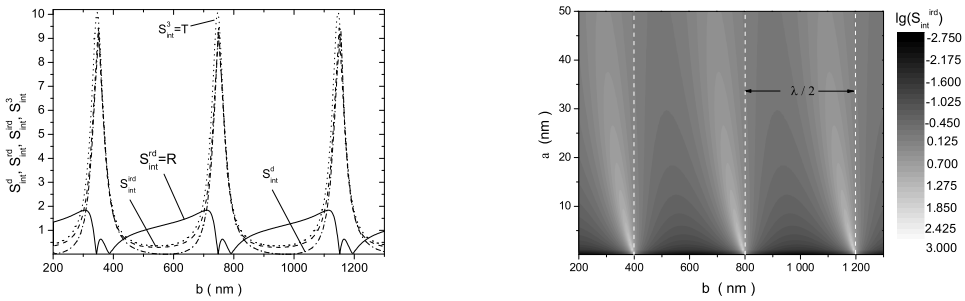
The studies [17, 18, 42, 59] have pointed out that the origin of anomalous scattering of light by a grating of slits (holes) can be better understood by clarifying the transmission and reflection properties of a single subwavelength slit. Along this direction, it was already demonstrated that the intensity of TM-polarized light resonantly transmitted through a single slit can be 10-10^3 times higher than the incident wave [3–5,59] and that the transmission coefficient versus wavelength possesses a Fabry-Perot-like behavior [17, 18, 59]. The high peak amplitudes (enhancement) demonstrated in in Refs. [3–5, 18, 59], however, are different from the low magnitudes (attenuation) predicted in the study [17]. Furthermore, the reflection properties of the slit have received no attention in the literature. The very recent study [41] only concerned the problem by regarding the total reflection of TE-polarized light by a grating of two-side-open slits to properties of the independent slit emitters.

In the present section, the forward and backward scattering of light by a two-side-open subwavelength slit is analyzed. To compare properties of the light reflection with the extraordinary transmission [17, 18, 59], we consider the scattering of TM-polarized light by a slit in a thick metallic film of perfect conductivity. From the latter metal property it follows that surface plasmons do not exist in the film. Such a metal can be described by the Drude model for which the plasmon frequency tends towards infinity. We show that the intensity of a wave localized in the near-field diffraction zone behind the slit is 10-10^3 times greater than the incident wave, but one order of magnitude smaller than the intra-cavity intensity. Interference of incident and resonantly backward-scattered light produces in the near-field diffraction zone, in front of the slit, a spatially localized wave whose intensity is also 10-10^3 times greater than the incident wave, but about ten times smaller than the intra-slit intensity. In addition, we show that the reflection coefficient versus wavelength possesses a Fabry-Perot-like dependence that is similar to the anomalous behavior of transmission reported in the study [Y. Takakura, Phys. Rev. Lett. **86**, 5601 (2001)]. The open slit totally reflects the light at the near-to-resonance wavelengths. The amplitude and phase of the resonant wave at the slit entrance and exit are different from that of a Fabry-Perot cavity. The material of

Let us start with the theoretical background for the scattering problem. It is well known that when a light wave is scattered by a subwavelength metallic object, a significant part of the incident light can be scattered forward or backward (reflected) whatever the object be reflecting or transparent. We will test whether a continuous wave can be resonantly enhanced and simultaneously localized in the near-field diffraction zone of a single two-side-open subwavelength metallic slit. To address this question, the energy flux in front of the slit and behind the slit is analyzed numerically for various regimes of the light scattering. In order to compare properties of the light scattering with that of the extraordinary (resonant) transmission [17, 18, 59], we consider the zeroth-order scattering of a time-harmonic wave of TM-polarized light by a slit in a perfectly conducting thick metal film placed in vacuum (Fig. 3). The energy flux $\vec{S}_I$ in front of the slit is compared with the fluxes $\vec{S}_{II}$ and $\vec{S}_{III}$ inside the slit and behind the slit, respectively. The amplitude and phase of the light wave at the slit entrance and exit are compared with that of a Fabry-Perot cavity. The electric $\vec{E}$ and magnetic $\vec{H}$ fields of the light are computed by using the traditional approach based on the Neerhoff and Mur solution of Maxwell's equations [3–5].

According to the model (Section 3), the electric $\vec{E}(x,z)$ and magnetic $\vec{H}(x,z)$ fields in front of the slit (region I), inside the slit (region II) and behind the slit (region III) are determined by the scalar fields $U_1(x,z)$, $U_2(x,z)$ and $U_3(x,z)$, respectively. The scalar fields are found by solving the Neerhoff and Mur integral equations. The magnetic field of the incident wave is assumed to be time harmonic and constant in the y direction: $\vec{H}(x,y,z,t) = U(x,z)\exp(-i\omega t)\vec{e}_y$. In front of the slit, the field is decomposed into $U_1(x,z) = U^i(x,z) + U^r(x,z) + U^d(x,z)$. The field $U^i(x,z)$ represents the incident field, which is assumed to be a plane wave of unit amplitude; $U^r(x,z)$ denotes the field that would be reflected if there were no slit in the film; $U^d(x,z)$ describes the backward diffracted field due to the presence of the slit. The time averaged Poynting vector (energy flux) $\vec{S}$ of the electromagnetic field is calculated (in CGS units) as $\vec{S} = (c/16\pi)(\vec{E}\times\vec{H}^* + \vec{E}^*\times\vec{H})$. The reflection coefficient $R = S^{rd}_{int}$ is given by the normalized flux $S^{rd}_n = S^{rd}/S^i$ integrated over the slit width $2a$ at the slit entrance ($z = b$), where S^{rd} is the z component of the backward scattered flux, and S^i is the incident flux along the z direction. The flux $S^{rd} = S^{rd}(U^r, U^d)$ is produced by the interference of the backward scattered fields $U^r(x,z)$ and $U^d(x,z)$. The transmission coefficient $T = S^3_{int}(b)$ is determined by the normalized flux $S^3_n = S^3/S^i$ integrated over the slit width at the slit exit ($z = 0$), where the flux z-component $S^3 = S^3(U^3)$ is produced by the forward scattered (transmitted) field $U^3(x,z)$. Notice that the definitions of the reflection R and transmission T coefficients are equivalent to the more convenient ones defined as the integrated reflected or transmitted flux divided by the integrated incident flux. In the following analysis, the reflection and transmission coefficients are compared to the fluxes S^d_{int} and S^{ird}_{int} obtained by integrating the normalized fluxes $S^d_n = S^d/S^i$ and $S^{ird}_n = S^{ird}/S^i$, respectively.

We analyzed the backward scattering of light for a wide range of scattering conditions determined by values of the wavelength λ, slit width $2a$ and film thickness b. As an example, the reflection coefficient $R = S^{rd}_{int}(b)$ as a function of the film thickness b computed for the wavelength $\lambda = 800$ nm and the slit width $2a = 25$ nm is shown in Fig. 12.



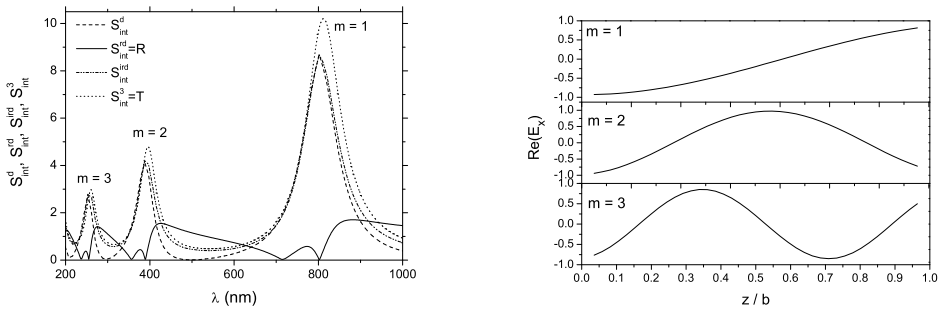
(a) (b)

Figure 12. (a) The reflection coefficient $R = S_{int}^{rd}(b)$, the transmission coefficient $T = S_{int}^{3}(b)$, and the integrated fluxes $S_{int}^{d}(b)$ and $S_{int}^{ird}(b)$ as a function of the film thickness b computed for the wavelength $\lambda = 800$ nm and the slit width $2a = 25$ nm. (b) The logarithm of the integrated flux $S_{int}^{ird}(a, b)$ as a function of the slit half-width a and film thickness b.

The transmission coefficient $T = S_{int}^{3}(b)$ and the integrated fluxes $S_{int}^{d}(b)$ and $S_{int}^{ird}(b)$ are presented in the figure for the comparison. We note the reflection resonances of $\lambda/2$ periodicity with the maxima $R_{max} \approx 2$. In agreement with the previous results [17, 18, 59], one can see also the transmission resonances having the same period, and the peak heights $T \approx 10$ ($T \approx \lambda/2\pi a$) at the resonances. It is worth to note the correlation between the positions of maxima and minima in the reflection and transmission. The resonance positions for the total reflection are somewhat left-shifted with respect to the transmission resonances. The maxima of the transmission coefficient correspond to reflection minima. In Fig. 12a, one can observe also many satellite peaks in reflection. For one broad minimum, it appears a local reflection maximum, which is characterized by a weak amplitude. The local maxima appear before 400, 800, and 1200 nm. The positions of the local maxima approximately correspond to the S_{int}^{d} maxima. To clarify a role of the fields U^i, U^r, U^d and U^3 in the resonant backward scattering, we compared the integrated flux $S_{int}^{ird}(U^i, U^r, U^d)$ with the fluxes $S_{int}^{d}(U^d)$ and $S_{int}^{3}(U^3) = T$. One can see from Fig. 12a that the flux S_{int}^{ird} produced in front of the slit by the interference of the incident field $U^i(x, z)$ and the backward scattered fields $U^r(x, z)$ and $U^d(x, z)$ is practically indistinguishable from that generated by the backward diffracted field U^d and forward scattered (transmitted) field U^3. The integrated flux $S_{int}^{ird}(a, b)$ as a function of the slit half-width a and film thickness b is shown in Fig. 12b. We notice that the widths and shifts of the resonances increase with increasing the value a. Analysis of Fig. 12a indicates that the difference between the integrated fluxes $S_{int}^{rd}(U^r, U^d) = R$ and $S_{int}^{3}(U^3) = T$ ($T \approx S_{int}^{d}(U^d)$) appears due to the interference of the backward diffracted field $U^d(x, z)$ and the reflected field $U^r(x, z)$.

The dispersion of the reflection coefficient $R(\lambda) = S_{int}^{rd}(\lambda)$ for the slit width $2a = 25$ nm and the screen thickness $b = 351$ nm is shown in Fig. 13a.

The integrated fluxes $S_{int}^{d}(\lambda)$, $S_{int}^{ird}(\lambda)$ and $S_{int}^{3}(\lambda) = T(\lambda)$ versus the wavelength are shown in the figure for the comparison. A very interesting behavior of the dispersion is that the coefficient R versus the wavelength λ possesses a Fabry-Perot-like dependence that is similar to the anomalous behavior of transmission $T(\lambda)$ reported in the studies [17,



(a) (b)

Figure 13. (a) The reflection coefficient $R = S^{rd}_{int}(\lambda)$, the transmission coefficient $T = S^{3}_{int}(\lambda)$, and the integrated fluxes $S^{d}_{int}(\lambda)$ and $S^{ird}_{int}(\lambda)$ versus the wavelength λ computed for the slit width $2a = 25$ nm and the screen thickness $b = 351$ nm. (b) The real part of the normalized electric field x-component $E_x(U_2) = E_x(x, z)$ versus the normalized distance z/b inside the slit cavity at $x = 0$, for the resonant wavelength $\lambda^1_r = 800$ nm, $\lambda^2_r = 389$ nm, and $\lambda^3_r = 255$ nm.

18, 37, 42, 59]. In agreement with the studies [3–5, 59], the height of the first (maximum) transmission peak is given by $T \approx \lambda^1_r/2\pi a$. The wavelengths corresponding to the resonant peaks $\lambda^m_r \approx 2b/m$ ($m = 1, 2, 3, \ldots$) are in accordance with the results [17]. The high peak amplitudes (enhancement), however, are different from the low magnitudes (attenuation) predicted in the study [17], but compare well with the experimental and theoretical results [3–5, 18, 59]. The difference is caused by the manner in which the Maxwell equations are solved. The study [17] uses a simplified approach based on the matching the cavity modes expansion of the light wave inside the slit with the plain waves expansion above and below the slit using two boundary conditions, at $z = 0$ and $|z| = b$. Conversely, the Neerhoff and Mur method performs the matching with five boundary conditions, at $z \to 0$, $z \to b$, $x \to a$, $x \to -a$, and $r \to \infty$. In contrast to the sharp Lorentzian-like transmission peaks, the slit forms very wide Fano-type reflection bands (see, Fig. 13a). For one broad minimum in reflection, it appears also a local reflection maximum, which is characterized by weak amplitude. At the near-to-resonance wavelengths of the transmission, the open aperture totally reflects the light. It is worth to note the correlation between the wavelengths for maxima and minima in the reflection $R(\lambda)$, transmission $T(\lambda)$, and the flux S^d_{int} (Table 1).

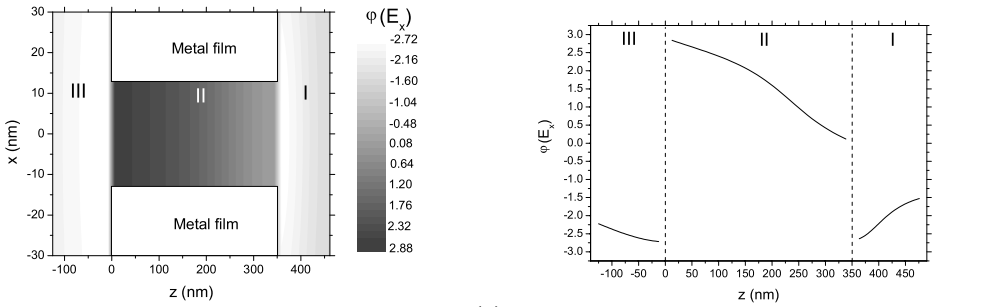
The resonance wavelengths for the main reflection maxima are red-shifted with respect to the transmission resonances. The wavelengths of both the transmission and reflection (main and little) resonances are red-shifted with respect to the Fabry-Perot wavelengths $\lambda^m_r = 702$ nm, 351 nm,... ($\lambda^m_r = 2b/m$, $m = 1, 2, 3, \ldots$). To understand the physical mechanism of the resonant backward scattering, we also compared the integrated flux $S^{ird}_{int}(U^i, U^r, U^d)$ with the fluxes $S^d_{int}(U^d)$ and $S^3_{int}(U^3) = T$. As can be seen from Fig. 13a, the integrated fluxes $S^d_{int}(\lambda)$, $S^{ird}_{int}(\lambda)$ and $S^3_{int}(\lambda) = T(\lambda)$ are practically indistinguishable also in the λ-domain (for the b-domain, see Fig. 12a). The difference between the integrated fluxes $S^{rd}_{int}(U^r, U^d) = R$ and $S^3_{int}(U^3) = T$ ($T \approx S^d_{int}(U^d)$) is caused by the interference of the backward diffracted field $U^d(x, z)$ and the reflected field $U^r(x, z)$

Table 1. The wavelengths for maxima and minima in the reflection $R(\lambda)$, transmission $T(\lambda)$, and the flux S^d_{int}.

λ (nm) of	R^{main}_{max}	R^{little}_{max}	R_{min}
	276	248	237
	426	377	255, m=3
	882	773	356
			389, m=2
			714
			802, m=1
	T_{max}	T_{min}	S^d_{max}
	260, m=3	226	253
	396, m=2	315	388
	812, m=1	542	802

in the energy flux $\vec{S} \sim (\vec{E} \times \vec{H}^* + \vec{E}^* \times \vec{H})$. The wavelengths of the little maxima of the reflection $R = R(U^d, U^r)$ correspond approximately to the high maxima S^d_{int}. Therefore, the little maxima can be attributed to the interference of the reflected field U^r with the dominant diffracted field U^d. The red shifts and the asymmetrical shapes of the reflection bands can be explained by a Fano analysis [36] of the scattering problem by distinguishing resonant and non-resonant interfering contributions to the reflection process. The resonant contribution is given by the field U^d and the non-resonant one is attributed to the field U^r. Other interesting interpretations of the shifts of resonant wavelengths in the transmission spectra from the values $2b/m$ can be found in the studies [17, 18, 20, 37, 41, 59]. It should be mentioned that the asymmetrical behavior of reflection was observed also in the case of a Fabry-Perot resonator [45, 46]. The conditions to achieve such an asymmetry are rested on the existence of dissipative loss in the resonator. There is no explicit loss in the present problem, but the dissipative loss can be substituted by radiative loss due to the diffraction by the slit.

After the analysis of Fig. 12a, it is not surprising that the maxima of the transmission are accompanied by the minima of the reflection also in the λ-domain (see, Fig. 13a). It should be noted in this connection that such a behavior of $R(\lambda)$ and $T(\lambda)$ is similar to that observed in the case of excitation of the surface plasmons in an array of slit in a thin metal film. [29] In the study [29], the minima in reflection spectra corresponding to the maxima in the transmission spectra were attributed to the redistribution of the energy of diffracted evanescent order into the propagating order. In the case of a thick film, we explain such a behavior by another physical mechanism, the interference of the backward diffracted field $U^d(x, z)$ and the reflected field $U^r(x, z)$. It can be noted that the correlation of positions of reflection minima and transmission maxima (see, Fig. 12a and Fig. 13a are consistent with that predicted by the study [41] for TE-polarized light scattered by a grating of two-side-open slits in a thick metal film. However, the values of $R(\lambda)$ and $T(\lambda)$ are in contrast to the relation $R(\lambda) + T(\lambda) = 1$ given in the study [41]. The difference can be explained by the fact that we examined light scattering by an infinite screen using local definitions of R

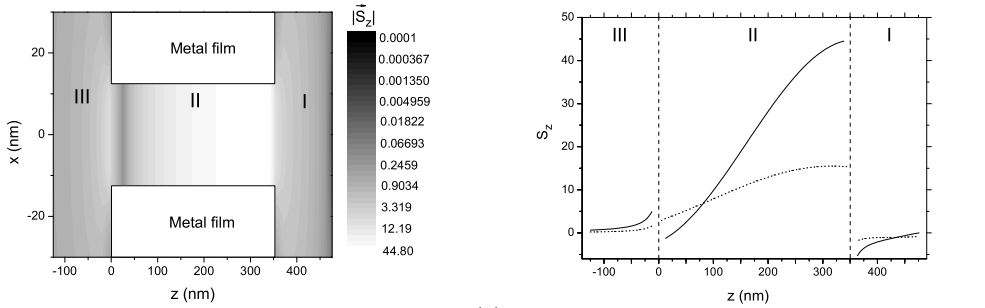


(a) (b)

Figure 14. (a) The phase distribution $\varphi(x,z)$ of the electric field x-component $E_x(x,z)$ inside and outside the slit. The field component $E_x(x,z)$ is given by $E_x(U^i,U^r,U^d)$, $E_x(U^2)$ and $E_x(U^3)$ in the regions I, II and III, respectively. (b) The phase distribution $\varphi(x,z)$ at $x=0$. The slit width $2a=25$ nm, the film thickness $b=351$ nm and the wavelengths $\lambda=800$ nm.

and T, while the study [41] analyzed the global reflection and transmission by a grating of finite size.

The dispersions $S^d_{int}(\lambda)$, $S^{ird}_{int}(\lambda)$ and $S^3_{int}(\lambda)=T(\lambda)$ shown in Fig. 13a indicate the wave-cavity interaction behavior, which is similar to that in the case of a Fabry-Perot resonator. The fluxes $S^d_{int}(U^d)$, $S^{ird}_{int}(U^i,U^r,U^d)$ and $S^3_{int}(U^3)=T$ exhibit the Fabry-Perot-like maxima around the resonance wavelengths $\lambda^m_r\approx 2b/m$. In order to understand the connection between the Fabry-Perot-like resonances and the total reflection, we computed the amplitude and phase distributions of the light wave at the resonant and near-resonant wavelengths inside and outside the slit cavity (see, Fig. 13b, Fig. 14 and Fig. 15). At the resonance wavelengths, the intra-slit fields possess maximum amplitudes with Fabry-Perot-like spatial distributions (Fig. 13b). However, in contrast to the Fabry-Perot-like modal distributions, the resonant configurations are characterized by antinodes of the electric field at each open aperture of the slit. Such a behavior is in agreement with the results [37,43,44]. It is interesting that at the slit entrance, the amplitudes E_x of the resonant field configuration possesses the Fabry-Perot-like phase shift on the value of π (Fig. 14). The integrated fluxes $S^{ird}_{int}(U^i,U^r,U^d)$, $S^d_{int}(U^d)$ and $S^3_{int}(U^3)$, at the first resonant wavelength λ^1_r, exhibit enhancement by a factor $\lambda/2\pi a\approx 10$ with respect to the incident wave (Fig. 15). For the comparison, the normalized resonant fluxes $S^{ird}_n(U^i,U^r,U^d)$ and $S^3_n(U^3)$ in the near-field zone ($z\approx -2a$) are about 5 times greater than the incident wave (see, Fig. 12 and Fig. 13). It should be stressed that the resonantly enhanced intra-cavity intensity $S^2_n(U^2)$ is about 10 times higher than the resonant fluxes $S^{ird}_n(U^i,U^r,U^d)$ and $S^3_n(U^3)$ localized in the near field zone in front of the slit and behind the slit, respectively (Fig. 15). The interference of the incident $U^i(x,z)$ wave and the backward scattered fields $U^r(x,z)$ and $U^d(x,z)$, at the resonant wavelengths $\lambda^m_r\approx 2b/m$, produces in the near-field diffraction zone a strongly localized wave whose normalized flux $S^{ird}_n(U^i,U^r,U^d)$ is $\lambda/2\pi a\approx$10-10^3 times greater than the incident wave, but about one order of magnitude smaller than the resonant intra-cavity intensity.



(a) (b)

Figure 15. (a) The spatial distribution $|\vec{S}_z(x,z)|$ of the absolute value of the normalized energy flux along the z direction inside and outside the slit. The distribution $|S_z(x,z)|$ is shown in the logarithmical scale. The flux S_z is given by the normalized fluxes $S^{ird}(U^i,U^r,U^d)/S^i(U^i)$, $S^2(U^2)/S^i(U^i)$ and $S^3(U^3)/S^i(U^i)$ for the regions I, II and III, respectively. (b) The energy flux distribution $S_z(x,z)$ at $x=0$. The slit width $2a=25$ nm, the film thickness $b=351$ nm, and the wavelengths $\lambda=800$ nm (solid line) and $\lambda=882$ nm (dotted line) corresponding to a transmission resonance and a little reflection resonance, respectively.

In the above-presented analysis showed that the interplay between the resonant enhancement and spatial uncertainty of a continuous wave leads to the following effects. The intensity of a wave localized in the near-field nm-scale zone behind the slit is 10-10 ^{3} times greater than the incident wave, but one order of magnitude smaller than the intra-cavity intensity. Interference of incident and resonantly backward-scattered light produces in the near-field zone, in front of the slit, a spatially localized (nm-scale) wave whose intensity is also 10-10^3 times greater than the incident wave, but about ten times smaller than the intra-slit intensity. In the model we considered an incident wave with TM polarization. Nevertheless, in our case of the perfectly conducting nonmagnetic screen placed in the vacuum, the symmetry of wave equations indicates that the large transmission coefficients (enhancement effect) and spatial localization of a continuous waves (Fourier's ω-components of a wave-packet (40)) do exist at the same experimental conditions for the TM and TE polarizations.

5.2. Resonant Enhancement and Near-Field Localization of FS Pulses

In Section 5.2., we consider the resonant enhancement and near-field localization of fs wave-packets. The material of the section is reprinted with permission from S. V. Kukhlevsky, M. Mechler, L. Csapo, K. Janssens, and O. Samek, Physical Review B, Vol. 70, 195428 (2004). Copyright (2004) by the American Physical Society.

Let us test whether a fs light pulse can be resonantly enhanced and simultaneously localized in space and time by a subwavelength nano-sized metal slit. To address this question, the spatial distribution of the energy flux of the transmitted pulse under various regimes of the near-field diffraction is analyzed numerically. The electric $\vec{E}$ and magnetic $\vec{H}$ fields of

the transmitted pulse in the near-field diffraction zone are computed by solving the equations (5-38) for each Fourier ω-component of the wave-packet (40). The amplitude of a FFT ω-component of the wave-packet transmitted through the slit depends on the wavelength $\lambda = 2\pi c/\omega$ (for details, see Sections 4.1. and 4.2.). Owing to the dispersion, the Fourier spectra of the transmitted wave-packet changes leading to modification of the pulse width and duration. The dispersion of a continuous wave is described by the normalized transmission coefficient $T_{cw}(\lambda)$, which is calculated by integrating the normalized energy flux S_z/S_z^i over the slit width (see, Eq. 41). In order to establish guidelines for the results of our numerical analysis, we computed the transmission coefficient $T_{cw}(\lambda, a, b)$ for Fourier's ω-components of a wave-packet as a function of the screen thickness b and/or wavelength λ for different values of slit width $2a$. Throughout the computations, the magnitude of the incident magnetic field is assumed to be equal to 1. We consider a perfectly conducting nonmagnetic film placed in vacuum ($\epsilon_1 = \epsilon_2 = \epsilon_3 = 1$). As an example, the dependence $T_{cw} = T_{cw}(b)$ computed for the wavelength λ = 800 nm and the slit width $2a$ = 25 nm is shown in Fig. 16. The dispersion $T_{cw} = T_{cw}(\lambda)$ for $2a$ = 25 nm and different values of the screen thickness b is presented in Fig. 17. In Fig. 16, we note the transmission resonances of $\lambda/2$ periodicity with the peak heights $T_{cw} \approx \lambda/2\pi a$ at the resonances.

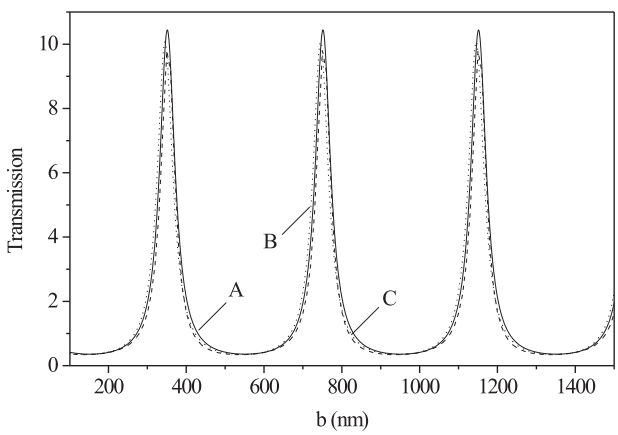


Figure 16. The transmission coefficient T_{cw} for a continuous wave (ω-Fourier component of a wave-packet) as a function of screen thickness b computed for the wavelength λ=800 nm and the slit width $2a$ = 25 nm. Curve A, B and C correspond to the numerical near-field formula (41), the analytical near-field formula (42) and the analytical far-field formula (Eq. (41) with fields given by Eqs. (35-38) and the integration performed in the region $[-\infty,\infty]$ at distance $z \gg a$), respectively. In the case of curve C the limits of the integral were chosen to -1000λ and 1000λ; the computation was performed at $z = -10\lambda$.

Notice that the resonance positions and the peak heights are in agreement with the results [4, 5] (see, also Section 4.1.). The resonance peaks appear with spacing of 400 nm, but at values somewhat smaller than 400, 800, 1200 nm, etc. In order to better understand

the shift of the resonant wavelengths from the naively expected values we have derived a simple analytical formula (42) for the transmission coefficient $T_{cw} = T_{cw}(a, b, \lambda)$ of a narrow slit (the fundamental mode only is retained):

$$T_{cw}(a, b, \lambda) = \frac{a}{2\frac{\pi}{\lambda}} \left[\left(\mathrm{Re}\left(D\left(b, \lambda \right) \right) \right)^2 + \left(\mathrm{Im}\left(D\left(b, \lambda \right) \right) \right)^2 \right], \tag{42}$$

where

$$D(b, \lambda) = \frac{4}{2i\frac{\pi}{\lambda}} \left[\left[\exp\left(2ib\frac{\pi}{\lambda} \right) \left(A(\lambda) - \frac{i}{2\pi/\lambda} \right) \right]^2 - \left(A(\lambda) + \frac{i}{2\pi/\lambda} \right)^2 \right]^{-1} \tag{43}$$

and

$$A(\lambda) = ia \left[H_0^{(1)}(\lambda) + \frac{\pi}{2} \left(\mathbf{H}_0(\lambda) \cdot H_1^{(1)}(\lambda) + \mathbf{H}_1(\lambda) \cdot H_0^{(1)}](\lambda) \right) \right]. \tag{44}$$

In Eq. 44, $H_0^{(1)}(\lambda)$ and $H_1^{(1)}(\lambda)$ are the Hankel-functions, while $\mathbf{H}_0(\lambda)$ and $\mathbf{H}_1(\lambda)$ are the Struve-functions. It is interesting to compare the "near-field" ($z << \lambda$) transmittance with the conventional "far-zone" ($z >> \lambda$) transmittance measured in experiments. Figure 16 compares the transmission coefficients $T_{cw} = T_{cw}(b)$ calculated by using the formula (42) with the values given by the numerical solution of the equations (28-31, 41). We notice that the results are practically indistinguishable. Analysis of the denominator of the formula (42) indicates that for $2a/\lambda$ small enough, the transmission $T_{cw} = T_{cw}(b)$ will exhibit the maximums around wavelengths $\sim \lambda/2$. The shifts of the resonance wavelengths from the value $\lambda/2$ are caused by the wavelength dependent terms in the denominator of Eq. 42.

The dispersion $T_{cw} = T_{cw}(\lambda)$ presented in Fig. 17, indicates the slit-wave interaction behavior, which is similar to those of a Fabry-Perot resonator. The transmission resonance peaks, however, have a systematic shift towards longer wavelengths. Analysis of the denominator of the formula (42) indicates that for the value $2a/\lambda$ small enough, the transmission $T_{cw} = T_{cw}(\lambda)$ will exhibit the Fabry-Perot like maximums around wavelengths where $\sin(kb)$ is zero. Such a behavior has already been described in refs. [17, 18, 38]. The shifts of the resonance wavelengths from the Fabry-Perot resonances are caused by the wavelength dependent terms in the denominator of Eq. 42. It should be noted that the values of the shifts calculated using Eq. 42 are in good agreement with the shifts calculated using the analytical formula (8) of the study [17]. Our computations showed that the peak heights at the main (strongest) resonant wavelength λ_0^R (in the case of Fig. 17, λ_0^R= 500 or 800 nm) are given by $T_{cw}(\lambda_0^R, a) \approx \lambda_0^R/2\pi a$. Notice that the Fabry-Perot like behavior of the transmission coefficient is in agreement with analytical and experimental results published earlier [17, 18]. The enhancement coefficient ($T_{cw} \sim 30$) calculated using Eqs. 5-38 and 42, however, is different from the attenuation ($T_{cw} < 1$) predicted by the study [17]. The one order difference between the experimental value and the enhancement calculated in the present article is attributed to the small transverse dimension D ($D \sim \lambda/2$) of the metal plates that form the slit. In our calculations we studied a slit in infinite transverse dimension (D = ∞).

Analysis of Fig. 17 indicates that a minimum thickness of the screen is required to get the waveguide resonance inside the slit at a given wavelength. The result is in agreement

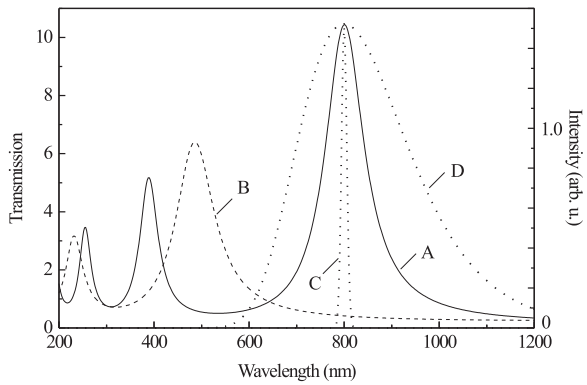


Figure 17. The dispersion $T_{cw} = T_{cw}(\lambda)$ for a continuous wave (ω-Fourier component of a wave-packet) computed for the slit width $2a$ = 25 nm and different values of the screen thickness b: A - 350 nm and B - 200 nm. The Fourier spectra (curves C and D) are presented for the comparison. Curves C and D show the Fourier spectra of incident wave-packets with central wavelength $\lambda_0 = 800$ nm and duration $\tau = 100$ fs and $\tau = 5$ fs, respectively, which were used in the computations presented in Fig. 18 and Fig. 19.

with the study [17], which demonstrated that there is no transmission peaks at the condition $b < \lambda/2$. Notice that the transmission enhancement mediated by surface plasmons does exist at considerably smaller thicknesses of the metal in comparison with the thickness required for the waveguide resonance. The surface plasmons/polaritons are excited in the skin, the depth of which is about the extinction length of the energy decay of electromagnetic wave in the metal. For instance, the extinction length in an aluminum screen is ~ 4 nm for λ = 800 nm. Therefore, for not too narrow slits ($2a$ = 25 nm) in thick screens (b = 200 and 350 nm) considered in the paper, the finite skin depth does not fundamentally modify the process of the extraordinary optical transmission. At the wavelength 800 nm, the 10-times enhancement ($T_{cw}(a, \lambda) \approx \lambda/2\pi a \approx 10$) is limited by the slit width ($2a$ = 25nm). In the far-infrared region ($\lambda \approx 100$ μm) several orders of magnitude enhancement can be achieved at the same experimental conditions. It should be noted that the optimal choice of parameters has been discussed in the literature and the obtained enhancement by the factors 10 and 1000 are typical for continuous waves in the optical and far-infrared spectral ranges. In the case of narrower and thinner slits, however, the influence of the finite conductivity effects on the transmission and localization of a pulse should be taken into account [5, 17].

The existence of transmission resonances for Fourier ω-components of a wave-packet leads to the question: What effect the resonant enhancement has on the spatial and temporal localization of a light pulse? Presumably, the high transmission at resonance occurs when the system efficiently channels Fourier-components of the wave-packet from a wide area through the slit. At resonance, one might assume that if the energy flow is symmetric about the screen, the pulse width and duration should increase very rapidly past the screen. Thus the large pulse strength associated with resonance could only be obtained at the expense of

the spatial and temporal broadening of the wave-packet. To test this hypothesis, the spatial distributions of the energy flux of a transmitted wave-packet were computed for different slit thicknesses corresponding to the resonance and anti-resonance position. The characteristic difference between the resonant and non-resonant transmissions could be understood better through a single figure of the field distributions in all regions I, II and III. However, it seems to be impossible to do this, because the visualization of a pulse by a single figure requires a 4-dimensional (U, x, z, t) coordinate system or a great number of 3-dimensional figures at different locations and suitable fixed times in all the regions. Therefore we present the eight 3-dimensional distributions (18a, 18b, 18c, 18d, 19a, 19b, 19c and 19d) only in the zone of main interest (region III). As an example, Figs. 18a,c and 19a,c show the energy flux of the anti-resonantly transmitted pulses. Figures 18b,d and 19b,d correspond to the

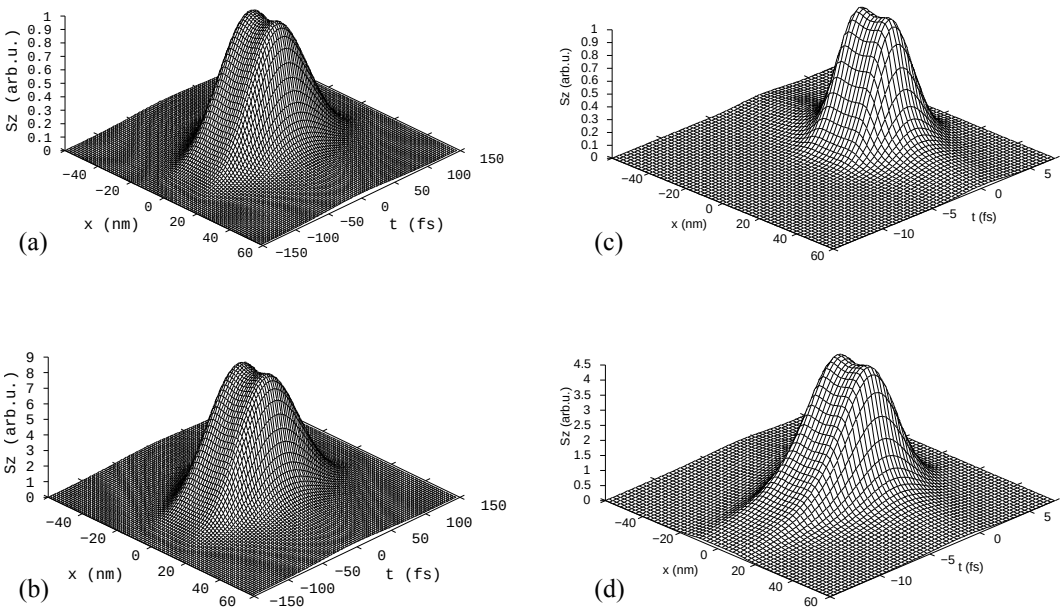


Figure 18. The energy flux of a transmitted pulse at the distance $|z| = a/2$. (a), (c) The non-resonant transmission by the slit ($2a$ = 25 nm, b = 200 nm; $\tau = 100$ fs and 5 fs, respectively). (b), (d) The resonant transmission by the slit ($2a$ = 25 nm, b = 350 nm; $\tau = 100$ fs and 5 fs, respectively). The central wavelength of the incident wave-packet is $\lambda_0 = 800$ nm.

case of the waveguide-mode resonance in the slit. Figures 18 and 19 show the transmitted pulses at the distances $|z| = a/2$ and a, respectively. The rigorous analysis [5] showed that the number of modes required for the accurate computation of the transmittance, near-field distribution, and other optical characteristics of the system is given by $N > 2a/z$, where N is the number of the waveguide modes and z is the distance from the screen. Therefore at the distances $|z| = a/2$ and a the calculations required at least 4 and 2 modes, respectively. The shape and intensity of an output pulse depends on the slit parameters and the spectral width of the pulse. For narrow slits, the spectral width of a 100-fs input pulse is smaller compared to the spectral width of the resonant transmission (see, Fig. 17). According to the figure, the resonant and non-resonant transmission coefficients are practically constant for

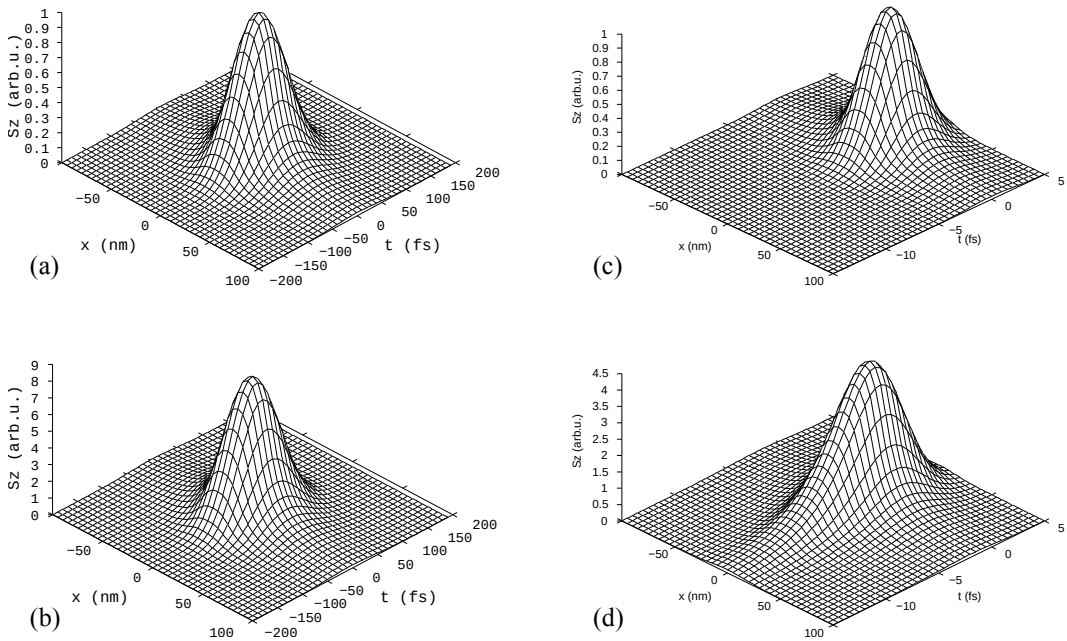


Figure 19. The energy flux of a transmitted pulse at the distance $|z| = a$. (a), (c) The non-resonant transmission by the slit ($2a$ = 25 nm, b = 200 nm; $\tau = 100$ fs and 5 fs, respectively). (b), d The resonant transmission by the slit ($2a$ = 25 nm, b = 350 nm; $\tau = 100$ fs and 5 fs, respectively). The central wavelength of the incident wave-packet is $\lambda_0 = 800$ nm.

all Fourier's spectral components of the wavepacket. The comparison of the flux distribution presented in Fig. 18a and Fig. 19a with those of Fig. 18b and Fig. 19a, Respectively, shows that, for the parameter values adopted, a transmitted wavepacket is enhanced by one order of magnitude and simultaneously localized in the 25-nm and 100-fs domains of the near-field diffraction zone. The shapes of the intensity distributions of the output pulses are very much the same off and on resonance. The figures differ only in the order of magnitude of S_z. Thus at the distance $|z| = a/2$, the slit resonantly enhances the intensity of the pulse without its spatial and temporal broadening. The result can be easily understood by considering the dispersion properties of the slit. For the screen thickness b = 200 nm, the amplitudes of the Fourier-components of the wave-packet, whose central wavelength λ_0 is detuned from the main (at 500 nm) resonance, are practically unchanged in the wavelength region near 800 nm (see, curve B in Fig. 17). This provides the dispersion- and distortion-free non-resonant transmission of the wave-packet (Fig. 18a and Fig. 19a). In the case of the thicker screen (b = 350 nm), the slit transmission experiences strong mode-coupling regime at the wavelengths near to 800 nm (see curve A of Fig. 17) that leads to a profound and uniform enhancement of amplitudes of all of the Fourier ω-components of the wave-packet (see curve C in Fig. 17). Thus, the slit resonantly enhances by one order of magnitude the intensity of the pulse without its spatial and temporal broadenings (Figs. 18b and 19b). Also, notice that at the distance $|z| = a$ (Figs. 19a and 19b), both the resonantly and anti-resonantly transmitted pulses experience natural spatial broadening in the transverse direction, while their durations are practically unchanged. The spectral width of a 5-fs in-

put pulse, however, is bigger than the spectral width of the resonant transmission (Fig. 17). In this case the durations of the resonantly transmitted 5-fs pulses are longer in comparison with the pulses transmitted off the resonance (Fig. 18c,d and respectively Fig. 18c,d). We also notice that the enhancement of the intensity of 5-fs pulses is approximately two times lower in comparison with the 100-fs pulses. The temporal broadening of the pulse and the decrease of the enhancement indicates a natural limitation for the simultaneous temporal and spatial localization of a pulse together with its enhancement.

By comparing the data for anti-resonant and resonant transmissions presented in Fig. 18 and Fig. 19 one can see that at the appropriate values of the distance $|z|$ and the wave-packet spectral width $\Delta\omega$ the resonance effect does not influence the spatial and temporal localization of the wave-packet. To verify this somewhat unexpected result, the FWHMs of the transmitted pulse in the transverse and longitudinal directions were calculated for different values of the slit width $2a$, central wavelength λ_0 and pulse duration $\tau{\approx}1/\Delta\omega_p$ as a function of screen thickness b at two particular near-field distances $|z| = a/2$ and a from the screen. It was seen that, at the dispersion-free resonant transmission condition $\Delta\omega_p < \Delta\omega_r$, the transmitted pulse indeed does not experience temporal broadening. Thus the temporal localization associated with the duration τ of the incident pulse remains practically unchanged under the transmission. The value of τ is determined by the dispersion-free condition $\tau{\approx}1/\Delta\omega_p > 1/\Delta\omega_r$, where $\Delta\omega_r = \Delta\omega_r(a)$ practically does not depend on the screen thickness b. We found that the energy flux of the transmitted wave-packet can be enhanced by a factor $T_{cw}(\lambda_0^R, a){\approx}\lambda_0^R/2\pi a$ by the appropriate adjusting of the screen thickness $b = b(\lambda_0^R)$, for an example see Fig. 18 and Fig. 19. Thus the wave-packet can be enhanced by a factor $\lambda_0^R/2\pi a$ and simultaneously localized in the time domain at the $\tau = \tau(a)$ scale. It was also seen that the FWHM of the transmitted pulse in the transverse direction depends on the wave-packet central wavelength λ_0 and the distance z from the slit. Nevertheless, the FWHM of the transmitted pulse can be always reduced to the value $2a$ by the appropriate decreasing of the distance $|z| = |z(a)|$ from the screen ($|z| = a/2$, in the case of Fig. 18). Thus high transmission can be achieved without concurrent loss in the degree of temporal and spatial localizations of the pulse. In retrospect, this result is reasonable, since the symmetry of the problem for a time-harmonic continuous wave (Fourier ω-component of a wave-packet) is disrupted by the presence of the initial and reflected fields in addition to the diffracted field on one side (Eq. 10). As the thickness changes, the fields U^d and U_3 change only in magnitude, but the field U_1 changes in distribution as well since it involves the sum of U^d with unchanging fields U^i and U^r. At resonance, the distribution of U_1 leads to channeling of the radiation, but the distribution of U_3 remains unaffected. By the appropriate adjusting of the slit-pulse parameters a light pulse can be enhanced by orders of magnitude and simultaneously localized in the near-field diffraction zone at the nm and fs scales.

The limitations of the above analysis must be considered before the results are used for a particular experimental device. The resonant enhancement with simultaneous nm-scale spatial and fs-scale temporal localizations of a light pulse by a subwavelength metal nano-slit is a consequence of the assumption of the screen perfect conductivity. The slit can be made of perfectly conductive (at low temperatures) materials. In the context of current technology, however, the use of conventional materials like metal films at a room temperature is more practical. Notice that a metal can be considered as a perfect conductor in the microwave

range. As a general criterion, the perfect conductivity assumption should remain valid as far as the slit width and the screen thickness exceed the extinction length for the Fourier ω-components of a wavepacket within the metal. The light intensity decays in the metal screen at the rate of $I_s = I_0 \exp(-b/\delta)$, where $\delta = \delta(\lambda)$ is the extinction length in the screen. The aluminum has the largest opacity ($\delta < 11$ nm) in the spectral region $\lambda > 100$ nm [20]. The extinction length increases from 11 to 220 nm with decreasing the wavelength from 100 to 50 nm. Hence, the perfect conductivity is a very good approximation in a situation involving a relatively thick ($b > 25$ nm) aluminum screen and a wave-packet of the duration $\tau \approx 1/\Delta\omega_p$ having the Fourier components in the spectral region $\lambda > 100$ nm. However, in the case of thinner screens, shorter pulses and smaller central wavelengths of wave-packets, the metal films are not completely opaque. This would reduce a value of spatial localization of a pulse due to passage of the light through the screen in the region away from the slit. Moreover, the phase shifts of the Fourier components along the propagation path caused by the skin effect can modify the enhancement coefficient and temporal localization properties of the slit.

The above analysis is directly applicable to the two-dimensional near-field scanning optical microscopy and spectroscopy. Although the computations were performed in the case of normal incidence, the preliminary analysis shows that in the case of oblique incidence the enhancement effect can be kept without further spatial and temporal broadening of the pulse. In a conventional 2D NSOM, a subwavelength ($2a < \lambda$) slit illuminated by a continuous wave is used as a near-field ($|z| < a$) light source providing the nm-scale resolution in space [5–7, 9]. The non-resonant transmission of fs pulses could provide ultrahigh resolution of 2D NSOM simultaneously in space and time [58, 59]. The above-described resonantly enhanced transmission together with nm- and fs-scale localizations in the space and time of a pulse could greatly increase the potentials of the 2D near-field scanning optical microscopy and spectroscopy, especially in high-resolution applications. However, as a a model for a NSOM-tip, a hole (quadratic, rectangular or circular) would be more appropriate than a slit-type waveguide. The consequences of the limitation of region II also in y-direction will result in faster attenuation of the amplitude of the waveguide modes compared to the slit-type waveguide modes [5]. This should affect the enhancement and spatial and temporal broadening of the pulse. It should be noted that the walls of the 2D slit have to be parallel over all the thickness to provide the resonance enhancement effect. Pulled NSOM 3D-tips, however, usually have conical ends. In the case of the tapered NSOM 3D waveguides, however, one should take into account the decrease of enhancement effect with increasing the waveguide taper. It should be also noted that the high transmission ($T_{cw}(\lambda_0^R, a) \approx \lambda_0^R/2\pi a$) of a pulse can be achieved without concurrent loss in the temporal and spatial localizations of the pulse only at short ($|z| = |z(a)|$) distances from the slit. The presence of a microscopic sample (a molecule, for example) placed at the short distance in strong interaction with NSOM slit, however, modifies the boundary conditions. In the case of strong slit-sample-pulse interaction, which takes place at the distance $|z| << 0.1a$, the response function accounting for the modification of the quantum mechanical behavior of the sample should be taken into consideration. The potential applications of the effect of the resonantly enhanced transmission together with nm- and fs-scale localizations of a pulse are not limited to near-field microscopy and spectroscopy. Broadly speaking, the effect concerns all physical phenomena and photonic applications involving a transmission of light by

a single subwavelength nano-slit, a grating with subwavelength slits and a subwavelength slit surrounded by parallel grooves (see the studies [3–9, 14–21, 24–35, 38, 40, 41, 54–60] and references therein). For instance, the effect could be used for sensors, communications, optical switching devices and microscopes.

In Section 4.2., we demonstrated the limitation of the spatio-temporal localization of an ultra-short pulse imposed by the spatio-temporal uncertainty of a wave-packet under the non-resonant transmission. The analysis presented in Section 5.2. indicates that the interplay between the resonant enhancement and the spatio-temporal uncertainty of a wave-packet leads to the additional limitation of the spatio-temporal localization of an ultra-short pulse. We showed that at the appropriate experimental conditions, the pulse can be enhanced by orders of magnitude and simultaneously localized in the near-field diffraction zone at the nm and fs scales.

6. Conclusion

In the present article we have considered the question whether a light pulse can be enhanced and simultaneously localized in space and time by a subwavelength metal nano-slit. To address this question, the spatial distributions of the electromagnetic field of an ultrashort (fs) pulse diffracted by a subwavelength (nanosized) slit in a thick metal film of perfect conductivity have been analyzed by using the conventional approach based on the Neerhoff and Mur solution of Maxwell's equations. The analysis of the spatial distributions for various regimes of the near-field diffraction demonstrated that the energy flux of a wave-packet can be enhanced by orders of magnitude and simultaneously localized in the near-field diffraction zone at the nm- and fs-scales. The extraordinary transmission, together with nm and fs scale localizations of the light pulse, make the nano-slit structures attractive for many photonic purposes, such as NSOM, sensors, communication and optical switching devices. We also believe that the addressing of the above-mentioned basic question gains insight into the physics of near-field non-resonant and resonant diffraction.

Acknowledgments

I wish to thank all my students who have been associated with the research described in this chapter. They are named as coauthors in the references below. The author is especially grateful to his colleagues K. Janssens and O. Samek for their substantial contributions to the present study. The support of the research by the Fifth Framework of the European Commission (Contract NG6RD-CT-2001-00602) is gratefully acknowledged. The study was supported in part by the Hungarian Scientific Research Foundation (OTKA, Contracts T046811 and M045644), and the Hungarian R&D Office (KPI, Contract GVOP-3.2.1.-2004-04-0166/3.0).

References

[1] L. Rayleigh, *Philos. Mag.* **XLIV**, 28 (1897).

[2] H.A. Bethe, *Phys. Rev.* **66**, 163 (1944).

[3] F. L. Neerhoff, G. Mur G, A*ppl. Sci. Res.* **28**, 73 (1973).

[4] R.F. Harrington, D.T. Auckland, *IEEE Trans. Antennas Propag* **AP28**, 616 (1980).

[5] E. Betzig, A. Harootunian, A. Lewis, M. Isaacson, *Appl. Opt.* **25**, 1890 (1986).

[6] E.A. Ash, G. Nicholls, *Nature (London)* **237**, 510 (1972).

[7] A. Lewis, K. Lieberman, *Nature (London)* **354**, 214 (1991).

[8] E. Betzig, J.K. Trautman, T.D. Harris, J.S. Weiner, R.L. Kostelak, *Science* **251**, 1468 (1991).

[9] D.W. Pohl, D. Courjon, Eds., *Near Field Optics*, *NATO ASI Series E* Vol. 242 (The Netherlands, Dordrecht: Kluwer, 1993).

[10] V.V. Fano, *Ann. Phys.* (N.Y.) **32**, 393 (1938).

[11] A. Hessel and A.A. Oliner, *Appl. Opt.* **4**, 1275 (1965).

[12] M. Nevière, D. Maystre, and P. Vincent, *J. Opt.* **8**, 231 (1977).

[13] D. Maystre and M. Nevière, *J. Opt.* **8**, 165 (1977).

[14] T.W. Ebbesen, H.J. Lezec, H.F. Ghaemi, T. Thio, P.A. Wolff, *Nature (London)* **391**, 667 (1998).

[15] H.J. Lezec, A. Degiron, E. Devaux, R.A. Linke, Martin-Moreno, F.J. Garcia-Vidal, T.W. Ebbesen, *Science* **297**, 820 (2002).

[16] J. A. Porto, F.J. Garcia-Vidal, J.B. Pendry, *Phys. Rev. Lett.* **83**, 2845 (1999).

[17] Y. Takakura, *Phys. Rev. Lett.* **86**, 5601 (2001).

[18] F.Z. Yang, J.R. Sambles, *Phys. Rev. Lett.* **89**, 063901 (2002).

[19] U. Schroter, D. Heitmann, *Phys. Rev. B* **58**, 15419 (1998).

[20] M.M.J. Treacy, *Appl. Phys. Lett.* **75**, 606 (1999).

[21] L. Salomon, F.D. Grillot, A.V. Zayats, F. de Fornel, *Phys. Rev. Lett.* **86**, 1110 (2001).

[22] A.E. Kaplan, P.L. Shkolnikov, *Phys. Rev. Lett.* **88**, 074801-4 (2001).

[23] M.I. Stockman, *Phys. Rev. Lett.* **84**, 1011 (2000).

[24] Q. Cao, P. Lalanne, *Phys. Rev. Lett.* **88**, 057403 (2002).

[25] I.I. Smolyaninov, A.V. Zayats, A. Stanishevsky, C.C. Davis, *Phys. Rev. B* **66**, 205414 (2002).

[26] A. Barbara, P. Quemerais, E. Bustarret, T. Lopez-Rios, *Phys. Rev. B* **66**, 161403 (2002).

[27] E. Altewischer, M.P. van Exter, J.P. Woerdman, *Nature (London)* **418**, 304 (2002).

[28] A.M. Dykhne, A.K. Sarychev, V.M. Shalaev, *Phys. Rev. B* **67**, 195402 (2003).

[29] J.M. Steele, C.E. Moran, A. Lee, C.M. Aguirre, N.J. Halas, *Phys. Rev. B* **68**, 205103 (2003).

[30] A. Naweed, F. Baumann, W.A. Bailey, A.S. Karakashian, W.D. Goodhue, *J. Opt. Soc. Am. B* **20**, 2534 (2003).

[31] Z.J. Sun, Y.S. Jung, H.K. Kim, *Appl. Phys. Lett.* **83**, 3021 (2003).

[32] X.L. Shi, L. Hesselink, R.L. Thornton, *Opt. Lett.* **28**, 1320 (2003).

[33] H.F. Schouten, T.D. Visser, D. Lenstra, H. Blok, *Phys. Rev. E* **67**, 036608 (2003).

[34] A. Nahata, R.A. Linke, T. Ishi, K. Ohashi, *Opt. Lett.* **28**, 423 (2003).

[35] G. Gomez-Santos, *Phys. Rev. Lett.* **90**, 077401 (2003).

[36] M. Sarrazin, J.P. Vigneron, and J.M. Vigoureux, *Phys. Rev. B* **67**, 085415 (2003).

[37] S. Astilean, Ph. Lalanne, and M. Palamaru, *Opt. Commun.* **175**, 265 (2000).

[38] A.P. Hibbins, J.R. Sambles and C.R. Lawrence, *Appl. Phys. Lett.* **81**, 4661 (2002).

[39] S.V. Kukhlevsky, M. Mechler, L. Csapo, K. Janssens, and O. Samek, *Phys. Rev. B* **70**, 195428 (2004).

[40] F.J. García-Vidal, H.J. Lezec, T.W. Ebbesen, and L. Martín-Moreno, *Phys. Rev. Lett.* **90**, 231901 (2003).

[41] A.G. Borisov, F.G. Garcia de Abajo, and S.V. Shabanov, *Phys. Rev. B* **71**, 075408 (2005).

[42] F.J. García-Vidal and L. Martín-Moreno, *Phys. Rev. B* **66**, 155412 (2002).

[43] J. Lindberg, K. Lindfors, T. Setala, M. Kaivola, and A.T. Friberg, *Opt. Express* **12**, 623 (2004).

[44] Y. Xie, A.R. Zakharian, J.V. Moloney, and M. Mansuripur, *Opt. Express* **12**, 6106 (2004).

[45] J.J. Monzón and L.L. Sánchez-Soto, *JOSA A* **12**, 132 (1995).

[46] R. Giust, J.M. Vigoureux, and M. Sarrazin, *JOSA A* **17**, 142 (2000).

[47] E. Popov, M. Nevière, S. Enoch, and R. Reinisch, *Phys. Rev. B* **62**, 16100 (2000).

[48] S.I. Bozhevolnyi, J. Erland, K. Leosson, P.M.W. Skovgaard, and J.M. Hvam, *Phys. Rev. Lett.* **86**, 3008 (2001).

[49] A. Bouhelier, M. Beversluis, A. Hartschuh, and L. Novotny,*Phys. Rev. Lett.* **90**, 013903 (2003).

[50] A. Dechant and A.Y. Elezzabi, *Appl. Phys. Lett.* **84**, 4678 (2004).

[51] W.J. Fan, S. Zhang, B. Minhas, K.J. Malloy, and S.R.J. Brueck, *Phys. Rev. Lett.* **94**, 033902 (2005).

[52] A.V. Zayats, I.I. Smolyaninov, and A.A Maradudin, *Phys. Rep.* **408**, 131 (2005).

[53] Y. Ben-Aryeh, *International J. Quantum Information* **3**, 111 (2005).

[54] F.J. Garcia-Vidal, L. Martin-Moreno, H.J. Lezec, T.W. Ebbesen, *Appl. Phys. Lett.* **83**, 4500 (2003).

[55] A. Dogariu, M. Hsu, L.J. Wang, *Opt. Comm.* **220**, 223 (2003).

[56] K.R. Li, M.I. Stockman, D.J. Bergman, *Phys. Rev. Lett.* **91**, 227402 (2003).

[57] M.I. Stockman, S.V. Faleev, D.J. Bergman, *Phys. Rev. Lett.* **88**, 067402 (2002).

[58] S.V. Kukhlevsky, M. Mechler, L. Csapo, K. Janssens, *Phys. Lett. A* **319**, 439 (2003).

[59] S.V. Kukhlevsky, M. Mechler, *Opt. Comm.* **231**, 35 (2004).

[60] P. N. Stavrinou, *Phys. Rev. E* **68**, 06604 (2003).

[61] A.G. Borisov, S.V. Shabanov, arXiv:physics/0312103 (2003).

[62] M. Labardi, M. Zavelani-Rossi, D. Polli, G. Cerullo, M. Allegrini, S. De Silvestri, and O. Svelto, *Appl. Phys. Lett.* **86**, 031105 (2005).

[63] M. Born, E. Wolf, *Principle of Optics* (Pergamon, Oxford, 1980).

[64] S. Akturk, X. Gu, P. Gabolde, R. Trebino, *Opt. Express* **13**, 8642 (2005). *Opt. Express* **13**, 8642-8661 (2005)

[65] P.M. Paul, E.S. Toma, P. Breger, G. Mullot, F. Auge, P. Balcou, H.G. Muller, P. Agostini, *Science* **292**, 1689 (2001).

[66] M. Hentschel, R. Kienberger, Ch. Spielmann, G.A. Reider, N. Milosevic, T. Brabec, P.B. Corcum, U. Heinzmann, M. Drecher, F. Krausz, *Nature* **414**, 509 (2001).

[67] L. Novotny, D.W. Pohl, P. Regli, *J. Opt. Soc. Am. A* **11**, 1768 (1994).

[68] O. Mitrofanov, M. Lee, J.W.P. Hsu, L.N. Pleiffer, K.W. West, J.D. Wynn, J. Federici, *Appl. Phys. Lett.* **79**, 907 (2001).

[69] O. Mitrofanov, M. Lee, L.N. Pleiffer, K.W. West, *Appl. Phys. Lett.* **80**, 1319 (2002).

[70] S.V. Kukhlevsky, M. Mechler, *J. Opt. A* **5**, 256 (2003).

[71] J.W. Weaver, C. Krafka, D.W. Lynch, E.E. Koch, "Optical Properties of Metals", *Phys. Daten* **18-2**, 74 (1981).

[72] G.A. Massey, *Appl. Opt.* **23**, 658 (1984).

[73] N.A. McDonald, *IEEE Trans. Microwave Theory Tech. MTT* **20**, 689 (1972).

[74] A. Drezet, J.C. Woehl, S. Huant, *Europhys. Lett.* **54**, 736 (2001).

[75] C. Bergemann, H. Keymeulen, J.F. van der Veen, *Phys. Rev. Lett.* **91**, 204801 (2003).

[76] C. Girard, O.J.F. Martin, A. Dereux, *Phys. Rev. Lett.* **75**, 3098 (1995).

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN: 1-60021-194-1
Editor: Steven B. Larkin, pp. 43-61

Chapter 2

SINGLE EXPERIMENTAL SETUP FOR HIGH SENSITIVE ABSORPTION COEFFICIENT AND OPTICAL NONLINEARITIES MEASUREMENTS

Debabrata Goswami*
Department of Chemistry, Indian Institute of Technology,
Kanpur, India

Abstract

Accurate knowledge of absorption coefficient of a sample is a prerequisite for measuring the third order optical nonlinearity of materials, which can be a serious limitation for unknown samples. We introduce a method, which measures both the absorption coefficient and the third order optical nonlinearity of materials with high sensitivity in a single experimental arrangement. We use a dual-beam pump-probe experiment and conventional single-beam z-scan under different conditions to achieve this goal. We also demonstrate a counterintuitive coupling of the non-interacting probe-beam with the pump-beam in pump-probe z-scan experiment.

I. Introduction

Development of high power laser sources has motivated an extensive research in the study of nonlinear optical properties and optical limiting behavior of materials. There exits a continued effort in making sensitive measurements on absorption coefficient and nonlinear coefficients, however, most of the experimental techniques are focused on measuring one or the other of these two important parameters. A variety of interferometric methods [1],[2], degenerate four wave mixing [3], nearly degenerate three wave mixing [4] and beam distortion measurement [5], have been used for measuring the nonlinear refractive index. One of the most important techniques to measure nonlinear refractive index was shown by Sheik Bahaei et al. [6]. This technique is simple and versatile yet is highly sensitive. However, an accurate knowledge of

* E-mail address: dgoswami@iitk.ac.in

absorption coefficient (α_0) is necessary for the use of this technique, which is a serious limitation for unknown samples. An easy way to measure α_0 is to use the Beer's law [7-8], which operates in the linear absorption regime and has limited sensitivity. More sensitive methods have been developed, of which the technique [9-15] using thermal lens (TL) effect is perhaps the simplest and the most effective. In this method, a lens focuses the laser beam into the sample resulting in a temperature gradient, which in turn produces a spatial gradient in refractive index. The relative change in transmittance of the laser beam can then be measured, after passing through an aperture, with the help of a detector [9-11]. Shen et al. [12-14] introduced a pump probe laser scheme under mode-mismatched and mode-matched conditions to improve sensitivity of the TL method. More recently, Marcano et al. [15] have used this method to measure the absorption coefficient of water with high accuracy.

A single experimental technique to measure both the parameters, however, is yet to emerge, which could be of significance in the study of new materials. In this paper we introduce a single experimental technique to measure both the parameters. Our aim has been to measure the absorption coefficient (α_0) as well as the real and imaginary parts of third-order optical nonlinearity ($\chi^{(3)}$) with high sensitivity in a transparent sample using a single experimental setup. We also show how mid-IR absorption in water (at 1560nm) manifests itself as a minute nonlinear absorption coefficient in the near-IR transmission window (at 780nm). Such high sensitive experiments have become possible due to the ultrahigh sensitivity provided by a stable ultrafast laser operating at both the above wavelengths.

Our technique is a modification of the well-known z-scan technique introduced by Shiek Bahaei et al. [6] in 1990, where one measures the change in transmittance of a focused laser beam through sample that is being moved through the focal point of the lens. Since we are dealing with Gaussian optics when the beam is passing through lens and sample, we will use the Gaussian optics formalism. Subsequently, we will discuss the z-scan theory and finally we will discuss the modifications that we have introduced in addition to the new experimental results and discussions.

II. Background

a. The Gaussian Beam in a Homogeneous Medium

In most laser application it is necessary to focus, modify or shape the laser beam by using lenses or other optical components. In general, laser beam propagation can be approximated by assuming that the laser beam has an ideal Gaussian intensity profile corresponding to TEM_{00} mode. Using Maxwell equation in an isotropic, charge free medium one can derive the wave equation [16],[17]:

$$\nabla^2 E + k^2 E = 0 \tag{1}$$

where $k^2 = \mu_0 \varepsilon_0 \omega^2 = 2\pi / \lambda$. Let us assume a solution whose transverse dependence is only on $r = \sqrt{x^2 + y^2}$, which will enable us to replace ∇^2 by $\nabla^2{}_t + \dfrac{\partial^2}{\partial z^2}$ in Eq.(1). We

consider nearly plane wave situation where the flow of energy is along a single direction (e.g. z), and therefore the electric field, *E*, is:

$$E = \Psi(x, y, z)e^{-ikz} \tag{2}$$

Substituting these in Eq.(1) we derive

$$\nabla^2{}_t \Psi - 2ik\frac{\partial \Psi}{\partial z} = 0 \tag{3}$$

where we have assumed that longitudinal variation is slow, such that $k\frac{\partial \Psi}{\partial z} \geq \frac{\partial^2 \Psi}{\partial z^2} \leq k^2 \Psi$ is valid. In the next step, we take Ψ of the form

$$\Psi = \exp\{-i[P(z) + \frac{1}{2}Q(z)r^2]\} \tag{4}$$

By substituting Eq.(4) in Eq.(3) we derive

$$-Q^2 r^2 - 2iQ - kr^2 Q' - 2kP' = 0 \tag{5}$$

If this equation is to hold true for all *r*, then the coefficients of different powers of *r* must be equal to zero, which leads to:

$$\begin{aligned} Q^2 + kQ' &= 0 \\ P' &= -i\frac{Q}{k} \end{aligned} \tag{6}$$

For solving this differential equation we introduce a function *S(z)*, such that

$$Q = k\frac{S'}{S} \tag{7}$$

Replacing the value of *Q* in Eq.(6) with the relation from Eq.(7), we get

$$(\frac{kS'}{S})^2 + k[\frac{kS''S - S'^2}{S^2}] = 0 \tag{8}$$

which implies $S'' = 0$, and consequently,

$$S' = a \text{ and } S = az + b \tag{9}$$

where *a* and *b* are arbitrary constant. Replacing the values from Eq.(9) in Eq.(7), we get

$$Q = k\frac{a}{az+b} \tag{10}$$

It is more convenient to deal with a parameter *q*, where $q(z) = \frac{k}{Q(z)}$. So that we can rewrite Eq.(9) as:

$$q = z + q_0 \tag{11}$$

where q_0 is a constant ($q_0 = b/a$). From Eq.(6) and (11) we have

$$P' = \frac{-i}{q} = \frac{-i}{z+q_0} \quad \Rightarrow \quad P(z) = -i\ln(1+\frac{z}{q_0}) \tag{12}$$

where the arbitrary constant of integration is chosen as zero. The constant of integration will modify the phase of the field solution, Eq.(2). Since the time origin is arbitrary, the phase can be taken as zero. Combining Eqs.(11) and (12) in Eq.(4), we obtain

$$\Psi = \exp\{-i[-i\ln(1+\frac{z}{q_0}+\frac{k}{2(q_0+z)}r^2]\} \tag{13}$$

We take the q_0 to be purely imaginary and express in terms of new constant w_0 as $q_0 = i\frac{\pi {w_0}^2}{\lambda}$. By substituting q_0 in Eq.(13) and defining following parameters:

$$w^2(z) = {w_0}^2[1+(\frac{\lambda z}{\pi {w_0}^2})^2] = {w_0}^2(1+\frac{z^2}{{z_0}^2}) \tag{14}$$

$$R(z) = z[1+(\frac{\pi {w_0}^2}{\lambda z})^2] = z(1+\frac{{z_0}^2}{z^2}) \tag{15}$$

$$\eta(z) = \tan^{-1}(\frac{\lambda z}{\pi {w_0}^2}) = \tan^{-1}(\frac{z}{z_0}) \tag{16}$$

where $z_0 = \frac{\pi {w_0}^2}{\lambda}$. We can write Eq.(2) as:

$$E = E_0\frac{w_0}{w(z)}\exp\{-i[kz-\eta(z)]-r^2[\frac{1}{w^2(z)}+\frac{ik}{2R(z)}]\} \tag{17}$$

and we can also write

$$\frac{1}{q(z)} = \frac{1}{R(z)} - i\frac{\lambda}{n\pi w^2(z)} \tag{18}$$

which is the fundamental Gaussian beam solution. The parameters $w(z), w_0$ are beam spot size and minimum spot size at *z=0* and the parameter *R(z)* is the radius of curvature of the spherical wavefronts at *z*. Our aim is to calculate spot size of the beam when it passes through a thin lens of focal length *f* as shown in the Fig.1. Since, at the input plane (1) of Fig.1, $w = w_{01}$ and $R_1 = \infty$, we can write using Eq.(18) the following relation:

$$\frac{1}{q_1} = \frac{1}{R_1} - i\frac{\lambda}{\pi {w_{01}}^2} = -i\frac{\lambda}{\pi {w_{01}}^2} \tag{19}$$

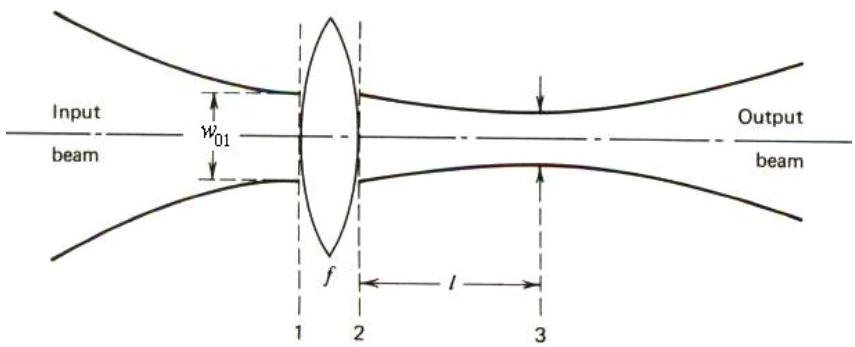


Figure 1. Schematic of a Gaussian beam of waist (w_{01}) propagating through a thin lens of focal length *f*. The beam focuses at distance *l* from the lens with a Rayleigh range of *RR*.

Similarly, at the output plane (2) of Fig.1, we get:

$$\frac{1}{q_2} = \frac{1}{q_1} - \frac{1}{f} = -i\frac{\lambda}{\pi {w_{01}}^2} - \frac{1}{f} \tag{20}$$

or

$$q_2 = \frac{1}{-\frac{1}{f} - i\frac{\lambda}{\pi w_{01}^2}} \tag{21}$$

Finally, at plane (3), $q_3 = q_2 + l$, and the output beam waist, $R_3 = \infty$. Thus, the location of new waist is

$$l = \frac{f}{1+\left(\frac{f}{z_0}\right)^2} \tag{22}$$

and the minimum spot size in focal point is equal to:

$$w_{03} = w_{01}\frac{\frac{f}{z_0}}{\sqrt{1+\left(\frac{f}{z_0}\right)^2}} \tag{23}$$

where $z_0 = \frac{\pi {w_{01}}^2}{\lambda}$. The other parameter of interest is the Rayleigh range (*RR*), which is the axial distance from the point of minimum beam waist (w_0) to the point where the beam diameter has increased to $\sqrt{2}\, w_0$ in the region of a Gaussian beam focus by a diffraction-limited lens. This is given by the expression:

$$RR = \frac{\pi w_0^2}{\lambda} \tag{24}$$

We are using a lens with focal length f=75cm and w_{01}=2.77 mm (which is measured by integrating the residual intensity that is measured by translating a knife edge across the beam (Fig.2). With this background on Gaussian optics, we now discuss the technique of Sheik Bahei et al. [6] in the following section.

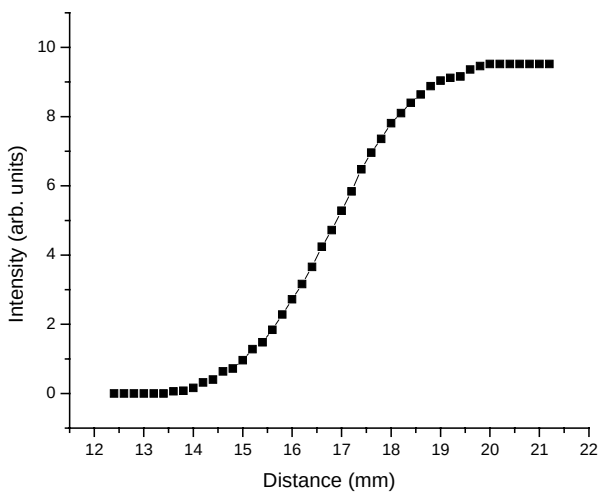


Figure 2. (continued on next page)

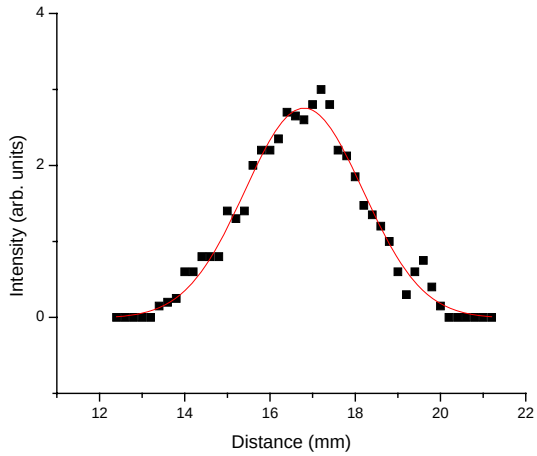


Figure 2. Measurement of laser beam size by translating a knife-edge across the laser beam falling on a photodiode. (a) A plot of the intensity of the light (along y-axis) coming to the photodiode as the translating knife-edge (distance along x-axis) lets out more and more of the incident laser beam into the photodiode. (b) A derivative of the data in Fig.2a gives the beam size in the lower plot, which fits to a Gaussian beam waist of 2.77mm.

b. The Z-scan Technique

The technique introduced by Bahaei et al. [6] is now popularly known as the z-scan technique [18] as it involves the motion of the sample in the sample across the focal point of laser beam along the direction of propagation of the laser beam (Fig.3). Assuming Gaussian beam optics as discussed in the previous section, this experiment allows an intensity scan of the irradiated sample, and provides information about the nonlinearity in the sample. The typical z-scan is performed by translating the sample along the z axis from one side of the focus to the other (Fig.3). This results in changing the spot size of the incident beam on the sample to a minimum at the focus and then increasing again on crossing the focus. Correspondingly, the intensity of incident light increases on approaching the focus till a maximum at the focus is reached and then reduces on moving away from the focus. Thus, the overall purpose of the experiment is to determine the variation in transmission as the incident intensity changes by translation along the z-axis. The change in the transmittance of the focusing Gaussian beam in a medium is recorded as a function of position of medium. The transmitted beam is collected either completely (which is called the open aperture case) or through a finite aperture (A) as shown in Fig.3.

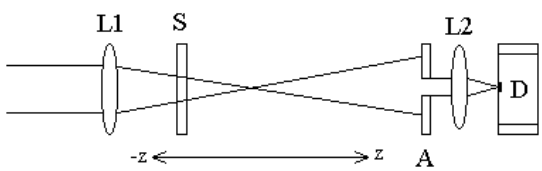


Figure 3. Schematic of typical z-scan experimental setup where a sample (S) is scanned across a laser beam that is being focused through a lens of focal length (L1) and is collected through an aperture (A) and a lens of focal length (L2) into the detector (D).

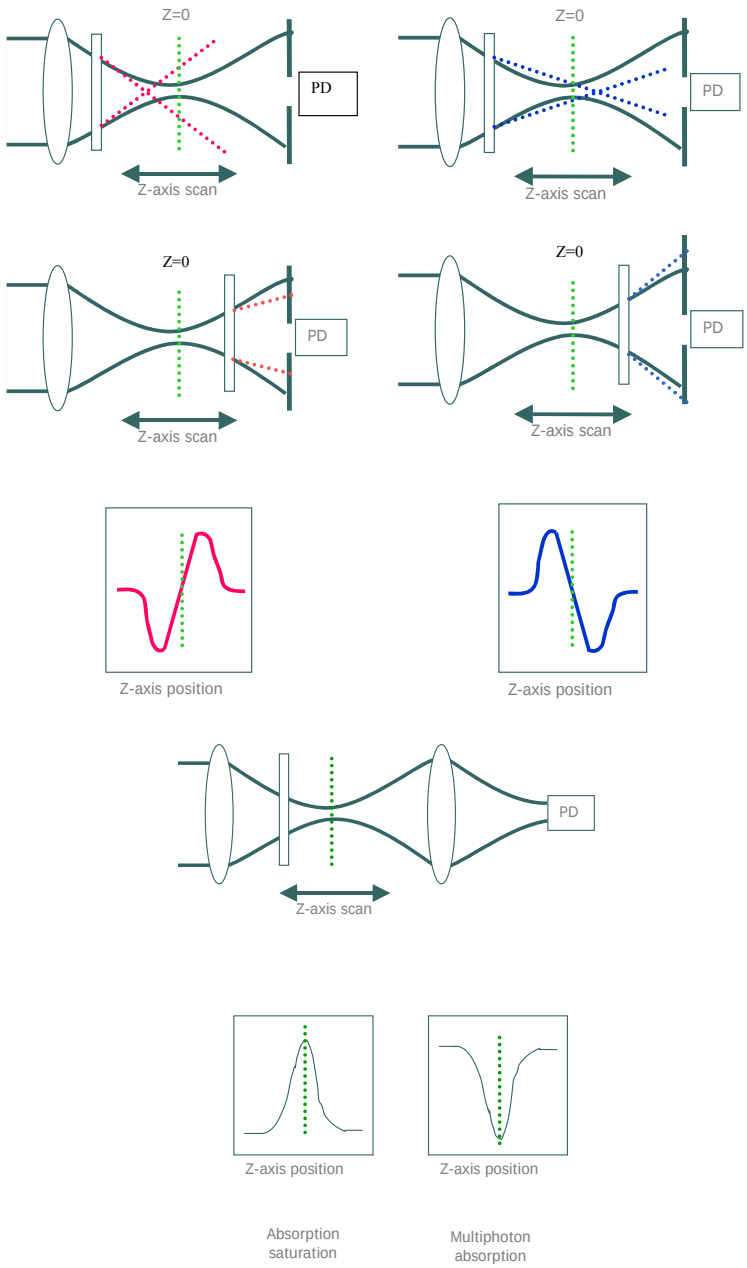


Figure 4. Typical experimental results of z-scan experiments performed under various conditions. (a) For materials with positive refractive index with 40% closed aperture in front of the detector, the typical increased signal followed by reduced signal at the detector is due to the characteristic beam bending as shown. (b) For materials with negative refractive index with 40% closed aperture in front of the detector, the results are opposite to that of 4a due to the opposite bending of the beam. (c) Experiments with open aperture result in a characteristic dip or a peak at the focal point position of the scan either due to multiphoton absorption or absorption saturation.

Let us first discuss schematically, a simple case of a thin sample with negative nonlinear refractive index when the aperture is closed (A=0.5, which means just 50% of the beam

passes through the aperture). When it moves in the z direction it can act as a thin lens with variable focal length. If we start the scan from –z (far from focal length), where the nonlinear refraction is negligible, the transmittance remains relatively constant. As the sample moves closer to the focus, the beam irradiance increases because of self-focusing of the beam will tend to collimate the beam and cause a beam narrowing at the aperture which results in an increase in the measured transmittance (Fig.4a). As the scan continues and sample passes the focal plane, the self-defocusing phenomena will occur. This will broaden the beam at the aperture and a corresponding decrease in transmittance will continue until the sample reaches +z (that is sufficiently far from focus) such that the transmittance becomes linear.

If we open the aperture (A=1) and do the same scan again from –z direction, the transmittance will increase till focal point and as discussed above, it will decrease to the linear case when the sample moves away from focal point to the +z direction (Fig.4c).

Thus, the open aperture case scan gives information on purely absorption nonlinearity while a close aperture case scan contains information about the absorption and dispersion nonlinearity. In case of materials with positive refractive index the story is the reverse of the above cases (Fig.4b and Fig.4d). Induced beam focusing and defocusing of this type have been observed during nonlinear refractive measurement of some semiconductors [10],[11].

Let us now consider the above qualitative discussion mathematically. We consider a sample with third order nonlinearity where the index of refraction is equal to:

$$n = n_0 + \frac{n_2}{2}|E|^2 = n_0 + \gamma I \tag{24}$$

where n_0 is the linear refraction index, E is the peak electric field (derived in Eq.(17)). I is the irradiance of the laser beam within the sample, n_2 and γ are related through the conversion formula $n_2(esu) = \frac{n_0 c}{40\pi}\gamma(m^2 / W)$ ($c(m/s)$ is speed of light).

Since our sample is thin we can approximate the Gaussian beam is parallel inside the sample. We want to calculate the phase shift of the beam when it passes through the sample. The amplitude $\sqrt{I}$ and phase of electric field in the slowly varying envelop approximation as a function of z' (propagation depth in the sample), are given by two pair equations [6]:

$$\frac{d\Delta\phi}{dz'} = \Delta n(I)k \tag{25}$$

$$\frac{dI}{dz'} = -\alpha(I)I \tag{26}$$

$\alpha(I)$ contains all linear and nonlinear absorption. Using Eq.(24) we can solve the coupled Eqs. (25) and (26) together to derive $\Delta\phi$ at the exit surface of the sample as a function of stage z and radial variation of incident beam.

$$\Delta\phi(z,r,t) = \Delta\phi_0(z,t)\exp(-\frac{2r^2}{w^2(z)}) \tag{27}$$

with

$$\Delta\phi_0(z,t) = \frac{\Delta\Phi_0(t)}{1+z^2/z_0{}^2}. \tag{28}$$

$\Delta\Phi_0$, the on axis (r =0) phase shift at focus (z =0) is defined as :

$$\Delta\Phi_0(t) = k\Delta n_0(t) L_{eff} \tag{29}$$

where $L_{eff} = (1-e^{-\alpha L})/\alpha$, L is sample length and α linear absorption coefficient and $\Delta n_0 = \gamma I_0$ (I_0, I at r=0 and z=0) . Now the electric field which is coming out from sample will look like

$$E_{out}(z,r,t) = E_{in}(z,r,t)e^{-\alpha L/2}e^{i\Delta\phi(z,r,t)} \tag{30}$$

where $E_{in}(z,r,t)$ is the same as in Eq.(17). Now we are going to derive the electric field in aperture. A method which is called "Gaussian decomposition" (GD) and is given by Weaire et al. [12] can be used to obtain the far field pattern of the electric field at the aperture plane. They decompose the E_{out} into a summation of Gaussian beams through a Taylor series expansion therefore from Eqs.(27) and (30)

$$e^{i\Delta\phi(z,r,t)} = \sum_{m=0}^{\infty} \frac{[i\Delta\phi_0(z,t)]^m}{m!} e^{-2mr^2/w^2(z)} \tag{31}$$

therefore

$$E_a(z,r,t) = E(z,r=0,t)e^{-\alpha L/2} \sum_{m=0}^{\infty} \frac{[i\Delta\phi_0(z,t)]^m}{m!} \frac{w_{m0}}{w_m} \times \exp(-\frac{r^2}{w_m{}^2} - \frac{ikr^2}{2R_m} + i\theta_m). \tag{32}$$

defining d as distance from sample to aperture and $g = 1 + d/R(z)$ (R(z) is defined in Eq.(15)) all parameters in Eq.(32) are expressed as [13]:

$$w_{m0}{}^2 = \frac{w^2(z)}{2m+1}, \quad d_m = \frac{kw_{m0}{}^2}{2}, \quad w_m{}^2 = w_{m0}{}^2[g^2 + \frac{d^2}{d_m{}^2}]$$

$$R_m = d[1 - \frac{g}{g^2 + d^2/d_m{}^2}]^{-1} \text{ and } \theta_m = \tan^{-1}[\frac{d/d_m}{g}] \tag{33}$$

The GD method is very useful for small phase distortions detected with Z-scan therefore only few terms of Eq.(32) are needed. Now we can calculate transmittance power through the aperture :

$$P_T(\Delta\Phi_0(t)) = c\varepsilon_0\pi\int_0^{r_a}\left|E_a(r,t)\right|^2 rdr \tag{34}$$

Including the pulse temporal variation, the normalized Z-scan transmittance can be calculated as

$$T(z) = \frac{\int_{-\infty}^{\infty} P_T(\Delta\Phi_0(t))dt}{S\int_{-\infty}^{\infty} P_i(t)dt} \tag{35}$$

where $P_i(t) = \pi w_0{}^2 I_0(t)/2$ is the instantaneous input power (within the sample) and $S = 1-\exp(-2r_a{}^2/w_a{}^2)$ is the aperture linear transmittance (w_a is the beam radius at the aperture). In above discussion we have assumed the effect of third order nonlinearity only and that no absorptive nonlinearity effects that arise from multiphoton or saturation absorption exist. Multiphoton absorption suppress the peak and enhance the valley, while saturation produce the opposite effect [6],[9].

c. The Dual-beam Technique

Shen et al. [12-14] introduced a pump probe laser scheme under mode-mismatched and mode-matched conditions to make sensitive TL measurements. In such dual beams experiments, one of laser beams is essentially probing the effect of the TL caused by the pump beam by scanning across its focus. This results in an effective z-scan of the probe beam across a TL generated by a focusing pump beam. The closed aperture case of this scenario is shown schematically in Fig.5 which has been used by Marcano et al. [15] to measure the absorption coefficient of water with high accuracy.

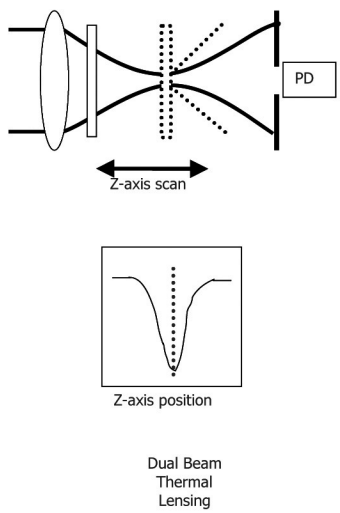


Figure 5. Typical experimental results of dual-beam thermal lens experiment where the probe beam is collected through a 40% closed aperture.

Mathematically, as in Ref. [15], we can also use the expression of Shen et al. [12-14], who have derived an expression for the TL signal using diffraction approximation for Gaussian beams in steady state case as:

$$S(z) = [1 - \frac{\theta}{2}\tan^{-1}(\frac{2mV}{1+2m+V^2})]^2 - 1 \tag{36}$$

where

$$\begin{aligned} m &= (\omega_p / \omega_o)^2, \\ V &= (z - a_p)/z_p + [(z - a_p)^2 + z_p{}^2]/[z_p(L - z)], \\ \omega_{p,o} &= b_{p,o}[1 + (z - a_{p,o})^2 / z^2{}_{p,o}]^{1/2}, \\ \theta &= -P_o \alpha_0 l(ds/dT)/\kappa\lambda_p \end{aligned} \tag{37}$$

z is sample position with respect to the focal point, a_p, a_o, z_p, z_o and b_p, b_o, are position of the waists, the confocal parameters and the beam radius for the probe and pump beams, respectively. λ_p is the wavelength of the probe beam, κ is thermal conductivity coefficient of the sample. L is the detector position, P_o is the total power of the pump beam and l is the sample thickness.

Both continuous and pulsed lasers have been effectively used for z-scan experiments that have relied on these mathematical principles discussed here [19]. These discussions in this section form the basis of our present work that we present hereafter. We explore the open-aperture dual beam TL experiments and achieve our single experimental setup to achieve high sensitive measurements.

III. Present Work

Our experiments are variation from the conventional z-scan discussed in the above section. We not only use the single beam technique as mentioned in the previous section, but with very simple changes in the experimental set-up, make measurements corresponding to the dual beam z-scan experiments. We will now concentrate more on our actual experimental scheme and the results and discussions arising thereafter.

a. Experiment

Our experimental scheme involves a sub-100 femtosecond mode-locked Er:doped fiber laser (IMRA Inc.) operating at a repetition rate of 50MHz and provides the fundamental (1560nm)

and its second-harmonic wavelength (780nm) simultaneously as a single output. The pulse characteristics of the laser pulses are shown in Fig.6. Either we use both the wavelengths from the laser simultaneously or separate the two copropagating beams with the help of a dichroic beamsplitter and use each of them independently. We scan the sample through the focal point of a 75cm focusing lens and this allows a smooth intensity scan for either/or both of the wavelengths. Care has been taken to make sure that there is no effect of the laser beams on the cuvette alone by conducting an empty cuvette experiment. A silicon photodetector (Thorlab: DET210) is used for the 780nm beam detection, while an InGaAs photodetector (Acton Research) is used for the 1560nm beam detection.

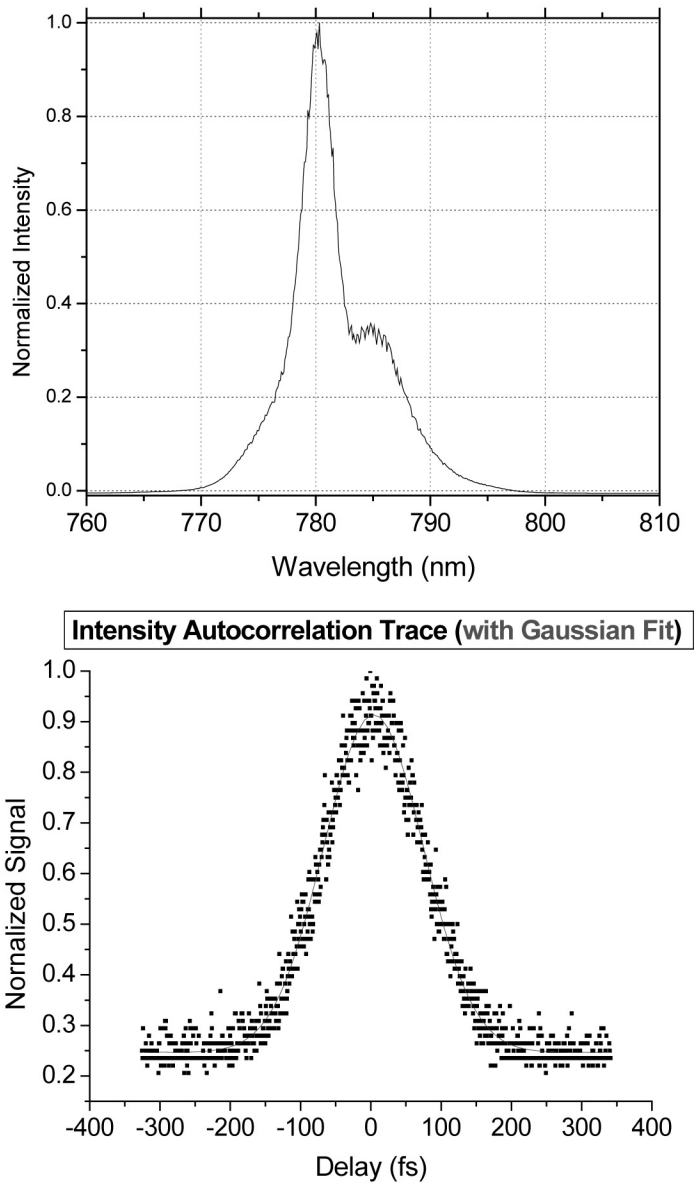


Figure 6. continued on next page

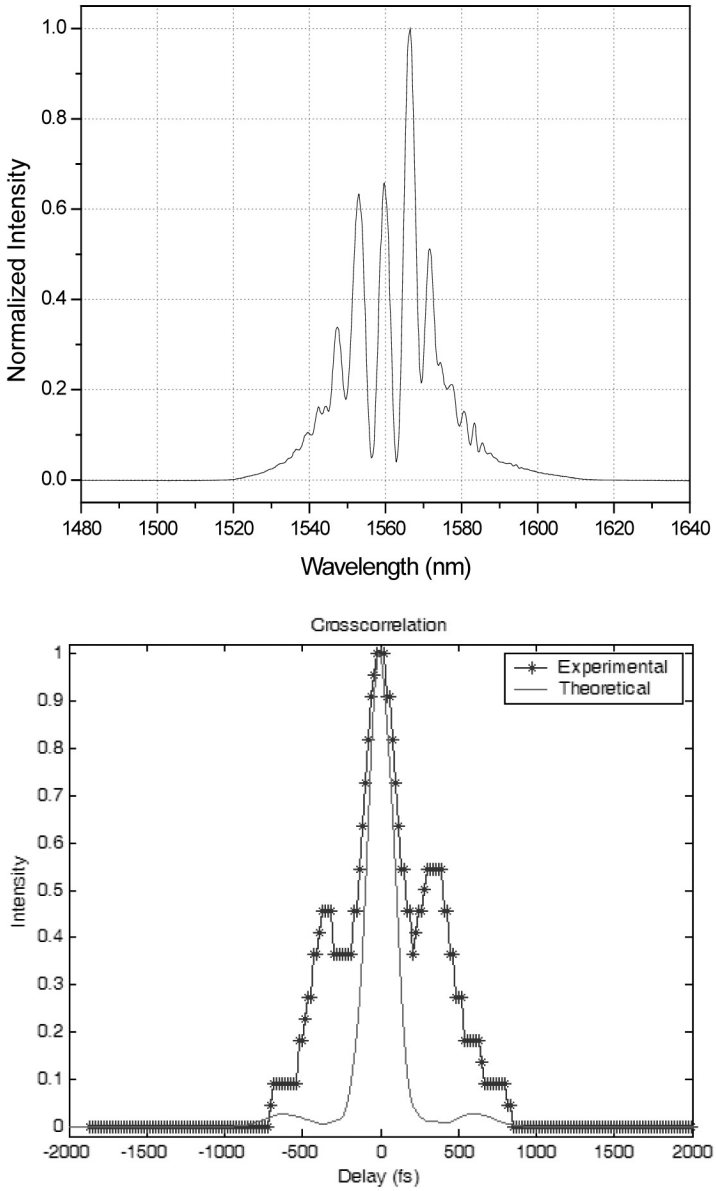


Figure 6. The dual wavelength femtosecond fiber laser (IMRA Inc., Femtolite-C) pulse characteristics as measured in our laboratory. (a) The spectra at the center wavelength of 780nm measured through the SP-150 monochromator (Acton Research Co.) into a silicon photodetector (Thorlab: DET210). (b) The pulsewidth of the 780nm pulse measured through a non-collinear autocorrelation using a speaker as the delay arm into a second-harmonic BBO crystal, which is detected into a silicon photodetector (Thorlab: DET210). The Gaussian fit to the autocorrelation trace provides a pulsewith of 90fs for the 780nm laser pulses. (c) The spectra at the center wavelength of 1560nm measured through the SP-150 monochromator into an InGaS detector (both Acton Research Co.) (d) The pulsewidth of the 1560nm pulse measured through collinear cross-correlating with the 780nm beam into a second-harmonic BBO crystal, which is detected into a silicon photodetector (Thorlab: DET210). Using a Gaussian fit and deconvoluting the 90fs pulse of 780nm results in a pulsewidth of 120fs for the 1560nm pulse.

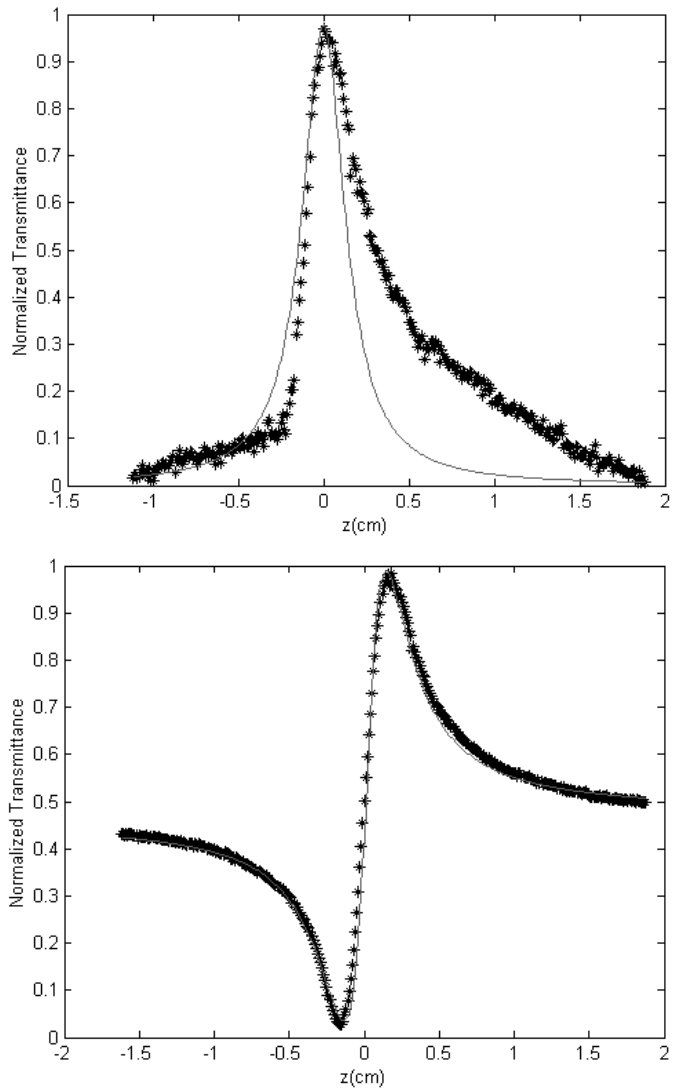


Figure 7. (a) Measured z-scan of a 16mm thick double distilled water using 120fs pulses at λ=1560nm (diamond) and theoretical fit (solid line) for fully open aperture (experimentally measured transmittance is normalized to unity).(b) Measured z-scan of a 16mm thick double distilled water using 120fs pulses at λ=1560nm (diamond) and theoretical fit (solid line) for 40% closed aperture (experimentally measured transmittance is normalized to unity).

We find that the 1560nm beam produces changes in the relative transmission of the laser beam at different intensities as the sample is scanned through the lens focus depending whether we collect all the light or only central 40% of the transmitted light (Fig.7). These results in Fig.7 essentially represent the z-scan technique of Sheik Bahaei et al. [6] to measure the real and imaginary parts of the third-order optical nonlinearity ($\chi^{(3)}$). However, the 780nm beam does not produce any effect even at our peak powers at the focal point of the laser as is expected from negligible absorption at 780nm (Fig.8). This enables us to use the 780nm wavelength as the non-interacting probe beam for the subsequent dual-beam

experiments where we use both the wavelengths from the laser simultaneously. Since our 75cm lens focuses the 780nm probe beam to its minimal spot size position 0.4mm ahead of the pump beam of 1560nm, this is a mode-mismatched pump-probe experiment. However, the focal spot size of 9μm for 780nm is 15μm smaller than the corresponding 1560nm spot size at its own focus and from Eq.(24) the Rayleigh range for 780nm is 0.32mm and for 1560mm it is 1.15mm. Thus, the 780nm laser volume is always confined within the 1560nm laser beam volume when both the beams are used simultaneously from the laser and can act as an effective probe.

b. Results and Discussion

Saturation leads to thermal lensing as is seen in the pump-probe data in water with 1560nm as the pump laser and 780nm as the probe beam. The absorption coefficient of water at 780nm is so small (Fig.8) that at our intensities, it is an effective probe beam. The thermal lens effect of the pump 1560nm beam results in a temperature gradient that in turn produces a spatial gradient in refractive index, which is depicted in the relative change in transmittance of the probe beam. Fig.9 shows results of the experiment when we collect only the 780nm probe beam by separating the 1560nm pump beam after passing through the sample at different intensities as the sample is scanned through the lens focus depending whether we collect only central 40% of the transmitted light (Fig.9a) or all the light (Fig.9b). The thermal lens effect of the pump beam resulting in a corresponding spatial gradient in refractive index is depicted in the relative change in transmittance of the probe beam in Fig.9a. Such thermal lensing (TL) effect can be used to determine the absorption coefficient (α_0) of the sample at the pump wavelength very accurately [15].

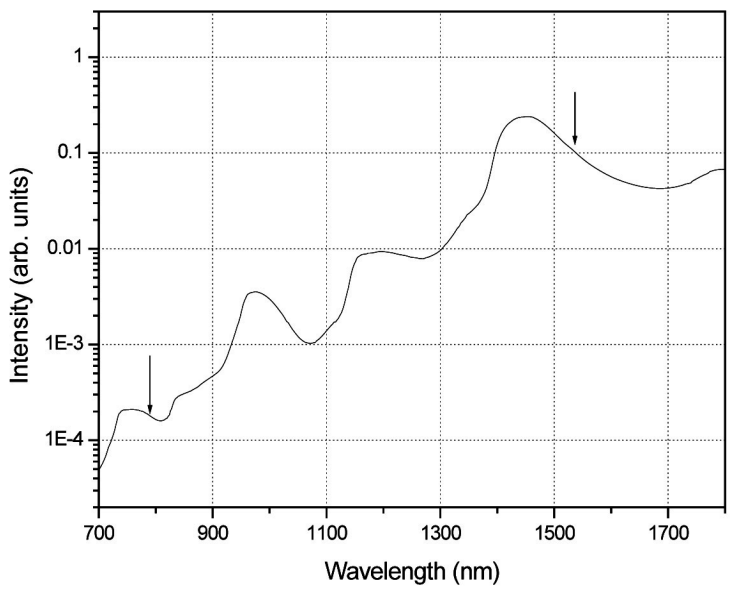


Figure 8. Water spectra covering the 780nm and 1560nm range of wavelengths as a semi-Log plot that shows the absorption differences between the two marked wavelengths.

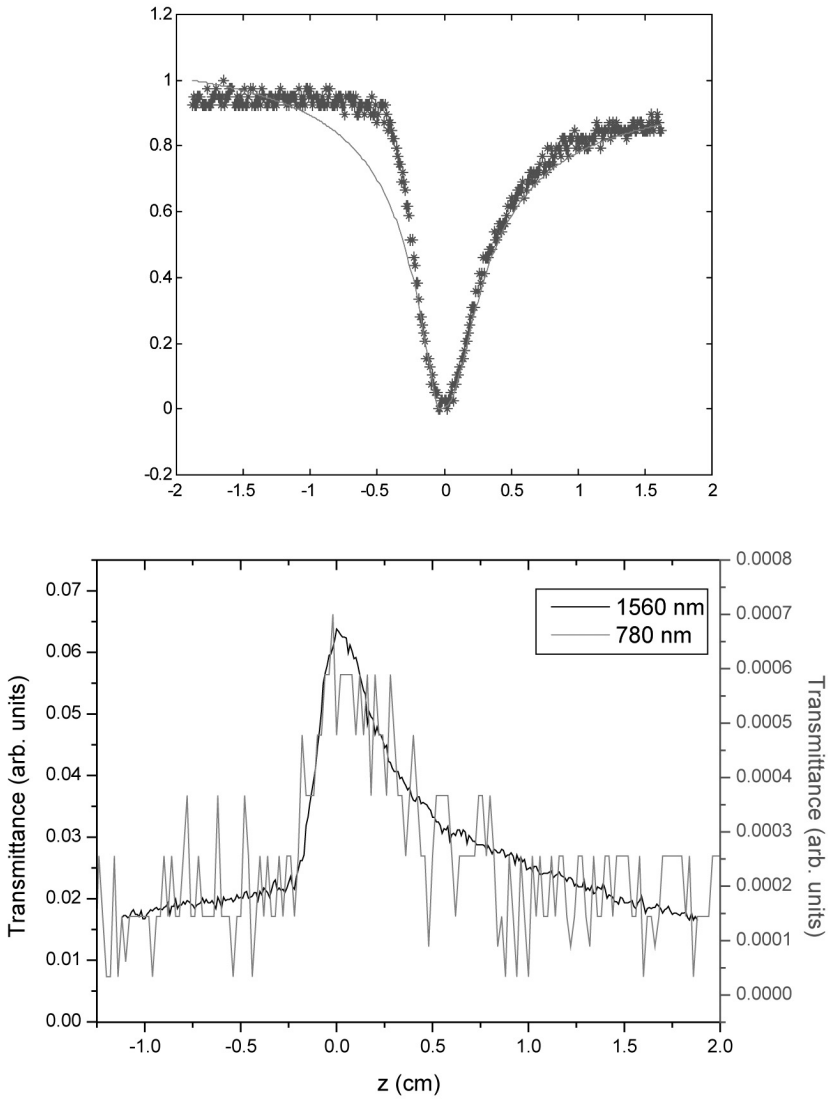


Figure 9. (a) Measured z-scan transmittance of 90fs pulses of λ=780nm as a probe through a 16mm thick double distilled water being irradiated with 120fs of λ=1560nm as pump (diamond) and theoretical fit (solid line) in 40% closed aperture (experimentally measured transmittance is normalized to unity). (b) Measured z-scan transmittance of 90fs pulses of λ=780nm as a probe through a 16mm thick double distilled water being irradiated with 120fs of λ=1560nm as pump (raw transmittance data is presented to illustrate sensitivity of our measurements). We replot the open aperture data for 1560nm alone from Fig.7a on this same plot to show that both essentially have same features except their two orders of magnitude difference in their signal levels and hence the same fit would suffice.

The solid line in the Fig.9a is the result of a theoretical fit to Eq.(36). This fit gives the value of phase shift, θ = 9.957, which when substituted in Eq.(37) with the parameters $ds/dT = -9.1\times10^{-5}\ K^{-1}$ and $\kappa = 0.598\times10^{-2} WK/cm$ for pure water [20], we get the calculated value for α_0 as 10.6327cm^{-1} for the 1560nm beam which is within 1% of reported literature data [21]. While this fairly large value of α_0 need not be measured with such a

sensitive technique, it serves as the proof-of-principle for the experimental setup and the accuracy of the experimental measurements.

Our experimental results discussed above also enables us to determine the nonlinear absorption coefficients of water. For nonlinear materials the index of refraction *n* is expressed in terms of nonlinear n_2 or γ through the relation: $n = n_0 + \frac{n_2}{2}|E|^2 = n_0 + \gamma I$, where n_0 is the linear index of refraction, *E* is the peak electric field (cgs) and *I* (MKS) is the intensity of the laser beam inside the sample. n_2 and γ are related to each other as: $n_2(esu) = \frac{cn_0}{40\pi}\gamma(m^2/W)$, where *c*(m/s) is the speed of light in vacuum. The third order optical susceptibility is considered to be a complex quantity: $\chi^{(3)} = \chi_R{}^{(3)} + \chi_I{}^{(3)}$. The real and imaginary parts are related to the γ and β respectively [22] where β is the nonlinear absorption coefficient and is defined as $\alpha(I) = \alpha_0 + \beta I$. We fit the fully open aperture data with 1560nm wavelength alone from Fig.7a (solid line in Fig.7a) to a theoretical expression derived by solving the differential equation [23] for the transmitted light through a sample of thickness *l*

$$T(z) = \eta + \frac{\beta I_0 l}{(1 + z^2 / z^2{}_0)} \tag{38}$$

where $z_0 = k w_0{}^2 / 2$, *k* and w_0 are the wave vector and the minimum spot size in the focal point respectively, while η and β are the fitting parameters. The best fit gives the value of β = -2.58 cm/GW.

In Fig.7b, the valley-peak structure representing the 40% closed-aperture data for 1560nm suggests a self-focusing effect inside the sample. The Ryleigh range (Z(r)) for 1560mm is 1.15mm. From Fig.7b, valley to peak separation at 1560nm is 4mm, which is 3.4×Z(r) indicating that all the effects at 1560nm are thermal in nature [24]. So we use the Gaussian Decomposition method to fit this closed aperture z-scan data quite convincingly (Fig.7b, solid line fit to the raw data), and we obtain γ=1.57× 10^{-3} cm^2 /GW, which is proportional to n_2=4.9×10^{-12} esu. Thus α_0, β and n_2 values of the water sample at 1560nm wavelength are determined.

As mentioned earlier, when only the 780nm beam passes through the sample, there is a constant signal and we do not see any peak or valley. The absorption coefficient of water at 780nm is so small that at our intensities, there is no saturation effect. However, when 1560nm beam is simultaneously present, it affects the propagation of the 780nm beam. As the 1560nm beam starts to saturate the sample at its focal point, the 780nm beam also experiences a saturated environment, whereby its transmittance increases at its focal point (Fig.9). This statement is further reinforced by the fact that the open-aperture plot of 1560nm beam alone from Fig.7a essentially has the same behavior as is evident when plotted along with the results from the 780nm probe results in Fig.9b. Essentially, as the 1560nm beam starts to saturate the sample at its focal point, the 780nm beam also experiences a saturated environment, whereby its transmittance increases at its focal point and shows an identical

transmission behavior although the signal level is two orders of magnitude lower. Such a result indicates the thermal capacity of water that can affect the spectroscopic behavior of water. Thus, this technique can be used to measure the nonlinear absorption coefficient of materials, whose linear absorption coefficients are extremely small. We can use the same fitting of the experimental data using a theoretical expression given by Eq.(38) to result in the expected value of β = -2.58 cm/GW as was found for the 1560nm data.

Acknowledgements

The author thanks the Ministry of Information Technology and the Dept. of Science and Technology, Govt. of India and the International SRF Program of the Wellcome Trust (UK) for funding the research results presented here.

References

[1] M.J. Weber, D. Milam and W.L. Smith, *Opt. Eng.* **17**, 463 (1978).
[2] M.J. Moran, C.Y. She, R.L. Carman, *IEEE J. Quantum Electron.* **QE-11**, 259 (1975).
[3] S.R. Fribrg and P.W. Smith, *IEEE J. Quantum Electron.* **QE-23**, 2089 (1987).
[4] R. Adair, L.L. Chase and S.A. Payne, *J. Opt. Soc. Am. B* **4**, 875 (1987).
[5] W.E. Williams, M.J. Soileau and E.W. Van Stryland, *Opt. Comm.* **50**, 256 (1984).
[6] M. Sheik-Bahaei, A.A. Said, T. Wei, D.J. Hagan, E.W. Van Stryland, *IEEE J. Quantum Electron.* **26**, 760 (1990).
[7] G.T. Fraser, A.S. Pine, W.J. Lafferty and R.E. Miller, *J. Chem. Phys.* **87**, 1502 (1987).
[8] H. Petek, D.J. Nesbitt, D.C. Darwin and C.B. Moore, *J. Chem. Phys.* **86**, 1172 (1987).
[9] N.J. Dovichi and J.M. Harris, *Anal. Chem.* **51**, 728 (1979).
[10] G.R. Long and S.E. Bialkowski, *Anal. Chem.* **56**, 2806 (1984).
[11] N.J. Dovichi and J.M. Harris, *Anal. Chem.* **53**, 106 (1981).
[12] J. Shen, M.L. Baesso and R.D. Snook, *J. Appl. Phys.* **75**, 3738 (1994).
[13] J. Shen, R.D. Lowe and R.D. Snook, *Chem. Phys.* **165**, 385 (1992).
[14] J. Shen, A.J. Soroka and R.D. Snook, *J. Appl. Phys.* **78**, 700 (1995).
[15] A.O. Marcano, C. Loper and N. Melikechi, *Appl. Phys. Lett.* **78**, 3415 (2001).
[16] E. Siegman, *Lasers* (University Science Books, 1986).
[17] Yariv, *Quantum Electronics* (John Wiley & Sons, 1989).
[18] J.A. Herman and R.G. McDuff, *JOSA B* **11**, 2056 (1993).
[19] T.D. Krauss and F.W. Wise, *Appl. Phys. Lett*. **65**, 1739 (1994).
[20] S. E. Bailkowski, *Photothermal Spectroscopy Methods for Chemical Analysis* (Wiley, New York, 1996).
[21] R.M. Pope and E.S. Fry, *Appl. Opt*. **36**, 8710 (1997).
[22] P. Gunter, *Nonlinear Optical Effects and Materials* (Springer, Berlin, 2000).
[23] S. Vijayalakshmi, F. Shen and H. Gerbel, *Appl. Phys. Lett*. **71**, 3332 (1997).
[24] S.M. Mian, S.B. McGee, N. Melikechi, *Opt. Comm*. **207**, 339 (2002).

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN 1-60021-194-1
Editor: Steven B. Larkin, pp. 63-119

Chapter 3

NONLINEAR OPTICS AS A METHOD FOR MATERIALS CHARACTERISATION

Chris J. Lee[1]***, Clare J. Strachan***[2]***, Thomas Rades***[2,*]***, Peter J. Manson***[1]
[1]Physics Department, University of Otago, PO Box 56, Dunedin, New Zealand
[2]School of Pharmacy, University of Otago, PO Box 913, Dunedin, New Zealand

Abstract

With the development of stable, compact and reliable pulsed laser sources the field of characterising materials, and in particular interfaces through their nonlinear optical response has bloomed. Second harmonic generation due to symmetry breaking at interfaces has provided a new spectroscopic tool of great utility. For example, it has been used to observe diffusion processes in semiconductors and to monitor the surface concentration of glucose oxidase in water. The surface response of various stained cellular features has been used as a source of illumination in microscopy, providing results similar to fluorescence microscopy but with less absorption and resulting damage.

The nonlinear optical response of various materials provides a very sensitive technique for measuring and monitoring some phase changes such as crystallisation from solution, micellisation of surfactants and polymorphic transitions of explosives. Our own work has focused on characterising and determining concentrations of pharmaceutically interesting bulk compounds, dispersions and emulsions. This work has potential application for in-line monitoring and quality control of pharmaceutical manufacturing.

In this chapter we present an extensive review of various spectroscopic techniques that make use of the nonlinear optical response of one or more media. We will also present the results of our own work in this field.

1. Introduction

Since the advent of the laser, nonlinear optics has been an area of active and fast paced scientific research. However, much of that work and associated publications have focused on device development [1–3]. In the minds of most laser physicists, nonlinear optics is

*E-mail address: thomas.rades@stonebow.otago.ac.nz

a way to generate coherent, laser like sources at wavelengths that are not accessible to traditional lasers. These developments have lead to many interesting spectroscopic devices in their own right [4,5]. We focus on using the nonlinear optical properties that are inherent to many materials and interfaces as a probe. Results are presented that show that the optical nonlinearity of molecules, interfaces and bulk materials can be probed to provide sensitive information about many molecular and bulk properties such as orientation, concentration and phase changes.

In general, nonlinear optical spectroscopy can refer to any measurement that probes the optical nonlinearity of the material of interest in order to gain some information about the material. Thus, one could consider probing such properties as the intensity dependent refractive index, Raman scattering, Brillouin scattering or four wave mixing as nonlinear optical spectroscopy. However, in this chapter we focus on second order optical nonlinearity which consists of sum and difference frequency generation (SFG and DFG) and the special case of SFG, second harmonic generation (SHG).

This chapter is laid out as follows. Section 2. describes some other spectroscopic techniques which can be considered as complementary to nonlinear optical spectroscopy. The techniques presented are examples of other spectroscopic techniques rather than a complete list. This part of the chapter is intended to set the background in which nonlinear optical techniques operate. Section 3. provides an outline of the theory behind nonlinear optics at a bulk, interfacial, microscopic and molecular level, it is through these theoretical results that one understands where SFG, SHG and DFG will be most sensitive and how to employ them successfully. Finally we present experimental and modelling results from various research groups including our own. Section 5. presents results that probe the second order optical nonlinearity at a molecular level while Sections 6. and 7. detail experimental measurements that probe surfaces, interfaces and adsorption. Bulk SHG measurements are presented in Section 8..

2. General Spectroscopic Methods

In this section various spectroscopic techniques are outlined with their concomitant advantages and disadvantages. Infrared and near-infrared spectroscopy are presented first because they are well established techniques that are part of the standard arsenal of tools used to characterise molecules and bulk materials. Raman spectroscopy is outlined also since it is complementary to infrared spectroscopy. Raman spectroscopic techniques are still under intense development and it is not so widely deployed as infrared spectroscopy. Nuclear magnetic resonance spectroscopy is briefly explained because it can be used to elucidate a very wide range of molecular, liquid and solid state properties. Finally a newly developed technique, terahertz pulsed spectroscopy is outlined.

2.1. Infrared Spectroscopy

Infrared (IR) spectroscopy, and in particular Fourier transform IR spectroscopy is one of the more widely used spectroscopic techniques. IR radiation coincides with most fundamental intramolecular vibrational modes, as well as some hydrogen bond vibrations. Thus,

it is very good for the identification and quantification of molecular compounds. It is extremely versatile and can be employed to identify gases, liquids, dissolved materials and solids [6,7]. In the case of solids the IR spectrum is usually collected by measuring changes in the diffuse reflected intensity and as such loses some sensitivity compared to IR spectroscopy performed on gases. IR spectroscopy of materials dissolved in liquid media may be hindered by the absorption spectrum of the solvent, especially if the solvent is polar in nature.

Physical changes such as polymorphic phase transitions, the formation of liquid crystals or particle size changes can be detected by the effect these changes have on the molecular vibrational spectrum. In principle, the absorption spectrum from a surface or interface may be different from those of the bulk material, however, the differences are small and the total absorption at the surface monolayer is too small to be detected.

For example, IR spectroscopy has been widely used to study polymorphism and crystallinity of pharmaceutical materials. Measurements are usually performed using reflection measurements, because sample compression, required for transmission measurements, may lead to formation of new crystal or amorphous forms [8,9]. Reflectance measurements may be in the form of diffuse reflectance Fourier transform infrared spectroscopy [9]. Several reviews [9–12] and books [8,13] have covered the use of Fourier transform IR spectroscopy to investigate the solid state forms of pharmaceutical materials. The IR spectrum of solid materials is sensitive to particle size variations as this changes the average reflectance of the sample and dilution of the sample is often necessary. Other methods such as attenuated total reflectance spectroscopy (ATR) can also be used to characterise the solid state. However, ATR is mainly sensitive to the surface of the bulk material and sample preparation often requires some compression which can lead to physical changes [9].

2.2. Near Infrared Spectroscopy

Near-infrared (near-IR) spectroscopy probes combinations and overtones of fundamental molecular vibrations observed in the mid-IR region. These combination and overtone modes are the result of anharmonicity of the fundamental vibrations. Many bands observed in the near-IR region are overtones and combinations for bonds involving hydrogen [14]. These bands are difficult to interpret and frequently overlap, and therefore this region has been slower than IR spectroscopy to be adopted by industrial sectors. A major advantage of near-IR spectroscopy over mid-IR spectroscopy is that the molar absorptivity is much lower. This allows spectra to be recorded in reflectance mode on undiluted solid samples, thus lending the technique to in line monitoring. Instrumental (for example, improved detectors, development of fast scan and Fourier transform instruments and development of fibre optic probes) and computational (statistical multivariate analysis techniques) developments led to a surge in interest in the technique for pharmaceutical monitoring in the 1990s [14].

Near-IR spectroscopy is used for both qualitative and quantitative purposes. Recently, the technique has become widely used for the rapid determination and quality control of raw materials including active drugs and especially excipients (materials other than the active drug used in drug products) [14]. In combination with multivariate analysis, near-IR

spectroscopy is becoming popular for quantitation of polymorphism and crystallinity during processing. Aaltonen et al. (2003) used near-IR spectroscopy to screen for sulphathiazole polymorphs and solid state changes induced by milling and compression [15]. Using near-IR spectroscopy, quantitative analysis of polymorphs in binary and multicomponent mixtures and pharmaceutical products has been performed [16–18]. Particle size dependent scattering is one of the largest factors affecting near-IR spectra. Although this is often considered a disadvantage, it has also been investigated as a means of determining particle size [19]. Near-IR radiation is strongly absorbed by water, and near-IR spectroscopy has been used to determine the moisture content of pure substances and formulated drug products [14]. However, the strong absorbance of water limits its application for characterising colloidal dispersions.

2.3. Raman Spectroscopy

Raman spectroscopy detects vibrations that involve a change in polarisability. Raman spectroscopy is complementary to IR spectroscopy because Raman spectroscopy is sensitive to aromatic systems and nonpolar bonds, while IR spectroscopy is more sensitive to polar bonds.

Difficulty in detecting the comparatively weak Raman scattering has limited its use in the industrial setting. However, the development of new detection techniques and better laser sources is making the technology more attractive as a routine diagnostic tool. One advantage of this weak scattering is that it can be used on undiluted samples without saturation, thus simplifying sample preparation and associated artefacts. Another advantage of Raman spectroscopy is that it is very insensitive to water. Therefore, it can be used to characterise processes and systems that involve water, for example suspensions and biological systems [20, 21]. Raman spectroscopy is very versatile and many new applications are currently under development.

Raman spectroscopy is limited by fluorescence, a common problem with organic molecules. A frequent solution is to illuminate the sample using near-IR laser sources, since the photon energy is then insufficient to excite electronic transitions. A factor that should be considered in Raman spectroscopy is that the spot size is typically less than 1 mm and the resulting average intensity can be sufficient to cause physical or chemical degradation. The high intensity radiation is necessary to excite sufficient Raman photons for detection of the spectra. The small sample area may also have consequences for quantitative analysis of inhomogeneous samples [22].

Traditionally Raman spectroscopy has been used to investigate molecules in solution or the gaseous or liquid states. In the solid state, Raman spectroscopy has found wide application in pharmaceutical research. For example, Tudor et al. (1993) used near-IR Fourier transform Raman spectroscopy to investigate and quantify the polymorphism of the drug, chlorpropamide [23], and amorphous and crystalline forms of fructose and glucose were investigated by Söderholm et al. (1999) using the same technique [24]. Raman spectroscopy setups have been developed to investigate polymorphism in microlitre volume fluids [25] and pharmaceutical dosage forms [26, 27]. In addition, Raman microscopy, also known as Raman mapping, has been used to map the drug, ibuprofen, in two dimensions in solid dispersions with polyvinylpyrrolidone and to determine that the drug was in an amorphous,

not crystalline, form [28].

2.4. Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy measures the interaction of an applied magnetic field with atomic nuclei. NMR spectroscopy has been used to elicit both chemical and physical information about materials and systems.

Solid state NMR spectroscopy is capable of extracting detailed structural information and can be used for crystal structure elucidation. The technique suffers the drawbacks of being expensive and slow. NMR spectroscopy is also insensitive to particle size, which may or may not be an advantage. The technique has been used in several investigations of pharmaceutical polymorphism and some reviews [9, 11, 29, 30] and books [8, 13] have covered this area.

Self diffusion NMR spectroscopy is widely used to characterise colloidal systems. The self diffusion coefficient depends not only on the molecular species, but also on the physical context. Self diffusion NMR can be used to provide much information on the colloidal system studied, including the structure of the colloids themselves and interactions in multicomponent systems [31].

2.5. Terahertz Pulsed Spectroscopy

Terahertz pulse spectroscopy (TPS) is a very recently developed technique, which has allowed much improved access to modes in the far-IR region. In materials, terahertz radiation can induce rotational transitions, low frequency bond vibrations, crystalline phonon vibrations, hydrogen-bonding stretches and torsional vibrations [32, 33]. In gases, rotational transitions predominate, in liquids some hydrogen bonding and other low energy bonding vibrations occur, and in solids the response is composed mainly of crystalline phonon vibrations [32–34]. Measurements can be in transmission or reflectance mode. The technique can be used either as a spectroscopic tool probing intermolecular bonds directly, or it can be used to image substances that are strongly scattering. Another advantage of the technique is that the energy used is lower than that of shorter wavelength techniques. However, the equipment for TPS is more expensive than for IR, near-IR and Raman spectroscopy, and the techniques has not yet been widely applied. In addition, terahertz radiation is strongly absorbed by water.

TPS has recently been used to investigate and quantify polymorphism and crystallinity of pharmaceutical materials [35–37]. TPS has also been used to image tablet layering [38], and identify cancerous tissue [39].

3. General Nonlinear Optical Theory

In this section we introduce the reader to basic concepts of nonlinear optics. These include topics such as the definition of the electric field susceptibility and some of its general properties, propagation of light in materials that exhibit a nonlinear response and the importance of dispersion and birefringence of nonlinear optical materials. This introduction is designed

to succinctly outline some of the more important results of nonlinear optics and not as a replacement for a text book. For further clarification of the material presented here, there are several good text books such as Shen [40] and Boyd [41].

3.1. Nonlinear Polarisation

The electric field susceptibility is defined in terms of the polarisation induced by an applied electric field is given by [42]

$$P_i(\mathbf{r},t) = \epsilon_0 \sum_{n=1}^{\infty} \int d\mathbf{r}'_1 \int dt'_1 \cdots \int d\mathbf{r}'_n \int dt'_n \times \chi^{u(n)}_{ij_1\cdots j_n}(\mathbf{r}-\mathbf{r}'_1, t-t'_1, \cdots, \mathbf{r}-\mathbf{r}'_n, t-t'_n) \times E_{j_1}(\mathbf{r}'_1, t'_1)\cdots E_{j_n}(\mathbf{r}'_n, t'_n) \quad (1)$$

where $\mathbf{r}$ is the position vector, the Einstein summation convention is used and it is assumed that $\mathbf{P}$ = 0 when no applied field is present. $\chi^{u(n)}_{ij_1\cdots j_n}(\cdots)$ is the electric field susceptibility, the linear version of which is given in standard electromagnetism texts by $\epsilon = 1 - \chi$, where ϵ is the electric field permittivity. The superscript, $u(n)$, is used to denote that the susceptibility of order n is not a symmetric tensor. Causality requires that $\chi^{u(n)}_{ij_1\cdots j_n}(\cdots)$ = 0, if any $t - t'_n < 0$. $\mathbf{P}$ is uniquely determined by equation 1; however, the electric field susceptibility is not unique. It makes physical and mathematical sense to introduce a symmetric representation of the susceptibility. This also has the desirable property of uniquely determining $\chi^{(n)}_{ij_1\cdots j_n}$ which is the symmetric representation of the electric field susceptibility and is given by [43]

$$\chi^{(n)}_{ij_1\cdots j_n}(\mathbf{r}-\mathbf{r}'_1, t-t'_1, \cdots, \mathbf{r}-\mathbf{r}'_n t-t'_n) = \frac{1}{n!}\sum_{\mathrm{P}} \chi^{u(n)}_{ij_{\mathrm{P}(1)}\cdots j_{\mathrm{P}(n)}}(\mathbf{r}-\mathbf{r}'_{\mathrm{P}(1)}, t-t'_{\mathrm{P}(1)}, \cdots, \mathbf{r}-\mathbf{r}'_{\mathrm{P}(n)}, t-t'_{\mathrm{P}(n)}) \quad (2)$$

where the summation is over all permutations of P. The symmetry properties of $\chi^{(n)}_{ij_1\cdots j_n}$ allow the interchange of the indicies on all but the first position vector and time as shown below

$$\chi^{(n)}_{ij_1\cdots,j_k,\cdots,j_l,\cdots,j_n}(\mathbf{r}-\mathbf{r}'_1, t-t'_1, \cdots, \mathbf{r}-\mathbf{r}'_k, t-t'_k, \cdots, \mathbf{r}-\mathbf{r}'_l, t-t'_l, \cdots, \mathbf{r}-\mathbf{r}'_n, t-t'_n) = \chi^{(n)}_{ij_1\cdots,j_l,\cdots,j_k,\cdots,j_n}(\mathbf{r}-\mathbf{r}'_1, t-t'_1, \cdots, \mathbf{r}-\mathbf{r}'_l, t-t'_l, \cdots, \mathbf{r}-\mathbf{r}'_k, t-t'_k, \cdots, \mathbf{r}-\mathbf{r}'_n, t-t'_n) \quad (3)$$

for any $k, l > 1$. Thus $\chi^{(n=2)}_{ijk}$ has 18 independent elements instead of 27, which is commonly expressed by removing one of the subscripts. Thus the second order susceptibility (a 3 x 3 x 3 tensor) is mapped to a 3 x 6 element tensor.

The number of independent elements of $\chi^{(n)}_{ij_1\cdots j_n}$ can be further reduced by considering the symmetry mapping operations, $\{S\}$ of the medium. A symmetry mapping operation requires that the properties of the medium remain invariant under that operation. In general, S is a second-rank three-dimensional tensor S_{lm}, and when applied to $\chi^{(2)}_{ijk}$ is given by [40]

$$(\hat{i}.\mathbf{S}^{\dagger})\cdot\chi^{(2)} : (\mathbf{S}\cdot\hat{j})(\mathbf{S}\cdot\hat{k}) = \chi^{(2)}_{ijk} \quad (4)$$

For instance, index exchange symmetry of equation 3 applied to the second order susceptibility shows that $\chi^{(2)}_{ijk} = \chi^{(2)}_{ikj}$. However, inversion symmetry ($\mathbf{S} \cdot \hat{e} = -\hat{e}$) requires that $\chi^{(2)}_{ijk} = -\chi^{(2)}_{ikj}$ which is only satisfied if $\chi^{(2)}_{ijk} = 0$. Symmetry considerations are described further in Section 4..

If the total electric field can be expressed as the sum of monochromatic waves, as in

$$E(\mathbf{r}, t) = \sum_i E(\mathbf{k}_i, \omega_i) \tag{5}$$

where $\mathbf{k}$ is the wave vector, then the Fourier transform of equation 1 gives [40]

$$\mathbf{P}(\mathbf{k}, \omega) = \mathbf{P}^{(1)}(\mathbf{k}, \omega) + \mathbf{P}^{(2)}(\mathbf{k}, \omega) + \cdots + \mathbf{P}^{(n)}(\mathbf{k}, \omega) \tag{6}$$

where

$$\mathbf{P}^{(1)}(\mathbf{k}, \omega) = \epsilon_0 \chi^{(1)}(\mathbf{k}, \omega) \cdot E(\mathbf{k}, \omega) \tag{7}$$

$$\mathbf{P}^{(2)}(\mathbf{k}, \omega) = \epsilon_0 \chi^{(2)}_{ij}(\mathbf{k} = \mathbf{k}_i + \mathbf{k}_j, \omega = \omega_i + \omega_j) \cdot E(\mathbf{k}_i, \omega_i) E(\mathbf{k}_j, \omega_j) \tag{8}$$

$$\mathbf{P}^{(3)}(\mathbf{k}, \omega) = \epsilon_0 \chi^{(3)}_{ijk}(\mathbf{k} = \mathbf{k}_i + \mathbf{k}_j + \mathbf{k}_k, \omega = \omega_i + \omega_j + \omega_k) \cdot E(\mathbf{k}_i, \omega_i) E(\mathbf{k}_j, \omega_j) E(\mathbf{k}_k, \omega_k) \tag{9}$$

where $\chi^{(n)}_{ij_1 \cdots j_n}$ is the Fourier transform of the symmetric form of the electric field susceptibility.

With the polarisation expressed as a function of frequency and wave vector it is clear that the higher order susceptibility terms result in momentum changes and frequency mixing. Thus, the second order electric field susceptibility, $\chi^{(2)}_{ijk}$, results in sum and difference frequency terms generated from two light fields. A special case of sum frequency generation (SFG) is second harmonic generation (SHG) where a single intense light field results in photons at twice the frequency of the illuminating light field. Since the electric field susceptibility is a function of molecular structure, higher order polarisation response can be used as a probe to examine bulk, surface and molecular structure.

In section 8. we review some recent experimental work that utilises the bulk nonlinearity of materials to monitor phase changes and measure mixture ratios of chemically similar substances.

3.2. Optical Nonlinearity at a Surface

In the previous section we outlined the general properties of the second order nonlinear optical susceptibility (also referred to as the electric field susceptibility) in the electric dipole approximation. Due to the presence of inversion symmetry, materials that are isotropic exhibit no second order interactions such as second harmonic generation. However, at the interface between two such media inversion symmetry is broken so the second order susceptibility is non zero. In this section we follow the approach used by Janner and Koopmans to outline some of the properties of the optical nonlinearity at a surface [44, 45].

Relaxing the electric dipole approximation to allow for contributions from electric quadrupole and magnetic dipoles, the effective second order polarisation , $\mathbf{P}^{(2)}_e$ can be written

$$\mathbf{P}^{(2)}_e(\omega = \omega_1 + \omega_2) = \mathbf{P}^{(2)}(\omega_1 + \omega_2) - \nabla \mathbf{Q}^{(2)}(\omega = \omega_1 + \omega_2) + \frac{c}{j2\omega} \nabla \times \mathbf{M}^{(2)}(\omega = \omega_1 + \omega_2) + \cdots \tag{10}$$

where $\mathbf{Q}$ and $\mathbf{M}$ are the electric quadrupole polarisation and magnetic dipole magnetisation. The electric quadrupole polarisation has a susceptibility, $\chi^{(2)}_{Q,ijkl}$ associated with it. The quadrupole susceptibility is fourth rank tensor and can be split into two components $\chi^{(2)}_{Q,ijkl} = \chi^{(2)}_{Q_p,ijkl} - \chi^{(2)}_{Q_q,ijkl}$. The magnetic dipole magnetisation has a third order susceptibility, $\chi^{(2)}_{M,ijk}$. Their respective contributions can be written

$$\mathbf{P}^{(2)}(\omega = \omega_1 + \omega_2) = \epsilon_0 \chi^{(2)}_{ijk} : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2) + \epsilon_0 \chi^{(2)}_{Q_p,ijkl}\big(\mathbf{E}(\omega_1)\nabla\mathbf{E}(\omega_2) + \mathbf{E}(\omega_2)\nabla\mathbf{E}(\omega_1)\big) \tag{11}$$

$$\mathbf{Q}^{(2)} = \epsilon_0 \chi^{(2)}_{Q_q,ijkl} : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2) \tag{12}$$

$$\mathbf{M}^{(2)} = \mu_0 \chi^{(2)}_{M,ijk} : \big(\mathbf{E}(\omega_1)\mathbf{E}(\omega_2)\big) \tag{13}$$

Note that the electric quadrupole susceptibility contributes in two quite different ways. In equation 11 it contributes directly to the second order polarisation through the gradient of the fundamental fields, while in equation 12 it contributes through the gradient of the polarisation which can be thought of as the gradient of the generated fields. The magnetic dipole contribution vanishes for materials that are non-magnetic and exhibit bulk inversion symmetry. Thus, assuming an isotropic, non-magnetic medium, the effective polarisation can be written as

$$\begin{aligned}\mathbf{P}^{(2)}_e(\omega = \omega_1 + \omega_2) &= \epsilon_0 \chi^{(2)}_{ijk} : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2) \\ &+ \epsilon_0 \nabla\big(\chi^{(2)}_{Q_q,ijkl}\big) : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2) + \epsilon_0 \chi^{(2)}_{Q_p,ijkl}\big(\mathbf{E}(\omega_1)\nabla\cdot\mathbf{E}(\omega_2) + \mathbf{E}(\omega_2)\nabla\cdot\mathbf{E}(\omega_1)\big)\end{aligned} \tag{14}$$

In bulk isotropic media the electric dipole contribution, $\chi^{(2)}_{ijk}$, is zero due to symmetry properties as is $\nabla(\chi^{(2)}_{Q_q,ijkl})$. However, the remaining electric quadrupole terms are not zero. Thus, the bulk nonlinear susceptibility of the isotropic medium is given by

$$\mathbf{P}^{(2)}_{bulk}(\omega = \omega_1 + \omega_2) = \epsilon_0 \chi^{(2)}_{Q_p,ijkl}(\mathbf{E}(\omega_1)\nabla\cdot\mathbf{E}(\omega_2) + \mathbf{E}(\omega_2)\nabla\cdot\mathbf{E}(\omega_1)) \tag{15}$$

At an interface, neither the gradient of the quadrupole susceptibility, nor the electric dipole contributions are zero. Thus, the surface polarisability is given by

$$\begin{aligned}\mathbf{P}^{(2)}_s(\omega = \omega_1 + \omega_2) &= \epsilon_0 \chi^{(2)}_{ijk} : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2) - \epsilon_0 \nabla_s\big(\chi^{(2)}_{Q_q,ijkl}\big) : \mathbf{E}(\omega_1)\mathbf{E}(\omega_2) \\ &+ \epsilon_0 \chi^{(2)}_{Q_p,ijkl}\big(\mathbf{E}(\omega_1)\nabla_s\mathbf{E}(\omega_2) + \mathbf{E}(\omega_2)\nabla_s\mathbf{E}(\omega_1) + \nabla_s\big(\mathbf{E}(\omega_1)\cdot\mathbf{E}(\omega_2)\big)\big)\end{aligned} \tag{16}$$

where the subscript s denotes that the gradient should be taken across the interface. Note that the quadrupole term allows contributions from both the surface and the bulk of the media which can complicate experiments. For the sake of clarity we assume that $\omega_1 = \omega_2 = \omega$, then the bulk polarisation is given by

$$\begin{aligned}\mathbf{P}^{(2)}_{bulk}(2\omega) &= \epsilon_0 \chi^{(2)}_{Q,xxyy}\mathbf{E}(\omega)(\nabla\cdot\mathbf{E}(\omega)) + \\ &+ \frac{1}{2}\epsilon_0 \chi^{(2)}_{Q,xyxy}\nabla(\mathbf{E}(\omega)\cdot\mathbf{E}(\omega)) + \epsilon_0 \chi^{(2)}_{Q,xxzz}(\mathbf{E}(\omega)\cdot\nabla)\mathbf{E}(\omega)\end{aligned} \tag{17}$$

where we have substituted explicit components for the various dipole and quadrupole contributions to the optical nonlinearity.

Since the divergence of the macroscopic electric field is zero the first term vanishes. The third term will only contribute when $\mathbf{E} \cdot \mathbf{k} \neq 0$ which is only true in the presence of multiple non-collinear plane waves. The second term is non zero and indistinguishable from the surface contribution. By careful manipulation of the experimental geometry, one can use the relationship between the second and third terms to separate out their individual contributions and thus obtain only the surface contribution to the signal [44]. In sections 6. and 7. we present a review of recent experimental results using surface optical nonlinearity to elucidate various physical phenomena.

4. Microscopic Theory of Nonlinear Susceptibility

The polarisation of the medium in response to the applied electric fields is a combination of the induced dipole moments of the constituent atoms or molecules. At the microscopic level, the response is usually expressed in terms of (hyper)polarisability tensors $\boldsymbol{\alpha}, \boldsymbol{\beta} \ldots$ which relate the induced molecular dipole moment $\boldsymbol{\mu}(t)$ to the electric field.

$$\boldsymbol{\mu}(t) = \boldsymbol{\alpha} \cdot \mathbf{E}(t) + \boldsymbol{\beta} : \mathbf{E}(t)\mathbf{E}(t) + \boldsymbol{\gamma} : \mathbf{E}(t)\mathbf{E}(t)\mathbf{E}(t) + \ldots \tag{18}$$

Techniques for calculating the hyper-polarisabilities vary depending on the microscopic model of the atom used. We will describe the solution based both on a classical charged particle in a potential well and on the solution to the equation of motion for the density matrix. The classical model is relevant because the solution is relatively straightforward and it shows many of the characteristics of the full density matrix model.

The simplest, classical model of an atom consists of an electron cloud centered on a fixed positive nucleus, with the electron assumed to move in an anharmonic potential well. An externally applied electric field causes a small displacement of the electron cloud, producing on average a separation of the positive and negative charges in the atom and hence an electric dipole. The combination of the microscopic electric dipoles produces the macroscopic polarisation of the medium.

In this model [40, 41], the electron is treated as a classical particle displaced by a distance $x(t)$ from the fixed nucleus. The equation of motion for the electron is

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x + ax^\eta = -eE(t)/m. \tag{19}$$

The natural frequency ω_0 of the harmonic motion is taken to be the frequency of the optical transition. The model also includes a damping force with coefficient γ which is related to the linewidth of the optical transition. The force driving the oscillation is due to the applied electric field $E(t)$, assumed to be a number of monochromatic waves. The anharmonic term, ax^η, is included [40, 41] to represent the nonlinearity added to the basic harmonic oscillator model. Without this term, the response of the oscillator is purely linear and no additional frequencies beyond those present in the incident field are generated. In the simplest model, the parameter η is set to 2 or 3, depending on the symmetry characteristics of the medium. A value of $\eta = 2$ corresponds to a cubic perturbation to the harmonic potential function $U(x)$, which is used for media without a centre of inversion symmetry. The choice $\eta = 3$ gives an x^4 addition to the harmonic potential function U, so $U(x) =$

$U(-x)$ as required for centrosymmetric media. These are the lowest order additions to the harmonic potential that display the required symmetry.

Equation 19 cannot be solved directly, so a perturbative solution technique is used under the assumption that the anharmonic modification to the potential energy is small. The result is a series of contributions $x^{(n)}(t)$ to the displacement, corresponding to successive orders in the perturbation. The details of the solution are given in Reference [41], so the results will be summarised here.

The equation for the lowest order $(n = 1)$ is equation 19 without the anharmonic term, so the lowest order response is equal to the linear response of the oscillator. If the driving field consists of a number of monochromatic waves of frequency ω_j, the linear response consists of an oscillatory term for each frequency

$$x(t) = -\frac{e}{m}\frac{E}{\omega_0^2 - \omega_j^2 - j\omega_j\gamma}e^{-j\omega_j t}. \tag{20}$$

This term is obviously resonant at ω_0, the natural frequency of the harmonic oscillator, and is proportional to the strength of the external field. The susceptibility is then calculated from the polarisation of the medium $\mathbf{P}(t) = Nex(t) = \epsilon_0\chi E(t)$

For $\eta = 2$ corresponding to a noncentrosymmetric medium, each subsequent order $x^{(n)}(t)$ with $n \geq 2$ is the solution to a modified form of equation 19, i.e.

$$\ddot{x}^{(n)} + \gamma\dot{x}^{(n)} + \omega_0^2 x^{(n)} = -a\sum_{\nu=1}^{n-1} x^{(\nu)}x^{(n-\nu)} \tag{21}$$

As the solution process proceeds through subsequent orders n beyond the linear response (equation 20), the products of lower order solutions appearing in the driving term generate new frequencies in each $x^{(n)}(t)$. For example [41], if the external field has two frequency components, the second order response $x^{(2)}(t)$ will consist of components at the sum and difference frequencies, the second harmonics of each input frequency and at DC. Thus each order in the perturbative solution corresponds to the same order in the (hyper-)polarisabilities (equation 18) and therefore the nonlinear optical susceptibilities. The anharmonic oscillator model therefore predicts all the standard second order nonlinear optical effects and allows the susceptibilities to be calculated. In the general case, each new frequency component includes all the resonant denominators from the lower order source terms, in addition to a resonance at the new harmonic of ω_0.

For a centrosymmetric medium, $\eta = 3$ a similar solution process is followed. In this case the modified equation of motion for $x^{(n)}(t)$ for $n \geq 3$ is

$$\ddot{x}^{(n)} + \gamma\dot{x}^{(n)} + \omega_0^2 x^{(n)} = -a\sum_{\nu=2}^{n-1}\sum_{\mu=1}^{\nu-1} x^{(\mu)}x^{(\nu-\mu)}x^{(n-\nu)} \tag{22}$$

Again, new frequency components are generated in each increasing order of the solution. However, the most significant difference from the previous case is that the equation for $x^{(2)}(t)$ has no driving force term, and hence the second order response $x^{(2)}(t)$ is identically zero. Thus the classical anharmonic oscillator model predicts the well-known result that

second harmonic generation and other second-order processes are forbidden in centrosymmetric media on symmetry grounds.

Although the classical anharmonic oscillator model predicts much of the form of the nonlinear optical response of an atom, it is a classical model and is expected to be incomplete. Perhaps the most obvious deficiency is that there is only a single resonance frequency, ω_0. The full quantum mechanical treatment overcomes these problems. The susceptibilities could be calculated by solving the Schrödinger equation for the atomic wave function, calculating the average dipole moment or polarisation and hence the susceptibility. The disadvantage of this approach is that it cannot treat relaxation processes accurately, although there are approximations such as allowing the resonance frequencies in the final results to be complex.

The most complete treatment involves solving the Liouville equation for the density matrix ($\hat{\rho}$) [40, 41, 46] and again calculating the average polarisation from $\hat{\rho}$. Again, a perturbative method is used, and the individual orders of (hyper-)polarisability relate directly to the individual perturbation orders.

The equation of motion for the density matrix, $\hat{\rho}$, is known as the Liouville equation. The basic form of the Liouville equation involves a commutator of the system Hamiltonian with the density matrix. In addition to this commutator, relaxation processes must be added phenomenologically. The terms added are selected to reflect the physical system being modelled. Perhaps the simplest model [41, 46, 47] includes relaxation of all density matrix components towards equilibrium values $\hat{\rho}^{\mathrm{eq}}_{nm}$ (i.e. in the absence of the external interaction) with characteristic rates γ_{nm}

$$\dot{\rho}_{nm} = \frac{-j}{\hbar}\left[\hat{H},\hat{\rho}\right]_{nm} - \gamma_{nm}\left(\hat{\rho}_{nm} - \hat{\rho}^{\mathrm{eq}}_{nm}\right) \tag{23}$$

The equilibrium state density matrix is generally assumed to include diagonal elements only. Clearly the relaxation processes in equation 23 stop when the system reaches the equilibrium state.

A more complete model of relaxation explicitly tracks the sources and sinks of relaxation processes for each diagonal element.

$$\dot{\rho}_{nn} = \frac{-j}{\hbar}\left[\hat{H},\hat{\rho}\right]_{nn} + \sum_{E_m>E_n}\Gamma_{nm}\hat{\rho}_{mm} - \sum_{E_m<E_n}\Gamma_{mn}\hat{\rho}_{nn} \tag{24}$$

$$\dot{\rho}_{nm} = \frac{-i}{\hbar}\left[\hat{H},\hat{\rho}\right]_{nm} - \gamma_{nm}\hat{\rho}_{nm} \tag{25}$$

The second and third terms of equation 24 respectively represent the decay of population from higher levels into level n and the loss of population from level n into lower levels. In each case, Γ_{ij} represents the probability per unit time of a decay from level i to level j occurring, and the result is exponential decay or increase respectively of the population in level n. These equations are more appropriate for modelling closed atomic systems [48], i.e. systems where the population may only exist within the set of levels included in the calculation. In contrast, equations 23 are appropriate when the level system is open, and the system relaxes towards a thermal equilibrium configuration which may be maintained by an additional process such as collisional or optical pumping. The second term of equation 25 represents the decay of the coherence between levels n and m. There are a number

of mechanisms for this decay, and it is worth noting that coherences generally decay at a different rate from the associated population decay rate.

Once the form of the Hamiltonian for the system is known, the set of coupled differential equations represented by equations 23 or 24 and 25 is solved to find the density matrix. To model the interaction of an optical field $\mathbf{E}(t)$ with an atom, the system Hamiltonian takes the form

$$\hat{H} = \hat{H}_0 + \hat{V}(t) = \hat{H}_0 - \hat{\boldsymbol{\mu}} \cdot \mathbf{E}(t) \tag{26}$$

where $\hat{H}_0$ is the atomic Hamiltonian and $\hat{V}(t)$ represents the interaction. Under the electric dipole approximation, the interaction takes the form shown in the second part of equation 26, where $\hat{\boldsymbol{\mu}}$ is the electric dipole operator for the atom. Because the solution process is simpler yet retains the character of the nonlinear susceptibilities, we will describe the solution process for equation 23.

The first step in the solution is to split the commutator into two terms using equation 26. Since the atomic states are eigenstates of $\hat{H}_0$, the matrix representation of $\hat{H}_0$ is diagonal and equation 23 becomes

$$\dot{\rho}_{nm} = -j\omega_{nm}\hat{\rho}_{nm} - \frac{j}{\hbar}\left[\hat{V}(t), \hat{\rho}\right]_{nm} - \gamma_{nm}\left(\hat{\rho}_{nm} - \hat{\rho}_{nm}^{\mathrm{eq}}\right), \tag{27}$$

where ω_{nm} is the frequency of the optical transition between levels n and m. In common with the solution to the anharmonic oscillator model described above, a perturbative technique is used to solve these equations, assuming that the interaction $\hat{V}(t)$ is weak. The result is a series of contributions $\hat{\rho}^{(n)}(t)$ to the density matrix.

The lowest order contribution $\hat{\rho}^{(0)}(t)$ satisfies a modified version of equation 23 which has no interaction term. This contribution is therefore just equal to the equilibrium state density matrix, while the higher orders represent small departures from the equilibrium state caused by the interaction. It is worth noting that the lowest order term in the perturbative expansion of $x(t)$ for the anharmonic oscillator was $n = 1$, since the equilibrium position was implicitly assumed to be $x = 0$. Contributions $\hat{\rho}^{(q)}$ to the density matrix for $q \geq 1$ also satisfy a modified version of equation 23

$$\dot{\rho}_{nm}^{(q)} = -\left(j\omega_{nm} + \gamma_{nm}\right)\hat{\rho}_{nm}^{(q)} - \frac{j}{\hbar}\left[\hat{V}, \hat{\rho}^{(q-1)}\right]_{nm} \tag{28}$$

which can be solved directly [41] to give

$$\hat{\rho}_{nm}^{(q)}(t) = \frac{-j}{\hbar}\int_{-\infty}^{t}\left[\hat{V}(t'), \hat{\rho}^{(q-1)}\right]_{nm} e^{(j\omega_{nm}+\gamma_{nm})(t-t')}dt'. \tag{29}$$

So in a similar fashion to the classical anharmonic oscillator solution, the source function for each order $\hat{\rho}^{(q)}$ comes from the previous order. Since the zeroth order solution is DC, the first order contribution will again consist of a term oscillating at each frequency of the incident field, which is again assumed to consist of a number of monochromatic waves. Higher order frequency components in the density matrix are then generated in subsequent orders because of frequency mixing with the oscillating interaction term.

To complete the solution, the details of the interaction operator $\hat{V(t)}$ must be included in equation 29. As indicated in equation 26, the interaction involves the atomic dipole moment

operator, $\hat{\boldsymbol{\mu}} = -e\hat{\mathbf{r}}$. The diagonal elements of this operator are generally assumed to be zero [41,49], although it is worth noting that systems with non-zero diagonal elements have been studied [50, 51, and references therein]. The off-diagonal elements of the dipole operator are related to the transition strength or oscillator strength of the transition [52]. Once the dipole operator and the incident field components have been included in equation 29, the integral can be evaluated and the sequence of density matrix contributions calculated.

The nonlinear optical susceptibility for each order is calculated from the polarisation $\mathbf{P}$ of the medium evaluated using the same order density matrix contribution

$$\mathbf{P}^{(q)} = N\langle\hat{\boldsymbol{\mu}}^{(q)}\rangle = N\,\mathrm{Tr}\left(\hat{\rho}^{(q)}\hat{\boldsymbol{\mu}}\right). \tag{30}$$

where N is the atomic number density. The polarisation vector thus depends on the external field through the interaction operator, and on a combination of *two* vector dipole operators through the interaction operator and the trace operation in equation 30. The double combination of dipole vectors produces the tensor nature of the susceptibility.

If the incident field is a single monochromatic wave of frequency ω, the linear susceptibility tensor can be written in the form [41]

$$\boldsymbol{\chi}^{(1)}(\omega) = \frac{N}{\epsilon_0\hbar}\sum_{nm}\rho^{\mathrm{eq}}_{mm}\left[\frac{\boldsymbol{\mu}_{mn}\boldsymbol{\mu}_{nm}}{\omega_{nm}-\omega-j\gamma_{nm}} + \frac{\boldsymbol{\mu}_{nm}\boldsymbol{\mu}_{mn}}{\omega_{nm}+\omega+j\gamma_{nm}}\right] \tag{31}$$

Although this is not the most compact way of writing the linear susceptibility, the format matches that used for higher order susceptibilities where the notation is designed to emphasise the symmetry of the susceptibility. The combinations of dipole moment vectors in equation 31 produce a dyadic, giving the tensor character of the susceptibility. This tensor character is more obvious if the susceptibility is written [41] in terms of Cartesian coordinates

$$\chi^{(1)}_{ij}(-\omega_\sigma;\omega) = \frac{N}{\epsilon_0\hbar}\sum_{nm}\rho^{\mathrm{eq}}_{mm}\left[\frac{\mu^i_{mn}\mu^j_{nm}}{\omega_{nm}-\omega-j\gamma_{nm}} + \frac{\mu^i_{nm}\mu^j_{mn}}{\omega_{nm}+\omega+j\gamma_{nm}}\right] \tag{32}$$

where the notation μ^i_{mn} represents the $i = x$, y or z Cartesian component of the atomic electric dipole vector associated with the transition between level n and level m. The susceptibility is again resonant but, in contrast to the anharmonic oscillator model, there are several resonant frequencies ω_{nm}; one for each pair of atomic levels labelled by m and n in the summation.

The frequency argument of $\chi^{(1)}_{ij}$ in equation 32 has been written in the conventional form that is often used for higher order susceptibilities, even though it is not particularly helpful in the linear case. When the medium polarisation is calculated from the density matrix and then combined with the electric fields equations 7 ... 9 etc to determine the susceptibilities, the resulting susceptibilities are independent of time, but do depend on the frequencies of the applied fields. The convention often used is to write the susceptibility in the form $\chi^{(n)}\left(-\omega_\sigma;\pm\omega_1,\pm\omega_2,\ldots\pm\omega_n\right)$ where $\omega_\sigma = \sum_{i=1}^{n}\pm\omega_i$. The frequencies of the incident fields are thus combined to produce sum and difference frequencies in the nonlinear response and the combination of signs chosen determines the physical process described. For example, $\chi^{(2)}\left(-\omega_\sigma;\omega_1,-\omega_2\right)$ refers to difference frequency generation. This notation

simplifies the statement of full permutation symmetry which was introduced in section 3. and is mentioned again below.

Calculation of the higher order susceptibilities proceeds in a similar fashion. The higher order density matrix contribution is calculated using equation 29 with the preceding order as the source term and with the appropriate number of field components included in the interaction. After calculating the polarisation using equation 30, the susceptibility tensor components are derived by comparison of the result with equations 8, 9, etc. The second order susceptibility is given by

$$\begin{aligned}
&\chi^{(2)}_{ijk}\left(-\omega_\sigma;\omega_1,\omega_2\right) = \\
&\quad\frac{N}{2\epsilon_0\hbar^2}\sum_{lmn}\rho^{\mathrm{eq}}_{ll}\left[\frac{\mu^i_{ln}\mu^j_{nm}\mu^k_{ml}}{\left(\omega_{nl}-\omega_1-\omega_2-j\gamma_{nl}\right)\left(\omega_{ml}-\omega_2-j\gamma_{ml}\right)}\right. \\
&\quad+\frac{\mu^i_{ln}\mu^k_{nm}\mu^j_{ml}}{\left(\omega_{nl}-\omega_1-\omega_2-j\gamma_{nl}\right)\left(\omega_{ml}-\omega_1-j\gamma_{ml}\right)} \\
&\quad+\frac{\mu^k_{ln}\mu^i_{nm}\mu^j_{ml}}{\left(\omega_{mn}-\omega_1-\omega_2-j\gamma_{mn}\right)\left(\omega_{nl}+\omega_2+j\gamma_{nl}\right)} \\
&\quad+\frac{\mu^j_{ln}\mu^i_{nm}\mu^k_{ml}}{\left(\omega_{mn}-\omega_1-\omega_2+j\gamma_{mn}\right)\left(\omega_{nl}+\omega_1+j\gamma_{nl}\right)} \\
&\quad+\frac{\mu^j_{ln}\mu^i_{nm}\mu^k_{ml}}{\left(\omega_{nm}+\omega_1+\omega_2+j\gamma_{nm}\right)\left(\omega_{ml}-\omega_2-j\gamma_{ml}\right)} \\
&\quad+\frac{\mu^k_{ln}\mu^i_{nm}\mu^j_{ml}}{\left(\omega_{nm}+\omega_1+\omega_2+j\gamma_{nm}\right)\left(\omega_{ml}-\omega_1-j\gamma_{ml}\right)} \\
&\quad+\frac{\mu^k_{ln}\mu^j_{nm}\mu^i_{ml}}{\left(\omega_{ml}+\omega_1+\omega_2+j\gamma_{ml}\right)\left(\omega_{nl}+\omega_2+j\gamma_{nl}\right)} \\
&\quad\left.+\frac{\mu^j_{ln}\mu^k_{nm}\mu^i_{ml}}{\left(\omega_{ml}+\omega_1+\omega_2+j\gamma_{ml}\right)\left(\omega_{nl}+\omega_1+j\gamma_{nl}\right)}\right]
\end{aligned} \tag{33}$$

where ω_1 and ω_2 are the frequencies of the two incident fields.

Interpretation of equation 33 can be simplified by considering both the arrangement of relevant atomic energy levels and the type of interaction with the fields. Firstly, the summation over l, m and n includes all combinations of atomic energy levels with which the fields may interact. However, the set of level combinations to include in the summation is substantially restricted by the dipole matrix elements in the numerators. Since all the diagonal elements of the dipole moment operator are equal to zero, only those combinations without repeated levels are allowed by the sets of subscripts on the dipole matrix elements.

Secondly, it should be noted that the field frequencies ω_1 and ω_2 may be positive or negative, with positive frequencies representing photon absorption by the medium and negative frequencies representing emission. The appropriate choice can generally be made based on the type of process considered and the relative energies of the levels involved. For example, frequency doubling in a ladder atomic system would correspond to positive frequencies for ω_1 and ω_2, while stimulated Raman scattering in a lambda system would require opposite signs for the two frequencies.

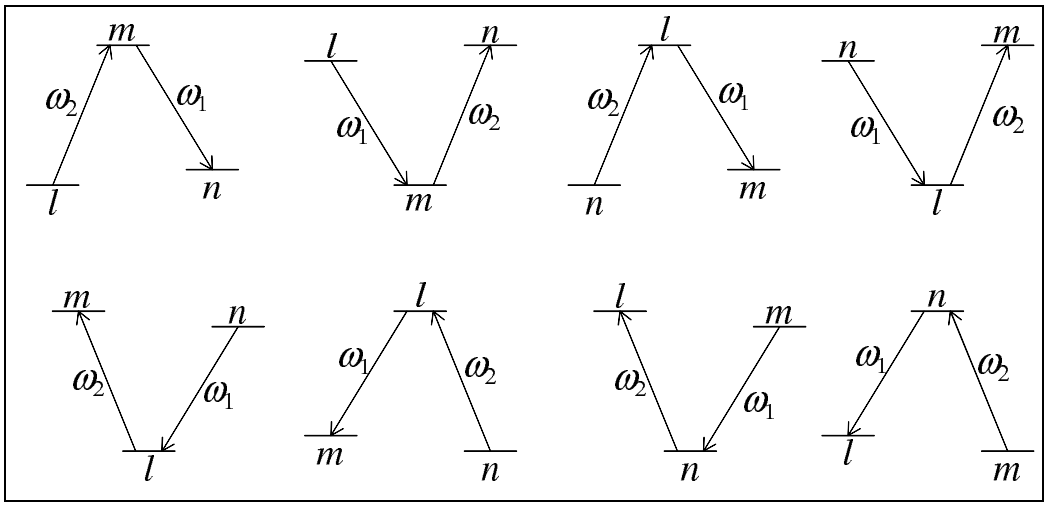


Figure 1. The combination of atomic levels and applied fields producing resonances in the second order nonlinear susceptibility corresponding to $\omega_1 < 0$ and $\omega_2 > 0$. The sequence of combinations corresponds directly to the sequence of terms in equation 33.

Clearly, each term in equation 33 consists of two resonant denominators. If the frequencies ω_1 and ω_2 are close to atomic level spacing frequencies, only resonant interactions between the medium and the fields need be considered. Many of the terms can then be eliminated since they are far from resonance. In order to eliminate non-resonant terms, the relative energy configuration of atomic levels l, m and n that would ensure both fields were resonant is established, based on the signs of the field frequencies. Boyd [41] shows diagrams of the level configurations appropriate for two positive frequencies (i.e. absorptions) while Figure 1 shows the corresponding arrangement for $\omega_1 < 0$ and $\omega_2 > 0$.

At this stage, the dummy labels l, m and n can be assigned to the levels under consideration, effectively eliminating the summation over these three indices in equation 33 and matching terms to specific level configurations. Some terms may then be eliminated if some of the levels are unpopulated. Since only dummy index l holds equilibrium population ρ_{ll}^{eq} in equation 33, any combination where label l was allocated to an unoccupied level may be eliminated. The second order susceptibility is then the sum of contributions from the remaining configurations.

Alternatively, if the incident fields are not resonant with any transitions in the medium, none of the denominators is small and none of the terms can be eliminated for this reason. However, some simplification is possible by recognising that the imaginary parts of all the denominators in equation 33 can be ignored. This allows some terms to be combined, reducing the eight terms in equation 33 to six [41].

$$\chi^{(2)}_{ijk}\left(-\omega_\sigma;\omega_1,\omega_2\right) = \frac{N}{2\epsilon_0\hbar^2}\sum_{lmn}\rho^{\mathrm{eq}}_{ll}\left[\frac{\mu^i_{ln}\mu^j_{nm}\mu^k_{ml}}{\left(\omega_{nl}-\omega_1-\omega_2\right)\left(\omega_{ml}-\omega_2\right)}+\frac{\mu^i_{ln}\mu^k_{nm}\mu^j_{ml}}{\left(\omega_{nl}-\omega_1-\omega_2\right)\left(\omega_{ml}-\omega_1\right)}\right.$$
$$+\frac{\mu^j_{ln}\mu^i_{nm}\mu^k_{ml}}{\left(\omega_{nl}+\omega_1\right)\left(\omega_{ml}-\omega_2\right)}+\frac{\mu^k_{ln}\mu^i_{nm}\mu^j_{ml}}{\left(\omega_{nl}+\omega_2\right)\left(\omega_{ml}-\omega_1\right)}$$
$$\left.+\frac{\mu^k_{ln}\mu^j_{nm}\mu^i_{ml}}{\left(\omega_{ml}+\omega_1+\omega_2\right)\left(\omega_{nl}+\omega_2\right)}+\frac{\mu^j_{ln}\mu^k_{nm}\mu^i_{ml}}{\left(\omega_{ml}+\omega_1+\omega_2\right)\left(\omega_{nl}+\omega_1\right)}\right] \quad (34)$$

Equations 33 and 34 both satisfy the expected symmetry properties. Intrinsic permutation symmetry is a general, rigorous property which applies to all susceptibility orders regardless of the material system. This principle states that any component of the susceptibility tensor $\chi^{(n)}_{ij_1j_2\ldots j_n}\left(-\omega;\omega_1,\omega_2,\ldots,\omega_n\right)$ is invariant under any permutation of the parameter pairs $\left(j_1,\omega_1\right),\left(j_2,\omega_2\right),\ldots,\left(j_n,\omega_n\right)$. For the second order susceptibility, the only possible permutation is the exchange of Cartesian indices j and k and of the two frequencies ω_1 and ω_2. Both equations 33 and 34 can be seen to be unchanged when these changes are made.

When all relevant optical frequencies (i.e. all single photon frequencies and combinations of frequencies appearing in the susceptibilities) are far from resonance with transitions in the medium, so-called full permutation symmetry applies. This principle adds $(i,-\omega)$, i.e. the first Cartesian coordinate and the combined frequency, to the previous list of parameter pairs that may be exchanged. In this case, it can be seen that the full expression for $\chi^{(2)}$ (equation 33) does not satisfy this symmetry while the non-resonant expression (equation 34) does.

Except in unusual cases where the incident electric field strength is comparable to the electric field strength within the atom, nonlinear optical effects are significantly weaker than their linear counterparts. Indeed, the series expansions equations 1 or 6 will not converge if this is not the case. However, in applications such as the use of nonlinear effects to generate new wavelengths or the measurement of optical or structural properties, increasing the magnitudes of the effects would be advantageous. The general form of the nonlinear susceptibility illustrated by equation 33 shows that, apart from the obvious technique of increasing the number density, this may be achieved by using the resonant enhancement associated with a transition in the medium. Resonant enhancement may be associated with any of the single-photon transitions in the medium, corresponding to the denominators in equation 33 with a single field frequency ω_1 or ω_2. In addition, multiphoton resonance is possible if the output frequency $\omega_\sigma = \pm\omega_1 \pm \omega_2 \ldots \pm \omega_n$ is close to a transition in the medium. Although they are not present in $\chi^{(2)}$ (equation 33), higher-order susceptibilities $\chi^{(n)}$ include resonant denominators for all intermediate processes up to $(n-1)$-photon interactions, and all may be used to enhance the desired effect. Resonant enhancement of nonlinearity has been used in many different experiments, including the generation of squeezed states of light [53, 54], investigations of surface relaxation dynamics [55] and the generation of second harmonic radiation using quantum dots and quantum wells [56, 57].

4.1. Propagation in Nonlinear Media

When performing nonlinear optical spectroscopy in bulk media it is important to understand how waves propagate in a nonlinear medium. This section presents some of the basic properties of propagation in a nonlinear medium.

The polarisation induced in a medium by an electric field, in turn, affects the electric field, so that Maxwell's equations must be solved including the polarisation. Maxwell's equations are

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{35}$$

$$\nabla \times \mathbf{H} = \frac{\partial}{\partial t}(\epsilon_0 \mathbf{E} + \mathbf{P}) + \mathbf{J} \tag{36}$$

$$\nabla \cdot (\epsilon_0 \mathbf{E} + \mathbf{P}) = 0 \tag{37}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{38}$$

where $\mathbf{J}$ is the current density, $\mathbf{B}$ is the magnetic flux density and $\mathbf{H}$ is the magnetic field.

A wave equation in an insulator ($\mathbf{J} = 0$) that includes the polarisation as a source term can be derived from equations 35-38 [40, 58]

$$\left(\nabla \times (\nabla \times) + \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) \mathbf{E}(\mathbf{r}, t) = -\frac{1}{\epsilon_0 c^2} \frac{\partial^2 \mathbf{P}(\mathbf{r}, t)}{\partial t^2} \tag{39}$$

Following Shen [40], we assume that $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{P}(\mathbf{r}, t)$ can be decomposed into a set of infinite plane waves and that the polarisation can be separated out into linear and nonlinear terms

$$\mathbf{E}(\mathbf{r}, t) = \sum_l \mathcal{A}_l \exp(\mathbf{k}_l \cdot \mathbf{r} - \omega_l t) \tag{40}$$

$$\mathbf{P}(\mathbf{r}, t) = \mathbf{P}^{(1)}(\mathbf{r}, t) + \mathbf{P}^{(NL)}(\mathbf{r}, t) \tag{41}$$

$$\mathbf{P}^{(1)}(\mathbf{r}, t) = \epsilon_0 \sum_l \chi^{(1)}(\omega_l) \cdot \mathcal{A}_l \exp(\mathbf{k}_l \cdot \mathbf{r} - \omega_l t) \tag{42}$$

$$\begin{aligned} \mathbf{P}^{(NL)}(\mathbf{r}, t) &= \sum_{n \geq 2} \mathbf{P}^{(n)}(\mathbf{r}, t) \\ &= \sum_m \mathbf{P}^{(NL)}(\mathbf{k}_m, \omega_m) \end{aligned} \tag{43}$$

where $\mathcal{A}_l$ is a slowly varying complex amplitude. With $\epsilon(\omega_l) \equiv 1 - \chi^{(1)}(\omega_l)$, substituting from equations 40 - 43 into equation 39 and taking the time derivatives explicitly leads to

$$\left(\nabla \times (\nabla \times) - \frac{\omega^2}{c^2} \epsilon \right) \mathbf{E}(\mathbf{k}, \omega) = \frac{\omega^2}{\epsilon_0 c^2} \mathbf{P}^{(NL)}(\mathbf{k}_m, \omega_m = \omega) \tag{44}$$

If $\mathbf{P}^{(NL)}$ is reduced to $\mathbf{P}^{(n)} = \epsilon_0 \chi^{(n)} \mathbf{E}_1 \cdots \mathbf{E}_n$ then there should be n similar wave equations which together with equation 44 form a set of $n + 1$ coupled equations. In the

case of the second order polarisation and three waves $\mathbf{E}_1(\mathbf{k}_1,\omega_1)$, $\mathbf{E}_2(\mathbf{k}_2,\omega_2)$ and $\mathbf{E}(\mathbf{k},\omega=\omega_1+\omega_2)$ then

$$\begin{aligned}
\left(\nabla\times(\nabla\times)-\frac{\omega_1^2}{c^2}\epsilon_1\right)\mathbf{E}_1(\mathbf{k}_1,\omega_1) &= \frac{\omega_1^2}{\epsilon_0 c^2}\chi^{(2)}(\omega_1=-\omega_2+\omega):\mathbf{E}_2^*(\mathbf{k}_2,\omega_2)\mathbf{E}(\mathbf{k},\omega)\\
\left(\nabla\times(\nabla\times)-\frac{\omega_2^2}{c^2}\epsilon_2\right)\mathbf{E}_2(\mathbf{k}_2,\omega_2) &= \frac{\omega_2^2}{\epsilon_0 c^2}\chi^{(2)}(\omega_2=\omega-\omega_1):\mathbf{E}_1^*(\mathbf{k}_1,\omega_1)\mathbf{E}(\mathbf{k},\omega)\\
\left(\nabla\times(\nabla\times)-\frac{\omega^2}{c^2}\epsilon\right)\mathbf{E}(\mathbf{k},\omega) &= \frac{\omega^2}{\epsilon_0 c^2}\chi^{(2)}(\omega=\omega_1+\omega_2):\mathbf{E}_1(\mathbf{k}_1,\omega_1)\mathbf{E}_2(\mathbf{k}_2,\omega_2)
\end{aligned} \tag{45}$$

where the subscripts for $\chi^{(2)}_{ijk}$ have been omitted for clarity. Assuming that all fields vary only along the propagation direction z, equations 45 become

$$\begin{aligned}
\left(\frac{\partial^2}{\partial z^2}-\frac{\omega_1^2}{c^2}\epsilon_1\right)\mathcal{A}_1(z)\exp(j(\mathbf{k}_1 z-\omega_1 t)) &= \frac{\omega_1^2}{\epsilon_0 c^2}\chi^{(2)}(\omega_1=-\omega_2+\omega):\\
&\quad \mathcal{A}_2^*(z)\mathcal{A}(z)\exp(j((\mathbf{k}-\mathbf{k}_2)z-(\omega-\omega_2)t))\\
\left(\frac{\partial^2}{\partial z^2}-\frac{\omega_2^2}{c^2}\epsilon_2\right)\mathcal{A}_2(z)\exp(j(\mathbf{k}_2 z-\omega_2 t)) &= \frac{\omega_2^2}{\epsilon_0 c^2}\chi^{(2)}(\omega_2=\omega-\omega_1):\\
&\quad \mathcal{A}_1^*(z)\mathcal{A}(z)\exp(j((\mathbf{k}-\mathbf{k}_1)z-(\omega-\omega_1)t))\\
\left(\frac{\partial^2}{\partial z^2}-\frac{\omega^2}{c^2}\epsilon\right)\mathcal{A}(z)\exp(j(\mathbf{k}z-\omega t)) &= \frac{\omega^2}{\epsilon_0 c^2}\chi^{(2)}(\omega=\omega_1+\omega_2):\\
&\quad \mathcal{A}_1(z)\mathcal{A}_2(z)\exp(j((\mathbf{k}_1+\mathbf{k}_2)z-(\omega_1+\omega_2)t))
\end{aligned} \tag{46}$$

Applying the slowly varying amplitude approximation ($|\partial\mathcal{A}^2/\partial z^2|\ll|\partial\mathcal{A}/\partial z|$), multiplying both sides by $\exp(j(\omega t-\mathbf{k}z))$ and some algebra reduces equations 46 to

$$\begin{aligned}
\frac{\partial\mathcal{A}_1(z)}{\partial z} &= -j\frac{\omega_1^2}{\epsilon_0 k_1 c^2}\chi^{(2)}(\omega_1=-\omega_2+\omega):\mathcal{A}_2^*(z)\mathcal{A}(z)\exp(j\Delta\mathbf{k}z)\\
\frac{\partial\mathcal{A}_2(z)}{\partial z} &= -j\frac{\omega_2^2}{\epsilon_0 k_2 c^2}\chi^{(2)}(\omega_2=\omega-\omega_1):\mathcal{A}_1^*(z)\mathcal{A}(z)\exp(j\Delta\mathbf{k}z)\\
\frac{\partial\mathcal{A}(z)}{\partial z} &= j\frac{\omega^2}{\epsilon_0 k c^2}\chi^{(2)}(\omega=\omega_1+\omega_2):\mathcal{A}_1(z)\mathcal{A}_2(z)\exp(j\Delta\mathbf{k}z)
\end{aligned} \tag{47}$$

where $\Delta\mathbf{k}$ is the wave vector mismatch

$$\Delta\mathbf{k}=\mathbf{k}-\mathbf{k}_2-\mathbf{k}_1 \tag{48}$$

The phase mismatch $\Delta\mathbf{k}$ prevents the efficient transfer of energy between the three fields involved. However, under some circumstances $\Delta\mathbf{k}$ = 0 (referred to as phase matched) can be achieved in which case energy transfer between the fields is much more efficient. Although phase matching is not usually an important consideration in most nonlinear optical spectroscopy work we present a short description of it here since it can affect the results of experimental work done on bulk media. Considering the case of SHG, the gain experienced by the second harmonic is a maximum when the phase mismatch $\Delta\mathbf{k}$, is zero. Normal

dispersion typically does not allow the phase mismatch to be zero when photon energy is conserved ($\omega = \omega_1 + \omega_2$). This can be illustrated by examining the phase matching condition for SHG. Energy conservation requires that $\omega_{sh} = 2\omega$ and phase matching requires that $\mathbf{k}_{sh} = 2\mathbf{k}$. The phase matching condition can be rewritten in terms of the angular frequency, $c/(n_{sh}\omega_{sh} = 2c/(n\omega)$, thus the phase mismatch will be zero when $n_{sh} = n$ and this cannot be achieved in an isotropic, dispersive medium since its refractive index increases with increasing frequency. However in anisotropic material where $\mathbf{k}$ depends on the direction and polarisation of the wave, phase matching can, in some cases, be achieved.

In anisotropic media the refractive index is represented by a 3 x 3 tensor n_{ij}. By choosing an appropriate coordinate transformation n_{ij} can be made diagonal and define the principal axes of the crystal. In this orientation each n_{ii} can be mapped to a particular n_x, n_y and n_z, which correspond to the refractive index for beams travelling in the x, y and z directions respectively. Crystals can be classified into isotropic ($n_x = n_y = n_z$), uniaxial ($n_x = n_y \neq n_z$) and biaxial ($n_x \neq n_y \neq n_z$) classes. Crystal classes are further distinguished as positive or negative depending on the size of each component of the refractive index. Thus a positive uniaxial crystal is defined by $n_x \geq n_y > n_z$ or more usefully $n^{\mathbf{e}} > n^o$, where $n^{\mathbf{e}}$ and n^o (see next paragraph) denote the refractive index for extraordinary and ordinary polarisations. Likewise a negative uniaxial crystal is defined by $n_x \leq n_y < n_z$ (or $n^{\mathbf{e}} < n^o$). The biaxial definition for a positive crystal is $n_x > n_y > n_z$. Biaxial crystals are not discussed here, a more comprehensive discussion of anisotropic media can be found in various reference texts (e.g., Sutherland [59] or Dmitriev [60]).

A uniaxial crystal has a single optical axis and a single principal plane. The optical axis is defined by the direction for which $n^{\mathbf{e}} = n^o$ and is labelled the z direction. The principal plane is defined the optical axis and the wave vector of the beam. The extraordinary ray is polarised in the principal plane and its refractive index varies with direction as given by

$$n^e(\theta) = n^o \sqrt{\frac{1 + \tan^2\theta}{1 + (\frac{n^o}{n^{\mathbf{e}}})^2 \tan^2\theta}} \tag{49}$$

where θ is the propagation angle relative to the optical axes. The ordinary ray is polarised orthogonal to the principal plane and its refractive index is n^oindependent of propagation angle (see Figure 2).

In a negative uniaxial crystal with input waves ($\omega_1 > \omega_2 \geq \omega_3$) phase matching can be achieved provided [1]

$$|n^o(\omega_1) - n^{\mathbf{e}}(\omega_1)| \geq \left|n^o\big(\omega_1 - \frac{n^o(\omega_2)\omega_2}{\omega_1} - \frac{n^o(\omega_3)\omega_3}{\omega_1}\big)\right| \tag{50}$$

The physical interpretation of equation 50 is that there will be a set of frequencies such that $\mathbf{k}_1^e$ is equal to $\mathbf{k}_2^o + \mathbf{k}_3^o$ when propagating collinearly at a particular angle θ_{pm}, relative to the optic axes. The frequencies ω_2 and ω_3 can be tuned by rotating the crystal relative to $\mathbf{k}_1^e$. This particular birefringent phase matching scheme is called Type I phase matching. If the crystal were a positive uniaxial crystal then ω_1 would be the ordinary wave while ω_2 and ω_3 would be extraordinary waves. It is also possible to phase match interactions where ω_2 and ω_3 are of mixed polarisation (e.g. $\mathbf{k}_1^e = \mathbf{k}_2^e + \mathbf{k}_3^o$). This so-called Type II birefringent phase matching requires less birefringence in the case of difference frequency generation and so

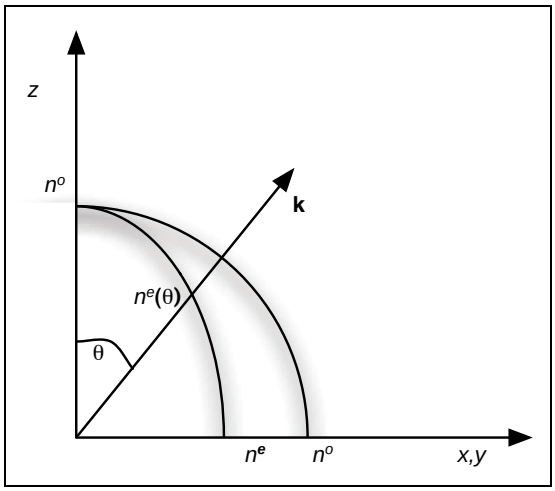


Figure 2. Index ellipsoids for a negative uniaxial crystal in the principal plane defined by the z axis and $\mathbf{k}$. The refractive index experienced by the extraordinary ray varies with direction (given by the partial ellipse). The refractive index experienced by the ordinary ray is independent of direction (given by the partial circle).

may be more readily achievable. Thus far we have limited the discussion to situations where all the wave vectors are parallel, referred to as collinear phase matching but $\Delta\mathbf{k}$ is a vector relationship so the interaction is not necessarily collinear. Phase matching in the case where the wave vectors are not collinear is called non-collinear phase matching.

In either collinear or non-collinear phase matching the length over which the interaction occurs is limited by either the angle between non-collinear components or by Poynting vector walk off. For the extraordinary wave $\mathbf{E} \times \mathbf{H}$ is not parallel to $\mathbf{k}$ so even though the wave vectors may be collinear the waves will no longer possess sufficient spatial overlap to continue interacting after a distance l_w, given by [61]

$$l_w = \frac{\sqrt{\pi}}{2}\frac{w_1}{\rho}\sqrt{\frac{w_1^2 + w_2^2}{w_1^2 + \frac{w_2^2}{2}}} \tag{51}$$

where w_n is the (1/e^2) radius of the n^{th} wave and ρ is the angle between the extraordinary and ordinary component, which is given by [60]

$$\rho(\theta_{pm}) = \pm \arctan\left(\left(\frac{n^o}{n^{\mathbf{e}}}\right)^2 \tan\theta_{pm}\right) \tag{52}$$

In general, spectroscopic applications of optical nonlinearity have interaction lengths that are very short compared to the interaction length given by equation 51, thus phase matching is not usually a consideration. However, phase matching can be significant in some forms of bulk spectroscopy where the aggregate interaction length can be quite long.

4.2. Optical Nonlinearity of Fluids and Weakly Ordered Systems

The optical nonlinearity of fluids, thin films and what might be described as weakly ordered systems (i.e., polymers in solution) are often measured using a technique called Hyper Rayleigh scattering (HRS). HRS is a method for measuring the first hyper-polarisability of a molecular substance. Unlike linear Rayleigh scattering, the laser light is not only scattered but it may also have its frequency changed by the nonlinear response of the induced electric dipole moment of the molecule. If the sample is illuminated by a monochromatic laser source of high intensity the dominant nonlinear scattering is at twice the frequency. A theoretical understanding of the molecular HRS response is given below.

Consider a molecule held at a fixed orientation to a laser beam. The electric dipole response of the molecule to the incident light field is given by [62]

$$\mu_i = \sum_j \alpha_{ij} E_j + \sum_{j,k} \beta_{ijk} E_j E_k \tag{53}$$

In this representation μ_i is the induced dipole moment along the i axis of the molecule. The linear polarisability is given by α_{ij} and the first hyper-polarisability is given by β_{ijk}. E_j and E_k are the applied fields which in the particular case of HRS are identical. The electric field at the molecule is given by

$$E_j = E_{0_j} \exp(j\omega_j t) \tag{54}$$

The bulk second order nonlinearity of the medium is given by a vector summation of β_{ijk} which in an isotropic medium has an average value of zero. However, orientation fluctuations will cause a non zero nonlinear susceptibility over a small volume and the local nonlinear susceptibility will be correlated only over distances short compared to the wavelength of light. Consider two volumes labelled 1 and 2. The intensity of light scattered at twice the frequency will be proportional to $\langle \chi^{(2)}_{ijk_1} \chi^{(2)}_{jmn_2} \rangle$. If we assume that each photon with energy twice that of the fundamental has undergone only a single scattering event (i.e. interacted with only a single molecule or aggregate) then we can replace the bulk nonlinear susceptibility with the molecular first hyper-polarisability. It is also reasonable to assume that for small molecules that are weakly interacting the moment of the first hyper-polarisability will be uncorrelated between molecules on average. Thus the HRS signal is generated from molecules that are correlated during the laser pulse, however, one would not expect those molecules to remain correlated. It should be noted that some of the most interesting results produced by HRS measurements come from systems where correlations must be considered [63, 64].

Since HRS is usually performed in liquid samples, the response must be averaged over the random orientations of the molecules. Thus the second harmonic output is given by

$$E(2\omega) \propto \langle \beta^2 \rangle E(\omega) E^*(\omega) \tag{55}$$

If one takes into account the imperfect light gathering ability of the apparatus used to perform the HRS experiment and the sensitivity of the detector then the recorded intensity of the second harmonic output will be

$$I(2\omega) = GN \langle \beta^2 \rangle I^2(\omega) \tag{56}$$

where G is referred to as the instrument factor. Separating out the response of both the solvent and solute gives

$$I(2\omega) = G(N_1\langle\beta_1^2\rangle + N_2\langle\beta_2^2\rangle)I^2(\omega) \tag{57}$$

where N_1 and β_1 are the number density and first hyper-polarisability of the solvent while N_2 and β_2 are the number density and first hyper-polarisability of the solute. For dilute solutions the change in number density of the solvent as the solute concentration is varied is very small and may be neglected. Changes in HRS intensity are only due to the change in solute concentration. Equation 57 can therefore be written

$$I(2\omega) = C + GN_2\langle\beta_2^2\rangle I^2(\omega) \tag{58}$$

where C is the solvent contribution. Thus the intensity of the second harmonic is linearly proportional to the concentration of the solute as long as the molecules that make up the solute are not spatially correlated.

Hendrickx et al. have also considered the effect of correlations in HRS [64]. Consider a laser beam of intensity I_w propagating in the u direction, where the subscript w indicates the polarisation of the beam. Then the second harmonic light is given by

$$I_{s,w} = GB\langle\beta_{www}^2\rangle I_w^2 \tag{59}$$

where B is given by

$$B = \frac{32\pi^2}{\epsilon_0^3 c\lambda^4 r^2} \tag{60}$$

and $\langle\beta_{www}\rangle$ is the first hyper-polarisability averaged over the direction cosines employed to transform from the molecular reference frame to the laboratory reference frame. Usually HRS is performed using a polarisation insensitive detector. Thus, the measured first hyper-polarisability is the sum of contributions from the two orthogonal polarisations.

In the macroscopic view, the medium consists of a large number of scattering molecules positioned over a region of space that is large compared to the wavelength of the illuminating laser. The polarisation of the molecules induced by the incident light will have a phase difference due to their position. If we assume a reference molecule labelled i at the origin and oriented in the u direction, the field scattered will be given by.

$$|\mathbf{E}_{i,p}(2\omega)| = |\mathbf{E}_i(2\omega)|\cos\delta_i \tag{61}$$

$$|\mathbf{E}_{i,q}(2\omega)| = |\mathbf{E}_i(2\omega)|\sin\delta_i \tag{62}$$

where equation 61 is the contribution that is in phase with the reference molecule and equation 62 is the contribution that is at a phase angle of 90 (i.e. in phase quadrature). The total in phase and phase quadrature fields are given by

$$|\mathbf{E}_p(2\omega)| = |\mathbf{E}_1(2\omega)|\sum_i^N \cos\delta_i \tag{63}$$

$$|\mathbf{E}_q(2\omega)| = |\mathbf{E}_1(2\omega)|\sum_i^N \sin\delta_i \tag{64}$$

where we have assumed that all the molecules in solution are identical thus the magnitude of their scattering contributions are identical and N is the total number of molecules. The total scattered intensity is given by summing equations 63 and 64 in quadrature

$$I_s = |\mathbf{E}_1(2\omega)|^2 \left[\Big(\sum_i^N \cos\delta_i \Big)^2 + \Big(\sum_j^N \sin\delta_j \Big)^2 \right] \tag{65}$$

For $i = j$ the cosine and sine squared terms sum to one so the incoherent contribution is given by

$$I_s \propto |\mathbf{E}_1(2\omega)|^2 N \tag{66}$$

which is just the result given in equation 56 with the constant factors removed. When the cross terms in the summations of equation 65 are included the intensity of the scattered light is given by

$$I_s \propto |\mathbf{E}_1(2\omega)|^2 \left[N + \sum_{j\neq i}^N \cos(\delta_i - \delta_j) \right] \tag{67}$$

When there are correlations between scattering molecules the second term cannot be ignored. Thus, in terms of the first hyper-polarisability the coherent term is given by

$$I_{s,w} = G I_w^2 \sum_{j\neq i}^N \langle \beta_{www,i} \beta_{www,j} \rangle \tag{68}$$

$$I_{s,u} = G I_w^2 \sum_{j\neq i}^N \langle \beta_{uww,i} \beta_{uww,j} \rangle \tag{69}$$

where we have ignored local field correction factors. These correlation terms can either be calculated explicitly using a two particle correlation function or one can replace each group of correlated molecules with a single entity (with a new first hyper-polarisability and a reduced concentration). The HRS signals from these entities are uncorrelated and standard HRS theory can be used.

In section 5. we review some recent, novel work using HRS to explore the dynamics of both liquid and solid media.

4.3. Optical Nonlinearity in Bulk Turbid Media

In much bulk nonlinear optical spectroscopy the sample is very disordered and strongly scattering. Thus the description of propagation, energy transfer and phase matching must be approached statistically. Here we present the response of a system of particles to monochromatic laser radiation. This is well understood and we follow the approach of Sutherland [65].The response of a single spherical particle, $I_{2\omega}$ goes as

$$I_{2\omega} \propto d^2 L_c^2 \sin\left(\frac{\pi r}{2 L_c} \right) \tag{70}$$

where $d = \frac{1}{2}\chi^{(2)}$, r is the particle radius and L_c is length over which constructive interference between the fundamental and its second harmonic occurs (called the coherence

length). The response of a system of particles with an average size $\langle r \rangle$ depends on whether the particles can be phase matched. In the case where the particles can be phase-matched, the response will be dominated by those particles for which the orientation of the crystal axis relative to the laser beam coincides with the phase matching angle. The coherence length becomes the average of

$$L_c = \frac{\lambda_\omega}{4(n^e_{2\omega} - n^o_\omega)\sin\theta_{pm}(\theta - \theta_{pm})} \tag{71}$$

where λ_ω is the fundamental wavelength, θ is the laser angle of incidence relative to the optical axis of the crystal and θ_{pm} is the phase-matching angle relative to the optical axis. Provided that $\langle r \rangle \gg \frac{\lambda_w}{2n_w \sin\rho \sin\theta_{pm}}$ the second harmonic response goes as

$$I_{2\omega} \propto d^2_{eff} L \Lambda_{pm} \tag{72}$$

where $\frac{\lambda_w}{2n_w \sin\rho \sin\theta_{pm}}$ is the length over which efficient SHG can occur and d_{eff} is the combination of d tensor elements in the phase-matching direction and L is the path length of the sample. Equation 72 shows that the particle size is not important but the length over which the laser beam can interact with the second harmonic is. When $\langle r \rangle \leq \frac{\lambda_w}{2n_w \sin\rho \sin\theta_{pm}}$, each particle emits some fraction of second harmonic which is partially correlated with that emitted by the others. The nonlinear polarisation can then be treated like a one dimensional random walk. The result is simply proportional to the size of particles

$$I_{2\omega} \propto d^2_{eff} \frac{L}{L_c} \frac{\langle r \rangle}{L_c} \tag{73}$$

If the particles cannot be phase-matched and are much larger than the coherence length, then the response of each particle has no phase coherence to any other. In this case one can simply sum the intensities by multiplying the single particle response (equation 70) by the number of particles. Since the orientation of the particles is random one must average d as well. The resulting response is

$$I_{2\omega} = \langle d^2 \rangle L \langle L_c \rangle \frac{\langle L_c \rangle}{\langle r \rangle} \tag{74}$$

where $\langle L_c \rangle$ is the average coherence length for the crystal type. For the case of $\langle r \rangle \leq \langle L_c \rangle$ the same argument as used for equation 73 can be applied but with d^2_{eff} replaced with $\langle d^2 \rangle$.

$$I_{2\omega} \propto \langle d^2 \rangle \frac{L}{L_c} \frac{\langle r \rangle}{L_c} \tag{75}$$

Thus the dependence on particle size of the response of a powdered sample is expected to be similar to that shown in Figure 3. In reality the response is smooth due to the spread of particle sizes in the sample.

The results presented in equations 72-75 all have three common assumptions. These assumptions are that the particles are all of a single species and polymorph, that the gaps between particles are much smaller than L_c and that the particles are all immersed in an index matching fluid. It is relatively simple to extend equations 72-75 to include multiple

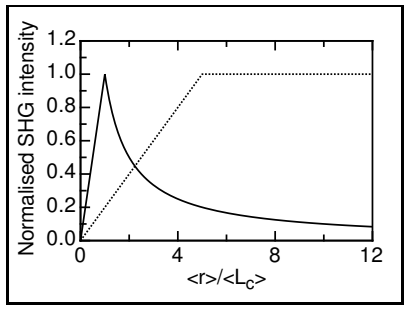


Figure 3. Theoretical response of phase-matched (dashed line) and non-phase-matched (solid line) particles as a function of size.

species. However, if the particles are not in an index matching fluid then L becomes a function of particle size, refractive index and experimental geometry. In section 8. we present experimental results showing some of the applications of nonlinear optical spectroscopy in characterising bulk turbid media.

5. Hyper Rayleigh Scattering

HRS is a method for measuring the first hyper-polarisability (β) of a molecule in solution [62]. In general a pulsed laser is used to illuminate a liquid sample (as shown in Figure 4), light is collected from the sample at right angles to the incoming laser light and is carefully filtered to remove any two photon fluorescence or residual pump light. The intensity of the second harmonic signal is proportional to the number of molecules in the solution and the size of the first hyper-polarisability of those molecules. One can measure β by stepwise diluting the concentration of the molecule of interest and measuring the second harmonic intensity. The second harmonic intensity at zero concentration is due to the first hyper-polarisability of the solvent and the slope is used to calculate the first hyper-polarisability of the molecule of interest. Examples of HRS measurements and calculations of β can be found throughout the literature (see e.g. [66–68]). Although this form of HRS measures the orientation average, $\langle\beta\rangle$, more precise information can be gathered by measuring the intensity of the s and p polarised second harmonic as a function of the polarisation of the illuminating laser light [69]. Some more subtle uses of HRS are presented in the remainder of this section.

HRS has been used to study inhomogeneous thin films of Pb $_x$($Zr_{0.53}Ti_{0.47}$)O_3 [63]. In these experiments Melnikov et al. measured the angular dependence of the HRS signal and found that there was a coherent SHG contribution at the angle of specular reflection. The incoherent HRS signal was found to be a broad peak with respect to the angle of measurement. However, by reducing the spot size of the illuminating laser beam a speckle pattern emerged from the broad peak. This structure was predicted by using a one dimensional Markov model to create the two particle correlation function required for solving equation 69. Appropriate corrections for the finite spot size of the laser beam were also included in the calculation. The model developed by Melnikov et al. has only two free parameters, the average cross section of the facets (i.e. the length over which the HRS signal is absolutely

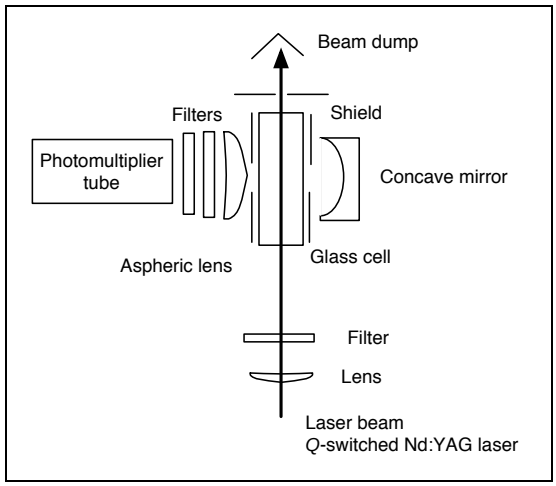


Figure 4. Generalised HRS experimental set up. The cell is constructed of glass to prevent SHG from the cell walls. Fluorescence from the cell walls and the beam dump are blocked by shields and the signal is further filtered through coloured glass and interference filters.

correlated) and the (anti) correlation between contributions from different facets.

The size dependence of CdS nanoparticles has been studied by HRS [70]. CdS nanoparticles were manufactured at size fractions between 2.8 and 5.0 nm. The solution was illuminated with a Q-switched Nd:YAG laser emitting pulses 10 ns in duration with pulse energies up to 3 mJ at a repetition rate of 10 Hz. Initially the first hyper-polarisability of the nanoparticles was measured. It was found to decrease as a function of volume to a minimum near 3.0 nm. However, the first hyper-polarisability increases slight for particles smaller than 3.0 nm. A possible explanation for this is due to the size dependence of the optical bandgap and oscillator strength, however, when Zhang et al. considered the two level model of the first hyper-polarisability they concluded that this is not so. Thus they concluded that the increased response is because the surface response of the particles starts to dominate the bulk response. When these results are expressed as a function of the number of molecules per particle it is found that the first hyper-polarisability monotonically decreases as a function of particle volume, which supports their conclusion that the measured increase in the first hyper-polarisability was due to the increased dominance of the nanoparticle surface.

HRS has been used to determine the critical micelle concentration (CMC) of surfactants [71]. Ghosh et al. attached chromophores to hydrocarbon chains to create surface active molecules. They then measured the HRS signal as a function of concentration and found that there was a peak in the HRS signal at the CMC (see Figure 5). It is notable that the dependence of HRS intensity on concentration is fairly linear below the CMC. Above the CMC the HRS signal initially drops and then increases linearly again. This indicates that the equilibrium constants governing the bulk concentration are different above and below the CMC. Ghosh et al. attribute the drop in HRS signal at concentrations slightly higher than the CMC to the formation of smaller aggregates such as dimers and trimers.

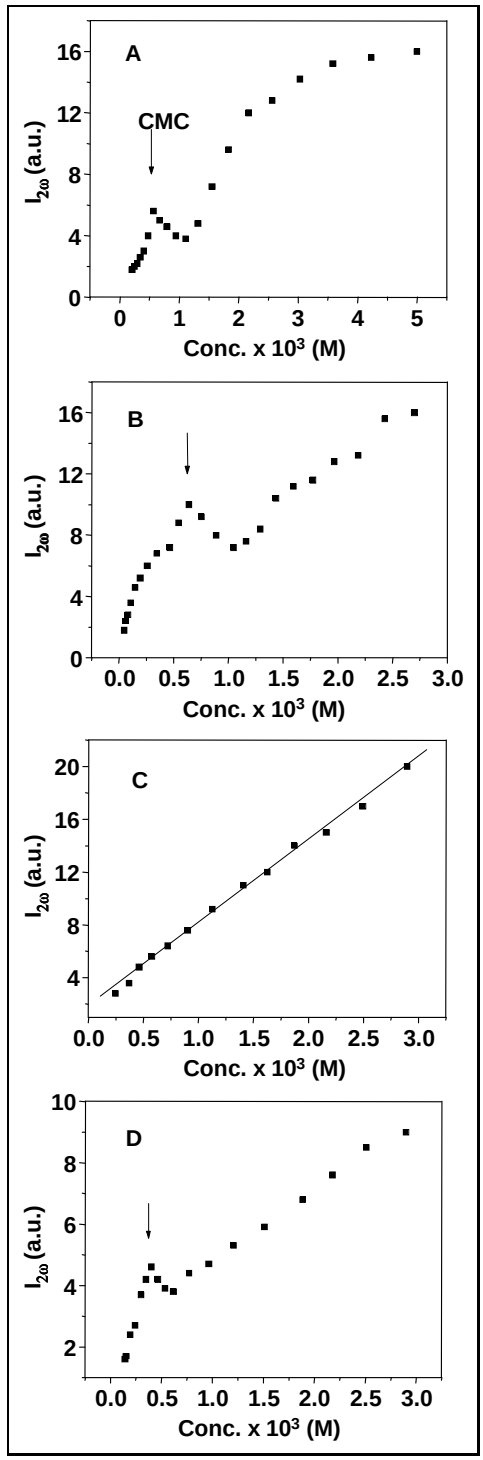


Figure 5. HRS signal as a function of surfactant concentration of four surfactants. [1]

[1]Reprinted with permission from S. Ghosh, A. Krishnan, P. K. Das, S. Ramakrishnan, Journal of American Chemical Society, **125**, pp 1602 (2003). Copyright (2003), American Chemical Society

The experiments by Ghosh et al. show that HRS is a viable method for studying the properties of surfactant molecules and that this may lead to a better understanding of the dynamical behaviour of such systems.

The HRS response of DNA and other biological macromolecules is also of interest. In this case we do not have adequate experimental data to present, however, what follows is the theoretical modelling which we will compare with our experimental results. In a physical experiment a sample of DNA would have its HRS response measured at fixed concentration. Electric fields would then be used to manipulate the ternary structure of the DNA. Since the first hyper-polarisability of a DNA strand is a geometric sum of the first hyper-polarisability of the monomer units we would expect from Markov statistics that $\beta \propto \sqrt{N}$ where N is the number of monomer units in the DNA strand. However, if electric fields are used to stretch the DNA out then the measured HRS signal will increase as the strand unfolds. For this to be useful in characterising DNA strands the behaviour of DNA under the influence of electric fields must be understood. Here we examine computationally the behaviour of DNA under the influence of electric fields. The model used is a very simple chain on a lattice where the DNA strand can assume any conformation provided it does not break the chain or have two monomers occupying the same lattice site. We then allow the DNA to move, and after each time step calculate the bounding ellipsoid and its orientation. The resulting simulations showed that the DNA stretches out at a fixed orientation to the field and that larger strands take longer to do so. An HRS experiment performed on DNA under the influence of an electric field should result in a time dependent signal which could then be deconvolved to obtain information about the DNA conformation and size [73].

HRS experiments on DNA strands also offer interesting possibilities for examining the molecular dynamics of long chain polymers. The first HRS measurements on ogilonucleotides (single stranded chains of nucleotides) were performed by Rinuy [74]. In these experiments the individual hyper-polarisability of each nucleotide was measured separately. Then single strands of varying monomer number (all of the same nucleotide) were manufactured and measured. It was also found that complementary strands allowed to combine into DNA resulted in a large drop in the HRS signal. The hydrogen bonding between complementary base pairs results in a reduced dipole moment for both monomer units and hence a reduced first hyper-polarisability. However, it was also found that the presence of intercalating dyes can substantially increase the measured first hyper-polarisability.

Although HRS is a well established experimental technique its use has mainly been limited to evaluating potential chromophores. Here, we have shown several novel experiments which use molecular nonlinear response to measure other interesting behaviour.

6. Second Harmonic Generation at a Surface

Second harmonic generation at a surface has been thoroughly explored in the last 10 years. It is an ideal technique for separating out the surface properties from bulk properties in centro-symmetric materials such as liquids and metals. A general experimental diagram is shown in Figure 6 to help illustrate the results from the following review of recent work.

Surface SHG has been used to examine the dynamic behaviour of an oscillating bubble [75]. The experiments were carried out using the second harmonic of a mode locked Nd:YAG laser (532 nm), thus giving a high intensity while reducing the Joule heating of

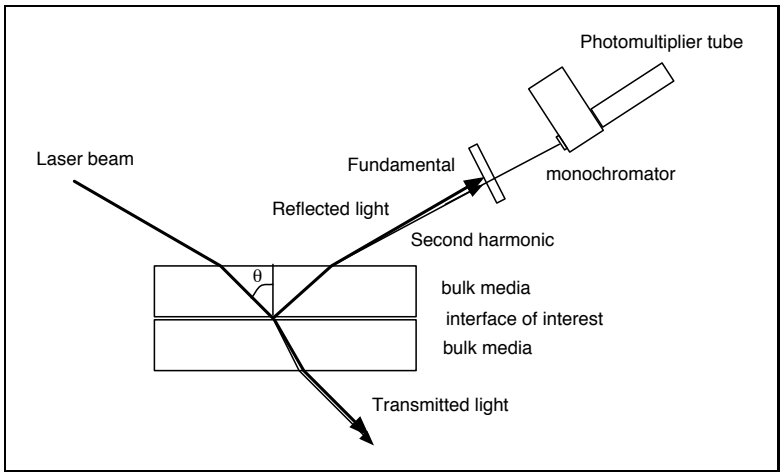


Figure 6. Generalised surface SHG experimental set up. The angle of incidence θ can be chosen such that total internal reflection occurs at the interface thus eliminating the transmitted beam.

the sample compared to using a Q switched laser. The resulting SHG signal (266 nm) was detected using a solar blind photomultiplier tube, thus reducing the potential for spurious signals directly from the illuminating laser. The solution from which the bubble was produced consisted of water and a surfactant that contained a strong chromophore. Thus the SHG signal was directly proportional to the surface concentration of the chromophore. The bubble itself was attached to the end of a capillary tube and the pressure in the capillary tube could be controlled in an arbitrary manner thus causing the size of the bubble to oscillate. The laser was triggered in such a way as to illuminate the bubble at a fixed phase in the pressure oscillation, thus the second harmonic signal as a function of surface tension could be obtained (see Figure 7). From these measurements Örtegren et al. were able to measure the effect of oscillation frequency on bulk and surface surfactant concentrations and the molecular exchange dynamics at the surface.

Second harmonic generation at interfaces has been employed as an imaging technique akin to fluorescence microscopy (see Figure 8). Campagnola et al. demonstrated this technique by adapting a laser scanning microscope [76]. The illuminating source was a mode-locked Ti:sapphire laser producing 100 fs pulses and both the two photon fluorescence and the second harmonic were imaged and their performance compared. SHG imaging with a mode-locked laser has the advantage of having very high intensity which is due to the very short laser pulse and is necessary for the observation of nonlinear optical activity while delivering very little energy to the specimen, thus photobleaching and the resulting phototoxicity of the chromophore dyes can potentially be reduced. It was found that the chirality of the chromophore had an unexpectedly large effect on the intensity of the SHG signal. Campagnola et al. speculate that the presence of a chiral molecule relaxes the electric dipole selection rule. Although it was found that the use of nonlinear optical imaging has

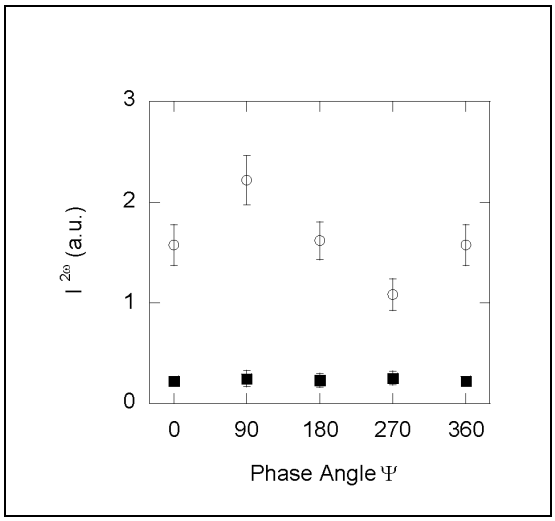


Figure 7. The dependence of SHG intensity on the phase and hence surface tension of an oscillating bubble. Circles are for a surfactant concentration of 10 μM while squares are for water only.[2]

no significant advantage in resolution over related techniques such as two-photon excited fluorescence, it does demonstrate increased sensitivity to changes in membrane potentials.

Molecular adsorption at various interfaces has been studied by analysing the intensity of the surface second harmonic generation. Rinuy et al. studied the adsorption of the glucose oxidase enzyme at an air water interface [77]. Glucose oxidase is a protein whose optical nonlinear activity is related to a flavin adenine dinucleotide which is bound to the protein through hydrogen bonding. The pump source for the experiment was an optical parametric oscillator (OPO) based on a β barium borate crystal and pumped by an Nd:YAG laser. The idler output of the OPO was tuned between 840 and 980 nm which allowed the second harmonic to be resonant with peaks in the UV-visible absorption spectrum. The OPO idler was focused on the air water interface at an angle sufficiently shallow for total internal reflection to occur. It was found that the peak in second harmonic intensity occurred at 450 nm which corresponded with one of the peaks in the absorption spectrum. The second harmonic intensity at a fixed wavelength and OPO output energy were directly proportional to the number of protein molecules at the interface. The bulk solution concentration was increased from 0 to over 500 nM while monitoring the intensity of the second harmonic radiation. The peak intensity was reached at 400 nM and thereafter decreased, indicating that the surface proteins were interacting with each other. This was confirmed by monitoring the change in depolarisation of the second harmonic above and below the concentration at which peak second harmonic generation occurred (see Figure 9). Marrucci et al. [78] have used a similar experiment to observe the molecular adsorption of a surfactant solubilised in a liquid onto a titanium dioxide solid substrate (see Figure 10). A mode locked Ti:sapphire laser with a pulse duration of 120 fs, a repetition rate of 1 kHz and set to a centre frequency

[2]Reprinted from the Journal of Colloid and Interface Science, Vol **279**, No 1, J. Örtegren, K.-D. Wantke, H. Motschmann, H. Möhwald, *A study of kinetic molecular exchange processes in the medium frequency range by surface SHG on an oscillating bubble* , pp 266-276, Copyright (2004), with permission from Elsevier

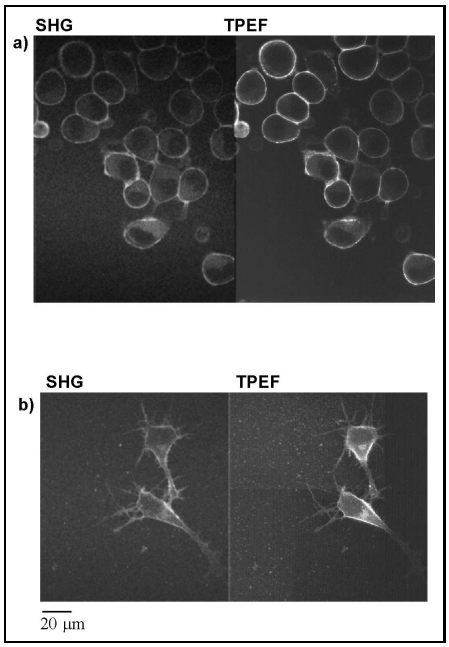


Figure 8. Images of undifferentiated (subfigure a) and differentiated (subfigure b) neuroblastoma cells. The left hand images are from the SHG signal and the right hand images are from two photon fluorescence. The cells were stained with JPW-2080. [3]

of 800 nm illuminated the electrode. The polarisation of the laser beam and the second harmonic were controlled and filtered before and after surface interrogation was performed. The intensity of the coherent SHG signal was proportional to the surfactant concentration and can be well fitted to existing models.

Ishida et al. have also used second harmonic generation to observe adsorption at a surface [79]. The goal of this study was to gain information about the surface electronic states of copper on exposure to benzene or carbon monoxide (see Figure 11). Normal techniques for studying surface electronic states generally involve exciting and measuring photon emission, however, due to its surface specificity second harmonic generation is also a good candidate. In studies on copper Ishida et al. employed an optical parametric generator and amplifier combination, based on β barium borate and pumped by an amplified, mode-locked Ti:sapphire laser. The resulting source was tunable from 470-700 nm and was p polarised. The p polarised SHG signal displayed a prominent peak at 251 nm which is attributed to the surface electronic states of the copper. This resonance was quickly suppressed by even a small amount of benzene (less than one monolayer). These data allowed Ishida et al. to determine the preferential sites for benzene adsorption and to observe previously unobserved electronic states.

[3]Reprinted from the Biophysical Journal, Vol **77**, P. J. Campagnola, M. Wei, A. Lewis, L. M. Loew,*High resolution nonlinear optical imaging of live cells by second harmonic generation* , pp 3341-3349, Copyright (1999), with permission from the Biophysical Society

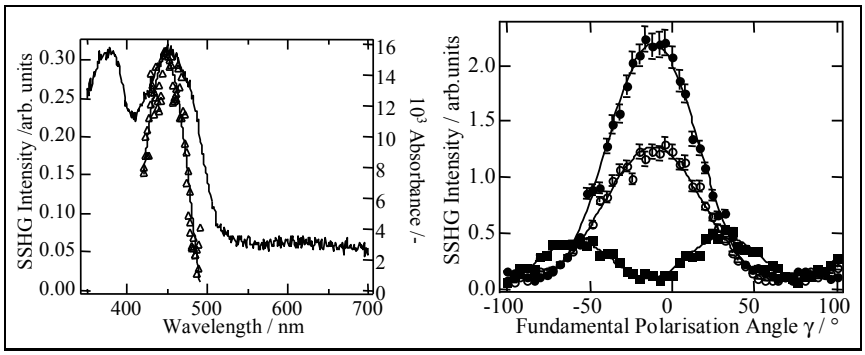


Figure 9. Subfigure a: Glucose oxidase SH spectrum. Subfigure b: Glucose oxidase polarisation dependence; filled and unfilled circles are for p-polarised light at concentrations of 300 nM and 550 nM respectively. The filled squares are s-polarised light for a concentration of 300 nM.[4]

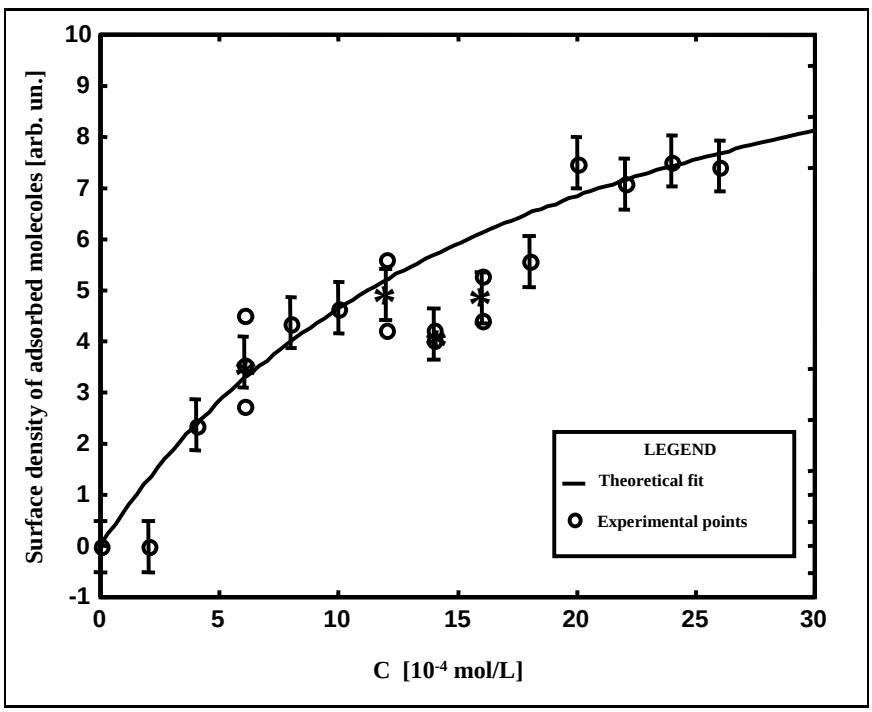


Figure 10. Surface density of quinolinium as a function of bulk concentration. [5]

[4]Reprinted from Biophysical Journal, Vol **77**, J. Rinuy, P. F. Brevet, H. H. Girault,*Second harmonic generation of glucose oxidase at the air/water interface* , pp 3350-3355, Copyright (1999), with permission from the Biophysical Society

[5]Reprinted from Optics and Lasers in Engineering, Vol **37**, No 5, L. Marrucci, D. Paparo, G. Cerrone, C. de Lisio, E. Santamato, S. Solimeno, S. Ardizzone, P. Quagliotto, *Probing interfacial properties by optical second-harmonic generation* , pp 601-610, Copyright (2002), with permission from Elsevier

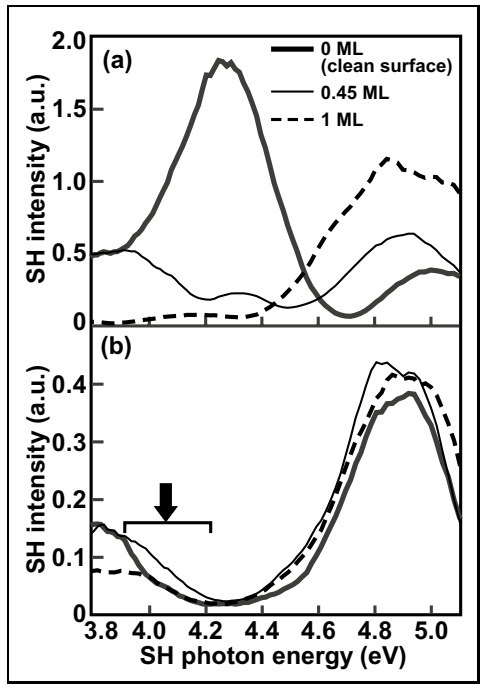


Figure 11. Surface absorption of carbon monoxide at the Cu(111) surface. Subfigure (a)[6].

SHG can be also used to monitor semiconductor production quality, in particular diffusion of copper at critical areas such as surfaces and interfaces [80]. The diffusion of copper around the interconnects is very important in device degradation and it is believed that surface diffusion is much faster than bulk diffusion [80] (and references therein). Thus a surface specific technique was desirable in order to measure the significance of surface and interface diffusion compared to the already known bulk diffusion properties. The measurements were performed using a Nd:YAG laser with a pulse energy of 50 mJ and a 5 ns pulse duration at a 10 Hz repetition rate. The second harmonic light was used to create an image of the semiconductor sample and it was found that significant surface diffusion of copper occurred. However, for complete characterisation, calibration using other techniques such as transient X-ray fluorescence was necessary.

SHG at surfaces and interfaces can be influenced by electric fields [81]. Aktsipetrov et al. performed electric field induced second harmonic generation (EFISHG) on a Si(001)-SiO_2 interface. A mode locked Ti:sapphire laser emitting pulses 120 fs in duration with an average power of up to 300 mW and a repetition rate of 76 MHz was used to probe the interface. The SHG signal was measured as a function of an applied DC field, these measurements show sensitivity to doping impurities which themselves influence the charge characteristics at the interface.

Kityk has used photoinduced SHG to monitor the superconductivity of Hg-1201 films and the properties of nanocrystallites [82, 83]. Photoinduced SHG requires two laser

[6]Reprinted from Surface Science, Vol **526**, No 1-2, Ishida, H., Mizoguchi, R., Onda, K., Hirose, C., Kano, S. S. and Wada, A.,*Second harmonics observation of Cu(111) surface in situ measurements during molecular adsorption*, pp 201-207, Copyright (2003), with permission from Elsevier

sources. The first laser pulse induces a small optical nonlinearity in the sample and a second laser probes that nonlinearity. In the case of superconducting films it is important that the laser pulses result in only a limited amount of heating. Kityk employed a nitrogen laser operating at 337 nm with a pulse duration of 650 ps with an intensity limited to 9 mJ/cm ^{2}. An Nd:YAG picosecond laser was employed as the probe laser. Kityk showed that the SHG signal began showing signs of the superconductivity transition before the measured resistance began to drop (see Figure 12). He also showed that under high nitrogen laser intensities, one can control the kinetics of the carriers and this controls the superconduction phase transition. In a similar experiment, SiC nanocrystallites were embedded in a polymatrix and subject to the experiment described above. The photoinduced SHG signal correlated well with interface, structural and bandgap calculations of the nanocrystallites.

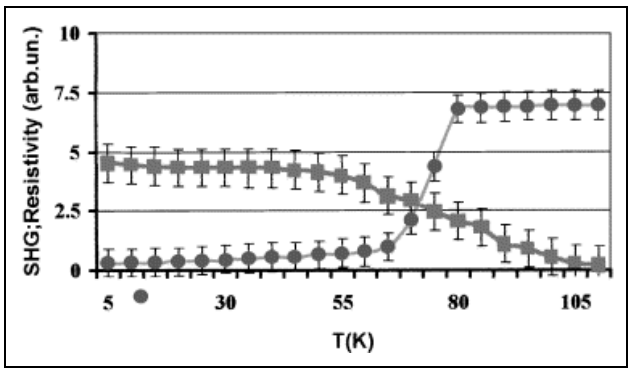


Figure 12. Photoinduced SHG signal as a function of superconductor temperature. [7]

Photoinduced SHG can also be used for spectroscopy on surface electronic and phonon structure [84]. GaAS (100 and 110) surfaces were illuminated by a modelocked Ti:sapphire laser emitting 35 fs pulses at a repetition rate of 80 MHz and a centre wavelength of 850 nm. This laser performed both the photoinduction and the probe duties. The probe was split off from the pump and passed through a variable delay line. The SHG intensity was measured as a function of delay time and this delay time was used to explore the spectroscopy of phonons that are coherently excited by the pump pulse. Tom et al. performed similar measurements to elucidate the surface electronic spectra of GaAs (110) surfaces. These measurements were slightly more complex since it was necessary to capture both the reflected pump pulse and the SHG signal which were then cross correlated. This cross correlation measurement can be used to calculate the dephasing rate ($1/T_2$) and bandgap of the material.

SHG at a surface in combination with polarisation information (called second harmonic ellipsometry) has also been demonstrated [85]. The laser source for the experiment was, for fixed frequency measurements a frequency doubled Nd:YAG laser, while tunable results were obtained using a dye laser tuned from 552 - 586 nm. Pulse energies between 10 and 15 mJ per pulse at a 10 Hz repetition rate were used. The polarisation of the incoming light was controlled with a half-wave plate and a Glan Taylor polariser. The polarisation of the

[7] Reprinted from Optics and Lasers in Engineering, Vol **40**, No 3, I. V. Kityk,*Photoinduced nonlinear optics methods as new methods of monitoring the superconducting films* , pp 179-187, Copyright (2003), with permission from Elsevier

reflected second harmonic was filtered such that only p-polarised light was detected. Initial tests were performed on an air-water interface which showed no polarisation dependence (all the molecules involved are achiral). The water was then replaced with binapthol, which is chiral. This caused the second harmonic to show a strong dependence on the polarisation. The s polarised fundamental generated circularly polarised second harmonic while the p polarised light generated plane polarised second harmonic. Thus, by combining traditional ellipsometry and SHG one can obtain information about the chirality and concentration of components present at interfaces.

Elliptically polarised SHG can be combined with photoinduction methods in the study of thin films [86]. Plucinski et al. measured the SHG response of SiN_xO_y films as a function of polarisation, photoinduction properties and varying nitrogen/oxygen content. Photoinduction was performed by a mode locked nitrogen laser emitting 1 - 5 ps pulses with an average power flux of 50 MW, the degree of elliptical polarisation of the nitrogen laser was controlled using a potassium titanyl phosphate crystal. The sample was then probed by a mode locked Nd:YAG laser emitting pulses of 10 - 50 ps with a pulse energy less than 2 μJ. The polarisation of the Nd:YAG laser was held parallel to the plane of incidence as was the detection apparatus for the SHG signal. Care was taken to measure and thus eliminate the hyper-Raman signal which can interfere with such experiments. Sample roughness was accounted for by measuring over several points on the surface and averaging the results. Plucinski et al. found that the nonlinear response of the films depended strongly on the N/O ratio and that electron-phonon interactions play an important role in the photoinduction process. For high N/O ratios SHG is more sensitive to the surface electronic states than extended X-ray absorption fine structure, however at low N/O ratios SHG is not so sensitive because the low nitrogen content allows the structure to relax into a centro-symmetric configuration.

What is clear from the review of recent experimental work is that SHG is a very good candidate for monitoring interactions at surfaces and interfaces. Q-switched or nanosecond pulsed lasers may be utilised in such experiments, however, it is preferable to utilise mode-locked lasers which produce much shorter pulses and deliver lower total energy per pulse. Such lasers produce pulses with a very high intensity while not providing enough energy to significantly heat and subsequently damage the sample. These lasers are now readily available at cheap prices thanks to the development of all solid state diode pumped laser systems. However, readily tunable sources are still the subject of intense research effort and have not yet attained the reliability and availability of fixed frequency laser sources. We expect to see surface SHG mature as an experimental technique and become a routine analytical tool for scientists studying surfaces and interfaces.

7. Sum and Difference Frequency Generation at a Surface

It is also possible to use sum and difference frequency generation to probe an interface between two centro-symmetric materials. In this case one generally employs one laser at a fixed frequency and a second laser that is tunable. Then as the second laser is tuned, either the fundamental or the DFG/SFG signal between the two lasers is tuned over a spectral region of interest. In general this is a more difficult experiment to perform since the

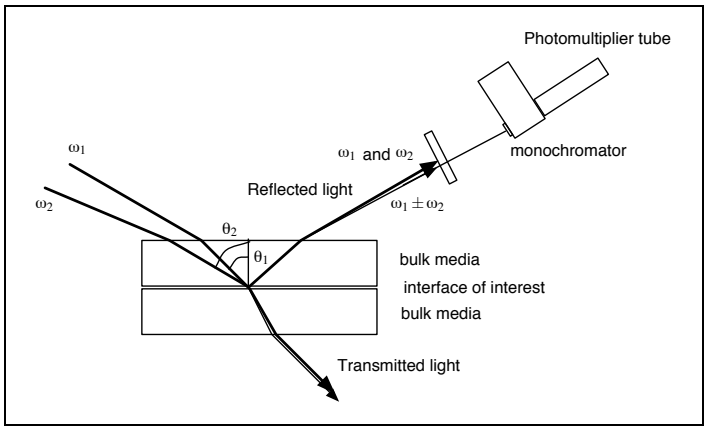


Figure 13. Generalised sum or difference frequency experimental set up. The angles θ_1 and θ_2 can be used to obtain total internal reflection of the sum or difference frequency or they may be used to control such experimental variables as the partial overlap of the beams.

DFG/SFG signal is proportional to $\mathbf{E}_1(\mathbf{r}_1, t)\mathbf{E}_2(\mathbf{r}_2, t)$, and is very sensitive to spatial and temporal overlap conditions. A generalised experimental set up is shown in Figure 13.

Because of the added difficulty in performing DFG/SFG spectroscopy it is usually employed to examine parts of the spectrum where SHG spectroscopy cannot be easily performed. Tadjeddine et al. performed such experiments on various interfaces using a free electron laser (FEL) [87]. In one example, they used DFG to study a gold interface in contact with a solution of $NaClO_4$ and pyridine [88]. These experiments elucidated the adsorption geometry of the pyridine molecule at the gold interface. In other experiments, both sum and difference frequency generation were employed to monitor the adsorption of cyanide at the surface of platinum, silver and gold electrodes [89]. The same apparatus was used to study the influence of hydrogen on the surface spectra of platinum, which is of great industrial interest [87]. In further SFG and DFG studies the self assembly of p-nitroaniline-dodecane thiol on a gold electrode was monitored. The experimental apparatus for these experiments consisted of a FEL emitting 1 ps pulses in 10 μs bunches with a bunch repetition rate of 25 Hz. The FEL is widely tunable so the second laser is a frequency doubled mode locked Nd:YAG laser emitting 70 ps pulses.

In many respects SFG and DFG spectroscopy is of more utility than SHG alone because one has the ability to choose which of the three wavelengths is to be tuned across a spectral feature of interest. The draw back is that one must have access to a widely tunable source of high intensity radiation. The most versatile of these sources is the FEL which can tune across a very large range, however, these are only available at a few places in the world. More practical laser sources for this are pulsed, line narrowed, Ti:sapphire lasers [90] or various optical parametric oscillator sources [91–93]. Typically these sources are not widely available commercially and are often manufactured in the laboratory.

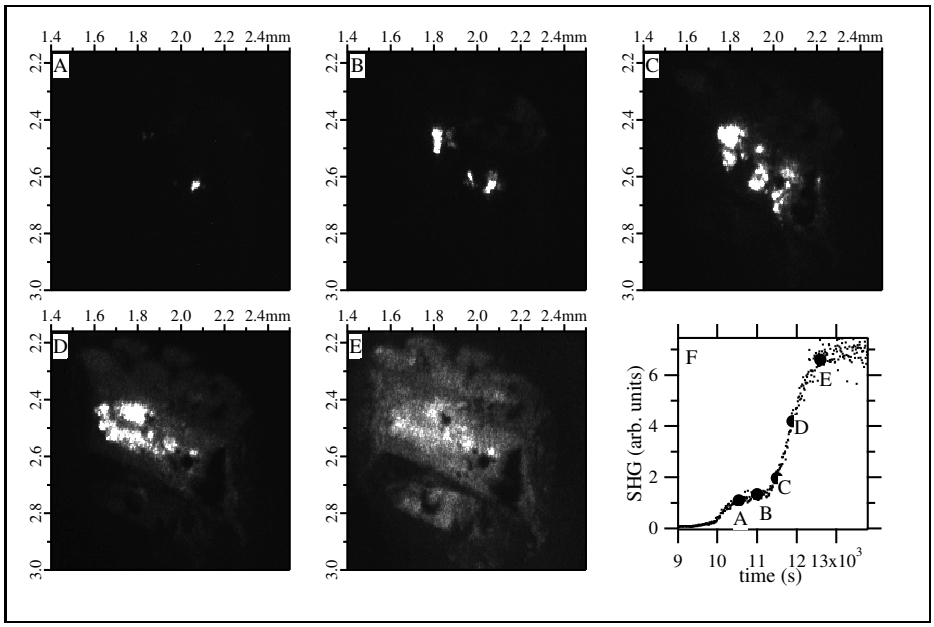


Figure 14. Monitoring a temperature induced phase transition in HMX using the bulk SHG signal.[8]

8. Second Harmonic Generation in Bulk

Second harmonic generation may also be used to monitor the phase and growth of a non centro-symmetric crystalline material [94]. LeCaptain et al. observed a supersaturated solution of dihydrogen phosphate (KDP), a commonly used nonlinear optical crystal, by illuminating a sample with a Q-switched Nd:YAG laser. In doing so they observed the spontaneous nucleation of the KDP crystals, the time taken to form the first seed crystals (the induction time) and the crystallisation dynamics as the solution desaturated. This is possible since the fluid produces no SHG response until crystals have begun to seed. Similar but more qualitative studies were performed on α-lactose and an amino acid (lysine m-HCL). In doing so, LeCaptain et al. demonstrated that the technique may have wide applicability across many industrial applications.

Henson et al. have demonstrated that SHG can be used to observe phase changes in bulk media [95]. The example that they chose was the β-δ phase transition of HMX (octahydro-1,3,5,7-tetranitro-1,3,5,7-tetraocine) which is an explosive. HMX is stable in the β form but is unstable in the δ phase. Heating β-HMX induces a phase change to the δ form which then rapidly decomposes, generating more heat. Henson et al. illuminated HMX samples with a Q-switched Nd:YAG laser and either imaged the SHG intensity with a CCD camera or recorded the integrated intensity with a photomultiplier tube (see Figure 14). The phase transition was induced in a controlled fashion by irradiating the samples with a CO_2 laser. Other explosives are thought to undergo a phase change before decomposition and this has since been observed by nonlinear optical techniques [96].

[8]Reprinted with permission from L. Smilowitz, B. F. Henson, B. W. Asay, and P. M. Dickson, Journal of Chemical Physics, **117**, pp 3789 (2002). Copyright 2002, American Institute of Physics

SHG is also used as a method to evaluate the potential of new nonlinear crystals [59]. Instead of growing bulk crystals for evaluation, a powder sample is produced and irradiated with a pulsed laser. Simultaneously, the laser irradiates a reference sample. The ratio between the two allows for the evaluation of the optical nonlinearity. In general, if a crystal has a geometry that allows for phase matching, those crystals that are close to the phase matching angle will dominate the response. Thus one can quickly evaluate a potential nonlinear crystal for nonlinearity, and also discover if the crystal can be phase matched. This method of evaluating nonlinear crystals is called the Kurtz powder method.

A modification of the Kurtz powder method allows for the evaluation of the larger nonlinear tensor elements whether or not they can be phase matched [59,97]. The experimental apparatus is nearly the same except that the laser light is coupled into the sample through a hemicylindrical prism. Reflection occurs at the interface between the sample and the prism and the resulting SHG signal that is generated in the evanescent wave will be phase matched in the prism at a particular angle of observation. Thus by measuring the SHG signal as a function of angle, one can evaluate all nonlinear tensor elements.

The dynamic properties of probe molecules associated with polymers has been measured using SHG in combination with fluorescence properties [98]. These experiments involved several fluorescent probe molecules coupled with high molecular mass polymers. The samples were placed on a slide and processed such that the sample thickness was of the order of 100 μm. A strong electric field was used to hold the probe and polymer in a fixed orientation. The sample was illuminated with a Q-switched Nd:YAG laser with a repetition rate of 10 Hz while the electric field was turned off. The SHG signal was measured as a function of time, while the fluorescence molecules re-oriented to obtain a more centrosymmetric form. The rate of SHG decay was a function of the mobility of the chromophores present in the polymer.

The phase and switching characteristics of liquid crystals can be monitored using bulk second harmonic generation [99]. A liquid crystal cell was irradiated with an OPO pumped by a Q-switched Nd:YAG laser emitting pulses 4 ns in duration. The transmitted SHG signal was measured as a function of applied voltage, polarisation and cell temperature. In doing so Fokin et al. were able to measure the critical exponents associated with the liquid crystal phase transition and determine that the phase transition is of second order. The non zero SHG signal with no applied voltage was thought to be due to an interaction between the liquid crystal and the cell wall which results in the liquid retaining some of the liquid crystal features.

We have also performed experiments using bulk SHG to measure the nonlinear optical response of binary mixtures [100, 101]. The purpose of these initial experiments was to evaluate the suitability of such techniques for in-line monitoring during drug production. As well as being of scientific and/or pharmaceutical interest, the physical characterisation of pharmaceutical materials and formulations is required by regulatory bodies. An additional complication in the development of such techniques is that any monitoring techniques used in the pharmaceutical industry must be formally validated according to international guidelines [102, 103]. The work presented here does not meet the requirements for formal validation.

The initial tests were performed on powdered glass and quartz mixtures. Quartz and glass are both primarily composed of silicon dioxide, SiO_2. Quartz consists almost exclu-

sively of SiO_2 and exists in several polymorphic forms, but the most common crystalline form, stable at room temperature and atmospheric pressure, is α-quartz [104]. α-Quartz is a non-centro-symmetric trigonal crystalline material, belonging to space group, $P3_12_1$ and is known to have a nonlinear response. Glass is amorphous and principally derived from SiO_2. Window or plate glass contains about 10% w/w sodium and calcium. Pyrex is a trademarked form of glass, differing from typical window glass, inter alia, by containing about 4% boron, which is present in a higher proportion than sodium, and, the absence of calcium [105]. Hereafter Pyrex glass will be referred to as Pyrex. Since all forms of glass are amorphous they are not capable of bulk SHG in the electric dipole approximation.

Quartz has been subjected to second harmonic measurements, and its nonlinear optical properties are well characterised [106, 107]. Quartz and glass have high laser damage thresholds. Since the Kurtz powder technique has not been used to quantify powder mixtures, quartz and Pyrex were used as proof-of-principle before pharmaceutical samples were investigated and quantified.

SHG measurements were performed using a Continuum PowerLite 700 *Q*-switched Nd:YAG laser. The laser emitted pulses of 10 ns (FWHM) duration with a pulse energy as high as 40 mJ and at a repetition rate of 10 Hz. The pulse energy was controlled using a half wave plate and polarising beam splitter combination. The laser pulse energy was monitored by detecting the scattered radiation from one of the beam steering mirrors using a silicon avalanche photodiode (2 ns rise time).

The wavelength dependence of light scattered by quartz and Pyrex between 392 and 692 nm were recorded (see Figure 15). The quartz spectrum shows a sharp peak at 532 nm, the second harmonic wavelength and a minimal baseline response at other wavelengths. Pyrex responds evenly over the spectral range, with no peak at 532 nm indicating no SHG. It was noted that the Pyrex response to a single laser pulse was larger that its average response to many laser pulses. This suggests that Pyrex exhibits some fluorescence, which is probably due to sodium and can be temporarily quenched by repeated irradiation.

Figure 16 shows the dependence of the second harmonic response of quartz and Pyrex on particle size. The response of quartz was not size dependent over the size ranges investigated. This may be because all particle sizes are much larger than the coherence length (L_c), where the particle size influence on the response becomes negligible according to the theoretical model for nonphase-matchable materials (Section 4.3. and Figure 3). For quantitative analysis of pharmaceutical polymorphism or crystallinity however, it is an advantage if the response is not dependent on particle size. As expected the SHG by amorphous Pyrex was negligible for all size ranges.

With a constant PMT bias voltage, the response is expected to increase linearly with increasing quartz concentration (Figure 17). At low quartz concentrations the response appears to be highly linear. However, for samples with quartz concentrations above 50% the standard deviation of the response increases. The reasons for this require further investigation, but it may be related to the amount of the sample contributing to the second harmonic response or multiple scattering.

The ability to distinguish enalapril maleate (EM) polymorphic forms I and II and polyvinylpyrrolidone (PVP) through their optical nonlinear response and then quantify binary mixtures of EM forms I and II followed by EM form II and PVP (amorphous) was

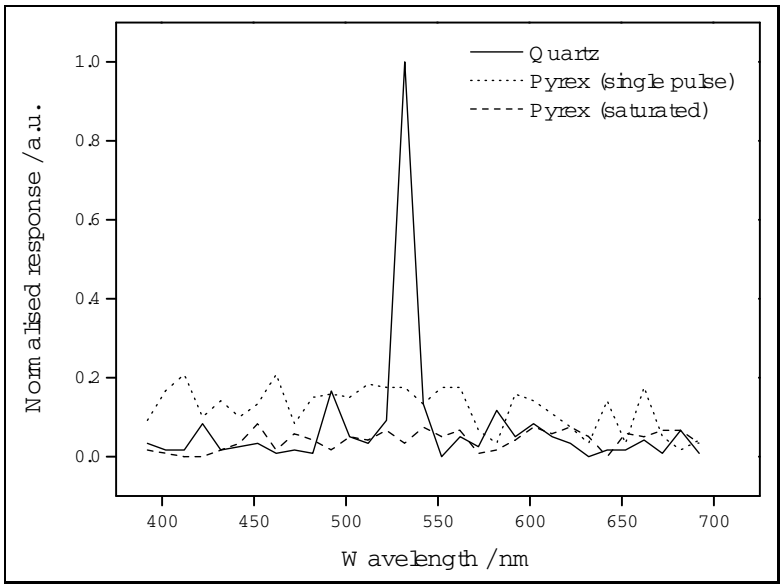


Figure 15. Response spectrum of quartz and Pyrex after a single pulse and after saturation. Each measurement is the average of ten scans and particle size was 125 - 180 μm.

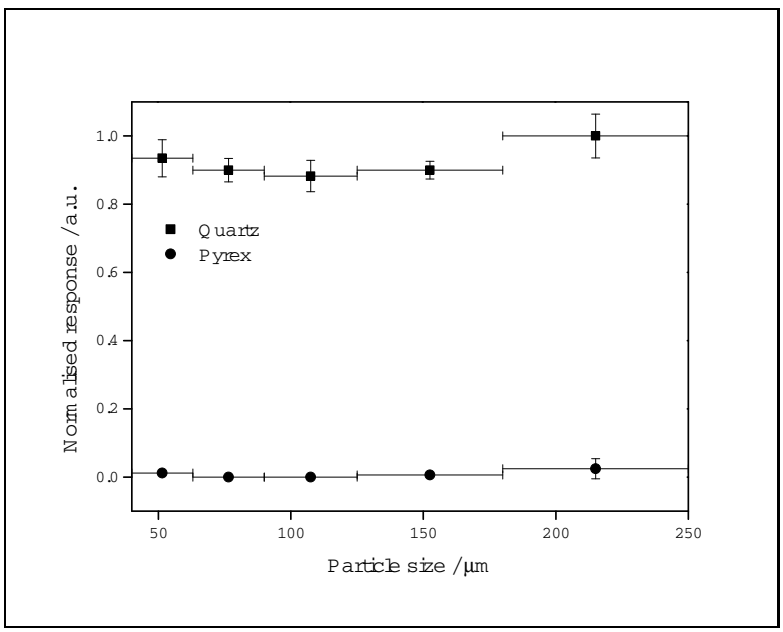


Figure 16. Second harmonic response of different particle size ranges of quartz and Pyrex (mean $\pm$ SD (y-axis), n = 3). x-axis error bars represent particle size ranges.

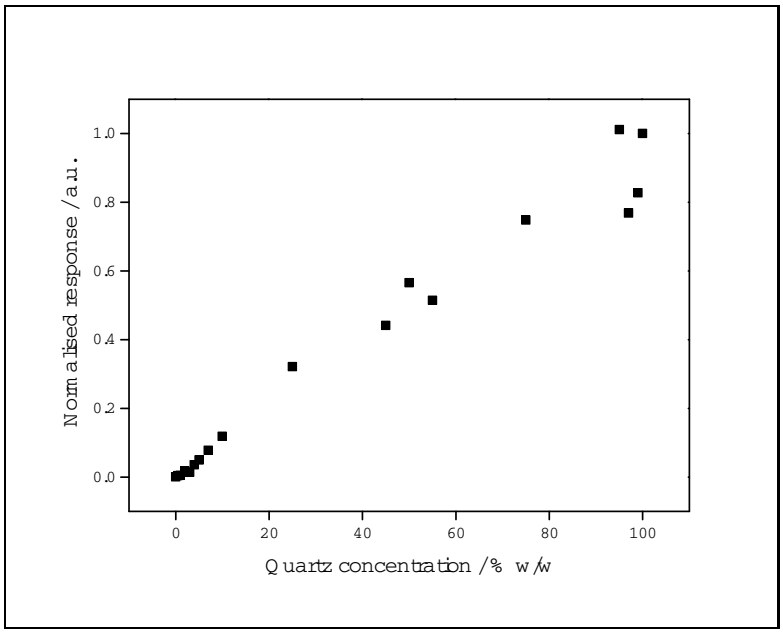


Figure 17. Second harmonic response of binary mixtures of quartz and Pyrex as a function of increasing quartz concentration (n = 1). Particle size was 90 - 125 μm and samples were irradiated for 30 s prior to measurement.

investigated. Mixtures of lactose polymorphs were also investigated.

EM is an orally active angiotensin converting enzyme inhibitor used to treat hypertension and heart failure [108]. It is known to exist in at least two polymorphic forms that exhibit similar stability [109, 110]. PVP is an amorphous pharmaceutical excipient commonly used as a dispersing and suspending agent, and as a tablet binding, granulating and coating agent [111]. It is also frequently used in the production of solid dispersions to stabilise poorly water-soluble drugs in the amorphous form [112]. Lactose is known to exist in four polymorphic forms and a hydrated form of which two were investigated: α-lactose monohydrate (LMH) and anhydrous β-lactose (βL). Lactose is a commonly used excipient in tablets and capsules.

The particle size dependence of the second harmonic response of EM forms I and II and PVP is shown in Figure 18. EM form I gave the most intense response followed by form II and then PVP. The effect of particle size on the responses of EM forms I and II and PVP was analysed using a one-way analysis of variance (ANOVA) combined with the Tukey method (95% confidence level) of mean comparisons.

The second harmonic response of EM form I was dependent upon particle size (P = 0.002, where P is the chance of the result occurring randomly). Particles in the size range 90 - 125 μm, with the largest mean second harmonic response, had a statistically significantly larger response than particles in the other size ranges. The response of EM form II was weakly dependent on particle size (P = 0.040). Particles between 90 and 125 μm had a larger response than particles in the remaining size ranges. However, these measurements display a larger uncertainty. The response of PVP was much smaller than that of EM forms

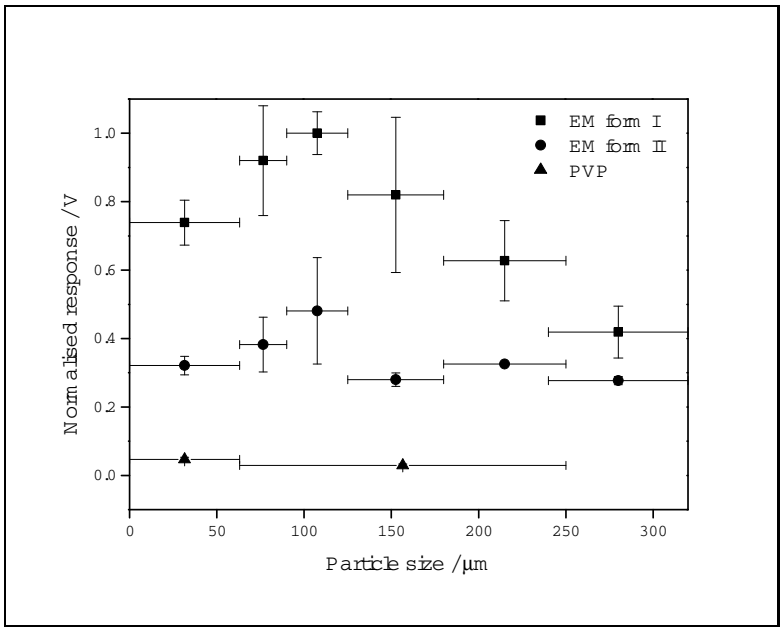


Figure 18. Second harmonic response of different particles size ranges of EM forms I and II and PVP. Data points and y-axis error bars are the mean and standard deviation of three measurements. X-axis error bars represent particle size ranges.

I and II as expected.

SHG from binary mixtures of EM forms I and II steadily increased with increasing form I concentration (Figure 19). Regression analysis showed a significant relationship between form I concentration and response ($P < 0.0001$). The non-zero intercept means that form II also had significant SHG in the model (P = 0.012). At low form I concentrations the standard deviations of the responses were small; however as the proportion of form I increased, the standard deviation of measurements increased. This is consistent with the quartz and Pyrex mixture measurements.

Calculation of the limit of detection (LOD) based on the standard deviation of the slope derived from a linear fit (response = 0.00776(% Form 1) + 0.136, R^2 = 0.974) resulted in an LOD of 23.5%. However, if the LOD is calculated based on the standard deviation of the three 0% form I values, the LOD becomes 5.1%. This decrease in LOD is probably due to the large standard deviations of the responses at high form I concentrations. Although the LODs are too high for most practical uses, these results suggest that with experimental improvement the technique may provide acceptable LODs for in-line process control.

EM form II and PVP mixtures are an example of mixtures of crystalline and amorphous forms. From Figure 20, one can see that the second harmonic response increased with increasing EM form I concentration ($P < 0.001$). In addition, there is no evidence of a non-zero intercept (P = 0.534).

The LODs are 23.0% and 0.13%, based on the standard deviation of the slope derived from a linear fit (response = 0.01064(% Form II) - 0.03165, R^2 = 0.987) and three 0% form II values respectively. The very great discrepancy between the two values suggests that use

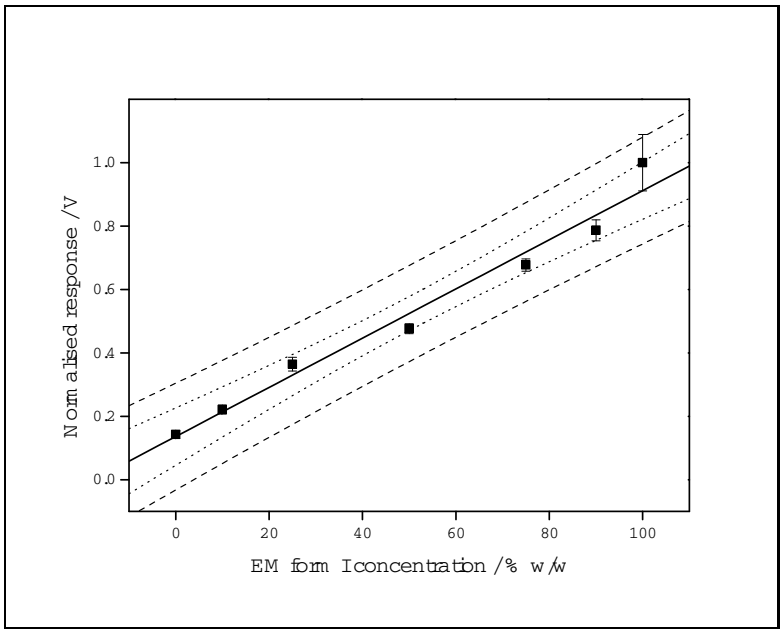


Figure 19. SHG from binary mixtures of EM forms I and II with 95% confidence bands (dots) and prediction bands (dashes) (mean $\pm$ SD, n = 3). The particle size range was $<$ 63 μm.

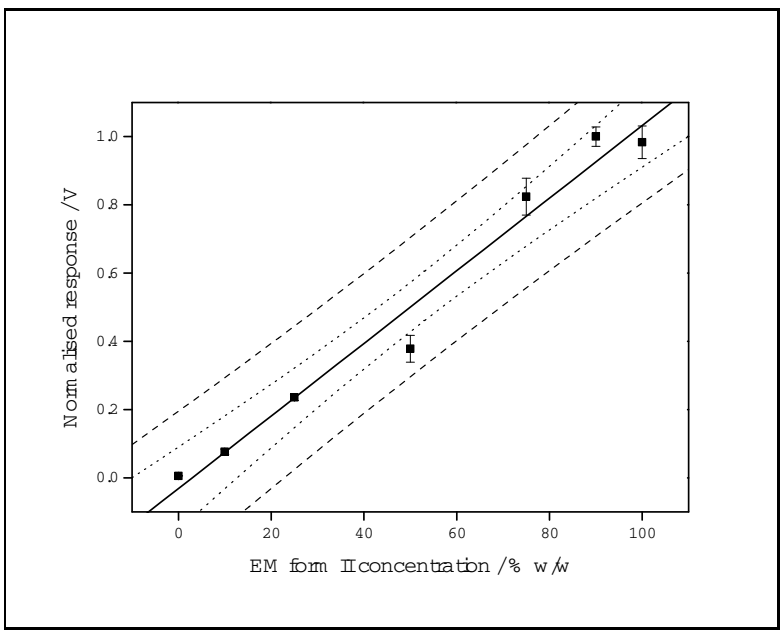


Figure 20. SHG from binary mixtures of EM form II and PVP with 95% confidence bands (dots) and prediction bands. The data points and error bars are the mean and standard deviation from three measurements. The particle size range was 0 - 63 μm.

of SHG is an excellent technique to detect small amounts of crystalline EM form II in an otherwise amorphous PVP medium. However, at higher levels of crystallinity the technique currently loses much of its sensitivity.

The LOD of 0.13% is substantially lower than levels obtained using most other techniques, such as near-IR spectroscopy and X-ray powder diffraction [17, 18, 113–115]. Although further investigations are required, the sensitivity in this example and the speed of the technique suggests great potential for the use of SHG to detect very low levels of crystalline material in an otherwise amorphous medium.

The second harmonic response of binary mixtures of βL and LMH is shown in Figure 21. There was a linear increase in SHG as the βL concentration increased from 0 to 90%. There was disproportionate increase in SHG at 100% βL which may be due to βL being a phase-matchable material. If L_c is much larger than r the introduction of the nonphase-matchable αLMH may have disturbed the coherent phase-matched response of βL leading to a sharp drop in SHG. Up to 90% βL it is likely that no phase-matching occurred. A linear fit in the 0 - 90% βL concentration range revealed a significant relationship between βL concentration and response ($P < 0.001$). There was a non-zero intercept ($P < 0.001$), which confirms that pure αLMH produces a significant second harmonic response.

Quantitative analysis of the data yielded LODs of 24.9% and 12.7% based on the standard deviation of the slope (0 to 90% βL) and 0% βL sample measurements respectively.

In future investigations, samples with between 90 and 100% βL will be measured to characterise the SHG response in the concentration range where phase-matching of βL appears to be occurring. If the sharp change can be well quantified, a lower LOD for βL than αLMH may be obtained in binary mixtures of the two forms.

In more recent work we have begun to perform measurements of the bulk SHG response as a function of both angle of incidence and angle of measurement (see Figure 22) [116]. These measurements along with control over the pump spot size are expected to further lower the limits of detection and quantification along with providing simultaneous particle size information.

Since the laser beam spot size was 1.5 mm at the sample face it is expected that the optical nonlinear response will be a smooth function of angle. Indeed this turns out to be so (see Figure 23), note that the shape of the function is independent of particle size while the absolute magnitude of the response is strongly dependent on particle size. However when the measurements take into account the polarisation of the SHG response we see that the angle of peak response is weakly dependent on particle size which was unexpected and is the subject of further investigation. It is expected that as the laser beam size is decreased a speckle pattern will be produced similar to that of found by Melnikov et al. [63] and that by finding a suitable correlation function one will be able to extract particle size data as well as mixture ratio information.

Further measurements on EM form II/PVP mixtures have been performed and some of the data are presented in Figure 24. The LOD (already excellent) has not been improved by this method, however the accuracy of quantification has been improved at higher percentages.

The biggest remaining problem with expanding the apparatus into a production line monitoring device is that it is only sensitive to binary ratios. This is because the SHG response is measured at only a single non-resonant wavelength which means that if the

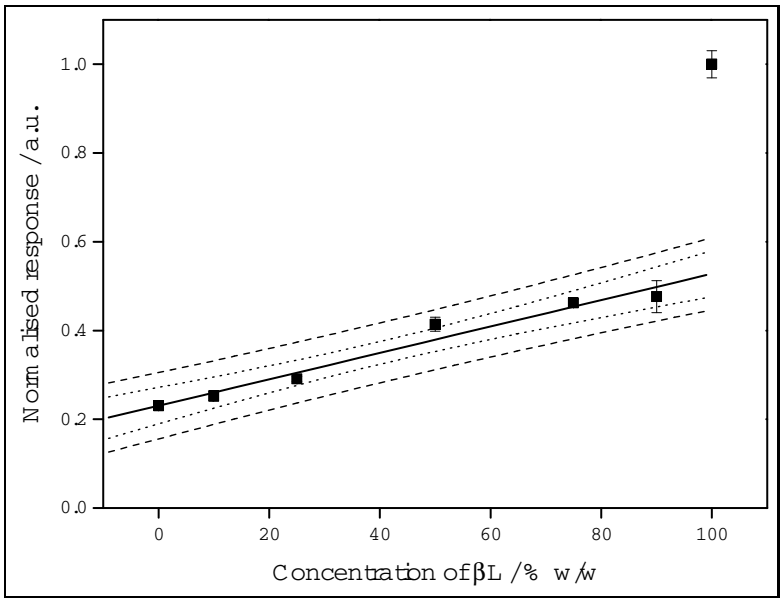


Figure 21. SHG from binary mixtures of βL in LMH with 95% confidence bands (dots) and prediction bands (dashes). The data points and error bars are the mean and standard deviation from three measurements. Particle size range was 63 to 125 μm. The linear fit was calculated using the second harmonic response from 0 to 90% βL.

mixture contains three components and all are nonlinear then the response is a non-unique function of concentration. To solve this problem we intend to use tunable SHG, where the fundamental or its second harmonic are tuned over resonances of the various compounds present. By obtaining a spectrum of SHG, along with angle data, it is expected that multi-component tablets could be characterised on the production line in a very short time.

SHG as method for characterising bulk materials has until recently been limited to the Kurtz method and the modified Kurtz powder method. In recent years there has been a growth in experimental effort devoted to using SHG as a method for aiding in the characterisation of bulk and often turbid media. SHG may be seen as complementary method to existing methods such as powder x-ray diffraction, differential scanning calorimetry and IR spectroscopy. These traditional techniques suffer from limitations in their practical application outside of the laboratory environment. X-ray diffraction and differential scanning calorimetry are sensitive to both phase changes and chemical constituents, while IR spectroscopy is only sensitive to chemical constituents and even then is more qualitative rather than quantitative. SHG on the other hand is sensitive to phase transitions and provides quantitative information on those phase transitions. By utilising a tunable laser source one can detect chemical constituents as well, however, it cannot be employed on every chemical system and one must have some *a priori* knowledge of the system in order to interpret the information received. Finally the processing speed of SHG experiments is very fast, limited only by the pulse repetition rate of the laser. Thus bulk SHG measurement techniques have many of the qualities desirable for industrial monitoring of chemical processing and we expect to see the technique utilised more widely in the coming years.

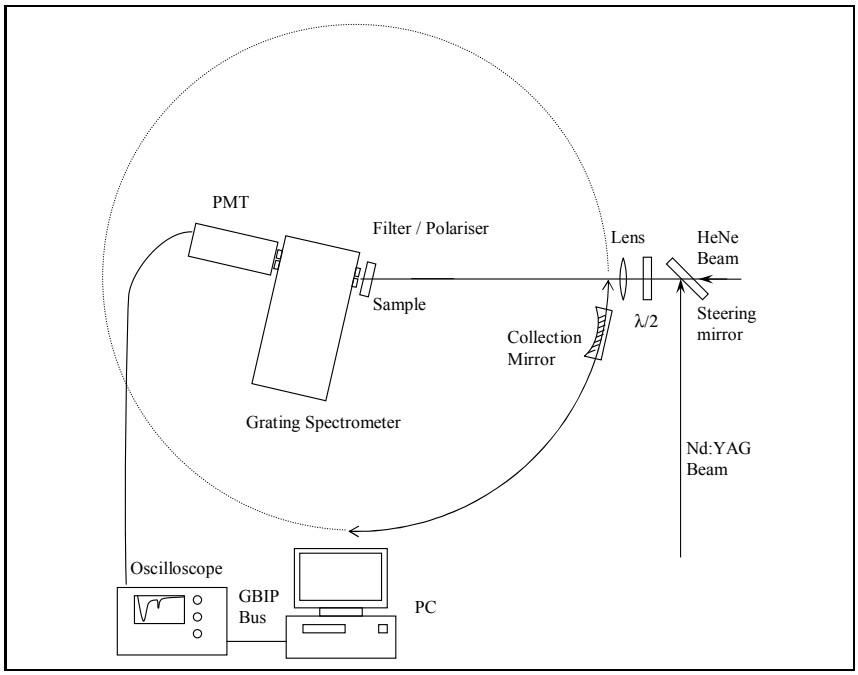


Figure 22. Apparatus for detecting optical scattering as a function of angle of incidence and angle of measurements. The spectrometer entrance slit and sample are both placed along the axis of rotation for the collecting mirror. The spectrometer and mirror rotate such that the mirror images light collected from the sample at the entrance slit of the spectrometer for any angle of measurement. The sample is mounted such that the laser beam maintains a fixed angle of incidence though out the measurement.

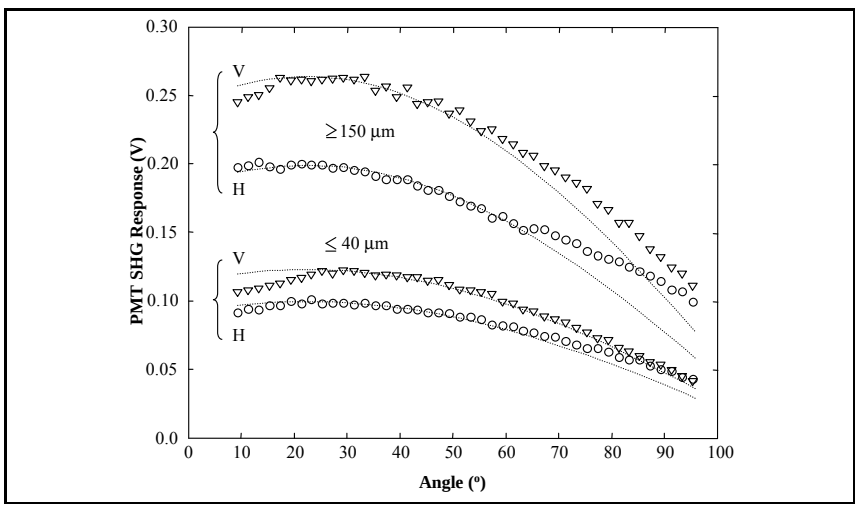


Figure 23. Particle size and polarisation dependence of SHG from lactose.The fitted curves are the expected angular dependence from purely geometric optics considerations.

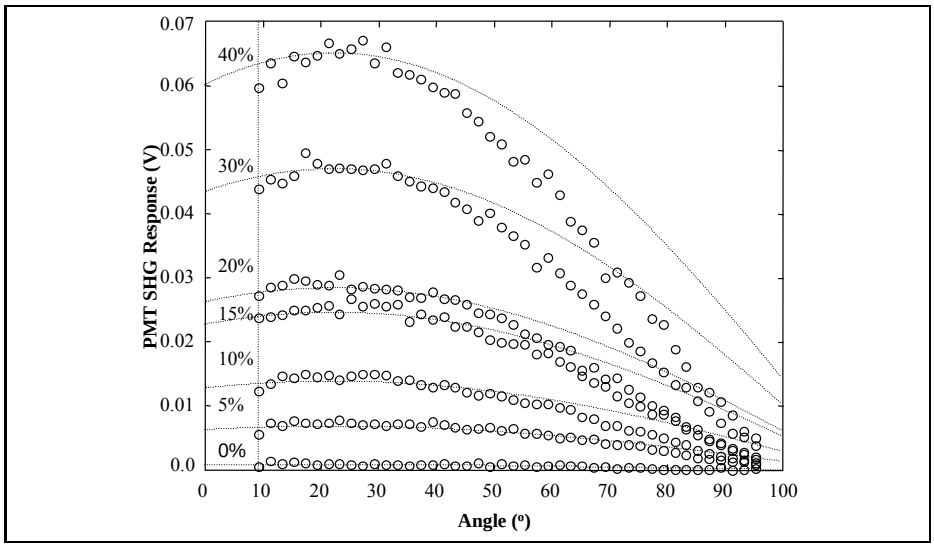


Figure 24. Scattering intensity for varying percentages of EM form II. The fitted curves are the expected angular dependence from purely geometric optics considerations.

9. Conclusion

A variety of spectroscopic and other techniques are used to characterise bulk and surface properties of materials and systems, including those in biology, medicine and pharmaceutical sciences. Nevertheless, many complex organic systems remain difficult to investigate and monitor, due to a lack of appropriate analytical techniques. Therefore, there is continued interest in improving existing spectroscopic methods and using novel spectroscopic approaches to characterise biological and pharmaceutical systems.

Second order nonlinear optical processes, including SHG, SFG, DFG and HRS, depend on symmetry and dipole moments of materials and molecules. Their theory is comparatively well developed, however they have been applied only to a very limited extent as diagnostic tools. SHG is ideally suited for characterising surfaces, and a few examples in this area have illustrated that monolayer adsorption of molecules to surfaces can be detected. It can also be used to investigate and quantify bulk properties including phase changes. SHG is ideally suited for detecting solid phases within liquid media, since the amorphous liquid phase does not generate a second harmonic response. HRS scattering is also well suited to investigating liquid systems, and it has been demonstrated that HRS can be useful to characterise colloidal systems. In addition, since nonlinear optical responses are instantaneous, they are ideal candidates for in-line quantitative control in industrial processes.

Our research and that of others others suggests that nonlinear optical responses are a viable tool for eliciting new information about systems as well as monitoring them. The development of second order optical nonlinear processes for diagnosis and monitoring of biological and pharmaceutical as well as other organic systems will be watched with interest.

References

[1] C. L. Tang and L. K. Cheng. Funadamentals of Optical Parametric Processes and Oscillators, volume 20 of Laser Science and Technology: an International Handbook. Harwood academic publishers, Amsterdam, 1995.

[2] S. Izumi, M. Sato, J. Suzuki, T. Taniuchi, and H. Ito. Periodically-poled $LiNbO_3$ optical parametric oscillator with 55% slope efficiency pumped by a Q-switched Nd:YAG laser. *Japanese Journal of Applied Physics Part 2-Letters* , **37**(11B):L1383–L1385, 1998.

[3] K. Akagawa, S. Wada, and H. Tashiro. High-speed optical parametric oscillator pumped with an electronically tuned Ti:sapphire laser. *Applied Physics Letters*, **70**(10):1213–1215, 1997.

[4] G. W. Baxter, Y. He, and B. J. Orr. A pulsed optical parametric oscillator, based on periodically poled lithium niobate (ppln), for high-resolution spectroscopy. *Applied Physics B-Lasers & Optics* , **67**(6):753–756, 1998.

[5] A. Miklos, C. H. Lim, W. W. Hsang, G. C. Liang, A. H. Kung, A. Schmohl, and P. Hess. Photoacoustic measuerement of methane concentrations with a compact pulsed optical parametric oscillator. *Applied Optics*, **41**(15):2985–2993, May 2002.

[6] N. B. Colthup, L. H. Daly, and S. E. Wiberley. *Introduction to Infrared and Raman Spectroscopy*. Academic Press, Inc., San Diego, 3rd edition, 1990.

[7] D. H. Williams and I. Fleming. *Spectroscopic Methods in Organic Chemistry* . McGraw-Hill Book Co., London, 5th edition, 1995.

[8] H. G. Brittain, editor. *Physical Characterization of Pharmaceutical Solids* . Marcel Dekker, Inc., New York, 1995.

[9] H. G. Brittain. Spectral methods for the characterization of polymorphs and solvates. *Journal of Pharmaceutical Sciences* , **86**(4):405–412, 1997.

[10] D. C. Fries, S. T. Rao, and M. Sundaralingam. Structural chemistry of carbohydrates. III. Crystal and molecular structure of 4-O-β-D-Galactopyranosyl-α-D-glucopyranose monohydrate (α-lactose monohydrate). *Acta Crystallographica* , **B27**:994–1005, 1971.

[11] D. E. Bugay. Characterization of the solid state: spectroscopic techniques. *Advanced Drug Delivery Reviews* , **48**:43–65, 2001.

[12] G. N. Kalinkova. Infrared spectoscopy in pharmacy. *Vibrational Spectroscopy* , **19**:307–320, 1999.

[13] J Bernstein. *Polymorphism in Molecular Crystals. International Union of Crystallography, Monographs on Crystallography* . Oxford University Press, Oxford, 2002.

[14] M. Blanco, J. Coello, H. Iturriaga, S. Maspoch, and C. de la Pezuela. Near-infrared spectroscopy in the pharmaceutical industry. *Analyst*, **123**:135R–150R, 1998.

[15] J. Aaltonen, J. Rantanen, S. Siiria, M. Karjalainen, A. Jorgensen, N. Laitinen, M. Savolainen, P. Seitavuopio, M. Louhi-Kultanen, and J. Yliruusi. Polymorph screening using near-infrared spectroscopy. *Analytical Chemistry*, **75**(19):5267–5273, 2003.

[16] M. Blanco and A. Villar. Polymorphic analsysis of a pharmaceutical preparation by NIR spectroscopy. *Analyst*, **125**:2311–2314, 2000.

[17] A. D. Patel, P. E. Luner, and M. S. Kemper. Quantitative analysis of polymorphs in binary and multi-component powder mixtures by near-infrared reflectance spectroscopy. *International Journal of Pharmaceutics* , **206**:63–74, 2000.

[18] A. D. Patel, P. E. Luner, and M. S. Kemper. Low-level determination of polymorph composition in physical mixtures by near-infrared reflectance spectroscopy. *Journal of Pharmaceutical Sciences* , **90**:360–370, 2001.

[19] M. C. Pasikatan, J. L. Steele, C. K. Spillman, and E. Haque. Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials. *Journal of Near Infrared Spectroscopy* , **9**(3):153–164, 2001.

[20] J. Ling, S. D. Weitman, M. A. Miller, R. V. Moore, and A. C. Bovik. *Applied Optics*, **41**(28):6006–6017, 2002.

[21] T. R. M. de Beer, G. J. Vergote, W. R. G. Baeyens, J. P. Remon, C. Vervaet, and F. Verpoort. Development and validation of a direct, non-destructive quantitative method for medroxyprogesterone acetate in a pharmaceutical suspension using ft-raman spectroscopy. *European Journal of Pharmaceutical Sciences* , **23**(4-5):355–362, 2004.

[22] L. S. Taylor and G. Zografi. The quantitiave analysis of crystallinity using FT-Raman spectroscopy. *Pharmaceutical Research* , **15**:755–761, 1998.

[23] A. M. Tudor, S. J. Church, P. J. Hendra, M. C. Davies, and C. D. Melia. The qualitative and quantitative analysis of chlorpropamide polymorphic mixtures near-infrared Fourier transform Raman spectroscopy. *Pharmaceutical Research* , **10**(12):1772–1776, 1993.

[24] S. Söderholm, Y. H. Roos, N. Meinander, and M. Hotokka. Raman spectra of fructose and glucose in the amorphous crystalline states. *Journal of Raman Spectroscopy* , **30**:1009–1018, 1999.

[25] P. A. Anquetil, C. J. H Brenan, C. Marcolli, and I. W. Hunter. Laser Raman spectroscopic analysis of polymorphic forms in microliter fluid volumes. *Journal of Pharmaceutical Sciences* , **92**(1):149–159, 2003.

[26] D. Pratiwi, J. P. Fawcett, K. C. Gordon, and T. Rades. Quantitative analysis of polymorphic mixtures of ranitidine hydrochloride by Raman spectroscopy and principal components analysis. *European Journal of Pharmaceutics and Biopharmaceutics* , **54**:337–341, 2002.

[27] L. S. Taylor and F. W. Langkilde. Evaluation of solid-state forms present in tablets by Raman spectroscopy. *Journal of Pharmaceutical Sciences* , **89**(10):1342-1353, 2000.

[28] J. Breitenbach, W. Schrof, and J. Neumann. Confocal Raman spectroscopy: analytical approach to solid dispersions and mapping of drugs. *Pharmaceutical Research* , **16**(7):1109–1113, 1999.

[29] T. L. Threlfall. Analysis of organic polymorphs - a review. *Analyst*, **120**(10):2435–2460, 1995.

[30] U. Holzgrabe, B. W. K. Diehl, and I. Wawer. NMR spectroscopy in pharmacy. *Journal of Pharmaceutical and Biomedical Analysis* , **17**:557–616, 1998.

[31] P. C. Griffiths, A. Y. F. Cheung, J. A. Davies, A. Paul, C. N. Tipples, and A. L. Winnington. Probing interactions within complex colloidal systems using PGSE-NMR. *Magnetic Resonance in Chemistry* , **40**:S40–S50, 2002.

[32] G. W. Chantry. *Submillimetre Spectroscopy: A Guide to the Theoretical and Experimental Physics of the Far Infrared* . Academic Press Inc. Ltd., London, 1971.

[33] K. D. Möller and W. G. Rothschild. *Far-Infrared Spectroscopy* . Wiley Series in Pure and Applied Optics. John Wiley & Sons, Inc., New York, 1971.

[34] P. M. A. Sherwood. *Vibrational Spectroscopy of Solids* , volume 1 of *Cambridge Monographs in Physical Chemistry 1* . Cambridge University Press, Cambridge, 1972.

[35] C. J. Strachan, T. Rades, D. A. Newnham, K. C. Gordon, M. Pepper, and P. F. Taday. Using terahertz pulsed spectroscopy to study crystallinity of pharmaceutical materials. *Chemical Physics Letters* , **390**(1-3):20–24, 2004.

[36] C. J. Strachan, P. F. Taday, D.A. Newnham, K.C. Gordon, J.A. Zeitler, M. Pepper, and T. Rades. Using terahertz pulsed spectroscopy to quantify pharmaceutical polymorphism and crystallinity. *Journal of Pharmaceutical Sciences* , **94**(4):837–846, 2005.

[37] P. F. Taday. Applications of terahertz spectroscopy to pharmaceutical sciences. *Philosophical Transactions of the Royal Society of London A* , **362**:351–364, 2004.

[38] A. J. Fitzgerald, B. E. Cole, and P. F. Taday. Non-destructive analysis of tablet coating thicknesses using terahertz pulsed imaging. *Journal of Pharmaceutical Sciences* , **94**(1):177–183, 2004.

[39] R. M. Woodward, V. P. Wallace, R. J. Pye, B. E. Cole, D. D. Arnone, E. H. Linfield, and M. Pepper. Terahertz pulse imaging of ex vivo basal cell carcinoma. *Journal of Investigative Dermatology*, **120**(1):72–78, 2003.

[40] Y. R. Shen. *The Principles of Nonlinear Optics*. John Wiley and Sons, New York, 1984.

[41] R. W. Boyd. *Nonlinear Optics*. Academic Press, San Diego, 1st edition, 1992.

[42] H. Orihara and Y. Ishibashi. A phenomenological theory of nonlinear dielectric response. II. Miller's rule and nonlinear response in nonferroelectrics. *Journal of the Physical Society of Japan*, **66**(1):242–246, January 1997.

[43] P. N. Butcher and D. Cotter. *The Elements of Nonlinear Optics*. Cambridge University Press, New York, 1990.

[44] B. Koopmans. *Interface and bulk contributions in optical second harmonic generation*. PhD thesis, University of Groningen, the Netherlands, 1993.

[45] A. M. Janner. *Second-Harmonic Generation, a Selective Probe for Excitons* . PhD thesis, University of Groningen, the Netherlands, 1998.

[46] N. Bloembergen and Y. R. Shen. Quantum-theoretical comparison of nonlinear susceptibilities in parametric media, lasers and Raman lasers. *Physical Review*, **133**(1A):A37 – A49, Jan 1964.

[47] Karl Blum. *Density Matrix Theory and Applications* . Physics of Atoms and Molecules. Plenum Press, New York, 1981.

[48] Marc D. Levenson and Satoru S. Kano. *Introduction to Nonlinar Laser Spectroscopy* . Academic Press, 1250 Sixth Avenue, San Diego, CA 92101, revised edition, 1988.

[49] B. R. Mollow. Power spectrum of light scattered by two-level systems. *Physical Review*, **188**(5):1969 – 1975, Dec 1969.

[50] R. Bavli, D. F. Heller, and Y. B. Band. Nonlinear-optical properties of two-level systems with permanent dipole moments. *Physical Review A*, **41**(7):3960 – 3968, Apr 1990.

[51] R. Bavli and Y. B. Band. Sum and difference frequency generation in a two-level system with permanent dipole moments. *Physical Review A*, **43**(9):5044 – 5048, May 1991.

[52] Robert C. Hilborn. Einstein coefficients, cross sections, f values, dipole moments, and all that. *American Journal Physics*, **50**(11):982 – 986, Nov 1982.

[53] R. E. Slusher, L. W. Hollberg, Bernard Yurke, J. C. Mertz, and J. F. Valley. Observation of squeezed states generated by four-wave mixing in an optical cavity. *Physical Review Letters*, **55**(22):2409 – 2412, Nov 1985.

[54] Mari W. Maeda, Prem Kumar, and Jeffrey H. Shapiro. Squeezing experiments in sodium vapour. *Journal of the Optical Society of America B*, **4**(10):1501 – 1513, Oct 1987.

[55] Susan M. Dounce, Minchul Yang, and Hai-Lung Dai. Surface-state relaxation dynamics on Ag{110} probed by temperature-dependent resonantly enhanced second-harmonic generation. *Physical Review B*, **67**(20):205410 – 205410–7, May 2003.

[56] S. Sauvage, P. Boucaud, T. Brunhes, F. Glotin, R. Prazeres, J.-M. Ortega, and J.-M. Gérard. Second-harmonic generation resonant with s-p transition in InAs/GaAs self-assembled quantum dots. *Physical Review B*, **63**(11):113312 – 113312–4, Mar 2001.

[57] D. Indjin, A. Mircetic, Z. Ikonic, V. Milanovic, and G. Todorovic. Resonantly enhanced bound-continuum intersubband second harmonic generation in optimized asymmetric semiconductor quantum wells. *Physica E*, **4**(2):119 – 127, Apr 1999.

[58] J. D. Kraus. *Electromagentics*. Mcgraw Hill International Editions, New York, 4th edition, 1992.

[59] R. L. Sutherland. *Handbook of Nonlinear Optics*. Marcel Dekker, Inc, New York, 1996.

[60] V. G. Dmitriev, G. G. Gurzadyan, and D. N. Nikogosyan. *Handbook of Nonlinear Optical Crystals*. Springer-Verlag, New York, 3rd edition, 1999.

[61] S. J. Brosnan and R. L. Byer. Optical parametric oscillator threshold and linewidth studies. *IEEE Journal of Quantum Electronics*, **QE-15**(6):415–430, June 1979.

[62] K. Clays and A. Persoons. Hyper-Rayliegh scattering in solution. *Physical Review Letters*, **66**:2980–2983, 1991.

[63] A. V. Melnikov, K. Nikulin, and O. A. Aktsipetrov. Hyper-Rayleigh scattering by inhomogeneous thin films of $Pb_x(Zr_{0.53}Ti_{0.47})O_3$: Disorder effects. *Physical Review B-Condensed Matter*, **67**:134104, 2003.

[64] E. Hendrickx, A. Vinckier, K. Clays, and A. Persoons. Evidence of octopolar symmetry in bacteriorhodopsin trimers by hyper-Rayliegh scattering from purple membrane suspensions. *Journal of Physical Chemistry*, **100**:19672–19680, 1996.

[65] R. L. Sutherland. *Handbook of Nonlinear Optics*, pages 207–233. Marcel Dekker, Inc, New York, 1996.

[66] P. S. Ray. The effects of π-conjugation on first hyperpolarisabilities of charged NLO chromophores. *Chemical Physics Letters*, **394**:354–360, 2004.

[67] B. R. Cho, S. B. Park, S. J. Lee, K. H. Son, S. H. Lee, M. J. Lee, J. Yoo, Y. K. Lee, G. J. Lee, M. Kang, T. I. Cho, and S. J. Jeon. 1,3,5-tricyano-2,4,6-tris(vinyl)benzene derivatives with large second-order nonlinear optical properties. *Journal of the American Chemical Society*, **123**:6421–6422, 2001.

[68] B. K. Spraul, S. Suresh, T. Sassa, M. À. Herranz, L. Echegoyen, T. Wada, D. Perahia, and D. W. Smith Jr. Thermally stable triaryl amino chromophores with high molecular hyperpolarisabilities. *Tetrahedron Letters*, **45**:3253–3256, 2004.

[69] G. J. T. Heesink, A. G. T. Ruiter, N. F. van Hulst, and B. Bölger. Determination of hyperpolarizabilty tensor components by depolarized hyper Rayleigh scattering. *Physical Review Letters*, **71**(7):999–1002, 1993.

[70] Y. Zhang, M. Ma, X. Wang, D. Fu, H. Zhang, N. Gu, J. Liu, Z. Lu, L. Xu, and K. Chen. Hyper-Rayleigh scattering of nanoscale CdS colloid and its formation process. *Inorganic Chemistry Communications*, **6**:427–430, 2003.

[71] S. Ghosh, A. Krishnan, K. P. Das, and S. Ramakrishnan. Determination of critical micelle concentration by hyper-Rayleigh scattering. *Journal of the American Chemical Society*, **125**:1602–1606, January 2003.

[72] D. S. D. Black. *Properties of Detergents(Amphiphiles)* . http://psyche.uthct.edu/shaun/SBlack/detergnt.html, 1998.

[73] C. J. Lee, T. C. A. Molteno, and P. J. Manson. Metropolis simulations of the manipulation of dna strands in solution. *Biophysical Journal*, In preparation, 2005.

[74] J. Rinuy. *Nonlinear optics of proteins at liquid interfaces and of DNA oligonucleotides in liquid phase*. PhD thesis, École de Polytechnique Fédérale de Lausanne, 2001.

[75] J. Örtegren, K. D. Wantke, H. Motschmann, and H. Möwald. A study of kinetic molecular exchange processes in the medium frequency range by surface SHG on an oscillating bubble. *Journal of Colloids and Interface Science*, **279**:266–276, 2004.

[76] P. J. Campagnola, M. Wei, A. Lewis, and L. M. Loew. High-resolution nonlinear optical imaging of live cells by second harmonic generation. *Biophysical Journal*, **77**:3341–3349, 1999.

[77] J. Rinuy, P. F. Brevet, and H. H. Girault. Second harmonic generation of glucose oxidase at the air/water interface. *Biophysical Journal*, **77**:3350–3355, December 1999.

[78] L. Marrucci, D. Paparo, G. Cerrone, C. de Lisio, E. Santamato, S. Solimeno, S. Ardizzone, and P Quagliotto. Probing interfacial properties by optical second-harmonic generation. *Optics and Lasers in Engineering*, **37**:601–610, 2002.

[79] H. Ishida, R. Mizoguchi, K. Onda, C. Hirose, S. S. Kano, and A. Wada. Second harmonic observation of Cu(111) surface: in situ measurements during molecular adsorption. *Surface Science*, **526**:201–207, 2003.

[80] S. H. Tey, K. Prasad, K. C. Tee, L. H. Chan, and E. G. Seebauer. Non-linear optical studies of copper diffusion at surfaces and interfaces of microelectronic interconnect structures. *Thin Solid Films*, **466**:217–224, 2004.

[81] O. A. Aktsipetrov, A. A. Fedyanin, A. V. Melnikov, E. D. Mishina, A. N. Rubtsov, M. H. Anderson, P. T. Wilson, M. ter Beek, X. F. Hu, J. I. Dadap, and M. C. Downer. DC-electric-field-induced and low-frequency electromodulation second-harmonic generation spectroscopy of Si(001) -SiO_2 interfaces. *Physical Review B-Condensed Matter* , **60**(12):8924–8938, September 1999.

[82] I. V. Kityk. Photoinduced nonlinear optics methods as new methods of monitoring the superconducting films. *Optics and Lasers in Engineering* , **40**:179–187, 2003.

[83] I. V. Kityk. Non-linear optical phenomena in the large-sized nanocrystallites. *Journal of Non-Crystalline Solids* , **292**:184–201, 2001.

[84] H. W. K. Tom, Y. M. Chang, and H. Kwak. Coherent phonon and electron spectroscopy on surfaces using time-resolved second-harmonic generation. *Applied Physics B-Lasers & Optics* , **68**:305–313, 1999.

[85] A. J. Timson, R. D. Spencer-Smith, A. K. Alexander, R. Greef, and J. G. Frey. Second harmonic ellipsometry. *Measurement Science and Technology* , **14**:508–515, 2003.

[86] K. Plucinski, I. V. Kityk, and A. Mefleh. Elliptically polarized light-induced second harmonic generation in SiN_xO_y. *Journal of Non-Crystalline Solids* , **2002**:143-154, 2000.

[87] A. Tadjeddine, O. Pluchery, A. Le Rille, C. Humbert, M. Buck, A. Peremans, and W. Q. Zheng. What can we learn from the non-linear optical investigation of the solid — liquid interface? *Journal of Electroanalytical Chemistry* , **473**:25–33, 1999.

[88] A. Tadjeddine, A. Le Rille, O. Pluchery, P. Hébert, W. Q. Zheng, and T. Marin. Adsorption of pyridine on gold, studied by different frequency generation (DFG) using the (CLIO-FEL). *Nuclear Instruments and Methods in Physics Research A* , **429**:481–484, 1999.

[89] A. Tadjeddine, A. Le Rille, O. Pluchery, F. Vidal, W. Q. Zheng, and A. Peremans. Sum and difference frequency generation at the electrochemical interface. *Physica Status Solidi A - Applied Research* , **175**:89–107, 1999.

[90] J. C. Diettrich, I. T. McKinnie, and D. M. Warrington. Novel single longitudinal mode near infrared lasers. In *Conference on Lasers and Eletro-optics/Europe* , Nice, France, 2000. Optical Society of America.

[91] C. J. Lee, G. W. Baxter, J. C. Diettrich, I. T. M. McKinnie, D. W. Warrington, P. C. R. Smith, G. W. Ross, and D. C. Hanna. Periodically poled lithium niobate optical parametric oscillator pump-tuned by a single-axial-mode Ti:sapphire laser. *Optical Engineering* , **41**(4):848–851, 2002.

[92] C. J. Lee, D. W. Warrington, I. T. M. McKinnie, P. G. R. Smith, G. W. Ross, and D. C. Hanna. High pulse energy, periodically poled lithium niobate optical parametric oscillator pumped by a grazing incidence coupled cavity Ti:sapphire laser. *Trends in Optics and Photonics* , **68**:182–188, 2002.

[93] P. Schlup, G. W. Baxter, and I. T. McKinnie. Single-mode near- and mid-infrared periodically poled lithium niobate optical parametric oscillator. *Optics Communications*, **176**:267–271, 2000.

[94] D. J. LeCaptain and K. A. Burglund. The application of second harmonic generation for in situ measurement of induction time of selected crystalization systems. *Journal of Crystal Growth*, **203**(4):564–569, 1999.

[95] B. F. Henson, B. W. Asay, R. K. Sander, S. F. Son, J. M. Robinson, and P. M. Dickson. Dynamic measurement of the HMX β-δ phase transition by second harmonic generation. *Physical Review Letters*, **82**(6):1213–1216, 1999.

[96] S. F. Son, B. W. Asay, B. F. Henson, R. K. Sander, A. N. Ali, P. M. Zielinski, D. S. Phillips, R. B. Schwartz, and C. B. Skidmore. Dynamic observation of a thermally activated struture change in 1,3,5-triamino-2,4,6-trinitrobenzene (TATB) by second harmonic generation. *Journal of Physical Chemistry B*, **103**:5434–5440, 1999.

[97] M. Kiguchi, M. Kato, N. Kumegawa, and Y. Taniguchi. Technique for evaluating second-order nonlinear optical materials in powder form. *Journal of Applied Physics*, **75**(9):4332–4339, 1994.

[98] J. C. Hooker and J. M. Torkelson. Coupling of probe reorientation dynamics and rotor motions to polymer relaxation as sensed by second harmonic generation and fluorescence. *Macromolecules*, **28**:7683–7692, 1995.

[99] Y. G. Fokin, S. V. Krupenin, T. V. Murzina, O. A. Aktispetrov, S. Soria, and G. Marowsky. Ferroelectric switching and phase transitions in thin cells of chiral smectic liquid crystals. *Surface Science*, **566-568**:783–788, 2004.

[100] C. J. Strachan, C. J. Lee, and T. Rades. Partial Characterisation of different mixtures of solids by measuring the optical nonlinear response. *Journal of Pharmaceutical Sciences*, **93**(3):733-742, 2004.

[101] C. J. Strachan, T. Rades, and C. J. Lee. Determination of the optical second harmonic response of pharmaceutical solid-solid mixtures. *Optics and Lasers in Engineering*, **43**:209-220, 2005.

[102] ICH Expert Working Group. ICH harmonised tripartite guideline. Specifications: test procedures and acceptance criteria for new drug substances and new drug products: chemical substances (Q6A). http://www.ich.org/MediaServer.jser?@_ID=430&@_MODE=GLB, 1999.

[103] ICH Expert Working Group. ICH harmonised tripartite guideline. Text on Validation of Analytical Procedures (Q2A). http://www.ich.org/MediaServer.jser?@_ID=418&@_MODE=GLB, 1994.

[104] P. J. Heaney, C. T. Prewitt, and G. V. Gibbs, editors. *Silica: physical behavior, geochemistry and materials applications*, volume 29 of *Reviews in minerology*. Mineralogical Society of America, Washington, D.C., 1994.

[105] E. C. Sullivan and W. S. Taylor. *Glass*. Number 1 304 623. United States Patent Office, 1919.

[106] J. P. Bachheimer and G. Dolino. Measurement of order parameter of alpha-quartz by second-harmonic generation of light. *Physical Review B*, **11**(8):3195–3205, 1975.

[107] D. A. Pinnick, S. A. Lee, X. Sun, and H. J. Simon. Optical second-harmonic-generation study of quartz up to 31 GPa. *Physical Review B*, **55**(13):8031-8033, 1997.

[108] K. Parfitt, editor. *Martindale: The Complete Drug Reference* , pages 863–864. Pharmaceutical Press, London, UK, 32nd edition, 1999.

[109] D. P. Ip, G. S. Brenner, J. M. Stevenson, S. Lindenbaum, A. W. Douglas, Klein S. D., and J. A. McCauley. High resolution spectroscopic evidence and solution calorimetry studies on the polymorphs of enalapril maleate. *International Journal of Pharmaceutics*, **28**:183–191, 1986.

[110] K. Florey, editor. *Analytical Profiles of Drug Substances* , volume 16, pages 227–230. Academic Press Inc., 1987.

[111] K. Parfitt, editor. *Martindale: The Complete Drug Reference* , page 1474. Pharmaceutical Press, London, UK, 32nd edition, 1999.

[112] A. H. Forster, J. Hempenstall, and T. Rades. Characterization of glass solutions of poorly water-soluble drugs produced by melt extrusion with hydrophilic amorphous polymers. *Journal of Pharmacy and Pharmacology* , **53**:303–315, 2001.

[113] D. Giron, P. Piechon, S. Goldbronn, and S. Pfeffer. Thermal analysis, microcalorimetry and combined techniques for the study of the polymorphic behaviour of a purine derivative. *Journal of Thermal and Analytical Calorimetry* , **57**:61–73, 1999.

[114] V. P. Tanninen and J. Yliruusi. X-ray powder diffraction profile fitting in quantitative determination of two polymorphs from their powder mixture. *International Journal of Pharmaceutics*, **81**(2-3):169–177, 1992.

[115] Z. D. Dong, E. J. Munson, S. A. Schroeder, I. Prakash, and D. J. W. Grant. Neotame anhydrate polymorphs II: quantitation and relative physical stability. *Pharmaceutical Research*, **19**(9):1259–1264, 2002.

[116] C. B. Rawle, C. J. Lee, C. J. Strachan, K. Payne, P. J. Manson, and T. Rades. Towards characterisation and identification of solid state pharmaceutical mixtures through second harmonic generation. *Journal of Pharmacuetical Sciences*, **95**(4):761-768, 2006.

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN: 1-60021-194-1
Editor: Steven B. Larkin, pp. 119-171

Chapter 4

ENVIRONMENTAL MONITORING BY LASER RADAR

Fiorani Luca*

University of Rome "Tor Vergata", Degree Course in Materials Science, Via della Ricerca Scientifica 1, 00133 Roma, Italy

Abstract

Since the discovery of laser, optical radars or lidars have been successfully applied to the monitoring of the three main environments of our planet: lithosphere, hydrosphere and atmosphere. The transmission of light depends on the medium: while in soil it does not propagate, in water and air it can typically travel for meters and kilometers, respectively. Analogously, three theoretical frameworks can be established: lidar equations for hard, dense and transparent targets. Such different behaviors drive the range of lidar applications in the three above mentioned environments. In the lithosphere, laser radars have been applied to three-dimensional scans of underground cavities. In the hydrosphere, lidar fluorosensors have been very effective in the bio-optical characterization of the first layers of sea waters. Such instruments are usually aboard planes and ships and can help filling the gap between in situ measurements and satellite imagery. The medium where lidars are unbeaten, at least for some purposes, is air. Atmospheric applications of laser radars range from troposphere (e.g. pollution monitoring in urban areas and wind speed measurements) to stratosphere (e.g. polar stratospheric cloud detection and ozone hole assessment), and even mesosphere (e.g. profiling of K and Na). The purpose of this paper is twofold: from one hand, the interested reader is introduced in lidar science and technology, to the other one, the researcher familiar with laser remote sensing is faced with some current investigations. The first aim is achieved by introducing the lidar principles, for hard, dense and transparent targets, and by illustrating selected case studies, taken from the experience of the author, in order to exemplify some relevant applications of such principles. The second one is pursued by describing in more detail the most recent results obtained by the author in the field of environmental monitoring by laser radar.

* E-mail address: fiorani@frascati.enea.it, Internet: www.afs.enea.it/fiorani, Tel.: +39(06)94005861, Fax: +39(06)94005312, Address: ENEA, Via Fermi 45, 00044 Frascati RM

1. Introduction

Lithosphere, hydrosphere and atmosphere play a major role in sustaining life on our planet. All the biogeochemical cycles take place on them or are part of the matter exchange among them. A thorough understanding of such global processes requires powerful tools.

Remote sensing can support a systematic investigation of natural processes, anthropogenic effects and their interactions. It can be either passive [Elachi 1987], using the Sun as a light source, or active [Measures 1992], being assisted by artificial light sources. The passive technique relies on observing the Earth, i.e. on measuring the sunlight backscattered by the planet surface at different wavelengths. It suits especially satellite borne sensors. The active method is usually based on the optical radar. It is applied mainly to air or ship borne sensors.

The revolutionary capabilities offered to environmental monitoring by light amplification by stimulated emission of radiation (laser) [Hecht 1985, Svelto 1998] have been firstly understood and exploited in the atmospheric field [Fiocco and Smullin 1963, Goyer and Watson 1963, Collis 1966]. Actually, the concept of light detection and ranging (lidar) has been introduced before the laser discovery, in the context of pulsed light detectors of clouds [Middleton and Spilhaus 1953]. Anyway, laser sources, because of high intensity, small divergence, good monochromaticity and the possibility of short pulse emission, are ideal tools for lidar and that word is today synonym of laser radar.

The measurement, up to many kilometers of height, of density, temperature and humidity of air, the detection of trace gases, the study of clouds, the observation of stratospheric aerosols, the probing of high atmosphere and the monitoring of pollutants are some examples among the possible atmospheric applications of lidar [Collis and Russel 1976, Measures 1992, Grant 1995].

Lidar has been successfully used also in the hydrosphere for bathymetric surveys in shallow waters, turbidity measurements, pollution detections, especially in case of oil spills, and phytoplankton mappings [Measures 1992, Grant 1995, Bunkin and Voliak 2001].

As far as lithosphere is concerned, lidar employment has been limited by the lack of light propagation in soil. In this case, the more relevant environmental applications of laser radar are related to the characterization of hollow spaces beneath the terrestrial surface, e.g. the three-dimensional scan of underground cavities by laser range-finder [Fiorani et al. 2000].

2. Lidar Principle

2.1. Hard Target (Lithospheric Applications)

A lidar is essentially composed of a transmitter (laser and beam shaping optics) and a receiver (telescope and signal detection electronics). Its principle of operation is illustrated in Figure 1: the target at the distance R from the system sends back part of the laser pulse toward the telescope surface. Consequently, the analysis of the detected signal as a function of t, time interval between emission and detection, allows one to measure R by the simple relation:

$$R = \frac{c\,t}{2}, \tag{1}$$

where c is the speed of light in the medium where the beam propagates, i.e. air in most cases.

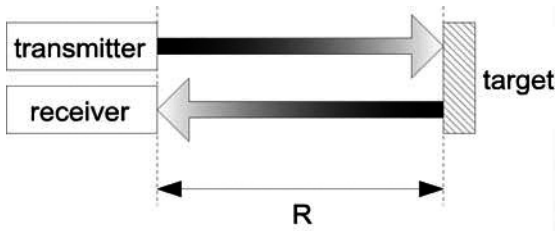


Figure 1. Lidar principle of operation.

If the target is Lambertian [Jelalian 1992], contains the laser footprint and the transmitter divergence is smaller than the receiver field of view, the received power is given by [Mamon et al. 1978]:

$$P = P_0 \frac{A\,\eta_t\,\eta_r\,\rho}{\pi\,R^2} \exp(-2\,\alpha\,R), \tag{2}$$

where P_0 is the transmitted power, A is the receiver area, η_t and η_r are the transmitter and receiver efficiency, respectively, ρ is the target reflectivity and α is the laser beam extinction coefficient of the involved air volume. In this case, α is regarded as constant simply because R is small.

If the laser pulse is Gaussian, the accuracy in the measurement of the distance between lidar and target is given by [Carmer and Peterson 1996]:

$$\sigma_R^2 = \left(\frac{c\,\tau_L}{2}\frac{1.2}{SNR}\right)^2 + \left(\frac{c\,\tau_D}{2}\right)^2, \tag{3}$$

where τ_L is the pulse duration, SNR the signal-to-noise ratio and τ_D the resolution of the device measuring t.

Equation (3) is obtained under the assumption that any bias in the measurement of t has been avoided (as an example of source of bias, let us consider a threshold discriminator: such device measures the occurrence of a higher pulse before that of a lower one).

Typically, the distance between optical radar and target is obtained measuring <R>, mean of R calculated averaging on N laser pulses. The corresponding accuracy is:

$$\sigma_{<R>} = \frac{\sigma_R}{\sqrt{N}}. \tag{4}$$

If, as usual, the signal is detected by a detector coupled to a preamplifier, SNR can be written as [Jelalian 1992]:

$$SNR = \frac{i_s}{\sqrt{i_{sn}^2 + i_{bn}^2 + i_{dn}^2 + i_{an}^2}}, \tag{5}$$

where i_s is the signal current (root mean squared) and i_{sn}, i_{bn}, i_{dn} and i_{an} are the noise currents (root mean squared) due to, respectively, signal fluctuation (shot noise), environmental radiation (background noise), dark current (detector noise) and preamplifier (amplification noise). More explicitly [Mamon et al. 1978]:

$$SNR = \frac{Q\,G\,P}{\sqrt{2\,e\,B\,G^2\,(Q\,P + Q\,P_b + i_d) + B\,i_a^2}}, \tag{6}$$

$$Q = \frac{e\,\lambda\,\eta}{h\,c}, \tag{7}$$

$$P_b = A\,\eta_t\,\Delta\lambda\,H\,\Omega\left\{\rho\exp(-\alpha\,R) + \frac{1}{4}\left[1 - \exp(-\alpha\,R)\right]\right\}\frac{1}{\pi}, \tag{8}$$

where G is the detector gain, e is the electron charge, B is the receiver bandwidth, i_d is the detector dark current, i_a is the preamplifier noise current, λ is the transmitted wavelength, η is the detector quantum efficiency, h is the Planck's constant, $\Delta\lambda$ is the receiver passband, H is the solar irradiance and Ω is the receiver solid angle.

2.2. Dense Target (Hydrospheric Applications)

Let us consider the case of a fluorescing target, very important for natural waters because it allows one to detect chromophoric dissolved organic matter (CDOM) and chlorophyll-a (chl-a), the main indicator of phytoplankton. In this case, the principle of operation is called laser-induced fluorescence (LIF) and the instrument lidar fluorosensor. In general, the experimental system is above the sea (Figure 2), at a range R_w from its surface, and the laser beam, after propagation in air, probes a water layer characterized by the extinction coefficient α_w. Also in this case, the extinction coefficients in air and water are regarded as constant because the distances are small.

With respect to equation (2), ρ/π has to be replaced by the product of N_F (number density of fluorescing molecules) times σ_F (fluorescence cross section). The effect of the air-water interface on the propagation of forward and backward photons is taken into account with φ, two-way transmission factor, and m, refractive index of water (both can be regarded as constant because their variation with the wavelength is small). Moreover, the dependence of

laser emission and transmitter/receiver efficiencies on the wavelength has to be considered but, once the transmitted and received wavelengths are fixed, all the related parameters can be included in the system constant k_F. If the transmitted beam is contained in the receiver field of view, $\alpha_w \times c \times \tau_L > 10$, $\tau_L/\tau_D < 5$ and $\tau_L > \tau$ (fluorescence decay time), hypotheses that usually hold for lidar fluorosensors aimed to CDOM and chl-a detection, the received energy is given by [Measures 1992]:

$$E_F(\lambda_F, R) = E_0 \frac{k_F \, A \varphi \, N_F(R) \sigma_F(\lambda, \lambda_F)}{R^2 \, m^2 \, [\alpha_w(\lambda) + \alpha_w(\lambda_F)]} \exp\{-[\alpha(\lambda) + \alpha(\lambda_F)] R_w\} \quad , \quad (9)$$

where λ_F is the fluorescence wavelength and E_0 the transmitted energy.

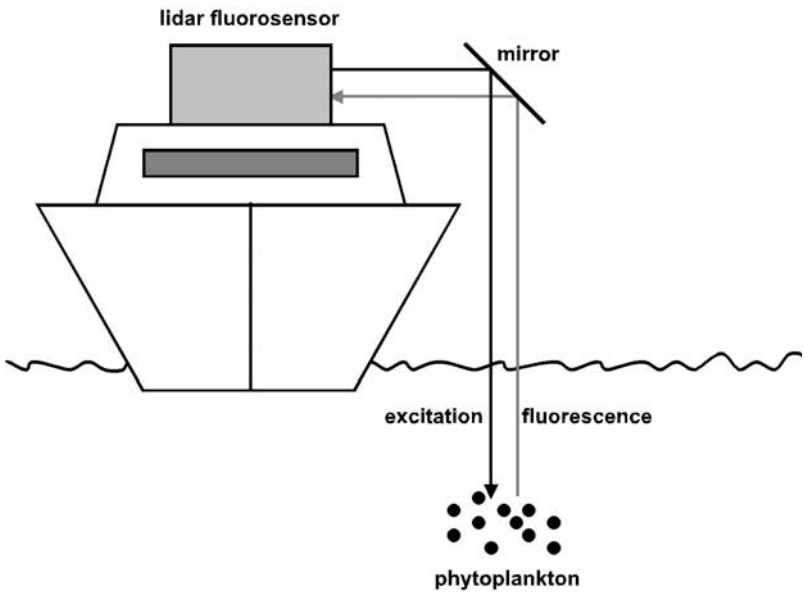


Figure 2. Lidar fluorosensor principle of operation.

A similar equation can be written for the signal coming from Raman scattering of water [Measures 1992]:

$$E_R(\lambda_R, R) = E_0 \frac{k_R \, A \varphi \, N_R \, \sigma_R(\lambda, \lambda_R)}{R^2 \, m^2 \, [\alpha_w(\lambda) + \alpha_w(\lambda_R)]} \exp\{-[\alpha(\lambda) + \alpha(\lambda_R)] R_w\} \quad , \quad (10)$$

where λ_R is the Raman-shifted wavelength, N_R the number density of water molecules (practically constant) and σ_R the Raman scattering cross section for the OH stretching vibrational mode of liquid water.

If fluorescence and Raman signal are acquired simultaneously, the following ratio (also called fluorescence in Raman units) can be calculated:

$$E^*(R) \equiv \frac{E_F(\lambda_F, R)}{E_R(\lambda_F, R)} = \frac{k_F \; N_F(R)\, \sigma_F(\lambda, \lambda_F)\, [\alpha_W(\lambda) + \alpha_W(\lambda_R)] \exp\{-[\alpha(\lambda) + \alpha(\lambda_F)] R_W\}}{k_R \; N_R(R)\, \sigma_R(\lambda, \lambda_R)\, [\alpha_W(\lambda) + \alpha_W(\lambda_F)] \exp\{-[\alpha(\lambda) + \alpha(\lambda_R)] R_W\}}. \quad (11)$$

If, as usual for CDOM and chl-a detection, the extinction coefficients ratio changes slowly and the exponentials ratio is close to unity, E* can be written as:

$$E^*(R) = k \frac{N_F(R)}{N_R}, \quad (12)$$

where k is a new system constant including also the cross sections. This latter formula implies that, in the above mentioned hypothesis, the fluorescence in Raman units is proportional to the density of fluorescing molecules. As a consequence, the absolute concentration of fluorescing molecules can be obtained calibrating the lidar fluorosensor with in situ measurements. The measurement error of the final result can be derived by the statistical parameters of that fit.

2.3. Transparent Target (Atmospheric Applications)

In case of atmospheric applications, the target and the medium before it are of the same nature. Usually, the target at range R is defined as the air layer defined by τ_D, i.e. delimited by the distances R and $R+c\times\tau_D/2$, because the photons detected in the time interval defined by t and $t+\tau_D$ come from that layer (Figure 3). Their number n is proportional to the thickness $c\tau_D/2$ and to the laser beam backscattering coefficient β of the involved air volume [Measures 1992]: in fact, their product take the place of ρ/π in equation (2):

$$n(R, \lambda) = n_0(\lambda)\, \zeta(\lambda) \frac{A}{R^2} \beta(R, \lambda) \frac{c\tau_D}{2} \exp\left[-2 \int_0^R \alpha(R', \lambda) dR' \right], \quad (13)$$

where n_0 is number of photons of the original pulse and the system efficiency ζ includes η_t and η_r.

Note that, in this case, α can not be regarded as constant because we are interested in long distance remote sensing and atmospheric layer with different optical density can be encountered by the laser beam.

The measurement errors will be discussed in section 5.

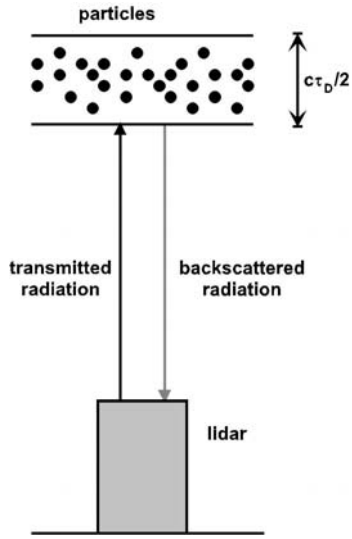


Figure 3. Atmospheric lidar principle of operation.

3. Lithospheric Applications

3.1. Laser Range-Finder for the Three-Dimensional Scan of Underground Cavities

The two main application fields of lithospheric laser range finders are mining [Measurement Devices 2005] and archaeology [Fiorani et al. 2000]. Mining systems have larger range (~ 100 m), smaller accuracy (~ 0.1 m) and can be useful in geology (territory assessment), civil engineering (structural inspection) and remote sensing in hazardous area (nuclear power plants). Various kinds of underground cavities (homes, temples, tombs, ovens, wells, tanks and aqueducts) can be found among the archaeological remains. A precise determination of the dimensions of such buildings (range ~ 10 m, accuracy ~ 0.01 m) is useful to guide possible excavations. For this reason, GEOLIDAR, a compact laser radar, has been developed. The scanning head of the system can be inserted in a small drilling and, rotating the beam, a three-dimensional image of the cavity can be obtained.

A laser provides the light pulse. The radiation is conducted in the cavity through an optical fiber. This is accomplished thanks to injection optics. The transferred beam is then collimated and aimed in a well-defined direction by the scan optics. Part of the radiation backscattered by the cavity wall, after having been collected and filtered by the reception optics, containing an interference filter (IF), reaches the photomultiplier (PM) where it is transformed in an electronic signal. Scan optics, reception optics and PM are integrated in the scanning head.

A Nd:YAG laser with frequency tripling has been chosen as source emitting short pulses of ultraviolet radiation. The heart of the scan optics consists of two prisms able to rotate independently around perpendicular axes: in this way, the laser beam can be aimed in nearly all the solid angle. The rotations are performed by two stepper motors.

The system is computer driven through two RS232 and a GPIB interfaces: the first RS232 is for the laser and the second for the controller of the stepper motors, the GPIB is for the oscilloscope. The trigger is provided to the oscilloscope by a PM that observes a very small part of the laser pulse reflected by two uncoated optical surfaces, transferred thanks to an optical fiber and transmitted through an IF. The trigger and receiver PMs are identical, in order to avoid any bias in the distance measurement that could arise from a difference in rise time. The operator can adjust their responsivities through potentiometers that control the voltage of the power supplies. If necessary, a preamplifier and an amplifier can be inserted after each PM. Then, the signal undergoes analog-to-digital conversion in the oscilloscope and is transferred to the computer where it is stored. Finally, the GEOLIDAR software measures the distance and controls laser, stepper motors and oscilloscope.

The theoretical accuracy of GEOLIDAR has been calculated substituting in equations (3, 4, 6 – 8) the fundamental constants, the characteristics of the system components and the typical values of other parameters (table 1), obtaining an accuracy of 0.015 m. This value is extremely encouraging if compared with the typical accuracy obtained by ground penetrating radars (0.3 – 0.5 m) and taking into account that the laser emits 100 pulses in only 10 s. The dominant source of error is the resolution of the device measuring t (digital oscilloscope). The noise comes from the signal fluctuation (about 2/3) and the optional preamplifier (about 1/3). The noise due to background and dark current is negligible.

Table 1. Values of the parameters used in the calculation of the GEOLIDAR accuracy.

parameter	**value**	**unit**
P_0	1.3×10^{7}	W
A	5×10^{-5}	m^2
η_t	1.6×10^{-2}	
η_r	0.4	
ρ	0.05	
α	1.2×10^{-3}	m^{-1}
R	10	m
c	3×10^{8}	$m\ s^{-1}$
τ_L	6×10^{-9}	s
τ_D	10^{-9}	s
G	9×10^{5}	
e	1.6×10^{-19}	C
B	5×10^{8}	Hz
i_d	5×10^{-10}	A
i_a	8×10^{-7}	A
λ	3.6×10^{-7}	m
η	0.06	
h	6.6×10^{-34}	J s
$\Delta\lambda$	1	nm
H	0.4	$W\ m^{-2}\ nm^{-1}\ sr^{-1}$
Ω	0.3	sr
N	100	

The accuracy has also been measured in laboratory mounting the scanning head on an optical bench and aiming the beam at a target at known distance (Figure 4). After acquisition

and analog-to-digital conversion, the output signals of the trigger and receiver PMs are transferred to the computer where an algorithm calculates the distance between GEOLIDAR and target.

Observing the output signal of the trigger and receiver PMs, one notes that their shapes are similar for a given pulse, even if the pulse shape changes slightly in time. Consequently, it is necessary to find an algorithm able to determine the relative delay of the two signals from their comparison. The more refined method is the calculation of their correlation as a function of their relative delay. Unfortunately, such method is quite slow and can hardly suit a system that should take many measurements in real time. Finally, after some trials, an excellent compromise between accuracy and speed has been found. Such procedure can be schematized as follows:

- the two signals are inverted (the output of the PMs is negative);
- the maximum of both signals is determined;
- the instants where each signal reaches half of its maximum are measured;
- the time interval between the aforementioned instants is calculated;
- steps 1 – 4 are repeated for N (100, in our case) laser pulses and, at last, the computer provides the mean value.

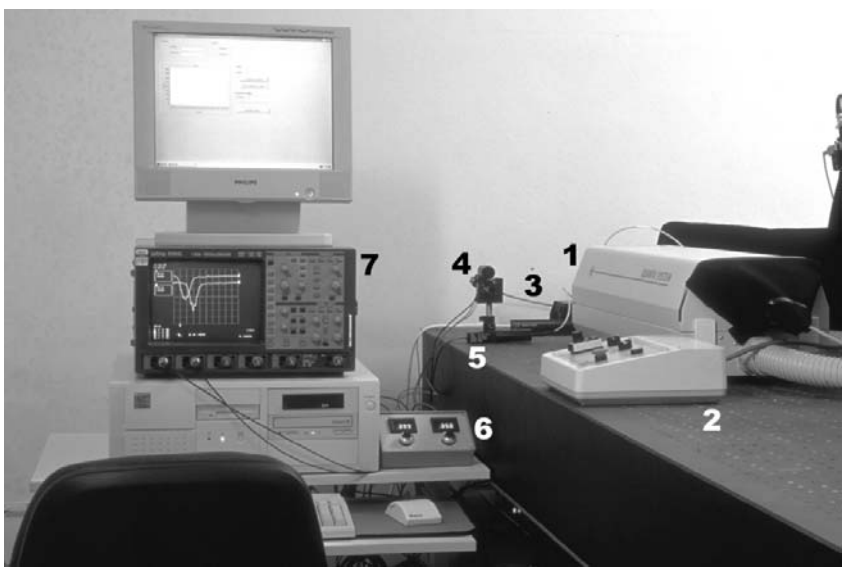


Figure 4. GEOLIDAR during laboratory tests: frequency-tripled Nd:YAG laser (1), laser control (2), optical fiber (3), scanning head (4), trigger PM (5), PMs control (6), oscilloscope (7).

Of course, we should subtract to such value the offset due to the dissimilar paths of the trigger and receiver signals caused by the different length of optical fibers and electronic cables. In order to solve this problem, we can determine the calibration curve of GEOLIDAR. From equation (1), we expect such curve to be a straight line with slope c/2 and intercept the aforementioned offset. Calculating the linear fit of the data, one obtains a result like:

$$d = m\,t + n\,, \tag{14}$$

where d is the correct distance and t is the measured time. The intercept is -1.860 m and the slope is, as expected, 1.500×10^8 m/s. The difference between points and curve is small, more precisely it is comprised between 0.02 m and 0.03 m. Such values are greater of a factor 1.5 – 2 with respect to the theoretical accuracy. This result is not surprising: the calculation did not take into account the uncertainty in the trigger instant determination that, in our case, is performed in a way similar to that of the reception instant.

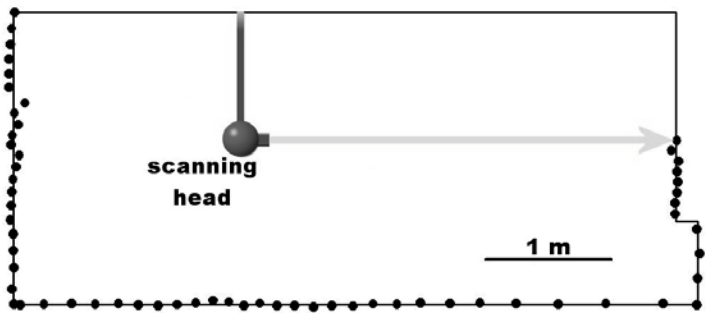


Figure 5. Example of the two-dimensional scans of a cavity acquired by GEOLIDAR. The points are the measurements of GEOLIDAR, the polygon is the shape of the cavity as it is retrieved by the software reconstruction.

The final test of GEOLIDAR has been performed by scanning an artificial cavity. The analysis of such scans showed that the position of the walls was determined with an accuracy of 0.01 – 0.02 m. This result does not contradict the range accuracy found during the calibration (0.03 m) because each determination of the position of a wall is the result of some measurements of the range at different angles. For the sake of simplicity, let us consider a two-dimensional scan of the cavity (Figure 5). The operator sets the angle step and the initial direction (horizontal in this case). Then, the scan continues up to the maximum allowed by the mechanical constraints. The three-dimensional scan is accomplished by repeating automatically the two-dimensional scan for different angles. If a zone has not been covered by a scan, the missing measurements can be taken after repositioning of GEOLIDAR. In conclusion, the test of GEOLIDAR proved that it was able to characterize the artificial cavity and confirmed the accuracy measured during its calibration.

4. Hydrospheric Applications

4.1. Lidar Fluorosensor

Lidar fluorosensors are probably the more common LIF instruments and have been extensively operated to monitor water bodies [Reuter et al. 1993]. LIF spectra contain signatures of phytoplankton pigments, CDOM and dispersed impurities, such as crude oils [Bristow et al. 1981, Hoge and Swift 1981, Koechler et al. 1992]. In a typical lidar fluorosensor, CDOM and chl-a are detected measuring their emission at around 450 and 680 nm, respectively, after excitation by a frequency-tripled Nd:YAG laser at 355 nm. This technique has been applied to air or ship borne sensors, thus supplying in situ measurements (sea truth). When operated in the pump-and-probe emission mode [Chekalyuk et al. 1993], a

lidar fluorosensor can determine the chl-a fluorescence yield, a parameter related to the photosynthetic electron transport, under actual environmental conditions (nutrient availability, sun illumination, salinity, pH, presence of toxic substances, etc.). Thanks to narrowband filtering and electronic gating, LIF signals do not need corrections for radiometric and spectral characteristics of solar irradiance and surface reflectance. Furthermore, due to the short distance from the target (few meters), atmospheric effects are negligible.

In this section, the main characteristics of the ENEA Lidar Fluorosensor (ELF) [Barbini et al. 1999] are reviewed and some recent results are described. ELF has been operational aboard the research vessel (RV) *Italica* during four Italian expeditions in Antarctica (13^{th}, 15^{th}, 16^{th} and 18^{th} in 1997-98, 2000, 2001 and 2003, respectively) [Barbini et al. 2001b] and the MIPOT (Mediterranean Sea, Indian and Pacific Oceans Transect) oceanographic campaign (2001-02) [Barbini et al. 2004]. As we will see in the next section, its data has been used for the validation and/or the calibration of space borne radiometers [Barbini et al. 2001a, 2003a], up to the calculation of new estimations of satellite sensed primary production (PP) [Barbini et al. 2003b, 2005] and CDOM.

ELF (Figure6) is part of a complete laboratory, including local and remote instruments for continuous monitoring and in situ sampling, lodged into an ISO 20' container. It is assisted by ancillary instruments: a lamp spectrofluorometer, a pulsed amplitude fluorometer (PAM), a solar radiance detector, measuring the photosynthetically available radiation (PAR), and a global positioning system (GPS).

The light source is a frequency-tripled Nd:YAG laser (355 nm) followed by a beam expander (BE). ELF transmits the exciting pulse and receives the generated radiation through an optical window and an external mirror in order to reach the sea surface at normal incidence. The optical signal, after collection by a telescope, traverses a dichroic filter, rejecting most of the laser beam, and is forwarded by a multi-arm optical fiber (OF) to IFs. After that spectral selection, PMs perform the detection. Their electronic output is digitized by analog-to-digital converters (ADCs). A personal computer (PC), embedded in a versa module eurocard (VME) bus, controls all the experimental settings, including the normal or pump-and-probe excitation, the laser transmitter energy, the PMs high voltage (HV) and gating time and the data acquisition parameters.

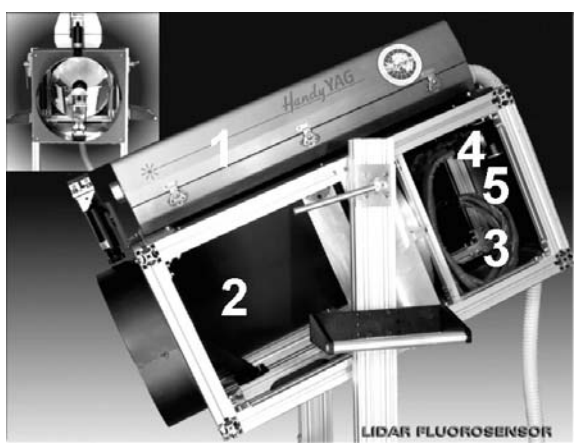


Figure 6. Side view of ELF (a front view is given in the insert). Transmitter: frequency-tripled Nd:YAG laser (1). Receiver: Cassegrain telescope (2). Detection: OF (3), IFs (4) and PMs (5).

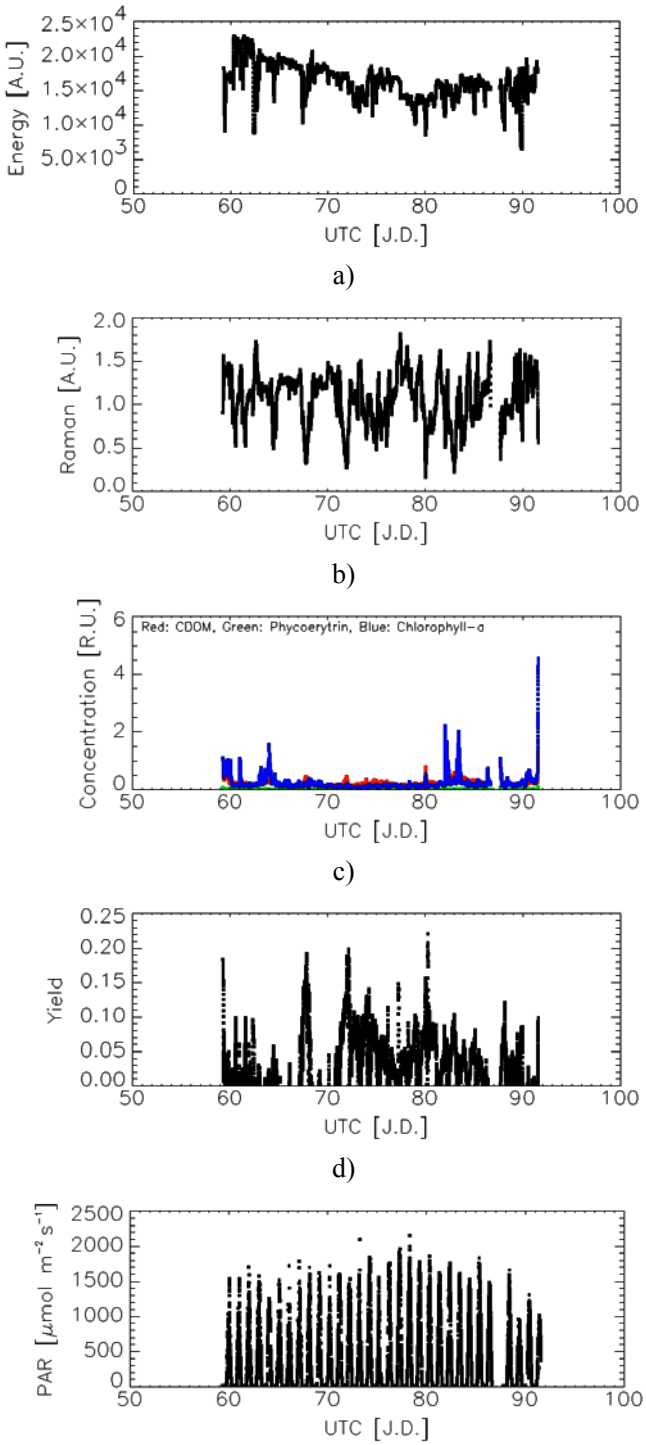


Figure 7. Simultaneous measurements of a) laser energy, b) Raman signal, c) concentrations of CDOM, phycoerythrin and chl-a, d) chl-a fluorescence yield and e) PAR during the New Zealand – Italy transect of MIPOT. The x-axis represents the universal time coordinated (UTC) in Julian Days of 2002.

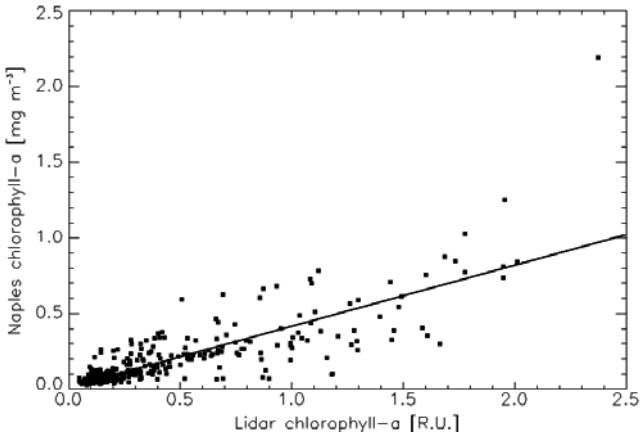


Figure 8. Calibration with conventional techniques of chl-a measured by ELF during MIPOT. Number of points: 332. Correlation coefficient: 0.82. Intercept: 0.0137 ± 0.0094. Slope: b = 0.405 ± 0.015.

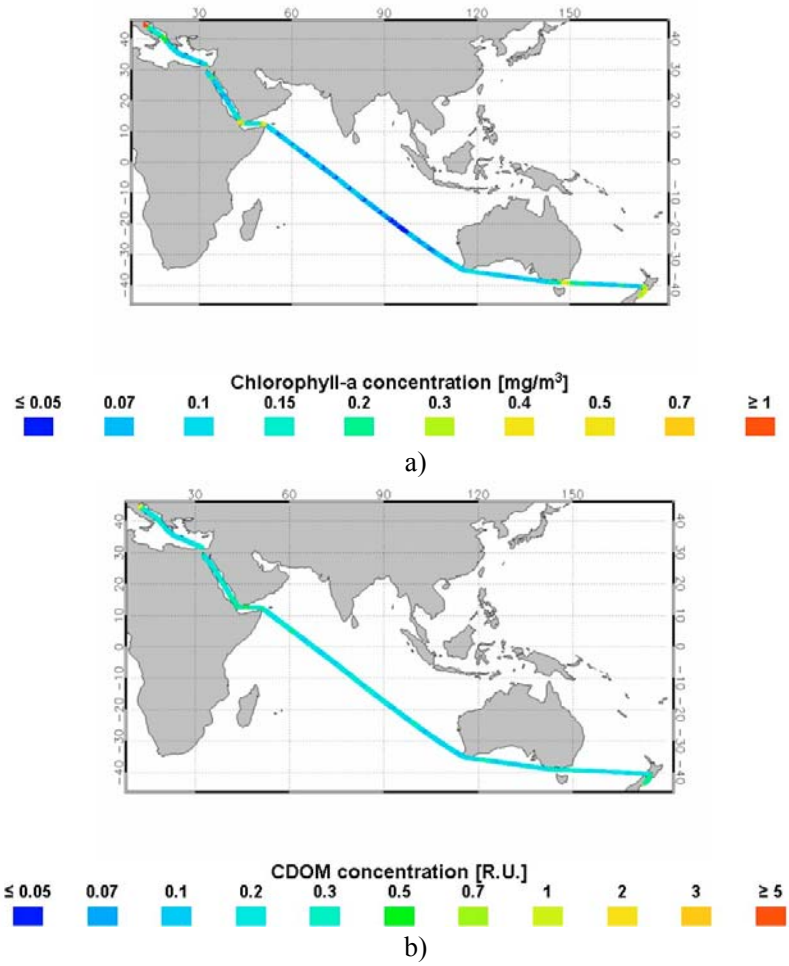


Figure 9. Thematic maps of a) chl-a and b) CDOM by ELF during the New Zealand – Italy transect of MIPOT.

Table 2. Typical specifications of ELF.

Transmitter	Laser	Nd:YAG
	Wavelength	355 nm
	Pump pulse energy	30 mJ
	Probe pulse energy	3 mJ
	Pulse duration	10 ns
	Pulse repetition rate	10 Hz
	Beam expansion	1 × – 20 ×
Receiver	Telescope	Cassegrain
	Clear aperture	0.4 m
	Focal length	1.65 m
	OF length	1 m
	OF diameters	Input 25.4 mm, output 7 mm
	IFs center wavelength	404, 450, 585, 680 nm
	IFs bandwidth	5 nm FWHM
	PMs	Hamamatsu R3896, R1477, R928
	Gated HV	100 – 200 ns
Electronics	Bus	ISA[a]-VME mixed bus
	ADC	Caen V265 (15 bit)
	PC	VME-CPU[b] embedded 486-100 MHz

[a] Industry standard architecture.
[b] Central processing unit.

The beam footprint on the sea surface is a circle of about 0.1 m. The depth of the probed layer (15 m) is determined by an electronic gate. The main spectral channels have a full width at half maximum (FWHM) of 5 nm and are centered at 404, 450, 585 and 680 nm, corresponding to water Raman backscattering, CDOM, phycoerythrin and chl-a fluorescence, respectively. An example of the data recorded by ELF (laser energy, Raman signal, concentrations of CDOM, phycoerythrin and chl-a, and chl-a fluorescence yield) and by the solar radiance detector (PAR) is given in Figure 7.

The consumption of the Nd:YAG flashlamp can be monitored observing the laser energy. The PAR measurement is essential to account for chl-a fluorescence variations related to changes in the downwelling light field [Hoge et al. 1998]. At first, the concentrations of CDOM, phycoerythrin and chl-a are released in Raman units. The transformation from Raman units to mg m^{-3} is performed by calibration with conventional techniques in some selected points: an example of such an experimental fit is given in Figure 8. Lamp spectrofluorometer and PAM provide some accessory measurements of natural waters. Eventually, the data are georeferenced with GPS. An example of thematic maps of chl-a and CDOM is shown in Figure 9.

The main characteristics of ELF are listed in table 2. More details on ELF, the ancillary instruments, the measurement principle, the calibration procedure and the results obtained with the pump-and-probe excitation were previously given [Barbini et al. 1999, 2001b].

4.2. Lidar Calibration of Satellite Imagery

The first ocean color radiometer was CZCS [Hovis 1980], operated on the Nimbus-7 satellite from 1978 to 1986. Four new ocean color radiometers, SeaWiFS [Hooker et al. 1992], two MODIS sensors [Esaias et al. 1998] and MERIS [Huot et al. 2002], were launched in 1 August 1997, 18 December 1999, 4 May 2002 and 1 March 2002, respectively. The first, aboard OrbView-2, is a joint venture of Orbital Sciences Corporation and the National Aeronautics and Space Administration (NASA). The two MODIS sensors, integrated on Terra and Aqua, respectively, are part of the NASA's Earth Observing System (EOS), a program including six spacecrafts: Aqua, Aura, ICESat, Landsat 7, SORCE and Terra. MERIS is one of the instruments aboard ENVISAT, the more important Earth observing satellite of the European Space Agency (ESA).

The modern ocean color radiometers have been designed mainly for the observation of the phytoplankton biochemistry: in particular, they are expected to provide surface chl-a concentration over the range 0.05 to 50 mgChl m^{-3}. The main characteristics of SeaWiFS, MODIS and MERIS are summarized in table 3.

Table 3. Main characteristics of SeaWiFS, MODIS and MERIS radiometers.

Radiometer	SeaWiFS	MODIS	MERIS
Scan Width	±58°.3[a], ±45°.0[b]	±55°	±34°.25
Scan Coverage	2,800 km[a], 1,500 km[b]	2,330 km	1,150 km
Nadir Resolution	1.13 km[a], 4.5 km[b]	0.25 km[c], 0.5 km[d], 1 km[e]	0.3 km[f], 1.2 km[g]
Bands	8 (412 – 865 nm)	36 (405 – 14,385 nm)	15 (412.5 – 900 nm)
Scan Period	0.167 s[h]	0.903 s[h]	0.044 s[i]
Tilt	20°, 0°, +20°	No	No
Digitization	10 bits	12 bits	12 bits
Acquisition start	4 September 1997	24 February 2000[j]	22 March 2002

[a] Local area coverage.
[b] Global area coverage.
[c] Bands 1 – 2
[d] Bands 3 – 7
[e] Bands 8 – 36
[f] Full resolution.
[g] Reduced resolution.
[h] From the period of the continuously rotating scan mirror.
[i] From the period of the CCD frame.
[j] For MODIS aboard Terra; 24 June 2002 for MODIS aboard Aqua.

4.2.1. Chlorophyll-a

Data processing for chl-a retrieval consists mainly of two steps: atmospheric corrections [Gordon and Wang 1994, Fiorani et al. 1998] and bio-optical algorithms [O'Reilly et al. 1998, Sathyendranath et al. 2001]. Atmospheric corrections are necessary to obtain the water-leaving radiance, removing from the sensor measurements the contributions of air molecules and aerosols, which typically represent about 80% of the total in the visible bands. Usually, they are based on the experimental observation that the water-leaving radiance is negligible for the longer wavelengths (red and near infrared). In these spectral regions the contributions

of air molecules and aerosols can be directly obtained from the sensor measurements. Atmospheric corrections for the longer wavelengths are then combined with the predictions of models and applied to visible bands. Bio-optical algorithms consist of some semi-empirical equations used to calculate chl-a concentrations from the water-leaving radiances in the visible bands. Sunlight is not merely reflected by the water surface, but – after entering the ocean – it is selectively absorbed and scattered by phytoplankton and then backscattered through the water surface. This approach permits quantitative estimates of chl-a concentrations within the upper tens of meters of the open ocean and for minor depths in coastal zones. Due to large differences in absorption, data processing must be calibrated through in situ measurements.

The temporal and spatial resolutions of ELF, from one hand, and of SeaWiFS, MODIS and MERIS, from the other one, are very different: while a laser pulse is emitted every 0.1 s and the beam footprint on the water surface has a diameter of about 0.1 m, a satellite swath is acquired in some minutes and a radiometer pixel sizes around 1 km. The temporal and spatial resolution of the space borne sensors degrades further passing from raw to processed imagery. Conversely, as far as depth is concerned, lidar and radiometers probe a similar layer of the order of magnitude of 10 m (depending on water turbidity).

To give an example of lidar calibration of satellite imagery, SeaWiFS Level 3 (L3) daily Standard Mapped Image (SMI) products gathered during the New Zealand – Italy transect of MIPOT will be considered here [Barbini et al. 2004]. Only non questionable pixels will be retained. L3 daily SMI products were chosen for the following reasons:

- L3 is more accurate than Level 2 (L2): from one hand, each L3 measurement is the result of the spatial and temporal average of some L2 data [Campbell et al. 1995], from the other one, problems missed during the review of L2 products can be identified in the L3 fields [McClain et al. 1992]; the final accuracy of the chl-a retrieval by SeaWiFS is of the order of 50% [O'Reilly et al. 2000];
- 1 day is the finer temporal resolution for L3 files;

To sum up, the choice of L3 daily SMI products led to a temporal resolution of 1 day and a geometrical resolutions of 5'16" × 5'16" (corresponding to a spatial resolution of about 9 km × 9 km). As a consequence, some processing was necessary to compare ELF and SeaWiFS data. At first, all the ELF measurements falling in a SeaWiFS pixel were averaged, thus representing a nearly straight track (length: ~ 10 km, width: ~ 0.1 m) acquired in about 20 minutes: in the following, we will name this track ELF granule. Eventually, the measurements carried out by lidar and SeaWiFS were considered concurrent if the in situ track was acquired in the same day as the remote sensed pixel. Unfortunately, the resolution of SeaWiFS and ELF differs:

- in time, since the acquisition interval of ELF is short compared to 1 day,
- in space, because the square of about 9 km × 9 km observed by SeaWiFS is larger than the track of about 10 km × 0.1 m sensed by ELF, even if the track is inscribed in the square.

In order to avoid those differences, the ship should have spanned a 9 km × 9 km area for 1 day. This solution has been discarded: it would have allowed only one concurrent measurement per day and would have led to unaffordable durations of the oceanographic campaign. That dissimilarity in temporal and spatial resolutions can be responsible for part of the disagreement between SeaWiFS and ELF.

In many occasions, ELF was able to carry out measurements when SeaWiFS was not: during the entire cruise, 9086 ELF granules were collected and only 784 of them were concurrent to at least one SeaWiFS pixel. This partial failure to obtain results is due to the cloud coverage lasting for many days, especially in the Indian Ocean. Moreover, the results of MIPOT [Barbini et al. 2004] suggests that ELF performed better in the detection of local anomalies, as one can expect from the higher spatial resolution of the ELF granule with respect to the SeaWiFS pixel. Anyway, although the values found by SeaWiFS are in general higher than those measured by ELF, the agreement among the curves corresponding to the two sensors (Figure 10) is rather good.

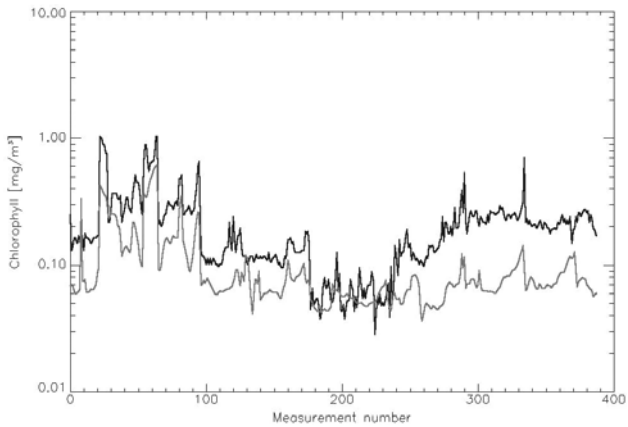


Figure 10. Simultaneous measurements of chl-a by ELF (grey) and SeaWiFS (black) during the New Zealand – Italy transect of MIPOT.

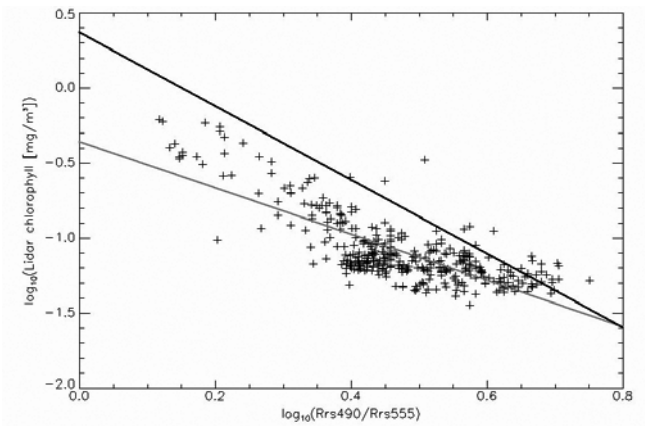


Figure 11. ELF (plus sign) chl-a concentration versus the SeaWiFS 490-555 band ratio during the New Zealand – Italy transect of MIPOT. The grey and black lines correspond to ELF-calibrated OC1 and standard OC1 bio-optical algorithms, respectively.

Examples of ELF calibration of the SeaWiFS bio-optical algorithm have been already published elsewhere [Barbini et al. 2001a, 2003a, 2004]: here we will only recall its main features. Although the standard SeaWiFS bio-optical algorithm (OC4) is based on a modified cubic polynomial applied to 4 bands we decided to use a linear approach applied to 2 bands (OC1) [O'Reilly et al. 1998]: in fact, Figure 11 shows a large dispersion of the ELF measurements and, as in earlier works [Mitchell and Holm-Hansen 1991, Mitchell 1992], it did not seem reasonable to use higher order polynomials and/or more bands, and therefore the ELF-calibrated bio-optical algorithm has been based on the linear fit of the log-log plot of the lidar concentration versus the SeaWiFS 490-555 band ratio (see below). Higher order polynomials would have led to an unstable fit, more bands would have introduced an unjustified complication. In our case, standard OC1 and standard OC4 are very close [Barbini et al. 2004] and thus the departure of ELF-calibrated OC1 from standard OC1 gives a good estimate of the discrepancy between ELF-calibrated and standard chl-a concentrations.

OC1 is expressed by [O'Reilly et al. 1998]:

$$C_{surf} = 10^{a_0 + a_1 R}, \tag{15}$$

where C_{surf} is the surface chl-a concentration in mgChl m^{-3} and:

$$R = \log_{10} \frac{Rrs(490\,\mathrm{nm})}{Rrs(555\,\mathrm{nm})}. \tag{16}$$

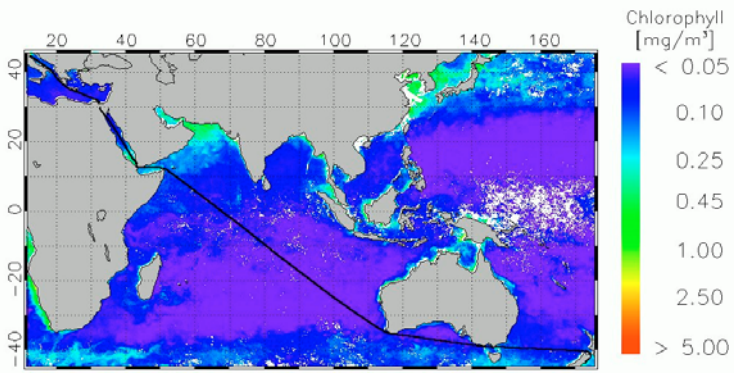
Rrs(490 nm) and Rrs(555 nm) are the remote sensing reflectance at 490 and 555 nm, respectively. Rrs(490 nm)/Rrs(555 nm) is called the 490-555 band ratio.

The parameters of the ELF-calibrated OC1 and standard OC1 are summarized in table 4.

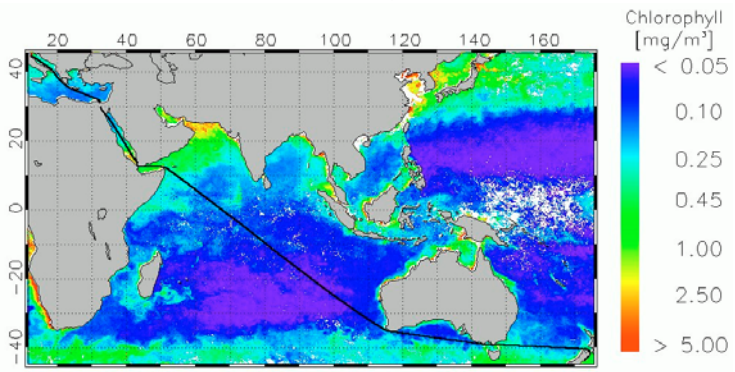
Table 4. Parameters of the bio-optical algorithms considered in this study (the standard deviations of the parameters of standard OC1 are not given in the literature).

Algorithm	**a_0**	**a_1**
Standard OC1	0.3734	-2.4529
ELF-calibrated OC1	-0.086 ± 0.025 (29%)	-2.076 ± 0.055 (3%)

Let us note that the difference between ELF-calibrated OC1 and standard OC1 is larger than the error bar (the standard deviations of the parameters of standard OC1 are not given in the literature). ELF-calibrated OC1 provides lower concentrations, thus correcting the tendency of SeaWiFS to overestimate chl-a. Eventually, it is possible to use ELF-calibrated OC1 to draw a map of concentration (Figure 12a) and to compare it to the standard SeaWiFS imagery (Figure 12b). Both results are based on the average of the satellite daily data gathered during the New Zealand – Italy transect of MIPOT [Barbini et al. 2004].



a)



b)

Figure 12. a) ELF-calibrated and b) standard SeaWiFS chl-a concentration during the New Zealand – Italy transect of MIPOT. Continuous line: ship track.

4.2.2. Primary Productivity

PP is the water parameter more related to the carbon cycle, probably the more important global phenomenon affecting the Earth climate. The Southern Ocean is one of the least studied but most interesting zones of our planet from the biological point of view, because of a series of circumpolar fronts that result in the definition of different oceanographic provinces, where the development of endemic phytoplankton is favored [Boyd 2002]. For this reason, we used the ELF measurements carried out during the expeditions in Antarctica to provide new estimates of PP of the Southern Ocean [Barbini et al. 2003b]. In their more recent form [Barbini et al. 2005], they are based on a Vertically Generalized Production Model (VGPM) [Behrenfeld and Falkowski 1997a, 1997b] tuned in Antarctica [Smith et al. 2000, Dierssen et al. 2000] and on the SeaWiFS chl-a calculated with ELF-calibrated algorithm. A discussion of the application of standard models in Antarctica and more details on the new model can be found elsewhere [Barbini et al. 2005]: here a short description and the main results of the new model will be given.

The new model is defined by the following relationships:

$$PP_{eu} = P^{B}_{opt}\ F\ Z_{eu}\ C_{surf}\ D\,, \tag{17}$$

$$P^{B}_{opt} = 1.09 \quad , \tag{18}$$

$$F = \frac{PAR}{PAR + 11.77} \quad , \tag{19}$$

$$Z_{eu} = 48.8\, C_{surf}^{\ -0.36} \quad , \tag{20}$$

where PP_{eu} is the daily PP within the euphotic zone per unit of surface [mgC m^{-2} d^{-1}], P^{B}_{opt} is the maximum chlorophyll-specific carbon fixation rate [mgC $mgChl^{-1}$ h^{-1}], observed within a water column and measured under conditions of variable irradiance during incubations typically spanning several hours, F is the relative fraction of potential photosynthesis lost within the euphotic zone due to light limitation [unitless], to first order equal to the average production of the water column divided by P^{B}_{opt}, Z_{eu} is the euphotic zone [m], i.e. the penetration depth of 1% surface irradiance and D is the daylength [h], also called photoperiod. PAR is measured by ELF and D is calculated with known formulas [Sellers 1965].

An example of the results of new and standard PP models is given in Figure 13, based on the ELF measurements carried out during the 16th Italian expedition in Antarctica (January 5th 2001 – February 26th 2001) [Barbini et al. 2005]. The present study indicates that usual PP models applied to standard chl-a concentrations can underestimate PP up to 50%.

Recently, ELF measurements have been used to calibrate the new PP model in some specific regions of the Ross Sea. The hydrographic regions of Terra Nova Bay (TNB), Cape Adare zone (CA) and center of the Ross Gyre (RG) are among the most interesting of the Ross Sea (Budillon et al. 2003):

- the TNB polynya is the area where the high salinity shelf water (HSSW) is generated by formation and removal of ice; it is strongly affected by ice melting and nutrient release from the Drygalsky Glacier (case II waters);
- the northward branch of HSSW mixes with the modified circumpolar deep water (MCDW) near CA and escapes in the northern continental shelf break;
- in RG nutrients are released into the upper ocean by water mixing processes, favoring the onset of intense phytoplankton blooms; RG is located in the Joides Basin, near the continental shelf break (case I waters).

The zones under study are defined as follows:

- TNB: -74.5 S – -75.25 S, 163 W – 166 W (0°.75 × 3°),
- CA: -71.5 S – -73 S, 170 W – 175 W (1°.5 × 5°),
- RG: -73.5 S – -74.5 S, 173 W – 177 W (1° × 4°).

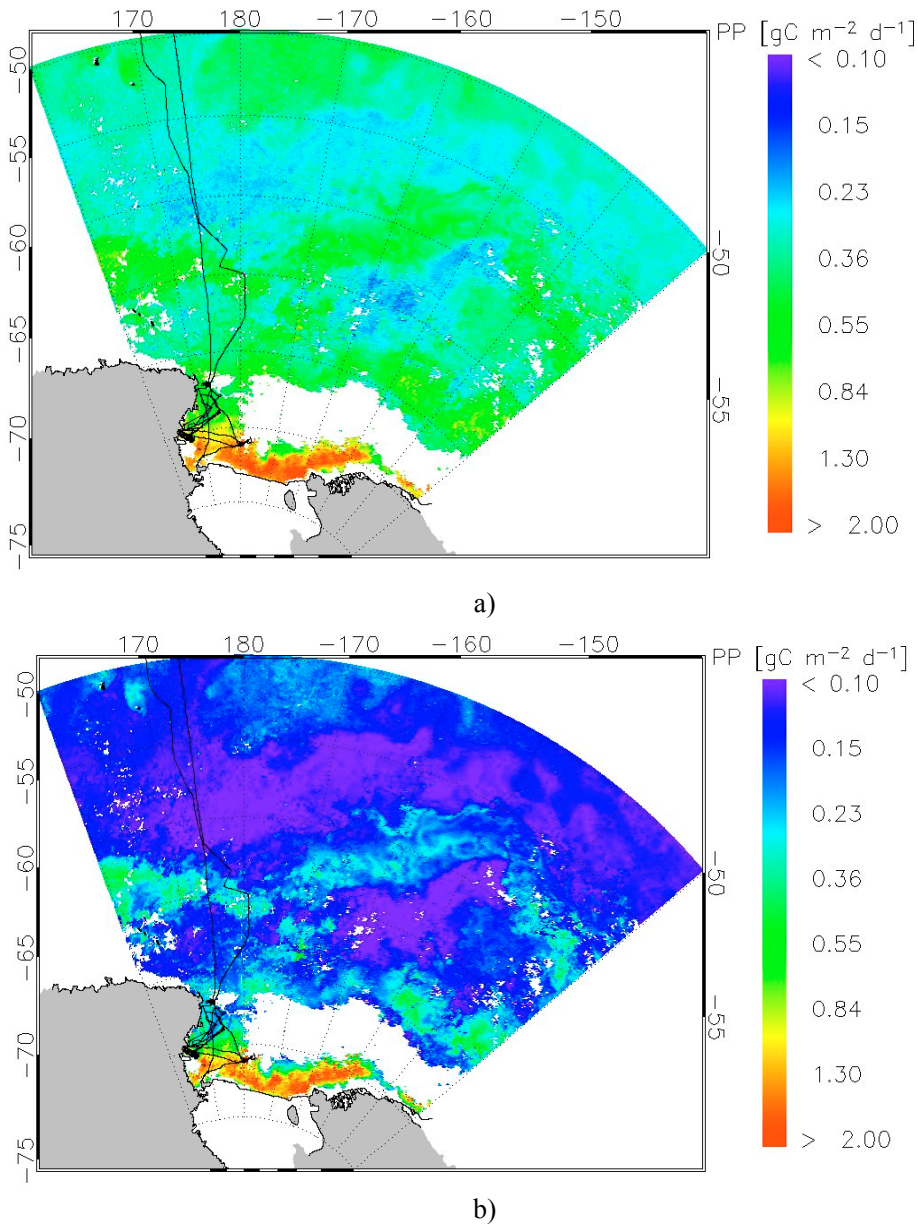


Figure 13. Average PP, based on the monthly products of January and February 2001, calculated with a) the ELF-calibrated SeaWiFS chl-a bio-optical algorithm and the new PP model, and b) the standard SeaWiFS chl-a bio-optical algorithm and the VPGM model by Behrenfeld and Falkowski. Continuous line: ship track.

Another interesting hydrographic region is the coastal belt from the Ross Island along the Ross Ice Shelf (RI). It has not been included because the ELF-calibration does not significantly improve the SeaWiFS chl-a bio-optical algorithm. Nevertheless, that large polynya area is a powerful supply for the Ross Sea PP.

The SeaWiFS chl-a bio-optical algorithm has been calibrated in TNB, CA and RG with the ELF measurements of the 16th Italian expedition in Antarctica. The procedure differs from

that already described only for the calculation of the linear fit. The details of the statistical treatment can be found elsewhere [Fantoni et al. 2005].

The results of the ELF calibration of the SeaWiFS chl-a bio-optical algorithm for TNB, CA and RG, as well as those obtained for the Ross Sea Region (RSR), i.e. the area delimited by the coast and a line (straight in the cylindrical equidistant projection) from a point near Cape Adare (72 S, 170 E) to a point near Cape Colbeck (76 S, 158 W), thus containing the highly productive RI, are resumed in table 5 and shown in Figure 14.

Table 5. Results of the ELF calibration of the SeaWiFS chl-a bio-optical algorithm in TNB, CA, RG and RSR.

Zone	a_0	a_1
TNB	0.09	-3.1
CA	0.56	-2.3
RG	0.78	-2.7
RSR	0.37	-1.4

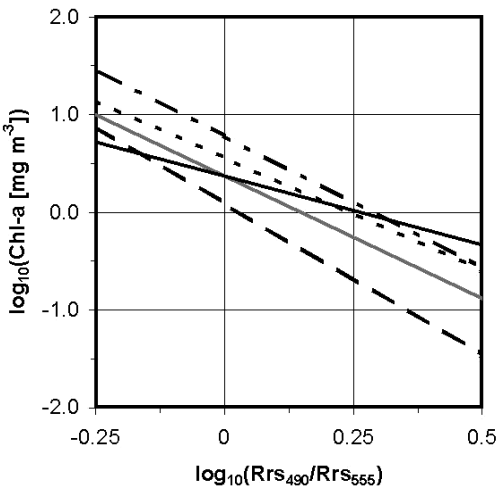


Figure 14. Comparison among different SeaWiFS regional chl-a bio-optical algorithms: standard OC1 (gray), ELF-calibrated OC1 during the 16th Italian expedition in Antarctica in RSR (continuous line), TNB (dashes line), CA (dotted line) and RG (dashed-dotted line).

Although RSR does not coincide exactly with the Ross Sea, its simple definition was a necessary compromise: the calculation of the bio-optical properties in RSR, rather than in the Ross Sea, saves computing time because the satellite products are usually delivered in the cylindrical equidistant projection and it is fast to determine whether a pixel is above or below a straight line.

In the measurement range of the 16th Italian expedition in Antarctica ($-0.25 < \log_{10}[\text{Rrs}(490\text{ nm})/\text{Rrs}(555\text{ nm})] < 0.5$) our results indicate that:

- in TNB, standard OC1 overestimate chl-a, especially at low concentrations (up to 75%),

- in CA, standard OC1 underestimate chl-a, especially at low concentrations (up to 70%),
- in RG, standard OC1 underestimate chl-a, especially at high concentrations (up to 50%),
- in RSR, standard OC1 overestimate high concentrations and underestimate low concentrations.

As expected:

- the algorithm calibrated in CA and RG (mostly open sea and only open sea, respectively) is the closest to OC1 (calibrated typically with measurement stations in the open sea),
- the algorithm calibrated in RSR, a sort of average algorithm of the Ross Sea, lies almost in the midst of those calibrated in TNB, CA and RG.

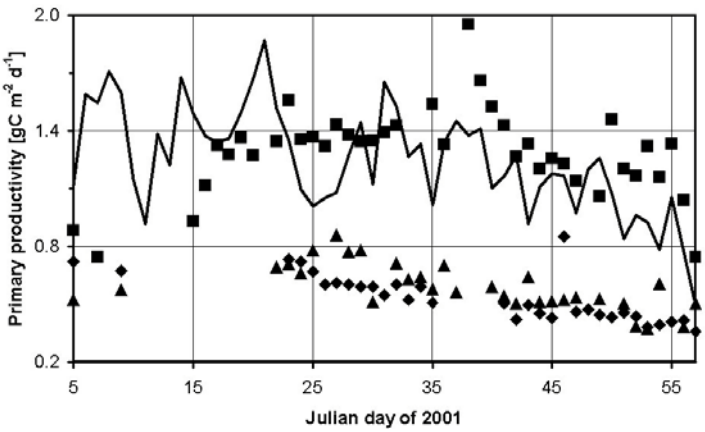


Figure 15. PP calculated with the ELF-calibrated SeaWiFS regional chl-a bio-optical algorithm and the new PP model in RSR (line), TNB (squares), CA (diamonds) and RG (triangles) during the 16th Italian expedition in Antarctica.

Eventually, PP in TNB, CA and RG has been calculated with the new model described above. The results are shown in Figure 15. In the first part of the oceanographic campaign, few radiometer data are available on TNB, CA and RG. The available measurements show a low productivity. From the 10th day, in TNB, and from the 17th and 18th day, in RG and CA, respectively, more radiometer data are available. While PP of CA and RG is about one half of that of RSR, TNB is characterized by a high PP, showing values larger than in RSR (note the peak reaching nearly 2 gC m^{-2} d^{-1} in the 38th day). Moreover, PP decreases more slowly in TNB than in RSR.

4.2.3 Chromophoric Dissolved Organic Matter

Lately, ELF measurements have been used to develop a bio-optical algorithm for the retrieval of CDOM concentration from SeaWiFS data. The interest in oceanic CDOM grew in the last years because some studies indicate that it is correlated with climate-driving factors, through a feedback process controlled by phytoplankton. Unfortunately, CDOM concentration is not

among the data products provided by ocean color satellite radiometers. This explains why a new satellite algorithm for CDOM concentration retrieval has been suggested.

The plausibility of a process that links oceanic dymethil sulfide (DMS) with climate has been discussed years ago [Carlson et al. 1987]. A recent research [Toole and Siegel 2004] demonstrated the connection between ultraviolet radiation fluxes in the open ocean and feedbacks in the global climate system, driven by the DMS emitted by phytoplankton cells. It would desirable to have information on the global distribution and dynamics of DMS but, unfortunately, the present ocean color satellite radiometers do not provide surface layer DMS concentrations. DMS concentration, however, is significantly correlated with the absorption coefficient for CDOM [Toole et al. 2002] and some algorithms have been suggested for the calculation of the CDOM absorption coefficient from ocean color data products [Aicken et al. 1995, Hoge et al. 1999, Barbini et al. 2002, Siegel et al. 2002, Johannessen et al. 2003]. Nevertheless, the debate on the satellite retrieval of CDOM is still open: e.g. a standard CDOM algorithm is not available for SeaWiFS, even if the SeaWiFS Project Office recognized its usefulness [McClain et al. 2004].

In this framework, we present an ELF-calibrated CDOM algorithm for satellite products. It is based on the data collected during the 18th Italian expedition in Antarctica, carried out from 5 January to 4 March 2003 in the Ross Sea, with particular emphasis on coastal zones and polynya areas. The approach of this study is similar to that followed for the ELF-calibrated chl-a algorithm (see above), but require the retrieval of the CDOM absorption coefficient from fluorescence measurements [Hoge et al. 1995].

Although the present results are preliminary, they provide valuable information on distribution and dynamics of CDOM in the Southern Ocean. The CDOM data are useful both for understanding the DMS distribution and dynamics, and the biogeochemical factors affecting PP. Finally, the information gathered on DMS and PP will improve our comprehension of the world climate.

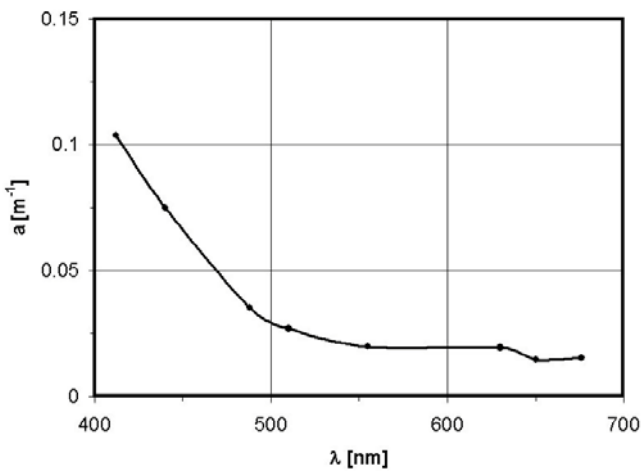


Figure 16. Average absorption of filtered seawater versus wavelength during the 18th Italian expedition in Antarctica. The points are the actual measurements, the line is the best fit.

As explained above, the CDOM fluorescence signal (f_{CDOM}) is released at first in Raman units. Afterwards, it is converted into CDOM absorption (a_{CDOM}) by calibration with

measurements carried out with the absorption and attenuation meter WET Labs ac-9 [Van Zee and Walsh 2004]. ELF and ac-9 operated simultaneously in some stations of the oceanographic campaign. Ac-9 determined spectral attenuation and spectral absorption of water over eight wavelengths: 412, 440, 488, 510, 555, 630, 650, and 676 nm. At 412 and 440 nm, a (absorption of filtered seawater) corresponds to a_{CDOM} because it is dominated by CDOM for $\lambda < 500$ nm [Aiken et al. 1995, Kahru and Mitchell 2001]. The average of a during the 18th Italian expedition in Antarctica is given in Figure 16.

As one can expect, the best correlation between f_{CDOM} and a_{CDOM} has been found at 440 nm: f_{CDOM} can traverse some meters of water before detection and, as a consequence, it is linked to absorption at 450 nm that is closer to absorption at 440 nm than to absorption at 412 nm. The final results of the ELF calibration are given in Figure 17 and table 6. The details of the statistical treatment can be found elsewhere [Fiorani et al. 2005].

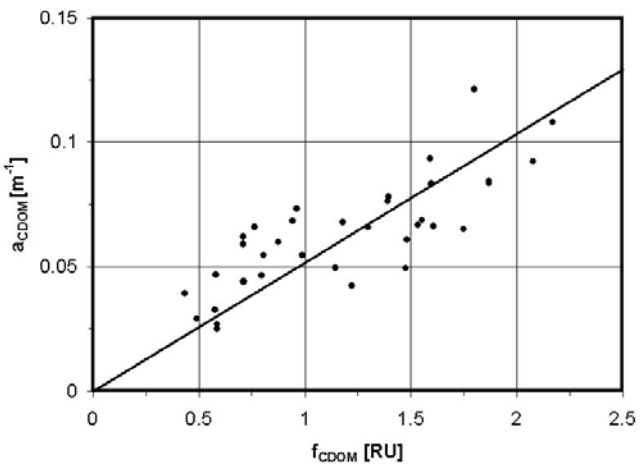


Figure 17. Calibration of ELF by ac-9.

Table 6. Results of the calibration of ELF by ac-9 (a_0 is the intercept, a_1 is the slope, σ_0 is the intercept error, σ_1 is the slope error and R is the correlation coefficient).

# of points	a_0	σ_0	a_1	σ_1	R
35	0	0	0.0517	0.0030	0.74

ELF CDOM data, now expressed in m^{-1}, can therefore be used for the ELF-calibration of the SeaWiFS CDOM algorithm, exactly in the same way ELF chl-a data have been used for the ELF-calibration of the SeaWiFS chl-a algorithm, as already described above. The only difference is that normalized water leaving radiance (Lwn) at 443 and 510 nm is employed, instead of that at 490 and 555 nm. The rationale of choosing those wavelengths is that a_{CDOM} is strong at 443 nm and weak at wavelengths greater than 500 nm [Kahru and Mitchell 2001]. The use of Lwn(412 nm) has been avoided because of problems with atmospheric correction at short wavelengths [Siegel et al. 2000]. Also in this case, only the final results of the SeaWiFS calibration are given (Figure 18 and table 7) and the reader is referred elsewhere for details of the statistical treatment [Fiorani et al. 2005].

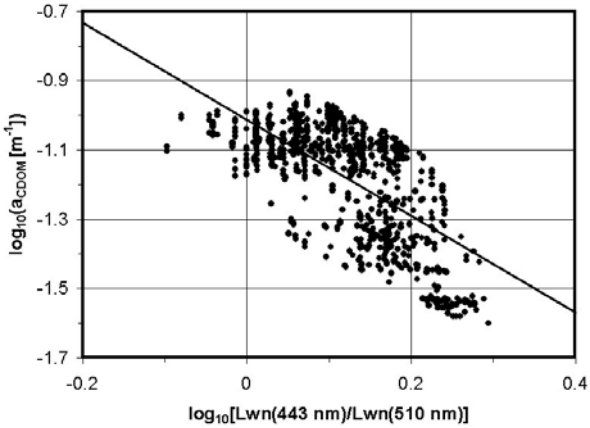


Figure 18. Calibration of SeaWiFS by ELF.

Table 7. Results of the calibration of SeaWiFS by ELF (a_0 is the intercept, a_1 is the slope, σ_0 is the intercept error, σ_1 is the slope error and R is the correlation coefficient).

# of points	a_0	σ_0	a_1	σ_1	R
899	-1.0115	0.0074	-1.393	0.054	0.69

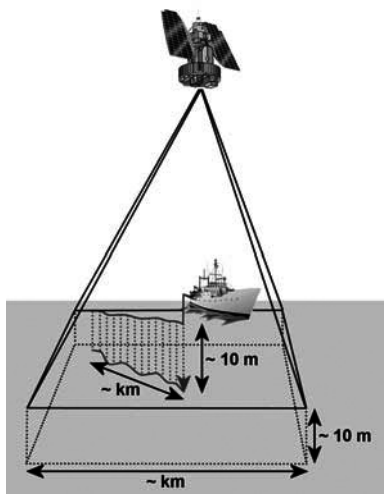


Figure 19. A SeaWiFS pixel covers a large area (size ~ 1 – 10 km) and integrates information from 0 to ~ 10 m. While an ac-9 sampling is a point, after averaging in a SeaWiFS pixel, ELF data represent a track (length ~ 1 – 10 km, width ~ 0.1 m) of water column measurements (from 0 to ~ 10 m).

In principle, it could have been possible to calibrate directly SeaWiFS by ac-9 but, in practice, the procedure followed here has two advantages:

- ELF data are more suitable to that purpose than ac-9 samplings because their geographic coverage and spatial resolution are closer to image extent and pixel size, respectively, of the SeaWiFS data products (Figure 19),

- the number of ELF measurements is much larger than that of ac-9 stations.

The results of the ELF-calibrated SeaWiFS algorithms for CDOM and chl-a, obtained during the 18th Italian expedition in Antarctica with L3 8-day SMI products are shown from Figure 20 to Figure 23, where the continuous line represents the ship track.

Missing satellite pixels are due to sea ice or to the cloud coverage on free waters. According to the observation in the visible band and in the microwave region by the National Oceanic and Atmospheric Administration (NOAA) satellites OLS (http://dmsp.ngdc.noaa.gov/html/sensors/doc_ols.html) and SSM/I (http://dmsp.ngdc.noaa.gov/html/sensors/doc_ssmi.html), respectively, the cloud coverage is the main obstacle to SeaWiFS imaging of the ocean at high latitudes, while sea ice dominates at low latitudes. By the way, it can be pointed out that ELF is not sensitive to clouds and carries out measurements also in polynyas. Phytoplankton blooms develop in ice free regions (near Terra Nova Bay and Ross Island) and in the Antarctic Divergence, an upwelling zone where nutrients are released. Austral summer 2002-2003 has been characterized by an exceptionally large ice coverage. This explains why only few pixels have been acquired in the Ross Sea from the half of February.

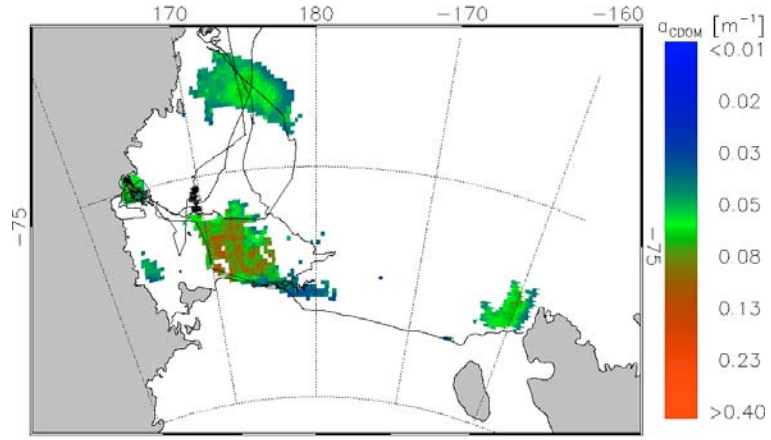
Two regions of interest have been defined in this study:

- the already mentioned RSR,
- the Ross Sea Sector (RSS) defined as the zone of Southern Ocean from the coast of Antarctica north to 50 S latitude in the 160 E – 130 W interval (the definition of RSS has been chosen in order to compare our study with literature data [Arrigo et al. 1998]).

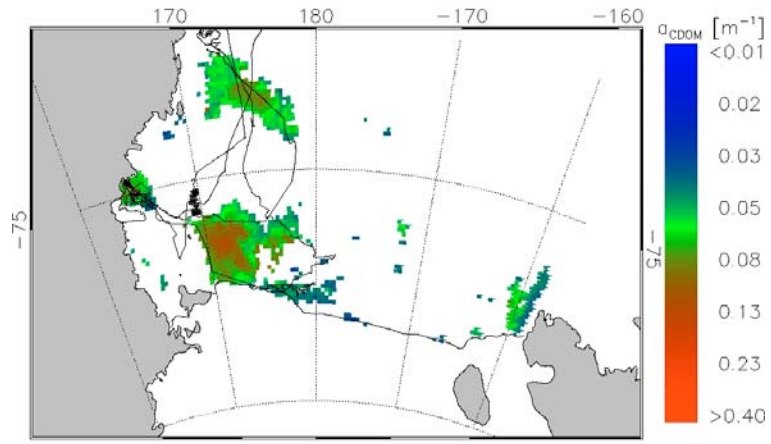
The relationship between CDOM and chl-a is noticeable both in RSR and RSS in the imagery shown from Figure 20 to Figure 23. If CDOM and chl-a are spatially averaged in RSR (Figure 24) and RSS (Figure 25) the correlation is evident (R is 0.59 in RSR and 0.97 in RSS). It is confirmed that RSR and RSS are quite productive in summer (1 – 2 mgChl m^{-3} in RSR and 0.5 – 1 mgChl m^{-3} in RSS). The average ratio between CDOM and chl-a is different in RSR (0.062 mg^{-1} m^2) and RSS (0.093 mg^{-1} m^2). This is probably due to the dissimilar taxonomic composition of RSR and RSS. On the other hand, a common feature can be observed, both in RSR and RSS: the ratio lowers when both CDOM and chl-a are high. This behavior could indicate two regimes: in the low ratio regime, CDOM could be dominated by the degradation product of phytoplankton metabolism, while in the high ratio regime, other sources could contribute to CDOM build up (e.g. ice melting near the Ross Ice Shelf).

In general, the values of a_{CDOM} found in this study (around 0.05 – 0.1 m^{-1}) compare quite well with literature data of the Ross Sea. In particular they are:

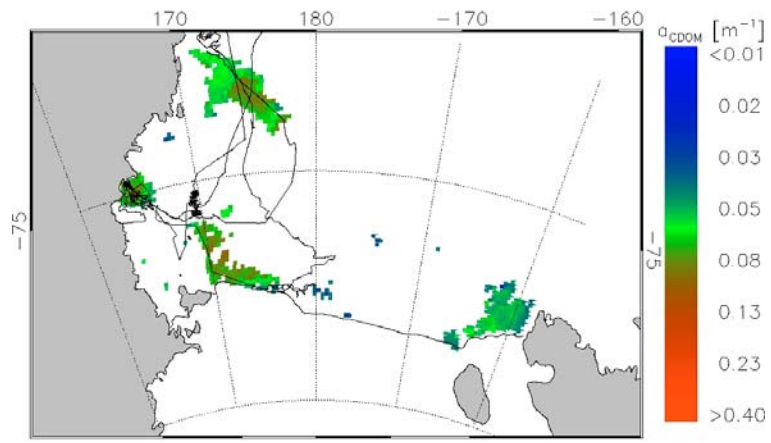
- higher than the measurements taken in a former oceanographic campaign carried out in 1997 [Mitchell 2003] (about 0.03 m^{-1}),
- comparable to the values of a semi-analytical algorithm applied to 1998 [Siegel et al. 2002] (about 0.1 m^{-1});
- lower than the results of an empirical algorithm relative to a 7-year period (1997 – 2004) [Fichot and Miller 2004] (around 0.2 m^{-1}).



a)

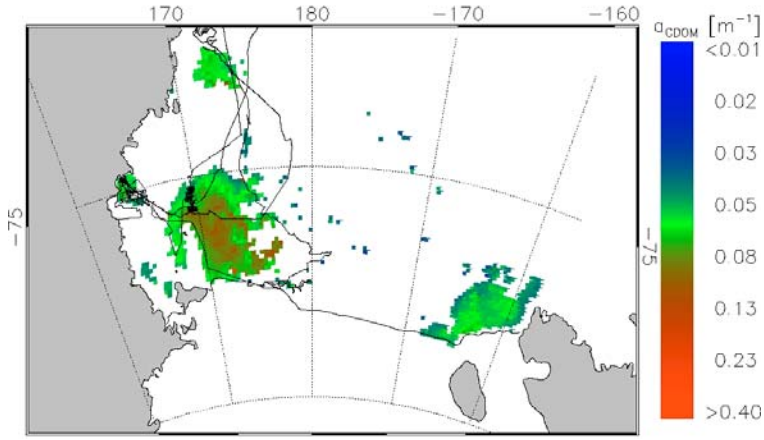


b)

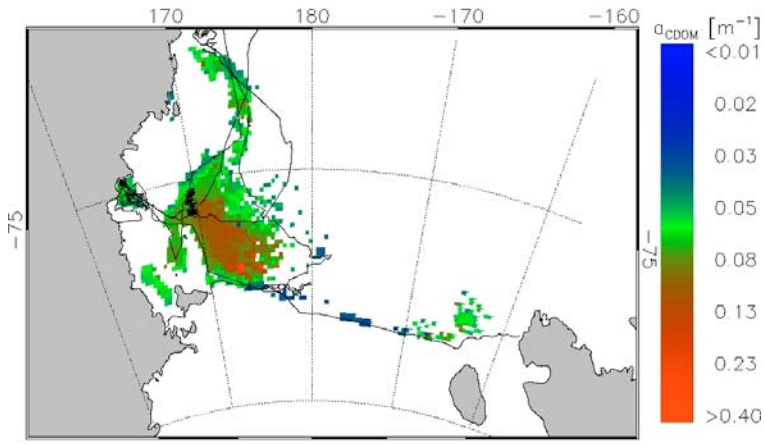


c)

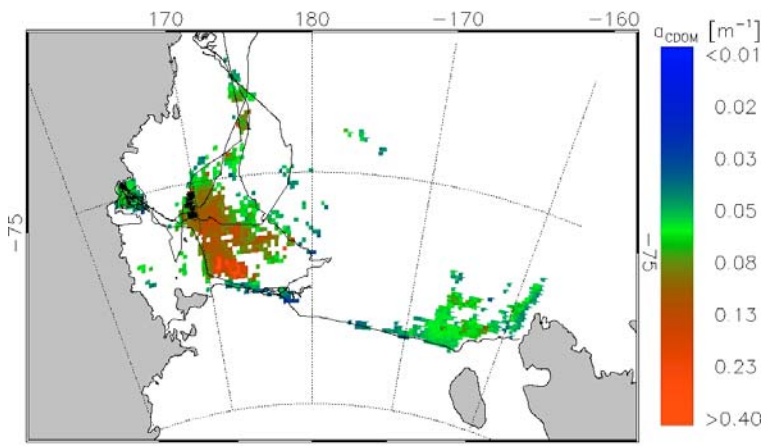
Figure 20 continued on next page



d)

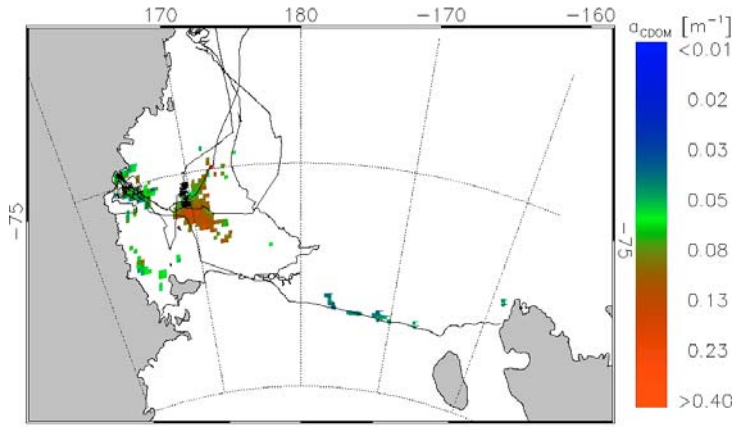


e)

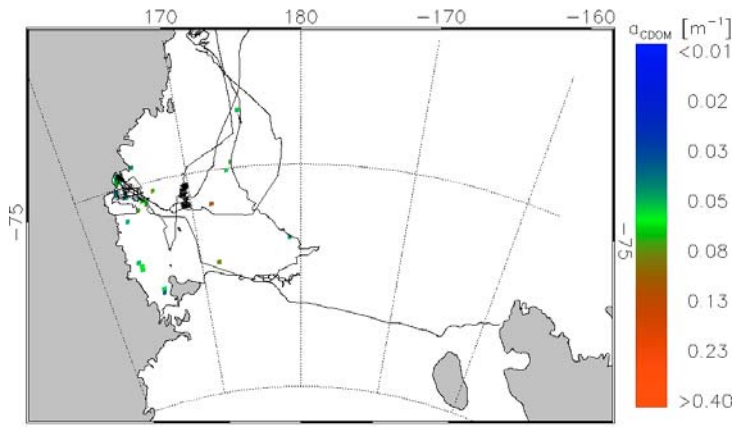


f)

Figure 20 continued on next page

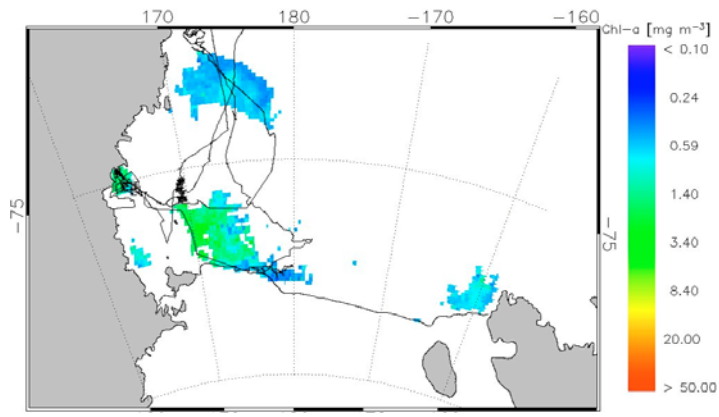


g)



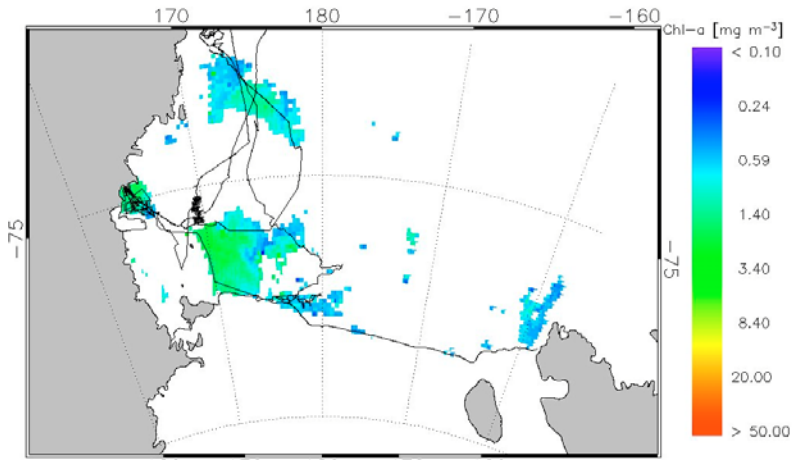
h)

Figure 20. CDOM measured with the ELF-calibrated SeaWiFS algorithm in RSR during the periods a) 01/01 – 08/01, b) 09/01 – 16/01, c) 17/01 – 24/01, d) 25/01 – 01/02, e) 02/02 – 09/02, f) 10/02 – 17/02, g) 18/02 – 25/02 and h) 26/02 – 05/03.

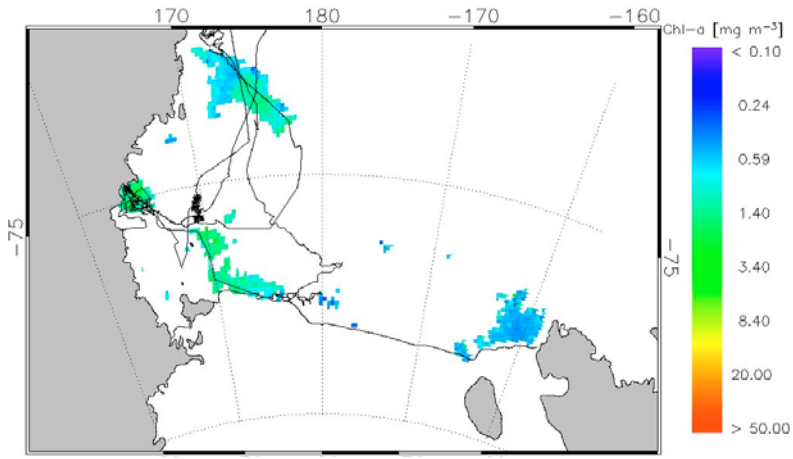


a)

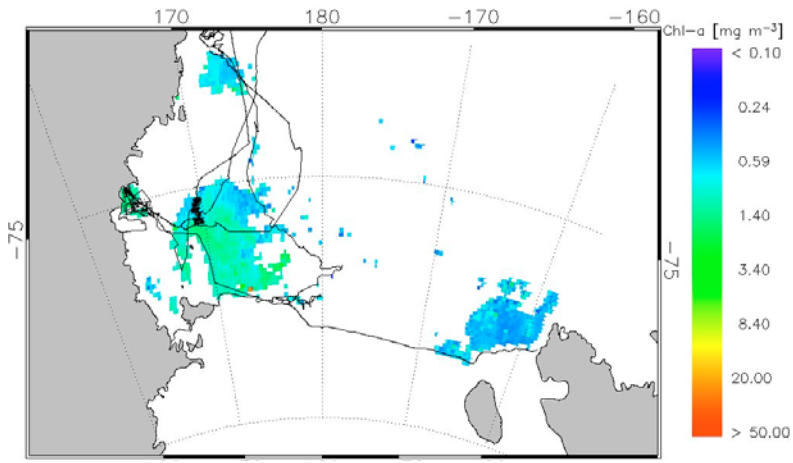
Figure 21 continued on next page



b)

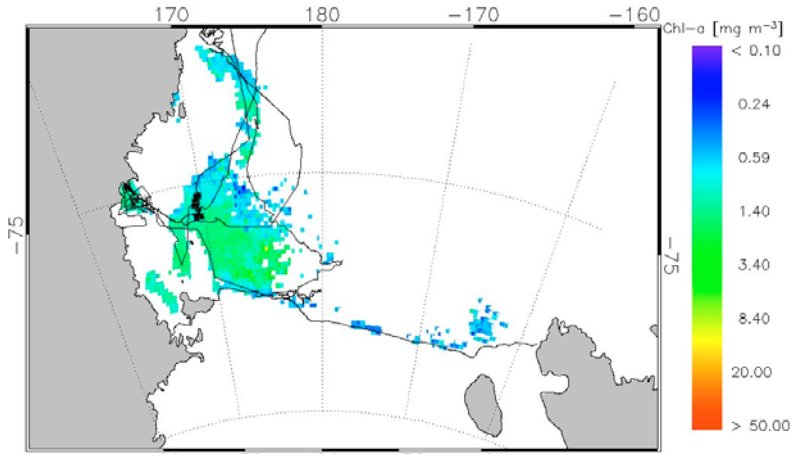


c)

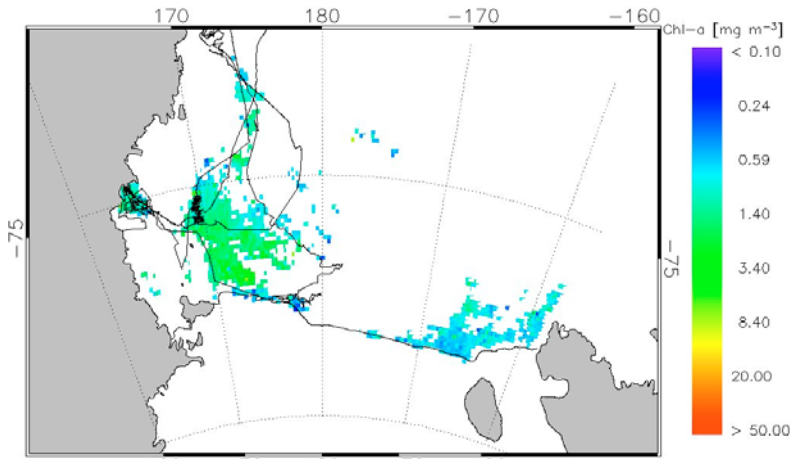


d)

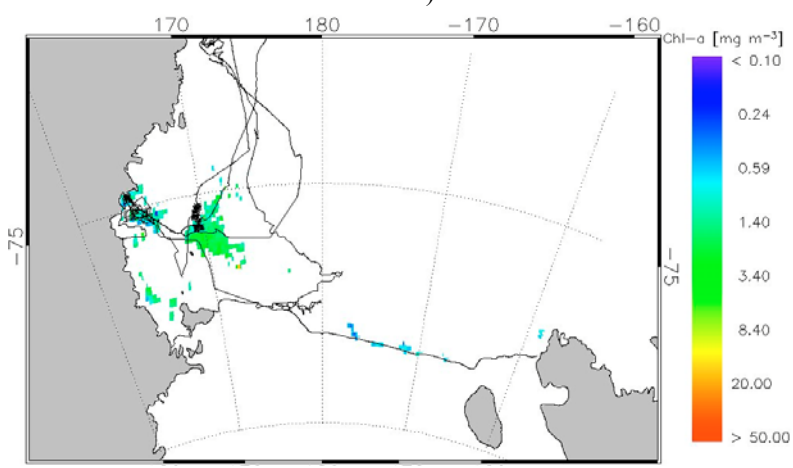
Figure 21 continued on next page



e)

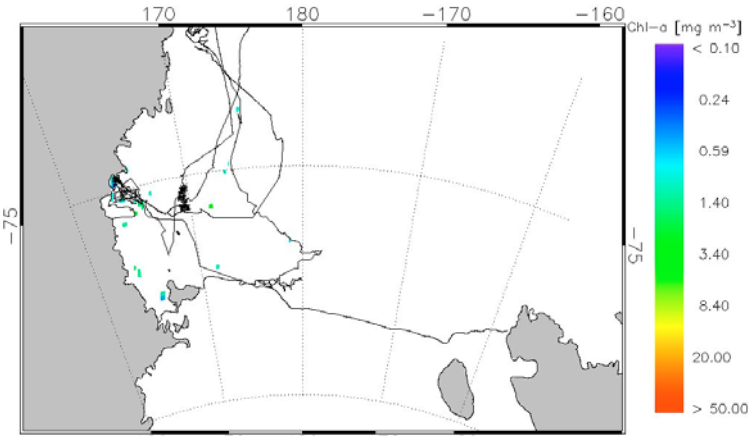


f)



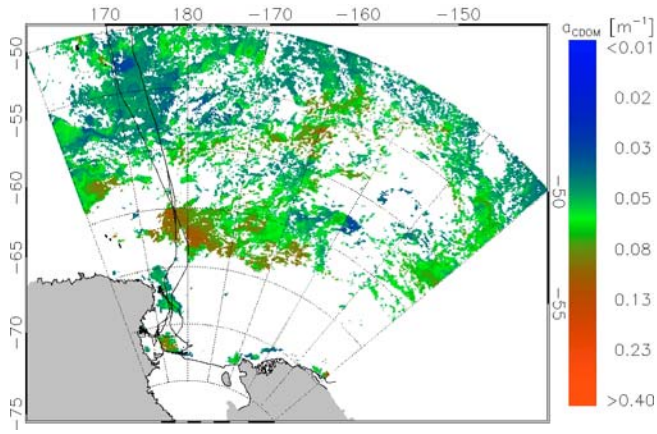
g)

Figure 21 continued on next page

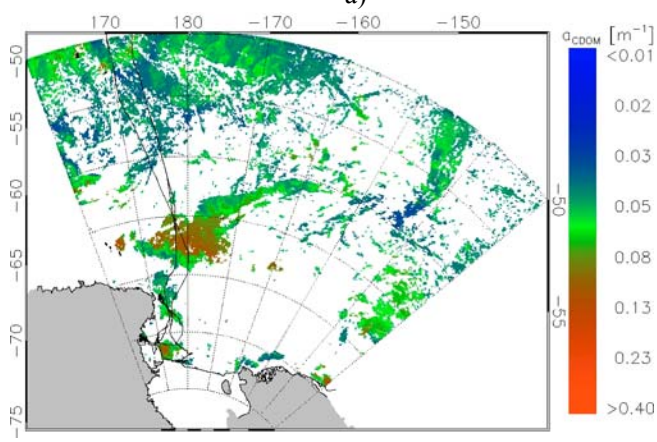


h)

Figure 21. Chl-a measured with the ELF-calibrated SeaWiFS algorithm in RSR during the periods a) 01/01 – 08/01, b) 09/01 – 16/01, c) 17/01 – 24/01, d) 25/01 – 01/02, e) 02/02 – 09/02, f) 10/02 – 17/02, g) 18/02 – 25/02 and h) 26/02 – 05/03.

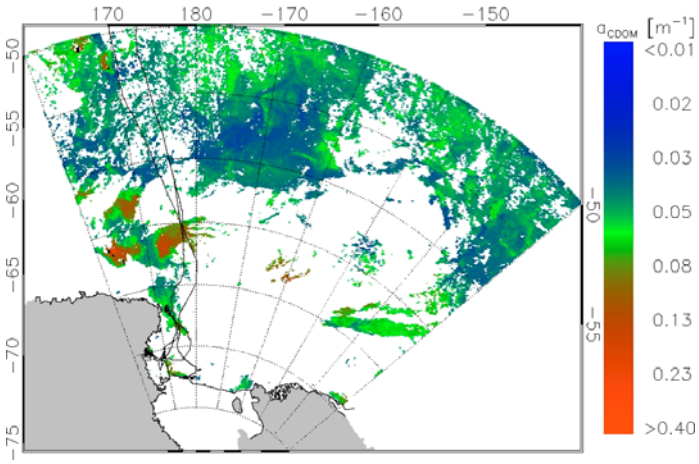


a)

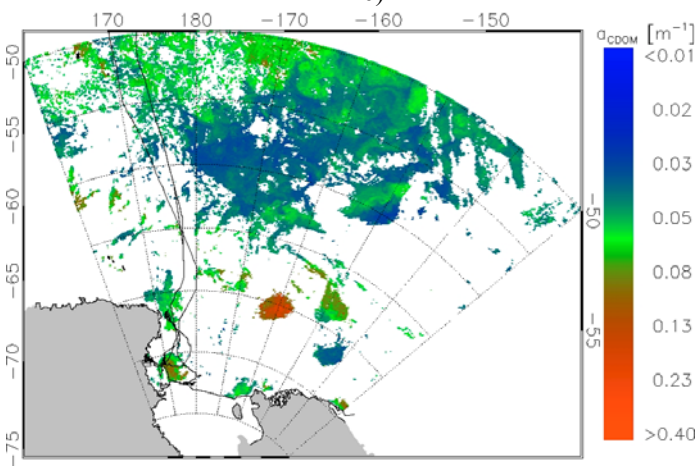


b)

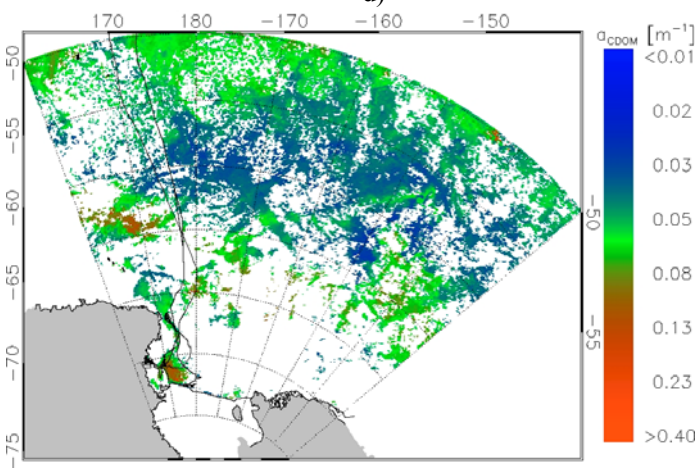
Figure 22 continued on next page



c)

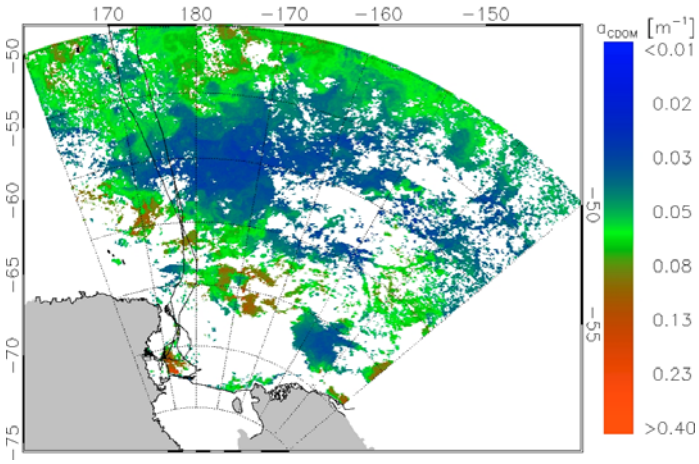


d)

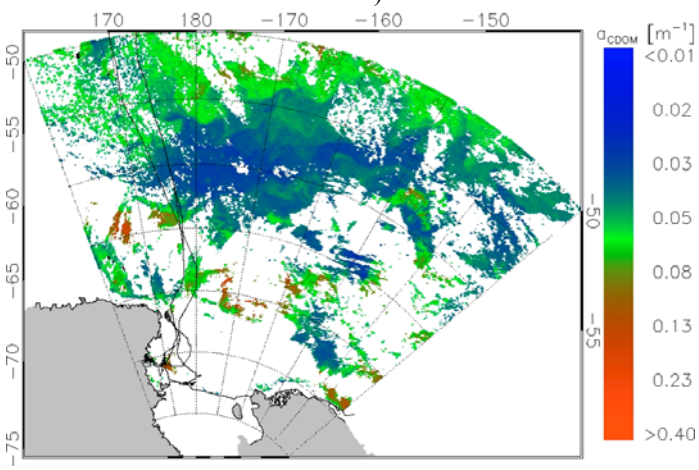


e)

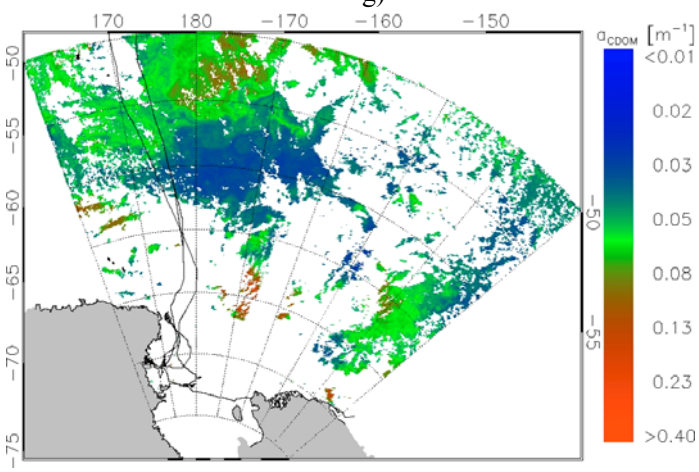
Figure 22 continued on next page



f)

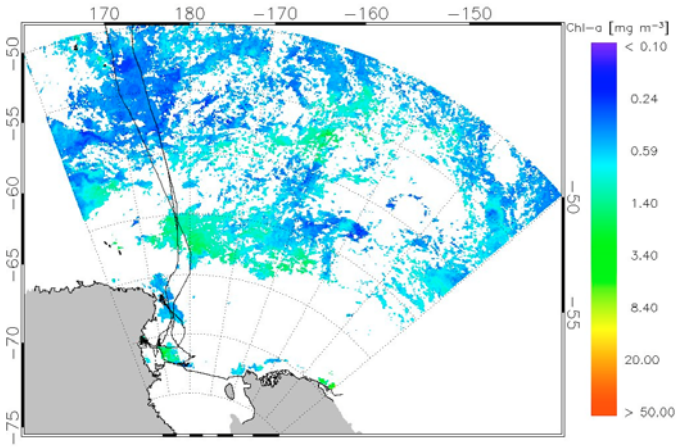


g)

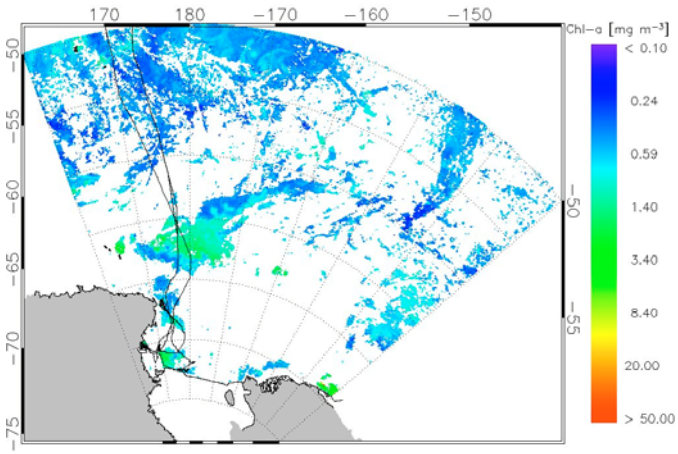


h)

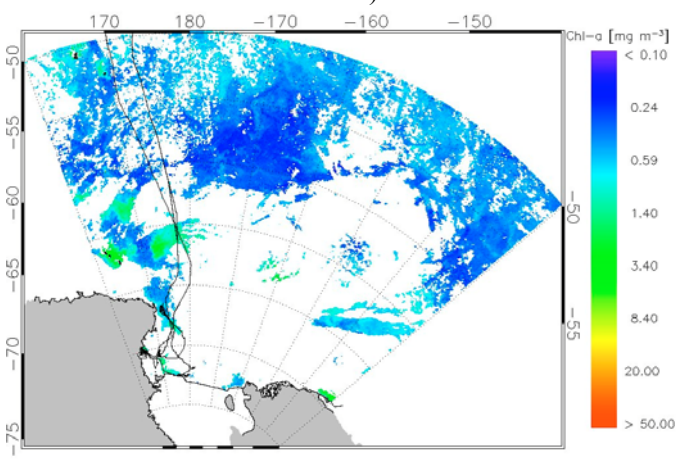
Figure 22. CDOM measured with the ELF-calibrated SeaWiFS algorithm in RSS during the periods a) 01/01 – 08/01, b) 09/01 – 16/01, c) 17/01 – 24/01, d) 25/01 – 01/02, e) 02/02 – 09/02, f) 10/02 – 17/02, g) 18/02 – 25/02 and h) 26/02 – 05/03.



a)

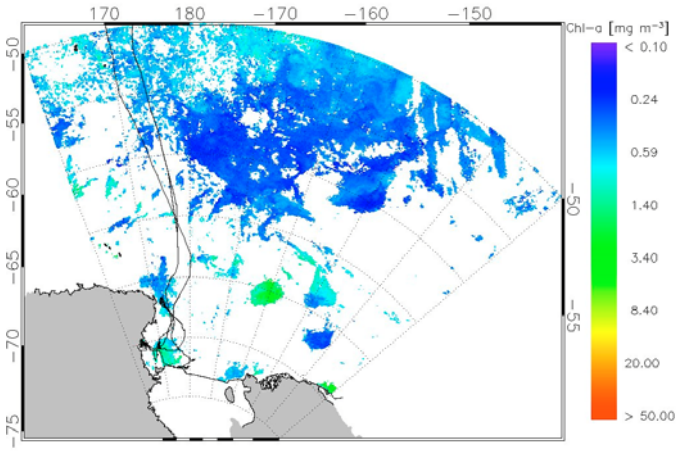


b)

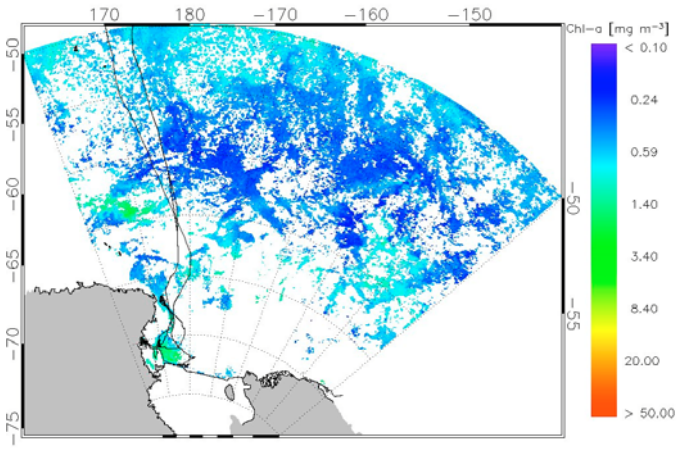


c)

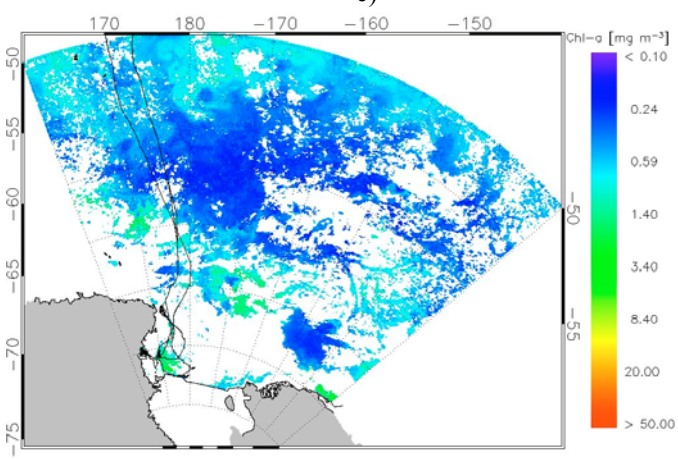
Figure 23 continued on next page



d)

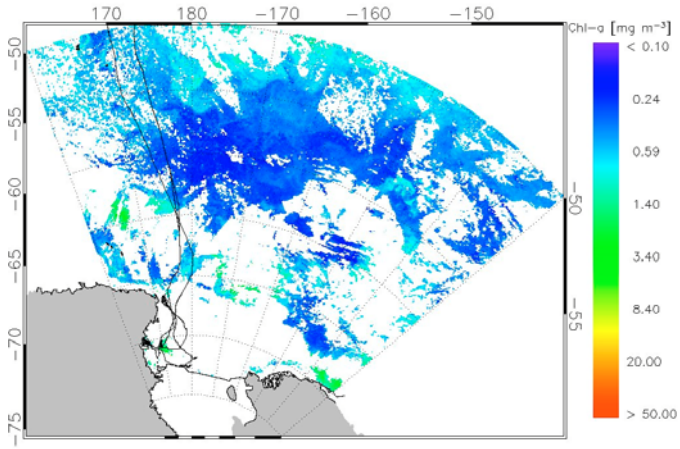


e)

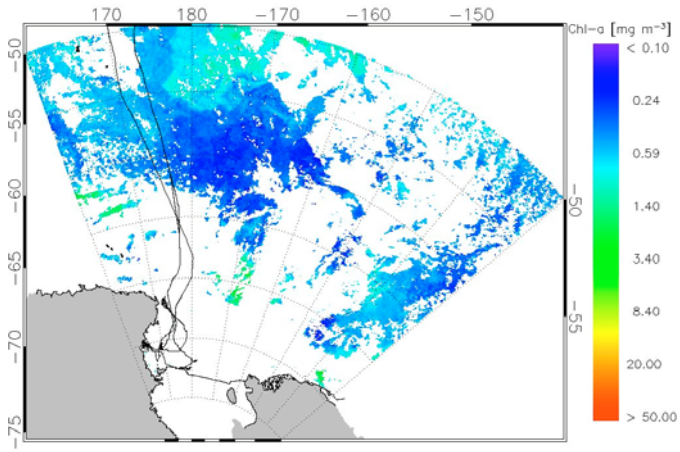


f)

Figure 23 continued on next page

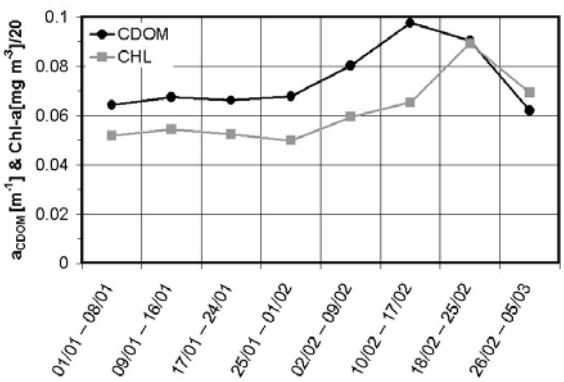


g)



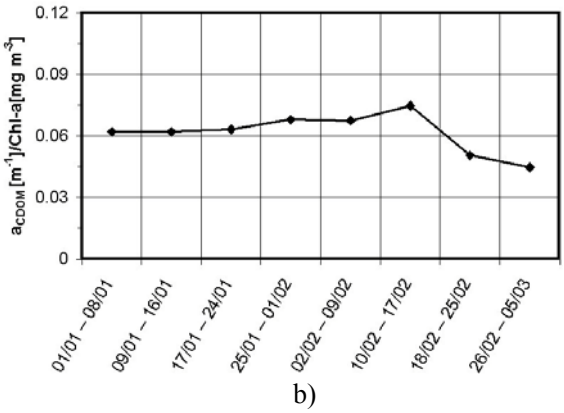
h)

Figure 23. Chl-a measured with the ELF-calibrated SeaWiFS algorithm in RSS during the periods a) 01/01 – 08/01, b) 09/01 – 16/01, c) 17/01 – 24/01, d) 25/01 – 01/02, e) 02/02 – 09/02, f) 10/02 – 17/02, g) 18/02 – 25/02 and h) 26/02 – 05/03.



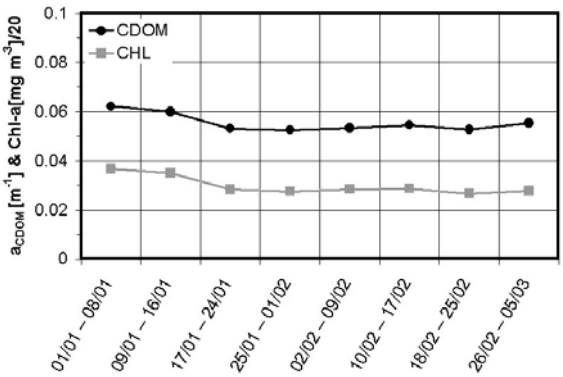
a)

Figure 24 continued on next page

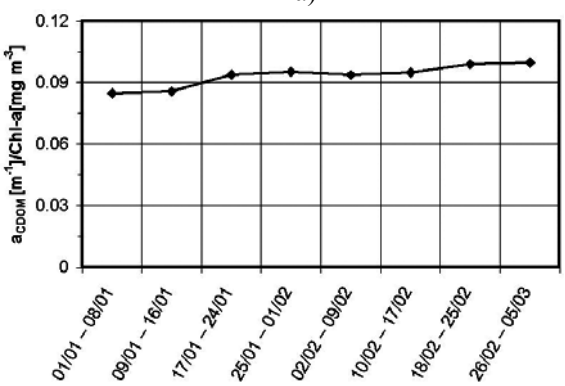


b)

Figure 24. a) Spatial averages of CDOM and chl-a in RSR and b) their ratio. Note that chl-a is scaled by a factor 20 in a).



a)



b)

Figure 25. a) Spatial averages of CDOM and chl-a in RSS and b) their ratio. Note that chl-a is scaled by a factor 20 in a).

5. Atmospheric Applications

5.1. Elastic Lidar

Section 2.3 made clear that the return of an atmospheric lidar is linked to the optical properties of the atmosphere along the laser beam. In general, the extinction and backscattering coefficient are determined mainly by the aerosols load. This parameter is important for its climatic effect, e.g. in Mediterranean towns [Papayannis et al. 1998]. Actually, α and β are both unknown, but some reliable methods have been elaborated to retrieve them from the signal, on the basis of some reasonable hypothesis [Klett 1981, Fernald 1984, Browell et al. 1985, Krichbaumer and Werner 1994]. In the following we will describe a scheme accepted by many researchers [Durieux and Fiorani 1997].

As we have observed, at first glance equation (2) contains two unknown quantities: α and β. Moreover, it is often very difficult to determine precisely the efficiency of the system. Consequently, the retrieval of α and β requires not only to establish a relation between them, but also to estimate a constant. As we will see, this latter evaluation will be replaced by the determination of the backscattering coefficient at a given distance.

In order to establish a relation between the extinction and backscattering coefficients, we distinguish the contributions of molecules and aerosols:

$$\alpha(R,\lambda) = \alpha_r(R,\lambda) + \alpha_a(R,\lambda), \tag{21}$$

$$\beta(R,\lambda) = \beta_r(R,\lambda) + \beta_a(R,\lambda), \tag{22}$$

and we remember that the terms concerning molecular scattering (α_r and β_r) can be modeled very precisely [Elterman 1968, Committee on Extension to the Standard Atmosphere 1976]. Finally, we recognize that α_a and β_a, the aerosol terms, are linked by the phase function P_π according to the equation:

$$\alpha_a(R,\lambda) = \frac{1}{P_\pi(R,\lambda)}\beta_a(R,\lambda). \tag{23}$$

Albeit the phase function varies with distance and time, in practice the final result is rather accurate even considering a constant P_π. The P_π value can be selected among those tabulated in the literature according to the aerosols kind (urban, rural, maritime, etc.) [Shettle and Fenn 1979].

After the definition [Browell et al. 1985]:

$$\mu(R,\lambda) \equiv \ln\left[n(R,\lambda)R^2\right] - 2\int_R^{R_M}\left\{\left[1 - \frac{3}{8\pi P_\pi(R',\lambda)}\right]\alpha_r(R',\lambda)\right\}dR', \tag{24}$$

where R_M is a given distance, the following differential equation is obtained from relations (2) and (21) – (23):

$$\frac{d\mu(R,\lambda)}{dR} = \frac{1}{\beta(R,\lambda)}\frac{d\beta(R,\lambda)}{dR} - \frac{2}{P_\pi(R,\lambda)}\beta(R,\lambda), \tag{25}$$

whose solution is [Klett 1981]:

$$\beta(R,\lambda) = \frac{\exp[\mu(R,\lambda) - \mu_M(\lambda)]}{\frac{1}{\beta_M(\lambda)} + \frac{2}{P_\pi(R,\lambda)}\int_R^{R_M}\exp[\mu(R',\lambda) - \mu_M(\lambda)]dR'}, \tag{26}$$

where:

$$\mu_M(\lambda) \equiv \mu(R_M,\lambda), \tag{27}$$

and β_M is the integration constant. Finally, from relations (21) – (23), we obtain the aerosols extinction and backscattering coefficients, i.e. two parameters closely linked to the aerosols load. Solution (26) has been formulated so that β_M is equal to $\beta(R_M)$. If a large R_M is chosen (corresponding to the free troposphere), the aerosols contribution to the backscattering coefficient is negligible and then β_M can be computed very precisely.

Figure 26 [Fiorani et al. 1998a] is an example of the lidar capability to analyze the dynamics of the aerosols load in the low troposphere. At the beginning of the experiment four aerosols layers are clearly visible, one above the ground and three near 500, 900 and 1400 m of altitude. Note that around sunset (18:07 Central Europe Time) the convective activity diminishes and consequently the upper layer bends down.

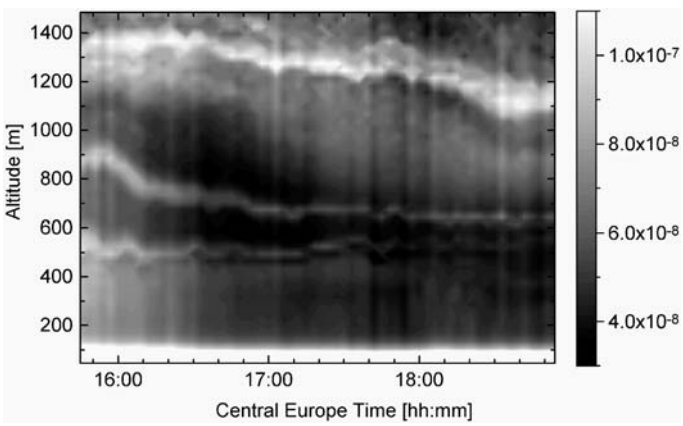


Figure 26. Aerosols backscattering coefficient [cm^{-1} sr^{-1}] measured as a function of time and altitude.

The analysis of the accuracy in the measurement of the aerosols extinction and backscattering coefficients is rather complex: we will summarize here only the main results, referring the reader to specific studies [Durieux and Fiorani 1997, Fiorani et al. 1997, Durieux and Fiorani 1998, Fiorani and Durieux 2001] for a more complete treatment. The statistical relative error of α and β is about equal to that of the lidar signal. This latter can be approximately written as [Schotland 1974, Pelon and Mégie 1982]:

$$s_n = \sqrt{n + n_B + n_D}\ , \tag{28}$$

where n_B and n_D are the number of photon counts originated by the brightness of the sky (solar background) and the electronic noise of the detector (dark current), respectively. The solar background can be computed from a model [Kneizys et al. 1988], the dark current is usually given by the manufacturer of the detector. Although equation (28) underestimate the statistical uncertainty [Durieux and Fiorani 1998, Fiorani and Durieux 2001], it gives a good rule of thumb estimate of it.

A lidar for the measurement of aerosols load could be realized with commercial components and would be compact enough to be transported. The transmitter is based on a Nd:YAG laser (1064 nm). This source has the advantage of being reliable and user-friendly. The receiver consists of a reflector telescope for light collection, of an IF for wavelength selection and of a PM (or an avalanche photodiode) for photon detection. The characteristics of the components listed in table 8 are closely inspired to those of commercial models.

Table 8. Characteristics of a lidar conceived for the measurement of aerosols load.

parameter	value	unit
Pulse energy	400	mJ
Optical efficiency (transmitter and receiver)	50	%
PM quantum efficiency	0.08	%
Filtering transmittance	50	%
Filtering bandwidth	0.5	nm
PM dark current	7000	count/s
Telescope diameter	50	cm

5.2. Correlation Lidar

With minor modification, the system just described can be used also for the measurement of wind speed. As we will see, all we need is the ability of steering the optics in any direction. In practice, the lidar can be aimed in two ways: either rotating together the transmitter-receiver couple or directing their optical axes by means of a pair of plain mirrors (like in an astronomic solar observatory).

According to equation (2), the number of photons backscattered by the atmosphere at a given distance from the instrument is proportional to the backscattering coefficient. This factor is a function of the density of aerosols responsible for the light scattering. Any change in its spatio-temporal distribution during a data acquisition will lead therefore to variations in the lidar returns from one laser shot to another. In particular, a wind flow along the beam axis

will be detected from the transport of the spatial inhomogeneities of β along the optical path [Eloranta et al. 1975].

In order to simplify the formalism underlying the measurement of wind speed, let us consider a discrete distance variable: in the following an index i will be incremented over the spatial bins. First of all the signal is multiplied by the range squared and normalized by the emitted energy on a shot-per-shot basis, to define a normalized signal $S(R_i)$ independent of the energy fluctuation and proportional to the backscattering coefficient:

$$S(R_i) = n(R_i)\frac{{R_i}^2}{E_0} \quad . \tag{29}$$

The mean normalized signal $S'(R_i)$ is obtained by averaging N successively recorded individual signals at successive times t_n as:

$$S'(R_i) = \frac{1}{N}\sum_{n=1}^{N} S(R_i, t_n) \quad . \tag{30}$$

The information about the shot-per-shot fluctuations of the signals with respect to $S'(R_i)$ is contained in the fractional deviations:

$$f(R_i, t_n) = \frac{S(R_i, t_n) - S'(R_i)}{S'(R_i)} \quad . \tag{31}$$

In order to determine the displacement of the backscattering inhomogeneities, a cross-correlation function $\rho(L, z_a, z_b)$ is defined with respect to an altitude interval (z_a, z_b) and to a lag value L corresponding to a discrete number of spatial points as:

$$\rho(L, z_a, z_b) = \frac{1}{\sigma\sigma_L}\sum_{i=a}^{b}\sum_{n=1}^{N-1} f(R_i, t_n) f(R_{i+L}, t_{n+1}) \quad , \tag{32}$$

where:

$$\sigma^2 = \sum_{i=a}^{b}\sum_{n=1}^{N-1} f^2(R_i, t_n) \quad , \tag{33}$$

$$\sigma_L^2 = \sum_{i=a}^{b}\sum_{n=1}^{N-1} f^2(R_{i+L}, t_{n+1}) \quad , \tag{34}$$

$$z_i = R_i \sin\theta \quad , \tag{35}$$

$$\Delta R = R_{i+1} - R_i \,, \tag{36}$$

$$\Delta t = t_{n+1} - t_n \,, \tag{37}$$

where θ is the elevation angle of the telescope.

The radial wind speed v_R in the altitude interval (z_a, z_b) is determined from the lag value L_{max} for which the correlation is maximum:

$$v_R = L_{max} \frac{\Delta R}{\Delta t} \,. \tag{38}$$

In the absence of vertical transport, the horizontal wind velocity is then estimated as:

$$v = v_R \cos\theta \cos\Phi \,, \tag{39}$$

where Φ is the difference in the azimuthal orientation of the wind vector and of the light path.

An example of wind speed measurement is given in Figure 27 [Fiorani et al. 1998b]. That plot provides also an idea of the accuracy that can be obtained in such experiments.

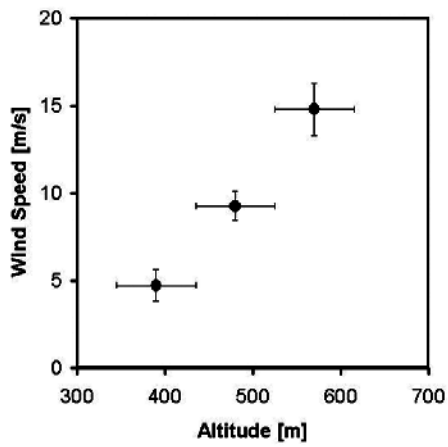


Figure 27. Wind speed as a function of altitude.

5.3. Differential Absorption Lidar

Laser remote sensing is widely used to measure pollutants concentration, usually by the differential absorption lidar (DIAL) technique [Measures 1992]. This method is based on the detection of the backscattered photons from laser pulses transmitted to the atmosphere at two different wavelengths. At one wavelength (λ_{OFF}), the light is almost only scattered by air molecules and aerosols, whereas at the other one (λ_{ON}), it is also absorbed by the pollutant under study. The difference between the two recorded signals is thus related to the pollutant

concentration. More precisely, on the basis of equation (2) the DIAL equation can be written [Measures 1992]:

$$C(R) = \frac{1}{2[\sigma(\lambda_{ON}) - \sigma(\lambda_{OFF})]} \times$$

$$\times \left\{ \frac{d}{dR} \ln\left[\frac{n(R, \lambda_{OFF})}{n(R, \lambda_{ON})} \right] - \frac{d}{dR} \ln\left[\frac{\beta(R, \lambda_{OFF})}{\beta(R, \lambda_{ON})} \right] - 2[\alpha(R, \lambda_{ON}) - \alpha(R, \lambda_{OFF})] \right\}, \quad (40)$$

where C and σ are, respectively, the concentration and the absorption cross-section of the pollutant molecule. If λ_{OFF} and λ_{ON} are close enough, equation (40) gets simpler:

$$C(R) = \frac{1}{2[\sigma(\lambda_{ON}) - \sigma(\lambda_{OFF})]} \frac{d}{dR} \ln\left[\frac{n(R, \lambda_{OFF})}{n(R, \lambda_{ON})} \right]. \quad (41)$$

In both cases, the analysis of the measurement error in pollutant concentration is rather complex: once more the reader will find it elsewhere [Durieux and Fiorani 1997, Fiorani et al. 1997, Durieux and Fiorani 1998, Fiorani and Durieux 2001]. Let us only recall here that the final accuracy ranges from 1%, at low altitudes, to 100%, at the end of the measurement range.

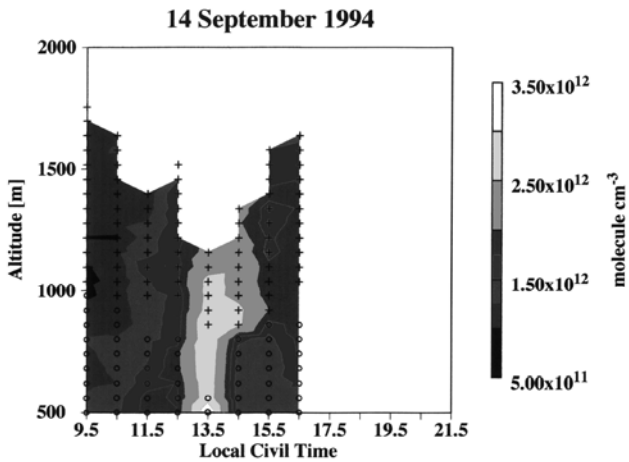


Figure 28. Ozone concentration as a function of time and altitude.

Figure 28 [Durieux et al. 1998] is an example of ozone concentration measurement during a photochemical smog episode in Athens. Ozone concentration was successively increasing and decreasing during 14 September 1994, up to values as high as three times the background level during polluted days. This increase was responsible for a shortening of the measurement range due to a larger absorption of the emitted light. This effect was partially compensated for at 12:30 by a change of the experimental parameters inducing an increase of

the SNR at a given altitude. However, this change was not sufficient to compensate for the important raise of the ozone concentration observed between 12:30 and 13:30. Consequently, the range at 13:30 was considerably lower than at 12:30. The photochemical smog episode was also visible from measurements performed with ground based punctual instruments installed near the lidar. The increase in ozone concentration was accompanied by a growth in the load of carbon black linked to a reduction of the visibility. These variations were due to a transport of pollutants from the center of the city to the site of measurement by a wind of about 0.8 ms^{-1} at the ground level blowing in that direction between 12.30 and 14.00. The wind direction then changes at 14.30, so that the pollutants are transported away from the site and the ozone concentration decreases down to its previous level.

A summary of the characteristics of the system that performed the above-mentioned measurements is given in table 9. Two excimer pumped dye lasers equipped with frequency doubling crystals are used to generate λ_{ON} and λ_{OFF} (272.2 and 291.65 nm, respectively: these values represent a compromise between maximizing the difference in the ozone extinction coefficient, in order to improve the sensitivity of the measurements and the necessity to minimize the corresponding difference in wavelength, as one wants to limit the uncertainties inherent in the correction for the differential extinction and backscattering by the aerosols; furthermore, they have been chosen so that there is no significant contribution of the SO_2 molecule to the differential absorption). Although the output energy from the XeCl excimer lasers reaches 400 mJ per pulse at a 100 Hz repetition rate, they are routinely set to provide about 300 mJ at 50 Hz for a better stability of the energy output over long periods of operation. Coumarin 153 and Rhodamine 6G are used for the generation of λ_{ON} and λ_{OFF} with a conversion efficiency near 15%. The throughput of the frequency doubling crystals is about 10%. The optical axes for the light emission and the light detection are separated by 430 mm. The geometrical position of the entry point of the emitted beam in the cone describing the volume of space seen by the detector defines an effective optical gate depending on the beam divergence and on the angle between the axis of the transmitter and of the receiver. At the exit of the Cassegrain telescope used for its collection, the light is injected through a lens into an optical fiber. The lidar returns are filtered by a spectrograph. An IF is placed after the slit selecting λ_{OFF} to further reject the solar background and to reduce the intensity of the signal to a value comparable to that obtained for λ_{ON}. Both signals illuminate approximately the same area of the active surface of a 12-stage PM. A normalization factor for the energy of each emitted light pulse is measured by integrating the current collected at the anode of a 5-stage PM. It is illuminated by less than 1% reflection off a plate. This reflection is attenuated by two neutral densities and one IF to avoid saturation.

Table 9. Characteristics of a lidar conceived for the measurement of pollutants concentration.

parameter	value	unit
Pulse energy	5	mJ
Optical efficiency (transmitter and receiver)	50	%
PM quantum efficiency	25	%
Filtering transmittance	50	%
Filtering bandwidth	2	nm
PM dark current	10	count/s
Telescope diameter	60	cm

6. Conclusion

The lidar technique has been reviewed, with special emphasis on its environmental applications on the three Earth domains were life develops: lithosphere, hydrosphere and atmosphere.

Lithospheric laser range finders can be very useful for remote sensing in hazardous area and archaeological prospecting. An example of such a system, GEOLIDAR, has been described, demonstrating that underground lidars can be accurate, fast, compact and user-friendly.

Hydrospheric applications are more common and give their valuable contribution in the thorough understanding of the biogeochemical cycles of the Earth system. The main characteristic of a lidar fluorosensor (ELF) devoted to bio-optical characterization of oceanic waters have been discussed. Its results show that lidar fluorosensors can be very useful in extreme environments, such as Antarctic basins, where the sea images acquired by space borne radiometers are deeply affected by cloud and ice coverage. Moreover, laser measurements are more suitable than in situ samplings to the accurate calibration/validation of the satellite sensors because, from one hand their geographic coverage and spatial resolution are closer to image extent and pixel size, respectively, of the ocean color data products, from the other one the number of laser measurements is much larger than that of in situ stations. ELF data have been used to calibrate the SeaWiFS algorithm for chl-a retrieval. The results show that standard algorithms (global) can misestimate chl-a in some space regions and/or time intervals. Moreover, those new estimates of chl-a have been employed to refine a known PP model in different regions of the Southern Ocean. Eventually, an original algorithm for a_{CDOM} retrieval has been developed. Although this study is rather fresh, it demonstrates that the information on spatial distributions and temporal dynamics of chl-a and a_{CDOM} improves our knowledge of the link between phytoplankton biomass and dissolved organic matter.

Atmospheric applications of optical radars have been introduced discussing the laser measurement of three important air parameters: aerosol load, wind speed and gas concentration with elastic, correlation and differential absorption lidars, respectively. Range- and time-resolved profiles of those parameters are powerful tools to describe the complex system represented by the Earth troposphere. With respect to usual air sampling techniques, lidars present many advantages: continuous retrieval of aerosols load, wind speed and pollutants concentration profile in a considerable range and with a good spatio-temporal resolution; probe-less measurement, thus eliminating the possibility of modifying the sample; integrated-path determination, less sensitive to local effects; capability of sweeping the complete hemisphere, thus allowing to follow the physico-chemical dynamics of the atmosphere.

The researches presented here demonstrate that laser radar can improve our understanding of the global cycles and, as a consequence, of our planet. It is desirable that, together with in situ samplers and satellite radiometers, they contribute to long-term records of Earth observations necessary to assess the coupled ocean-atmosphere general circulation models that will help the progression of climate science from description to prediction.

Acknowledgements

The author is deeply grateful to E. Durieux and coworkers (EPFL, Lausanne), N. Spinelli and coworkers (INFM, Naples), S. Vetrella and coworkers (CORISTA, Naples), and R. Barbini, F. Colao, R. Fantoni and A. Palucci (ENEA, Frascati) for the fundamental involvement and the constant encouragement.

The contribution of L. De Dominicis (data acquisition during MIPOT), P. Aristipini and D. Ferrante (technical design), D. Del Bugaro and R. Giovagnoli (mechanical parts), I. Menicucci (electronics), and E.S. Artamonov, M. Galli and I.G. Okladnikov (data processing) is kindly acknowledged.

A special thank is addressed to S. Fonda Umani and coworkers, L. Lazzara and coworkers, V. Saggiomo and coworkers, and G. Spezie and coworkers for providing unpublished data.

This work has been supported by the Italian Antarctic Research Program (PNRA) – Technology Sector, 5b1 and 11-5 Projects, PRNA – Oceanographic Sector, 8.3 Project and by three ENEA fellowships (one to E.S. Artamonov and two to I.G. Okladnikov).

The author would like to thank the SeaWiFS Project (Code 970.2) and the Distributed Active Archive Center (Code 902) at the Goddard Space Flight Center, Greenbelt, MD 20771, for the production and distribution of these data, respectively. These activities are sponsored by NASA's Mission to Planet Earth Program.

References

Aiken J., Moore G.F., Trees C.C., Hooker S.B. and Clark D.K., 1995, The SeaWiFS CZCS-type pigment algorithm, in *SeaWiFS Technical Report Series (NASA Technical Memorandum 104566)*, Eds. S.B. Hooker and E.R. Firestone, NASA, Greenbelt, Vol. 29.

Arrigo K.R., Worthen D., Schnell A. and Lizotte M.P., 1998, Primary production in Southern Ocean waters, *Journal of Geophysical Research C* **103**, 15587-15600.

Barbini R., Colao F., De Dominicis L., Fantoni R., Fiorani L., Palucci A. and Artamonov E.S., 2004, Analysis of simultaneous chlorophyll measurements by lidar fluorosensor, MODIS and SeaWiFS, *International Journal of Remote Sensing* **25**, 2095-2110.

Barbini R., Colao F., Fantoni R., Fiorani L., Okladnikov I.G. and Palucci A., 2005, Lidar calibrated satellite sensed primary production in the Southern Ocean, *Journal of Optoelectronics and Advanced Materials* **7**, 1091-1102.

Barbini R., Colao F., Fantoni R., Fiorani L. and Palucci A., 2001a, Remote sensing of the Southern Ocean: techniques and results, *Journal of Optoelectronics and Advanced Materials* **3**, 817-830.

Barbini R., Colao F., Fantoni R., Fiorani L. and Palucci A., 2002, Lidar tuning of the SeaWiFS algorithms in the Antarctic Ross Sea: preliminary results, in *E-proceedings of the 7th International Conference on Remote Sensing for Marine and Coastal Environments*, Ed. B. Petoskey, VERIDIAN, Miami, paper C-3.

Barbini R., Colao F., Fantoni R., Fiorani L. and Palucci A., 2003a, Lidar fluorosensor calibration of the SeaWiFS chlorophyll algorithm in the Ross Sea, *International Journal of Remote Sensing* **24**, 3205-3218.

Barbini R., Colao F., Fantoni R., Fiorani L., Palucci A., Artamonov E.S. and Galli M., 2003b, Remotely sensed primary production in the western Ross Sea: results of in situ tuned models *Antarctic Science* **15**, 77-84.

Barbini R., Colao F., Fantoni R., Palucci A. and Ribezzo S., 1999, Shipborne laser remote sensing of the Venice lagoon, *International Journal of Remote Sensing* **20**, 2405-2421.

Barbini R., Colao F., Fantoni R., Palucci A. and Ribezzo S., 2001b, Differential lidar fluorosensor system used for phytoplankton bloom and seawater quality monitoring in Antarctica, *International Journal of Remote Sensing* **22**, 369-384.

Behrenfeld M.J. and Falkowski P.G., 1997a, A consumer's guide to phytoplankton primary productivity models, *Limnology and Oceanography* **42**, 1479-1491.

Behrenfeld M.J. and Falkowski P.G., 1997b, Photosynthetic rates derived from satellite-based chlorophyll concentrations, *Limnology and Oceanography* **42**, 1-20.

Boyd P.W., 2002, Environmental factors controlling phytoplankton processes in the Southern Ocean, *Journal of Phycology* **38**, 844-861.

Bristow M., Nielsen D., Bundy D. and Furtek R., 1981, Use of Raman emission to correct airborne laser fluorosensor data for effects of water optical attenuation, *Applied Optics* **20**, 2889-2906.

Browell E.V., Ismail S. and Shipley S.T., 1985, Ultraviolet DIAL measurements of O_3 profiles in regions of spatially inhomogeneous aerosols, *Applied Optics* **24**, 2827-2836.

Budillon G., Pacciaroni M., Cozzi S., Rivaro P., Catalano G., Ianni C. and Cantoni C., 2003, An optimum multiparameter mixing analysis of the shelf waters in the Ross Sea, *Antarctic Science* **15**, 105-118.

Bunkin A.F. and Voliak K.I., 2001, *Laser Remote Sensing of the Ocean*, John Wiley & Sons, New York.

Campbell J.W., Blaisdell J.M. and Darzi M., 1995, Level-3 SeaWiFS Data Products: Spatial and Temporal Binning Algorithms, in *SeaWiFS Technical Report Series (NASA Technical Memorandum 104566)*, Eds. S.B. Hooker and E.R. Firestone, NASA, Greenbelt, Vol. 32.

Carlson R.J., Lovelock J.E., Andrea M.O. and Warren S.G., 1987, Oceanic phytoplankton, atmospheric sulfur, cloud albedo and climate, *Nature* **326**, 655-661.

Carmer D.C. and Peterson L.M., 1996, Laser radar in robotics, *Proceedings of the IEEE* **84**, 299-320.

Chekalyuk A.M. and Gorbunov M.Yu., 1993, Lidar in situ study of sun light regulation of phytoplankton photosynthetic activity and chlorophyll fluorescence, *Proceedings of SPIE* **1922**, 421-427.

Clifford S.F., Kaimal J.C., Lataitis R.J. and Strauch R G., 1994, Ground-based remote profiling in atmospheric studies: an overview, *Proceedings of the IEEE* **82**, 313-355.

Collis R.T.H., 1966, Lidar: a new atmospheric probe, *Quarterly Journal of the Royal Meteorological Society* **92**, 220-230.

Collis R.T.H. and Russel P.B., 1976, Lidar measurements of particles and gases by elastic backscattering and differential absorption, in *Laser Monitoring of the Atmosphere*, Ed. E.D. Hinkley, Springer-Verlag, Berlin, pp. 71-151.

Committee on Extension to the Standard Atmosphere, 1976, *United States Standard Atmosphere*, United States Government Printing Office, Washington.

Derr V.E. and Little C.G., 1970, A comparison of remote sensing of the clear atmosphere by optical, radio, and acoustic radar techniques, *Applied Optics* **9**, 1976-1991.

Dierssen H.M., Vernet M. and Smith R.C., 2000, Optimizing models for remotely estimating primary production in Antarctic coastal waters, *Antarctic Science* **12**, 20-32.

Durieux E. and Fiorani L., Data processing, 1997, in *Instrument Development for Atmospheric Research and Monitoring*, Eds. J. Bösenberg, D. Brassington, P.C. Simon, Springer-Verlag, Berlin, pp. 89-116.

Durieux E. and Fiorani L., 1998, Measurement of the lidar signal fluctuation with a shot-per-shot instrument, *Applied Optics* **37**, 7128-7131.

Durieux E., Fiorani L., Calpini B., Flamm M., Jaquet L., Van den Bergh H., 1998, Tropospheric ozone measurements over the Great Athens Area during the MEDCAPHOT-TRACE campaign with a new shot-per-shot DIAL instrument. Experimental system and results, *Atmospheric Environment* **32**, 2141-2150.

Elachi C., 1987, *Introduction to the Physics and Techniques of Remote Sensing*, John Wiley & Sons, New York.

Eloranta E.W., King J.M. and Weinman J.A., 1975, The determination of wind speeds in the boundary layer by monostatic lidar, *Journal of Applied Meteorology* **14**, 1485-1489.

Elterman L., 1968, *UV, Visible and IR Attenuation for Altitudes to 50 km – AFCRL-68-0153*, United States Air Force, Hanscom.

Esaias W.E., Abbott M.R., Barton I., Brown O.B., Campbell J.W., Carder K.L., Clark D.K., Evans R.H., Hoge F.E., Gordon H.R., Balch W.M., Letelier R. and Minnett P.J., 1998, An overview of MODIS capabilities for ocean science observations, *IEEE Transactions on Geoscience and Remote Sensing* **36**, 1250-1265.

Fantoni R., Fiorani L., Palucci A. and Okladnikov I.G., 2005, *Regional models for satellite measurements of primary productivity in the Ross Sea*, ENEA, Frascati.

Fernald F.G., 1984, Analysis of atmospheric lidar observations: some comments, *Applied Optics* **23**, 652-653.

Fichot C.G. and Miller W.L., 2004, CDOM on the global scale: implementation of an improved algorithm (SeaUV) to7 years of SeaWiFS data, in *E-proceedings of Ocean Optics XVII*, Ed. S. Ackleson, ONR, Fremantle, Paper 223.

Fiocco G. and Smullin L.D., 1963, Detection of scattering layers in the upper atmosphere (60-140 km) by optical radar, *Nature* **199**, 1275-1276.

Fiorani L., Armenante M., Capobianco R., Spinelli N. and Wang X., 1998a, Self-aligning lidar for the continuous monitoring of the atmosphere, *Applied Optics* **37**, 4758-4764.

Fiorani L., Bortone M., Mattei S., Ruocchio C., Salomé A. and Vetrella S., 2000, GEOSCOPE and GEOLIDAR: integrated instruments for underground archaeological investigations, *Subsurface Sensing Technologies and Applications* **1**, 305-319.

Fiorani L., Calpini B., Jaquet L., Van den Bergh H. and Durieux E., 1997, Correction scheme for experimental biases in differential absorption lidar tropospheric ozone measurements based on the analysis of shot per shot data samples, *Applied Optics* **36**, 6857-6863.

Fiorani L., Calpini B., Jaquet L., Van den Bergh H. and Durieux E., 1998b, A combined determination of wind velocities and ozone concentrations for a first measurement of ozone fluxes with a DIAL instrument during the MEDCAPHOT-TRACE campaign, *Atmospheric Environment* **32**, 2151-2159.

Fiorani L. and Durieux E., 2001, Comparison among error calculations in differential absorption lidar measurements, *Optics and Laser Technology* **3**, 371-377.

Fiorani L., Mattei S. and Vetrella S., 1998, Laser methods for the atmospheric correction of marine radiance data sensed from satellite, *Proceedings of SPIE* **3496**, 176-187.

Fiorani L., Palucci A. and Okladnikov I.G., 2005, *Remote sensing of chromophoric dissolved organic matter by ocean color radiometers calibrated by laser-induced fluorescence*, ENEA, Frascati.

Gordon H.R. and Wang M., 1994, Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm, *Applied Optics* **33**, 443-452.

Goyer G.G. and Watson R., 1963, The laser and its application to meteorology, *Bulletin of the American Meteorological Society* **44**, 564-570.

Grant W.B., 1995, Lidar for atmospheric and hydrospheric studies, in *Tunable Laser Applications*, Ed. F.J. Duarte, Dekker Publishing, New York, pp. 213-305.

Hecht J., 1985, *Laser Pioneers*, Academic Press, San Diego.

Hoge F.E. and Swift R.N., 1981, Airborne simultaneous spectroscopic detection of laser induced water Raman backscatter and fluorescence from chlorophyll a and other naturally occurring pigments, *Applied Optics* **20**, 3197-3205.

Hoge F.E., Vodacek A., Swift R.N., Yungel J.K. and Blough N.V., 1995, Inherent optical properties of the ocean: retrieval of the absorption coefficient of chromophoric dissolved organic matter from airborne laser spectral fluorescence measurements, *Applied Optics* **34**, 7032-7038.

Hoge F.E., Wright C.W., Lyon P.E., Swift R.N. and Yungel J.K., 1999, Satellite retrieval of inherent optical properties by inversion of an oceanic radiance model: a preliminary algorithm, *Applied Optics* **38**, 495-504.

Hoge F.E., Wright C.W., Swift R.N. and Yungel J.K., 1998, Airborne laser-induced oceanic chlorophyll fluorescence: solar-induced quenching corrections by use of concurrent downwelling irradiance measurements, *Applied Optics* **37**, 3222-3226.

Hooker S.B., Esaias W.E., Feldman G.C., Gregg W.W. and McClain C.R., 1992, An overview of SeaWiFS and ocean color, in *SeaWiFS Technical Report Series (NASA Technical Memorandum 104566)*, Eds. S.B. Hooker and E.R. Firestone, NASA, Greenbelt, Vol. 1.

Hovis W.A, 1980, Nimbus-7 coastal zone color scanner: system description and initial imagery, *Science* **210**, 60-63.

Huot J.-P., Tait H., Rast M., Delwart S., Bézy J.-L. and Levrini G., 2002, The optical imaging instruments and their applications: AATSR and MERIS, *ESA Bulletin* **106**, 56-66.

Jelalian A. V., 1992, *Laser Radar Systems*, Artech House, Norwood.

Johannessen S.C., Miller W.L. and Cullen J.J., 2003, Calculation of UV attenuation and colored dissolved organic matter absorption spectra from measurements of ocean color, *Journal of Geophysical Research* **108(C9)**, 3301.

Kahru M. and Mitchell B.G., 2001, Seasonal and nonseasonal variability of satellite-derived chlorophyll and colored dissolved organic matter concentration in the California Current, *Journal of Geophysical Research* **106(C2)**, 2517-2529.

Klett J.D., 1981, Stable analytical inversion solution for processing lidar returns, *Applied Optics* **20**, 211-220.

Kneizys F.X., Anderson G.P., Shettle E.P., Gallery W.O., Abreu L.W., Selby J.E.A., Chetwynd J.H. and Clough S.A., 1988, *Users Guide to LOWTRAN 7 – AFGL-TR-88-0177*, United States Air Force, Hanscom.

Koechler C., Verdebout J., Bertolini G., Gallotti A., Zanzottera G. and Fiorina L., 1992, Determination of aquatic parameters by time resolved laser fluorosensor operating from helicopter, *Proceedings of SPIE* **1714**, 93-107.

Krichbaumer W. and Werner C., 1994, Current state-of-the-art of LIDAR inversion methods for atmospheres of arbitrary optical density, *Applied Physics B* **59**, 517-523.

Mamon G., Youmans D.G., Sztankay Z.G. and Mongan C.E., 1978, Pulsed GaAs laser terrain profiler, *Applied Optics* **17**, 205-221.

McClain C.R., Esaias W.E., Barnes W., Guenther B., Endres D., Hooker S.B., Mitchell B.G. and Barnes R., 1992, SeaWiFS Calibration and Validation Plan, in *SeaWiFS Technical Report Series (NASA Technical Memorandum 104566)*, Eds. S.B. Hooker and E.R. Firestone, NASA, Greenbelt, Vol. 3.

McClain C.R., Feldman G.C. and Hooker S. B., 2004, An Overview of the SeaWiFS project and strategies for producing a climate research quality global ocean bio-optical time series, *Deep-Sea Research II* **51**, 5-42.

Measurement Devices, 2005, *Cavity Autoscanning Laser System*, Measurement Devices, Malton (available at www.mdl.co.uk).

Measures R.M., 1992, *Laser Remote Sensing*, Krieger Publishing, Malabar.

Middleton W.E.K. and Spilhaus A.F., 1953, *Meteorological Instruments*, University of Toronto, Toronto.

Mitchell B.G., 1992, Predictive bio-optical relationships for polar oceans and marginal ice zones, *Journal of Marine Systems* **3**, 91-105.

Mitchell G., 2003, AESOPS/NBP97-8, in *US JGOFS Process Study Data 1989-1998 (CD-ROM)*, Ed. US JGOFS Data Management Office, WHOI, Woods Hole, Vol. 1.

Mitchell B.G. and Holm-Hansen O., 1991, Bio-optical properties of Antarctic Peninsula waters: differentiation from temperate ocean models, *Deep-Sea Research Part A* **38**, 1009-1028.

O'Reilly J.E., Maritorena S., Mitchell B.G., Siegel D.A., Carder K.L., Garver S.A., Kahru M. and McClain C., 1998, Ocean color chlorophyll algorithms for SeaWiFS, *Journal of Geophysical Research C* **103**, 24937-24953.

O'Reilly J.E., Maritorena S., O'Brien M.C., Siegel D.A., Toole D., Menzies D., Smith R.C., Mueller J.L., Mitchell B.G., Kahru M., Chavez F.P., Strutton P., Cota G.F., Hooker S.B., McClain C.R., Carder K.L., Müller-Karger F., Harding L., Magnuson A., Phinney D., Moore G.F., Aiken J., Arrigo K.R., Letelier R. and Culver M., 2000, SeaWiFS Postlaunch Calibration and Validation Analyses, Part 3, in *SeaWiFS Postlaunch Technical Report Series (NASA Technical Memorandum 2000-206892)*, Eds. S.B. Hooker and E.R. Firestone, NASA, Greenbelt, Vol. 11.

Papayannis A., Bais A., Balis D., van den Bergh H., Calpini B., Durieux E., Fiorani L., Jaquet L., Ziomas I. and Zerefos C.S., 1998, Role of urban and suburban aerosols on solar UV radiation over Athens, Greece, *Atmospheric Environment* **32**, 2193-2201.

Pelon J. and Mégie G., 1982, Ozone monitoring in the troposphere and lower stratosphere: evaluation and operation of a ground-based lidar station, *Journal of Geophysical Research* **87**, 4947-4955.

Reuter R., Diebel D. and Hengstermann T., 1993, Oceanographic laser remote sensing: measurements of hydrographic fronts in the German Bight and in the northern Adriatic Sea. *International Journal of Remote Sensing* **14**, 823-848.

Sathyendranath S., Cota G., Stuart V., Maass H. and Platt T., 2001, Remote sensing of phytoplankton pigments: a comparison of empirical and theoretical approaches, *International Journal of Remote Sensing* **22**, 249-273.

Schotland R.M., 1974, Errors in the lidar measurement of atmospheric gases by differential absorption, *Journal of Applied Meteorology* **13**, 71-77.

Sellers W.D., 1965, *Physical Climatology*, University of Chicago Press, Chicago.

Shettle E.P. and Fenn R.W., 1979, *Models for the Aerosols of the Lower Atmosphere and the Effects of Humidity Variations on Their Optical Properties – AFGL-TR-79-0214*, United States Air Force, Hanscom.

Siegel D.A., Maritorena S., Nelson N.B., Hansell D.A. and Lorenzi-Kayser M., 2002, Global distribution and dynamics of colored dissolved and detrital organic materials, *Journal of Geophysical Research* **107**(C12), 3228.

Siegel D.A., Wang M., Maritorena S. and Robinson W., 2000, Atmospheric correction of satellite ocean color imagery: the black pixel assumption, *Applied Optics* **39**, 3582-3591.

Smith R.C., Baker K.S., Dierssen H.M., Stammerjohn S.E. and Vernet M., 2000, Bio-optical modeling of primary production from SeaWiFS ocean color data for the western Antarctic Peninsula region, in *E-proceedings of Ocean Optics XV*, Eds. S. Ackleson and J. Marra, Musée Océanographique, Monaco, Paper 79.

Svelto O., 1998, *Principle of Lasers*, Plenum, New York.

Toole D.A. and Siegel D.A., 2004, Light-driven cycling of dimethylsulfide (DMS) in the Sargasso Sea: closing the loop, *Geophysical Research Letters* **31**, L09308.

Toole D.A., Siegel D.A., Slezak D., Kiene R.P., Nelson N.B. and Dacey, J.W., 2002, A light driven upper-ocean dimethylsulfide (DMS) biogeochemical cycling model for the Sargasso Sea, in *E-proceedings of the AGU Fall Meeting*, Ed. R. Wesson, AGU, San Francisco, paper 210.

Van Zee H. and Walsh I., 2004, *Ac-9 Specification Sheet*, Wet LABS, Philomath.

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN: 1-60021-194-1
Editor: Steven B. Larkin, pp. 173-192

Chapter 5

THERMAL EFFECTS AND POWER SCALING OF DIODE-PUMPED SOLID-STATE LASERS

Xiaoyuan Peng
Light Age, Inc., 500 Apgar Drive, Somerset, NJ 08873,USA
Anand K. Asundi
School of MAE, Nanyang Technological University, Singapore

Abstract

Thermal effects of diode-pumped solid-state (DPSS) lasers were theoretically analyzed and modeled for Nd doped lasing materials. The models determined the thermal lens and separated the end effect from the thermal effects. The knowledge of thermal lensing in the laser cavity is critical to scale the laser output power with the diffraction-limit laser beam output and to optimize the design of DPSS lasers. To validate the model, a new modified Twyman-Green interferometer is proposed to measure the thermal effects directly and accurately. Good agreement with the analytical model was obtained. Based on these pre-determined thermal lens effects, a power-scaling model was setup to guide the development of DPSS lasers. Within the fracture limits of the lasing crystals under high-power end-pumping, together with the design aims of the TEM_{00} mode output and the conversion efficiency, the model optimized the mode-to-pump ratio, doping concentration and dimensions of the lasing crystals. It also predicted the output power of DPSS lasers. The model showed a good agreement with the experiment later on. Based on the model, a practical diode-pumped Nd:YVO_4 laser system with a simple linear cavity was demonstrated. The laser produced an output power of 9.8W in the TEM_{00} mode under a pump power of 28W where the rod surface acting as one of end mirrors in the linear plano-concave resonator. In conclusion, the models are quite useful for broadly guiding the design and development of most DPSS lasers, including rare earth ions doped YAG, YVO_4, and YLF, etc.

1. Introduction

In diode-pumped solid-state lasers, the combination of volumetric heating of laser materials by the absorbed pump radiation and surface cooling required for heat extraction leads to a non-uniform temperature distribution in the rod, which results in a distortion of the laser beam due to temperature- and stress-dependent variations of the refractive index. In general, there are two typical pump configurations in diode-pumped solid-state (DPSS) lasers. End-pumped lasers take advantage of the properties of the laser diode and solid-state laser resonator to concentrate the pump power within the mode volume of the active medium longitudinally. The efficiency of end-pumped lasers is typically higher than that of side-pumped lasers in CW TEM_{00} operation [[1]], [[2]]. Recently, Minassian etc. demonstrated a highly efficient side-pumped Nd: YVO_4 by using a grazing-incident bounce geometric slab [[3]], but the roundness of the laser beam is still an issue, which needs to be improved for the requirements of most applications. However end-pumped lasers have more serious thermal lensing effects than side-pumped. The thermal effects consist of three parts: index parts, end effect, and stress effect [[4]], [[5]]. For Nd: YVO_4 and Nd:YAG rods, the index parts introduce a lens whose power has the same sign as the change of refractive index with temperature gradient; while for Nd:YLF, there are of different signs. The non-uniform temperature distribution leads to strain and stress in the crystal, which also changes the refractive index, leading to stress-induced birefringence. The crystal expands with temperature, which results in the rod expansion along the pump axis. The end bulge forms a lens which has the same sign as the expansion of the crystal along the resonator axis. As a consequence, the crystal forms a thermally induced lens, called a thermal lens, which significantly influences the laser performance in high-power diode-pumped solid-state lasers. The knowledge of thermal lensing in the laser cavity is crucial to the scaling of laser power and optimal design of laser systems [[6]]. In power scaling, besides the thermal lens, the maximum incident pump power is restricted by thermal fracture of the laser crystals [[7]]. For example, the fracture stress of YLF is as low as 40MPa, which is about five times lower than that of Nd:YAG. In a recent study, it was found that the thermal fracture stress of vanadate is almost three times lower than that of Nd:YAG [[6]]. For a 0.5 at.% Nd doped vanadate crystal and a pump beam radius of 0.4mm, the maximum pump power is limited by thermal fracture to approximately 36W for a single-end-pump geometry [[6]]. Therefore, one of the most important features in laser design is to avoid thermally induced fracture, especially for the YVO_4 and YLF crystals.

In this chapter, based upon the lattice structure of the crystals, the thermal effects of diode-end-pumped Nd: YVO_4, Nd: YLF and Nd: YAG lasers are discussed using an analytical model. To validate the theoretical models, a new Twyman-Green interferometer is developed that measures thermal lensing effects for of Nd:YVO_4 and Nd:YAG lasers. Based on the determined thermal lensing effect, a power-scaling model was setup to guide the development of DPSS lasers. Within the fracture limit of lasing crystals, and the design aims of the TEM_{00} mode output and the conversion efficiency, the model optimized the mode-to-pump ratio, doping concentration and dimensions of the lasing crystals. It also predicted the output power of DPSS lasers. The model showed a good agreement with the experiment later on. Based on the model, a practical diode-pumped Nd:YVO_4 laser system was demonstrated.

2. Modelling of Thermal Lens

The fundamental assumptions in this analysis are:

- laser diodes emit at a central wavelength of 808nm with a spectrum bandwidth of 3-4nm, which coincides with the strong absorption bands of neodymium doped YAG, YLF, and YVO_4;
- single-end-pump geometry is used;
- air-cooled end faces and conduction-cooled sides of the crystal (edge-cooled structure);
- Use of a stable half-symmetric resonator with one directly coated end mirror, e.g. the laser rod surface as one of the cavity end mirrors.

For a Nd doped crystal (r_0-mm-radius) end-pumped by a CW pump source, a side and an end view of the laser crystal and heat sink geometry are shown in Figure 1. We assume that the temperature of the laser crystal is controlled by the heat sink surrounding the laser rod. In this geometry, most of the heat is generated inside a small volume near the center of the crystal and removed from the radial surface via conduction. This results in a radial temperature gradient inside the rod.

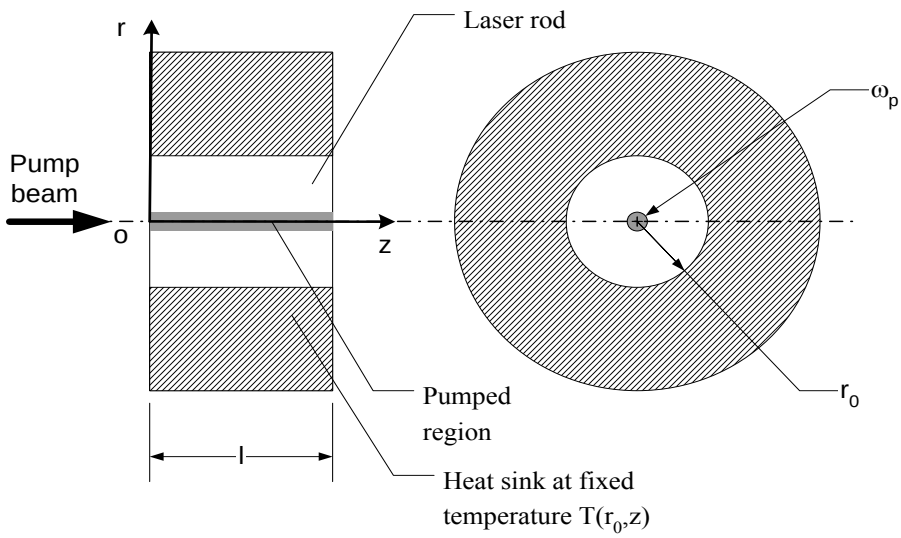


Figure 1. Side view and end view of an end-pumped laser. The rod is peripherally cooled by a water-cooled copper heat sink. The radius of the rod is r_0, and the $1/e^2$ pump spot size of the pump beam is ω_p.

Assuming that heat radiation along the z axis decays exponentially, the heat flow equation can be described in a two-dimensional form as [[8]]

$$\frac{1}{r}\cdot\frac{\partial}{\partial r}\cdot\left(r\cdot\frac{\partial T(r,z)}{\partial r}\right)+\frac{Q(r,z)}{K}=0 \tag{1}$$

where K is the thermal conductivity of the laser rod, Q(r, z) represents the thermal power density of the heat radiation, namely, heat generated by the incident pump power per unit volume.

The radiation from an end-pumping source can be characterized as a circular Gaussian spatial function with a beam waist ω_p [[9]]

$$Q(r,z) = \frac{2\eta_h \cdot P_{in} \cdot \alpha}{\pi \cdot \omega_p{}^2} \cdot \exp(-\frac{2r^2}{\omega_p{}^2}) \cdot \exp(-\alpha \cdot z) \tag{2}$$

where P_{in} is the incident pump power, η_h is the heat conversion coefficient (or fractional heat coefficient) resulting from fluorescence efficiency, upconversion, and quantum defect, α is the absorption coefficient denoting the axial absorption of pump power. The pump beam ω_p has a Gaussian spatial distribution expressed by

$$\omega_p{}^2(z) = \omega_{p0}{}^2 \cdot \left[1 + (\frac{M^2 \cdot \lambda_p \cdot z}{\pi \cdot n_0 \cdot \omega_{p0}{}^2})^2 \right] \tag{3}$$

where n_0 is the refractive index of the laser rod, ω_{p0} is the beam waist of the pump laser, λ_p is the center wavelength of pump laser. Substituting Eqn. 2 and 3 into Eqn. 1, the analytical solution to the differential equation is given by [[10]]

$$T(r,z) - T(0,z) = -\frac{\eta_h \cdot P_{in} \cdot \alpha \cdot e^{-\alpha z}}{4\pi K} \cdot \frac{2r^2}{\omega_p{}^2} \tag{4}$$

T(r,z) is the temperature (°K) of the rod as a function of radius and length. The temperature profile is parabolic, with the highest temperature at the center of the rod.

The combined effects of the temperature and stress distributions resulting the refractive index change and end bulging, lead to the optical distortions. For a paraxial coherent beam propagating in the longitudinal direction of the rod, Koechner expressed the space-resolved optical path difference (OPD) induced by the thermal effects as [[4]]

$$\Delta OPD = \int \frac{dn}{dT} \cdot [(T(r,z) - T(0,z)]dz + n_0 \Delta l(r) + \sum_{i,j=1}^{3} \int \frac{\partial n}{\partial \varepsilon_{ij}} \cdot \varepsilon_{ij} dz \,, \tag{5}$$

where the first term is caused by the thermal dispersion $\frac{dn}{dT}$, the second term results from the change induced by the axial end bulging Δl, and the third term represents the strain-induced birefringence with the strain tensor ε_{ij}, n_0 is the refractive index of the rod, $\frac{dn}{dT}$ is the thermo-optic coefficient, and $\frac{\partial n}{\partial \varepsilon_{ij}}$ is the elasto-optic coefficient.

Here, the thermally induced lens is assumed to be an ideal thin lens. Based on the theorem of light transmission optics, the focal length (f_{th}) and focal power (D) can be determined by the optical path difference of the wave passing the lens as [[11]]

$$f = -\frac{r^2}{2\Delta OPD}, \tag{6}$$

$$D = -\frac{2\Delta OPD}{r^2}, \tag{7}$$

The thermal lens power under continuous wave (cw) lasing condition by considering energy transfer upconversion (ETU) is given by [[12]]

$$D_0 = D_{th} - D'_{th} + \frac{P_{in}\xi_Q}{2\pi\omega_p{}^2 K}\left(\frac{dn}{dT} + \alpha_T(n-1)\frac{l_b^*}{l_b}\right), \tag{8}$$

$$D'_{th} = \frac{P_{th}\xi_Q}{2\pi\omega_p{}^2 K}\left(\frac{dn}{dT} + \alpha_T(n-1)\frac{l_b^*}{l_b}\right) \tag{9}$$

where P_{th} is the threshold pump power, and $D'_{th}=D_{th}$ when the upconversion is very weak, which applies to YAG and YVO_4 lasers. The inversion density and ETU are clamped at the threshold value, therefore, D_{th} is expressed as

$$D_{th} = D_{\max}\left[1 - \frac{2(1-\xi_Q)}{\beta}\left(\left(2(\sqrt{1+\beta}-1\right) + \ln\left(\frac{4}{\beta}\frac{\sqrt{1+\beta}-1}{\sqrt{1+\beta}+1}\right)\right)\right], \tag{10}$$

where D_{max} is the maximum dioptric power under operating conditions where upconversion is very strong (i.e., $\beta\to\infty$) so that all of the absorbed pump is converted to heat, β is a dimensionless parameter given by

$$\beta = \frac{8WP_{th}\alpha\tau_f{}^2}{\pi\omega_p{}^2 h\nu_p}, \tag{11}$$

where τ_f is the fluorescence lifetime of the upper laser level, W is the upconversion rate parameter, and ν_p is the pump laser frequency. The absorption coefficient is determined by the central pump wavelength of laser diodes and the Nd doping concentration, which is given by

$$\alpha = \alpha_0 \frac{\rho}{\rho_0}, \tag{12}$$

where α_0 and ρ_0 are the standard absorption coefficient and doping concentration, respectively.

3. Interferometric Measurments

The OPD was measured using a modified Twyman-Green laser interferometer as shown in Figure 2. A frequency stabilized He-Ne laser was aligned as an interferometric probe beam at 632.8nm passing through the laser crystal along the optical axis of resonator, and perpendicular to the end face of crystals. The thermal beam distortions in the crystals were quantitatively measured from interferograms. The interference occurred between the wave fronts reflected from the front (F) and the back (B) face of the crystal. Beamsplitters (BS1 and BS2) with high transmission at 1064 nm and partial reflectivity at the wavelength of 632.8 nm were used to separate the input probe beam and the output interferometric beam. The Pellicle splitters (Thorlabs, BP133) are 33:67 at 632.8 nm and over 82% transmission at 1064 nm. A 10x beam expander (BE, Edmund R55-578) was employed to expand and collimate the probe beam. The pre-CCD short pass filter (Thorlabs FES0700) with the cut-off wavelength range (420-700nm) was used to attenuate infrared irradiation of diode laser beam (808nm) and solid-state laser beam (1064nm) because the camera CCD chip is sensitive to the light in those wavelength range. The generated interference patterns were imaged by a digital CCD camera connected to a PC for calculation and analysis. The captured fringes gave the spatial distribution of wavefront distortions. The power meter (Ophir 30A-P-SH) was placed after to beam splitters to measure the IR power.

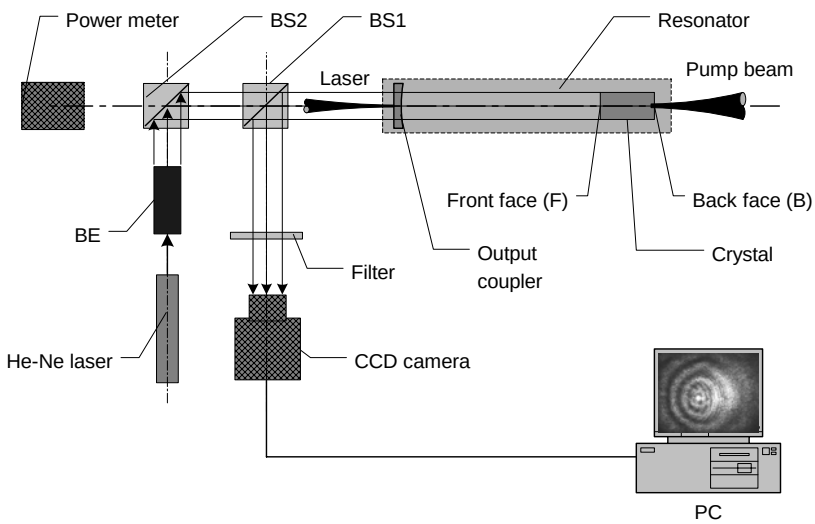


Figure 2. Schematic layout of the modified Twyman-Green laser interferometer

To verify the model, an 1.1 at.% Nd:YAG and 0.5 at.% Nd:YVO_4 crystals [[13]], [[14]], with almost the same peak absorption coefficients were chosen as examples. Therefore, over 99% of the pump power at the central wavelength of 808nm is absorbed by the crystals with the same length. The size of cylindrical Nd:YAG and cubic Nd:YVO_4 are ϕ3 mm x 5 mm and

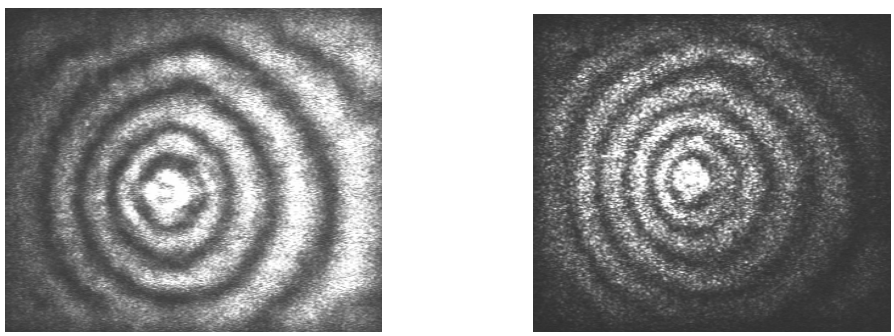
3x3x5 mm^3, respectively. Both of the crystals had a 1.06 μm AR-coating at one end, a 1.06 μm HR-coating and an 808 nm AR-coating at the pump end. The pump end face of the crystals acts as one of the cavity mirrors. The rods were end-pumped by a 30W fiber-coupled diode laser (FAP) from Coherent Inc. The fiber core is 0.8-mm-diameter with a N.A. of 0.16. The fiber output was collimated and focused into the rod in a spot size of 0.8-mm-diameter. The laser cavity length was 100 mm. The reflectivity of the output coupler was 85%. The crystals, wrapped in indium foil, were mounted on a water-cooled copper block.

Typical fringe patterns taken from the series of interferograms are shown in Figure 3 (a) and (b), for the Nd:YVO_4 and Nd:YAG lasers under lasing conditions with a pump power of 15W, respectively. It is apparent that the Nd:YAG crystal has more serious thermal lensing effects than the Nd:YVO_4 crystal because it shows higher order of fringes at the same pump power and pump beam spot size.

The measured interferograms were used to determine the space-resolved OPD in the crystals by counting the number of fringes. The total optical path difference (OPD) induced by changes in both refractive index and physical length L of the crystals can be expressed as

$$OPD_{a/c} = n_{a/c}\Delta L_{a/c} + L\Delta n_{a/c} \tag{13}$$

where $\Delta n_{a/c}$ and $\Delta L_{a/c}$ are the temperature-dependent refractive index change and end bulging along a and c axis, respectively. For symmetric crystals like Nd:YAG, the OPD will be symmetric as well. Apparently, the measured space-resolved OPD presented a combined effect of index parts and end effect influencing the wavefront. The circular fringes emerged one by one from the center of pump region of the crystal when increasing pump powers. The spacing of two adjacent fringes is closer and becomes thinner to the edge.



(a) (b)

Figure 3. Captured interferograms of the end-pumped edge-cooled Nd:YVO_4 and Nd:YAG with a probing wavelength of 632.8nm. (a) Nd:YVO_4; (b) Nd:YAG

A numerical filter was implemented during our post image processing to reduce speckle noise. Figure 4 shows the improved interferogram and optical density distributions of fringe patterns across a and c axis of the Nd:YVO_4 crystal. By calculating the optical density distributions of the fringe pattern, the space-resolved OPD_Ms along a and c axes can be determined.

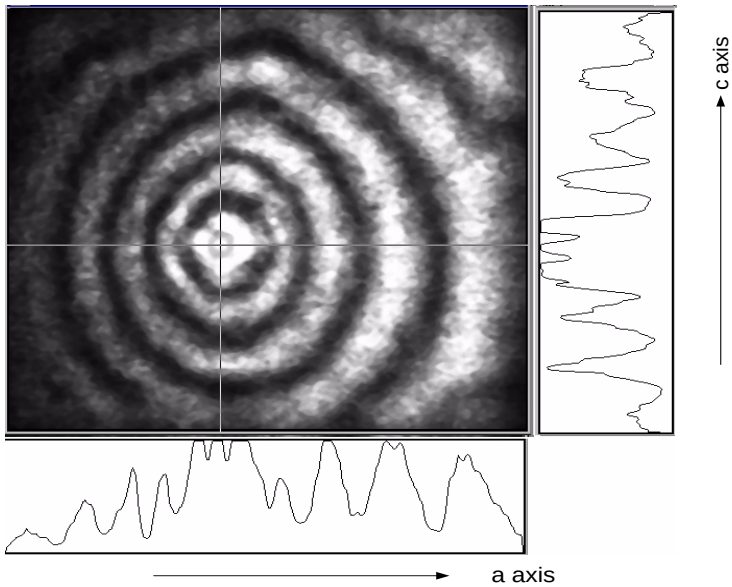


Figure 4. Optical density distributions of fringe patterns across a and c axes for Nd:YVO_4

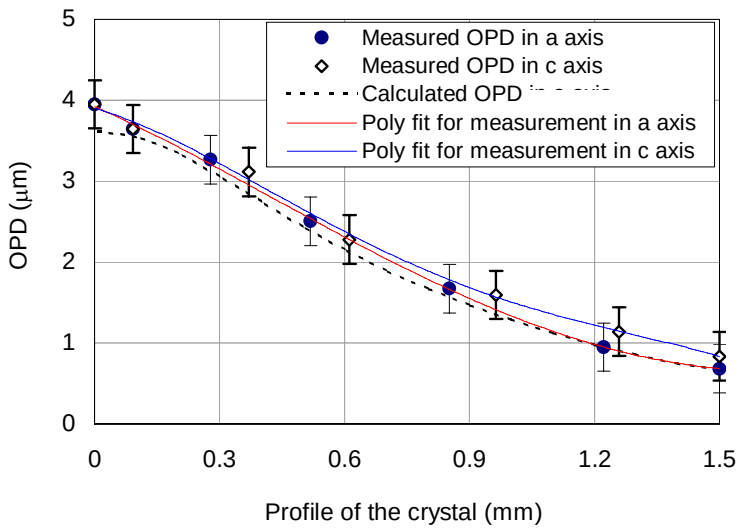


Figure 5. Measured space-resolved OPD_M (•) and (◊) compared with those calculated OPD_T (- - -) for Nd:YVO_4 (0.5 at.% Nd doped, 3x3x5mm) under pump power of 15W and pump radius of 0.4mm (measured OPD_Ms are fitted in 4 -order polynomial) for Nd:YVO_4

For Nd:YVO_4, the measured OPD_Ms obtained from Figure 4 are fitted to a 4th-order polynomial as solid lines in Figure 5. To facilitate the comparison with the calculated OPD_Ts, the measured OPD_Ms were converted from the fringe changes. The calculated OPD_Ts are also shown in Figure 5 as dash lines. It can be seen that the calculated OPD_Ts agreed well with the OPD_Ms. Similarly, we also examined the Nd:YAG modeling by using the interferometer and similar good agreement was found. Therefore, this gives enough confidence that the

analytical model produces realistic results. The reason, that measured OPD_Ms are slightly larger than the calculated OPD_Ts, results in part from neglecting the stress-dependent refractive index variation in the calculations.

The calculated thermal lensing effects as a function of pump power are shown in Figs. 6 and 7, respectively. Similar to the results of interferograms, the thermal focal powers of Nd:YAG are greater than those of Nd:YVO_4 under the same pump conditions.

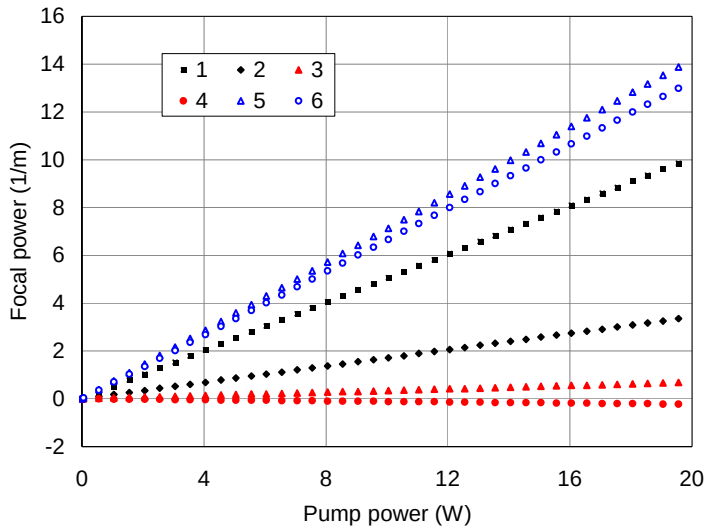


Figure 6. Focal power as a function of pump power for Nd:YAG (1.1 at.% Nd doped, ϕ3x5mm). 1- temperature-dependent focal power, 2- end effect induced focal power, 3 and 4- stress-dependent focal power along r and θ direction, 5 and 6- total focal power along r and θ direction

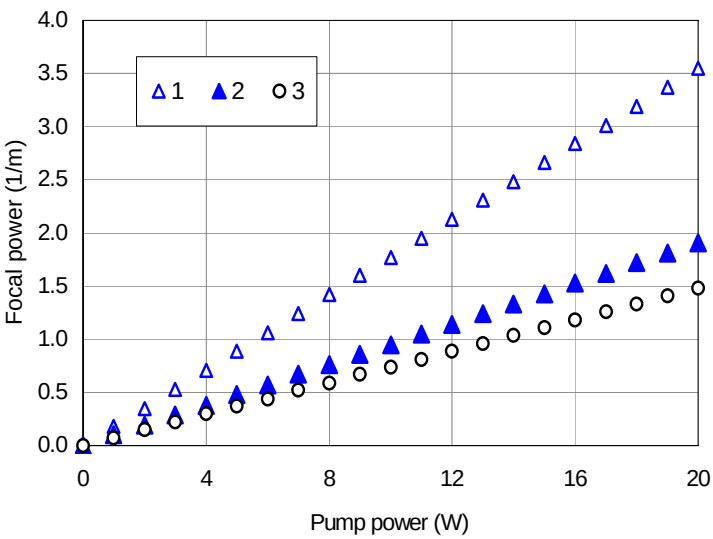


Figure 7. Focal power as a function of pump power for Nd:YVO_4 brick. 1, 2, and 3 are total focal power, index parts, and end effect, respectively

Here, in order to express the difference more clearly, we introduced a parameter called thermal lensing coefficient, "β", as the ratio of the focal power to the pump intensity [[15]]

$$\beta = \frac{D}{\frac{P_{in}}{\pi \cdot \omega_p^{\,2}}}, \tag{14}$$

where D is the thermally induced focal power. The thermal lensing coefficient for a laser material is a constant and can serve as a figure of merit for the thermal properties. The maximum TEM_{00} mode power is closely related to this coefficient β. The calculated results showed that the thermal lensing coefficients for Nd:YVO_4 and Nd:YAG were 0.21 and 0.29, respectively. Under the same pump condition, the thermal lensing of Nd:YAG is 27% more than that of Nd:YVO_4. Meanwhile we noticed that these calculated values were larger than those recently obtained by Hodgson [[16]]. This is due to the fact that only the temperature-dependent thermal focal lensing was considered in Hodgson's calculations; while the end effect was neglected. The thermal lensing effect for other crystals can thus be solved in the same way. The thermal lensing effect is an essential prerequisite for optimizing and scaling diode-pumped solid-state lasers.

4. Power Scaling

4.1. Fracture Limit

Power scaling of diode-pumped solid-state YLF and YVO_4 lasers has been hindered by the material's rather low fracture limit, which is ~4-5 times lower than that of Nd:YAG. Moreover, ETU can be a serious problem in end-pumped Nd:YLF lasers, where the combination of a tightly focused pump beam and a short absorption length for the pump light leads to very high excitation densities. The subsequent multi-photon decay following ETU leads, in some cases, to very dramatic increase in thermal loading and hence much stronger thermal effects [[12]]. Therefore, the optimal design at high pump powers for Nd:YLF and Nd:YVO_4 must take into account their inherent low fracture limit. The strategies employed to overcome the low fracture limit are to use a long crystal and distribute the pump absorption over its length either by detuning of the pump wavelength from the absorption peak or by use of a crystal with lower Nd concentration, and to reduce the radial temperature gradient by use of larger pump beam spot size.

Simply, we estimate the maximum stress by using a plane stress solution, which is given as [[17]]

$$\sigma_{\max} = \alpha_T E \frac{\eta_h P_{in} \alpha}{4\pi K}\left[1 - \frac{1}{2}\left(\frac{\omega_p}{r_b}\right)^2\right], \tag{15}$$

where K is the thermal conductivity, P_{in} is the incident pump power, E is Young's modulus, α_T is the thermal expansion coefficient, α is the absorption coefficient for the pump, and r_b is the radius of the crystal. Before implementing eqn.13, it has to be mentioned that the expression is only accurate for symmetrical crystals. For anisotropic crystals such as Nd:YLF and Nd:YVO_4, the analytical expression of the maximum stress is more intricate because of the asymmetric thermal and mechanical properties of the crystals. Normally, the maximum stress can be solved through a numerical model, which is descried in our previous work [[6]]. Nevertheless, if we can use larger values of α_T,α, and ξ, smaller value of K at the doping concentration of 1 at.% (the actual doping concentration used in our lasers will be less than 1 at.%), this will over estimate the maximum tensile stress. Consequently, if the designed maximum stress is lower than the fracture limit, the laser will operate safely for a certain pump power and beam spot size.

4.2. Optimal Design

For the optimal design of a diode-pumped solid-state laser, a key parameter, which determines the laser efficiency and output power, is the spatial overlap between the pump beam and the laser mode in the active medium. However, thermal effects introduce both thermal focal lensing and spherical aberration in the resonator. As a consequence, the laser mode is changed and diffraction losses are increased with the higher pump power. Therefore, mode matching is a critical issue for the optimal design. In addition, an optimal doping concentration exists for high-power pumped solid-state lasers [[18]]. In this chapter, the dependence of the optimal doing concentration on the pump power was investigated by using a space-dependent rate equation analysis taking into account thermally induced focal lensing and diffraction losses.

Considering the end bulging effect and index part, we modify Clarkson's expression [[19]] of mode-to-pump ratio based on the beam quality degradation, and then optimize the mode-to-pump ratio. Assuming a Gaussian pump beam, the optimum mode-to-pump ratio (m) can be simply expressed by

$$\left(\frac{\omega_l}{\omega_p}\right)^4 = M_q{}^2 \frac{K\lambda\sqrt{2}}{2P_{abs}\xi_Q\left|\frac{dn}{dT}+\alpha_T(n-1)\frac{l_b^*}{l_b}\right|}, \tag{16}$$

where, ω_l is the laser beam mode radius, λ is the laser beam, n is the reflective index, l_b^* is the length of the end section over which the expansion occurs, and M_q^2 is the degradation factor in beam quality due to the quadratic phase aberration, which is expressed by

$$M_f{}^2 = \sqrt{\left(M_i{}^2\right)^2 + \left(M_q{}^2\right)^2}\ . \tag{17}$$

It means a laser beam with a Gaussian intensity profile and initial beam quality factor M_i^2 suffers a degradation in beam quality with resultant beam quality factor of M_f^2 after

propagating through a lens which produces a phase distortion. In Eqn. 16, the temperature-dependent refractive index of Nd:YLF has an opposite sign to the end bulging. But for Nd:YVO_4 and Nd:YAG, the two effects have the same sign.

The optimization of doping concentration is based on the space-dependent equation analysis, in which we take into account thermally induced focal lensing, aberrations, diffractive losses, and the effect of ETU. The degradation of output power caused by diffractive losses strongly depends on thermally induced spherical aberration in diode-end-pumped lasers. δ, the total round-trip losses inside the cavity, can be written as

$$\delta = \delta_d + \delta_f + T, \tag{18}$$

where δ_d is the diffraction loss, T is the transmission of the output coupler, δ_f is the non-diffractive round-trip losses (such as impurity absorption, coating losses and scattering), which is defined as

$$\delta_f = \delta_o + \Delta\delta_b l_b, \tag{19}$$

where δ_o is the intracavity losses from the surface coatings of mirrors and rod, and $\Delta\delta_b$ is the scattering loss per unit length of the lasing crystals. According to the diffractive theory of aberrations [[18]], the fraction of intensity at the center of the reference sphere can be described by the Strehl intensity ratio (SR) of an incident Gaussian beam with a radius of ω_l in the presence of aberrations. With the knowledge of the SR, the thermally induced diffractive losses can be calculated by [[18]]

$$\delta_d = 1 - SR = 1 - \frac{\left| \int_0^{r_b} \exp(i\Delta\phi) \exp(-\frac{r^2}{\omega_l^2}) r dr \right|^2}{\left| \int_0^{\infty} \exp(-\frac{r^2}{\omega_l^2}) r dr \right|^2}, \tag{20}$$

where the phase shift $\Delta\phi$ induced by the wave aberration is given by [[18]]

$$\Delta\phi = -\frac{\pi D(r) r^2}{\lambda} = -\frac{\pi D_0 r^2}{\lambda} + \frac{3\pi \omega_p^2 D_0 \gamma}{2\lambda} (\frac{r}{\omega_p})^4, \tag{21}$$

where D(r) is the focal power, D_0 is the effective focal lensing power, and γ is the pump profile shape factor. The quartic phase term (fourth power of the radius) leads to deterioration of the beam quality.

Based on the space-dependent rate equations, the output power (P_{out}) as a function of incident pump power (P_{in}) and threshold (P_{th}) can be expressed by

$$P_{out} = \frac{T}{\delta}\eta(P_{in} - P_{th}) \tag{22}$$

and

$$P_{th} = \frac{I_{sat}}{\eta_p l}\frac{\delta}{2}V, \tag{23}$$

where η is the excitation efficiency, I_{sat} is the saturation intensity, V is the effective mode volume, and η_p is the pump efficiency. η is given by

$$\eta = \eta_p \eta_m, \tag{24}$$

where η_m is the mode matching efficiency and η_p is expressed by

$$\eta_p = \eta_{abs}\eta_t\eta_{st}\eta_Q, \tag{25}$$

where η_{abs} is the absorption efficiency of the pump radiation, η_t is the optical transfer efficiency of pump light from the pump source to the active gain medium, η_{st} is Stokes' efficiency, which is equal to the ratio of the pump wavelength λ_p to the laser wavelength λ, and η_Q is the quantum efficiency.

The overlap efficiency η_m related to the spatial distribution of absorbed pump density and laser mode profile inside the active medium is simply given by [[20]]

$$\eta_m = \frac{\zeta(\alpha,\omega_l)^2}{\zeta\left(\alpha,\frac{\omega_l}{\sqrt{2}}\right)}, \tag{26}$$

$$\zeta(\alpha,\omega_l) = \frac{1}{2}\frac{\alpha}{1-e^{-\alpha l_b}}\int_0^{l_b}\left(\frac{\omega_l}{\omega_p(z)}\right)^2\left[1-e^{\left(-\frac{2\omega_p(z)}{\omega_l}\right)^2}\right]e^{-\alpha z}dz. \tag{27}$$

On the basis of the paraxial approximation, $\omega_p(z)$ is given by

$$\omega_p(z) = \omega_p + \frac{\theta_p}{n}\left|z - z_0\right|, \tag{28}$$

where θ_p is the far-field half angle of pump beam, n is the refractive index of gain medium and z_0 is the focal plane.

The mode-to-pump ratio (m) is optimized for maximizing the output power in TEM_{00} mode. The optimal doping concentration of the rod and extraction efficiency is directly

related to this ratio. We will discuss the optimal model by implementing to practical laser design in the next section.

5. Results

In this section, two designs of diode-pumped Nd:YLF and Nd:YVO_4 lasers are demonstrated, where we design two typical end-pumping configurations: single- and double-end-pumping. First is a doubly end-pumped Nd:YLF laser while the second is a singly end-pumped Nd:YVO_4 laser.

If a 2.5-mm-radius Nd:YLF rod is doubly end-pumped at each end by 30W of diode laser power, the maximum stress can approximately be given by a Gaussian beam with $2\omega_p$=1-2.8mm, rod length l_b=20-40mm, and α=5cm^{-1} at the pump wavelength of 797 nm, α_T=13.3x10^{-6}/K, E=85GPa, K=5.3W/m/K, $\xi=\xi_Q+\xi_E$, in which we assume ξ_Q =24% resulting from quantum defect heating and ξ_E =8% from ETU at a cw mode [[19]]. Figure 8 shows the calculated maximum stress as a function of the pump beam radius at varied Nd doping concentration of 0.4%, 0.6%, 0.8%, and 1%, respectively. We can see that the crystal will facture due to the tensile stress close to even higher than the fracture stress when the doping concentration is equal to or higher than 0.6%. Besides low doping contraction, a larger spot size of pump beam helps to reduce the maximum tensile stress. From the effective absorption point of view, a larger pump beam associates with a smaller divergence angle, which provides a better overlap efficiency between the pump and laser modes. In our case, the large pump beam has to be restricted by the ratio of $\omega_p/r_b \leq ½$ for avoiding excessive aperture losses. As long as known the rod radius of 2.5mm, the pump beam size can be chosen as 1.25-mm-radius. Since the rod doping concentration has to be less than 0.6%, we design a rod longer than 20mm to absorb over 97% of the pump powers.

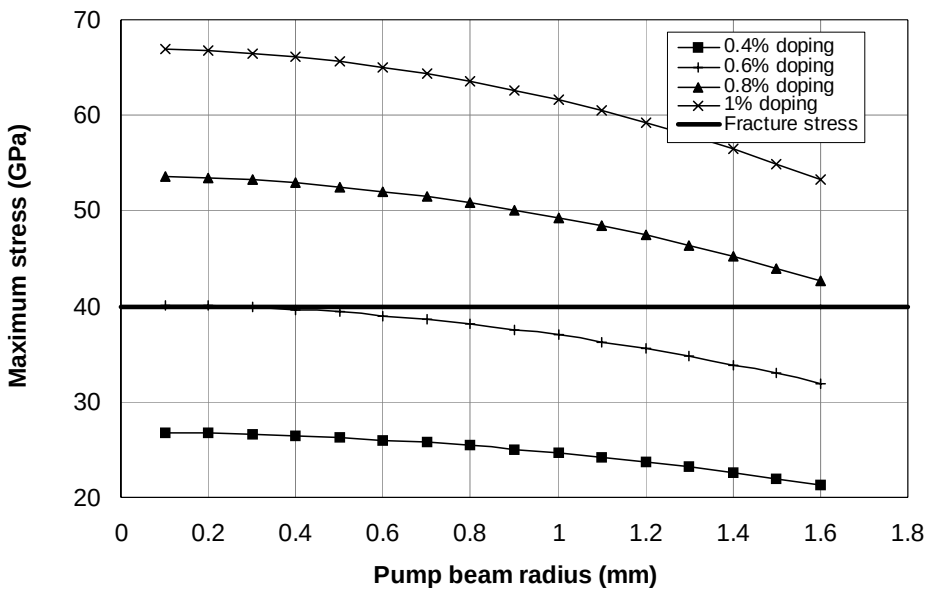


Figure 8. Maximum stress as a function of pump beam radius at different doping concentrations

If the initial beam quality factor M_i^2 is equal to 1 and a final degraded beam quality factor M_f^2 is 1.1, an estimated pump-to-mode ratio (m) can be determined to be 1.05. Therefore we can design the laser mode size of 1.2-mm-radius in the active medium corresponding to a pump beam radius of 1.25mm.

Assuming an output coupler of 12% transmission at 1053nm, δ_o=0.2%, and γ=0.1, the output power as a function of doping concentration for different rod lengths can be analytically calculated by using (18)-(28). The results were shown in Figure 9. It can be apparently seen that the doping concentration from 0.35% to 0.45% is an optimal range for Nd:YLF lasers with a rod length over 30 mm under a pump power of 60W. Based on the optimal calculations, we designed a 0.4% Nd-doped YLF rod with a length of 30mm configured with a dual-pumped geometry, by which the rod works safely due to the maximum stress relatively below the fracture stress. The laser is theoretically capable of producing about 28W output at the 1053 line in TEM_{00} mode.

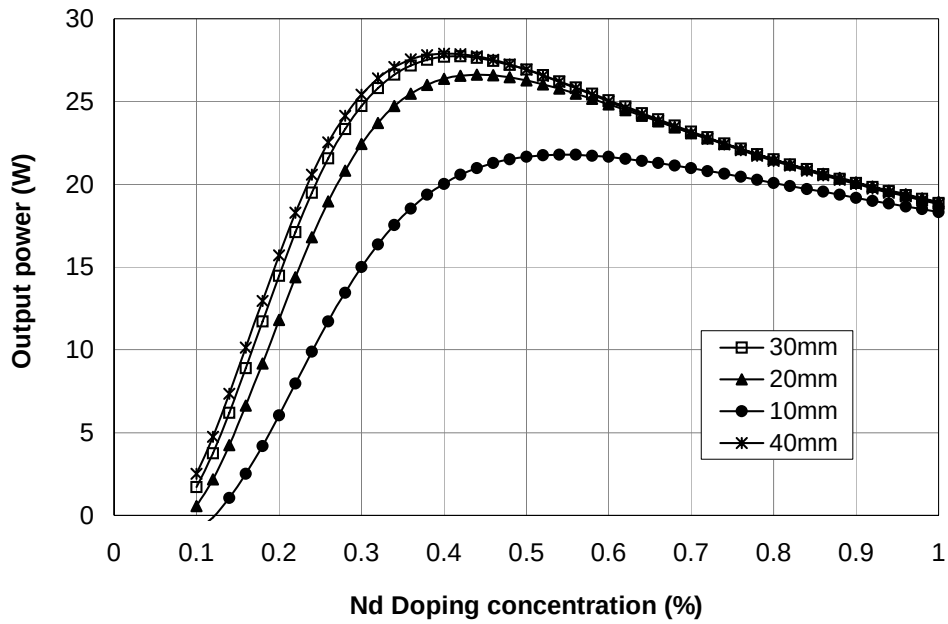


Figure 9. Output power as a function of Nd doping concentration for different rod lengths

Second example is the design of a singe-end-pumped Nd:YVO_4 laser. Assuming a single 30W diode-pumped Nd:YVO_4 laser, similarly the fracture-limited pump power as a function of doping concentration can be calculated and the results were shown in Figure 10. If the pump power is changed from 3 to 30W, correspondingly the doping concentration from 2.5 at.% to 0.25 at.% is an optimized range for vanadate lasers. Also, the vanadate crystal with 0.5 at.% Nd^{3+} doping produces the optimized output power shown in Figure 11 where the crystal is under a pump power of 30W before thermal fracture occurs. The slope efficiency under 0.5 at.% is the best doping concentration of Nd^{3+} under the pump power of 30W.

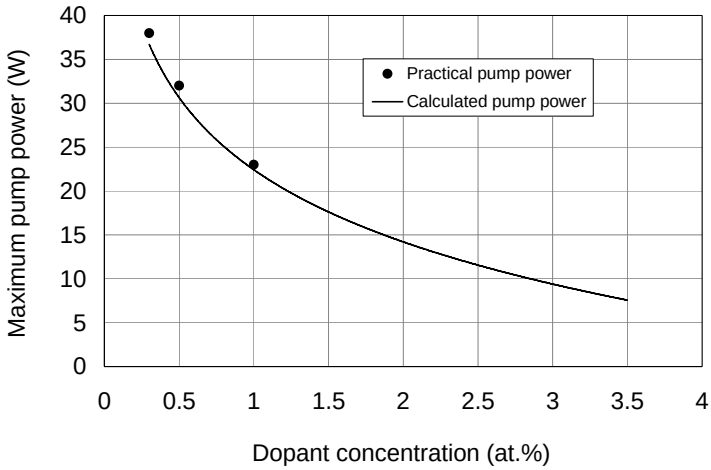


Figure 10. Calculated fracture-limited pump power as a function of dopant concentration

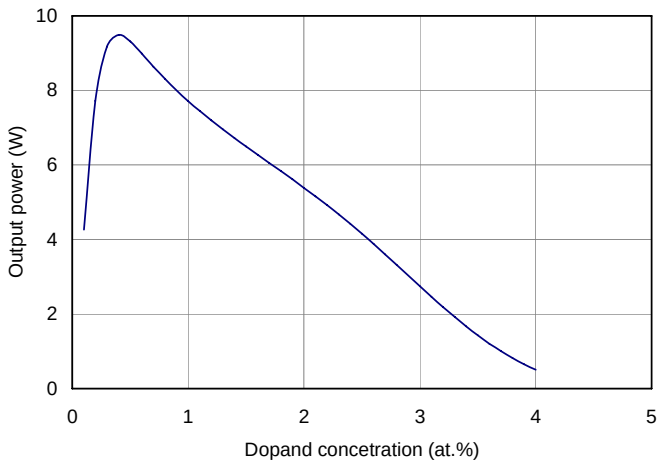


Figure 11. Output power as a function of doping concentration under the pump power of 30W

A 5-mm-long Nd:YVO_4 crystal is sufficient to absorb over 985 pump power. The longer crystal is beneficial to heat removal, but it also contributes to the intracavity absorption and scattering losses.

Figure 12 were the optimized mode-to-pump ratio as a function of pump power for the diode-end-pumped edge-cooled vanadate lasers with a 15% output coupler. The curve showed a decreasing trend with the pump power. In addition, it is worthy to note that the optimum mode-to-pump ratio is less than unity when the pump power is higher than 3W, because central portion of the highly aberrated thermal lens presents less loss to the TEM_{00} mode. It can be explained that the radial heat flow outside of the pump beam results in a strong deviation of the thermal lens from an ideal lens. Consequently, greater diffraction losses will be introduced if a mode-to-pump ratio is larger than the optimum value.

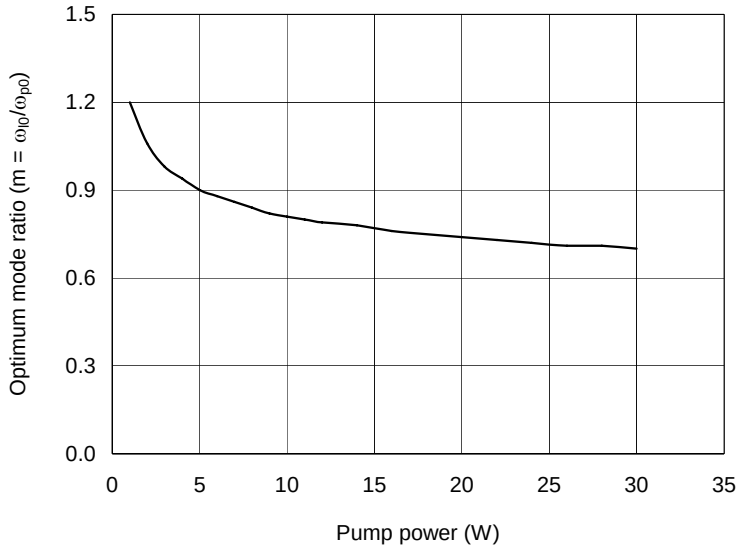


Figure 12. Calculated optimum mode-to-pump ratio as a function of pump power for vanadate lasers with a 15% output coupler

Therefore, considering a maximum pump power of 30W and 0.5 at.% Nd^{3+} doped vanadate crystal, the optimum mode-to-pump ratio corresponding to the highest slope efficiency is 0.75. The input-output curve of the vanadate laser was calculated and shown in Figure 13.

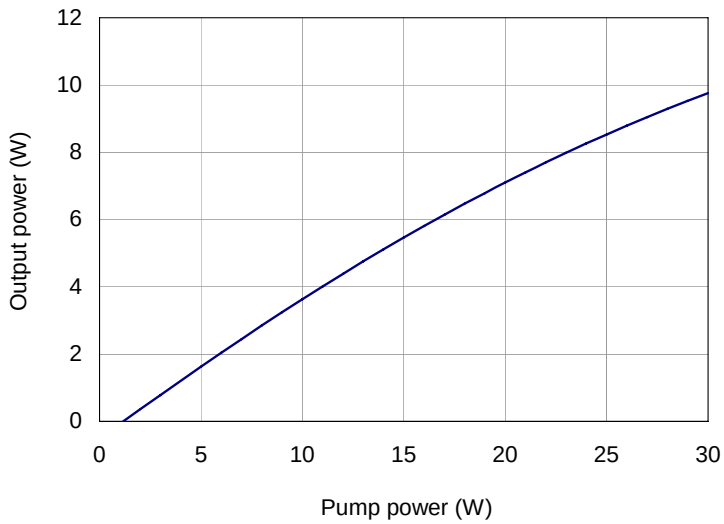


Figure 13. Calculated input-output curve for optimized vanadate lasers in TEM_{00} mode

The calculated results showed that the maximum output power of vanadate lasers in the single transverse mode (TEM_{00} mode) can extract almost 10W through a 15% transmission output coupler with the pump power of 30W by using a simple linear plano-concave resonator with the rod surface acting as an end mirror. Then, we implemented the theoretical design to a

practical vanadate laser system. The cavity length of the resonator was 100mm, and the reflectivity of the flat output coupler was 85%. An output power of 9.8W was obtained in the TEM_{00} mode with the pump power of 28W. For the multimode, 12.4W of output power in M^2 factor of 4 was achieved, and the optical-to-optical efficiency was 45%. In addition, as an example, an optimal model for a 1 at.% doped Nd:YVO_4 with a pump beam radius of 0.37 based on the investigation of Chen was developed [[21]]. The pump power of 15W was below the maximum pump limit. The mode-to-pump ratio was about 0.8. The output power was 7W for the pump power of 15W. There was good agreement with the theoretical predictions.

Conclusion

In conclusion, a theoretical model to scale and optimize the diode-pumped solid-state lasers is developed. Thermal lensing effect of the gain medium is shown to play a critical role on scaling the laser output power with the diffraction-limited laser beam output and optimizing the design of DPSS lasers. The thermal effects of diode-pumped solid-state (DPSS) lasers were theoretically analyzed and modeled for Nd doped lasing materials like YVO_4, YAG and YLF. The models determined the thermal lens and separated the end effect from the thermal effects. To validate the model, a new modified Twyman-Green interferometer was developed to measure the thermal effects directly and accurately. The thermal lensing for both isotropic (Nd:YAG) and anisotropic (Nd:YVO_4) crystals are measured. Similar to the analytical results, the thermal lensing of Nd:YAG is 27% more than that of Nd:YVO_4 under the same comparable conditions. The model showed good agreement with the measurement. Based on the calculated thermal lensing, a power-scaling model was setup to guide the development of DPSS lasers. With the consideration of the fracture limit of lasing crystals, TEM_{00} mode output, ETU, and pump power, the model optimized the mode-to-pump ratio, doping concentration and dimensions of the gain medium, and eventually predicted the output power of DPSS lasers. We demonstrated two typical examples of designing diode-pumped solid-state lasers with single- and double-end-pumping configuration. We designed a doubly end-pumped Nd:YLF laser with a total pump power of 60W. We optimized the doping concentration in a range of 0.35% - 0.45%, rod size of ϕ5mmx30mm, and pump-to-mode ratio of 1.05 to scale the fundamental output power over 28W where the laser mode size in the rod was 1.2-mm-radius. Also we designed a singly end-pumped Nd:YVO_4 laser with a total pump power of 30W. The optimized range of dopant concentration for the vanadate lasers was from 0.25 to 2.5 at.% . Consequently, the maximum pump power could be as high as 40W for a pump beam radius of 0.4 mm and crystal dimension of $3x3x5mm^3$. The optimum mode-to-pump ratio was reduced while the pump power increased. Besides, the optimum mode-to-pump ratio is less than unity for pump powers greater than 3W. For the pump power of 30W and 0.5 at.% Nd^{3+} doping vanadate crystal, the optimum mode-to-pump ratio was 0.75 corresponding to the highest slope efficiency. The maximum output power of the vanadate laser in the TEM_{00} mode was 10W with the pump power of 30W for a simple linear cavity when one of the rod surfaces acted as a resonator mirror. A practical diode-pumped Nd:YVO_4 laser system was demonstrated based on one of above modeling examples. A 9.8W output power in the TEM_{00} mode is obtained for a pump power of 28W. The theoretical analyses are in good agreement with the experimental results. To be concluded,

the models are useful for generally guiding the design and development of most DPSS lasers, including rare earth ions doped YAG, YVO_4, and YLF, etc.

References

[1] D. L. Sipes, Highly efficient neodymium yttrium aluminum garnet laser end-pumped by semiconductor laser array, *Appl. Phys. Lett*., **47**(2), pp.74-76, 1985.

[2] S. R. Scheps, Efficient laser diode pumped Nd lasers, *Appl. Opt.*, **28**(1), pp. 89-91, 1989.

[3] Minassian, B. Thompson, and M. J. Damzen, Ultrahigh-efficiency TEM_{00} diode-side-pumped Nd:YVO_4 laser, *Appl. Phys. B: Lasers and Opt.*, **76**, pp. 341-343, 2003.

[4] W. Koechner, *Solid-state Laser Engineering* (Third Edition), Springer-Verlag, 1992.

[5] X. Peng, A. Asundi, Y. Chen, and J. Xiong, "Heating measurements in diode-end-pumped Nd:YVO_4 lasers," *Opt. Eng.*, vol. 40, no. 6, pp. 1100–1105, 2001.

[6] X. Peng, A. Asundi, Y. Chen, and Z. Xiong, Study of the mechanical properties of Nd:YVO_4 crystal by using laser interferometry and finite-element analysis, *Appl. Opt.*, **40**(9), pp. 1396-1403, 2001.

[7] S. C. Tidwell, J. F. Seamans, M. S. Bowers, and A. K. Cousins, Scaling CW diode-end-pumped Nd:YAG lasers to high average powers, *IEEE J. of Quan. Electron*., **28** (4), pp.997-1009, 1992.

[8] H. S. Carslaw, and J. C. Jaeger, *Conduction of Heat in Solids* (2^{nd} edition), Oxford: Clarendon, 1986.

[9] T. M. Baer, W. L. Nighan, and M. S. Keirstead, Modeling of end-pumped, solid-state lasers, CLEO'93, *OSA Tech. Dig. Series*, **11**, pp.638, 1993.

[10] M. E. Innocenzi, H. T. Hura, C. L. Fincher, and R. A. Fields, Thermal modeling of continuous-wave end-pumped solid-state lasers, *Appl. Phys. Lett.*, **56** (19), pp.1831-1833, 1990.

[11] R. Weber, B. Neuenschwander, M. Macdonald, M. B. Roos, and H. P. Weber, Cooling schemes for longitudinally diode laser-pumped Nd: YAG rods, *IEEE J. of Quan. Electron.*, **34** (6), pp. 1046-1053, 1998.

[12] W. A. Clarkson, "Thermal effects and their mitigation in end-pumped solid-state lasers", *J. Phys. D: Appl. Phys*. **34**, 2381-2395 (2001).

[13] SYNOPTICS, Synthetic crystals and optical products catalog, 2000.

[14] Casix, Crystal guide'99, 1999.

[15] X. Peng, A. Asundi, Y. H. Chen, and Z. Xiong, Studies of thermal effects of diode-end-pumped vanadate lasers, *Asia Journal of Phys.,* **10** (1), pp. 79-86, 2001.

[16] N. Hodgson, K. D. Griswold, W. A. Jordan, S. L. Knapp, A. A. Peirce, C. C Pohalski, E. A. Cheng, J. Cole, D. R Dudley, A. B. Petersen, and W. L. Nighan, High power TEM_{00} mode operation of diode-pumped solid-state lasers, *SPIE proceedings*, vol.3611, 119-131(1999).

[17] L. Yan, and C. H. Lee, Thermal effects in end pumped Nd:Phosphate glasses, *J. Appl. Phys.* **75** (3) 1286-1292 (1994).

[18] X. Peng, L. Xu, A. Asundi, Power scaling of diode-pumped Nd:YVO_4 lasers, *IEEE J. Quantum Electron.* **38** (9), 1291-1299 (2002).

[19] W. A. Clarkson, Thermal effects and their mitigation in end-pumped solid-state lasers, *J. Phys. D: Appl. Phys*. **34**, 2381-2395 (2001).

[20] Y. F. Chen, T. M. Huang, C. L. Wang, L. J. Lee, S. C. Wang, Theoretical and experimental studies of single-mode operation in diode-pumped Nd:YVO_4/KTP green laser: influence of KTP length, *Opt. Commun*. **152**, 319-323 (1998).

[21] Y. Chen, T. Huang, C. Wang, and L. Lee, Compact and efficient 3.2-W diode-pumped Nd:YVO_4/KTP green laser, *Appl. Opt.*, **37** (24), pp. 5727-5730, 1998.

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN: 1-60021-194-1
Editor: Steven B. Larkin, pp. 193-212

Chapter 6

CATASTROPHE OPTICS IN THE STUDY OF SPREADING OF SESSILE DROPS

Nengli Zhang
Ohio Aerospace Institute at NASA Glenn Research Center,
Cleveland, OH 44135, USA
David F. Chao
Microgravity Division, NASA Glenn Research Center,
Cleveland, OH 44135, USA
John M. Sankovic
Science Division, NASA Glenn Research Center,
Cleveland, OH 44135, USA

Abstract

Optical catastrophe occurring in far field of waves of drop-refracted laser beam implies a wealth of information about drop spreading. When a parallel laser beam passes through a sessile drop on a slide glass, the rays are refracted by the drop and produce a new wave field in space after the drop. Very interesting optical image patterns occur on a screen far from the drop. Stable spreading-and-evaporations of sessile drops on an isothermal and isotropic slide-glass surface give nearly perfect circular far-field sectional images of the drop-refracted laser beam on the distant screen with a bright-thick ring and a set of fringes. The image configurations are time dependent and sensitive to the drop properties. In some cases of evaporating drops, rapidly varying optical image patterns occur on the screen. Classical geometric optics cannot correctly interpret these optical image patterns. An analysis based on catastrophe optics reveals and interprets the formation of these optical image patterns. The circular caustic line manifested as the bright-thick ring on the screen implies that the caustic is the lowest hierarchy of optical catastrophes, called fold caustic, which is produced from the initial wavefront with a local third-order surface given by the drop refraction. The set of fringes next to the fold caustic is the caustic diffraction instead of an aperture effect misunderstood by many investigators. The caustics and their diffraction fringes on the far-field sectional images of the drop-refracted laser beam can sensitively detect the interface instability of sessile drops, measure real-time local contact angles, estimate the drop-foot heights, visualize thermo-capillary convection flows, and investigate the spreading

characteristics of evaporating liquid drops, such as drop profiles, effects of evaporation and thermo-capillary convection flows on the spreading, etc. The successful applications of catastrophe optics to the study of drop spreading convincingly show that catastrophe optics is not only able to qualitatively explain optical phenomena in nature but also able to quantitatively measure physical parameters of the objects of study, such as spreading drops.

Introduction

Apply catastrophe theory to geometrical optics to understand ubiquitous phenomena of natural focusing of rays results catastrophe optics. As for catastrophe theory, it is a new mathematical method to describe the evolution of forms in nature, which is particularly applicable where gradually changing forces produce sudden effects called catastrophes [1]. As an elementary dynamic system, geometrical optics based on Fermat's principle is one of the typically applicable fields of catastrophe theory. The objects of catastrophe optics are light caustics produced by focusing of light rays in the absence of symmetry. Differing from the perfect symmetry focusing, caustics are structural stable and generic. The conventional studies of focusing concentrate on perfect symmetry to produce perfect point focus that never exists in practice [2]. In any real system of optical focusing, because of inevitable manufacturing defects, the rays focus so as to generate a bright surface, which is treated as aberration in classical geometrical optics but is recognized as a caustic in catastrophe optics. Caustics have been awkward to fit into optical theory while the short-wave theory fails to predict the unphysical intensity divergences of caustics. The catastrophe optics dramatically changes this unsatisfactory state of affairs [3]. The establishment of catastrophe optics is mainly attributed to the theoretical and experimental studies performed by Berry [3, 4] and Nye [5]. A comprehensive discuss on the catastrophe optics has been given by Nye [6]. Although catastrophe optics successfully interprets many optical phenomena in nature but seems rare to be applied to quantitatively measure physical parameters of objects of study.

Recently, the catastrophe optics has been successfully applied to investigate spreading of evaporating liquid sessile drop [7-9]. The results of this study have demonstrated that catastrophe optics is also able to be used for quantitative measurements of physical parameters of objects of study. In this chapter, we introduce and explain the catastrophe optics applied in study of spreading of sessile drops and give examples of the application.

Optical Catastrophe of Drop-Refracted Laser Beam

As a generalized case, evaporating liquid sessile drops will be considered only, because sessile drops without evaporation are a special case of the former. When a parallel laser beam passes through an evaporating drop sitting on a slide glass, the rays are refracted by the drop and the slide to produce a new wave field in space after the slide. A bright-thick ring at the outmost rim and a set of fringes next to it can be observed on a distant screen, as shown in Figure 1 (slide-glass is not shown). Stable spreading-and-evaporations of a sessile drop on an isothermal and isotropic slide surface give a nearly perfect circular far-field sectional image (FFSI) of the drop-refracted laser beam on the screen with a bright-thick circular ring and a set of fringes. However, the image configurations are time dependent and sensitive to the drop shape.

Typical photos of the FFSIs of the drop-refracted laser beam are shown in Figure 2. The FFSIs of the laser beam refracted by a cyclohexane drop with initial volume V_o = 5 μl on an isothermal and on a non-isothermal slide-glass surfaces are shown in Figure 2a and 2b, respectively. The former was taken at τ = 0.6 for t_f = 88.1 s and the latter at τ = 0.61 for t_f = 85.3 s. Here τ is the reduced time, defined as the spreading-and-evaporation time, t, over the lifetime of drop, t_f, i.e. $\tau = t/t_f$. For some liquids, such as R-113, n-pentane etc., evaporation induces thermocapillary convection flows (TCCFs) in the drops. For example, an R-113 drop with V_o = 5 μl on an isothermal slide-glass surface gives a rapidly varying optical pattern all over the image with the circular bright-thick ring at outmost rim and the corresponding fringes, as shown in Figure 2c, which was taken at τ = 0.60 for t_f = 30.2 s. Adding a small amount of aluminum powder will destroy both drop symmetry and stability of spreading, as shown in Figure 2d, which was taken at τ =0.63 for t_f = 29.0 s. However, the bright-thick ring and the fringes are still clearly observed.

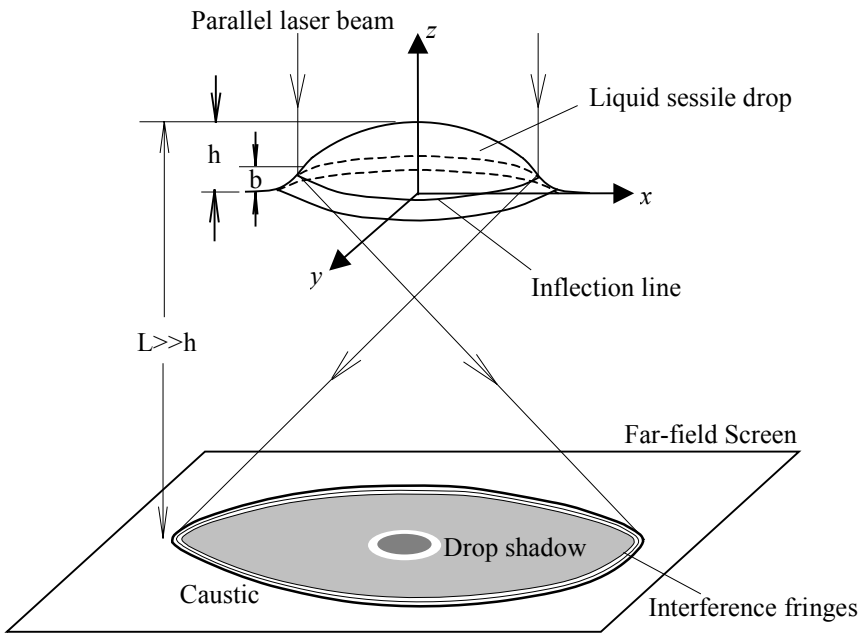


Figure 1. Far-field image of drop-refracted laser beam

Conventional geometry optics cannot appropriately explain the formation of the outmost bright-chick ring with fringes and the rapidly varying optical image patterns on the distant screen. In fact, according to the catastrophe optics, both the outmost ring with fringes shown in Figure 2 and the varying optical image patterns appearing within the ring shown in Figs. 2c and 2d are the sections of the caustic surfaces formed by focusing of the drop-refracted laser rays in far field of waves. In other words, the refracted laser rays occur optical catastrophes in far field and form three-dimensional caustic surfaces, whose sectional images are shown on the screen.

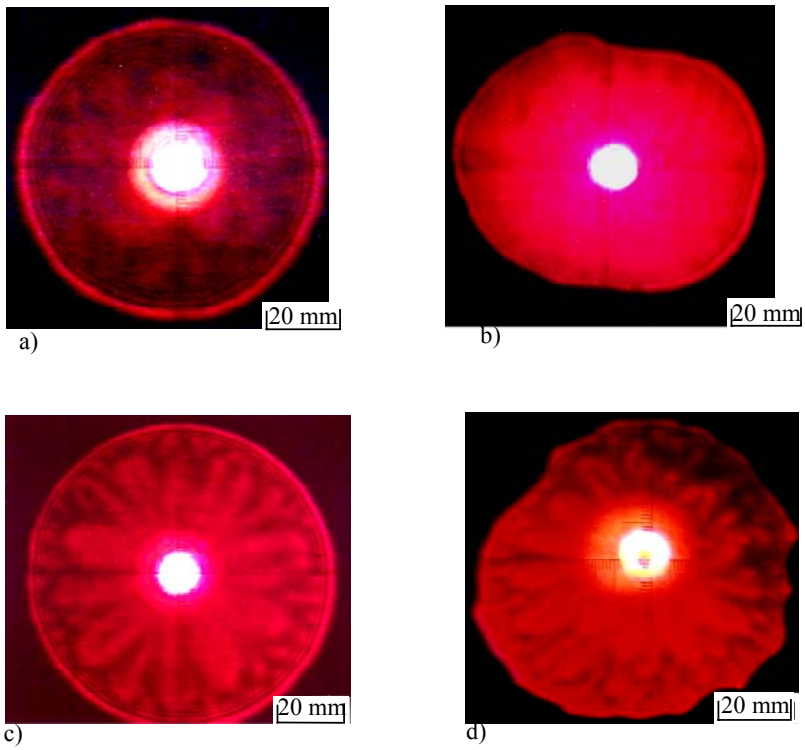


Figure 2. FFSIs of parallel laser beam refracted by: a) cyclohexane drop on an isothermal slide-glass; b) cyclohexane drop on a non-isothermal slide-glass; c) R-113 drop on an isothermal slide-glass; d) R-113 drop with aluminum powder.

Condition of Caustic Formation

Although optical catastrophes are structurally stable and generic, caustics are formed only when light goes to focus. It should be pointed out emphatically again that the focusing discussed in catastrophe optics is different from the focusing in classical geometrical optics. The unnatural focusing of light concerned by classical geometrical optics requires a point focus as the prerequisite for forming an image, while catastrophe optics studies the natural focusing that is envelopes formed by converging rays without the symmetry requirement for optical imaging. For an originally uniform plane wavefront of light, as long as it becomes a wavy wavefront after deformed by a phase screen (a refracting medium), caustics will be formed by focusing of the light. However, neither every ray would participate in the forming of caustics and nor caustic exists in everywhere.

Figure 3 shows that a uniform plane wavefront of a monochromatic light (parallel laser beam) travels downwards to pass through a sessile liquid drop sitting on a slide glass (not shown in the figure). The parallel laser beam is refracted by the drop-slide system (a phase screen), and so produces a new field of waves. A new initial wavefront of the refracted beam emerges immediately after the slide. If both the drop and the slide are homogeneous and the slide surface is smooth enough and isotropic, the initial wavefront, W_o, has the similar shape as the drop surface. A general point on the initial wavefront, P, has coordinates (x, y, f) in

Cartesian coordinate system or coordinates (r, f) in cylindrical coordinate system, while the corresponding point in the observation space has coordinates (X, Y, Z) or (R, Z). Generally, in the cases of axisymmetric system, cylindrical coordinates are more convenience. In view of the sessile drops studied in drop spreading are very small, the paraxial approximation and far field observation reasonably give the following conditions, respectively:

$$[Z\text{-}f(r)]^2 >> (R\text{-}r)^2 \tag{1}$$

$$Z >> |f(r)| \tag{2}$$

The distance between point P on W_o with coordinates (r, f) and the point Q in observation space with coordinates (R, Z) is

$$L = [(R - r)^2 + (Z\text{-}f)^2]^{1/2} \tag{3}$$

Consider the approximations given by Eqs. (1) and (2),

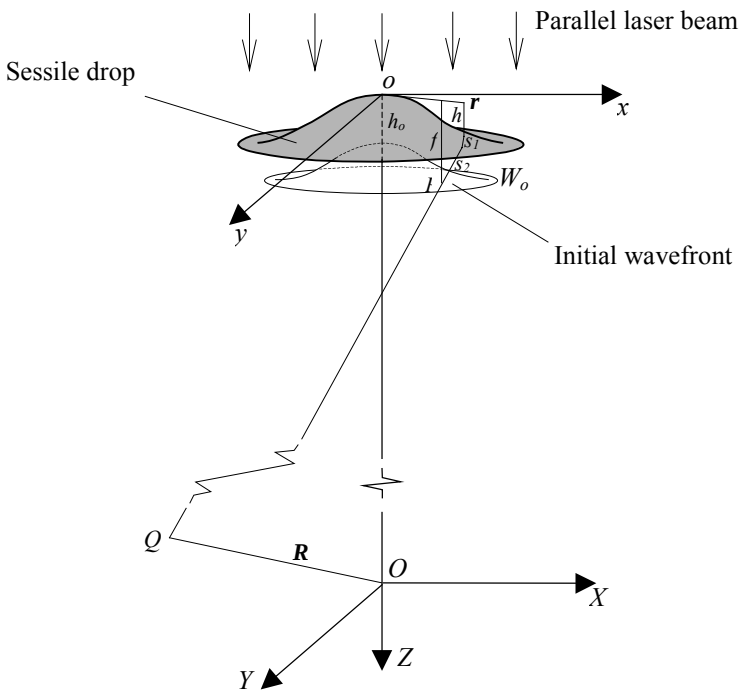


Figure 3. Initial wavefront traveling to far field.

$$L \approx Z - f + \frac{(R\sin\theta - r\sin\vartheta)^2 + (R\cos\theta - r\cos\vartheta)^2}{2Z} \tag{4}$$

where R and r are the magnitude of R and r, respectively, θ and ϑ are the angle from x-axis to r and from X-axis to R, respectively. Simple algebra leads Eq. (4) reducing to

$$L \approx Z - f + \frac{R^2 + r^2 - 2Rr\cos(\theta - \vartheta)}{2Z} \tag{5}$$

According to Fermat's principal, the condition of rays passing through a given point in observation space (R, Z) is $\partial L/\partial r = \partial L/\partial \vartheta = 0$. From Eq. (5) the ray conditions give

$$\frac{\partial L}{\partial r} = -\frac{\partial f}{\partial r} + \frac{r}{Z} - \frac{R}{Z}\cos(\theta - \vartheta) = 0 \tag{6}$$

$$\frac{\partial L}{\partial \vartheta} = -\frac{\partial f}{\partial \vartheta} - \frac{Rr}{Z}\sin(\theta - \vartheta) = 0 \tag{7}$$

The condition to form a caustic is to exist a set of displacements dr, $d\vartheta$ of the rays for which $dR = d\theta = 0$. In order to find this set of dr, $d\vartheta$, differentiate the ray conditions Eqs. (6) and (7) giving

$$\frac{\partial^2 f}{\partial r^2}dr - \frac{1}{Z}dr + \frac{\partial^2 f}{\partial r \partial \vartheta}d\vartheta + \frac{R}{Z}\sin(\theta - \vartheta)d\vartheta = 0 \tag{8}$$

$$\frac{\partial^2 f}{\partial \vartheta \partial r}dr + \frac{R}{Z}\sin(\theta - \vartheta)dr + \frac{\partial^2 f}{\partial \vartheta^2}d\vartheta - \frac{Rr}{Z}\cos(\theta - \vartheta)d\vartheta = 0 \tag{9}$$

Obviously, the condition existing a solution of the coupled equations (8) and (9) for the set of dr, $d\vartheta$ is that the determinant of coefficients vanishes; i.e.

$$K = \begin{vmatrix} \dfrac{\partial^2 f}{\partial r^2} - \dfrac{1}{Z} & \dfrac{\partial^2 f}{\partial r \partial \vartheta} + \dfrac{R}{Z}\sin(\theta - \vartheta) \\ \dfrac{\partial^2 f}{\partial r \partial \vartheta} + \dfrac{R}{Z}\sin(\theta - \vartheta) & \dfrac{\partial^2 f}{\partial \vartheta^2} - \dfrac{Rr}{Z}\cos(\theta - \vartheta) \end{vmatrix} = 0 \tag{10}$$

For a caustic at infinity, i.e. when $Z \to \infty$, the determinant reduces to

$$K = \begin{vmatrix} \dfrac{\partial^2 f}{\partial r^2} & \dfrac{\partial^2 f}{\partial r \partial \vartheta} \\ \dfrac{\partial^2 f}{\partial r \partial \vartheta} & \dfrac{\partial^2 f}{\partial \vartheta^2} \end{vmatrix} = 0 \tag{11}$$

In order to better understand the implication of Eq. (11), an alternative analysis can be made. In the cases of $Z \to \infty$, only the directions of the rays and waves contribute to the infinite far field of the waves from W_o [4]. In other words, the directions of the rays would form so-called directional caustics at infinity. The normal to the initial wavefront W_o at point (r, f) is the direction

$$\Omega(r) = -\nabla f(r) = -\frac{\partial f}{\partial r} r - \frac{1}{r}\frac{\partial f}{\partial \vartheta} \theta \quad . (12)$$

Consequently, Eq. (11) also represents that the Jacobian determinant of the mapping from r to Ω vanishes, $K = |d\Omega/dr| = 0$. Where K is also considered as the Gaussian curvature of W_o at r, which equals to the product of two principal curvatures k_1 and k_2, i. e. $K = k_1 k_2$. In this way, a clearly physical meaning of Eq. (11) is that the caustics produced by the rays traveling to infinity from W_o are the singularities of the mapping r to Ω. The caustics are the images in Ω of the lines on W_o where the Gaussian curvature vanishes, called as K-vanishing line.

Interpretation of Images of Drop-Refracted Laser Beam

Most liquid sessile drops studied in drop spreading have a macroscopic portion and a microscopic portion connected asymptotically by a crossover region. The macroscopic portion usually takes the shape of a spheroid cap while the microscopic portion is a precursor film at the very edge of the liquid front with typical thickness of the order of 0.01 μm [9-11]. This kind of drop profiles certainly gives such initial wavefronts of the drop-refracted laser beam that produce caustics because of an inflection line as the K-vanishing line existing on the initial wavefronts.

It can be seen from Figure 3 that if the liquid refractive index varies in the sessile drop we have

$$f(r, \vartheta, n) \approx \int_0^{h_0} n(r, \vartheta) dh(r, \vartheta) - \int_h^{h_0} n(r, \vartheta) dh(r, \vartheta) + h_0 - h(r, \vartheta) \quad (13)$$

Here, the optical role of the slide glass is omitted because its top and bottom surfaces are considered exact parallel.

When n = constant, Eq. (13) reduces to

$$f(r, \vartheta, n) \approx (n-1)h(r, \vartheta) + h_0 \quad (14)$$

If the sessile drop is axisymmetric and homogeneous, while the slide-glass surface is isotropic, the drop surface will take a shape of a cylindrically symmetric Gaussian hill with a circular inflection line on the drop surface near the three-phase line (TPL). One of the principal curvatures of the drop surface vanishes at the circular inflection line:

$$\kappa_1 = \frac{\partial^2 h}{\partial r^2} = 0 \tag{15}$$

While the axis-symmetry of the smooth drop surface gives

$$\frac{\partial^2 h}{\partial r \partial \vartheta} = 0 \tag{16}$$

It is obvious from Eq. (14) that the initial wavefront will have a corresponding circular inflection line, where the principal curvature of the initial wavefront $k_1 = \partial^2 f/\partial r^2 = 0$ and $\partial^2 f/\partial r \partial \vartheta = 0$. It means that the direction of vanishing curvature is always radial without touching to the inflection line. Consequently, a circular directional caustic is inevitably produced at infinity. According to catastrophe optics, smooth caustic curves at infinity are called fold caustics which may be interrupted only by at cusp points where the caustic curve reverses direction and crosses its tangent [4]. It can be easy understand from that differentiate Eq. (12) and set up a local coordinates whose origin is set at the *K*-vanishing line with $\partial^2 f/\partial r^2 = 0$ and $\partial^2 f/\partial r \partial \vartheta = 0$, i.e. let the local coordinate r be parallel to the direction of the vanishing principal curvature. Moving along the *K*-vanishing line we always have

$$d\Omega(r) = -\theta \frac{1}{r} \frac{\partial^2 f}{\partial \vartheta^2} d\vartheta \tag{17}$$

Once $d\vartheta = 0$ and so $d\Omega(r) = 0$ at a point, it implies that the direction of the rays is unchanged at this point, manifested as a cusp on the corresponding caustic at infinity, which is called cusp caustic.

Actually, caustics may occur in full three-dimensional space after the initial wavefront without the limitation of "at infinity". Under the assumptions given by Eqs. (1) and (2), the condition to form the caustics is described by Eq. (10). In other words, even with a finite distance the rays still define a mapping between the initial wavefront and the observation space. Now not only the directions of the rays but also the traveling distances of rays contribute to the formation of caustics. On a screen placed at a distance from the drop, a two-dimensional sectional image of the caustics can be observed, just as shown in Figure 2. An almost perfect circular caustic line appears in Figs. 2a and 2c, while a non-circular caustic line occur in Figs. 2b and 2d. The perfect circular caustic line indicates the initial wavefront is shaped like the perfect Gaussian hill by the drop refraction and consequently the drop surface is inferred to be the shape of a perfect Gaussian hill. Theoretically, the inflection line on the initial wavefront maps to infinity to form a circular directional caustic. In fact, the circular caustic images on the distant screen shown in Figs. 2a and 2c are produced by a circular line in close vicinity to the inflection line on the initial wavefront with $\partial^2 f/\partial r^2 = 1/Z$, which can be called quasi-inflection line. In other words, the *K*-vanishing line is no longer the inflection line but varies with the distance between the initial wavefront and the FFSI. It can be seen clearly from Eq. (10). For the axisymmetric initial wavefront, $\theta = \vartheta + \pi$, $\partial^2 f/\partial r \partial \vartheta = 0$, and $\partial^2 f/\partial \vartheta^2 + Rr/Z \neq 0$. Therefore, the condition to form caustic is $\partial^2 f/\partial r^2 - 1/Z = 0$.

When moving the distant screen up or down similar caustic images are observed just with different diameters due to the different distance between the initial wavefront and the screen, on which the caustics are mapped from different *K*-vanishing lines. Obviously, the caustics are of the spatial asymptotes. In the cases of sessile drops the caustics are horn-like surface opening out toward infinity. The caustics in shape of a smooth surface in space are classified as fold caustics by the catastrophe theory [1-6], whose sectional images on the distant screen are smooth lines, including non-circular lines, just as shown in Figure2a to c. The cusps on the fold caustics are classified as cusp caustics, which can also be observed at the screen, as shown in Figure 2d.

When a liquid drop is placed on a non-isothermal slide surface, the drop spreads with different rates at different direction. The drop surface is no longer the shape of a Gaussian hill but a deformed hill shape. The FFSI shown in Figure 2b indicates that the drop has an elliptic quasi-inflection line in the crossover region of the drop surface. In view of the fact that the caustic is manifested as a smooth curve on the distant screen, it is considered a fold caustic. According to catastrophe theory, the initial wavefront mapping to far-field space to produce a fold caustic must locally be a third-order surface. Therefore, the FFSIs shown in Figs. 2a to 2c clearly indicate that the tested drops have local three-order surfaces at their quasi-inflection lines. However, Figs. 2c and 2d that imply higher hierarchy caustics are mapped from more complex initial wavefronts.

The TCCFs in a liquid drop induced by evaporation changes the local liquid densities with time due to the varying temperature field and so the drop-refracted laser beam creates a rapidly varying initial wavefront. Because density and refractive index of liquids are temperature dependent, when the temperature of a liquid drop locally rises, the drop becomes locally less dense and the refractive index locally decreases. From Eq. (13) we can see that the initial wavefront profile is a function not only of the drop profile but also of the refractive index. It is inevitable that the refractive index will locally vary with time in the cases of TCCFs occurring in the drops. Consequently, in these cases the initial wavefronts rapidly change their profiles with time. It is the varying wavefronts that map the rapidly moving caustics to the observation space, indirectly displaying the cellular flow patterns, as shown in Figs. 2c and 2d. The effects of the locally varying refractive index on the initial wavefronts must be soft because large change of refractive index is impossible, especially in the “drop foot” region where the liquid is very thin. Therefore, the local variations of the wavefronts only occur at the portion above the quasi-inflection line and just slightly deform the wavefronts. That is why the caustics displaying the flow pattern occur only within the outmost circular caustic line, which is mapped from the quasi-inflection line where gives maximum refracted angle. It is also noted from Figure 2d that several cusps occur on the outmost caustic line. The cusps indicate that the smoothness of drop surface is damaged near the quasi-inflection line, producing several locally higher-order surfaces. The rapidly varying caustics must be some higher singularities for their four dimensions in control space, i.e. the codimension is four (the dimensionality of space-time) [6]. Unfortunately, we cannot determine their exact type of optical catastrophes because of unknown wavefront function. However, comparing Figs. 2c and 2d, we can confidently give the following conclusions that are useful to the study of drop spreading:

(1) The TCCFs locally changes the refractive index with time and so changes the initial wavefront with time but does not affect on the drop profile.
(2) The small amount of aluminum powder does not retard the occurrence of TCCFs induced by evaporation but greatly changes the drop spreading stability and the drop profile.
(3) The aluminum powder breaks the isotropy of the drop spreading and the drop-surface smoothness, which are manifested as the non-circular caustic line and occurring of several cusps.

Caustic Diffractions

The information of caustic diffractions is also provided by the images on the distant screen. From the view point of catastrophe optics, each catastrophe of geometrical optics is accompanied by its own characteristic diffraction pattern [6]. The fringes next to the caustic line on the screen, as shown in Figure 2, are the caustic diffractions. They are quite different from the aperture effect caused by sharply cut-off of the initial wavefront because the TPL of sessile drops and so the initial wavefront created by the drop refraction are asymptotic profiles instead of sharp edges. Diffraction occurs because of a lack of flatness in the initial wavefront, rather than the cut-off edge. The caustic diffraction can also be well described using catastrophe theory. The catastrophe modeling procedure gives a canonical diffraction catastrophe [2].

At the vicinity of caustics the wave intensity shows dramatic diffraction effects. For simplicity, just consider one-dimensional wavefronts. The wave disturbance at a given point in observation space can be expressed as

$$\psi = C\int_0^{\infty} L^{-1/2} e^{ikL}\, dr \tag{17}$$

where C is a complex constant, $k \equiv 2\pi/\lambda$ while λ is the wavelength.

For fold caustic-diffraction catastrophe induced by fold caustics, where the corresponding initial wavefront is a local third-order surface, represented by $f = pr^3/3 + \xi r$. It has been proved that the wave disturbance, ψ, near the caustic is proportional to the Airy function defined as

$$Ai(s) = \frac{1}{\pi}\int_0^{\infty} \cos(\tfrac{1}{3}t^3 + st)dt \tag{18}$$

where $s = k^{2/3}p^{-1/3}\xi$ and $t = (kp)^{1/3}\, r$, while p is a constant and $\xi = R/Z$. Here, it is noted that for the mapping between an axisymmetric initial wavefront and the FFSI on the screen, the angle difference $\theta - \vartheta = \pi$ in Eq. (6), and so $\xi = \partial f/\partial r - r/Z = R/Z$.

The modulus squared of the Airy function is an oscillation curve with the parameter s, as shown in Figure 4. The theoretical oscillations of intensity on the bright side of the caustic explain the light and dark fringes seen running parallel to the fold caustics [6].

Higher Hierarchy of Optical Catastrophe

Different hierarchies of caustics imply different spreading characteristics of liquid drops. In the study of spreading of evaporating sessile drops, the lowest two hierarchies of caustics, i.e. fold and cusp caustics, are most concerned. Cusp caustics imply the lack of flatness in the initial wavefront, usually manifested as a saw-tooth-like caustic line consisting of a lot of cusps. Typical photos of the sectional images of the cusp caustics are shown in Figure 5. When the cusps only occur at the outmost caustic line, as shown in Figure 5a, the drop profile and the corresponding profile of the initial wavefront ripple only at the quasi-inflection line. If the cusps occur both at and within and/or out of the outmost caustic line, such as shown in Figure 5b, it can be inferred that the drop ripples both at the quasi-inflection line and other places in the macroscopic portion of the drop surface above the quasi-inflection line.

More complex optical catastrophe patterns will occur in highly unstable spreading, or in the cases of the drops with TCCFs induced by evaporation, just as mentioned above.

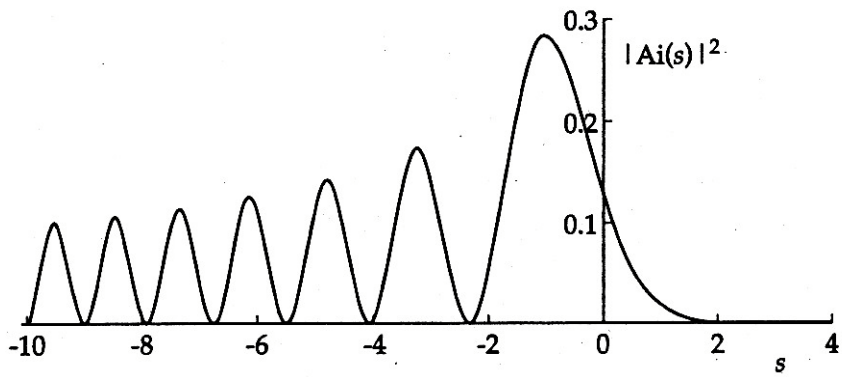
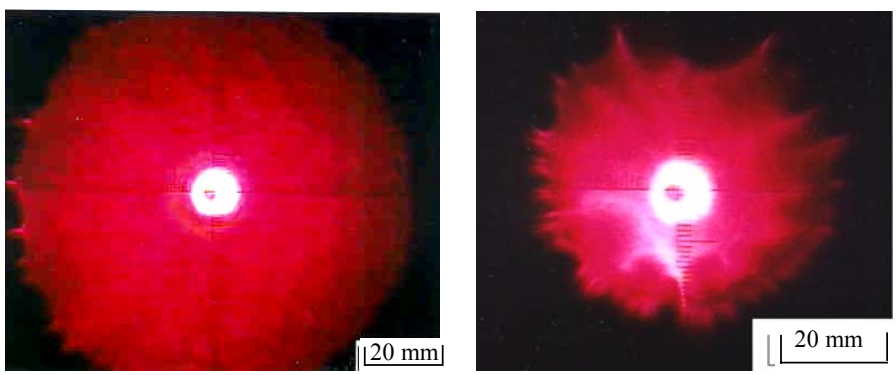


Figure 4. The modulus squared of the Airy function



a) b)

Figure 5. FFSIs of parallel laser beam refracted by an ethanol drop of 2.5 μl on an isothermal slide glass (t_f = 107 s) at: a) τ = 0.01; b) τ = 0.3.

Determine Sessile Drop Profiles by Catastrophe Optics

For different liquids and/or at different times during spreading-and-evaporations of liquid drops, the profiles of the drops are different. Four basic profiles have been revealed by experiments: spheroid cap with precursor film, named as C-P-type profile; spheroid cap with irregular 'protruding foot', as C-I-type profile; spheroid cap with smooth 'protruding foot', as C-PF-type profile; and perfect spheroid cap, as P-C-type profile, as shown in Figure 6. Spreading of an evaporating ethanol drop gives a good example of the changing profiles as the spreading-and-evaporation evolves.

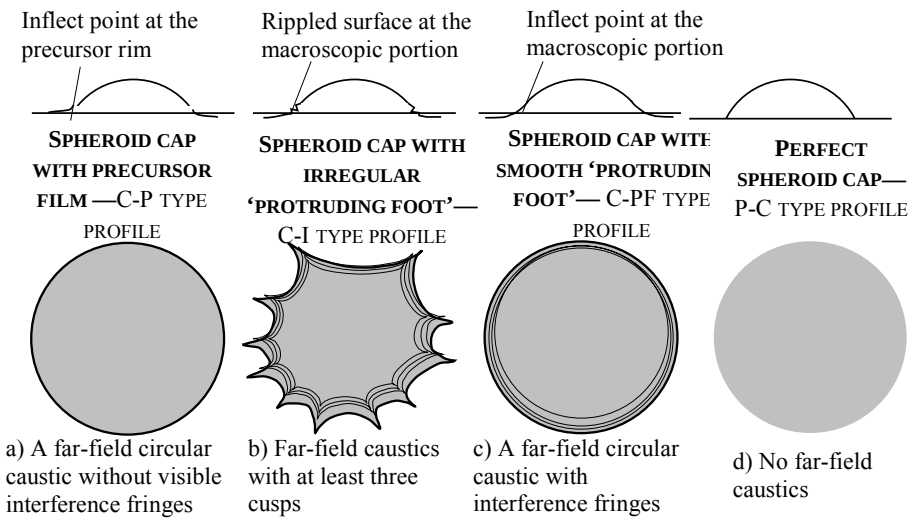


Figure 6. Different shapes of initial wavefront and their corresponding caustic images in distant screen.

At the beginning of the spreading, the FFSI of the ethanol drop-refracted laser beam shows a bright ring at outmost periphery without visible interference fringes as shown in Figure 7a that were taken at $\tau = 0.001$ for $t_f = 107$ s. Meanwhile the magnified top view of the drop, which can be recorded by a video camera placed over the drop [7], shows a perfect circle. It implies that the sessile drop takes a spheroid cap with precursor film at this moment. The quasi-inflection line exists at the crossover zone between the precursor film and the macroscopic wedge of the drop, as shown in Figure 6a. The curvature of the drop surface changes its sign only in the radial direction and therefore a far-field caustic without cusps appears on the screen. As the curvature of the drop surface changes very sharply at the crossover zone, the caustic diffraction fringes are too close to be identified, merging into the caustic. Very soon the ethanol-drop surface becomes instable because of the high value of the dielectric coefficient [13]. The drop surface ripples at the macroscopic potion of the drop near the TPL, and therefore, the cusp caustics appear on the FFSI, as shown in Figure 5a. At this moment the drop has the shape of a spheroid cap with an irregular 'protruding foot', i.e. the C-I type profile shown in Figure 6b.

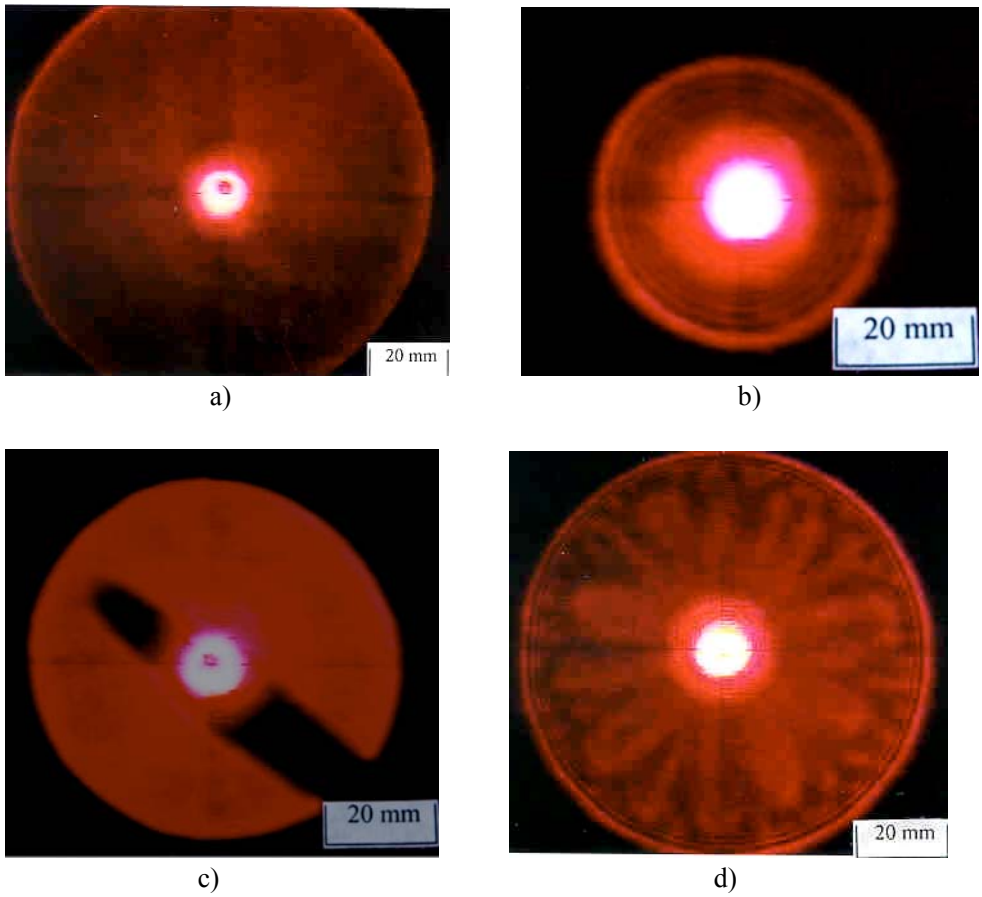


Figure 7. FFSIs of parallel laser beam refracted by: a) ethanol drop of 2.5 μl at $\tau = 0.001$ for $t_f = 107$ s; b) ethanol drop of 2.5 μl at $\tau = 0.76$ for $t_f = 107$ s; c) R-113 drop of 2.5 μlat $\tau = 0$ for $t_f = 16.1$ s; d) R-113 drop of 2.5 μl at $\tau = 0.19$ for $t_f = 16.1$ s.

After then, the macroscopic portion of the ethanol drop surface evaporates while its bottom continues to spread. Consequently, the drop is deformed and gradually becomes a crater-like profile and then a pancake-like profile, producing higher hierarchy caustics in far field, as shown in Figure 5b. Usually, this kind of drop profiles has minor interest and too difficult to quantitatively be measured in the study of liquid spreading, and therefore, it has not been listed in the basic profiles. At the last stage of the spreading-and-evaporation, the ethanol drop undergoes a retraction process. The drop recedes from the pancake-like profile to a new tiny spheroid cap with a smooth 'protruding foot'. A typical photo of the caustic with its diffraction on the FFSI of the drop-refracted laser beam are shown in Figure 7b, which were taken at $\tau = 0.76$. The circular caustic line with its diffraction fringes are clearly observed as a family of concentric circles in the distant screen. The drop profile is inferred as C-PF type profile shown in Figure 6c. At this moment, a regular protruding foot is formed by the combined actions of retraction and evaporation.

The sessile-drop shape of a perfect spheroid cap (P-C type), as shown in Figure 6d, is rare and may exist only at the very beginning of spreading of some liquid drops. For example, we observed that when an R-113 drop spreads in its unsaturated vapor environment, at the

moment when the micro-syringe tip was removed from the drop body, the FFSI of the drop-refracted laser beam was a circle with homogeneous brightness, as shown in Figure 7c, which was taken at $t = 0\ s$. At this moment, no caustic at all appeared in the far field of waves. It implies that the drop profile was a P-C type profile shown in Figure 6d. However, a bright caustic line immediately appeared at the outmost rim of the FFSI and then its diffraction fringes came out. This process implies that a precursor film formed around the sessile drop (C-P type profile) first and then a quasi-inflection line occurred at the macroscopic potion of the drop, indicating that a protruding foot was formed due to the spreading (C-PF type profile).

In the spreading-and-evaporation of an R-113 sessile drop in open air, TCCFs in the drop is induced by the strong evaporation, but the drop-surface profile still takes the shape of a spheroid cap in the macroscopic potion with a smooth 'protruding foot'. It retains the C-PF type profile from beginning to end during the spreading-and-evaporation process. Both the caustics produced by the varying refractive index field due to the TCCFs and the caustic produced by the local third-order surface at the quasi-inflection line on the drop surface with its diffraction fringes can be clearly observed on the screen. A typical picture of the FFSI is shown in Figure 7d taken at $\tau = 0.19$ for $t_f = 8.2\ s$. Because of the strong evaporation, C-P type profile cannot be formed on R-113 drops.

As for nonvolatile liquid drop, such as silicone oil (50 cSt.), a different changing process of the drop profile was observed. At the very beginning of the spreading the FFSI of the drop-refracted laser beam shows the same pictures as the FFSI of R-113 drop-refracted laser beam shown in Figure 7c. It indicates that the drop took a perfect spheroid-cap shape (P-C type) at this moment. Then the drop shape immediately changes to C-PF type and finally to C-P type profile, as implied by Figs. 8a and 8b, which were taken *at* $t = 5.03\ s$ and $t = 23.07\ s$, respectively.

Measurements of Drop Characteristic Parameters

Besides qualitatively determine the drop profiles and so ascertain the drop spreading stability, the caustics combined with their diffractions can also be used to quantitatively measure two important characteristic parameters of sessile drops: the real-time local contact angles and the foot height of the drops. As long as combine FFSIs of the drop-refracted laser beam with the corresponding magnified top view of the drop, these two measurements can be realized.

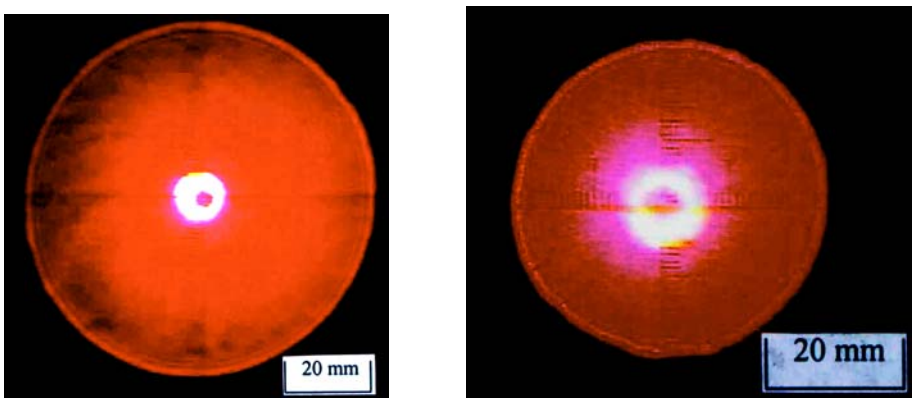
Contact Angles

As shown in Figure 9, the local contact angle of the drop at the right side in h-direction, θ_r, can be determined through measuring geometric parameters, p and L_r, and the liquid refractive index, n. Here, p is the distance between the drop base and the distant screen which is much greater than the drop height δ, L_r is the distance from the right rim of the drop projection on the screen to the left end of the caustic ring. Simple geometric relationship, $\tan\theta_r'' = L_r / p$, and application of Snell's law to the air-liquid and liquid-glass interfaces, i.e. $\sin\theta_r = n\sin\theta_r'$ and $\sin\theta_r'' = n\sin(\theta_r - \theta_r')$, yield the local contact angle as

$$\theta_r = \arctan \frac{\sin \arctan(L_r / p)}{\sqrt{n^2 - \sin^2 \arctan(L_r / p)} - 1} \tag{19}$$

where θ_r' and θ_r'' are the refracted angle at air-liquid interface and the emerging angle at glass-air interface, respectively.

Meanwhile, the local contact angle of the drop at the left rim, θ_l, can also be calculated using Eq. (19) as long as L_r is replaced by the distance from the left rim of the drop projection on the screen to the right end of the caustic ring, L_l. Similarly, the local contact angle in any selected direction i, such as the left contact angle $\theta_{i,l}$ and the right contact angle $\theta_{i,r}$ can be calculated also using Eq. (19) after L_r is replaced by the corresponding distances, $L_{i,l}$ and $L_{i,r}$, respectively.



a) b)

Figure 8. FFSIs of the laser beam refracted by a silicone oil (50 cSt) drop at different times from the beginning of spreading: a) at t = 5.03 s; b) at 23.07 s.

If a sessile drop has stable interfaces and spreads and evaporates on an isotropic surface, the FFSI of the drop-refracted laser beam would exhibit a circle, such as shown in Figs. 7 and 8. In these case, the contact angles in all directions should be the same and the Eq. (19) can be written as

$$\theta = \arctan \frac{\sin \arctan((D+d)/2p)}{\sqrt{n^2 - \sin^2 \arctan((D+d)/2p)} - 1} \tag{20}$$

where D is the diameter of the outmost circular caustic line on the FFSI, and d is the drop contact diameter that can be accurately measured from the magnified top view.

In practical operations, the thick caustic lines blur the measurements of D or L. In fact, based on the analysis of the cause of caustic formation mentioned above, it can be concluded that the inner rim of the caustic rings should be the measuring end.

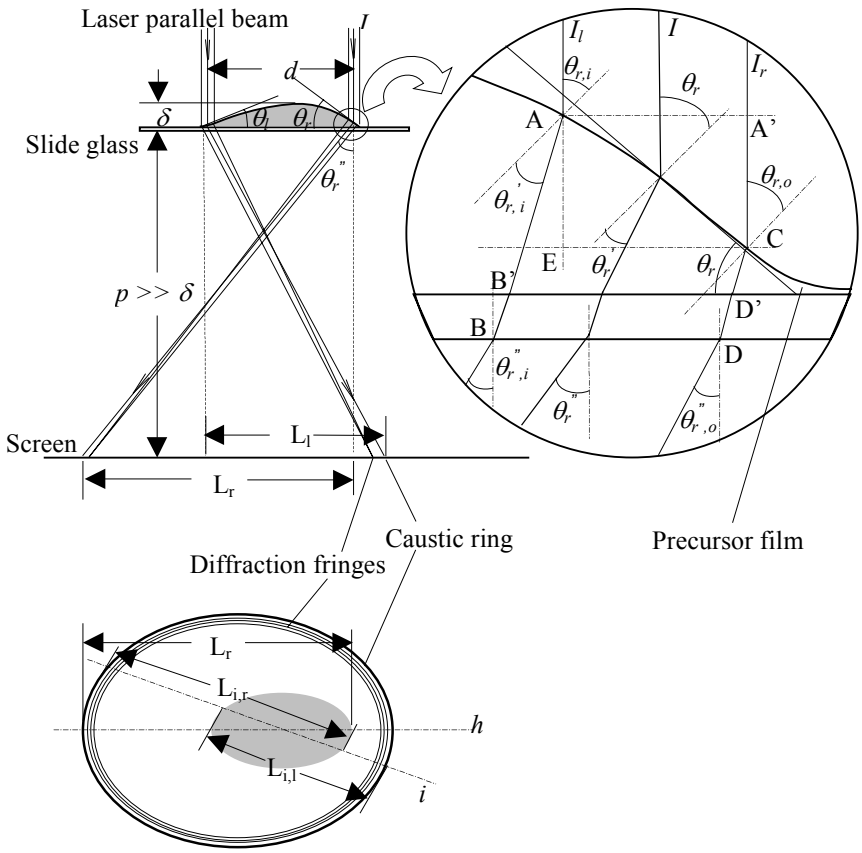


Figure 9. Schematic of optical path of drop-refracted laser-beam image-graph.

Foot Height

As analyzed above, it is the caustic produced by the rays at a quasi-inflection line (ray *I* shown in Figure 9) that forms the outmost bright-thick ring on the FFSI. The caustic diffraction arises from a lack of flatness in the initial wavefront, rather than from the aperture effect. It is the initial wavefront itself that starts out with a phase variation. The most elementary way of treating the diffraction is in terms of interfering rays, known as the semi-classical approximation. Each ray is characterized by amplitude and phase, so that they may interfere with each other. Based on this approximation, the caustic diffraction can be considered as a result produced by the interference between the rays at both side of the quasi-inflection line, I_l and I_r, as shown in Figure 9. The region around the quasi-inflection line is so-called 'drop foot' with a height of *b*. As mentioned above, when a collimated laser beam passes through the drop, the angle of incidence reaches maximum at the quasi-inflection line, which is considered as the macroscopic contact angle of the drop whose local value at the right rim is symbolized as θ_r, as shown in Figure 9. The angles of incidence of the rays at both sides of the quasi-inflection line (one on each side), $\theta_{r,i}$ and $\theta_{r,o}$, gradually reduce and the two rays next to the quasi-inflection line at the both sides will interfere with each other if the

difference in the optical paths between them is an integral time of the wavelength of the laser beam. Based on this relation the height of the inflection point can be estimated.

As shown in Figure 9, the difference in the optical paths between the two rays located on both side of the quasi-inflection line can be estimated as $\Delta=n(AB')+n_g(B'B) - (A'C) - n(CD)-n_g(D'D) \approx n(AB') - (AE) - n(CD)$, where n_g is the refractive index of the slide-glass. For generality, we consider a generalized point at the quasi-inflection line and move the subscript r out. It leads to

$$\Delta = (AE)\left[\frac{n}{\cos(\theta_i - \theta_i^{'})} - 1\right] \tag{21}$$

where θ_i and θ'_i are the incidence angle and the refracted angle of the ray on air-liquid interface at inner side of the quasi-inflection line, respectively. The condition to interfere between these two rays is $\Delta = N\lambda$, where N is a positive integer, λ is the laser beam wavelength. Since the drop profile changes rapidly from the macroscopic region to the microscopic region at the TPL region, only a finite number of the interference fringes can be produced on the FFSI, except an uncollimated beam is used. The length of (AE) approximates the foot height b, ignoring the precursor-film thick of the order of 0.01 μm. Hence,

$$b = N \lambda \cos(\theta_{i,m}-\theta'_{i,m})/[n-\cos(\theta_{i,m}-\theta'_{i,m})] \tag{22}$$

N is the number of frings, $\theta_{i,m}$ is the incidence angle of the ray at the inner side of the quasi-inflection line that produces the inmost interference fringe (innermost-side ray), and $\theta'_{i,m}$ is the refracted angle of the innermost-side ray. The incidence angle of the innermost-side ray, $\theta_{i,m}$, can be calculated using Eq. (20) in which D is replaced by the innermost fringe diameter $D_{i,m}$. The refracted angle of the innermost-side ray, $\theta'_{i,m}$, is determined by

$$\theta_{i,m}^{'} = \arcsin\frac{\sin\theta_{i,m}}{n} \tag{23}$$

Example of Applications

The simultaneous records of the magnified top view of the drops and the FFSIs of the drop-refracted laser beam provide not only the information of spreading, such as evolutions of contact angle and contact diameter, but also the information on the effects of both evaporation rate and TCCFs in drops on the spreading characteristics and the TPL nature. Additionally, for the drops having a profile of spheroid cap, the real-time volume of the drops can be estimated using a simple geometric relationship [13].

Three typical volatile liquids, cyclohexane, n-pentane, and R-113 were tested in open air at room temperature (23°C) and relative humidity of 44%. The FFSIs of the laser beam refracted by the sessile drops of these liquids were nearly perfect circle shape, and therefore, the instant contact angles of the drop were determined by Eq. (20).

Test results show that both evaporation rate and TCCFs induced by evaporation strongly affect the spreading. Although the average evaporation rates ($w = V_o/t_f$) of the tested liquids differ from each other by an order of magnitude, in all cases, the capillary spreading dominates the initial stage and then is overtaken by the effect of evaporation, as shown in Figure 10. Four stages are observed: (1) initial spreading; (2) spreading-evaporation balance; (3) evaporation dominant contraction; and (4) rapid contraction. The duration of each stage is determined by a combination of liquid spreading power, evaporation rate, and TCCFs, upon occurrence. It can be seen that although the evaporation rate of the n-pentane greater than one of the cyclohexane drop by almost an order of magnitude, the n-pentane drop has a longer initial spreading duration and much longer spreading-evaporation balance duration, see Figs. 10a and 10b. It implies that the n-pentane has stronger spreading power than cyclohexane.

The TCCFs induced by evaporation have been found to profoundly affect on the spreading. However, convection in sessile drops can be induced by evaporation only under certain conditions, depending on the properties of the volatile liquids. Of the three liquids, the cyclohexane drops with initial volumes of 1.5 μl to 5 μl exhibit no convection throughout their lifetime, while in the case of n-pentane drops the convection occurs near the end of the second stage (the spreading-evaporation balance stage). For example, for the n-pentane drop with initial volume of 1.61 μl, convection appears at about 1 second after the drop begins to spread, referring to Figure 10b. Before then there is no convection in the drop at all. As for R-113 drops, the TCCFs induced by evaporation occur from very beginning of the drop spreading and run very strong all the time. It is the convection that promotes evaporation role impeding the drop spreading. Consequently, the initial spreading duration of R-113 drops becomes very short though the evaporation rate of R-113 drops (0.12 $\mu l/s$) is much lower than one of n-pentane drop (0.22 $\mu l/s$). It is also due to the strong convection that observable spreading-evaporation balance stage does not exist for the R-113 drops, as shown in Figure 10c. It is noted that the final rapid contraction stage always starts from $\tau \approx 0.81$ and $d \approx 2$ mm for all three drops.

In contrast to the contact diameter, the evolutions of contact angle of sessile drops of various liquids are quite different from each other. For drops with lower evaporation rates, such as cyclohexane drop, the contact angle decreases in the initial spreading and spreading-evaporation balance stages; subsequently it increases linearly during the evaporation-dominant contraction stage and then rapidly grows in the final rapid-contraction stage, as shown in Figure 10a. Obviously, evaporation increases the contact angle because of a large evaporative loss near the drop edge. This result qualitatively agrees with the theoretical predictions of Anderson and Davis [14] and Hocking [15]. For the drops with higher evaporation rates, the contact angle decreases monotonically. Although evaporation generally increases the contact angle, but when a stronger evaporation consumes the liquid volume of the drop large enough, the contact angle starts to decrease. The evolution characteristics of both contact diameter and contact angle shown in Figs. 10b and 10c also agree well with the theoretical prediction made by Anderson and Davis [14].

When the drop foot heights of spreading drops are concerned, by the aid of the caustics and their diffraction fringes on the FFSIs of the drop-refracted laser beam, the foot heights can be easily estimated. As examples, the foot heights of the drops, refracted by which the laser beam's FFSIs are shown in Figs. 2a and 2c, were calculated using Eq. (22) combined with Eqs. (20) and (23) at 5.91 and 12.40 μm, respectively. Here the following parameter

values were used: p = 630.5 mm, λ = 0.6328 μm, for cyclohexane drop D = 78.67 mm, $D_{i,m}$ =73.5 mm, N = 4, d = 6.54 mm; for R-113 drop D = 91.26 mm, $D_{i,m}$ = 74.37 mm, N =7, d = 4.22 mm.

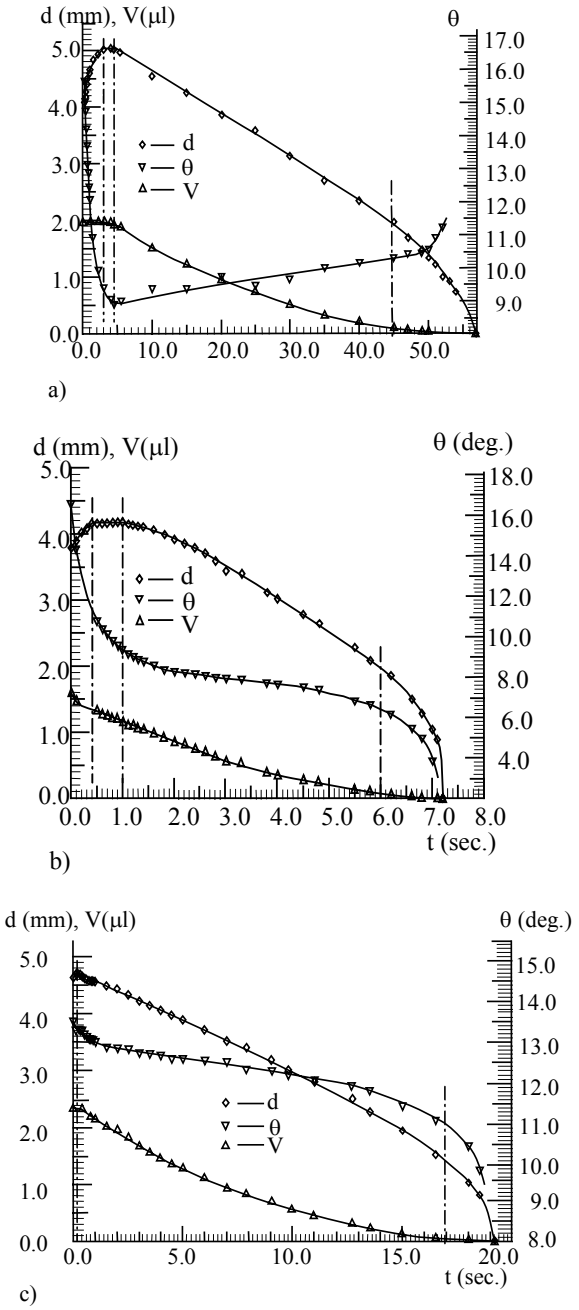


Figure 10. Time-history of contact angle and contact diameter of: (a) a cyclohexane drop with V_o = 1.91 μl, and w = 0.034 $\mu l/s$, (b) an n-pentane drop with V_o = 1.61 μl, and w = 0.22 $\mu l/s$, (c) an R-113 drop V_o = 2.36 μl and w = 0.12 $\mu l/s$.

Conclusion

Catastrophe optics is not only able to qualitatively explain optical phenomena in nature but also able to quantitatively study liquid spreading.

Caustics have been successfully applied to qualitatively determined drop profiles through the caustics on the far-field sectional images of the drop-refracted laser beam and so the drop spreading stability is ascertained. It is with the aid of catastrophe optics that four basic drop profiles in drop spreading are revealed.

The caustics and caustic diffractions have been used to quantitatively measure some important characteristic parameters of sessile drops, including real-time local contact angles and heights of drop foot. These two measurements are the unique functions of catastrophe optics, which cannot be realized by conventional methods in the study of drop spreading.

High hierarchy caustics have also been proved an exclusive means of flow visualization, which sensitively indicate the thermo-capillary convection flows occurring in the evaporating sessile drops that cannot be detected by other non-intrusive methods.

References

[1] Zeeman, E. C. in *Catastrophe Theory and Applications*; Sinha, D. K.; Ed.; John Wiley & Sons: New York, NY, 1981, pp.1-10.
[2] Wright, F. *Phys. Bull*. 1988, 39, 313-316.
[3] Berry, M. V. in *Progress in Optics*; Wolf, E.; Ed.; Elsevier North-Holland, Inc.: New York, NY, 1980, Vol. 18, pp. 257-346.
[4] Berry, M. V. *Adv. Phys*. 1976, 25, 1-26.
[5] Nye, J. F. *Proc. Roy. Soc. Lond*. 1978, A361, 21-41.
[6] Nye, J. F. *Natural Focusing and Fine Structure of Light*; IOP Publishing Ltd, Bristol, UK, 1999.
[7] Zhang, N.; Chao, D. F. *Optics & Laser Technology* 2003, 35, 155-161.
[8] Zhang, N.; Chao, D. F. (2003) CD Proceedings of 7th *Int. Symp. on Fluid Control*, Measurement and Visualization, Sorrento, Italy.
[9] Zhang, N.; Chao, D. F.; Wang, X. F. (2005*) Proceedings of 8th International Symposium on Fluid Control*, Measurement and Visualization, Chengdu, China.
[10] Gennes, P. G. *Rev. of Modern Phys*. 1985, 57, 827-863.
[11] Leger, L.; Joanny, J. F. *Rep. Prog. Phys*. 1992, 55, 431-486.
[12] Pompe, T.; Herminghaus, S. *Phys. Rev. Letters* 2000, 85, 1930-1933.
[13] Zhang, N.; Yang, W. J. *J. Heat Transfer*. 1983, 104, 656-662.
[14] Anderson, D. M.; Davis, S. H. *Phys. Fluids* 1995, 7, 248-265.
[15] Hocking, L.M. *Phys. Fluids* 1995, 7, 2950-2955.

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN: 1-60021-194-1
Editor: Steven B. Larkin, pp. 213-251

Chapter 7

RECENT ADVANCES IN TEA CO_2 LASER TECHNOLOGY

D.J. Biswas, J.P. Nilaya, M.B. Sai Prasad, and A. Kumar*
Laser & Plasma Technology Division, Bhabha Atomic Research Centre,
Mumbai – 400085, India
*AFFF, Bhabha Atomic Research Centre, Tarapur – 401502, India

Abstract

The new results obtained in the area of repetitive and single longitudinal mode operation of TEA CO_2 lasers have been presented. Helium-free operation of a conventional TEA CO_2 laser excited by a novel pulser circuit and its advantages in repetitive and single longitudinal mode lasing has been described. Much higher detunability of the single longitudinal mode achieved by making use of a traveling wave cavity in conjunction with a saturable absorber is discussed. A latch-proof pulser driven by an indigenously developed rotating dielectric spark gap switch makes the repetitive operation possible.

1. Introduction

By virtue of its high quantum efficiency, easily achievable selective excitation and high-pressure operation, CO_2 laser has been recognized as one of the most powerful lasers in the mid-infrared region [1]. Its utility in material processing [2], isotope separation [3], generation of coherent sources of higher wavelength [4] etc. are well researched and documented. Many of these applications require the laser to be operated repetitively and/or with high degree of spectral purity. In this article we address these two aspects viz., the repetitive operation and the single mode operation of TEA CO_2 lasers, along with the new results obtained by us in this area.

The role of helium, that constitutes a majority of the gas mixture in a pulsed transversely excited atmosphere (TEA) CO_2 laser, is to stabilize the discharge [1]. This expensive and scarce gas, with its very low electron affinity, facilitates the occurrence of an arc free discharge at higher operating pressures and hence has been indispensable in the conventional

operation of TEA CO_2 lasers. Efforts have been expended in the past to achieve helium-free operation of CO_2 lasers, although, with limited success. The work on helium-free operation of a conventional TEA CO_2 laser where the drawbacks of the earlier methods have largely been overcome is described here. He-free operation is possible under widely varying operating conditions viz., active volume, geometry of the electrodes, location and type of the preioniser, gas mixture etc. The electro-optic efficiency of the helium-free laser when operated on TEM_{00} mode is always found to be higher as compared to its operation with conventional gas mixture. This increased TEM_{00} mode filling factor in the TEA section is exploited to improve considerably the operating efficiency of a hybrid CO_2 laser which can prove advantageous in many applications e.g., laser isotope separation, LIDAR, optically pumped mid and far infrared lasers, nonlinear optical interaction in semiconductors and gases etc, where the emission from a TEA CO_2 laser is required to be on a single longitudinal and transverse mode. A more efficient method of obtaining single longitudinal mode (SLM) lasing is by the elimination of spatial hole burning effect. This is achieved by placing an appropriate intra-cavity saturable absorber that ensures unidirectional oscillation in the ring cavity leading to lasing on a single mode with much wider detunability.

The fact that expensive and scarce helium is not a constituent of the laser gas mixture makes it an ideal system for high repetition rate operation. The pulser required to operate a repetitive laser employs a high voltage high current switch, generally, a thyratron. Thyratrons are expensive and have limited life. Operation of a thyratron-less repetitive pulser for energising such TE lasers is also described in this article. The heart of this novel TE laser pulser is a rotating dielectric spark gap whose unique geometry is fully exploited to obtain diode-less resonant charging operation of the pulser and also to drive two high repetition rate TE lasers either simultaneously or with a variable delay.

1.1. Repetitive Operation

The two major considerations in the repetitive operation of a TEA CO_2 laser are the removal of the contaminated gas from the inter-electrode region between two successive pulses and the usage of a high voltage, high current, and fast switch that is capable of withstanding the repetitive operation. As will be discussed at length in section 1.1.3, following a discharge, the gas in the inter-electrode region is hot and rich in dissociation products that need to be replaced before the arrival of the next pulse. The rate of removal of the contaminated gas from the active region thus increases with increasing repetition rate. In the following section, the role that helium plays in the atmospheric pressure operation of pulsed CO_2 lasers and the strong bearing it has in the repetitive operation of these lasers is discussed.

1.1.1. Role of Helium

It is well known that helium is normally used in large proportion with CO_2 and N_2 for the operation of atmospheric pressure CO_2 lasers. Obtaining self-sustained discharge at atmospheric pressure with N_2 and CO_2 gases alone is a difficult proposition for the following reason. In the discharge, due to electron impact, CO_2 gets dissociated into CO and O_2.

$$2CO_2 \rightarrow 2CO + O_2$$

The O_2 molecules form negative ions by dissociative attachment-

$$O_2 + e \rightarrow O + O^-$$

And by three body attachment processes-

$$O_2 + e + M \rightarrow M + O_2^-$$

The formation of a negative ion is thus at the expense of an electron in the discharge. Not only this negative ion is unable to participate effectively in the process of ionization, it also prevents the electron to which it got attached from continuing this process. Various oxides of Nitrogen formed in the discharge also have higher electron affinity and therefore contribute more to the removal of electrons from the discharge rather than to their production. The discharge thus cannot be self sustained for long and tends to grow into an arc soon after its initiation. As a matter of fact in N_2-CO_2 mixture at atmospheric pressure, glow to arc transition occurs within few tens of nanoseconds. This precludes the operation of a conventional TEA CO_2 laser with molecular gases alone. For example a CO_2 laser capable of delivering 1 J of energy per pulse needs to be energized by a 20 nF capacitor charged to ~30 kV. For a typical TEA laser (with discharge impedance of ~10 Ω) the discharge time constant is ~200nsec. Within three time constants most of the electrical energy stored in the capacitor gets coupled to the discharge in the form of electronic kinetic energy. The electrons present in the discharge, in turn, transfer part of this energy as internal energy of CO_2 molecules through collisions. The duration of the discharge must, therefore, be few hundreds of nanoseconds while for a CO_2-N_2 mixture, the glow to arc transition occurs within few tens of nanoseconds. The operation of a conventional TEA CO_2 laser is thus not possible if the laser gas mixture constitutes only CO_2 and N_2.

Helium has no electron affinity. The collisions that electrons in the discharge undergo with helium are, in general, elastic. Helium, when present in large proportion with N_2 and CO_2 mixture, creates condition suitable for electron multiplication in the discharge and, at the same time, reduces the possibility of their removal through attachment processes described above. This, in turn, extends the glow to arc transition time making the operation of the conventional CO_2 laser possible.

Helium, which constitutes majority of a CO_2 laser gas mixture, is expensive and scarce. This has strong bearing on the repetitive operation of TEA CO_2 lasers. The expensive gas, that gets contaminated during a discharge owing to the formation of O_2, cannot be thrown out of the laser head and is, therefore, recycled after passing through heat exchanger and catalytic re-converter where recombination of O_2 and CO back to CO_2 takes place. This re-circulatory loop not only makes the repetitive laser bulky but also increases its basic cost. Understandably, therefore, there have been a number of attempts in the past to obtain helium free operation of CO_2 lasers although with limited success. These methods and their limitations are briefly described below-

1.1.1.1. Low Pressure CO_2 Laser

Helium-free operation has been successfully achieved in TE CO_2 lasers [5, 6] at sub-atmospheric pressures or low-pressure synchronous longitudinal discharges [7] or cw operation of low pressure CO_2 lasers [8]. While the variations in the excitation circuits allow

helium free operation in TE systems albeit below atmospheric pressure, the low operating pressure in CW systems intrinsically ensures glow mode operation of the discharge in absence of helium.

Due to the sub-atmospheric pressure operation of these systems, the maximum coherent power that can be obtained from them is always lower than TEA systems with similar active volume.

1.1.1.2. Rapid Discharge Technique [9-13]

This approach takes advantage of a very rapid discharge (few tens of nsec as against hundreds of nsec in a conventional operation) to realise helium free operation, as the glow to arc transition in the absence of helium is very fast. Thus the discharge extinguishes before arcing can set in.

Helium free operation by rapid discharge technique can be effected only in specially designed mini laser systems that inherently offer low discharge loop inductance. Such operation, therefore, restricts the active volume and hence the maximum obtainable energy output from the system. Rapid excitation invariably results in the emission of optical pulses of short duration. Conventional long pulse operation is therefore not possible by this method. It is to be noted here that Marchetti and Simili [13] obtained very short pulses from a helium free multi-atmospheric CO_2 laser by making use of a double-sided four fold LC inversion circuit originally used by Stuart et al [14] to operate continuously tunable high pressure CO_2 laser.

1.1.1.3. Seeding the Laser Gas Mixture with Low Ionisation Potential (LIP) Additives

In the absence of helium, the electrons in the discharge are lost largely by negative ion attachment processes giving rise to the formation of an arc discharge. The addition of LIP hydrocarbons increases the primary photoelectron density thereby compensating the loss of electrons in absence of helium leading to arc free operation [5].

The LIP additives seeded in the laser gas mixture undergo dissociation in an electric discharge and tend to settle on the optics, electrodes, and the internal surface of the laser head degrading the performance of the laser rapidly.

1.1.1.4. Preconditioning the Inter-electrode Volume by Electrons from an External Source

Loss of electrons in absence of helium can be overcome by deluging the active volume with electrons produced externally, as in case of an electron beam controlled CO_2 lasers, resulting in arc free operation [15].

In this method, helium-free operation of TEA CO_2 lasers calls for an external source of electrons thereby making the system more complicated, expensive and bulky. In addition, this is achieved at the expense of the wall plug efficiency.

1.1.1.5. Modified Excitation Circuit

Helium-free operation of a conventional TEA CO_2 laser has, of late, been obtained by suitably modifying the excitation circuit [16,17]. This method of helium free operation, achieved in systems where the active volume ranges from ~1cc to 200cc, is irrespective of i) the geometry of electrodes, viz., profiled, un-profiled, or cylindrical, ii) the type of ballasting used for preionisation viz., resistive, inductive, or capacitive, and iii) the location of the

preioniser, viz., beneath the semi-transparent electrode or along the side/sides of the electrodes [18]. This has been rendered possible by integrating the spiker and sustainer-like actions in to a single pulser network by making use of a coupling inductance. Another important advantage of this excitation circuit is its simplicity. A single source powers and a single switch controls the preioniser, the spiker, and the sustainer discharges.

1.1.1.5.1. Experimental System

Although helium-free operation by this method has been achieved in a wide variety of systems, a detailed description is provided here for the operation of a medium sized TEA CO_2 laser with contoured electrodes. A 3-dimensional view of the laser head is shown in Figure 1. It consists of two Ernst profiled [19] electrodes which define a discharge of cross section 1.5 × 1.5 cm^2 and length 30 cm. Preconditioning of the laser gas mixture is accomplished by creating auxiliary spark discharges between the brass preionising pins (14 pairs on either side) placed at a regular interval along the length of the discharge and at a distance of 5 cm from its centre. The entire assembly is housed in a leak tight Perspex chamber (36 cm× 12cm× 15 cm) the ends of which are 'O'ring sealed with a concave 4m ROC gold coated mirror and a ZnSe Brewster window. A 70% reflective ZnSe plane output mirror together with the gold-coated mirror defines the 65 cm long optical cavity. The flow of current through each of the auxiliary sparks is limited by connecting a small capacitance in series with it.

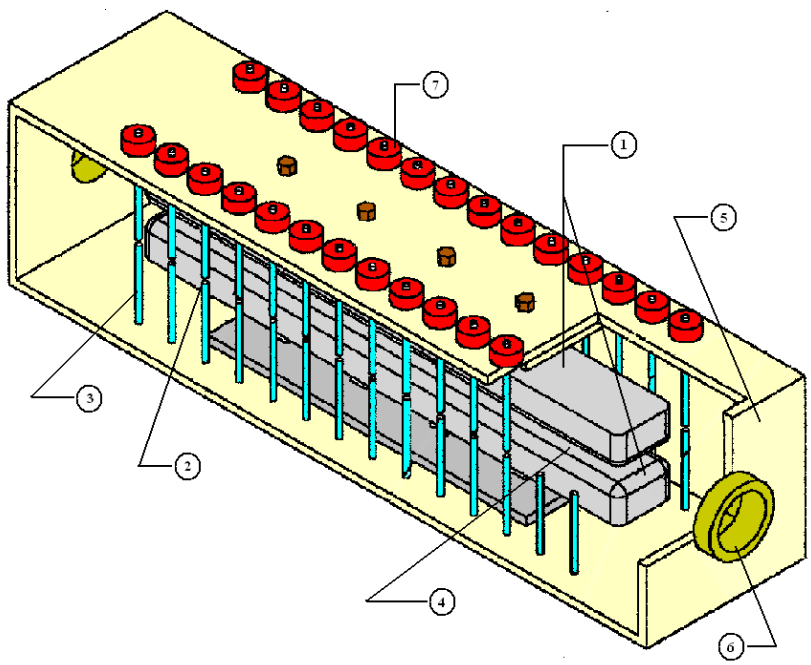


Figure 1. Three dimensional view of the laser head. 1- uniform field electrodes 2- preionisation gap, 3-preionisation pins, 4-inter-electrode gap, 5-Perspex housing, 6-cavity mirror, 7-preionisation capacitors.

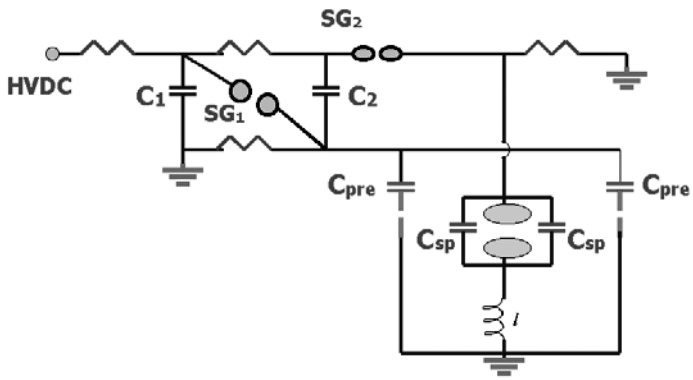


Figure 2. Schematic of the excitation circuit for helium-free operation of a TEA CO_2 laser.

The schematic of the excitation technique employed to operate the TEA CO_2 laser wherein the spiker and sustainer like actions have been integrated into a single pulser network with the help of a coupling inductance (*l*) is shown in Figure 2. Such an integration has made possible a single source to power and a single switch driven pulser to control all the three discharges viz., the pre, the spiker, and the sustainer discharges as against the conventional spiker sustainer excitation scheme that requires two sources and two switches to achieve the same [18]. The usage of a two stage Marx generator allows the condenser C_1, on closure of switch SG_1, to power the pre discharge. Alongside the preionisation, on closure of SG_2, the main condensers C_1 and C_2 come in series and charge transfer through the inductance causes the voltage across the spiker condenser C_{sp} to build up rapidly as shown in Figure 3 (a). Figure 3 (b) represents the current flowing from the main condensers following the closure of the spark gap SG_2. The first forward cycle of the current pulse charges up the spiker condenser C_{sp} to its peak voltage. As this high voltage impulse is impressed across the inter electrode gap, it closes leading to the flow of the spiker current. The presence of inductance in the spiker charge-up loop automatically delays the main discharge with respect to the pre discharge.

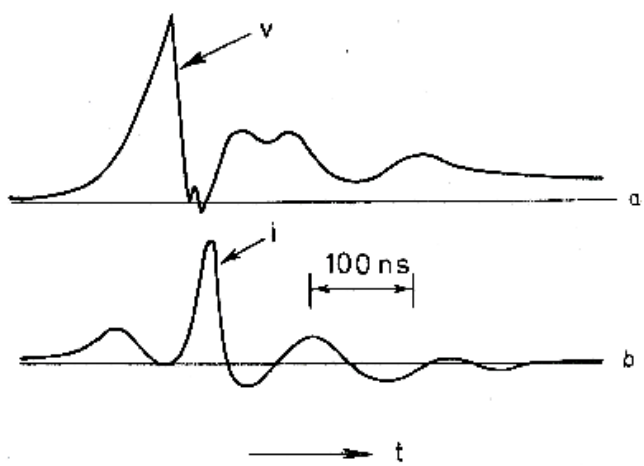


Figure 3. a) the voltage across the main discharge, b) the current in the sustainer loop.

This small delay can be readily measured from the temporal wave-forms of Figure 3 and has also been corroborated by adopting the delay measuring method of Kumar et al [20] based on the collection of light originating from the preioniser and main discharges separately. The initial voltage to which the main condensers are charged and the values of C_1, C_2, C_{pre} and C_{sp} are so chosen that after the switching of the Marx Bank, the voltage across the preconditioned inter-electrode gap results in an E/N appropriate for the initiation of the discharge. The reduced voltage across the main condenser, by virtue of its powering the preioniser discharge and spiker discharges, in conjunction with the inductance (*l*) maintains an E/N condition suitable for the sustenance of the discharge. Decoupling of the two discharges viz., the spiker and the sustainer and the tailoring of the sustainer pulse by making use of the inductance results in a condition where glow discharge can be reliably obtained even in the absence of helium under wide range of operating conditions. The inductance plays a very crucial role as it controls the delay between pre and spiker as well as spiker and sustainer discharges in addition to deciding the rate of rise of voltage across the electrodes.

1.1.1.5.2. Performance of the Laser

The energy, peak power, and the FWHM value of the emitted laser pulse are monitored as a function of the partial pressure of CO_2 gas for a particular value of the inductance. The results of these studies are summarized in Figure 4 and Figure 5. As can be seen, the most optimised performance of the laser in terms of its output energy occurs for equal concentration of CO_2 and N_2 (Figure 4). The electro-optic efficiency is estimated to be ~7% after accounting for the residual energy in the condenser at the end of the discharge (see Figure 3 (a)). The electro-optic efficiency when the same laser operates with conventional gas mixture (N_2:CO_2: He=1:1:5) is measured to be 7.8% for multi mode operation. To be noted here that the efficiency would be higher in both the cases if the fraction of the stored energy expended in the preconditioning is also taken into account. Besides contributing to the stabilisation of the discharge, helium also helps in depopulating the lower vibrational level (01^10) allowing thereby the participation of the same CO_2 molecule in the lasing process more than once. Although this process is slowed down in the absence of helium, however owing to the large concentration of CO_2 molecules (50 %) in the gas mixture, the electro optic efficiency reduces only marginally. With the addition of a small amount of hydrogen to the gas mixture (N_2:CO_2:H_2=1:1.2:0.1), the multimode efficiency increases to ~8.8%. Thus helium free operation has been achieved with an increased multi-mode efficiency, although with the addition of some amount of hydrogen. It should be noted that in the TEM_{oo} mode operation, however, the efficiency of helium free operation always exceeds that obtained with conventional gas mixture. This is due to the fact that in the absence of helium, width of the glow discharge is less [9] and hence TEM_{oo} mode filling is better. A logical conclusion of this finding is that the emission from helium-free TEA CO_2 laser will be less divergent as compared to the emission from a laser operated with conventional gas mixture.

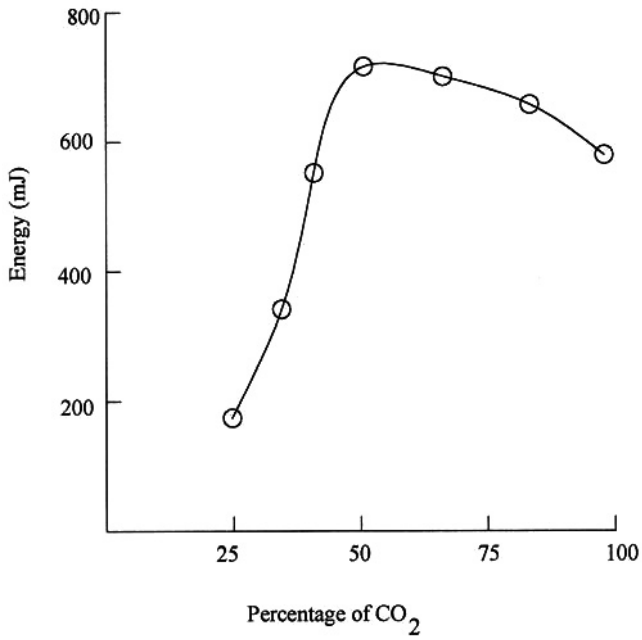


Figure 4. Output energy of the laser as a function of the partial pressure of CO_2 gas.

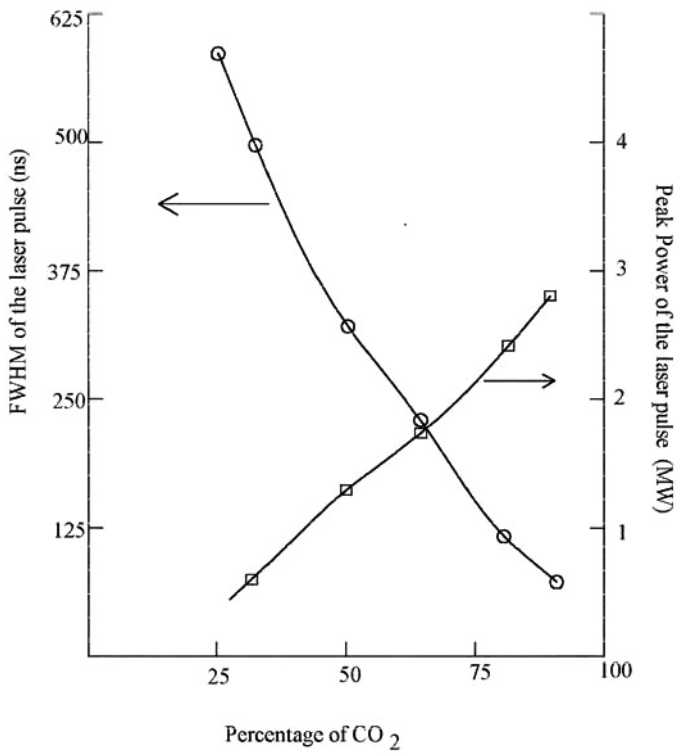


Figure 5. Peak power and the temporal width of the laser pulse as a function of the partial pressure of the CO_2 gas.

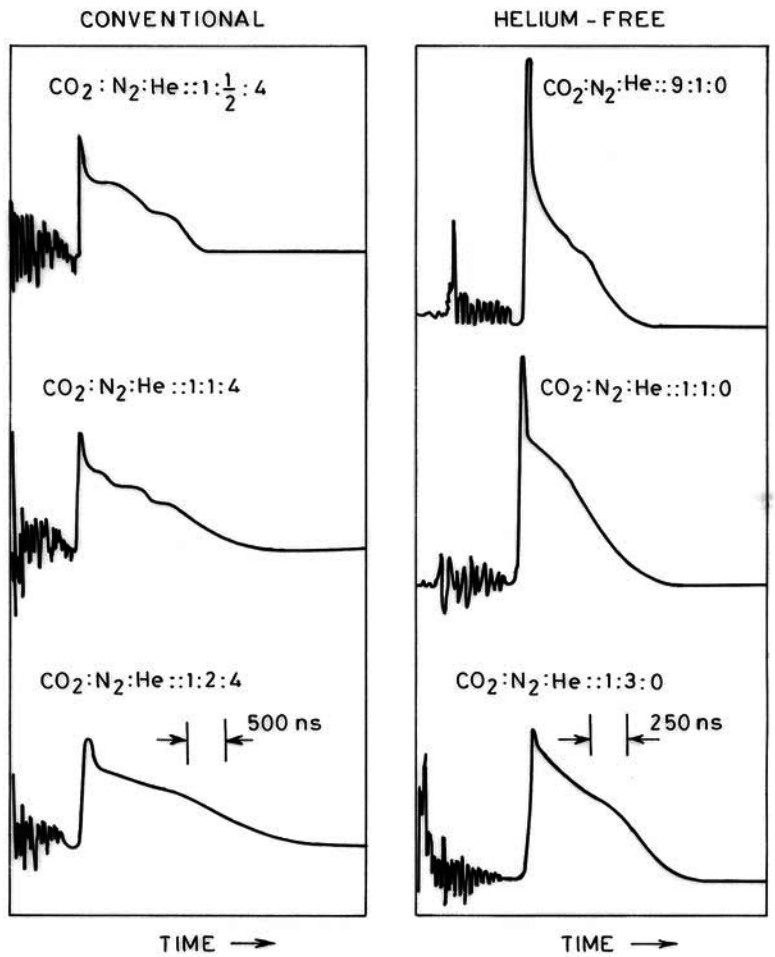


Figure 6. Temporal shape of the laser pulse for conventional and helium-free operation as a function of gas composition.

The peak power, duration and energy of the laser pulse can be varied by varying the partial pressures of N_2 and CO_2 as shown in Figure 5. As the concentration of CO_2 reduces from 92 % to 27 %, the FWHM value of the laser pulse increases (from 80 nsec to 590 nsec) by a factor of ~7.5 while the peak power drops by a factor of ~5.25 (from 2.84 MW to 0.54 MW). The peak power increases monotonically with increase in CO_2 concentration indicating higher gain at higher density of active molecules. The FWHM value of the optical pulse, on the other hand, increases with reduction in the concentration of CO_2. N_2 can hold vibrational energy for longer duration due to metastable nature of its vibrational levels and any increase in its concentration, therefore, leads to the flattening of the pulse. Such large variation of the peak power and pulse width is a direct consequence of utilising only molecular gases for the operation of this laser. The same laser when operated with the conventional gas mixture (consisting of helium) exhibits much smaller variation in the pulse duration with changing partial pressures of the molecular gases. As can be seen from Figure 6, the FWHM value of the laser pulse when the present system is operated with conventional gas mixture changes at most by a factor of ~1.6 (from 800 ns to 1350 ns). It is also seen from Figure 6 that the

maximum peak power obtainable in this case is in general lower than that obtainable with helium free operation. The maximum value of small signal gain for this system has been measured to be ~2.7 %/cm. The optimum value of the inductance depends on the partial pressure of CO_2 in the gas mixture and ranges from 0.5 to 6.5 μH.

The helium free operation of TEA CO_2 lasers has been achieved in systems with widely varying active volumes, electrode geometries and preionising conditions. The fact that expensive and scarce helium is not a constituent of the laser gas mixture makes it an ideal system for high repetition rate operation as the gas re-circulatory loop consisting of the heat exchanger and the catalytic re-converter, a mandatory requirement for conventional systems to recycle helium gas, can be readily dispensed with here. In the next section the construction and operation of a TE laser pulser that performs the crucial job of subjecting the gaseous medium to a transverse electric discharge leading to the inversion of population is described. The important features of such a pulser that would make it ideal for repetitive applications and also the work carried out by us till now in this area are also described.

1.1.2. The Repetitive Pulser

The function of a pulser in the operation of a pulsed gas laser begins with the drawing of energy from the source and ends with the realisation of most of this as the internal energy of the (lasing) gas. Undoubtedly, therefore, the overall efficiency of the laser depends quite strongly on the performance of the pulser. In the operation of pulsed gas lasers, energy is initially stored in a condenser, which is then made to discharge rapidly into the laser load with the help of a fast high voltage high current switch. The performance of the pulser during the charging process thus dictates the wall plug efficiency of the laser. Lesser is the energy expended by the pulser while drawing energy from the source, the better is the efficiency of the laser.

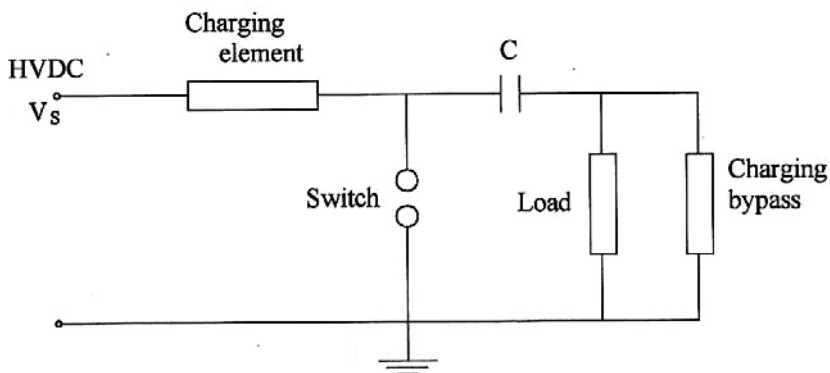


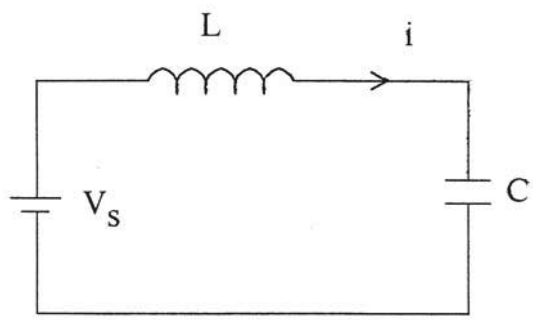
Figure 7. A typical pulse generator for a TE gas laser.

A typical pulse generator for a TE laser is shown in Figure 7. A DC high voltage supply charges up a condenser through a charging element, normally a resistance or an inductance. The charging bypass provides a path for the charging current. Once the condenser is charged to the required voltage, the rapid closure of the high voltage and high current switch enables the condenser to deliver its stored energy into the laser load before glow to arc transition can occur. The charging bypass must offer an impedance that is many times more than that of the

laser load lest this should eat up a significant fraction of the energy stored in the condenser lowering, thereby, the plug in efficiency of the laser. At the same time its impedance should be much less than that of the charging element so that the current flowing through the conducting switch from the source following a discharge can be kept low for a given repetition rate. In the single shot operation, the condenser is normally charged resistively and a spark gap is traditionally used as a switch. For repetitive operation, however, more efficient charging by means of inductance is employed and a thyratron replaces the spark gap. Following a discharge when the switch is still in a state of conduction, all the charging current delivered by the power supply gets diverted through this conducting switch. This current must be kept lower than the hold over current of the switch to ensure its recovery. Low charging current limits the maximum achievable repetition rate from the pulser. An ideal pulser would be one that allows recovery of the switch however high the charging current, and, in turn, the repetition rate is. A so-called 'ideal TE gas laser pulser', the repetitive operation of which is not hindered by the recovery problem of the switch, is discussed at length here. It is imperative that an account of this work is preceded by a brief review of the D-C resonant charging scheme on which the operation of a conventional repetitive TE gas laser pulser is normally based. Interested readers are referred to a review article [18] for a deeper insight to the various aspects of charging and discharging processes normally operative in a TE-laser. To be noted here that for short-lived discharges, that are typical of TE gas lasers, it has been shown analytically and experimentally that the recovery of the switch is independent of the energy deposited into the load [21].

1.1.2.1. Direct - Current (D-C) Resonant Charging

The main condenser can be charged to the required voltage, ranging within tens of kV depending on the type of the TE laser load, normally in two different ways: resistively or resonantly by a DC source although resonant charging of a condenser by an AC source is also not uncommon [18]. Resistive charging not only suffers from poor charging efficiency but also offers low charging frequency as the recovery problem of the switch assumes greater significance here [18]. The charging of the condenser through an inductance, commonly known as D-C resonant charging, finds wide application in the repetitive operation of TE gas lasers because of its inherent high plug-in efficiency [22] and high repetition rate capability [23].



a

Figure 8 continued on next page

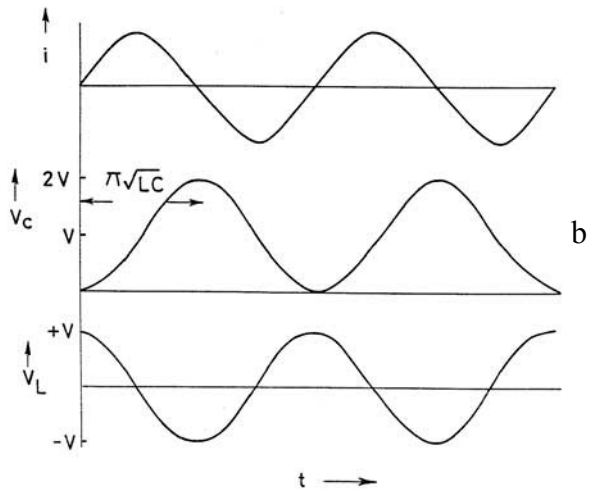


Figure 8. a) A typical D-C resonant charging network, b) the charging current (i), the voltage across the capacitor (V_c), and the voltage across the inductance (V_L) as a function of time (t).

The Kirchoff's loop equations in a typical D-C resonant charging circuit (as shown in Figure8a), considering an ideal case where there is no resistance in the charging loop and a D-C source (V_s) charges the capacitance C through an inductance L, can be written as follows:

$$V_S - L\,(di/dt) - (q/C) = 0 \tag{1}$$

Substituting i = (dq/dt), we obtain

$$(d^2q/dt^2) + q/(LC) - V_S/L = 0 \tag{2}$$

The solution of the above equation can be shown to be

$$q = V_S C\,(1 - \cos(\omega t)) \tag{3}$$

Where ω, the resonant frequency of the circuit, is given by

$$\omega = 1/(\sqrt{(LC)}) \tag{4}$$

The expressions for current (i) and voltages across the condenser (V_C) and the inductance (V_L) can be worked out to be,

$$i = dq/dt = V_S C\omega \sin(\omega t) \tag{5}$$

$$V_C = V_S\,(1\text{-}\cos(\omega t)) \tag{6}$$

$$V_L = V_S \cos(\omega t) \tag{7}$$

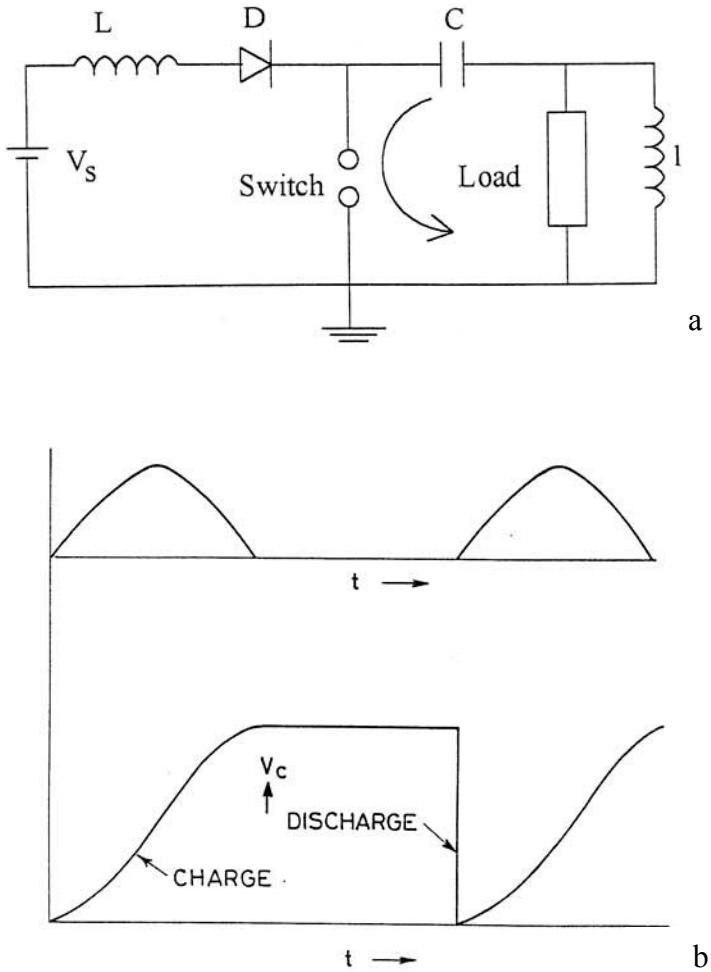


Figure 9. a) A typical resonantly charged TE laser pulser, b) the charging current and voltage across the condenser.

The current and the voltages represented by the above equations are illustrated in Figure 8b. Understandably, the sum of voltages across the inductance and the condenser at any instant equals the supply voltage V_S. It can be seen from this Figure that during a half cycle when the charging current is in the forward direction, the condenser acquires a voltage twice that of the supply. During the next half cycle, the current flows in the reverse direction and the condenser loses all its charge. If, however, a diode were introduced in the circuit (Figure9a) to arrest the flow of reverse current the condenser would then retain its voltage. This Figure also shows the discharge path of the condenser that includes the switch and the load. As the switch closes, condenser discharges through it into the load and the voltage across it drops to zero. The current flows in the forward direction once again and the condenser acquires twice the supply voltage and so on (Figure 9b).

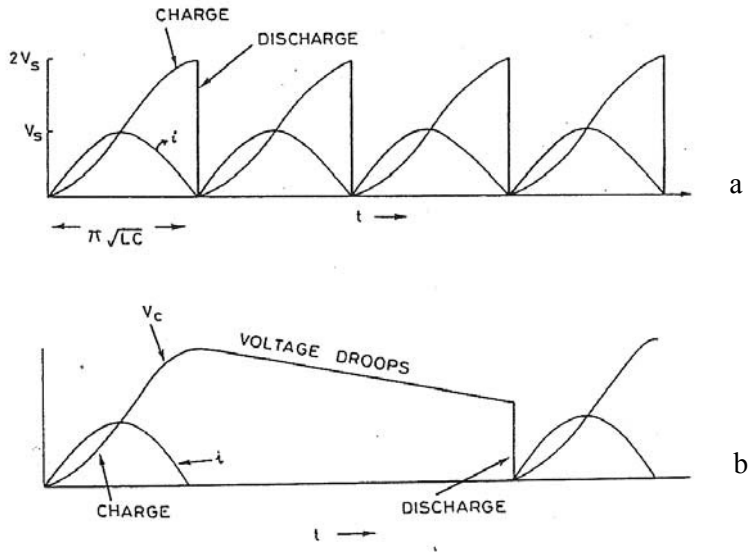


Figure 10. Few charging and discharging waveforms and the charging current (i) in a resonantly charged pulser. a: operation at 100% duty cycle. b: Operation at a lower duty cycle, the droop in the voltage across the condenser is apparent here.

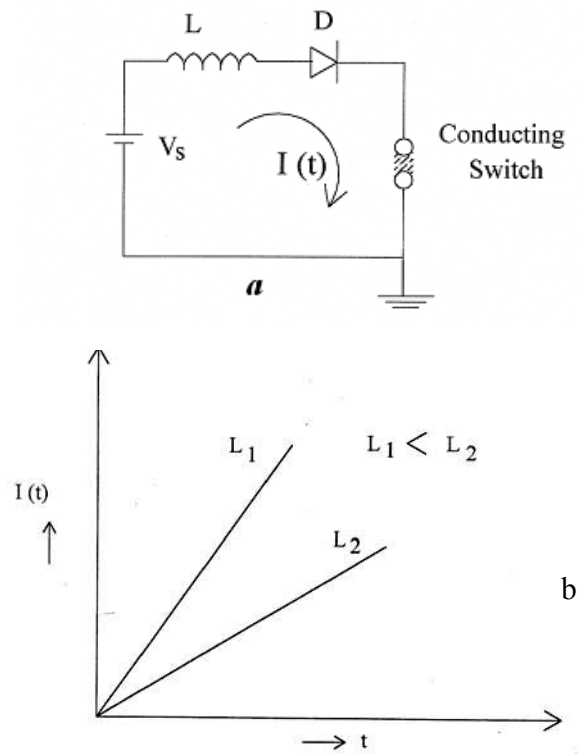


Figure 11. a: the power supply making a short through the conducting switch, b: short circuit current as a function of time for two different values of charging inductance L.

The time required for the condenser to get charged to $2V_S$ is $\pi\sqrt{(LC)}$, half the period of oscillation. Thus at 100% duty cycle, when the condenser discharges as soon as its acquired voltage is $2V_S$, the repetition rate is double the resonant frequency of the D-C resonant charging network (Figure10a). If, however, the network is operated at a lower duty cycle, i.e., the condenser discharges long after acquiring the peak voltage, the voltage across it would then droop (Figure 10b). This is because current, though small, may flow for longer duration through the voltage measuring circuitry. The voltage may also fall as a result of the flow of current due to minority carriers in the reverse biased diode. This effect would assume significance in the operation at a very low duty cycle.

Immediately following a discharge, as the switch is still in the state of conduction, the power supply makes a short through it (Figure11a). However, unlike in the case of R-C charging, where the peak charging current ($i=V_S/R$) appears immediately after the discharge, the charging current here builds up in time from zero value and can be expressed by the following equation:

$$\Delta i\,(t) = V_S \times \Delta t/L \tag{8}$$

The build up of the short circuit current (shown in Figure11b for two values of charging inductance) should be such that it does not exceed the hold over current of the switch until it's complete recovery. Higher is the inductance, slower is the rise of the current and better is the chance of the recovery of the switch although at the expense of the maximum achievable repetition rate, f_{max}, which is given by

$$f_{max} = 1/[2\pi\sqrt{(LC)}]. \tag{9}$$

As is seen from eqn.5, the same recovery condition of the switch can also be met for smaller values of V_S. However, V_S is fixed from the consideration of the laser load and cannot be utilised to control rise of the short circuit current. Thus for a particular load and a given switch, the value of L needs to be appropriately fixed to ensure recovery of the switch.

The efficiency (η) of resonant charging is given by the following expression [23].

$$\eta = 1 - \pi/(4Q), \tag{10}$$

Where the Quality factor Q is defined as

$$Q = 2\pi f_{max} L/R \tag{11}$$

R being the total ohmic component of the charging loop impedance and L the charging inductance. If $R/(f_{max}L) \ll 1$, the condition generally met while designing a resonant pulser, the charging efficiency approaches 100%. Thus almost the entire energy drawn from the source is deposited into the laser load contributing to the increased plug in efficiency compared to the case of resistive charging. As the condenser is charged to twice the supply voltage, the requirement of voltage from the source is lowered by a factor of half. The charging current arrives at a low rate starting from a zero value if the operating conditions are

chosen properly. This helps in the recovery of the switch that, in turn, allows high pulse repetition frequency (prf).

This scheme too suffers from few shortcomings. If the prf is considerably lower than the resonant frequency of the network, i.e., the pulser is operated at a low duty cycle; the discharge voltage may droop considerably. Secondly, the value of L cannot be made arbitrarily small as then the rapid build up of the short circuit current would prevent the conducting switch to go into the off state following a discharge. This effect, which limits the maximum achievable repetition rate, has been the subject of investigation in a number of studies [24,25] aimed at enhancing the repetition rate capability of a resonant charging network. The central point of these studies is to isolate the power supply during a discharge so that flow of short circuit current through the conducting switch is prevented. The most effective means of controlling the rise of short circuit current without jeopardising the repetition rate capability of the pulser network is to introduce a second switch in the charging loop, [26-28] commonly known as command resonant charging.

1.1.2.2. Command Resonant Charging

Schematic diagram of a pulser based on command resonant charging is as shown in Figure 12. A second switch S_2 is now introduced in the charging loop. The operation of the circuit can be explained as follows. When a trigger pulse T_2 arrives in the charging switch S_2, it closes and the condenser C gets charged. As the charging current becomes zero, this switch automatically turns off isolating the power supply from the rest of the circuit. At this point, a second trigger T_1 arrives at the discharge switch S_1, which then closes enabling the condenser

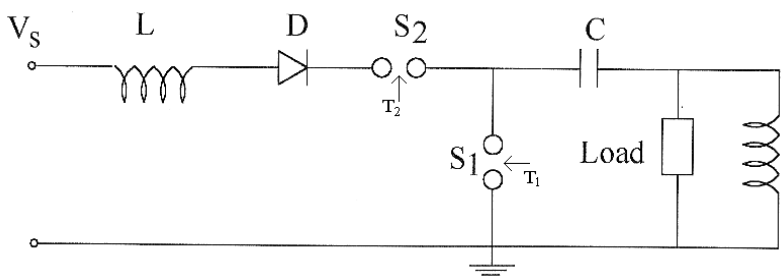


Figure 12. A typical command resonant charged TE laser pulser network.

to discharge through it into the load. During and following a discharge, the switch S_2, which is in the off state, forbids the flow of any short circuit current enabling S_1 to recover within its deionisation time. Thus in this mode of operation, maximum achievable repetition rate is not limited by the switch latch up problem. Simply lowering the value of L can reduce charge-up time of the condenser. Few charging and discharging waveforms in this mode of operation are shown in Figure13. As the condenser is charged on command here, this network can be operated at any duty cycle with almost no voltage droop. The condenser can be made to charge just before it has to be discharged. This increases the life of both the discharge switch as well as the storage capacitor as the voltage appears across them briefly. The droop free operation of this pulser has been illustrated in Figure 14 for a duty cycle of 10%. The resonant frequency shown here is 1kHz while the operating frequency is 100 Hz. 8 ms after a discharge the trigger arrives in S_2 allowing the condenser to be fully charged within 1 ms. The

condenser holds the charge for 1 ms when the second trigger arrives in S_1 causing it to discharge into the load. The voltage droop in a similar operation with conventional resonant charging network can be seen in Figure 10b.

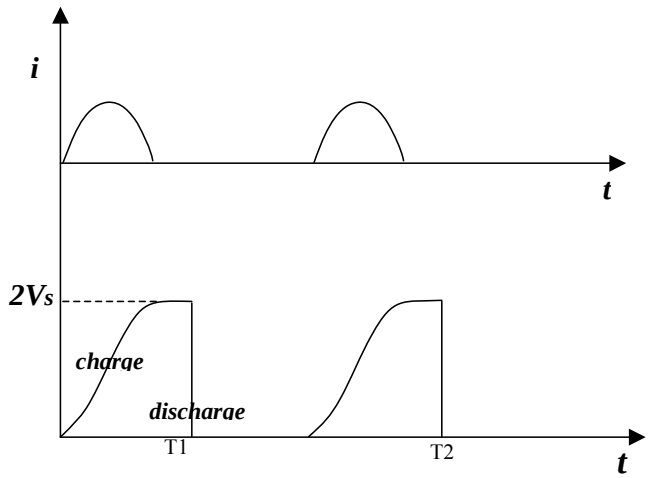


Figure 13. The charging current (*i*) and few charging and discharging waveforms in a command resonant charged pulser network.

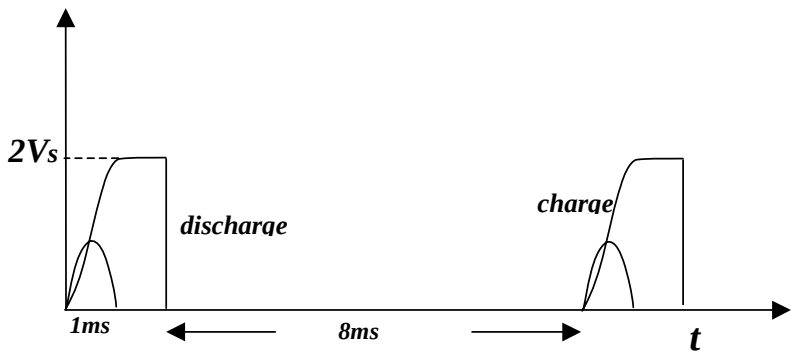


Figure 14. Droop-free operation in a command charged pulser network at a low duty cycle operation.

This method of charging too, however, is not devoid of disadvantages. Firstly, both the anode and the cathode of the charging switch have to be maintained at high voltages. High voltage isolation transformer is, therefore, required for the filaments. Much art is also needed in the design of the grid and bias circuits. False triggering of the charging switch also cannot be ruled out which can be caused by the electromagnetic noise associated with the high voltage high current discharge pulse initiated by the closure of the discharge switch. This would mean that both the switches are simultaneously in conduction once again bringing in the switch latch up problem with a much graver consequence as the value of charging inductance is kept low to make the command resonant charging high repetition rate compatible.

1.1.2.3. An Ideal Repetitive TE Laser Pulser

It is thus seen that short circuit proof operation cannot be guaranteed even for a pulser based on command resonant charging and the switches, normally two thyratrons, are thus vulnerable to damage. Thyratrons are expensive and have limited life. In the following section a novel, inexpensive, and simple switch that as a driver of a resonantly charged TE laser pulser offers complete latch proof operation is discussed. By exploiting the unique geometry of this device, a number of advantages have been achieved that even a command resonant charged pulser cannot match [29-34]. The heart of this device is a suitably conFigured circular dielectric plate that rotates between the electrodes of an ordinary spark gap. Such rotation intrinsically isolates the power supply from the rest of the circuit during a discharge and, as explained below, is instrumental in making a TE-laser pulser driven by this switch a near ideal one.

1.1.2.3.1. Rotating Dielectric Spark Gap

The rotating disc, the most important part of this switch, is shown in Figure 15. The disc is a circular plate of ~20 cm diameter with even number of equidistant holes (~6mm diameter) drilled along a circle close to its periphery. By mounting the disc on a motor, it can be so rotated that the holes pass symmetrically between the electrodes of the spark gap. The operation of this switch can be easily understood by considering a resistively charged pulser as shown in Figure 16. Within the travel time between two adjacent holes the condenser C gets charged close to the supply voltage $V_{s,}$ which is more than the air breakdown voltage of the spark gap. Every time a hole appears between the electrodes of the spark gap, it closes allowing the condenser to discharge into the load. Following a discharge, the appearance of the dielectric between the electrodes truncates the flow of any charging current through the switch causing thereby its forced recovery. The charging current thus can assume very high value reducing thereby the charging time of the condenser. If the rotation speed of the dielectric plate is made compatible with the charge up time of the condenser, the repetition

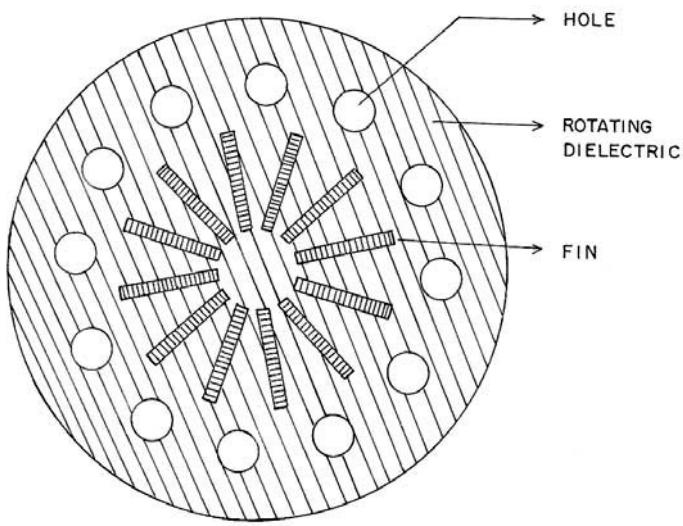


Figure 15. Schematic diagram of a rotating dielectric disc.

rate would then be accordingly enhanced. In contrast, the repetitive operation capability of a conventional spark gap when used with this pulser is restricted, as the charging current has to be less than the holdover current of the switch. In the operation with rotating dielectric spark gap as the switch, ~2kW of power was dissipated into a dummy load at a repetition rate of 300Hz [29]. The fins on the dielectric disc, when rotated, blow air jet into the spark gap and facilitate, thereby, such high repetition rate operation of the switch. In the un-triggered mode of operation, the large jitter associated with the fluctuation of the breakdown voltage marred the performance of the pulser. The triggering of the switch was accomplished by mounting a light emitting diode and a photo detector face to face on either side of the rotating dielectric in such a way that whenever a hole passed between them another hole also passed between the electrodes of the spark gap (Figure17). At every such coincidence, the detector received light from the emitter and gave out a pulse, which after processing was used to trigger the spark gap. In the triggered mode of operation the jitter significantly reduced to ~ 25 nsec [30].

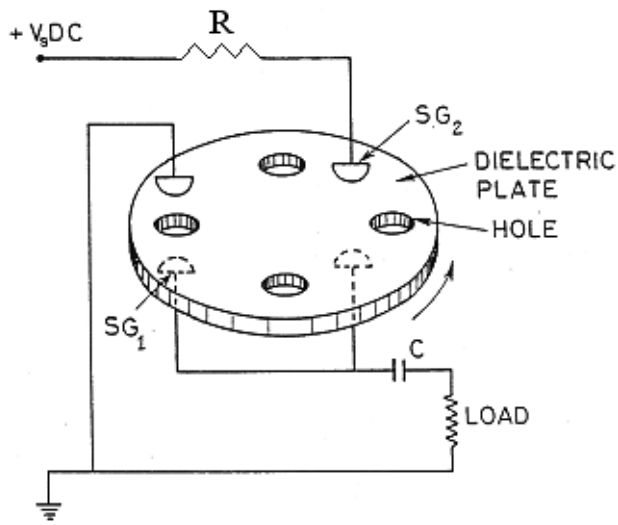


Figure 16. Resistively charged pulser network driven by a rotating dielectric spark gap.

As demonstrated in ref 31, a rotating dielectric spark gap also has latch proof operation capability when used with resonantly charged pulsers. However, such an operation is not possible utilising a simple circuit as of Figure 16 where the charging resistance is replaced by an inductance. This is because following a discharge the short circuit current (i) can build up during the travel time of the hole (t) between the electrodes of the conducting switch. For a TE laser load, this current and hence the energy stored in the charging inductor can be appreciable [31]. As the flow of this current is intercepted by the moving dielectric, the inductor releases the stored energy causing the dielectric plate to ignite. The energy stored can be reduced by increasing the value of L (as $E = \frac{1}{2} (Li^2) = (V_s^2 t^2)/(2L)$, from eqn 6). This, however, is not a practical solution as it would be at the expense of the achievable repetition rate.

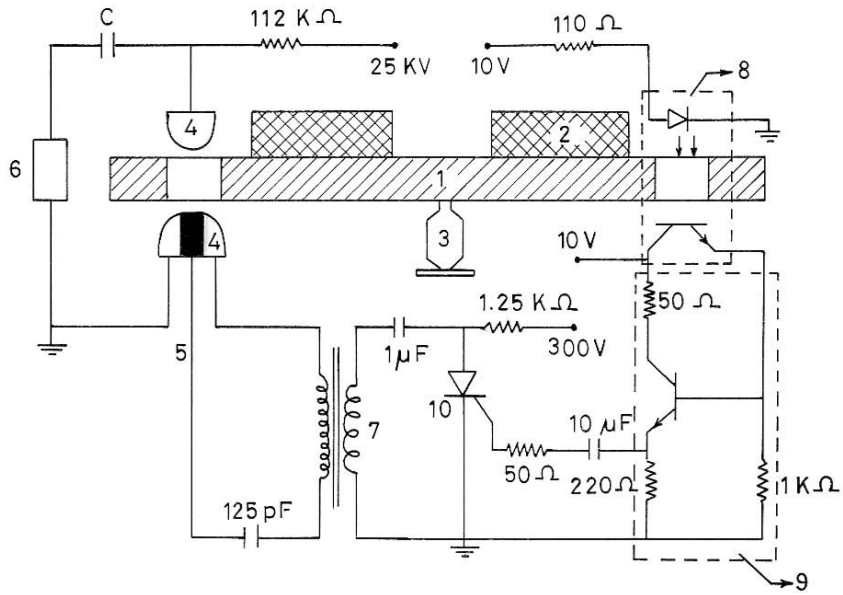


Figure 17. Schematic diagram of the rotating dielectric spark gap driven resistively charged pulser along with optical sensor based triggering circuitry. 1- rotating dielectric disc, 2- fins, 3-motor, 4- triggerable spark gap, 5- trigger, 6- load, 7- pulse transformer, 8- opto-electronic sensor, 9- current amplifier, 10- SCR.

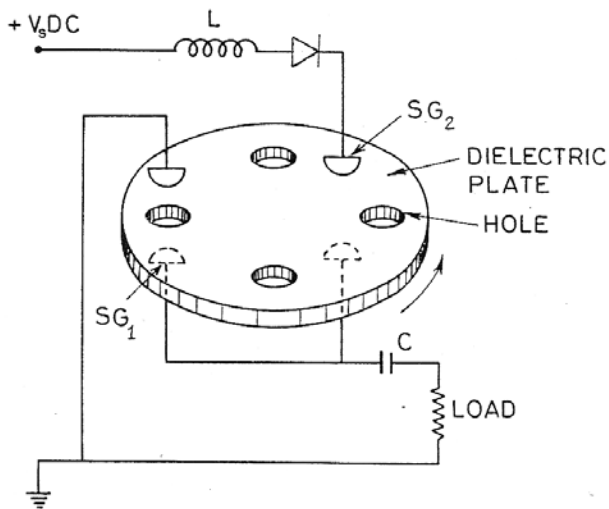


Figure 18. Short-circuit proof operation of the resonantly charged pulser when driven by a rotating dielectric spark gap.

An elegant solution to this problem is to incorporate a second spark gap in the charging loop and rotate the same dielectric plate between the electrodes of both the spark gaps in such a way that holes appear in them exactly out of phase (Figure18). The operation of this circuit can be explained in the following manner. When a hole gets aligned with the charging spark gap SG_2, it closes allowing the condenser to be charged to $2V_s$ in a time ($\pi\sqrt{(LC)}$ which is

made smaller than the travel time of the hole inside the gap by proper choice of L. After the passage of the hole, the appearance of the dielectric in the gap of this switch virtually cuts off the power supply from the rest of the circuit. A hole now appears between the electrodes of the discharge spark gap (SG_1) when it closes allowing the condenser to discharge into the load. During and immediately following the discharge, which is normally very short lived for a TE laser load, the open switch SG_2 prevents any short circuit current from flowing through SG_1. The discharge switch thus readily recovers, as the inductor now does not store any energy during its conduction. The condenser would get charged once again when a second hole gets aligned with SG_2 and so on. In this mode of operation the pulser delivered ~3.2kW of power at a repetition rate of 200 Hz into a dummy load which resembled a typical TEA CO_2 laser. Few charging and discharging waveforms shown in Figure19 conform to the above description. The jitter in the operation of both SG_1 and SG_2 is apparent from this Figure. When this pulser is used to drive a laser, the fluctuation in the closing of charging switch can be ignored as it does not affect the performance of the laser. Therefore triggering of the discharge switch alone suffices which is accomplished using the optical sensor based mechanism described in ref 29. Proper positioning of the holes with respect to the spark gaps such that the condenser discharges soon after acquiring the peak voltage ensures droop free operation with this device.

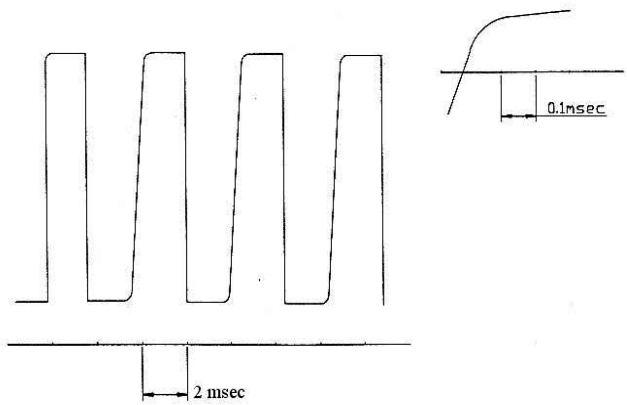


Figure 19. Few charging and discharging waveforms of the short-circuit proof operation of the resonantly charged pulser shown in Figure 18. The inset shows the charging of the condenser in an expanded time scale.

The rotating dielectric spark gap switch can also drive simultaneously two high repetition rate lasers [32]. The two lasers can be operated synchronously or with a delay that can be as large as a millisecond. Schematic diagram of the circuit is shown in Figure 20. The same dielectric is rotated between the electrodes of SG_1 and SG_2 in such a way that holes appear in them simultaneously. SG_1, triggered by the optical sensor based technique, causes the condenser C_1 charged to the supply voltage to discharge into the load R_1 while a pulse derived from this first pulser triggers SG_2 causing C_2 to discharge into the load R_2. A delay up to a few microseconds between the two discharges has been obtained by varying the value of '*l*'. The diodes D_1 and D_2 are required to isolate the two discharge circuits. A larger delay, ranging from several microseconds to more than a millisecond, has also been obtained by

positioning SG_1 and SG_2 with respect to the dielectric such that when a hole arrives in SG_1, another hole is yet to arrive in SG_2. The time interval between the arrivals of the two holes between the electrodes of their respective spark gaps is the delay between the two discharges. Triggering both SG_1 and SG_2 has considerably reduced the jitter in such operation. The voltage enhanced pulse from the optical sensor triggers SG_1 directly and SG_2 after being delayed by a delay generator. The performance of this device has been tested by switching a total of 2.5 kW of power at 200Hz into two identical dummy loads resembling typical TE lasers in terms of resistance. Though resistive charging has been employed in this operation, resonant charging can be used for better efficiency. It is noted here that this technique can be, in principle, employed for synchronisation of more than two lasers.

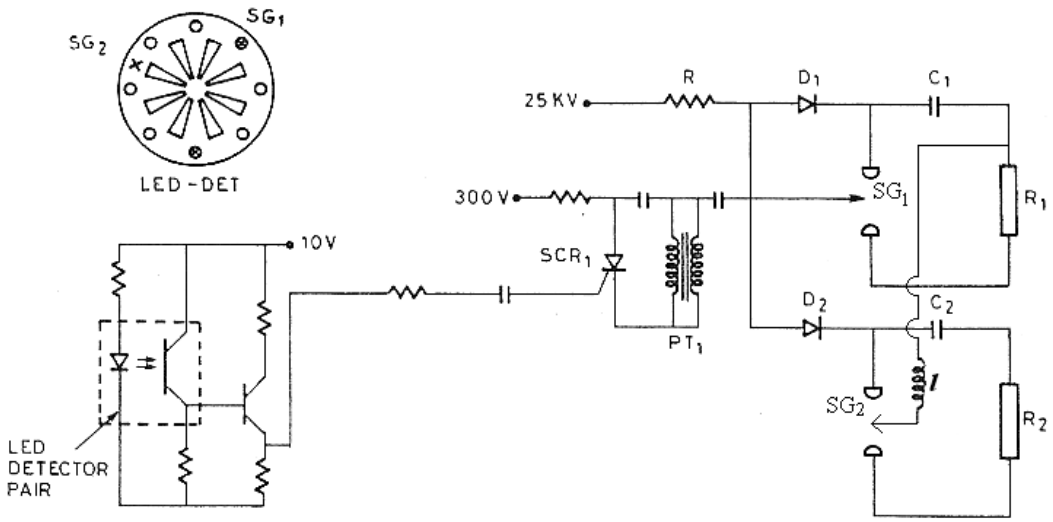


Figure 20. Rotating dielectric spark gap as a driver of two resistively charged high repetition rate lasers. Inset shows the positions of the holes with respect to the discharge gap SG_1 and SG_2.

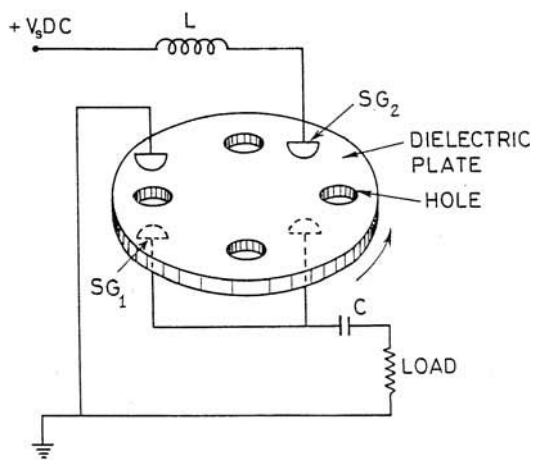


Figure 21. Diode-less operation of a resonantly charged pulser network driven by RDSG.

The unique geometry of the rotating dielectric spark gap also allows diode-less operation of a command resonant charging network [33]. In a conventional switch, thyratron, spark gap or SCR driven resonant charging pulser, the presence of a diode is mandatory to arrest the flow of reverse current so as to maintain the voltage on the condenser (refer to Figure 9a). These diodes, which should be capable of withstanding high voltages and high currents, are expensive. Moreover, they are prone to damage when used in TE laser pulsers meant for repetitive operation; in particular, in the event of a short circuit. The principle of diode less operation can be understood by referring to Figure 21. SG_1 and SG_2 are so located that when the dielectric rotates, holes appear in them exactly out of phase. As a hole gets aligned with SG_2, it closes allowing the condenser to get charged through L. If the time of passage of the hole between the electrodes of SG_2 exactly equals the time taken by the condenser to get charged fully ($=\pi\sqrt{(LC)}$) the appearance of the moving dielectric in the gap thereafter forces the switch to go into the off state preventing the flow of any reverse current thus rendering the usage of a diode superfluous. As a second hole gets aligned with SG_1, it closes and the condenser discharges into the load. A pulser has been operated in this mode at a repetition rate of 600 Hz with a dummy load resembling a typical TE laser. Few charging and discharging waveforms at this repetition rate are shown in Figure 22. It would be seen that the voltage of the condenser has dropped to about 90% of its initial value at the time of a discharge. This indicates that some reverse current had flown through SG_2 before the moving dielectric appeared between its electrodes and blocked it. Such a situation can be overcome by judicious adjustment of the charge up time of the condenser. This is possible by making use of a variable choke or alternately by adjusting the rotation speed of the motor.

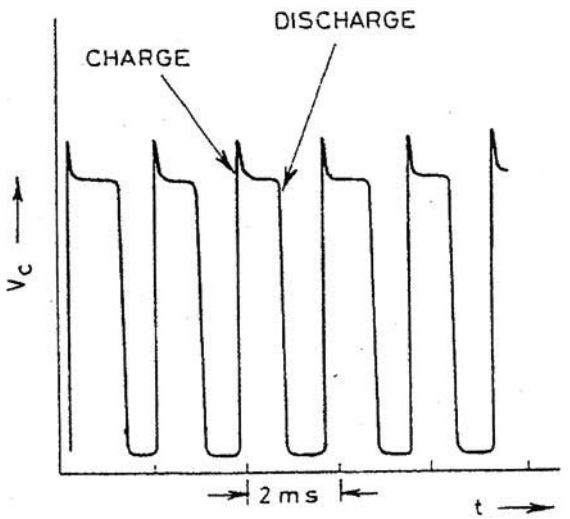


Figure 22. Few charging and discharging waveforms for the diode-less pulser.

The geometry of this switch allows an easy scalability of the maximum achievable repetition rate. The repetition rate (f) in Hz here can be written as

$$f = n \times s \tag{12}$$

where ‘n’ is the number of holes on the disc and ‘s’ is the number of rotations per second. Increasing ‘n’ or ‘s’ or both, therefore, can increase the repetition rate, however, up to a certain limit. If holes are too close the device no longer remains compatible with resonant charging. On the other hand, the speed of rotation increases at the expense of the mechanical stability of the device. Further increase in the repetition rate is possible by increasing the number of discharge gaps [34]. The schematic diagram of the circuit where a repetition rate of 1.2 kHz has been achieved with a rotating dielectric switch utilising two pairs of discharge gaps SG_2 and SG_3 is shown in Figure 23. Reliable short circuit proof operation was achieved with resonant charging by staggering the holes into an inner and an outer circle. The outer holes were aligned with the charging gap (SG_1) while the inner holes were aligned with the discharge gaps. The condensers C_1 and C_2 get charged whenever SG_1 conducts and C_1 discharges through SG_2 and C_2 through SG_3 with a delay determined by the location of SG_2 and SG_3 with respect to the passing holes. Diodes $D_1$and D_2 prevent the condensers from discharging through the same gap.

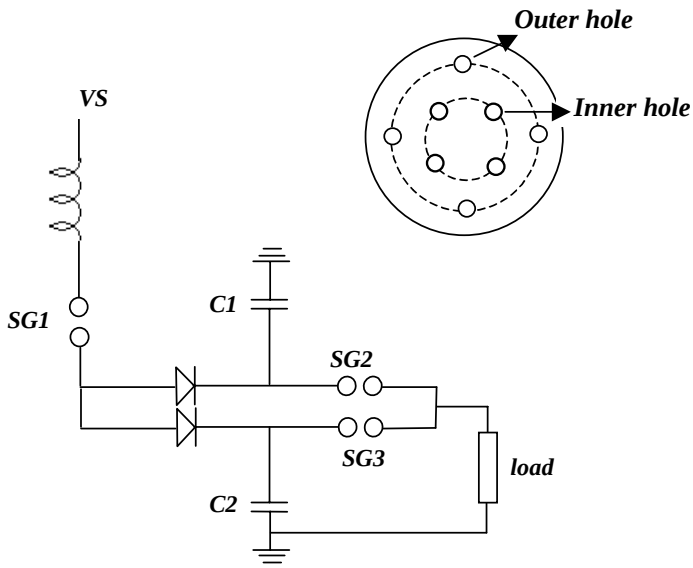


Figure 23. Schematic diagram of the pulser used to obtain kHz repetition rate with rotating dielectric spark gap. The inset shows the staggered configuration of the holes.

1.1.3. Repetitive Operation of a Helium Free Mini TEA CO_2 Laser

The single shot mode of helium free operation described in section 1.1.1.5 has been extended to the repetitive mode [35] by making use of a repetitive pulser based on the principle described in section 1.1.2.3. In this operation the inexpensive molecular gases, contaminated following a discharge, are flown out of the inter-electrode volume before the onset of the next discharge. At 50 Hz repetition rate, the laser produces an average power of ~ 1.75 W @ 35 mJ per pulse. The dispensation of the expensive and bulky re-circulatory loop consisting of heat exchanger, blower, and catalytic re-converter considerably simplifies the repetitive operation in this case.

1.1.3.1. Laser Head and Excitation Circuit

The laser head comprises of a pair of cylindrical electrodes defining a discharge of cross section 15mm x 3 mm and a length 80mm. Pre-conditioning of the laser gas mixture is accomplished by creating auxiliary spark discharges between five pair of stainless steel pins placed at a regular interval along the length of the discharge on one side at a distance of 4 cm from its center. The flow of current through each of the auxiliary sparks is limited by connecting a small capacitance in series with it. The entire assembly is housed in a leak tight Perspex chamber, the ends of which are "O" ring sealed with a concave one meter ROC gold coated mirror and a 95% reflecting Zn-Se output mirror which also formed the optical cavity of length 165 mm. Gas flows at the required rate in a direction transverse to the axis of the laser through the inter-electrode region.

The schematic of the excitation technique employed to operate the TEA CO_2 laser wherein the spiker and sustainer like actions have been integrated into a single pulser network with the help of a coupling inductance (*l*) and is as shown in Figure 2. As explained in section 1.1.1.5.1, decoupling of the two discharges viz., the spiker and the sustainer and the tailoring of the sustainer pulse by making use of the inductance results in a condition where glow discharge could be reliably obtained even in the absence of helium under wide range of operating conditions.

1.1.3.2. Experimental Results

In the repetitive operation of a TEA CO_2 laser, the maximum achievable repetition rate depends upon the rate of gas flow across the inter-electrode gap and energy loading in the discharge. For a constant flow, the pulse repetition rate can be increased until the pulsed glow discharge becomes unstable leading to the formation of arc. As major part (~90%) of the energy coupled to the discharge is realized as heat, with increase in repetition frequency the temperature of the gas increases resulting in reduced output and increased discharge instability. There are several mechanisms that transform a glow discharge into an arc discharge with increasing repetition rate [35]. The minimum gas flow rate necessary for arc free discharge as a function of energy loading and repetition frequency has been studied. It is evident from Figure 24 that the requirement of flow increases as the energy loading in the discharge increases. The pulse energy and pulse shape of the laser at different flow rates of

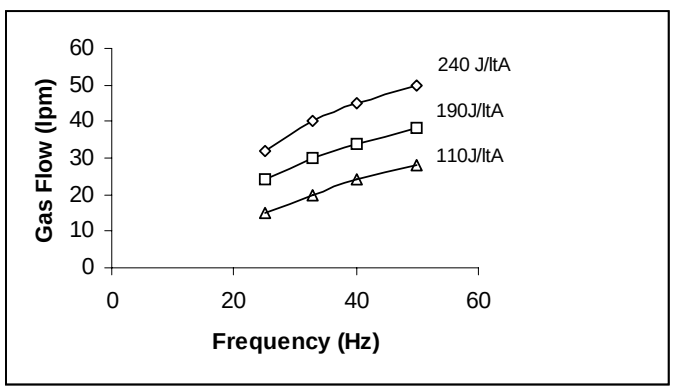
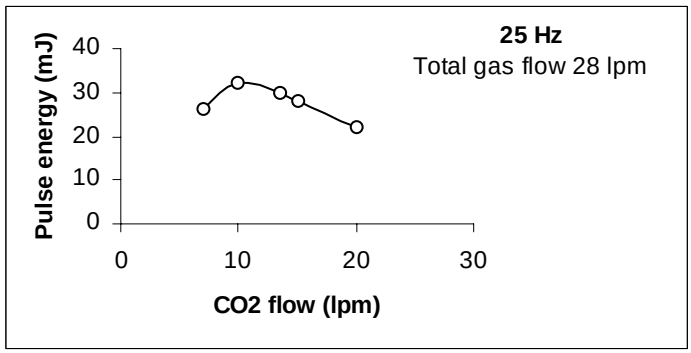
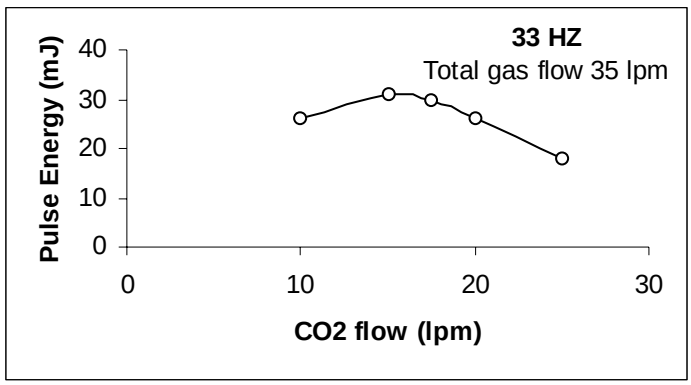


Figure 24. Minimum gas flow required for arc free discharge as a function of repetition rate for different energy loading in the discharge.

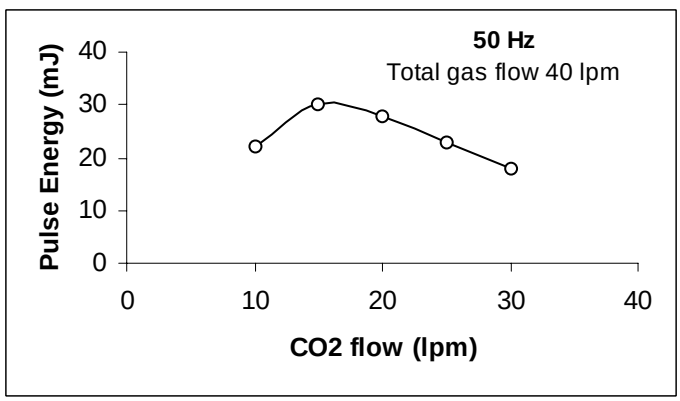
CO_2 and N_2 keeping the total flow rate constant at different frequencies were also monitored. Figures 25 and 26 show respectively the variation of pulse energy and pulse shapes with CO_2 concentration in the gas mixture. The pulse energy shows a maximal behaviour with increase in CO_2 flow rate. At lower partial pressure of CO_2, the concentration of active species is less while for its higher partial pressure, the energy storable in the vibrational manifold of N_2 reduces resulting in the reduction of output energy in either cases. The electro optic efficiency of the laser is 7.2 % up to 30 Hz but reduces to 6.8% at 50 Hz repetition rate.



a



b



c

Figure 25. Variation of laser pulse energy with CO_2 flow. a: Operation at 25Hz, b: 33 Hz, c: 50 Hz.

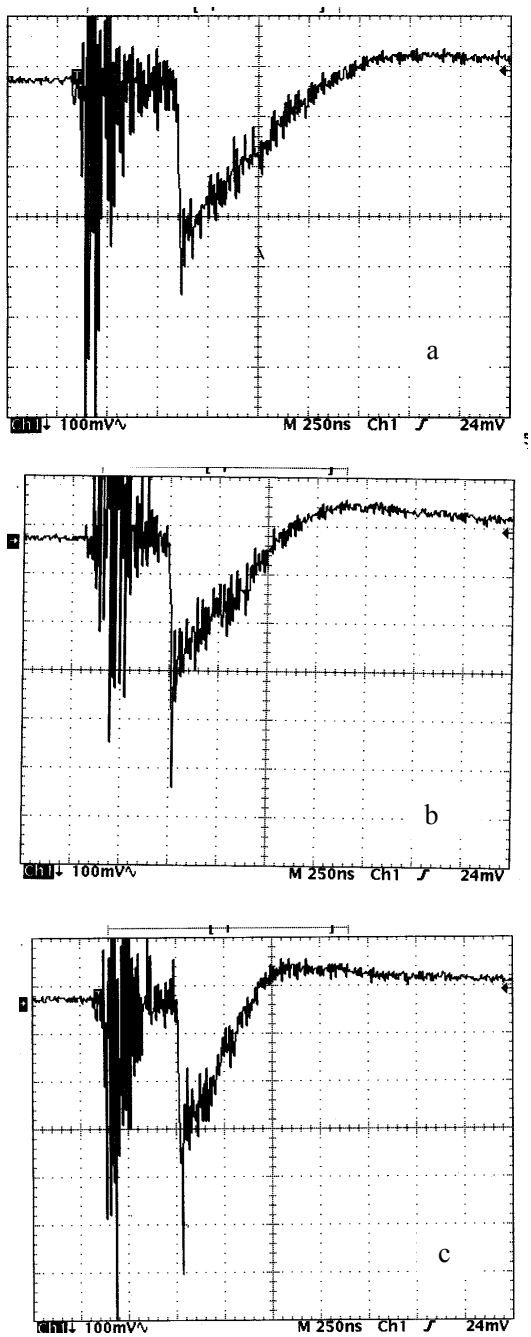


Figure 26. Temporal profile of the laser pulse obtained for different gas composition a- CO_2:N_2::1:3, b- CO_2:N_2::1:1, c-CO_2:N_2::3:1.

1.1.4. Repetitive Operation: Switch-Less Pulser

As has been discussed, in the operation of a TEA CO_2 laser, the main energy storage condenser normally discharges its energy into the load with the help of a fast, high current, high voltage switch, conventionally a spark gap or a thyratron. In the repetitive operation,

thyratron is known to be a better choice as spark gaps operate in the arc mode and suffer from recovery problem. Thyratrons, however, are expensive and have limited life. As a result there is a growing interest in replacing these switches by all-solid-state-exciters (ASSE) in conjunction with magnetic pulse compression (MPC) [36]. Although such systems have long life and high degree of reliability, they suffer from low wall plug efficiency and bulkiness [36]. Efforts have, therefore, been expended to eliminate the main discharge switch altogether in the operation of CO_2 lasers [37-39]. In these methods the main discharge condenser is directly connected across the laser electrodes and is charged to a voltage below the self-breakdown level of the laser gas mixture in the inter electrode volume. Automatic switching of this condenser occurs when the inter electrode volume is preconditioned by UV [37] or X-ray [38] photons or electrons from an external source [39]. In all these methods although the main discharge functions without a switch, the preconditioning of the inter-electrode volume is initiated by a switch. The operation of a UV preionised TEA CO_2 laser, which does not require the service of any external switch, is discussed in detail here.

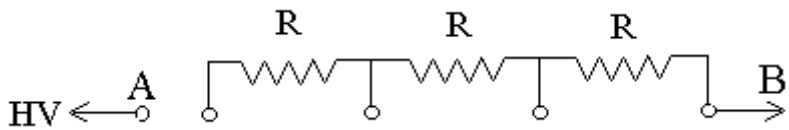


Figure 27. The configuration of the spark array cum switch.

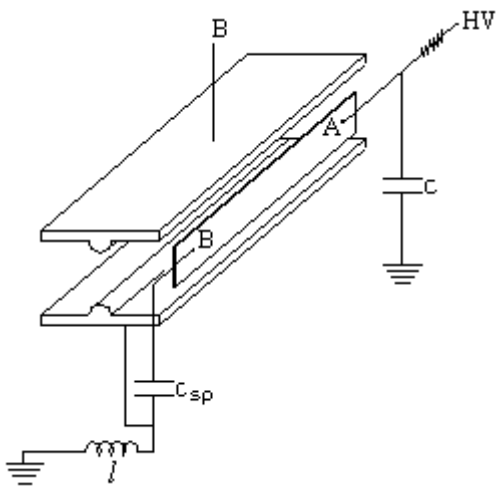


Figure 28. The laser head, the spark array cum switch and the pulser circuit.

In a typical UV preionised TEA CO_2 laser, the UV photons emanating from the spark channels placed along the length of the discharge precondition the inter-electrode region. In actual operation with a sequential spark type preioniser, all these spark channels are overvolted sequentially leading to their closure with the help of an external spark gap. The spark gap itself operates in arc mode and can be a source of UV radiation when in conduction. It is intended to make use of the UV radiation emanating from the spark gap itself for preconditioning the inter-electrode volume. This has been achieved by segmenting the main

spark gap into smaller adjustable gaps (see Figure27) and placing this integrated spark array along the length and to one side of the discharge as shown in Figure28. A resistance (R) is connected across each of the gaps. A mini TEA CO_2 laser has been successfully operated, albeit in single shot mode, where this spark array served the dual purpose of a switch as well as a source of UV photons for preconditioning of the inter-electrode volume [40].

1.1.4.1. Laser Head and the Excitation Circuit

The laser head comprises of a pair of cylindrical electrodes that defines a discharge of cross-section 0.3 cm × 1.1 cm and length 8.0 cm. The electrodes and the spark array are housed in a leak tight Perspex chamber, the two ends of which are O'ring sealed with a gold coated copper mirror (1m ROC) and a 90% reflective ZnSe output mirror defining a resonator cavity of length 16.5 cm.

The pulser circuit used for the excitation of the laser is shown in Figure 28. A condenser C is resistively charged to a suitable voltage 'V'. The resistance 'R' ensures that the entire voltage 'V' appears across the first gap of the spark array leading to its closure. The voltage then appears across the second gap resulting in closure of this gap and subsequently of all the remaining gaps in a cascade. The UV photons emanating from these sparks precondition the gaseous medium in the inter-electrode volume. Alongside preconditioning, the preionisation current also charges up the spiker capacitor 'C_{sp}' to an elevated voltage, which is impressed across the discharge electrodes. This high voltage impulse causes the breakdown of the preconditioned inter-electrode gap. The inductance 'l' provides the required delay between the pre and the spiker discharges. The remaining energy in the condenser 'C', the voltage across which is reduced as a result of powering the pre and the spiker discharges, now appears in the discharge at a rate decided by the value of 'l' causing its sustenance and excitation of the active medium.

1.1.4.2. Results and Discussion

The laser has been initially operated with a gas mixture of CO_2: N2: He:: 1:1:3 and the values of C, C_{sp}, and '*l*' were optimised respectively to 2.7 nF, 300 pF, and 3.3 μH by monitoring the output energy of the laser. While optimising the gas mixture it has been found that the operating voltage of the laser too changed. This is because the operating voltage of the laser is determined by the break down voltage of the first gap in the spark array and as the array is exposed to the gas mixture any change in its composition also changed the breakdown voltage and, in turn, the operating voltage of the laser. The most optimised and reliable performance in terms of energy (viz., 38mj @ 7 % electro-optic efficiency) occurs for a gas mixture of CO_2: N2: He:: 1:1:2.5 for which the operating voltage is ~20 kV. To be noted here that the energy expended in preionisation has not been accounted for while calculating the efficiency. By virtue of its short cavity length, the laser emission is expected to occur on a single longitudinal mode, which is corroborated by the absence of any beating at the cavity round trip time (t_R) in the temporal profile. In order to remove the dependence of the operating voltage on the gas composition, the preioniser cum switch needs to be isolated from the laser head by making use of appropriate windows for coupling the uv radiation into the inter-electrode volume.

One obvious disadvantage of this sequential spark preioniser array is that the entire discharge current flows though the preioniser. This would limit the life of the array in repetitive operation. A possible remedy would be to operate the laser by making use of a

parallel spark array preionser where the main discharge current is shared by all the spark channels. Simultaneous closure of a number of parallel gaps without the need of an external switch can be achieved by ballasting the gaps with mutually coupled inductances [41].

1.2. Single Mode Operation

A TEA CO_2 laser emits, normally, on a number of longitudinal and transverse modes [42]. The large homogeneous gain broadening at atmospheric pressure results in near equal gain on a number of modes close to the line centre. Therefore, the laser, when energised, aided by the spatial hole burning effect that is inherently associated with a Fabry-Perot cavity [43], emits on a number of longitudinal modes. For many applications e.g., laser isotope separation, LIDAR, optically pumped mid and far infrared lasers, nonlinear optical interaction in semiconductors and gases etc, the emission from a TEA CO_2 laser is required to be on a single longitudinal and transverse mode. Usage of an intra-cavity aperture of appropriate size ensures operation on single transverse mode. To achieve SLM operation, however, a number of techniques have been developed in the past [42, 44-49]. The most widely used methods viz., hybridization [48] and injection locking [42, 49], make use of the narrow gain broadening in a low pressure section to control the nature of emission of the high pressure section. In case of injection locking, the stimulated emission in the high-pressure laser is initiated by the SLM output of a low-pressure laser. An important precondition to obtain SLM output here is that the cavity lengths of both the low and high-pressure lasers have to be matched and the seed photon intensity needs to exceed a threshold value. A hybrid laser, on the other hand, consists of low and high-pressure sections sharing the same resonator cavity. Oscillation sets in on the longitudinal mode that is supported by both the sections and, understandably, the emission of the hybrid laser occurs on SLM. Since the useful discharge cross-section of the TEA section is restricted by the cavity mode determined by the narrow bore tube of the cw section, the efficiency of a hybrid laser is generally poor. Efforts have been made in the past, although with limited success, to increase the efficiency of a hybrid laser by using short, wide bored, RF excited cw gain tube in place of the conventional narrow bore discharge tube [50].

1.2.1. Single Mode hybrid CO_2 Laser with Increased Efficiency

Almost 50% improvement in the operating efficiency of a hybrid CO_2 laser has been achieved by restricting the energy deposited in the TEA discharge into a narrower central zone [51]. The energy that appears otherwise in the unutilized volume of the TEA discharge earlier, now appears in the region seen by the oscillating TEM_{00} mode, increasing, thereby, the output energy and hence the operating efficiency of the laser. The restriction in the width of the discharge has been possible by making use of molecular gases viz., CO_2 and N_2, alone in the laser gas mixture. In presence of helium, the discharge tends to be wider due to the decreased sensitivity of the net electron growth rate to electric field variation [52]. Helium free operation thus makes discharge cross-section narrower resulting, in turn, in a better TEM_{00} mode-filling factor. To be noted here that the helium free operation of a conventional TEA CO_2 laser was discussed in section 1.1.1.5.

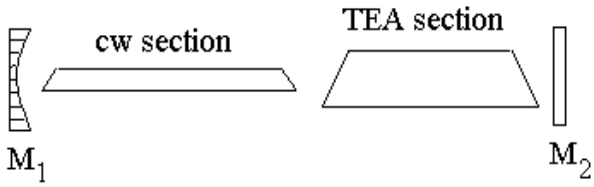


Figure 29. Schematic diagram of a hybrid CO_2 laser.

The schematic diagram of the hybrid laser is shown in Figure 29. A 60 cm long water-cooled Pyrex tube of internal diameter 10mm, filled with CO_2:N_2: He:: 1:1:5 to give a total pressure of ~ 15 mbar, the two ends of which are 'O' ring sealed with ZnSe Brewster windows, forms the low pressure active section. Two Ernst profiled electrodes defining a discharge of cross-section 1.5x1.5 cm^2 in case of conventional gas mixture (and ~ 1.5 x 1.2 cm^2 with helium-free gas mixture) and length 30 cm housed in a leak-tight Perspex chamber form the TEA laser head. Pre-conditioning of the laser gas mixture is accomplished by creating auxiliary spark discharges between brass pins placed at a regular interval along the length (14 pairs on either side) of the discharge and at a distance of 5 cm from its centre. A gold coated mirror M_1 (~100% R, 6 m ROC) and a ZnSe output coupler M_2 (70% R, plane) separated by 154 cm defines the resonator cavity. An aperture of diameter ~6mm placed near the plane mirror ensures oscillation on TEM_{00} mode. The gas composition in the TEA section that is devoid of any helium is CO_2:N_2:: 1:1.

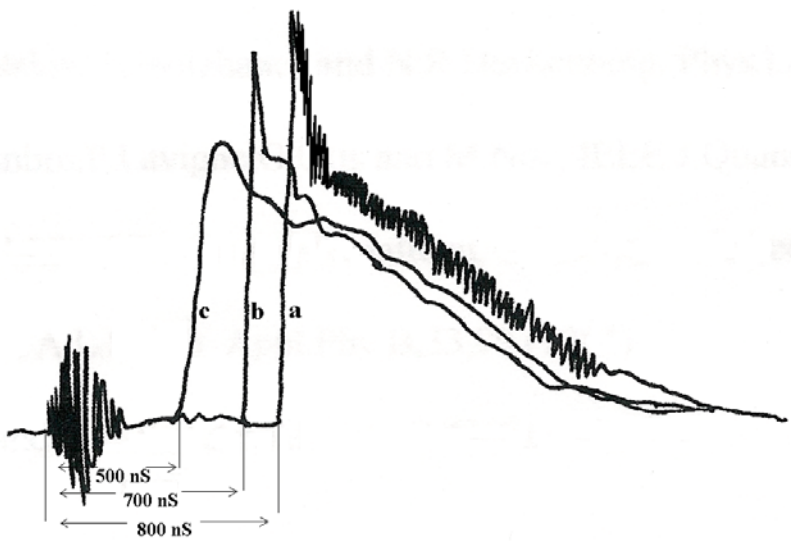


Figure 30. Temporal profile of the emission from a hybrid laser. a) cw section off, b) cw section below threshold, c) cw section above threshold.

The TEA section is energised by a novel spiker-sustainer scheme described in detail in section 1.1.1.5.1. The emission from the TEA laser when the cw section was switched off is shown in Figure 30a. The period of oscillation is found to be 3.5 nsec (~t_r/3) indicating the simultaneous lasing of four cavity modes. The laser pulse is delayed by ~800 nsec with respect to the current pulse. As expected, the temporal shape of the emitted pulse changes

when the cw section is also energised. In case of it's below threshold operation (I<3 mA), lasing occurs with a delay of ~700 nsec with respect to the current pulse and is devoid of any beating indicative of SLM lasing (Figure 30b). When the cw section is operated above threshold (I>3 mA), the pulse, with its peak flattened, starts much sooner (delay is ~ 500 nsec, Figure 30c).

The output of the hybrid laser is ~ 40 mJ/pulse as against ~27 mJ per pulse when it is operated under identical conditions but with a gas mixture containing helium as well in the TEA section. This is due to the fact that the discharge width in the TEA section spreads in presence of helium and a part of it only is filled by the oscillating mode. The volume of this oscillating mode, which is decided by the length and the inner diameter of the cw section, in turn, decides the output energy of the laser. The delay between the excitation pulse and the onset of lasing in case of operation with helium-free gas mixture is always considerably less as compared to the operation with conventional gas mixture for the same operating conditions indicating the enhancement of gain in the former case.

Figure 31. Spatial profile of the emission of a hybrid laser as recorded on a thermal sensitive paper when a) TEA section operates with conventional gas mixture, b) TEA section operates without helium.

The fact that the discharge cross-section in the TEA laser changes significantly with the presence/absence of helium in the gas mixture is corroborated by the burn patterns of the TEA laser taken on a thermal sensitive paper. Figure 31a shows the burn pattern of the TEA CO_2 laser emission with conventional gas mixture while 31b shows the burn pattern for the laser utilising only CO_2 and N_2 gases. It is evident from these burn patterns that the discharge width is much more confined to the central zone of the inter-electrode volume of the TEA laser in case of helium free operation.

A hybrid CO_2 laser is, therefore, operated with a ~50% improvement in its efficiency by utilizing helium-free gas mixture in the TEA section that resulted in a higher TEM_{00} mode filling factor and, in turn, improved efficiency. The single mode obtained in this manner, however, suffers from limited detunability owing to the narrow broadening of gain in the low-pressure section of the hybrid laser. In the following section a more elegant method of obtaining single mode emission from a TEA CO_2 laser wherein the detunability of the single mode is limited by the free spectral range of the cavity alone is described

1.2.2. Single Mode Lasing From a TEA CO2 Laser by the Elimination of Spatial Hole Burning Effect

As discussed in section 1.2, a TEA CO_2 laser emits on a number of axial modes. SLM operation, however, is an essential prerequisite for a number of applications and many

techniques, e.g., usage of intra-cavity etalons [44], saturable absorbers [45], modified interferometers in Fabry-Perot and ring cavities [53], short cavity lasers [47], hybridisation [54] and injection locking [42] etc have been evolved to force lasing on SLM. Most of these techniques rely on either increasing the gain of the central mode or inflict loss on all the neighbouring modes with respect to the one on which lasing is sought. Hybridisation and injection locking, that have emerged as the most popular methods of obtaining SLM lasing, make use of an additional low-pressure gain medium that adds to the complexity of the system. Further, the detunability offered by such systems is limited by the narrow broadening in the low-pressure gain cell. While the presence of a low pressure cell drastically reduces the operating efficiency of a TEA laser in the hybrid configuration due to the aperture effect, the synchronisation of the master and slave cavity lengths in an injection locked system requires the passive stabilisation of the two cavities in addition to the provision of fine tuning of their lengths.

Spatial hole burning effect, associated with a standing wave cavity, is known to aid simultaneous oscillation on a number of ro-vibrational transitions in a CO2 laser [55]. The same effect is also believed to be responsible for the multi-mode emission from a high pressure pulsed CO_2 laser [43] although no such observation, either theoretical or experimental, has been reported so far in the literature. Spatial hole burning helps sustain oscillation on a number cavity modes lying within the lasing domain as they, in presence of this effect, can interact with inverted population unutilised by the others. Elimination of spatial hole burning effect has been shown to automatically cause a TEA CO2 laser to emit on SLM by virtue of its predominantly homogeneous broadening nature of the gain [56]. This has been possible by making use of a ring cavity in which unidirectionality is imposed by locating inside the cavity a saturable absorbing medium of optimised pressure and length. The uniform field along the length of the gain medium in a traveling wave cavity precludes the formation of spatial holes in the inverted population by the oscillating mode. The neighbouring modes are automatically suppressed as the inverted population is uniformly burnt by the mode on which oscillation sets in first. This, to our knowledge, is the first experimental demonstration of the fact that spatial hole burning is responsible for multi-axial mode lasing in a high pressure pulsed CO_2 laser. The only other effort to this end by Lyon and George [57] has remained inconclusive as acknowledged by the authors themselves. This is due to the fact that oscillation in the ring cavity used by them was not forced into traveling wave condition and consequently the spatial effect could not be eliminated resulting in the laser emitting on two longitudinal modes.

The method of generation of SLM described here does not require the incorporation of any extraneous active element and the detunability of the emitted mode here is limited by the free spectral range (FSR) of the cavity alone. To be noted here that intra-cavity saturable absorber has been used to eliminate spatial effect in the past in order to study the rich dynamics of single mode cw CO_2 ring lasers [58, 59].

The schematic of the experimental system is as shown in Figure 32. A concave grating G (6m ROC, 150 lines/mm), a gold coated mirror (M_1) and a 70%R ZnSe output mirror (M_2) form the ring cavity of length 1.95 m. This dispersive ring cavity is aligned to lase on 10P(16) line by making use of a reference cw CO_2 laser and a He-Ne laser. The procedure of alignment of this ring cavity is described in reference 60. The TEA CO_2 gain medium is placed in the arm containing the grating and the gold mirror. The laser head comprises of a pair of Ernst profiled electrodes defining a discharge of cross-section 2.5×1.5 cm^2 and length

40 cm and is located inside a leak-tight Perspex chamber with Brewster windows at the two ends. The detailed description of the laser can be found in reference 61. An SS cell of length 10 cm filled with SF_6 gas at controllable pressure and 'O' ring sealed at the two ends by ZnSe AR/AR windows is placed in the arm containing the output coupler and the gold mirror.

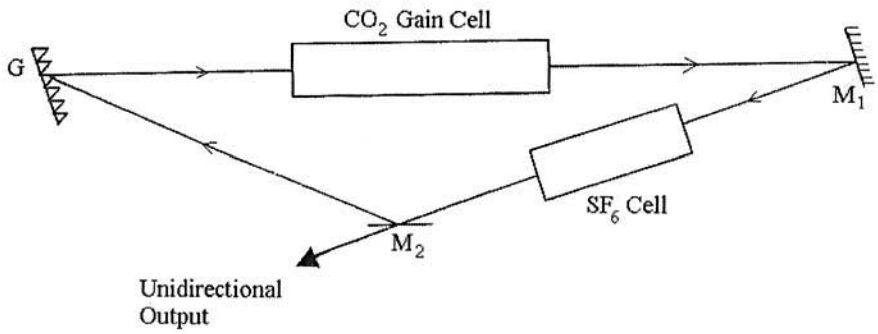


Figure 32. Schematic diagram of the ring cavity single mode CO_2 laser.

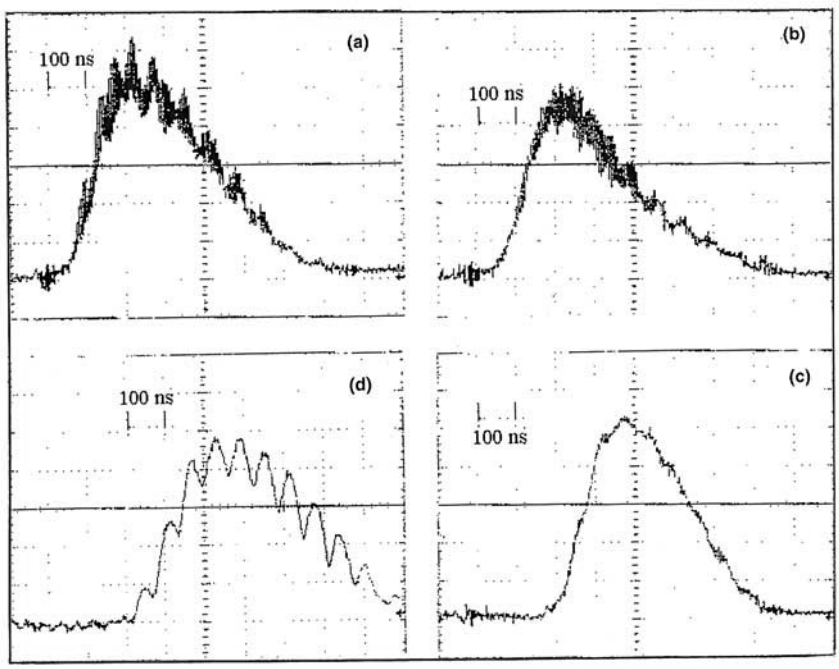


Figure 33. The temporal profiles of the laser emission, a) in the absence of both intra-cavity aperture and saturable absorber, b)intra-cavity aperture present, saturable absorber absent, c) intra-cavity aperture present, saturable absorber present, d) intra-cavity aperture absent, saturable absorber present.

In the absence of any absorber in the cell, the laser is found to lase bi-directionally on multi-axial modes as is evident from the beating occurring at a period of cavity round trip time (t_r) in the temporal profile of its emission (Figure 33a). The modulation at a much longer period (~$10t_r$) seen in this trace is due to the beating of transverse modes. This can be

removed by the insertion of an intra-cavity aperture that forces lasing on a single transverse mode (Figure 33b). At this point, SF_6 is introduced into the intra-cavity cell. At a cell pressure of ~0.6 mbar, the emission is found to be totally devoid of any beating- a signature of single longitudinal and transverse mode lasing (Figure 33c). In addition, the lasing now occurs only in one direction. The total energy output, however, remains nearly the same. With fully opened aperture SLM but multi transverse mode lasing is observed (Figure 33d).

The experimental observations can be explained in the following manner. In the absence of any intra-cavity saturable absorber, as the gain medium is energised, the two waves traveling in opposite directions grow equally giving rise to standing wave formation resulting in multimode bi-directional lasing. However, in presence of SF_6 beyond a certain pressure, the wave M_1GM_2 that enters the cell after suffering diffraction and transmission losses at the grating and the output coupler respectively is much weaker compared to the wave propagating in the opposite direction (M_2GM_1) and is unable to saturate the absorber leading to its annihilation. This results in intra-cavity traveling wave condition giving rise to the observed unidirectional SLM and single transverse mode lasing (Figure 33c). The removal of the intra-cavity aperture at this juncture results in SLM but multi-transverse mode output as is evident from long beat period in the temporal shape of the emission (Figure 33d). This clearly shows that the transverse modes, that anyway utilise different regions of the active volume by virtue of their different spatial field profiles, are unaffected by the formation of a traveling wave cavity. The interaction volume of different longitudinal modes, on the other hand, overlaps entirely along the axis of the cavity. The uniform axial field intensity in a traveling wave cavity precludes the formation of spatial holes in the inverted population. The longitudinal mode possessing the highest gain, therefore, alone survives the competition and gets to lase. Reversal in the direction of lasing can be easily achieved by placing the absorber cell in the arm connecting the grating and the laser. To be noted here that when the SF6 cell is placed in the arm containing the grating and the output coupler, bi-directional lasing is observed for the same operating parameters of the cell viz., the length and the absorber pressure. This is because both the counter-propagating waves, in this case, experience loss; one at the grating and the other at the output coupler before entering the SF_6 cell. The intensity of the counter propagating waves even after experiencing loss is such that the absorber saturates in both directions resulting in bi-directional lasing. Since the losses in the two directions are not identical, to obtain unidirectional SLM lasing, the pressure of the saturable absorber needs to be increased such that one of the waves with higher intensity alone can saturate it.

In the next set of experiments the dependence of the discrimination power of the absorber on its pressure is studied. In the absence of SF_6 in the cell, the ring cavity laser produces bi-directional multi-longitudinal mode output of nearly equal energies (Figure 34a). The large modulation depth seen in the temporal profiles of the two emissions is indicative of the near equal strengths of the counter propagating waves. With gradual increase in the absorber pressure, the energy emitted in one direction (along M_1M_2) increases at the expense of the output in the other direction (i.e., along GM_2, Figure 34b) until a pressure is reached when the lasing finally becomes unidirectional (Figure 34c). The reduction in the depth of modulation in the temporal profile (Figure 34b) too points to the gradual weakening of one of the waves while the clean temporal profile of Figure 34c indicates the complete annihilation of one of the waves resulting in unidirectional emission. The wave traveling in the direction GM_2 is unable to lase as its intensity is not enough to saturate the absorber at this pressure. The

disappearance of longitudinal mode beating in a traveling wave cavity, therefore, clearly provides the first experimental evidence in support of the fact that spatial hole burning effect, unavoidable in a standing wave cavity, is responsible for multi-longitudinal mode emission from a TEA CO_2 laser. To be noted here that the SLM emission from the ring cavity occurs on 10P(16) line. The TEA CO_2 laser is tuned to this transition as SF_6 exhibits a strong absorption at this frequency. To obtain SLM lasing on any other CO_2 laser transition a saturable absorber with suitable absorption characteristics needs to be chosen.

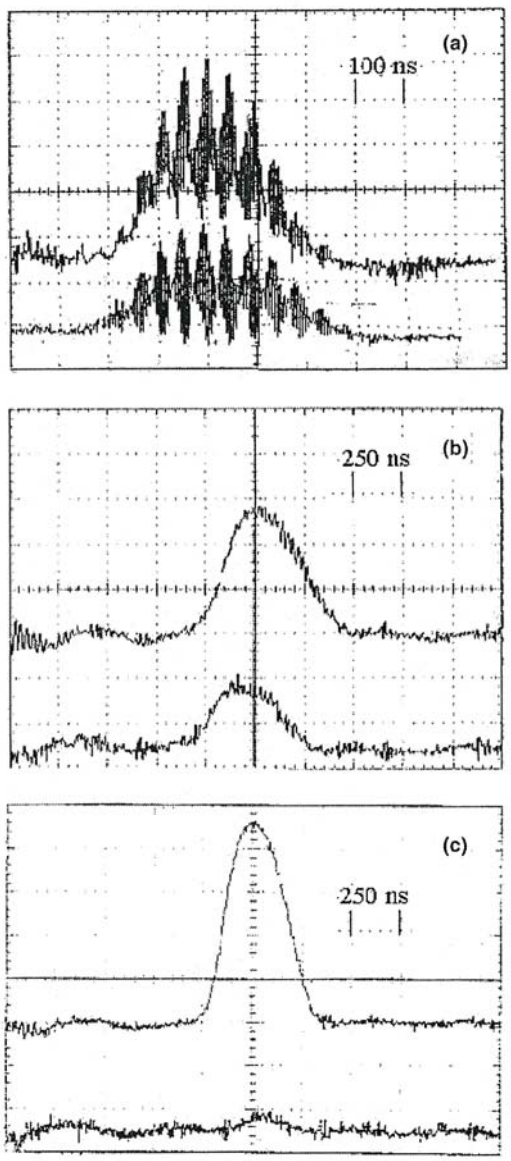


Figure 34. Temporal power profiles of the bi-directional emission of the ring laser with increasing pressure of the intra-cavity saturable absorber. a: no saturable absorber, multi-axial mode emission, b: saturable absorber at an intermediate pressure, stronger emission in one direction, c: saturable absorber with appropriate pressure, unidirectional emission.

An efficient alternative method of obtaining SLM emission from a TEA CO_2 laser by making use of a traveling wave ring cavity has been described in this section. This observation is an experimental demonstration of the fact that spatial hole burning effect associated with standing wave cavity is responsible for the multi-longitudinal mode emission from a high pressure CO_2 laser. Since no cw section is made use of here, the aperture effect and the problems related to synchronisation of two high voltage discharges are automatically eliminated. In addition, a wider detunability of the mode compared to hybrid or injection-locking methods can be achieved simply by tuning the cavity length as it is governed by the FSR of the cavity alone.

1.3. Conclusions

The two aspects of TEA CO_2 lasers dealt primarily in this article are the repetitive and SLM operation in the light of new results obtained by us in these areas. The helium-free operation of a conventional TEA CO_2 laser has been possible by the clever integration of the spiker and the sustainer discharges with the help of a simple coupling inductance. Helium-free operation has a strong bearing in the repetitive application as the inexpensive molecular gases, contaminated following a discharge, can be flown out of the inter-electrode volume before the onset of the next discharge. The total dispensation of the expensive and bulky re-circulatory loop consisting of heat exchanger, blower, and catalytic re-converter, an integral part of the high repetition rate TEA CO_2 laser using conventional gas mixture, is now possible which considerably simplifies its repetitive operation. The electro-optic efficiency of the helium-free laser when operated on TEM_{00} mode is always found to be higher as compared to its operation with conventional gas mixture. This increased TEM_{00} mode filling factor in the TEA section has been exploited to improve considerably the operating efficiency of a hybrid CO_2 laser commonly used for obtaining single mode emission. A more efficient and versatile method of obtaining single mode lasing would be by the elimination of spatial hole burning effect. The first ever experimental realization of this has been possible by placing an appropriate intra-cavity saturable absorber that ensured unidirectional oscillation in the ring cavity leading to lasing on a single mode with much wider detunability.

References

[1] D. C. Tyte, “Carbon dioxide laser” Ed D. W. Goodwin, Vol 1, Academic Press, New York (1970)
[2] “Lasers in Surface Engineering” Ed: N. B. Dahotre, Publ: ASM international, (1998)
[3] Los Alamos Science, Vol 3, No. 1, p-2 (1982)
[4] R. V. Ambartsaumian, Z. A. Grasiuk, A. P. Dyadkin, N. P. Furzikov, V. S. Letokhov, B. I. Vasilev, *Appl Phys* **15**, 27 (1978)
[5] R. Marchetti, E. Penco, G. Salvetti, *IEEE J. Quantum Electron* **19**, 1488 (1983)
[6] P. K. Bhadani, A. Sylvan, R. G. Harrison, *Rev Sci Instrum* **63**, 71 (1992)
[7] S. Marchetti, R. simili, *Opt Las Technol* **35**, 583 (2003)
[8] C. K. N. Patel, in: A. K. Levine (Ed), *Lasers – A series of Advances*, Vol 2, Mercel Dekker, New York, (1968)

[9] P. E. Dyer, B. L. Tait, *Appl Phys Lett* **41**, 506 (1982)
[10] P. E. Dyer, B. L. Tait, *J. Phys E* **16**, 467 (1983)
[11] M. Trtica, P. Vujkcvic Cvijin, I. Mendas, *Opt Quantum electron* **16**, 511 (1984)
[12] S. Marchetti, R. simili, M. Giorgi, *J. Phys (Paris)* **48**, C7-51 (1987)
[13] S. Marchetti, R. simili, *Opt Commun* **89**, 429 (1992)
[14] G. C. Stuart, H. Houtman, J. Meyer, *Rev Sci Instrum* **58**, 261 (1987)
[15] High power gas laser amplifier, *US Patent No. 4264868*
[16] D. J. Biswas, J. P. Nilaya, A. Kumar, *Opt Commun* **248**, 521 (2005)
[17] US Patent No.**6**,950,453
[18] D. J. Biswas, J. P. Nilaya, *Prog Quantum Electron* **26**, pp1-63 (2002)
[19] G. J. Ernst, *Opt Commun* **49**, 275 (1984)
[20] Aniruddha Kumar, R. B. Bhatt, D. J. Biswas, N. S. Banerjee, A. Mokhriwale, U. Nundy, *Meas Sci Technol* **12**, 1739 (2001)
[21] D. J. Biswas, J. P. Nilaya, U. K. Chatterjee, *Rev Sci Instrum* **69**, 4242 (1998)
[22] P. K. Bhadani, *Rev Sci Instrum* **60**, 605 (1989)
[23] J. V. Lebacqz, H. J. White, *Pulse generators*, ed: G. N. Glasoe and J. V. Lebacqz, Vol 5, p-275, McGraw Hill, New York, (1948)
[24] R. C. Sze, E. Seegmiller, *IEEE J Quantum Electron* **QE 17**, 81 (1981)
[25] S. Black and T. R. Borkes, Digest of Technical Papers, 2nd International Pulsed Power Conference, Lubbock, Texas, June 1979
[26] K. R. Rickwood, J. MsInnes, *Rev Sci Instrum* **53**, 1667 (1982)
[27] T. Kan, D. Ball, E. Schmitt, J. Hill, *Appl Phys Lett* **35**, 676 (1979)
[28] G. J. Scoles, B. P. Newton, *Proc of 11th modulator symposium, Sept 1973 (technical reprint* No. **78**, English Electric Valve Co., Chelmsford, U.K.)
[29] J. P. Nilaya, D. J. Biswas, B. S. Narayan, U. K. Chatterjee, *Rev Sci Instrum* **65**, 3590 (1994)
[30] J. P. Nilaya, Ph.D Thesis, University of Mumbai (2001)
[31] D. J. Biswas, J. P. Nilaya, U. K. Chatterjee, *Rev Sci Instrum* **66**, 4813 (1995)
[32] D. J. Biswas, J. P. Nilaya, U. K. Chatterjee, *Opt Eng* **36**, 588 (1997)
[33] D. J. Biswas, J. P. Nilaya, *Rev Sci Instrum* **72**, 2505 (2001)
[34] J. P. Nilaya and D. J. Biswas, *Proc of National Laser Symposium*, p-48, PRL, Ahemdabad, India (1998)
[35] icol reference
[36] H. Tanaka, H. Hatanaka, and M. Obara, *Rev. Sci. Instrum.* **61**, 2092 (1990)
[37] Sylvan, P. K. Bhadani, and R. G. Harrison, *Meas. Sci. Technol.* **3**, 200 (1992)
[38] K. Jayaram and A. J. Alcock, Appl. *Phys. Lett.* **46**, 636 (1985)
[39] V. P. Singal, R. Vijayan, B. S. Narayan, D. J. Biswas, U. Nundy, *Inf. Phys. Technol* **44**, 69 (2003)
[40] D. J. Biswas, J. P. Nilaya, M. B. Sai Prasad, and P. Raote, "Switch-less operation of a TEA CO_2 laser" *Opt Express* **13**, 9636 (2005)
[41] M. Kumar, P. Choudhary, K. Rama Rao, L. B. Rana, C. P. Paul and A. K. Nath, NLS 1997, p-95
[42] D. M. Tratt, A. K. Kar, and R. G. Harrison, *Prog Quant Electronics* **10**, pp229-266 (1985)
[43] W. T. Silfvast, *Laser Fundamentals*, Cambridge University Press, pp 382 (1999)
[44] P. Mathieu and J. R. Izatt, *IEEE J Quant Electron* **QE13**, 465 (1977)

[45] P. E. Dyer and D. J. James, *Appl Phys Lett* **26**, 331 (1975)
[46] G. Giuliani, E. Palange, and G. Salvetti, *Opt Lett* **9,** 393 (1984)
[47] P. E. Dyer and B. L. Tait, *J. Phys E: Sci Instrum* **16**, 467 (1983)
[48] N. R. Heckenberg and J. Mayer, *Opt Communications* **16**, 54 (1976)
[49] K. Kar, D. M. Tratt, and R. G. Harrison, *Opt Commun* **43**, 274 (1982)
[50] G. Scott and A. L. Smith, *Appl Phys B* **33**, 99 (1984)
[51] Aniruddha Kumar, J. P. Nilaya, D. J. Biswas *Rev Sci Instrum* **75**, 5203 (2005)
[52] P. E. Dyer and B. L. Tait, *Appl Phys Lett* **41**, 506 (1982)
[53] G. Kovar, M. Piche, and P. A. Belanger, *Optics Commun* **60**, 74 (1986)
[54] Gondhalekar, N. R. Heckenberg, and E. Holzhauer, *IEEE J Quant Electron* **11**, 103 (1975)
[55] D. J. Biswas, J. P. Nilaya, S. K. Sehegal and U. K. Chatterjee, *J. Appl. Phys*. **76**, 1340 (1994)
[56] A Kumar, J. P. Nilaya, and D. J. Biswas, *Optics Commun* **245**, 289 (2005)
[57] D. L. Lyon, E. V. George, Quarterly Progress Report, Research Laboratory of Electronics, MIT 15 April, P-55 (1971)
[58] Y. Liu and J. R. Rios Leite, *Opt Commun* **88**, 391 (1992)
[59] Y. Liu, L. C. Barbosa, and J. R. Rios Leite, *Phys Lett A* **203**, 189 (1995)
[60] Aniruddha Kumar, J. Padma Nilaya, and D. J. Biswas, *Optics and Laser Technol* **37**, 1 (2005)
[61] Aniruddha Kumar, J. P. Nilaya, and D. J. Biswas, *Infrared Physics and Technology* **44**, 121 (2003)

In: Lasers and Electro-Optics Research at the Cutting Edge ISBN: 1-60021-194-1
Editor: Steven B. Larkin, pp. 253-271

Chapter 8

NOVEL BISMUTH-ACTIVATED GLASSES WITH INFRARED LUMINESCENCE

Mingying Peng[1,2*] and Jianrong Qiu[1,3]

[1] Shanghai Institute of Optics & Fine Mechanics, Chinese Academy of Sciences, Shanghai 201800, China

[2] Graduate School of the Chinese Academy of Sciences, Beijing 100039, China

[3] Department of Materials Science and Engineering, Zhejiang University, Hangzhou 310027, China

Abstract

Broadband infrared luminescence covering the optical telecommunication wavelength region of O, E, S, C and L bands (1200~1600nm), with FWHM (Full Width at Half Maximum) of 200~400nm and fluorescent lifetime of 200~700μs, was observed from bismuth-doped SiO_2-Al_2O_3, GeO_2-Al_2O_3, GeO_2-B_2O_3, GeO_2-Ga_2O_3, GeO_2-Ta_2O_5, SiO_2-B_2O_3, SiO_2-Ta_2O_5, Al_2O_3-B_2O_3, Al_2O_3-P_2O_5 and ZnO-Al_2O_3-SiO_2 glass systems at room temperature in the case of 808nm excitation. In the absorption spectra of these glasses, a strong absorption peak near ~500nm was always detected in all samples, while absorption peaks at ~700, ~800 and ~1000nm were only in some of hosts such as GeO_2-M_2O_x glasses where M stood for B, Al, Ga and Ta. As usual, the emission and excitation peaks for Bi^{3+}-doped luminescent materials are located in ultraviolet or visible wavelength region, and the radiative lifetime seldom exceeds 10μs. Thus, the luminescent properties of these Bi-doped glasses are distinguishably different from the Bi^{3+} doped luminescent materials previously reported in literatures. Compared with the experimental results on BiO molecule in gas phase reported by Fink *et al* and Shestakov *et al*, we tentatively assign the broadband infrared luminescence to the BiO molecule dissolved in these bismuth doped vitreous matrices. The observed absorptions at ~500nm, 700nm, 800nm and 1000nm could then be attributed to the transitions of $X_1 \rightarrow$(H, I), $X_1 \rightarrow A_2$, $X_1 \rightarrow A_1$ and $X_1 \rightarrow X_2$ respectively, and the emission peak at 1300nm could be to $X_2 \rightarrow X_1$. These glasses may have potential applications in widely tunable laser and super-broadband optical amplifier for the optical communications.

Keywords*:* Infrared luminescence; glasses; bismuth

* E-mail address: m_ypeng@yahoo.com

1. Introduction

The popularization of computer and the internet techniques speeds up the development of information society that in turn requires the information transport with excellent flexibility and larger transmission capacity at much faster rates (In the year of 2010, 3000 channels will be needed at a rate of 5-100Tbps in the DWDM system according to a report from Japanese Photonic Network Research Committee.) [1]. At the present time, there are three different ways to increase the transmission capacity: 1) improving the data transmission rate of every communication channel; 2) minimizing the channel spacing to increase the channel number in the current limited bandwidth; 3) broadening the gain bandwidth of the amplifier and the laser source. Case 1 will increase the dispersion effect; and case 2 will increase the nonlinear effect and result in the possibility of the interchannel signal crosstalk, moreover the smaller channel space calls for the better wavelength stabilization of system component and further increases the building cost of the telecommunication system. Compared with the cases 1 and 2, case 3 is no doubt an attractive way to expand the number of transmission channels, to fully utilize the whole optical transmission window of the silica fiber and to achieve efficient wavelength division multiplexing transmission network. Considerable efforts have been devoted to the achievement of broadband amplification by choosing rare-earth (e.g. Pr^{3+}, Tm^{3+} or Er^{3+}) doped fiber amplifiers (REDFAs) [1, 2]. Unfortunately, the application bandwidths of REDFAs are hardly beyond 100nm because the f-f transitions of rare earth ions are confined in the inner-shell, which are insensitive to local environmental, even though the development of the wavelength division multiplexing (WDM) technology based on REDFAs promotes the progress of the optical communication during the past decades [1, 2]. Fiber Raman amplifiers (FRAs) can overcome the drawbacks of REDFAs and realize the broadband amplification. However, they require multi-wavelength pumping schemes, high-power consumption and more complicated structures [3]. In the current status of WDM system, multiple number of amplifiers including REDFA and FRA must be employed at the same time in order to fully utilize the wide-window resource of silica fiber in the range of 1.2~1.6μm, because there is still no efficient and broadband optical amplifier that can cover all the optical windows of silica glass fiber. If extra-broadband optical amplification could be realized with high gain efficiency by single-wavelength pumping scheme, a drastic evolution in the WDM technology might be anticipated. In order to develop superbroadband amplifiers, exploration of the superbroadband luminescent materials especially covering the 1.2~1.6μm telecommunication window becomes one of the key points. As a matter of fact, attempts to explore such luminescent materials have been made by selecting the transition-metal ion (e.g. Cr^{4+}, or Ni^{2+}, or V^{2+}) doped glasses and some promising results have also been achieved [4, 5, 6]. For example, Cr^{4+}-doped calcium aluminate glass was reported to exhibit a 1.3μm-peaked emission with FWHM larger than 200nm while Ni^{2+}-doped zinc aluminosilicate glass showed a broadband emission at ~1400nm with FWHM even wider than 300nm upon 976nm excitation at room temperature [4,5].

A new kind of luminescent materials activated by bismuth was recently reported to exhibit the superbroadband fluorescence with FWHM larger than 300nm and lifetime of several hundred microseconds, whose optical properties are distinguishably from the normal Bi^{3+}-doped optical materials [7-17]. The bismuth-doped optical glasses with near infrared broadband emissions are promising as gain mediums for broadband fiber amplifiers.

Unfortunately, it is still not very clear where the infrared emissions originate [8, 9]. This is impeding further design and development of new optical materials with optimizing properties for practical applications. Here, we first summarize the recent research development on bismuth-doped glasses, and then study the infrared luminescent mechanism and finally outlook the prospect of such new materials on the basis of our researches.

2. Experimental Section

All glass samples were synthesized using a conventional melting-quenching method in air by use of high-purity alumina or platinum crucibles. High-purity GeO_2 (>99.999%), Ga_2O_3 (>99.999%), Ta_2O_5 (>99.999%), and analytic-purity reagents H_3BO_3, SiO_2, TiO_2, Li_2CO_3, $NaNO_3$, $BaCO_3$, $NH_4H_2PO_4$, $SrCO_3$, Sb_2O_3, $Al(OH)_3$ and Bi_2O_3 were used as raw materials. SiO_2-, GeO_2-, P_2O_5- or B_2O_3-based glasses were prepared as described in Refs 8--17. The detailed glass compositions were found in following text. All the bismuth-doped glasses where the near infrared luminescence could be detected were bubble-free and their body colors changed with bismuth concentration.

The polycrystalline powder sample of $Sr_{0.95}B_4O_7$: 0.05Bi was prepared as follows: stoichiometric amounts of $SrCO_3$ and H_3BO_3 together with 3mol% excess of H_3BO_3 to compensate for B_2O_3 evaporation were mixed, ground, preheated at 700^{o}C in alumina crucible in air for 5h and then sintered at 850^{o}C for 5h twice after reground.

The optical absorption spectra of the samples were recorded using a JASCO V-570 spectrophotometer and the infrared fluorescence spectra were acquired using a ZOLIX SBP300 spectrofluorometer with an InGaAs detector upon the excitation of the 808nm InGaAs laser diode (LD) with maximum power of 135mW. The diffuse reflection spectra of the powder samples were obtained using a Hitachi U-4100 spectrophotometer. The emission and the excitation spectra in the ultraviolet and visible region were recorded using a JASCO FP-6500 fluorescence UV-Vis spectrofluorometer. The sample of $Sr_{0.95}B_4O_7$: 0.05Bi was checked by a Bruker D4 x-ray diffractometer at 40 kV/20 mA with CuKα1 (λ=1.5405 Å) as a radiation source and identified on the basis of JCPDS standard card 15-801 for SrB_4O_7. The fluorescence lifetime was measured by exciting the sample with the modulated 808nm LD. The signals detected by an InGaAs photodetector in the TRIAX550 spectrofluorometer were recorded using a storage digital oscilloscope (Tektronix TDS3052). All the measurements were carried out at room temperature.

3. Results and Discussions

3.1. Current Research Status of Bi-doped Glasses with near Infrared Luminescence

3.1.1. Bi-doped SiO_2-Al_2O_3 Glasses

In 1999, Fujimoto and Nakatsuka in Osaka University firstly observed an abnormal near infrared luminescence (NIRL) from $92SiO_2{\cdot}7.0Al_2O_3{\cdot}1.0Bi_2O_3$ (SAB) glass with broadband feature at room temperature, which is totally different from those of trivalent and divalent

bismuth ion doped materials previously reported in literature [7]. Figure 1 lists the absorption spectrum of SAB glass. As can be seen from Figure1, three absorption peaks appear at 500, 700 and 800nm respectively as well as a strong band edge absorption at ~300nm. As excited by 800nm, a strongest emission was found at 1250nm with a FWHM of ~300nm. As the excitation wavelength changing into 500nm, two emissions were detected at 750 and 1140nm respectively and the latter was much stronger than the former. With the 500nm excitation scheme, the lifetime of the emission at 1140nm was measured to be about 630μs at room temperature [7].

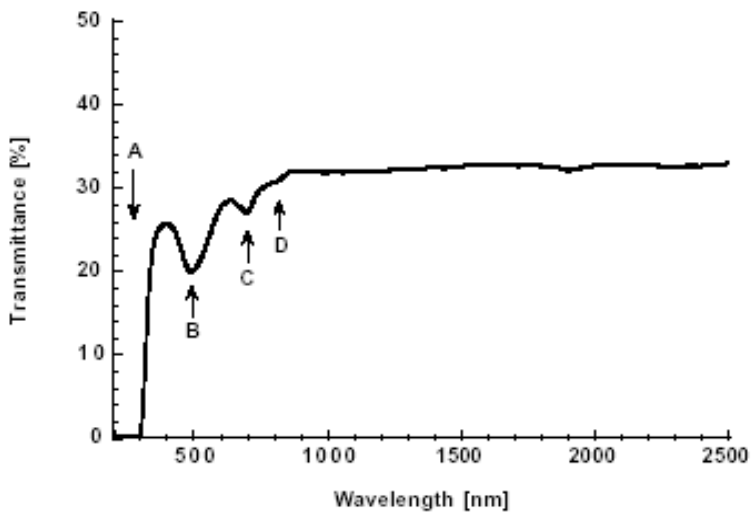


Figure 1. Transmittance spectrum of SAB glass [7].

3.1.2. Bi-doped GeO_2-Al_2O_3 Glasses

From Figure 1, we can clearly see that the transmittance of the SAB glass is only about 30% at maximum in the region of 1000~2500nm because of many bubbles included in the glass, even though the glass sample was prepared at a high temperature up to 1760 ^oC [7]. The existence of bubbles highly affects its practical applications. Comparing with SiO_2-based glass, GeO_2-based glass can be more easily prepared by conventional melting technique because its melting temperature is evidently lower (Melting temperatures are 1720 and 1115^oC for pure SiO_2 and GeO_2 glasses, respectively. [18]). In consideration of the closest structural analogy between glassy GeO_2 and SiO_2 [18], we would expect similar broadband infrared luminescence in bismuth doped GeO_2–Al_2O_3 glasses.

Figure2a shows the transmittance spectrum of $96GeO_2{\cdot}3Al_2O_3{\cdot}1Bi_2O_3$ (GAB) glass sample, consisting of five peaks at below 370nm, 500nm, 700nm, 800nm and 1000nm [8]. The former four peaks are similar to, but the last one is absent in SAB glass [7]. Since the band-gap absorption of glassy GeO_2 is at 5.63eV (221nm) and the absorption edge of the Bi_2O_3-based glass is at about 475nm, the band at below 370nm is therefore assigned to the contributions from them [19-21].

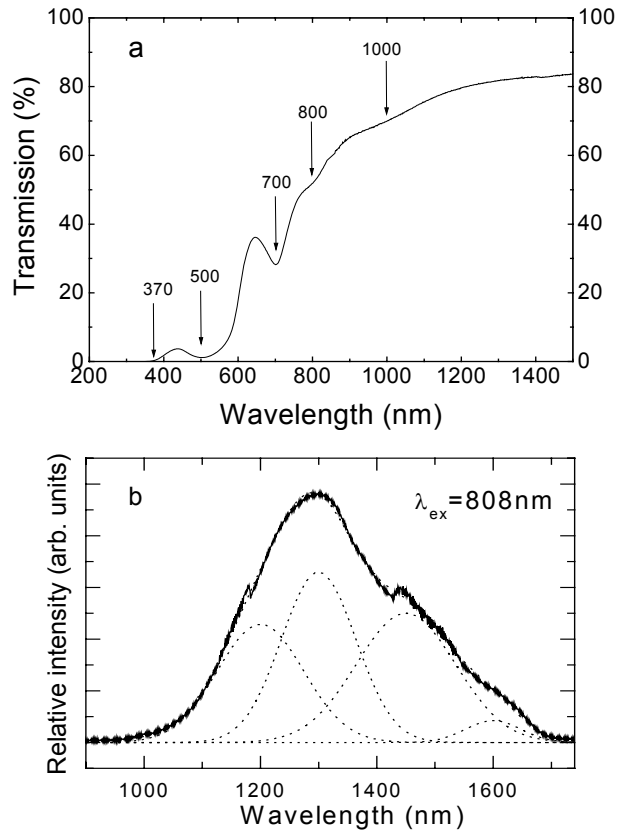


Figure 2. Absorption and emission spectra of GAB glasses at room temperature.

Figure 2b shows the fluorescence spectrum of GAB glass sample when pumped by a GaAlAs semiconductor laser diode with wavelength of 808nm [8]. A broad band with a maximum at about 1300nm appears with a FWHM of 320nm. In view of the asymmetric feature of the spectrum, it could be decomposed into four gaussian peaks, viz. at 1200nm, 1300nm, 1450nm and 1600nm with FWHM of 148nm, 128nm, 170nm and 94nm respectively. The decay curve of the ~1300nm emission was collected by monitoring the emission at 1300nm at room temperature in the case of the excitation of commercial 808nm LD, which demonstrated the lifetime of the ~1300nm emission was about 254μs [8].

Luminescent properties of Bi-doped GeO_2-Al_2O_3 glasses are greatly dependent on glass compositions [13]. Figure 3 shows the effects of bismuth concentration on the spectroscopic properties. As we have observed during the experimental process, the sample color turns gradually from nearly colorless, through pink to deep reddish-brown with the increment of bismuth concentration. In concert with the body color changing, their absorption and emission spectra correspondingly exhibit the strong dependence on bismuth concentration. Figure 3A and Figure 3B show the absorption and fluorescent spectra of (97-x)$GeO_2 \cdot xBi_2O_3 \cdot 3Al_2O_3$ (x=0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 2.0) glasses, respectively. As can be seen from Figure 3A, there are four absorption peaks at about 500, 700, 800 and 1000nm respectively just as shown in Figure 2a. With the bismuth concentration increasing, the peak of the ~500nm band shifts towards shorter wavelength from about 560nm gradually to 500nm while the peak of the

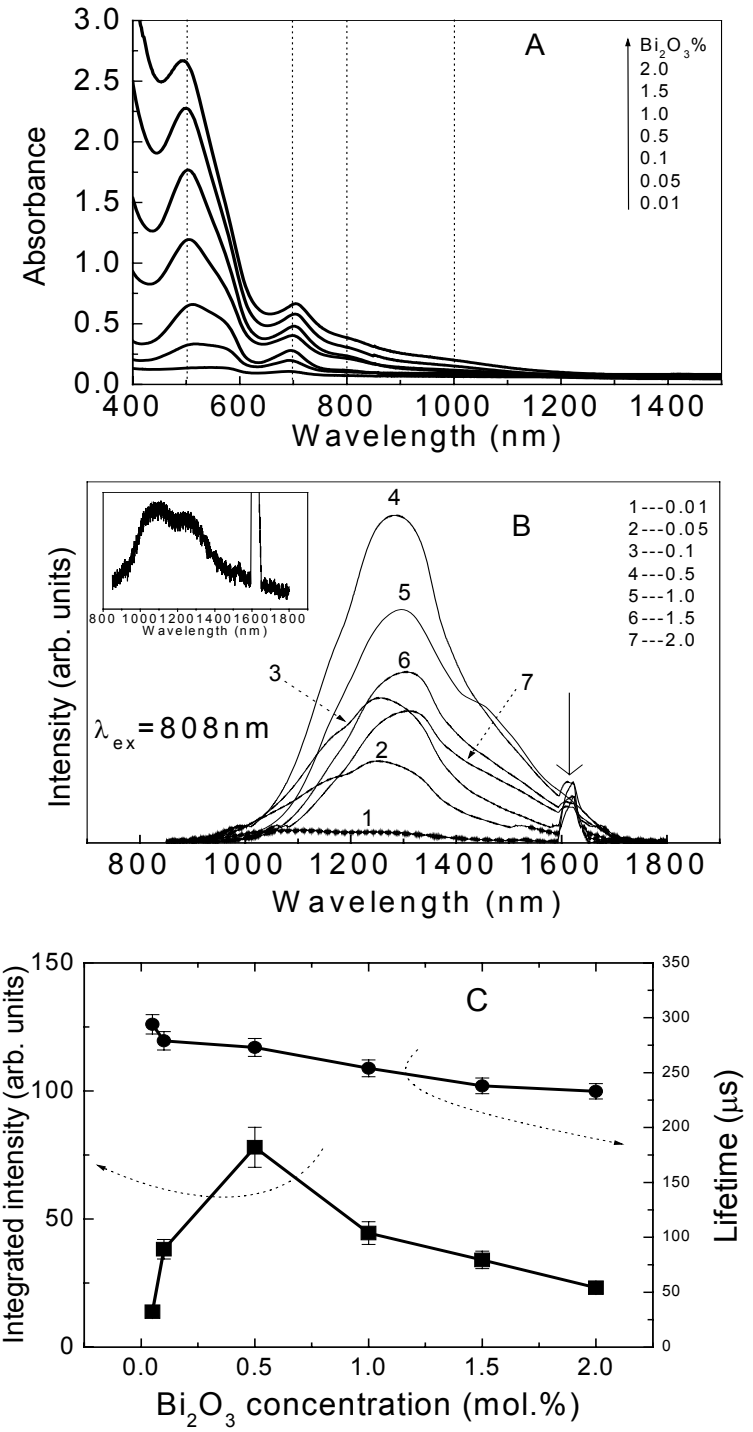


Figure 3. (A) Absorption spectra, (B) emission spectra and (C) integrated emission intensity and lifetime of $(97\text{-}x)GeO_2 \cdot 3Al_2O_3 \cdot xBi_2O_3$ (x=0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 2.0) glasses. The peaks indicated by arrow are due to the second order diffraction of 808nm-laser. Lines are drawn as guides for the eye [13].

~700nm band shifts towards the reverse direction [13]. And the latter two bands remain almost un-shift. At the same time, as shown in Figure 3B, the strongest emission peak moves towards longer wavelength, viz. 1100nm for x=0.01 → 1254nm for x=0.05 → 1260nm for x=0.1 → 1284nm for x=0.5 → 1300nm for x=1.0 → 1305nm for x=1.5 → 1310nm for x=2.0.

Furthermore, with increasing x value, the fluorescent intensity monotonically increases until x=0.5, and then decreases, showing that the critical dopant concentration is about 0.5 mol.%. As the Bi_2O_3 content increases in the glass host, the interaction between bismuth ions also increases which in turn leads to the decrease of the lifetime as shown in Figure 3C. All the previous works have clearly demonstrated that as for the single doping of Al or Bi ions in the GeO_2 and SiO_2 glasses, no any infrared luminescence in the region of 900nm-1700nm could be observed in our experiments [7-9]. Therefore, codoping of aluminum into bismuth-doped glasses is indispensable for the occurrence of the broadband infrared luminescence [7-9].

However, it is still unknown up to now what role the aluminum ion plays in the generation of infrared luminescence. For the sake of study on the possible role of aluminum in infrared luminescence of GeO_2: Bi, Al glasses, we systematically investigated the effects of aluminum concentration on the luminescent properties [13]. Figure 4 shows the integrated fluorescent intensity and lifetime as function of Al_2O_3 concentration. As y value increasing, the fluorescent intensity continuously increases until y=5.0, and then decreases. Meanwhile, the emission peak shifts from 1320nm (y=1.0), 1300nm (y=3.0), 1300nm (y=5.0) to 1286nm (y=8.0). Obviously, the influence of Al_2O_3 addition on the fluorescent peak position seems slighter than that of the Bi_2O_3 addition.

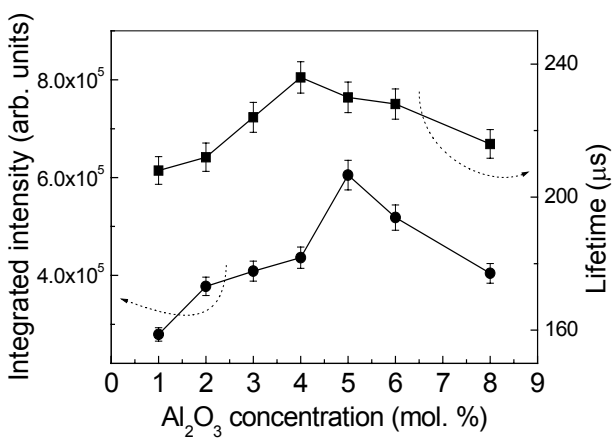


Figure 4. The integrated fluorescent intensity (●) and lifetime (■) of (99-x) $GeO_2 \cdot xAl_2O_3 \cdot 1Bi_2O_3$ (x=1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0) [13]. Lines are drawn as guides for the eye.

The coordination number of aluminum ion increases with increasing Al_2O_3 content in Al_2O_3-SiO_2 glass [22]. As a small amount of Al_2O_3 is introduced into the Al_2O_3-SiO_2 glass, Al^{3+} ions will be mainly in the fourfold coordination [22-27]. In view of the much more structural resemblance between glassy SiO_2 and GeO_2, the $[AlO_{4/2}]^-$ tetrahedrons will be more favorable in GeO_2-based glasses because the ion radius of Ge^{4+} (R_{Ge}^{4+}(CN=4) = 0.39Å where CN is the coordination number) matches that of Al^{3+} (R_{Al}^{3+} (CN=4) = 0.39Å) better than that of Si^{4+} (R_{Si}^{4+} (CN=4) = 0.26Å) [28]. When the Al_2O_3 concentration is as lower as

y<4.0, the formed $[AlO_{4/2}]^-$ tetrahedrons coupled with $[GeO_4]$ tetrahedrons will surround the bismuth-related infrared-emission centers and enlarge the center-to-center spacing in the bonding association of Al-O-Bi, which is similar to the dispersing effect of the Nd^{3+} ions by Al^{3+} ions in Nd^{3+} and Al^{3+} codoped silica glasses [22-25]. Consequently, the interaction between these emission centers will be weakened and the energy loss due to non-radiation transitions will be reduced. Therefore the fluorescent intensity and lifetime increases as indicated in Figure 4. However, when the Al_2O_3 concentration becomes as higher as 5.0<y<8.0, the five- or six-coordinated Al^{3+} ions, coexisting with four-coordinated ones, appear and increase [22, 25]. The five- or six-coordinated Al^{3+} ions play as the glass network modifiers and couldn't efficiently disperse the infrared-emission centers. Thus, the interaction between the infrared-emission centers becomes stronger, and the fluorescent intensity and lifetime decrease as shown in Figure 4.

In the preparation process of GeO_2: Bi, Al glass, we observed that the glass molten becomes more easily poured out of the alumina crucibles with increasing the Al_2O_3 concentration, which might be attributed to the increasing non-bridging oxygen ions with increasing network modifiers of five- or six-coordinated Al^{3+} ions. When the aluminum content is up to 10mol.%, no transparent glass could be obtained.

3.1.3. Bi-doped P_2O_5-Al_2O_3 and B_2O_3-Al_2O_3 Glasses

After the observation of near infrared emission from Bi and Al co-doped SiO_2 and GeO_2 glasses [7, 8, 13], we wonder whether the analogous results could be obtained in other Bi and Al co-doped glass systems, such as phosphate and borate glasses. It is really the case. We successfully observed the broadband near infrared emissions at ~1300 and 1250nm in the range of 1000~1600nm from bismuth doped P_2O_5-Al_2O_3 and B_2O_3-Al_2O_3 glasses respectively [16-17]. The fluorescence lifetime for the former glass is about 500μs while it is 370μs for the latter. Unlike SAB and GAB glasses, the glasses have only two absorptions at ~460nm and ~700nm, respectively [16-17]. Figure 5 shows the absorption and emission spectra of bismuth-doped P_2O_5-Al_2O_3 glasses.

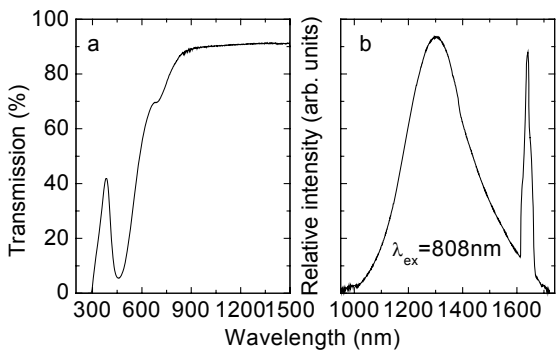


Figure 5. Absorption (a) and emission spectra of Bi-doped P_2O_5-Al_2O_3 glasses at room temperature.

3.1.4. Bi-doped MO_2-N_2O_X (M=Ge, Si; N=Ga, B, Ta) Glasses

The previous investigations showed that codoping of aluminum in bismuth-doped glasses is essential for the broadband infrared luminescence [7, 8, 13]. Since codoping of Bi with Al

into GeO_2 glass can lead to the infrared luminescence, we suspect that the analogous results can be obtained when codoping of Bi with other foreign cation, e.g. Ga or B, which are located in the same IIIA group and usually show the similarity in chemical and physical properties to Al [9, 12, 15].

Figure 6 shows the absorption spectra of $96GeO_2{\cdot}3Ga_2O_3{\cdot}1Bi_2O_3$ (simplified as GGB) and $96GeO_2{\cdot}3B_2O_3{\cdot}1Bi_2O_3$ (GBB) glass samples. Four absorption peaks can be observed at 500, 710, 800 and 1000nm for GGB, and 530, 705, 800 and 1000nm for GBB respectively [9]. These peaks are similar to those observed in GAB glass with peaks at 500, 700, 800 and 1000nm [8]. Figure 7a and Figure 7b show the fluorescence spectra of GGB and GBB glass samples, pumped by a GaAlAs semiconductor laser diode with wavelength of 808nm. In the fluorescent spectrum of GGB, there is a broad band with FWHM of 345nm peaked at about 1325nm [9]. And in the fluorescent spectrum of GBB, there is also a broad band with FWHM of 355nm peaked at about 1315nm. The measured lifetimes are 230μs for the 1325nm emission of GGB and 237μs for the 1315nm emission of GBB, respectively [9].

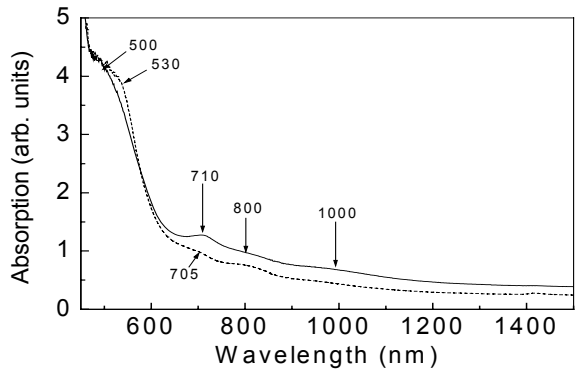


Figure 6. Absorption spectra of $96GeO_2{\cdot}3Ga_2O_3{\cdot}1Bi_2O_3$ (GGB) (solid line) and $96GeO_2{\cdot}3B_2O_3{\cdot}1Bi_2O_3$ (GBB) (dashed line) glasses.

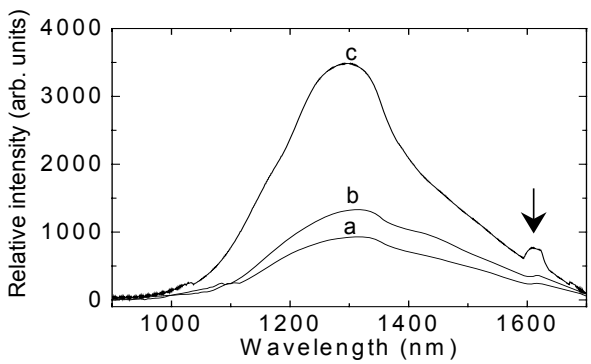


Figure 7. Fluorescence spectra of $96GeO_2{\cdot}3M_2O_3{\cdot}1Bi_2O_3$ (M = B (a), Ga (b), Al (c)) glasses when pumped by 808nm LD. The peaks indicated by arrow are due to the second order diffraction of 808nm-laser.

From Figure 7, we can also find out that the luminescence intensity of $96GeO_2{\cdot}3M_2O_3{\cdot}1.0Bi_2O_3$ (M = B, or Al, or Ga) increases along with the decrement of ion

radius difference between M^{3+} and Ge^{4+} (R_{Ge}^{4+}(CN = 4) = 0.39Å where CN is coordination number; R_B^{3+}(CN = 4) = 0.11Å; R_{Ga}^{3+} (CN = 4) = 0.47Å; R_{Al}^{3+} (CN = 4) = 0.39Å.) [28]. This indicates that the smaller ion radius difference between Ge^{4+} and the foreign cation codoped with Bi ion will benefit the infrared luminescence of bismuth in the germanium oxide glasses. The small ion radius difference between the Ge^{4+} and the M^{3+} leads to the small structural distortion of the glass network, and results in the good distribution and stability of Bi emission centers, which contributes to the infrared luminescence.

As codoping tantalum into Bi-doped GeO_2 or SiO_2 glass, similar absorptions and emissions to GAB and SAB glasses have been obtained [12]. Here, we just take Bi-and-Ta-codoped GeO_2 glass for example. The transmission spectra of $96GeO_2{\cdot}3Ta_2O_5{\cdot}1Bi_2O_3$ (GTB) and $96GeO_2{\cdot}3.0\ Al_2O_3{\cdot}1.0Bi_2O_3$ (GAB) glass samples are shown in Figure 8A. Four absorption peaks at 508nm, 712nm, 800nm and 1000nm can be observed in the absorption spectrum of GTB glass, which is similar to those in GAB glass (i.e. 500nm, 700nm, 800nm and 1000nm) and those in SAB glass (i.e. 500nm, 700nm and 800nm) [7, 8, 12]. Figure 8B shows the fluorescence spectrum of GTB glass under the 808nm LD excitation. In the fluorescence spectrum of GTB, there is a broad band with FWHM of 402nm peaked at about 1310nm, which is even larger than those of SAB and GAB glasses [7, 8, 12]. In addition, the ~1310nm emission is located at the longer wavelength as compared with the 1250nm one observed in SAB glass, also reflecting the influence of principle glass composition on the luminescent properties of Bi-doped glasses [7, 12].

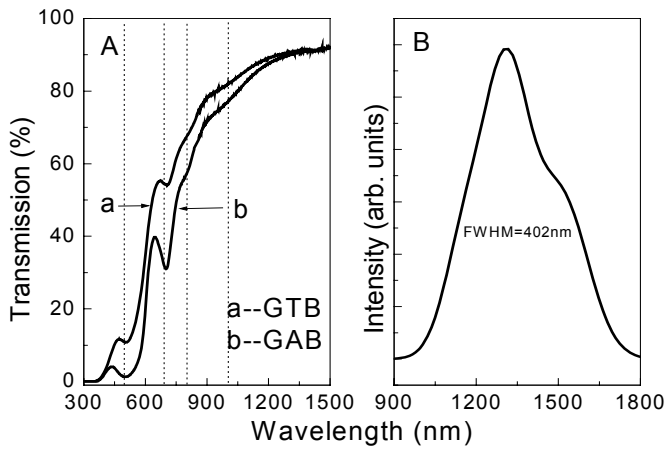


Figure 8. (A) Transmission spectra of (a) GTB and (b) GAB glasses; (B) Fluorescence spectrum of GTB glass when pumped by 808nm LD [12].

In boron and bismuth co-doped SiO_2 glass, the analogous broadband near infrared emission was found at 1340nm. But its notable absorptions are at about 500 and 800nm, which are slightly different from GAB, SAB and GTB glasses [15].

On the basis of the works presented above, it can be concluded that codoping of Ga, B and Ta instead of Al into Bi-doped GeO_2 and SiO_2 glasses can give the birth of the near infrared luminescence.

3.1.5. Whether Does the near Infrared Luminescence come from Cr^{4+} or from Bismuth in Aluminosilicate Glass Codoped with Cr_2O_3 and Bi_2O_3?

Recently, Batchelor *et al* reported the enhancement of Cr^{4+} emission in Li_2O-Al_2O_3-ZnO-SiO_2 (LAZS) glasses through the Bi_2O_3 addition at room temperature [29]. And they ascribed the broad emissions in 1000~1600nm region to the Cr^{4+} ions with the excitations of 514nm or 800nm [29]. But when studying their reported data, it can be easily found that no absorptions from Cr^{4+} but only absorptions from Cr^{3+} are observed in LAZS:0.01mol.%Cr_2O_3 glass while an absorption at ~500nm and an emission peaked at ~1300nm appear in the spectral range of 1000nm-1600nm in LAZS: 0.01.%Cr_2O_3, 1mol.%Bi_2O_3 glass [29]. Comparing these results with our works on bismuth-doped glasses, we don't believe the broadband near infrared luminescence is from tetravalent Cr^{4+} ions [8-17, 29]. Thus we prepared the glass samples with the same compositions as reported by Batchelor *et al* and carefully examined their luminescent properties. Our experimental results clearly indicate that the broad emissions should be assigned to bismuth rather than Cr^{4+} ion.

There are three broad bands at about 350, 440 and 650nm in the spectrum of LAZS: 0.01 mol.% Cr_2O_3 (LAZSC) glass (see Figure 9 (b)). Since the blank glass sample shows no absorptions in these regions (see Figure 9 (a)), these absorption bands should be assigned to chromium ions. The former one is due to Cr^{6+} ions and the latter two are due to Cr^{3+} ions [30, 31]. In the absorption spectrum of LAZS: 1 mol.% Bi_2O_3 (LAZSB) glass shown as Figure 9 (c), there are two notable peaks at around 485 and 712nm attributed to bismuth, which are similar to those of bismuth doped SiO_2 and GeO_2 glasses (i.e. at 500 and 700nm) [7-8, 10-11] and to those of bismuth doped borate and phosphate glasses (i.e. at ~465 and 700nm) [16-17]. While adding 0.01 mol.% Cr_2O_3 into the LAZSB glass, the 485nm peak keeps almost un-shifted whereas the longer wavelength peak of 712nm becomes flattened due to the absorption overlap between bismuth and chromium ions as shown in Figure 9 (d). At the excitation of 808nm LD, the blank sample and the LAZSC glass don't give out any detectable fluorescence while LAZS: 0.01 mol.% Cr_2O_3, 1 mol.% Bi_2O_3 (LAZSCB) and LAZSB glasses exhibit broad emissions peaked at around 1300nm with a 1400nm shoulder in 1000~1600nm region at room temperature (see Figure 10). That is sufficient to indicate that the 1300nm emission is not from Cr ions but from the luminescent centers correlated to the Bi_2O_3 addition, which is also the reason why Batchelor *et al* couldn't observe the Cr^{4+} absorptions in LAZSCB glass [29]. Interestingly, compared with the 1250nm emission observed in bismuth-doped silica glass at the excitation of 800nm [7], the strongest emission peak shifts to longer wavelength (i.e. 1300nm) in LAZSB glass, which demonstrates that the luminescent properties of Bi-doped glasses depend on the glass composition [10-11].

As the excitation source is changed into the 532nm laser, the strongest peak moves to the shorter wavelength of around 1200nm, implying the dependence of the infrared fluorescence on the excitation sources (see Figure 10 (f)). Similar phenomena have also been observed in the bismuth doped silica glass [7]. Batchelor *et al* assigned the emissions at ~750 or ~1150nm to the Cr^{3+} or Cr^{4+} ions at high crystal-field sites in LAZSCB glass respectively in the case of 514nm excitation [29]. But in bismuth-doped silica glass without Cr_2O_3 addition, the emissions at 750 and 1140nm were also observed at the 500nm excitation [7]. Even though the emission spectrum at the excitation of 514nm cannot be obtained due to the apparatus limits, the comparison between Batchelor's and Fujimoto's results and the observation of the ~1200nm emission at the 532nm excitation in LAZSB glass support at least that the ~1150nm emission corresponding to the 514nm excitation in Ref. 29 correlates to the Bi_2O_3 addition.

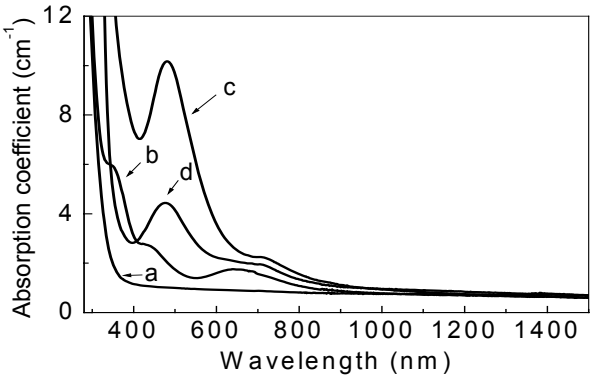


Figure 9. Absorption spectra of (a) LAZS, (b) LAZS: 0.01 mol.% Cr_2O_3, (c) LAZS: 1 mol.% Bi_2O_3 and (d) LAZS: 0.01 mol.% Cr_2O_3, 1 mol.% Bi_2O_3 glasses.

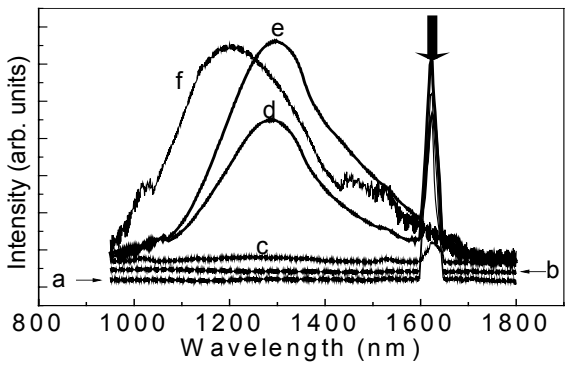


Figure 10. Fluorescence spectra of (a) LAZS, (b) LAZS: 0.01 mol.% Cr_2O_3, (c) LAZS: 0.04 mol.% Cr_2O_3, 1 mol.% Bi_2O_3, (d) LAZS: 0.01 mol.% Cr_2O_3, 1 mol.% Bi_2O_3 and (e) LAZS: 1 mol.% Bi_2O_3 glasses when pumped by 808nm laser; (f) fluorescent spectrum of LAZS: 1 mol.% Bi_2O_3 when pumped by 532nm laser. The peaks indicated by arrow are due to the second order diffraction of 808nm-laser.

From Figure 10 (c), (d) and (e), another phenomenon can be easily noticed increasing Cr_2O_3 concentration can quench the 1300nm emission, which leads to the reduction of fluorescence lifetime from 583μs for LAZSB to 530μs for LAZSCB [10-11]. That might be interpreted as follows. First, increasing Cr_2O_3 concentration doesn't favor the forming of bismuth centers responsible for the infrared luminescence, which can be verified by the decrement of absorption coefficient in near infrared region as Cr_2O_3% increases from 0% for LAZSB glass to 0.01% for LAZSCB glass (see Figure 10 (c) and (d)). Second, upon the 808nm-LD excitation, part-pumping light is absorbed by Cr^{3+} ion and consumed through non-radiation processes, since LAZSC glass cannot give out any light upon this excitation scheme.

3.2. Studies on Infrared Luminescence Mechanism

In the past years, luminescent properties of Bi^{3+} doped luminescent materials including crystals and glasses have been extensively studied [32-36]. The ground state of the Bi^{3+} ion is $^1S_0$, whereas the 6s6p excited states give rise to $^3P_0$, $^3P_1$, $^3P_2$ and $^1P_1$ in order of increasing

energy. Because the $^1S_0 \rightarrow {}^3P_0$ and $^1S_0 \rightarrow {}^3P_2$ transitions are strongly forbidden, the Bi^{3+} transitions between the $^1S_0$ ground state and the $^3P_1$ or $^1P_1$ excited state are usually observed. The excitation peaks of Bi^{3+} in both crystals and glasses appear in the ultraviolet region, while the emission peaks of Bi^{3+} occurs in the ultraviolet or visible wavelength region depending on host materials [32-36]. But no infrared luminescence from Bi^{3+} and no fluorescent lifetime of Bi^{3+} longer than 5μs in crystals or glasses have been reported up to now [32, 36]. Contrary to Bi^{3+}-doped luminescent materials, Bi^{2+}-doped ones were paid little attentions [37, 38]. To the best of our knowledge, only the luminescent properties of Bi^{2+} in crystals were reported. The Bi^{2+} transition between the ground state of $^2P_{1/2}$ and the first excited state of $^2P_{3/2}$ is parity forbidden, but becomes allowed when $^2P_{1/2}$ and the excited states of $^2S_{1/2}$ and $^2P_{3/2}$ are mixed by the uneven crystal-field terms, whereas the transition between $^2P_{1/2}$ and $^2S_{1/2}$ is strongly allowed [37, 38]. Therefore, three excitation peaks of Bi^{2+} in SrB_4O_7 corresponding to the transitions from $^2P_{1/2}$ to $^2P_{3/2}$ split into two by crystal fields and to $^2S_{1/2}$ were observed at 575, 470 and <312nm, and one emission peak of $^2P_{3/2} \rightarrow {}^2P_{1/2}$ was at 586nm [37]. And the fluorescent lifetime of SrB_4O_7: Bi^{2+} was 10.6μs at 300K [37].

Compared with the information on the Bi^{3+} or Bi^{2+} doped materials mentioned above, the peak-positions of the absorptions and emissions in GAB, SAB, GGB and GBB glasses are greatly different from those in the Bi^{3+} or Bi^{2+} doped materials [7-17, 37-38]. Four absorption peaks of GAB glasses cover the large wavelength region from visible to infrared, and their emission peaks lie in the infrared region. Furthermore the fluorescent lifetimes of GAB glasses are more than 200μs, which are one order or two order longer than those of the aforementioned Bi^{2+} or Bi^{3+} doped materials.

Fujimoto et al ascribed the absorption and the emission spectra of SAB to the Bi^{5+} transitions between the ground state of $^1S_0$ and the excited states of $^3D_{3,\ 2,\ 1}$ and $^1D_2$ [7]. However, at higher temperature, Bi_2O_3 will readily dissociate into the suboxide BiO [18]. We suspect that the infrared luminescence with the lifetime longer than 600μs might originate from the lower-valence bismuth but not from Bi^{5+} ion. In order to identify where the infrared luminescence came from, we prepared the $30Na_2O{\cdot}10Al_2O_3{\cdot}59SiO_2{\cdot}1.0Bi_2O_3$ (NASB), $63CaO{\cdot}36.5Al_2O_3{\cdot}0.5Bi_2O_3$ (CAB) glasses and the light-pink polycrystalline sample of $Sr_{0.95}B_4O_7$: 0.05Bi.

As well known, high-valent Bi^{5+} ion often exists in the compounds, e.g. $NaBiO_3$ or $KBiO_3$, containing alkali oxides with higher basicity [39-40]. The optical basicity calculated from the empirical formula proposed by J.A. Duffy [41-42] is 0.612 for the NASB glass, 0.745 for CAB glass and 0.491 for SAB glass. According to the optical basicity theory [42], the upper oxidation state of dopant is usually favorable in glass with higher basicity. Thus, the higher valent Bi^{5+} ions should be more preferred in NASB and CAB glasses than in SAB. If the infrared luminescence comes from Bi^{5+} ions, the infrared emission should also be observed from NASB and CAB glasses as well as from SAB glass. However, no broad infrared emission but only the typical absorption and emission due to Bi^{3+} ions could be detected in the NASB glass (see Figure 11) and CAB glass with higher basicity. So the infrared luminescence is not from Bi^{5+} and Bi^{3+} ions.

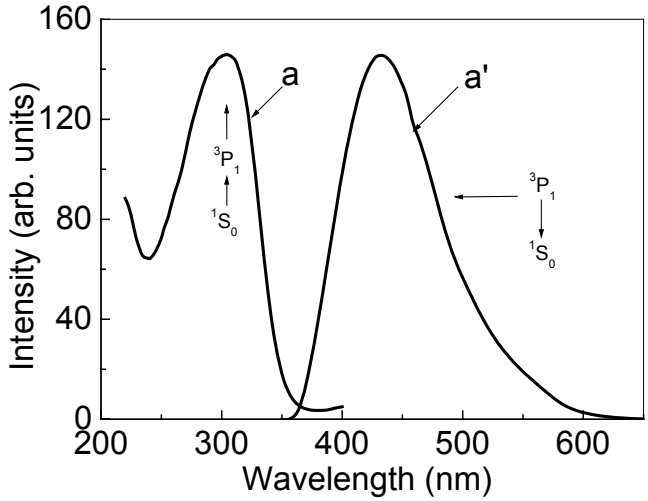


Figure 11. Excitation and emission spectra of NASB glass (a: λ_{em}= 433nm; a': λ_{ex}= 304nm).

The compound of SrB_4O_7 possesses the strong reducing ability and the trivalent RE^{3+} ions (RE=Eu, Sm, Yb, Tm, Nd) built into this matrix can be easily reduced to RE^{2+} even though SrB_4O_7: RE is prepared in an oxidizing atmosphere of air (The trivalent Sm, Tm or Nd ions are usually difficult to be reduced to their divalent form in most of other host compounds even prepared in N_2+H_2.) [43-45]. So it is possible that the Bi^{3+} ions doped into SrB_4O_7 are partially reduced to the lower-valence forms such as divalent ions. Blasse *et al* reported the co-existence of Bi^{3+} and Bi^{2+} in SrB_4O_7, but they didn't study the luminescence in the near infrared region [37]. By comparing the excitations of SrB_4O_7: Bi^{2+} with the absorption spectra of SAB, GAB, GGB and GBB glasses, one can easily find that the 575nm and 470nm excitation peaks for SrB_4O_7: Bi^{2+} fall exactly in the region where the strongest absorption bands of Bi-doped glasses appear. Is it possible the Bi-doped strontium tetraborate compound can emit the near infrared light? Thus, we prepared an $Sr_{0.95}B_4O_7$: 0.05Bi crystal and examined its luminescence in both visible and near infrared region. Three excitation peaks at 240nm, 480nm and 545nm were observed in the $Sr_{0.95}B_4O_7$: 0.05Bi. This is consistent with the observation by Blasse et al. [37], although their positions are slightly different. When sample was excited by a 240nm, 480nm or 545nm beam, a 587nm emission can always be observed. In addition to the Bi^{2+} and Bi^{3+} emissions as reported by Blasse et al [37], a weak near infrared broad emission at 1265nm was also detected in the $Sr_{0.95}B_4O_7$: 0.05Bi in the case of 808nm-excitation at room temperature (Figure 12). If the energy-level scheme proposed by Blasse et al is correct [37], it is difficult to justify the assignment of the observed broad 1265nm emission to Bi^{2+}.

Recently, Fink *et al* and Shestakov *et al* observed $X_2\ ^2\Pi_{3/2} \rightarrow X_1\ ^2\Pi_{1/2}$ emissions in the range of 1.0~1.6μm with a radiative lifetime of 480±100μs in the BiO molecule in gas phase [46-47]. These emissions are similar to our observation in the GAB (~1300nm and 254μs) [8], $96GeO_2{\cdot}3Ga_2O_3{\cdot}1Bi_2O_3$ (~1325nm and 230μs) [9], $96GeO_2{\cdot}3B_2O_3{\cdot}1Bi_2O_3$ (~1315nm and 237μs) [9], GTB (~1310nm and 222μs) [12] and SAB (~1140nm and 630μs) [7] glasses.

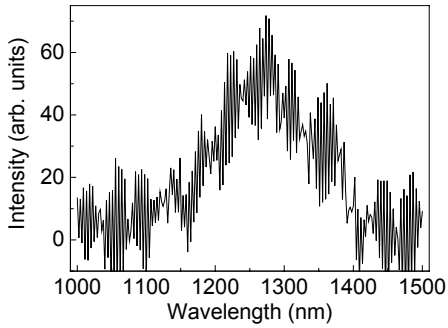


Figure 12. Near infrared emission spectrum of $Sr_{0.95}B_4O_7$: 0.05Bi when pumped by 808nm.

If the above near infrared luminescence originated from the BiO molecules dissolved in these glasses, the observed absorptions at ~500, 700, 800 and 1000nm could be assigned to the $X_1 \rightarrow$(H, I), $X_1 \rightarrow A_2$, $X_1 \rightarrow A_1$ and $X_1 \rightarrow X_2$ respectively, and the emission peak at 1300nm could be due to the $X_2 \rightarrow X_1$ transition [46]. In addition, BiO in gas phase also has an emission at around 750nm due to the $A_2 \rightarrow X_1$ transition, whose lifetime is about 9.3±1.5μs [46]. Correspondingly, a 750nm emission with the same order of lifetime (3.62μs) has also been observed in the Bi-doped SiO_2 glass [7, 46]. Therefore, BiO molecule would be a better candidate than the Bi^{2+} ion to be used to explain the emissions in the $SrBO_7$: Bi. If so, the emissions at 1265nm and 586nm can be attributed to the transitions of $X_2 \rightarrow X_1$ and $H \rightarrow X_1$ and the excitation peaks at 480nm and 545nm can be to $X_1 \rightarrow$(I, H) and $X_1 \rightarrow A_2$ [46]. The excitation peak at 240nm cannot be assigned at the present time since the electronic states of BiO higher than 30000cm^{-1} have not been reported. Moreover, the lifetime of the 586nm emission is 10.6μs at 300K [37], which coincides with the lifetime of the excited H state of BiO (τ_H=15±3μs) [46].

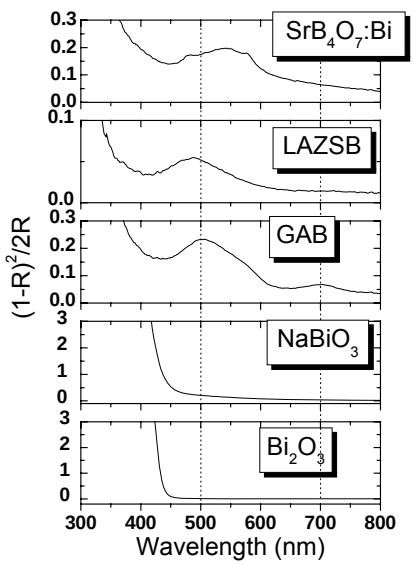


Figure 13. Diffuse reflection spectra of $NaBiO_3$ (Bi^{5+}), Bi_2O_3 (Bi^{3+}), GAB glass, LAZSB glass and SrB_4O_7: Bi (Bi^{3+}+BiO).

Diffusion reflection measurements in several valent bismuth compounds and glasses further confirm the above assignment. Figure 13 shows the diffuse reflection spectra of $NaBiO_3$ (Bi^{5+}), Bi_2O_3 (Bi^{3+}), SrB_4O_7: Bi (Bi^{3+} and BiO), LAZSB and GAB glass. In the spectra of $NaBiO_3$ and Bi_2O_3, there are no absorptions in the range of 450~800nm, while the similar broad absorptions are clearly seen in the same range in the SrB_4O_7: Bi, LAZSB and GAB. This suggests that Bi^{5+} and Bi^{3+} ions are not responsible for the absorptions and emissions as observed in the LAZSB and GAB glasses. Therefore, we believe that the abnormal infrared emission in bismuth glasses originates from Bi^{2+}, and probably in the form of BiO.

3.3. Future Development of Bismuth-Doped Optical Materials

3.3.1. Figure-of-Merits of Bandwidth and Gain of Bi-doped Glasses

Are these new bismuth-activated optical materials truly useful in broadband fiber amplifiers? In order to elucidate the doubt, we take the $94.5GeO_2 \cdot 5Al_2O_3 \cdot 0.5Bi_2O_3$ (GAB1) glass for example.

The absorption cross section (σ_{ab}) at 808nm for GAB1 glass was calculated to be $1.61\times10^{-20}cm^2$. The stimulated emission cross-section (σ_{em}) at 1280nm was estimated to be $1.55\times10^{-20}cm^2$ by Füchtbauer-Landenburg equation:

$$\sigma_{em}=\frac{\lambda^4}{8\pi n^2 c\cdot\Delta\lambda}\cdot\frac{1}{\tau_{rad}}\cdot g(\lambda)\ ,$$

where λ is wavelength, $g(\lambda)$ is the normalized spontaneous emission shape function, n is the host refractive index, c is light velocity, $\Delta\lambda$ is the FWHM of emission and τ_{rad} is the emission lifetime [13]. In this work, $\Delta\lambda$=322nm, τ_{rad}=273μs and n=1.6166 respectively for GAB1 glass.

As well known, the products of $\sigma_{em}\times$FWHM and $\sigma_{em}\times\tau$ are the important parameters to evaluate the figure-of-merit of bandwidth and gain of broadband amplifier, respectively [5, 48]. The bigger the products, the better the bandwidth and gain properties of the amplifiers. The parameters of $\sigma_{em}\times$FWHM and $\sigma_{em}\times\tau$ for the GAB1 glass are about $499\times10^{-20}cm^2\cdot nm$ and $4.23\times10^{-24}cm^2\cdot s$ respectively, which are bigger than that for Er^{3+}-doped silicate glass ($\sigma_{em}\times FWHM= 22\times10^{-20}cm^2\cdot nm$) [48], that for Cr-doped foresterite ($\sigma_{em}\times\tau= 2.85\times10^{-24}cm^2\cdot s$) [49] and that for Ti-doped sapphire ($\sigma_{em}\times\tau= 1.4\times10^{-24}cm^2\cdot s$) [5]. Therefore, it can be suggested from the above comparison that the Bi-doped glass might be the promising host material for the super-broadband amplifiers.

3.3.2. Excited State Absorption of Bi-doped Glasses

The excited state of X_2 with longer lifetime in the magnitude of several hundred microseconds and the existence of the (H, I) excited states above the X_1 level suggest that the excited state absorption (ESA) possibly appear from X_2 to (H, I) [46, 47]. If the ESA phenomenon is distinguished in these bismuth-doped glasses, any potential amplification will be killed and these glasses will become useless even though they have the super-broad infrared luminescence. In the case of the 808nm-LD excitation, two photons are needed to lift

one electron from X_1 up to (H, I) in view of the energy matching condition. However, the log-log dependence of the infrared emission intensity on the pumping power for the 808nm excitation is nearly linear, which demonstrates only one incident photon is involved in the infrared emission and the ESA can be neglected in the bismuth-doped glasses. Furthermore, the 1.3μm optical amplification has been successfully realized in Bi-doped silica glass in the case of the 0.8μm excitation [50].

3.3.3. Bi-doped Glasses: The Promising Candidates as Gain Mediums of Amplifiers Working at Around 1400nm

The strong absorption of hydroxyl ions in silica optical fibers lies around 1400nm and divides the 1100-to-1600nm region into two low loss windows, viz. the second window centered around 1310nm and the third window centered around 1550nm. Recently, Lucent Technologies succeeded in eliminating almost all water molecules from the silica glass fibers with a new ultrahigh purifying process to realize the ultra-low attenuation at around 1400nm in silica fiber and provided a possibility for using more than 100nm additional bandwidth for optical communication [51]. It is necessary to explore luminescent materials in the wavelength range covering the E+S bands in order to develop the fiber amplifier working around 1400nm (E+S bands). However, there have been few reports on such luminescent materials up to now [5, 9]. The bismuth-doped glasses show the super-wide luminescence with FWHM of larger than 300nm covering not only the E+S bands but also the O band, furthermore tailing into the C (1530-1565nm) + L (1565-1625nm) bands. These excellent properties make it possible to directly amplify the optical signal in the whole optical communication wavelength region only with one fiber amplifier.

3.3.4. Problems Impeding the Development of Bi-doped Glasses

Along with discovering a series of Bi-doped glasses, some problems have also been found, which we think might impede the further development of Bi-doped glasses. The first one is on the infrared luminescence mechanism of Bi-doped glasses. Even though we propose the near infrared luminescence originates from BiO, there is still lack of the direct evidences to support the suggestion [8-17]. The second is regarding the real role of codoping ions, such as Al, Ga, B and Ta, in the generation of near infrared emission centers and the enhancement of the emission intensity [8-17]. The former works show the strong dependence of the optical properties of the Bi-doped glasses on hosts [13]. Therefore the third problem is what kind of host can stabilize the Bi emission centers and facilitate the near infrared luminescence. The last is whether the optical amplification at different wavelengths in 1200~1600nm can be realized in Bi-doped glasses and whether the optical amplification devices can be successfully fabricated. It is no doubt the very important issue to resolve these problems for the further research and development of such luminescence materials.

4. Conclusion

The superbroadband luminescence in 1000~1600nm range, which is greatly different from Bi^{3+}-doped optical materials previously reported in literatures, can observed not only from bismuth and aluminum codoped germanate glass but also from Bi-and-Al-codoped phosphate and borate glasses at room temperature when pumped by 808nm-LD. As codoping of Bi with

Ga, or B, or Ta ions instead of Al ions into GeO_2 and SiO_2 glasses respectively, similar absorptions and emissions to GAB and SAB glasses were found. The luminescent properties of Bi-doped glasses show strong dependence on principle glass compositions, glass preparation temperatures and atmospheres. The abnormal near infrared broadband emission from Bi-doped glasses is proposed to be due to BiO molecules dissolved in the glass hosts after careful comparisons with the data on BiO molecules in gas phase reported by Fink *et al*. The observed absorptions at ~500nm, 700nm, 800nm and 1000nm can be attributed to the transitions from ground state of X_1 to the excited states of (H, I), A_2, A_1 and X_2 respectively, and the emissions at 750nm and ~1300nm can be to the $A_2 \rightarrow X_1$ and $X_2 \rightarrow X_1$ transitions respectively.

The broad infrared luminescence observed in LAZS: LAZS: 0.01 mol.% Cr_2O_3, 1 mol.% Bi_2O_3 glass is not initiated from Cr^{4+} ions but from bismuth, which is remarkably different from the results reported by Batchelor *et al*.

In summary, the ~1300nm emission from Bi-doped glasses with FWHM of 200-400nm covers O, E, S, C and L bands (1260~1625nm) and provides one possibility to directly amplify the optical signals in the whole telecommunication wavelength region with single fiber amplifier simultaneously. The excitation scheme of the commercially available 808nm semiconductor laser as well as the fluorescent lifetime of 200-700μs provides a promising route for developing the high-power laser sources. The attractive properties of the Bi-doped glasses make them useful for applications in optical communication.

References

[1] Tanabe, S. Glastechn. Ber. 2001, 74C, 67.
[2] Yamjada, M.; Ono, H.; Ohishi, Y. *Electron. Lett*. 1998, 34(15), 1490.
[3] Gapontsev, D.V.; Chernikov S.V.; Taylor, J.R. *Optics Comm*.1999, 166, 85.
[4] Tanabe, S.; Feng, X. *Appl. Phys. Lett*. 2000, 77(6), 818.
[5] Suzuki, T.; Ohishi, Y. *Appl. Phys. Lett*. 2004, 84(19), 3804.
[6] Zharikov, E.V.; Smirnov, V.A. Wide-gap luminescent materials: theory and applications, Rotman, S. R.; Ed.; Kluwer academic publishers, 1997; pp. 13-137.
[7] Fujimoto, Y.; Nakatsuka, M. Jpn. *J. App. Phys*. 2001, 40, L279.
[8] Peng, M.; Qiu, J.; Chen, D.; Meng, X.; Yang, L.; Jiang, X.; Zhu, C. *Opt. Lett*. 2004, 29(17), 1998.
[9] Peng, M.; Meng, X.; Qiu, J.; Zhao, Q. *Chem. Phys. Lett*. 2005, 403, 410.
[10] Peng, M.; Qiu, J.; Chen, D.; Meng, X.; Zhu, C. *Opt. Express* 2005, 13(18), 6892.
[11] Peng, M.; Meng, X.; Chen, D.; Qiu, J. *Appl. Phys. Lett*. 2005, 87(6), 066103-1.
[12] Peng, M.; Qiu, J.; Chen, D.; Meng, X.; Zhu, C. *Opt. Lett*. 2005, 30 (18), 2433.
[13] Peng, M.; Wang, C.; Chen, D.; Qiu, J.; Jiang, X.; Zhu, C. *J. Non-Cryst. Solids* 2005, 351(30-32), 2388.
[14] Peng, M.; Chen, D.; Qiu, J.; Jiang, X.; Zhu, C. *Opt. Mater*. 2006, in press.
[15] Meng, X.; Peng, M.; Chen, D.; Yang, L.; Jiang, X.; Zhu, C.; Qiu, J. *Chin. Phys. Lett*. 2005, 22(3), 615.
[16] Meng, X.; Qiu, J.; Peng, M.; Chen, D.; Zhao, Q.; Jiang, X.; Zhu, C. *Opt. Express* 2005, 13(5), 1635.

[17] Meng, X.; Qiu, J.; Peng, M.; Chen, D.; Zhao, Q.; Jiang, X.; Zhu, C. *Opt. Express* 2005, 13(5), 1628.
[18] Volf, M. B. Chemical approach to glass, in "Glass Science and Technology", Vol. 7; Elsevier Science Publishing Company, 1984; pp 406-410 and pp 465-469.
[19] Rvindra, N.M.; Weeks, R.A.; Kinser, D.L. *Phys. Rev. B* 1987, 36(11), 6132.
[20] Fu, J. *J. Non-Cryst. Solids* 1996, 194, 207.
[21] Chen, D.; Jiang X.; Zhu, C. *Acta Phys. Sinica* 2001, 50(8), 1501.
[22] Sen, S.; Stebbins, J. *J. Non-Cryst. Solids* 1995, 188, 54.
[23] Fujiyama, T.; Yokoyama, T.; Hori, M.; Sasaki, M. *J. Non-Cryst. Solids* 1991, 135, 198.
[24] Thomas, I.; Payne, S.; Wilke, G. *J. Non-Cryst. Solids* 1992, 151, 183.
[25] Arai, K.; Namikawa, H.; Kumata, K.; Honda, T.; Ishii, Y.; Handa, T. *J. App. Phys*. 1986, 59(10), 3430.
[26] Risbud, S.; Kirkpatrick, R.; Taglialavore, A.; Montez, B. *J. Am. Ceram. Soc*. 1987, 70(1), C-10.
[27] Shimbo, M.; Tai, S.; Tanzawa, K. *J. Am. Ceram. Soc*. 1986, 69(1), 23.
[28] Shannon, R.D. *Acta Cryst.* 1976, A32, 751.
[29] Batchelor, C.; Chung, W.; Shen, S.; Jha, A. *Appl. Phys. Lett*. 2003, 82(23), 4035.
[30] Paul, A.; Douglas, R. Phys. *Chem. Glasses* 1968, 9(1), 27.
[31] Boulon, G. Mater. *Chem. Phys*. 1987, 16, 301.
[32] Blasse, G.; Brill, A. *J. Chem. Phys*. 1967, 47(6), 1920.
[33] Blasse, G.; Brill, A. *J. Chem. Phys*. 1968, 48(1), 217.
[34] Blasse, G. *Prog*. *Solid State Chem*. 1988, 18, 79.
[35] Van Der Voort, D.; Blasse, G. *J. Solid State Chem*. 1992, 99, 404.
[36] Parke, S.; Webb, R.S. *J. Phys. Chem. Solids* 1973, 34, 85.
[37] Blasse, G.; Meijerink, A.; Nomes, M.; Zuidema, J. *J. Phys. Chem. Solids* 1994, 55(2), 171.
[38] Srivastava, A.M. *J. Lumin*. 1998, 78, 239.
[39] Retoux, R.; Studer, F.; Michel, C.; Raveau, B.; Fontaine, A.; Dartyge, E. *Phys. Rev. B* 1990, 41(1), 193.
[40] Salem-Sugui, S.; Jr.; Alp, E.E.; Mini, S.M.; Ramanathan, M.; Campuzano, J.C.; Jennings, G.; Faiz, M. *Phys. Rev. B* 1991, 43(7), 5511.
[41] Duffy, J.A.; Ingram, M.D. *J. Non-Cryst. Solids* 1976, 21, 373.
[42] Duffy, J.A. *J. Non-Cryst. Solids* 1996, 196, 45.
[43] Pei, Z.; Su, Q.; Zhang, J. *J. Alloy Compd*. 1993, 198, 51.
[44] Peterson, J.; Xu, W.; Dai, S. *Chem. Mater*. 1995, 7(9), 1686.
[45] Xu, W.; Peterson, J. *J. Alloy Compd*. 1997, 249, 213.
[46] Fink, E.H.; Setzer , K.D.; Ramsay, D.; Vervloet, M. *Chem. Phys. Lett*. 1991, 179, 103
[47] Shestakov, O.; Breidohr, R.; Demes, H.; Setzer, K.D.; Fink, E.H. J. *Mol. Spectrosc*. 1998, 190, 28.
[48] Yang, J.; Dai, S.; Zhou, Y.; Wen, L.; Hu, L.; Jiang, Z. *J. Appl. Phys*. 2003, 93(2), 977.
[49] Murata, K.; Fujimoto, Y.; Kanabe, T.; Fujita, H.; Nakatsuka, M. *Fusion Eng. Des*. 1999, 44, 437.
[50] Fujimoto, Y.; Nakatsuka, M. *App. Phys. Lett*. 2003, 82(19), 3325.
[51] Keiser, G. *Optical Fiber Communications*, 3rd Ed., the McGraw-Hill Companies, 2000, pp 10.

INDEX

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z