High Speed Optoelectronics: Photodiodes, Q-Switched Laser Diode and Photoconductive Sampling

Thesis by

Joel Paslaski

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1990

(Submitted May 30, 1990)

Acknowledgments

I would like to express my sincere gratitude to my advisor, Professor Amnon Yariv, for his encouragement and support, and the opportunity to participate and learn in his research group. His keen physical intuition and talents as a scientist have been an inspiration as well as an education. It has been an enjoyable and rewarding privilege to be a member of his pioneering quantum electronics group.

I would like to thank the following people for their valuable collaboration in this work: Professor Yasuhiko Arakawa, Dr. Anders Larrson, Professor Hadis Morkoç, and Dr. Howard Chen. I would also like to thank following for their collaboration on the many rewarding projects not presented here: Kazuhisa Kaede, Yoshikazu Hori, Dr. T. R. Chen, Dr. Pam Derry, Dr. Mark Cronin - Golomb, Dr. Kam Lau, and Steve Sanders. I would especially like to thank Dr. Maobin Yi for his patient and enthusiastic assistance, and Professor Kerry Vahala who got me started in some of my first research work and has constantly been available to assist, advise and provide invaluable discussion throughout my graduate career.

I would also like to thank Ali Ghaffari for growing many of the wafers needed for this work and Larry Begay for his expert assistance in building mechanical apparatus for many experiments. I would also like to express my deepest appreciation to Desmond Armstrong for building much of the instrumentation for experiments, as well as patiently sharing his expertise when I wanted to learn to build instruments myself.

I would like to also thank my fellow students for many valuable discussions as well as providing a friendly and enjoyable atmosphere to work in: Dr. Hal Zarem, Dr. Michael Mittelstein, Sidney Kan, Lars Eng, and Thomas Schrans. I would also like to thank Dr. Steve Smith, Dr. L. C. Chiu, and Dr. Christoph Harder for their friendly support when I joined the group. I also thank Jana Mercado for her

assistance, encouragement and good humor.

I also gratefully acknowledge the financial support received from the California Institute of Technology, the National Science Foundation, and the Office of Naval Research.

Finally, I wish to express my deepest gratitude to my parents for their support and encouragement.

Abstract

In this thesis, a variety of topics related to high speed optoelectronic devices and measurement techniques using ultrafast optical pulses are presented.

Following a brief introduction, the second chapter describes a Q-switched semiconductor laser using a multi-quantum well active layer both for gain and as an intracavity loss modulator. While Q-Switching does not produce as short a pulse as modelocking, it does offer the advantage of adjustability of the repetition rate making it attractive as a source for digital communication links. It is also found to be preferred to the similar approach of gain switching due to less demanding requirements on the rf modulation power level and waveform. Results include a pulse width of $\sim$ 20 ps which is fairly independent of the repetition rate, and a limiting repetition rate of 3.2 GHz. The onset of an irregular pulse train which limits the maximum modulation frequency, is analyzed by a graphical approach.

The potential for optical interconnects has motivated a marriage between the two technologies of Si VLSI and GaAs optoelectronics. Direct integration by the growth of GaAs on Si had been impossible, but the MBE and MOCVD techniques now enable the growth of such layers and of a quality suitable for devices. The third chapter describes the operating characteristics of GaAs-on-Si lasers and photodiodes with particular attention to their high speed performance. Both the lasers and photodiodes show comparable high speed performance to similar structures fabricated on GaAs, with most of the shortcomings being in their dc characteristics.

In the fourth chapter, a novel approach to improving the resolution of photoconductive sampling is presented, called differential sampling. This technique obviates the need for carrier lifetime reduction usually used to improve temporal resolution, and is in principal only limited by a small (few ps) RC circuit time. An analysis of the minimum detectable signal voltage shows the technique does quite

well compared with lifetime reduction techniques which also tend to reduce mobility and dark resistance. An experimental demonstration of this technique is presented in chapter five. Using a two gap sampler, accurate measurement (10 ps resolution) of a 60 ps pulse response from a photodiode is achieved using photoconductors with a recovery time of only 150 ps. Performance near the fundamental Johnson noise limit is also attained, though the minimum detectable signal is higher than predicted due to low response of the photoconductors(probably due to poor contacts).

Finally, in chapter six, the possibility of retrieving an impulse response from its autocorrelation is explored. The use of the logarithmic Hilbert transform for phase retrieval has been discounted in the literature since most such work is concerned with imaging problems for which it is not appropriate due to their symmetric nature. However, causality and the decay nature of transient phenomena make this technique very suitable for use with the impulse response of passive devices. Conditions for the validity of this technique for temporal problems are presented. Simulated retrieval of two functions with similar autocorrelations is demonstrated with sufficient clarity to distinguish them, as well as showing good agreement with the original. Practical limitations and aspects — such as noise, finite time domain, etc. — are also simulated and discussed.

Table of Contents

Chapter 1
Introduction

In 1981 the colliding pulse modelocked (CPM) dye laser was introduced, capable of generating 90 fs optical pulses which could be reduced to 30 fs using newly developed pulse compression techniques. Later, optimization of the cavity design resulted in 27 fs pulses which were subsequently compressed to an incredible 8 fs which represents only 4 periods of the optical frequency. At the same time, synchronously pumped systems were developed and offered commercially. Achievable pulse widths were a more modest 1-5 ps but they offered greater versatility in the way of tunability, available wavelength ranges through dye selection, as well as more convenient operation. The availability of such sources has proven a very valuable tool for studies of ultrafast phenomena in a diversity of fields including biology, chemistry, solid state physics, and especially in high speed electronics and optoelectronics.

In this last category, the emergence of the ultrashort pulse technology was complemented by new fabrication technologies such as molecular beam epitaxy (MBE), metal-organic chemical vapor deposition (MOCVD), and submicron lithography. These techniques have spawned a multitude of novel high speed devices including high electron mobility transistors (HEMT), resonant tunneling diodes, quantum well lasers and optical modulators, etc. Ultra short optical pulse capabilities have been instrumental in the characterization of most of these high speed devices, serving both as an ideal source for impulse excitation of devices, as well as forming the basis for new sampling techniques. While several such sampling schemes have been developed, the two most notable have been electro-optic sampling — with an exceptional resolution of $\sim$ 300 fs — and photoconductive sampling with a resolution of 2-10 ps but much better sensitivity.

Prior to the development of these techniques, high speed measurement in the time domain was basically limited to conventional sampling oscilloscopes (resolution of 25-35 ps) as well as more exotic sampling using Josephson junctions which had better resolution, but were inconvenient and not readily available. It is difficult to imagine how present day high speed optoelectronics could have developed without the measurement capabilities afforded by new optical pulse techniques.

In this thesis, a variety of topics related to high speed optoelectronic devices and measurement techniques using ultrafast optical pulses are presented.

While modelocked dye lasers are a must for the shortest optical pulses, a laser diode source would be much preferred due to its convenience, compactness, and efficiency. The second chapter describes a Q-switched semiconductor laser using a multi-quantum well active layer both for gain and as an intracavity loss modulator. While Q-Switching does not produce as short a pulse as modelocking, it does offer the advantage of adjustability of the repetition rate making it attractive as a source for digital communication links. It is also found to be preferred to the similar approach of gain switching due to less demanding requirements on the rf modulation power level and waveform. Results include a pulse width of $\sim$ 20 ps and a limiting repetition rate of 3.2 GHz.

In the early days of GaAs development, many speculated that Si technology would be replaced by the "superior" GaAs. It is now clear that Si is here to stay particularly in applications utilizing the highly developed VLSI capabilities. However, the potential for optical interconnects has motivated at least a marriage between the two technologies. Direct integration by the growth of GaAs on Si had been impossible, but the MBE and MOCVD techniques now enable the growth of such layers and of a quality suitable for devices. The third chapter describes the operating characteristics of GaAs-on-Si lasers and photodiodes with particular

attention to their high speed performance. Both the lasers and photodiodes show comparable performance to similar structures fabricated on GaAs.

In the fourth chapter, a novel approach to improving the resolution of photoconductive sampling is presented, called differential sampling. This technique obviates the need for carrier lifetime reduction usually used to improve temporal resolution, and is in principal only limited by a small (few ps) RC circuit time. An analysis of the minimum detectable signal voltage shows the technique does quite well compared with lifetime reduction techniques which also tend to reduce mobility and dark resistance. An experimental demonstration of this technique is presented in chapter five. Using a two gap sampler, accurate measurement (10 ps resolution) of a 60 ps pulse response from a photodiode is achieved using photoconductors with a recovery time of only 150 ps. Performance near the fundamental Johnson noise limit is also attained, though the minimum detectable signal is higher than predicted due to low response of the photoconductors (probably due to poor contacts).

Finally, in Chapter 6, the possibility of retrieving an impulse response from its autocorrelation is explored. The use of the logarithmic Hilbert transform for phase retrieval has been studied in the literature mostly in connection with imaging problems for which it is not appropriate due to their symmetric nature. However, causality and the decay nature of transient phenomena make this technique very suitable for use with the impulse response of passive devices. Simulated retrieval of two functions with similar autocorrelations is demonstrated with sufficient clarity to distinguish them, as well as showing good agreement with the original.

Chapter 2
Q-Switched Quantum Well Laser

2.1 Introduction

The technique of Q-switching to produce short, high power pulses from a laser is almost as old as the laser itself [1]. The basic principle is to raise the cavity losses of the laser to a high level to suppress lasing, and thus enable the gain to reach a high level without being clamped by stimulated transitions. Then the losses are quickly reduced leaving the laser far above threshold which results in a rapid build-up of photons and subsequent depletion of the gain. The result is a pulsed output containing most of the energy stored in the gain medium and with a width limited by the photon lifetime of the laser cavity. It has probably achieved its greatest success in solid state systems such as Nd:YAG where a long spontaneous lifetime (5.5×10^{-4}sec) enables the accumulation of a large quantity of energy in the gain medium which is then discharged in a short pulse (~15 ps) resulting in extremely high peak powers of $> 10^7$ W for a typical laboratory laser. In semiconductor lasers, a much shorter spontaneous lifetime (~1 ns) precludes the build-up of much stored energy with typical pumping currents, and the output is limited to pulses of about 30 ps with peak power of at most 1 W. Although high powers are not achieved, Q-switching is still of interest in semiconductor lasers for its ability to produce short pulses for use in high speed digital communications. While the technique of modelocking can produce even shorter pulses, it typically requires an external cavity and the pulse repetition rate is constrained to a fixed value which essentially precludes the encoding of information. Conversely, Q-switching can produce an almost arbitrary sequence of pulses limited only by a maximum pulse rate. The related technique of gain switching also offers this capability but typically requires

large amplitude (10 - 15 V) electrical pulses of short duration (<100 ps) in order to push the gain sufficiently far above threshold before the pulse forms. As will be shown, successful Q-switching is achieved using sinusoidal modulation with an amplitude of a few volts.

The crucial element to any Q-switched laser is the element which is used to adjust the cavity losses. This element must satisfy a few basic requirements to be effective for Q-switching. In order to attain a gain far above threshold, the overall losses of the cavity must be varied from some minimum value to a value greater than the desired initial gain maximum. As well as providing a large threshold variation, the loss modulator must also do so in a time which is short compared to the pulse forming process. Finally, the modulator must be insertable in the laser cavity. For laser diodes, this means the loss modulator must be a semiconductor device which can be integrated in the laser chip. Proposals for integrated loss modulators have included electrooptically switched [2] and acoustooptically switched [3] distributed feedback gratings to effectively vary the mirror reflectivity. An electroabsorption modulator based on the Franz-Keldysh effect has also been proposed and demonstrated for Q-switched operation [4]-[6]. In this chapter, a quantum well version of this device is described with regard to Q-switching.

2.2 Device Structure and DC Characterization

The laser structure — shown in Fig. 2.1— consists of an amplifier section and an electroabsorption modulator section which are based on a multiquantum well (MQW) active layer grown by molecular beam epitaxy. A ridge waveguide is used for optical confinement while SiO_2 blocking layers confine current injection to the top of the ridge. The separate gain and loss sections are defined solely by separate electrical contacts which can be independently biased. The top p^+-GaAs contact layer is etched away in the 5 μm gap between the contacts, but otherwise

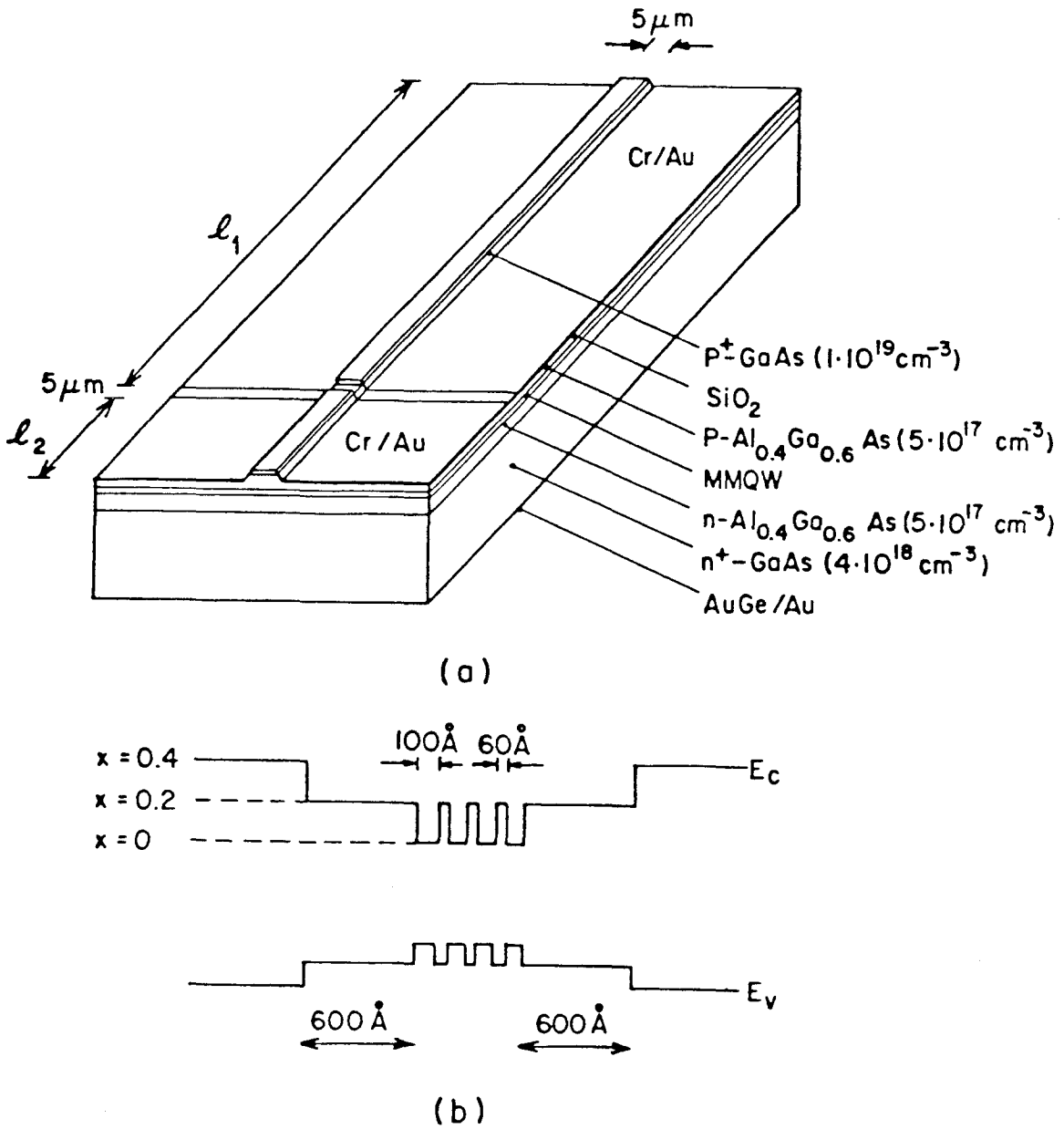


Fig. 2.1 (a) Diagram of the two segment quantum well laser with a ridge waveguide for optical confinement. The lengths of the gain section l_1 and the modulator section l_2 were 250 μm and 50 μm, respectively. (b) The band gap diagram of the multiple quantum well active layer.

the waveguide is continuous at this point so optical reflections are negligible. The measured resistance between the two contacts was 5kΩ which is sufficient to enable independent biasing. The lengths of the gain section, l_1, and loss modulator, l_2, were 250 μm and 50 μm respectively.

The effectiveness of this structure for loss modulation comes from a combination of properties in both the gain and loss sections which are due to the quantum well active layer. The loss modulator is based on the quantum confined Stark effect [7], [8] which is similar to the Franz-Keldysh effect seen in bulk material. When an electric field is applied perpendicular to the quantum well layers, the bottom of the well becomes triangular instead of square. The states of this new well tend to "sink into the corners" and the energies relative to the center of the well decrease for both electrons and holes. Thus the absorption edge corresponding to the ground state transition shifts to lower energy as an electric field is applied. Furthermore, the confinement of the well prevents excitons from being ionized as they are in bulk material, and excitonic resonances are observed at room temperature with large applied fields ($> 10^5$ V/cm). The binding energy of the exciton is reduced somewhat by an external field but the net change in the exciton transitions is still toward lower energy. At high fields, the exciton peaks are broadened and reduced due to tunneling through the confinement barrier which reduces the exciton lifetime. The net result is that the absorption edge of the modulator section can be red-shifted by $\sim$20 meV with the application of an external electric field. Complementing this loss modulation, the band gap of the gain section is reduced by the carrier induced band shrinkage effect which is significant at the high carrier densities required for lasing. This effect is further enhanced in quantum well lasers compared to conventional bulk lasers, and the lasing photon energy is typically reduced by about 20 meV [9]. Thus with the same active layer being used for both the gain and loss sections, the

cavity losses at the lasing wavelength can be varied over a wide range.

The effectiveness of the loss modulator is demonstrated in Fig. 2.2 which shows the DC threshold current (I_{th} for the gain section as a function of the bias voltage V_b applied to the loss modulator. The threshold current is normalized by the value of $I_{th}(V_b = 0)$ which is 115 mA. As can be seen the threshold is increased by about 2.7 times as the modulator voltage is varied from 1 V to -3 V. Due to the built in field of the *pn* junction, the flat field condition occurs at a positive applied voltage (~1.4 V) and loss modulation is still effective at $V_b = 0$.

2.3 Q-Switched Operation

Q-switched operation was obtained as shown in Fig. 2.3, by applying both a dc bias voltage and a microwave signal to the loss section while the gain section was quasi-dc biased with a current pulse (2 μs at 50 kHz rep rate). True dc operation was not possible due to the rather high currents required to get far above threshold. However, for the frequencies applied to the loss modulator, the 2 μs pulsewidth should be long enough for steady state to be established. The microwave modulation was sinusoidal with a typical level of 17 dBm ($\pm$2.24 V) and was varied from 100 MHz to 5 GHz. The laser output was detected by a high speed *pin* photodiode which was connected to a microwave spectrum analyzer for observation of the fundamental modulation frequency as well as higher harmonics indicating short pulses. Signal strength was not high enough, nor is the photodiode fast enough for resolved observation of the output in the time domain. Instead, conventional autocorrelation measurements of the pulses were made using second harmonic generation in a $LiIO_3$ crystal.

Fig. 2.4 shows the autocorrelation trace obtained with a modulation frequency of 1.5 GHz, a dc modulator bias of $V_b = 0$, and a gain section current of 170 mA ($= 1.5I_{th}(V_b = 0)$). The autocorrelation full width at half maximum (FWHM) is

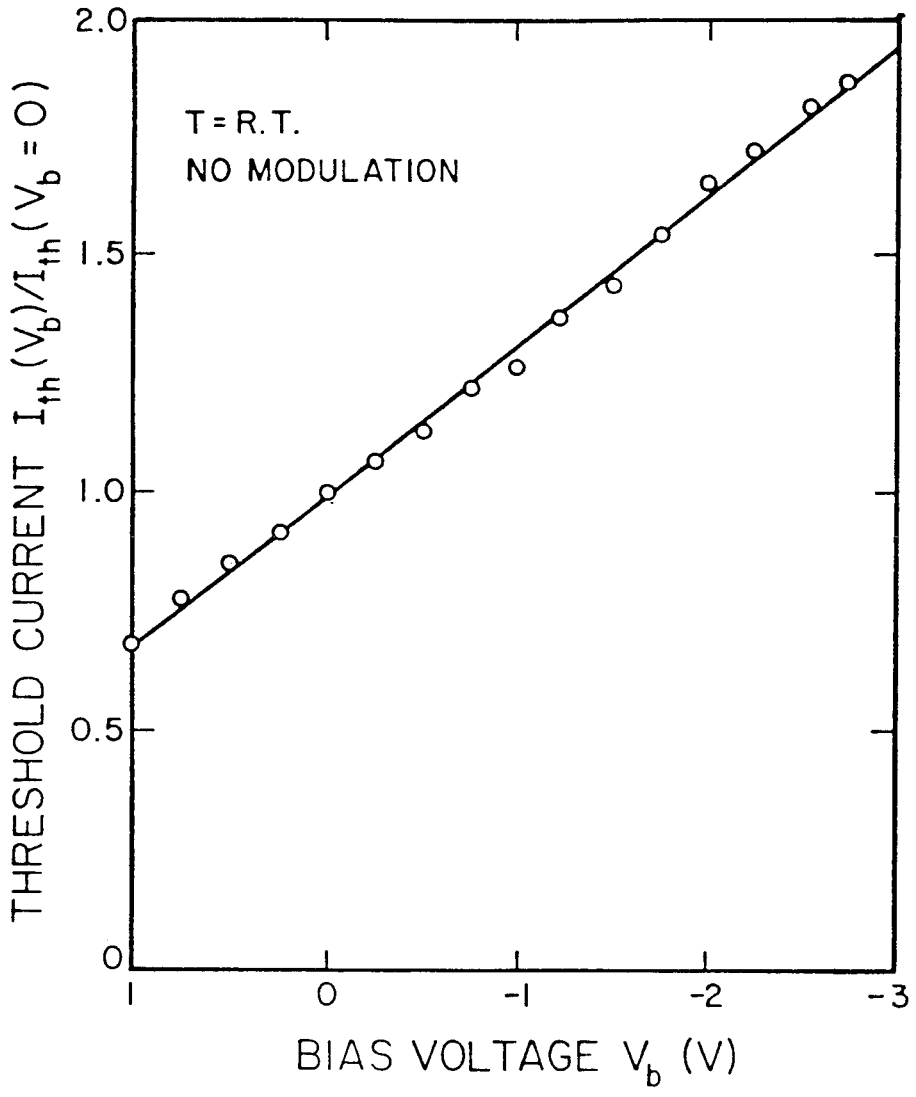


Fig. 2.2 Lasing threshold current I_{th} as a function of the modulator bias voltage V_b, demonstrating effective loss modulation. I_{th} is normalized by $I_{th}(V_b = 0)$ which is 115 mA.

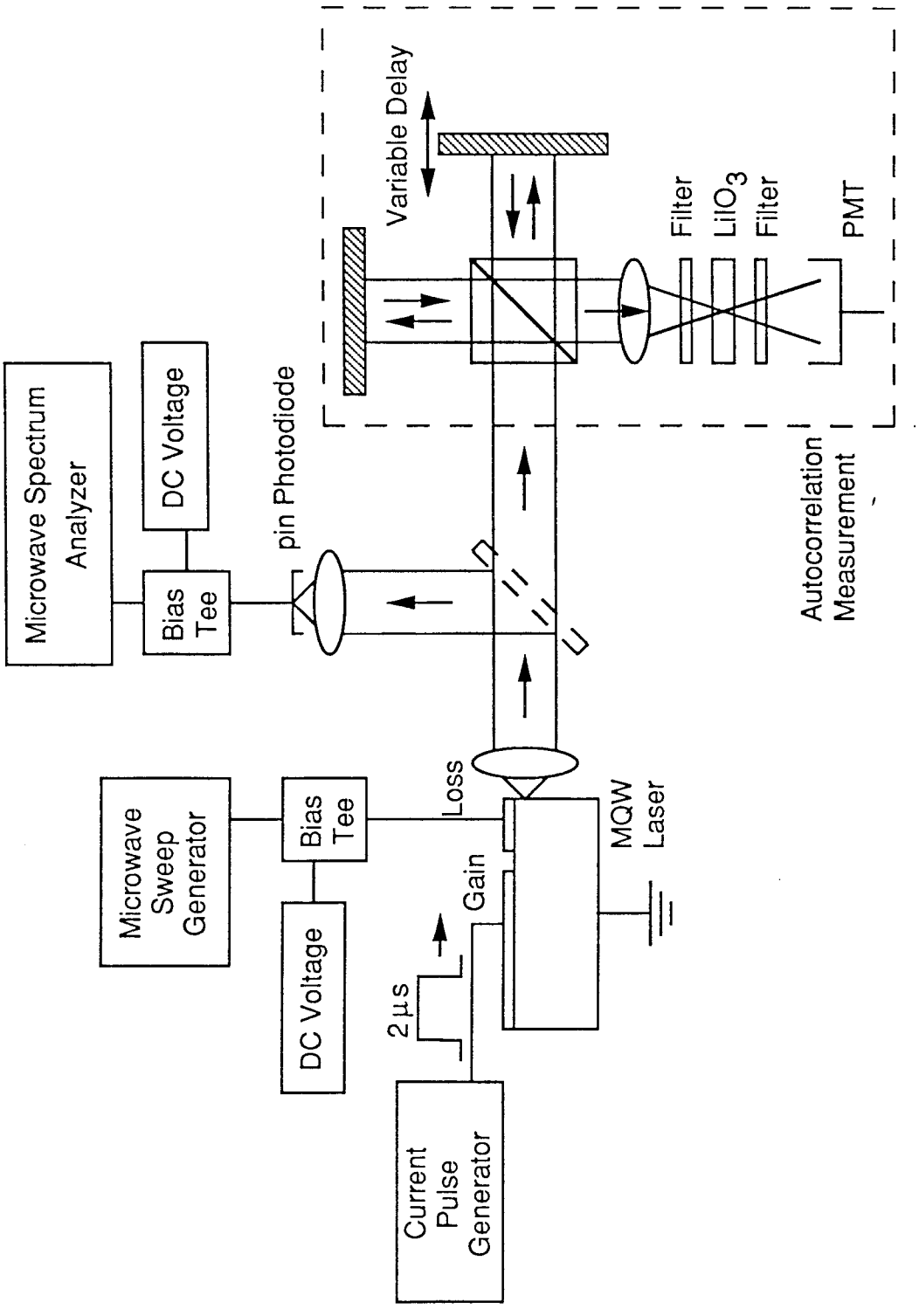


Fig. 2.3 Experimental set-up for Q-switching of two section MQW laser. Pulsed output was measured both in the frequency domain with a photodiode and microwave spectrum analyzer, and in the time domain with an autocorrelation measurement.

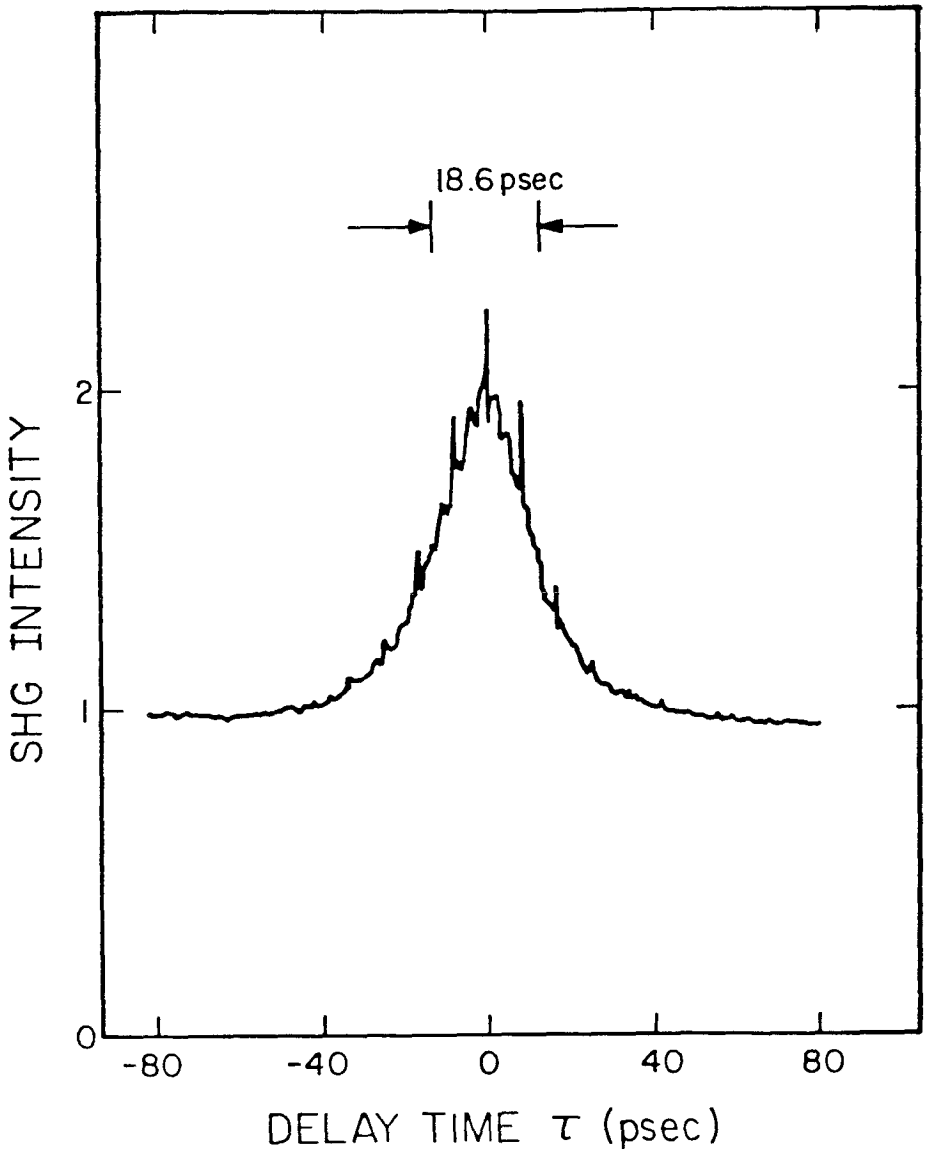


Fig. 2.4 Intensity autocorrelation trace of the Q-switched output at a modulation frequency of 1.5 GHz and I_{in} =170 mA. The autocorrelation FWHM is 26 ps which corresponds to a pulse FWHM of 19 ps if a Gaussian pulse shape is assumed.

26 ps which corresponds to an actual pulse width of 19 ps if a Gaussian pulse shape is assumed. The pulse width is virtually independent of the modulation frequency, changing only from 19 ps to 22 ps as the frequency was varied from 500 MHz to 3.2 GHz. This is consistent with the Q-switching mode of operation whereby pulse formation is governed by the dynamics of rapid photon build-up and subsequent gain depletion, and is relatively independent of the modulation waveform. The pulse width *is* strongly dependent on the current to the gain section as shown in Fig. 2.5. Again this is easily understood in terms of the pulse forming dynamics: a higher initial inversion leads to a more rapid build-up of photons as well as a higher peak photon density, which in turn gives a high stimulated recombination rate to deplete the gain quickly and terminate the pulse.

The autocorrelation shows regularly spaced "coherence spikes" which are well known in such measurements [10]. These are not real features in the intensity, but are related to the coherence of the underlying optical wave which is self-coherent at integer multiples of the laser cavity round-trip time. The narrowness of these spikes indicates low coherence which is corroborated by the optical spectrum which was very multimode and covered an overall width of about 50 Å.

The ability to produce 20 ps pulses suggests that such a laser could be used to transmit at bit rates approaching 25 Gbit/sec. However, the long lifetime of the gain (~2 ns) leads to pattern effects for frequencies above a few hundred MHz, since the amplitude of a given pulse depends on those which preceded it within the gain lifetime. For a fixed frequency modulation, this should lead to a reduced pulse amplitude at higher repetition rates as the gain has less time to recover between pulses. Such modulation frequency dependent effects were investigated by observing the microwave spectrum of the intensity with a photodiode as the loss modulation frequency was varied. Regular pulse generation was observed at the modulation

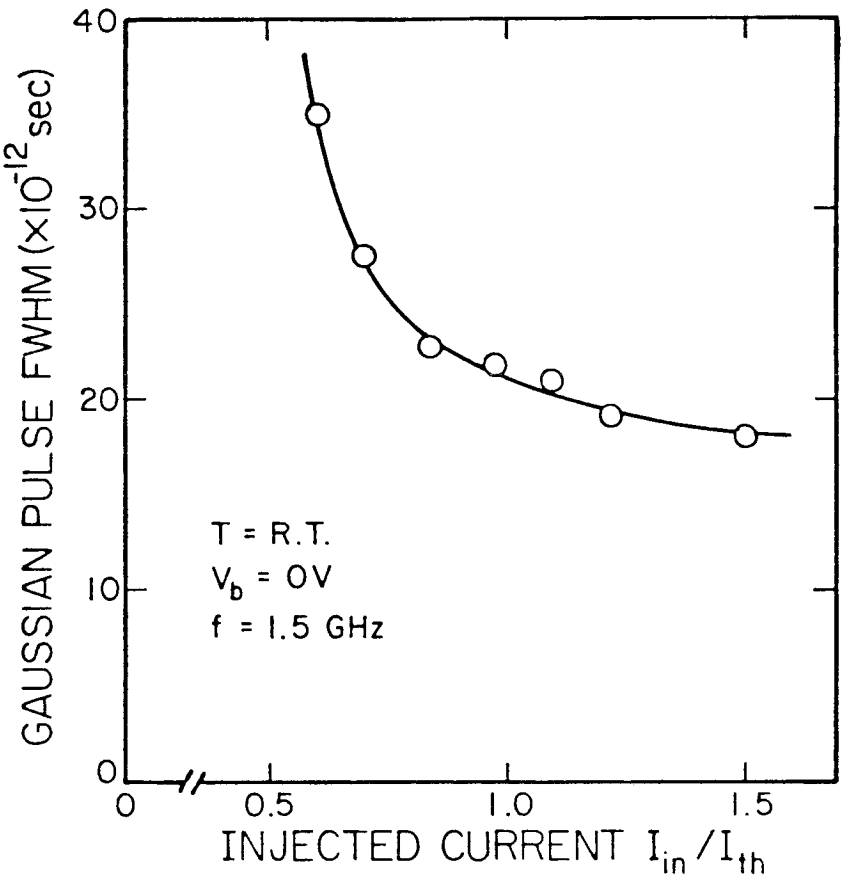


Fig. 2.5 Measured pulse width, assuming a Gaussian pulse shape, versus the injected current to the gain section. Modulation frequency is 1.5 GHz and I_{in} is normalized to the threshold current with no modulation voltage.

frequency up to a maximum frequency of 3.2 GHz as indicated by a spectrum consisting of the fundamental and several higher order harmonics. This is shown in the microwave spectrum traces of Fig. 2.6 for a laser modulated at 800 MHz. The fundamental frequency is accompanied by all integer harmonics out to 12 GHz with a fairly flat amplitude to about 6 GHz, which is the response rating of the photodiode used. Much of the amplitude variation is due to electrical reflections between the sweep oscillator and the high resistance loss modulator which is poorly matched to a 50Ω system. As expected, some reduction in the amplitude of the fundamental was observed at higher modulation frequency due to lower energy per pulse.

As the modulation frequency was increased beyond 3.2 GHz, subharmonics appeared — first at half-integer and then quarter-integer multiples — indicating that the pulse train was no longer regular. The onset of this behavior is shown in the sequence of four microwave spectrum traces of Fig. 2.7. The gain section is biased at $1.5I_{th}$ while the absorber is dc biased at -1 V and modulated with 17 dBm of rf power. At a modulation frequency of 2.24 GHz the spectrum shows only the fundamental frequency and integer harmonics (not visible in the range of 1.7 - 4.1 GHz). As the modulation frequency is increased to 2.35 GHz a signal appears at 3.52 GHz, i.e., 1.5 times the fundamental, as well as at other half-integer multiples of the fundamental. At this point, the half-integer harmonics are well below the fundamental and rather broad indicating that the period doubling behavior is not yet consistent over long times. At a modulation of 2.56 GHz, the half integer harmonic (3.84 GHz) is only 2 dB below the fundamental and quite narrow. Finally, at higher modulation frequencies, further sub-harmonics appear spaced at one-fourth the fundamental. The last trace of Fig. 2.7 shows fundamental modulation of 3.5 GHz, the first half-harmonic at 1.75 GHz, and a signal at 2.63 GHz

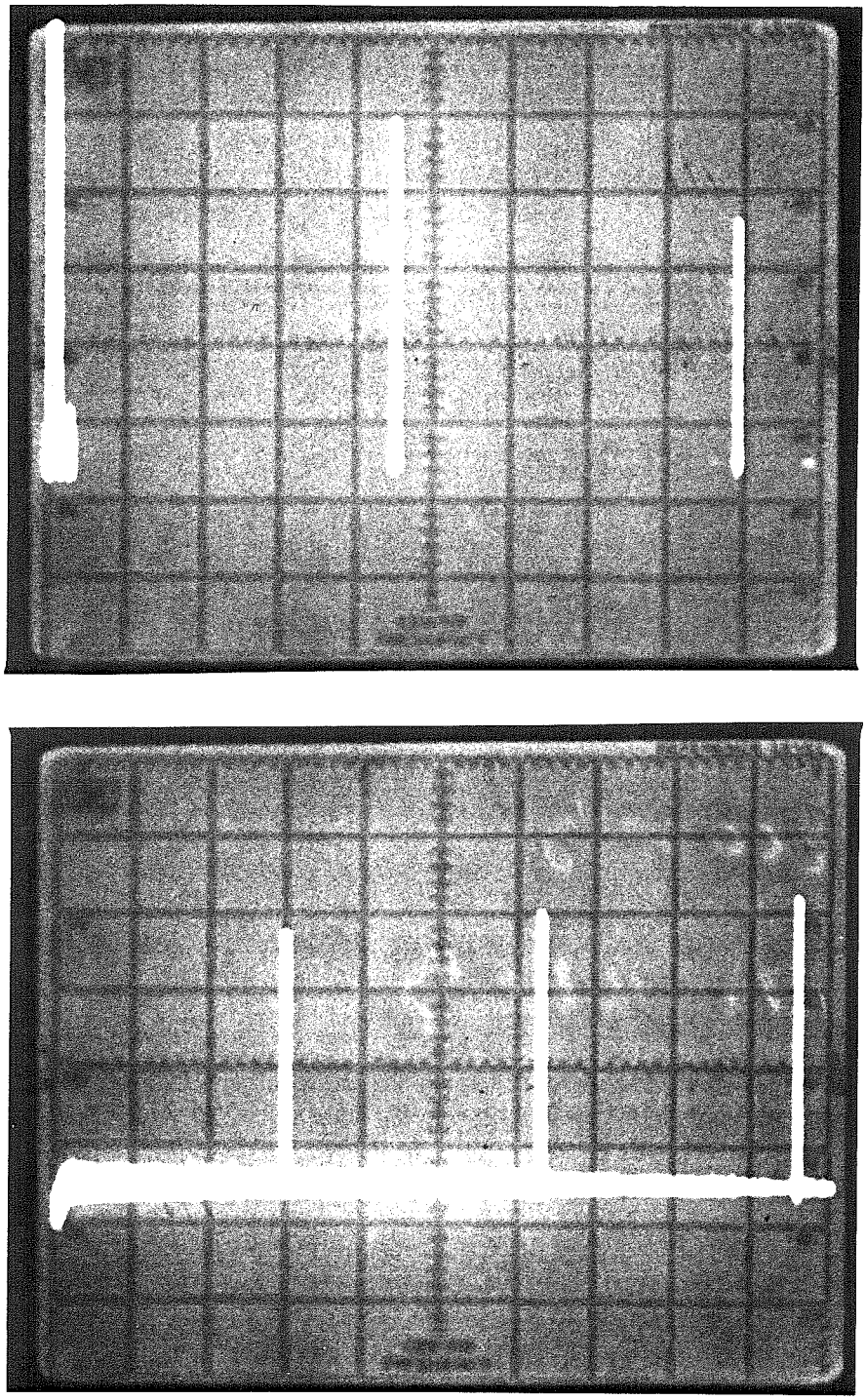


Fig. 2.6 Microwave spectrum analyzer traces of the optical intensity output of Q-switched laser modulated at 800 MHz, showing extensive harmonic content. Top trace covers the frequency range of .01 - 1.8 GHz, and bottom trace covers 1.7 - 4.1 GHz.

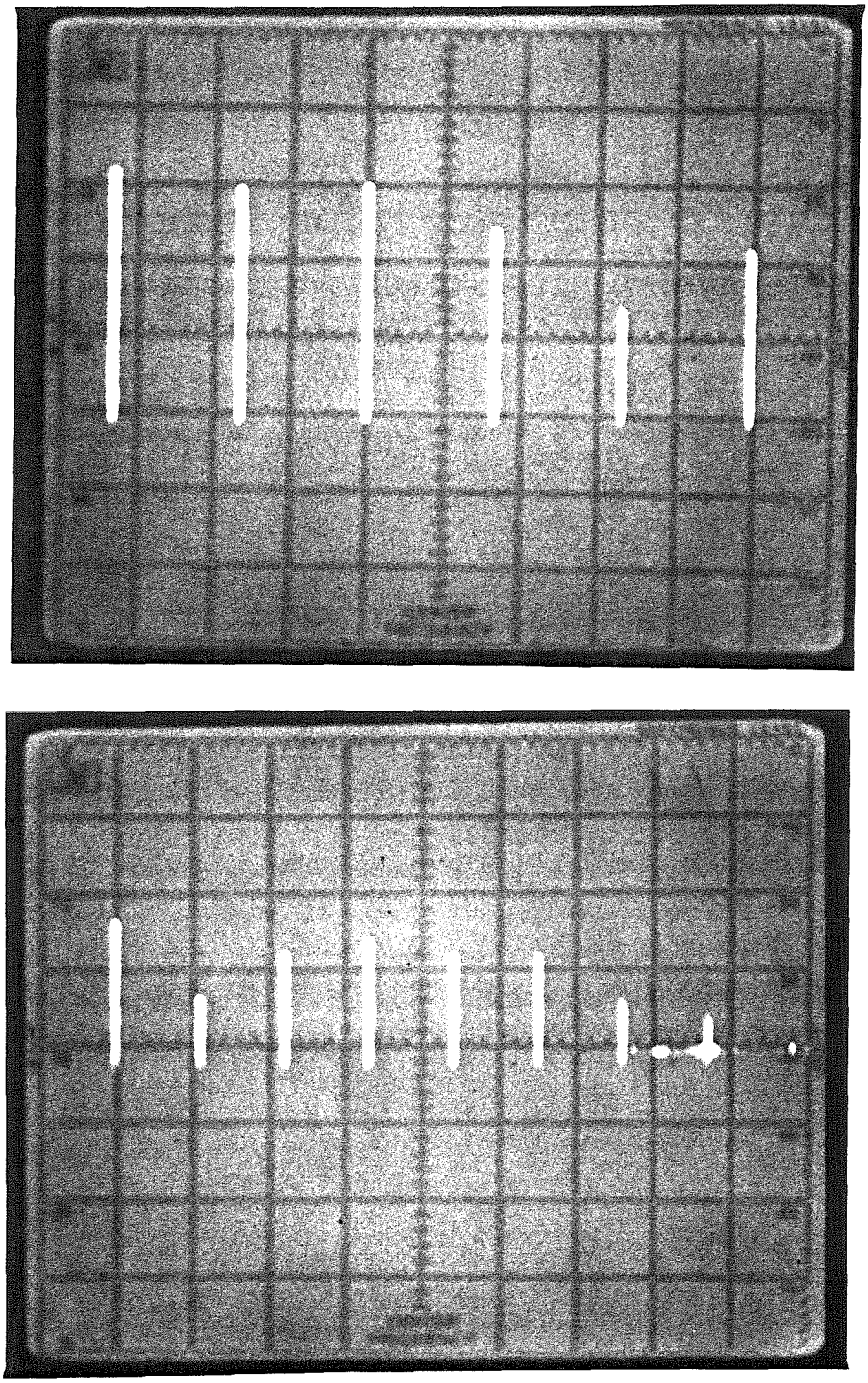


Fig. 2.6 (cont.) Further microwave spectra of Q-switched laser covering frequency ranges of (top) 3.8 - 8.5 GHz and (bottom) 5.8 - 12.9 GHz.

which is 3/4 the fundamental. Though not visible in the limited frequency range, similar signals appear at all quarter-integer harmonics. This quarter-harmonic signal is broad and relatively weak as in the onset of the half harmonic. However, its strength does not increase at higher modulation frequencies. The multiple peaks of this signal are due to cable length related resonances caused by poor impedance matching of the absorber. This chaotic behavior is analyzed in the next section.

2.4 Analysis of Q-Switching Including Chaotic Behavior

When half integer harmonics appear in the intensity spectrum, it is because the pulse train consists of alternating large and small pulses. Qualitatively, this is plausible since a large pulse will better deplete the gain resulting in a low inversion for the next pulse which will then be small resulting in a high inversion for the next pulse which is then large again. Explaining why this mode *should* occur and only at certain frequencies requires a more quantitative analysis. A complete numerical analysis by Tsang et al. [5], including amplified spontaneous emission and spatial variation of the gain and optical field, showed this chaotic, alternating pulse amplitude behavior at higher frequencies. Other analyses of Q-switching of semiconductor lasers are generally numerical models based on the rate equations and successfully predict dependencies of pulse width, delay time, energy utilization factor, etc., on the initial inversion as well as various recombination behaviors and spontaneous emission levels [2]. In spite of their successes, such numerical models generally do not reveal the essential elements of anomalous behavior — information which would be instrumental in deciding if and by what means it might be remedied.

We now develop a more physical — albeit less exact — picture of the Q-switching operation, with particular attention to the chaotic pulse train behavior. The analysis is based on the following rate equations for the average photon density,

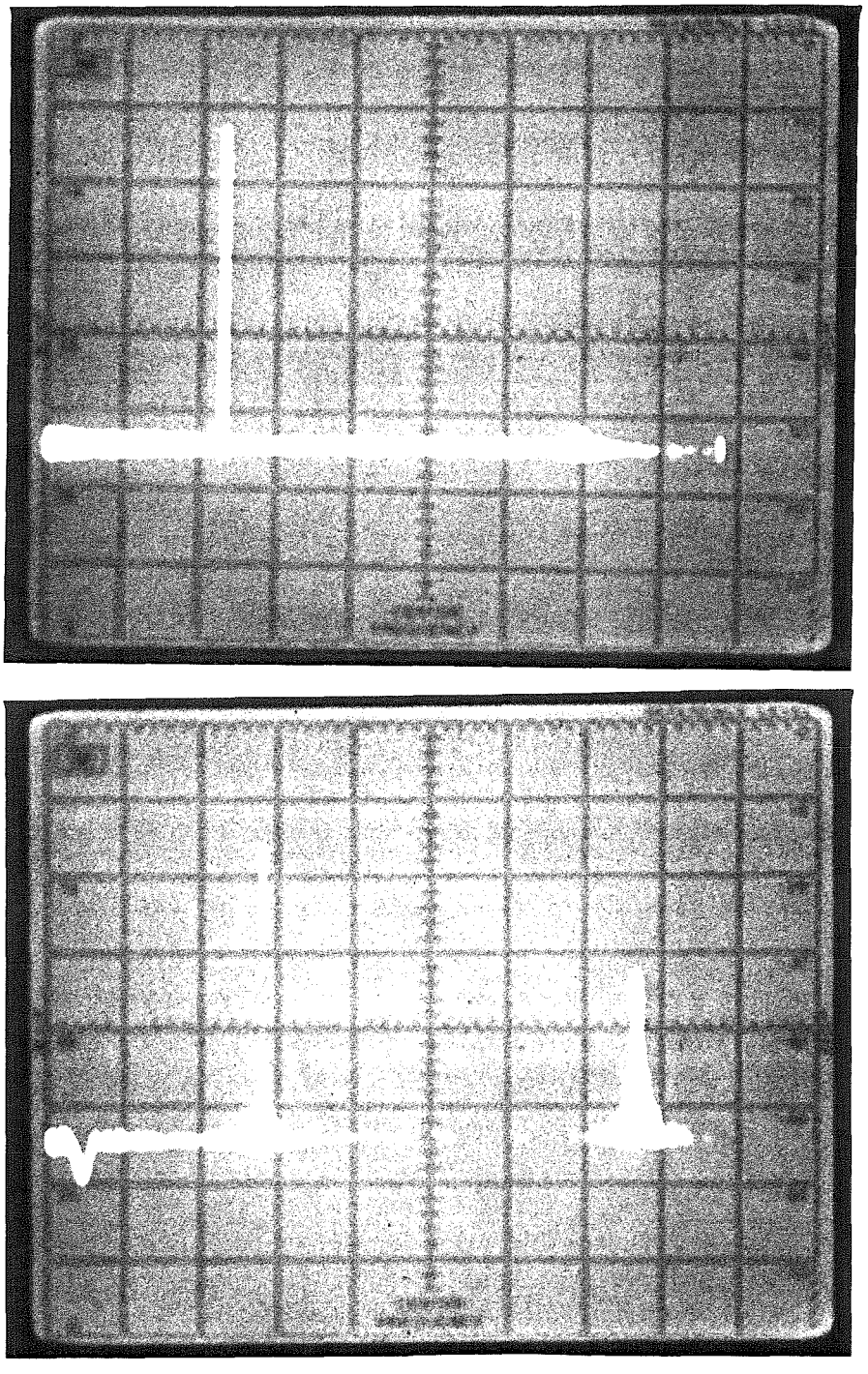


Fig. 2.7 Microwave spectra of Q-switched laser showing onset of sub-harmonics. Frequency range is 1.7 - 4.1 GHz and vertical scale is 10 dB/div. At a modulation frequency of 2.24 GHz (top), only the fundamental appears, while at 2.35 GHz a signal has appeared at 3.52 GHz which is 1.5× the fundamental.

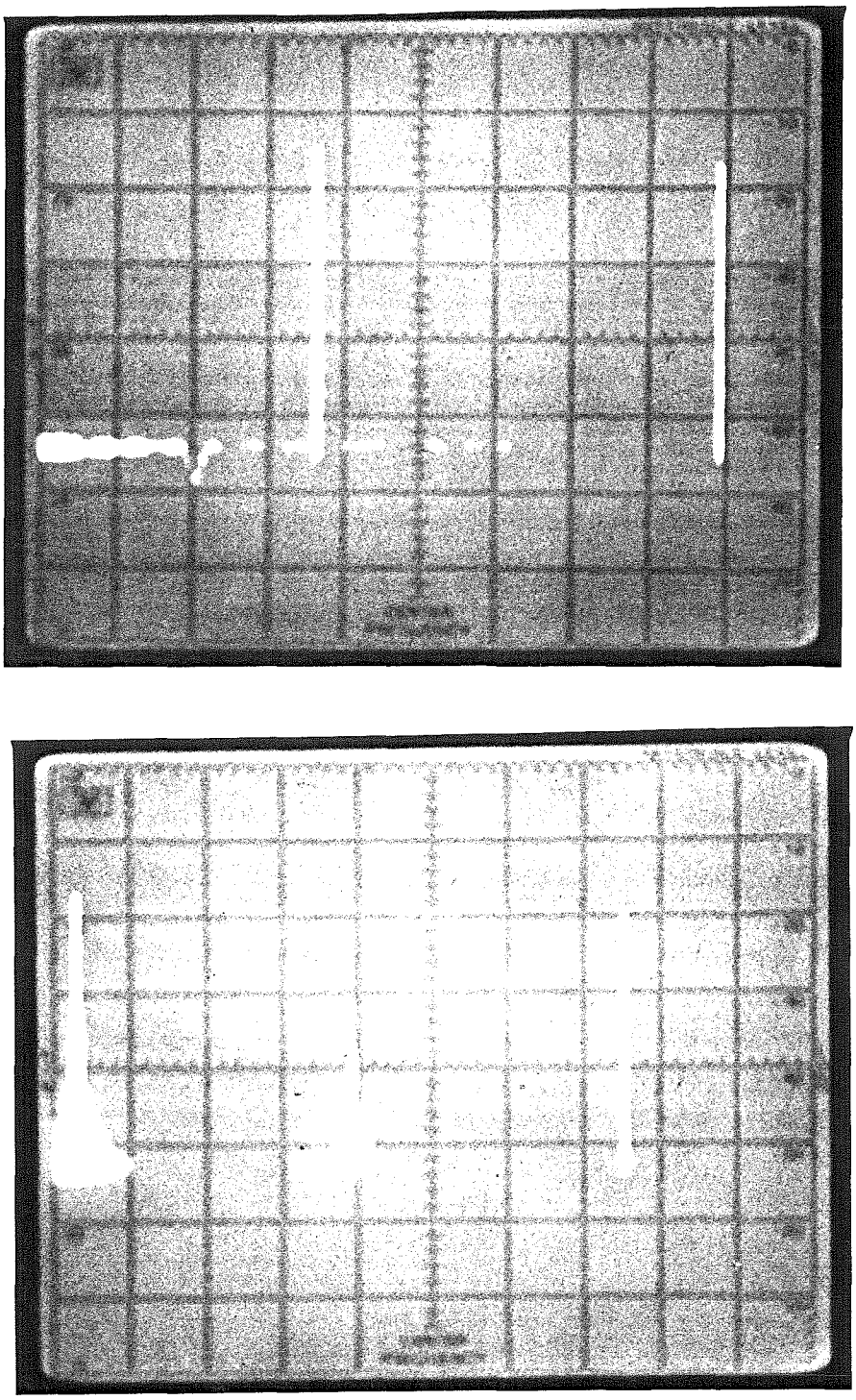


Fig. 2.7 (cont.) At a frequency of 2.56 GHz (top), the half-integer harmonic at 3.84 GHz is very narrow and only 2 dB below the fundamental. Further increase to 3.5 GHz results in the appearance of a signal at 3/4 the fundamental (2.63 GHz) as well as the half-integer harmonic at 1.75 GHz.

p, and gain section carrier density, n, of the laser:

$$\frac{dn}{dt} = \frac{J}{q\,\mathrm{Vol}} - p\,g(n) - \frac{n}{\tau_s} \tag{2.4.1}$$

$$\frac{dp}{dt} = \Gamma p\,g(n) - \frac{p}{\tau_{ph}} + \Gamma\beta\frac{n}{\tau_s}, \tag{2.4.2}$$

where J is the injected pump current, q is the electronic charge, Vol is the active layer volume, τ_s is the spontaneous lifetime (linear recombination assumed), Γ is the optical confinement factor, and β is the spontaneous emission factor. The photon lifetime is defined by $\tau_{ph} \equiv \frac{1}{v_{gr}(\alpha+\frac{1}{L}\ln(\frac{1}{R})}$, where v_{gr} is the photon group velocity, α is the distributed loss constant, L is the laser cavity length, and R is the mirror reflectivity. The dependence of gain on carrier density will be assumed to be of the linear form $g(n) = v_{gr}A(n - n_0)$ where n_0 is the transparency level.

Based on the experimentally observed short pulse operation, the approximation is made to break up the analysis into the two distinct processes of pulse formation and pumping. Each period of the modulation is assumed to consist of a time of zero light while the inversion is pumped to some high level, followed by a short time when the pulse is formed. By establishing relationships between the final and initial inversion for each stage, a steady state solution is found by requiring that the inversion repeat itself every cycle or possibly every few cycles.

Between pulses, the stimulated terms are neglected and the pumping of the inversion is governed by

$$\begin{aligned}\frac{dn/n_{th}}{dt} &= \frac{J}{qn_{th}\mathrm{Vol}} - \frac{n/n_{th}}{\tau_s} \\ &= \frac{1}{\tau_s}[J/J_{th} - n/n_{th}],\end{aligned} \tag{2.4.3}$$

where n_{th} is a threshold carrier density defined below and J_{th} is the corresponding threshold current. Denoting the carrier density immediately after a pulse as n_a

then the value immediately before the next pulse, n_b, is given from the solution of (2.4.3) by

$$\frac{n_b}{n_{th}} = \frac{n_a}{n_{th}} e^{-1/f\tau_s} + \frac{J}{J_{th}}(1 - e^{-1/f\tau_s}), \tag{2.4.4}$$

where the pumping time is taken as the inverse of the modulation frequency f.

The pulse formation relations are now found from the rate equations [11], while neglecting the pumping and spontaneous terms which is valid for short pulses. The modulator losses are assumed to be constant at their minimum value during the entire pulse. Normalizing time by the photon lifetime with $t' = t/\tau_{ph}$, the rate equations are then

$$\frac{dp}{dt'} = p(\Gamma g(n)\tau_{ph} - 1) = p\left(\frac{n - n_0}{n_{th} - n_0} - 1\right) \tag{2.4.5}$$

$$\frac{dn}{dt'} = -p\, g(n)\tau_{ph} = -\frac{1}{\Gamma} p \frac{n - n_0}{n_{th} - n_0}, \tag{2.4.6}$$

where n_{th} is the threshold carrier density implicitly defined by (2.4.5). Dividing these two equations results in

$$\frac{dp}{dn} = \Gamma\left(\frac{n_{th} - n_0}{n - n_0} - 1\right). \tag{2.4.7}$$

Denoting initial and final quantities by i, f respectively, the solution to this is

$$\frac{1}{\Gamma}(p_f - p_i) = (n_{th} - n_0)\ln\left(\frac{n_f - n_0}{n_i - n_0}\right) - (n_f - n_i). \tag{2.4.8}$$

Assuming the optical pulse to build-up from and decay to a small value, then $p_i \simeq 0 \simeq p_f$, and the gain values before and after the pulse are related by

$$\frac{(n_a - n_0)/n_{th}}{(n_b - n_0)/n_{th}} = \exp\left[\frac{(n_a - n_0)/n_{th} - (n_b - n_0)/n_{th}}{1 - n_0/n_{th}}\right]. \tag{2.4.9}$$

The solution to this transcendental equation as well as the pumping curves — Eqn. (2.4.4) — for various values of the frequency parameter $f\tau_s$ are plotted in Fig. 2.8.

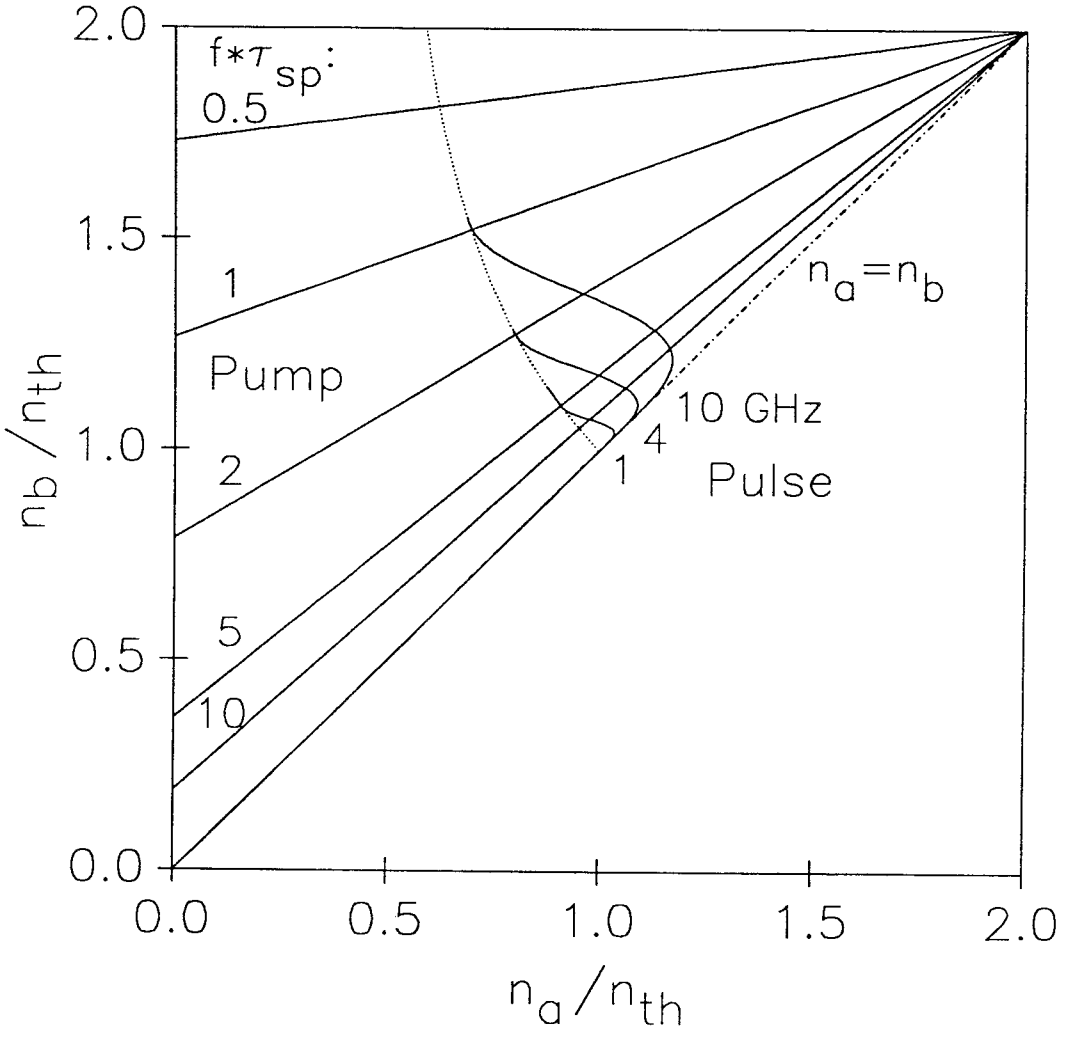


Fig. 2.8 Plots used for graphical analysis of repetitive Q-switching. The axes represent carrier densities before (n_b) and after (n_a) a Q-switched pulse. Pump curves, labeled by $f\tau_s$, give prepulse carrier density as a function of the density following the previous pulse. Pulse curves give $n_a(n_b)$ due to the depleting action of the Q-switched pulse. The dotted-line pulse curve assumes that the full Q-switched pulse occurs without interference by the return of modulator losses. Approximate modified pulse curves including this effect are labeled by frequency.

Also shown in this plot are pulse curves which have been modified to include the effect of clipping of the pulse by the return of high losses in the modulator before depletion of the gain has terminated the pulse. In the limit of high modulation rates, very little depletion or modulation of the carrier density occurs and the output power is just modulated as if by an external modulator. Thus the clipping of the Q-switched pulses by recovering losses can also be regarded as a transition between full Q-switched operation at low modulation frequencies and small signal modulation at high frequencies. This effect has been roughly calculated here by estimating a pulse width and a time window of positive gain, and using this to estimate what fraction of the total expected pulse photons are emitted and hence what fraction of the expected gain depletion occurs. The pulse width, τ_{pw}, is estimated from the ratio of the total photons emitted to the peak rate at which they are lost

$$\tau_{pw} = \frac{\Gamma(n_b - n_a)}{p_{max}/\tau_{ph}} = \frac{n_b - n_a}{(n_{th} - n_0)\left[\ln \frac{n_{th}-n_0}{n_b-n_0} - \frac{n_{th}-n_b}{n_{th}-n_0}\right]}\tau_{ph}. \tag{2.4.10}$$

The peak photon density occurs at $n = n_{th}$ and is evaluated by inserting this in Eqn. (2.4.8). Assuming a sinusoidal modulation of the losses such that the threshold carrier density varies by Δn_{th}, then the window for positive gain is approximately given by

$$\tau_{gw} = \frac{1}{\pi f}\cos^{-1}\left(1 - \frac{n_b - n_{th}}{\Delta n_{th}/2}\right).$$

For low initial gain, the Q-switched pulses are fairly symmetric so we take a normalized pulse shape of $[1 - \cos(\frac{\pi t}{\tau_{pw}})]/2$. Finally we make the approximation that only that fraction of the Q-switched pulse which falls within the gain window is actually emitted, and consequently the actual gain depletion is the same fraction of the total depletion expected from Q-switching:

$$(n_b - n_a)_{actual} = (n_b - n_a)_{\text{full}}\left[\frac{\tau_{gw}}{2\tau_{pw}} - \frac{1}{2\pi}\sin(\pi\,\tau_{gw}/\tau_{pw})\right], \qquad \tau_{gw} < 2\tau_{pw}.$$

The resulting modification to the pulse curves is now frequency dependent due to the frequency dependence of the gain window τ_{gw}. While this approach is very approximate and neglects some important effects such as the time delay from pulse turn-on to peak, it shows the basic behavior that a more exact calculation shows and sets a best case limit to this behavior. The essence of this modification is that at low n_b and high frequency, the pulse gets clipped before it gets large so little depletion occurs and $n_a \simeq n_b$. At higher n_b, Q-switched pulses are shorter — see Fig. 2.4 — due to faster photon build-up and higher peak stimulated rates. Thus the pulse will be unaffected by the modulator and the curve returns to the original full Q-switching curve.

The significance of this modification is apparent when considering the stability of steady state operating points corresponding to the intersection of the pump and pulse curves. This is illustrated graphically in Fig. 2.9. In the top figure, hypothetically consider a value of n_a ($\simeq 0.75$) different from the intersection point. After pumping, the value of n_b is given by the point "a" which leads to an n_a given by point "b" after the pulse. This process continues along the path indicated, and converges to the intersection of the pump and pulse curves showing that this is a stable, steady state solution. Conversely, for the situation in the lower figure, the sequence of points "abcde" diverges away from the intersection and reaches another steady state indicated by the rectangle "ABCD." A check of an initial point outside this rectangle shows that this path is stable. This path is that of a period doubled output since the pre-pulse carrier density— and hence the pulse amplitude — alternates between the values of n_b at the points "A" and "C," with pulses produced at the modulation frequency.

What then distinguishes the stable and unstable operating points associated with the intersection points? A little playing around with such diagrams shows that

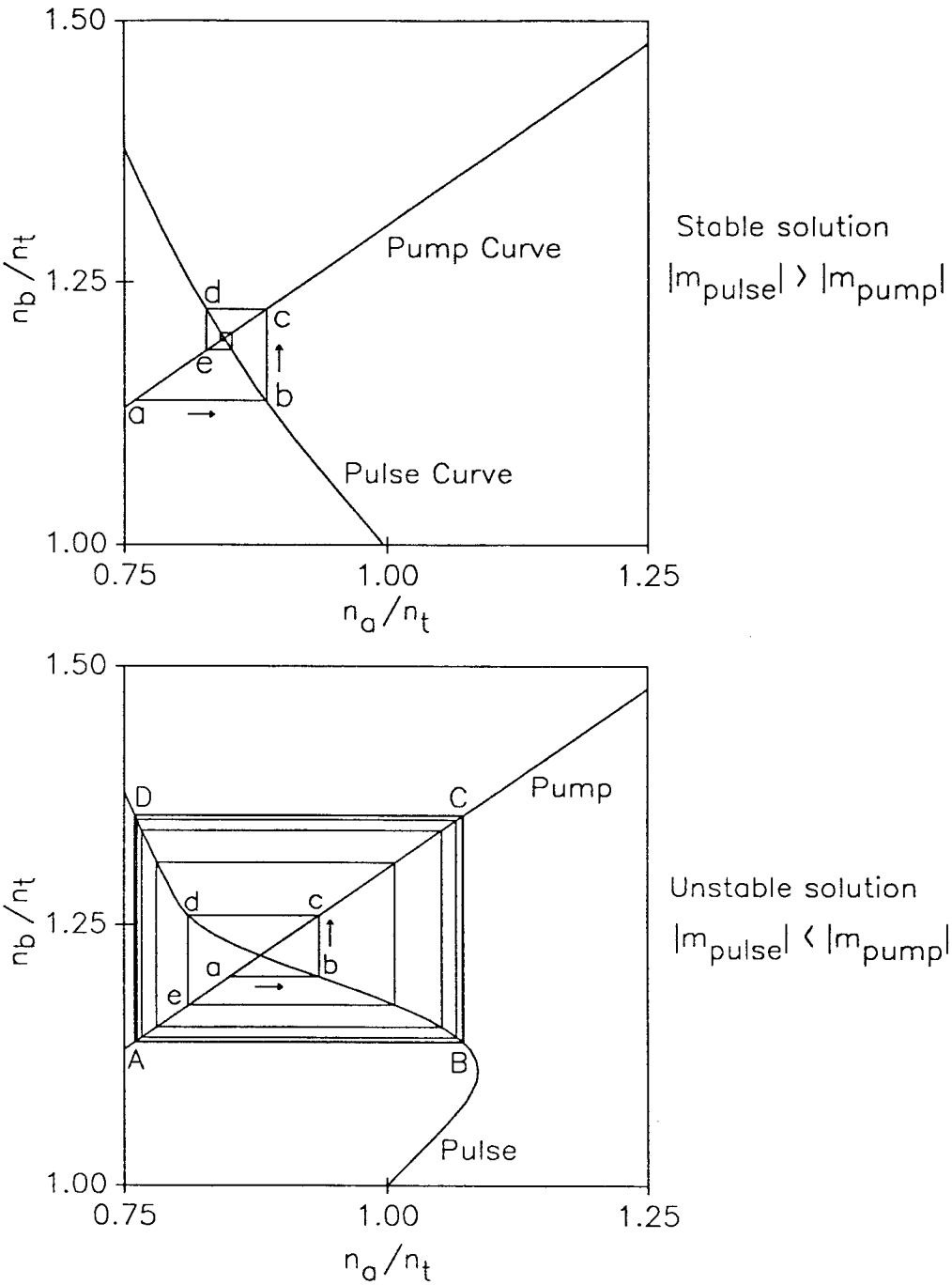


Fig. 2.9 Illustration of the stability of repetitive Q-switching and period doubling. The steady state solution associated with the intersection point is *stable* in the top figure, but *unstable* in the lower figure. In the latter case, pulses are still produced at the modulation frequency, but their energies alternate between two values. The stability condition is that $|m_{\mathrm{Pulse}}| > |m_{\mathrm{Pump}}|$.

the condition for stability is that

$$|m_{\text{Pulse}}| > |m_{\text{Pump}}|, \tag{2.4.11}$$

where the m's are the slopes — dn_b/dn_a — of the pulse and pump curves at their intersection point.

Now referring back to Fig. 2.5, the pump curves all have slopes $m_{\text{Pump}} < 1$, while the unmodified pulse curve has its minimum $|m_{\text{Pulse}}|$ at $n_b = n_a = 1$, where it can be shown that $m_{\text{Pulse}} = -1$. Thus, the Q-switching process itself (i.e., not including the clipping effect of the loss modulator) is stable at all frequencies and would not lead to chaotic pulse trains. The inclusion of the spontaneous emission term and bimolecular recombination processes does not alter this conclusion, but actually alters the curves toward greater stability.

Thus the inclusion of clipping of the pulse by the modulator is required to explain the chaotic pulse train behavior; and this will only occur when the intersection point lies on a limited portion of the upward facing section of the altered curve. At low frequencies, the modification to the pulse curve is small and the pump curve is nearly flat so the intersection lies in a very stable region. As the frequency is increased, the intersection moves down the pulse curve while the clipping "bump" becomes more prominent until the intersection point lies in an unstable region and the period doubled pulse train is expected. As the frequency is further increased, the intersection will move toward the peak of the bump and operation will again be stable. At this point, there is very little depletion of the gain per pulse since n_a is only slightly less than n_b, and the operation is essentially that of a sinusoidally modulated output rather than Q-switching.

As the frequency is increased beyond the onset of period doubling, the point "B" in Fig. 2.10 moves down the pulse curve toward the point of maximum n_a.

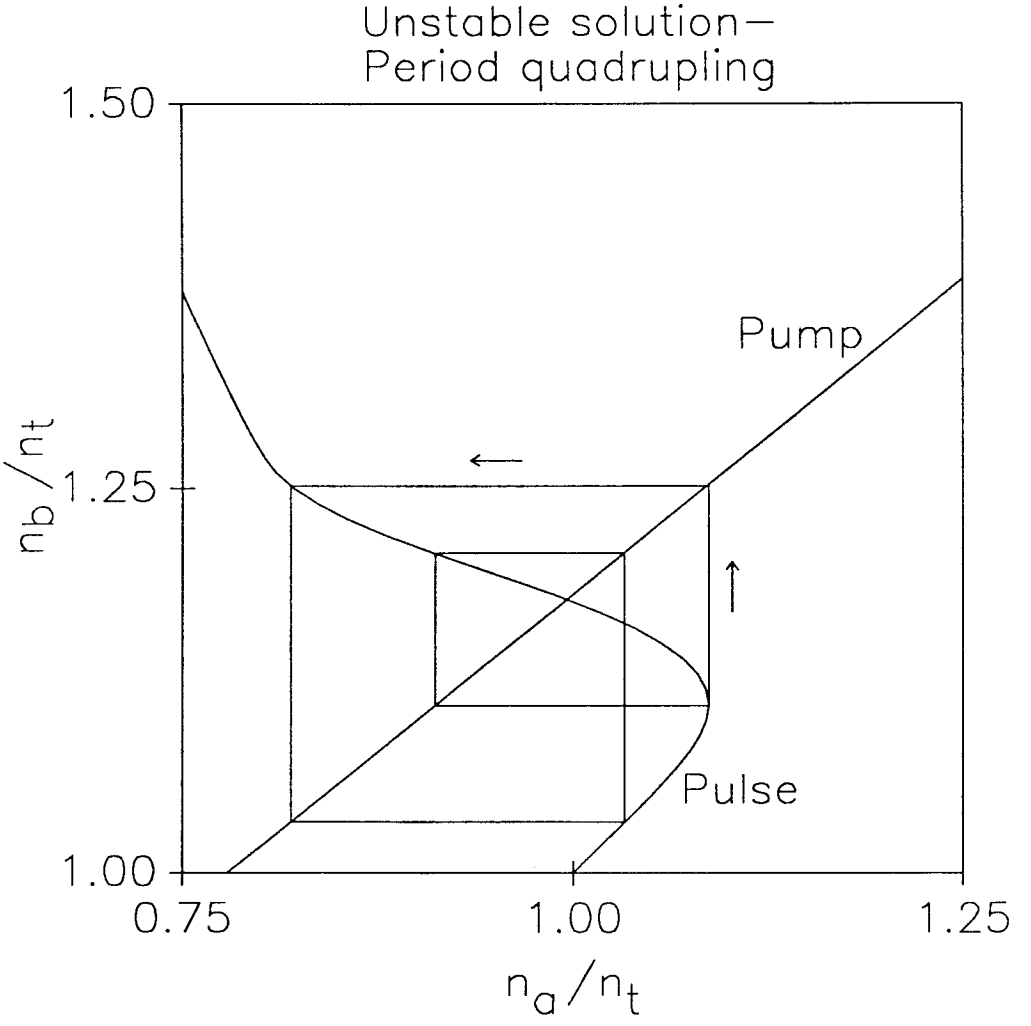


Fig. 2.10 Illustration of steady state solution corresponding to period quadrupling.

When it reaches this point, another type of solution occurs which is depicted in Fig. 2.10, and represents period quadrupling. As the frequency is further increased, the corners of the two rectangles approach each other and period doubling occurs again just before subharmonics disappear altogether. Thus, in addition to explaining the appearance of subharmonics in the intensity spectrum, this simple model predicts that these will be limited to half-integer and quarter-integer values. As this is a consequence of the pulse curve shape and not the actual values, it should continue to hold when a more accurate calculation of this curve is made including pulse delay, spontaneous emission, etc.

The actual frequency at which subharmonics appear will depend on the values of a number of parameters. For comparison with the experimental results of the last section, a pumping level of twice the low loss threshold is a good approximation assuming the low loss threshold is about that at $V_b = +1$ V. The spontaneous lifetime is generally taken to be 1 - 3 ns for bulk material, but a somewhat lower value is usually more appropriate for lasers due to diffusion from the active layer. This is especially true for structures without lateral carrier confinement, as with the ridge waveguides used here. We therefore assume a value of 1 ns which is convenient since the pump curve labels then correspond to frequency in GHz. With these assumptions, Fig. 2.5 indicates that subharmonics should appear somewhere between 2 - 4 GHz, probably close to 3 GHz. This corresponds very well to the observed value of 3.2 GHz especially considering the approximations made along the way.

Finally, we briefly consider how this problem might be eliminated or at least pushed to a higher frequency limit before it appears. It doesn't seem likely that the problem could be eliminated altogether since eventually the loss modulation will interfere with the finite width Q-switched pulses. Reduction of the spontaneous

lifetime would be effective but is undesirable in that it would raise the threshold current to levels where ohmic heating becomes a problem. Probably the best solution is to pump the gain section harder which moves the point at which the pump curves converge further up the line $n_b = n_a$. This is confirmed by the experimental data in Fig. 2.11 which shows an increase in the maximum modulation frequency for regular pulse generation as the pump current to the gain section is increased. According to Fig. 2.5, a 50% increase in the pumping level (from $2n_{th}$ to $3n_{th}$) would be sufficient to push the onset of subharmonics to a modulation frequency of greater than 10 GHz. Unfortunately, the lasers used would not survive this pumping level, but the trend of Fig. 2.11 indicates that this prediction would not be achieved and instead the frequency limit would be around 4.5 GHz. This discrepancy may be due to the simplified use of a linear gain versus carrier density in the calculations. Inclusion of a sublinear gain — which is characteristic of quantum well active layers — would mean that at high pump levels, the peak gain does not continue to increase as much and thus pulse widths are not as short as predicted with a linear gain model. Consequently, the frequency at which the loss modulation starts to interfere with the completion of the Q-switching process would also be reduced. A possible cure for this may be to use more quantum wells to enable higher gains to be achieved before it is saturated. A potential drawback to increased pumping of the gain is that the loss modulation might not be adequate to suppress continuous lasing at lower modulation rates due to the high levels of gain which would be reached. This is not a serious problem though, particularly if operation were to be restricted to a single frequency at a time.

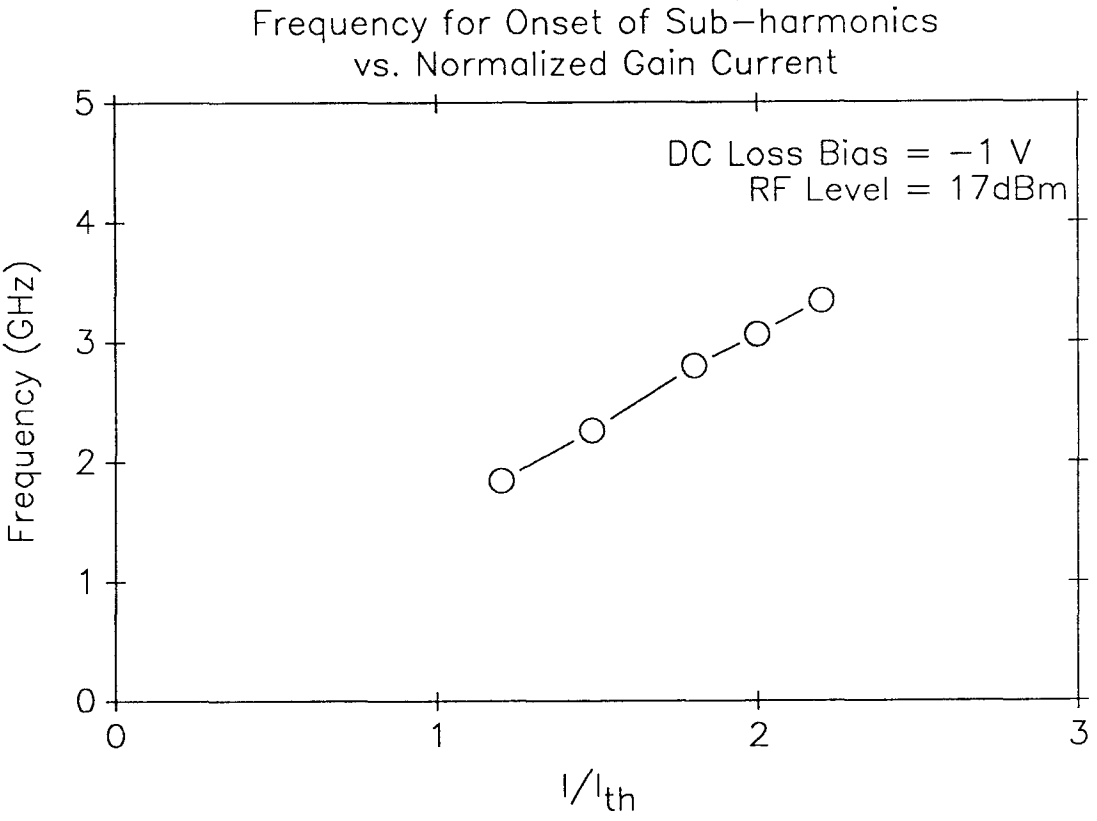


Fig. 2.11 Experimental data showing the frequency at which sub-harmonics of the modulation frequency appear as a function of the current to the gain section. Current is normalized to the threshold current at which any output first occurs while modulating the loss.

References

[1] F. J. McClung and R. W. Hellwarth, J. Appl. Phys.,**33**, 828 (1962).

[2] T. Tsukada and C. L. Tang, IEEE Jour. Quant. Elec. **QE-13**, 37 (1977).

[3] M. Yamanishi, K. Ishii, M. Ameda, and T. Kawamura, Jpn. J. Appl. Phys. Suppl. **17**, 359 (1978).

[4] D. Z. Tsang, J. N. Walpole, S. H. Groves, J. J. Hsieh, and J. P. Donnelly, Appl. Phys. Lett. **38**, 120 (1981).

[5] D. Z. Tsang and J. N. Walpole, IEEE Jour. Quant. Elec. **QE-19**, 145 (1983).

[6] D. Z. Tsang, J. N. Walpole, Z. L. Liau, S. H. Groves, and V. Diadiuk, Appl. Phys. Lett. **45**, 204 (1984).

[7] D. A. B. Miller, D. S. Chemla, T. C. Damen, A. C. Gossard, W. Wiegmann, T. H. Wood, and C. A. Burrus, Phys. Rev. Lett. **53**, 2173 (1984).

[8] J. S. Weiner, D. A. B. Miller, D. S. Chemla, T. C. Damen, C. A. Burrus, T. H. Wood, A. C. Gossard, and W. Wiegmann, Appl. Phys. Lett. **47**, 1148 (1985).

[9] S. Tarucha, H. Kobayashi, Y. Horikoshi, and H. Okamoto, Jpn. J. Appl. Phys. **23**, 874 (1984).

[10] H. A. Pike and M. Hercher, Jour. Appl. Phys. **41**, 4562 (1970).

[11] A. Yariv, *Quantum Electronics, Second Ed.*, Wiley, New York (1975).

Chapter 3

GaAs on Si: High Speed Laser Diodes and p-i-n Photodiodes

3.1 Introduction

The development of molecular beam epitaxy (MBE) has ushered in a new generation of exotic semiconductor devices which rely on its capability to control growth composition with atomic layer precision. Examples include the high electron mobility transistor (HEMT), quantum tunneling devices, low threshold quantum well lasers, and quantum confined Stark effect optical modulators. MBE has also opened up another technology of extreme practical importance with the successful growth of epitaxial layers of GaAs on Si substrates. The most significant aspect of this is the potential for monolithic integration of GaAs and Si devices, thus providing a marriage between the optoelectronic devices of GaAs and the highly developed VLSI processing technology of Si. Practical considerations also motivate the growth of solitary GaAs devices on Si simply due to the superiority of Si as a substrate material–it's lighter, stronger, cheaper, and is available in larger wafer sizes.

The major obstacles to success are a large lattice mismatch and a difference in thermal expansion between the two materials. The lattice constants of Si and GaAs are 5.43 Åand 5.65 Årespectively which can lead to defect densities as high as 10^{13} cm^{-2} at their interface. The mismatch in thermal expansion results in significant stress in the epitaxial layers when the wafer is cooled down from the elevated temperatures used for MBE growth. This often results in visible cracking of the GaAs layers particularly if they are too thick. Techniques involving the use of tilted Si substrates and superlattice buffer layers have significantly improved the

quality of GaAs layers obtainable, though not to the level of that obtained with growth on GaAs substrates. Nevertheless, successful devices have been numerously reported, particularly those which are solely electronic in nature [1]. Optoelectronic devices have been more difficult, presumably because they demand a higher material quality, but cw operation of lasers with moderate thresholds has been accomplished [2]-[4].

For application in high speed communication, an important characteristic of any laser diode is its modulation response. The first section of this chapter describes microwave current modulation measurements of GaAs-on-Si lasers to characterize their applicability to high speed systems.

Among optoelectronic devices, lasers have received the most attention since a Si laser is virtually impossible due to its indirect band structure. Photodetectors on the other hand, can be successfully made with Si and have consequently received less attention. However, high speed considerations favor the use of GaAs primarily due to a much higher (10X) absorption coefficient at GaAs laser wavelengths. This is discussed in section 3 along with a presentation of results on the performance of high speed GaAs-on-Si *pin* photodiodes grown by molecular beam epitaxy.

3.2 High Speed Modulation of GaAs/Si Lasers

The basic MBE growth structure used for these lasers is the same as was used to achieve the first room temperature cw operation of a GaAs/Si laser [5], [6]. In those cw experiments, broad area lasers were used to reduce the effects of current leakage in determining the current density necessary for lasing—a standard measure of quality for laser diodes. For high speed performance, lateral confinement of the lasing mode is desired to enable a high photon density, and thereby a high stimulated recombination rate. This is reflected in the expression for the modulation corner

frequency, f_r, of a laser [7]:

$$f_r = \frac{1}{2\pi}\sqrt{\frac{Ap_0}{\tau_p}} \tag{3.2.1}$$

where A is the differential gain times the group velocity of light, p_0 is the average photon density, and τ_p is the photon lifetime of the cavity. The optimum approach to achieving optical confinement is with an index guided structure which requires a crystal regrowth step. Such regrowth processes have so far not been successful due to the tendency of the Si substrate to completely dissolve in the GaAs melt of a liquid phase epitaxy (LPE) process. Consequently, a ridge waveguide was used for the guiding structure as illustrated in Fig. 3.1. The 10μm ridge is formed by wet chemical etching and a top contact of Cr/Au is evaporated and patterned by a liftoff. Layers of SiO_2 restrict the injection of current to the top of the ridge. The second contact is made to the bottom of the n^+ Si substrate so current must flow through the Si-GaAs interface. This then directly demonstrates the potential for the desired integration. The active layer in these lasers is a single quantum well graded refractive index separate confinement heterostructure (SQW GRINSCH) similar to that used in the first cw GaAs-on-Si lasers.

The lasers were mounted on a commercial H-mount which fits into a special microwave package for high-frequency modulation [8]. Wire bonding to the top contact caused some additional practical problems due to the nature of the GaAs-on-Si layers. First, the devices were more susceptible to damage by the bonding tool action than ordinary GaAs lasers, presumably due to the stress and defects in the GaAs-on-Si material. This was cured by making the bond as far from the laser ridge as possible and reducing the bonding tool pressure somewhat. Both of these were limited in extent since parasitic capacitance considerations prohibit a very large metal bonding pad and the bond pressure must be sufficient to make the bond work at all. In fact the second difficulty was that the bonds did not stick to

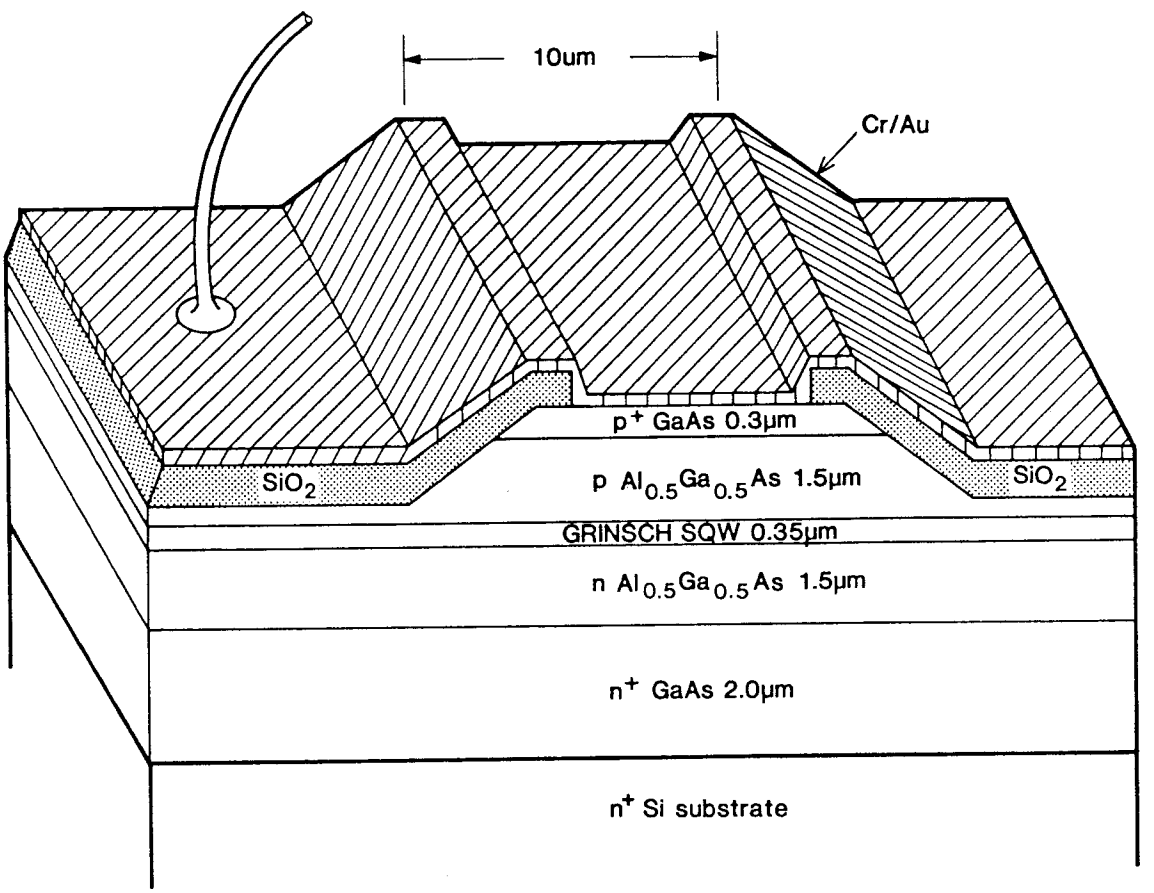


Fig. 3.1 MBE growth structure for GaAs-on-Si lasers. The ridge waveguide provides lateral optical confinement and the active layer uses a single quantum well for low threshold.

the bond pad as well as they usually do with GaAs, even before the bond pressure was reduced. The reason for this is not known for sure but may be related to the rough surface of the GaAs-on-Si and consequently the thin layer gold bond pad also. In the end, yield was not high and success was as much due to persistence as to technique.

After mounting, the high frequency modulation response is measured as diagrammed in Fig. 3.2. A quasi-dc bias and the microwave modulation are applied to the laser through a bias T. Unfortunately, these lasers would not survive a true dc bias and so had to be operated at a much reduced duty cycle. However the pulse bias width of 5 μs is several times the period of the minimum modulation frequency and so can be considered as dc for practical purposes. The laser light output is then detected with a high speed *p-i-n* photodiode and the modulation amplitude is measured with a microwave spectrum analyzer.

While the photodiode used had a rated 6 GHz corner frequency, it was still calibrated for this measurement and all results were normalized with this calibration. The calibration measurement was done by measuring the photodiode's frequency response with a picosecond light source as described in the next section.

Ordinarily, modulation measurements are performed with a network analyzer which automatically normalizes to the input rf signal strength and enables continuous variation of the frequency with full data collection in less than a minute. Due to the necessity of operating at a reduced duty cycle, the signal strength was not high enough for this technique and a spectrum analyzer was used instead. This enabled the use of an extended accumulation of the signal at a fixed modulation frequency. As a result, the measurements were limited to a discrete set of frequencies rather than a continuous sweep.

The measurement results are shown in Fig. 3.3 with a Bode plot of the response

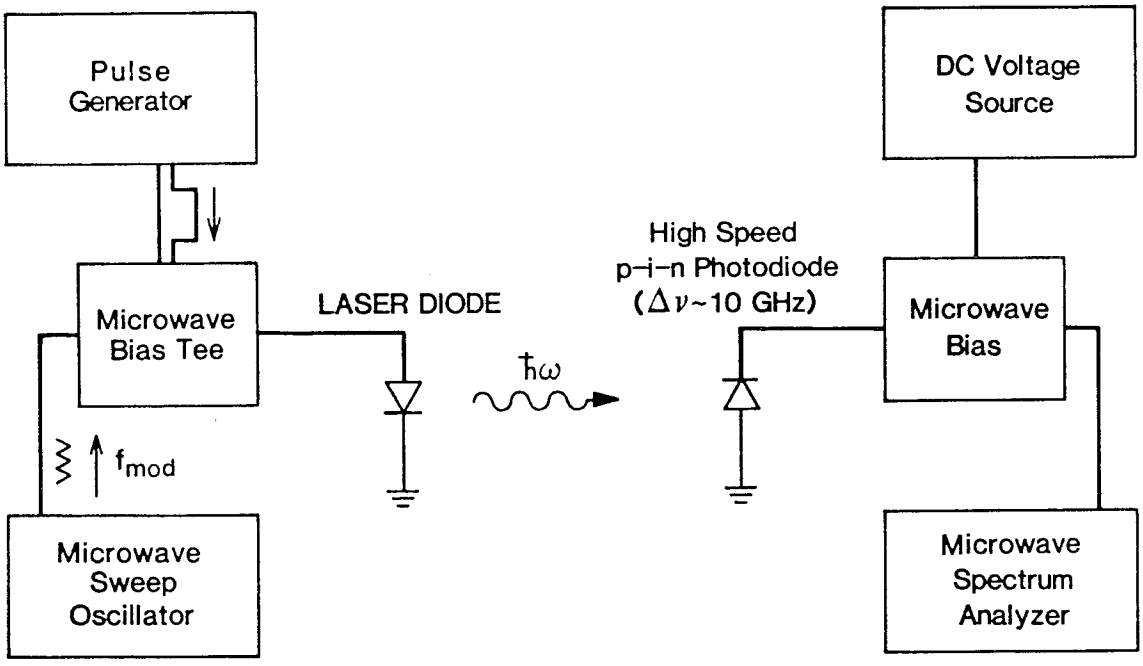


Fig. 3.2 Set-up used for measuring the microwave modulation response of GaAs-on-Si lasers.

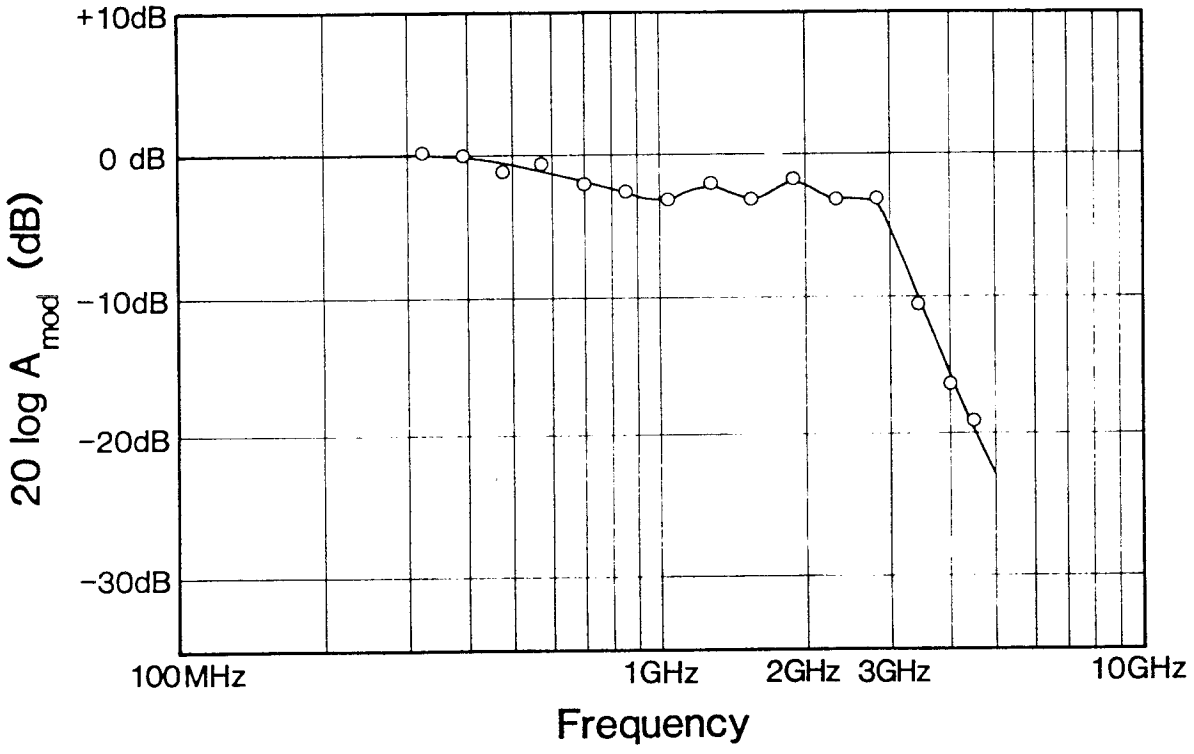


Fig. 3.3 Normalized frequency response of laser under microwave modulation. High frequency roll-off is about 48 dB/decade and the 3 dB corner frequency is between 2 and 3 GHz.

data normalized to 0 dB at low frequencies. The response was measured out to a frequency of 4.5 GHz and a corner frequency of about 2.5 GHz was determined with a high frequency roll-off of 48 dB per decade. Modeling of the laser dynamics predicts a high frequency roll-off of 40 dB/dec. However, higher values of about 48 dB/dec are usually measured and are attributed to parasitic effects [7].

As already mentioned, the ridge waveguide used for these lasers is not ideal for high speed performance. In addition the growth on a conductive substrate leads to excess parasitic capacitance and the fastest lasers are typically grown on semi-insulating substrates. Thus while GaAs lasers with corner frequencies >12 GHz have been reported, those with structures similar to that used here are typically limited to 2-3 GHz [7]. It seems then that the GaAs-on-Si lasers are limited more by restrictions on the structure than by deficiencies in the material quality.

3.3 High Speed GaAs-on-Si p-i-n Photodiodes

At GaAs laser wavelengths ($\sim 0.85\mu m$), Si is absorptive ($\lambda_{gap} \simeq 1.1\mu m$) and can be used directly as a detector medium without the need for GaAs-on-Si growth. However, GaAs does offer advantages (e.g., high mobilities and absorption coefficient, and a large band gap energy) which make it a superior material for high speed optical detectors. In particular, the large difference in absorption depths for light at GaAs laser wavelengths ($\simeq 10\mu m$ for Si versus $\simeq 1\mu m$ for GaAs) [9] has direct consequences in the potential gain-bandwidth products for photodiodes. Since gigahertz response usually requires carrier transit regions of only a few μm, a fast Si photodiode will be less sensitive than a similar GaAs photodiode, as well as being much more prone to diffusion tail effects which can seriously degrade the frequency response. In this last regard, the high mobility of GaAs (hence high diffusion constants) and the possibility of band gap tailoring with AlGaAs also favor GaAs with techniques for reducing these diffusion tail effects. Furthermore, at wavelengths

greater than 1.1μm (Si is transparent), low band gap ternary and quaternary III-V compounds grown on Si become very important for integrated photodetectors.

So far, reports of GaAs-on-Si photodetectors have been rather limited. High speed GaAs-on-Si photoconductors have been demonstrated with an impulse response of 60 ps using a short (4μm) contact spacing [10]. Avalanche photodiodes, with a *p-i-n* structure similar to that used here, have been investigated regarding dc operation with only a moderate gain (7×) being achieved due to large leakage currents at high reverse bias [11]. For the fiber optic communications wavelengths at 1.3 μm and 1.55 μm, GaInAs photodiodes grown by metalorganic chemical vapor deposition on Si substrates have also been reported [12]. However, the anticipated high speed performance of *p-i-n* photodiodes remains to be verified for any III-V compound grown on Si substrates.

The detector structure used was a conventional *p-i-n* layer sequence with a mesa-defined active area as shown in Fig. 3.4. The growth of the GaAs/Si interface (the most critical point) is the same as was used previously for lasers and is described in detail elsewhere [2],[3]. After a specialized growth start procedure to establish good quality GaAs at the GaAs/Si interface, the growth sequence proceeds with 2 μm n^+ GaAs, an undoped GaAs layer of thickness d, and a 0.1 μm p^+ GaAs contact layer. The top contact layer is kept very thin to reduce potential diffusion tails by reducing both the diffusive decay time as well as the number of carriers available for such a tail.

Mesas (70 × 100μm^2) were defined by photolithography and etched with a solution of H_2O_2:H_3PO_4:H_2O(1:3:40) to about 1 μm below the intrinsic layer. Bond pads (50μm Dia.) of Cr/Au were then applied to the mesas by liftoff and a back contact of AuGe/Au was made to the substrate. In order to demonstrate the capability for direct integration to Si circuits, all devices were fabricated with one

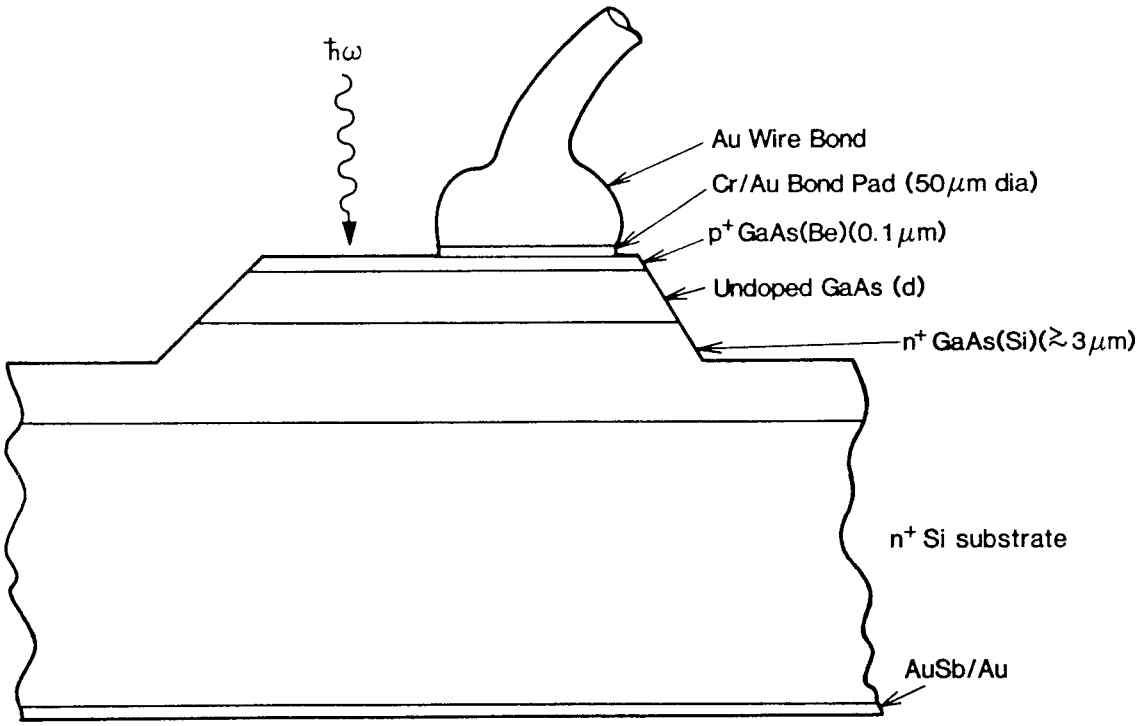


Fig. 3.4 Diagram of the GaAs/Si *p-i-n* photodiode structure. Mesa area is $70\times100\ \mu m^2$ and three different undoped layer thicknesses were used (d =1, 2, and 3 mum).

contact on the back of the Si substrate as opposed to a second topside contact to the GaAs itself which would allow the potentially troublesome GaAs-Si interface to be bypassed. In order to gauge the performance of the material and not the structure itself, similar devices were fabricated with GaAs grown on a GaAs substrate for comparison. For convenience, we will refer to GaAs-on-Si-substrate and GaAs-on-GaAs-substrate as just GaAs/Si and GaAs/GaAs respectively, throughout the rest of this chapter. In all, three GaAs/Si structures (d=1μm, 2μm, and 3μm) and two GaAs/GaAs structures (d=1μm and 3μm) were used in this investigation.

Current-voltage characteristics for both a GaAs/Si and GaAs/GaAs diode (d=1μm) are shown in Fig. 3.5. As can be seen, the GaAs/Si diode has a considerably lower voltage and "softer" reverse breakdown than the GaAs/GaAs device. These large reverse leakage currents are a typical problem in GaAs/Si diodes[11] and have been attributed to defect assisted tunneling or conduction through metallic precipitates situated around dislocations [13]. However, the GaAs/Si diode can still be safely biased at the levels typical for high speed operation, namely a few volts. The forward characteristic indicates that series resistance is limited to less than 10Ω which is not appreciable considering the detector typically drives a 50Ω transmission line. The GaAs/GaAs detectors required an alloying of the contacts (30 sec at 300 °C) in order to achieve a satisfactory forward current-voltage characteristic with low resistance. Conversely, the GaAs/Si detectors showed good forward characteristics without any alloying; furthermore, any attempt to alloy them resulted in a severe degradation of the reverse bias leakage and consequently, none of the GaAs/Si detector contacts were alloyed.

For high speed measurements, the detectors were mounted in a microwave 50Ω photodiode package [14]. The mounting process involved ultrasonic bonding directly on the mesa which apparently degraded the reverse breakdown characteristic of the

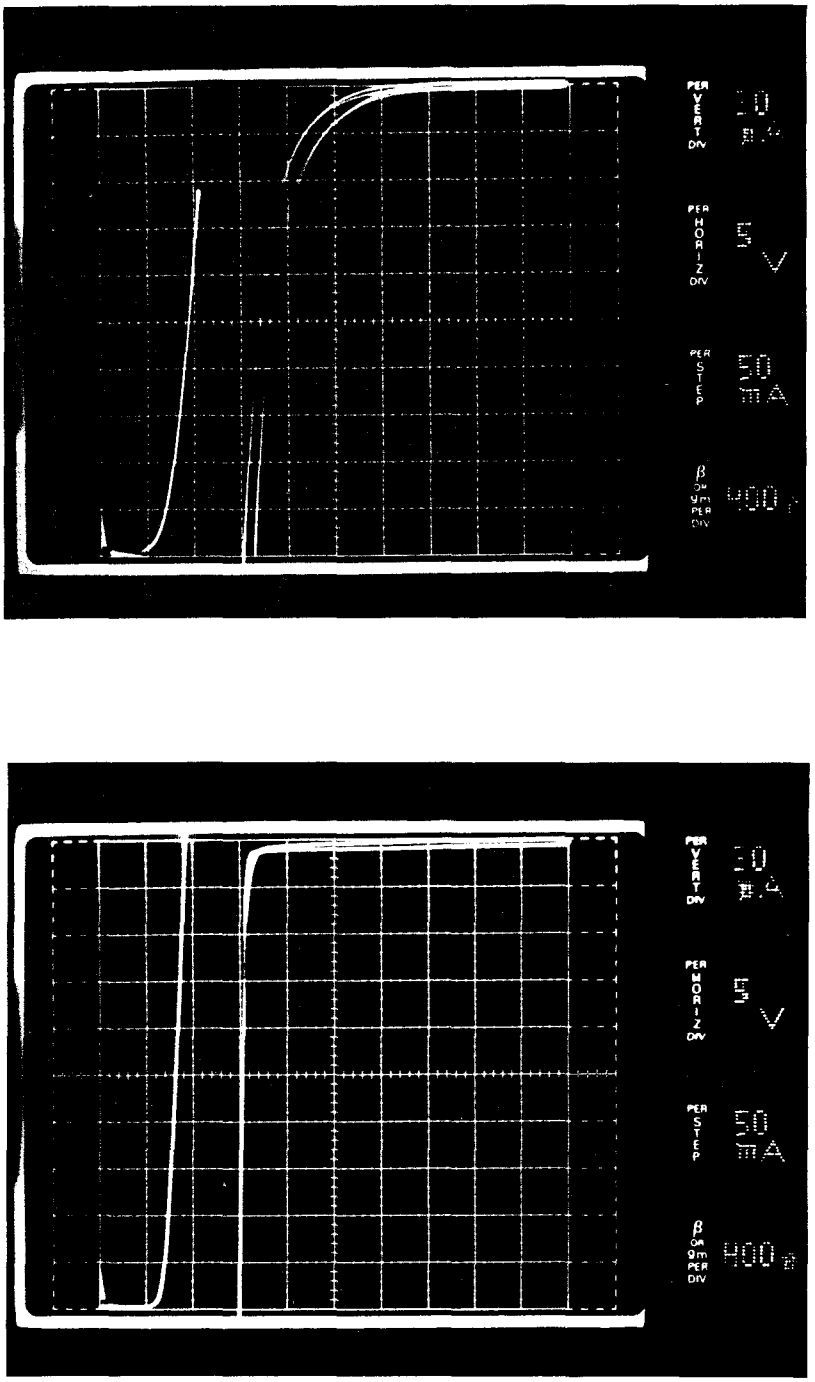


Fig. 3.5 Current-voltage characteristics for (top) GaAs/Si photodiode ($d = 1$ μm) and (bottom) GaAs/GaAs photodiode ($d = 1$ μm). In each figure, the left curve is the forward bias with scales of 1 V/div (horiz.) and 10 mA/div (vert.), and the right curve is reverse bias with scales of 5 V/div and 20 μA/div.

GaAs/Si diodes, but leakage currents at the operating voltages were still low. For the GaAs/Si detector described below, reverse leakage current was measured to be 70nA at an applied bias of -3V.

The capacitance of the mounted detector was also measured to evaluate the RC limit (as opposed to transit time effects) to the high speed performance. Assuming a depleted I-region of 2 μm and a mesa area of $70 \times 100 \mu m^2$, the capacitance should be 0.4 pf. At zero bias, capacitance was measured to be 0.56 pf which is significantly more than the estimated value. However, this quickly dropped to 0.37 pf with an applied reverse bias of 2 V indicating a low level of residual doping in the "intrinsic" region which must be depleted by such a reverse bias for high speed operation. Based on this value and a load resistance of 50Ω, the RC time constant of this detector should be $\simeq$20 ps. Assuming a saturated carrier velocity of 1×10^7cm/s, the transit time is also estimated to be 20 ps.

For high speed measurements, the mounted detector was biased through a microwave bias tee and illuminated with 5 ps optical pulses from a synchronously pumped modelocked dye laser (λ = 600nm) at a 100 MHz repetition rate. The output of the detector was then measured both with a sampling oscilloscope and a microwave spectrum analyzer. Fig. 3.6 shows the measured impulse response of both a GaAs/Si(d=2μm) and a GaAs/GaAs(d=3μm) detector at their optimum bias voltages. Both exhibit a full width at half maximum (FWHM) of approximately 45 ps, and are almost indistinguishable except for a somewhat longer tail in the GaAs/Si response. This is suggestive of a diffusion tail which would be more prominent with a thinner *i*-region diode. However, at the wavelength used here, the short absorption depth ($\simeq$0.3 μm) should preclude such tails altogether in either a 2 μm or 3 μm depletion layer photodiode. While the source of this tail is not known, its small relative size indicates it is not a significant concern for high

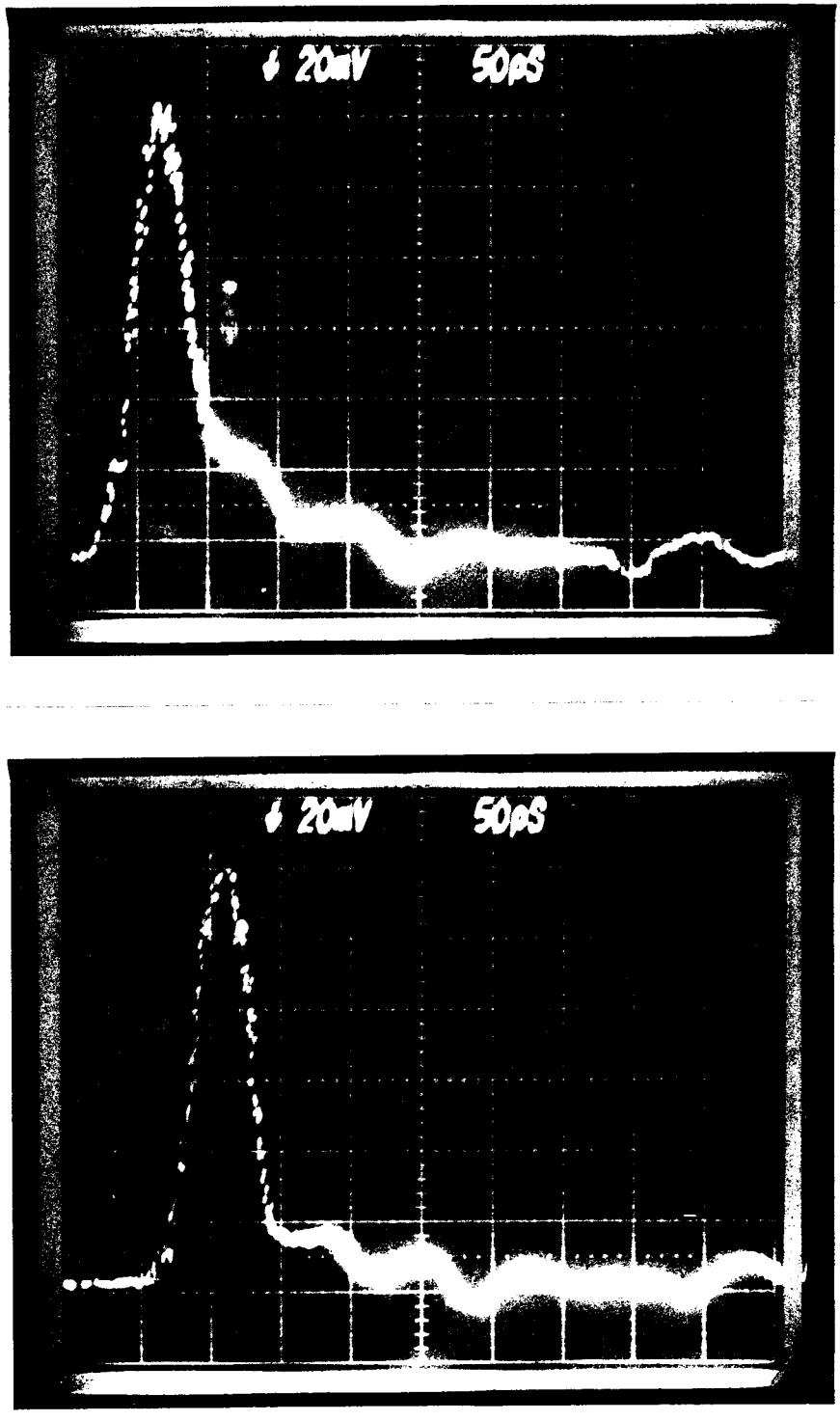


Fig. 3.6 Impulse response measurements of (top) GaAs/Si photodiode ($d = 2\mu$m) and (bottom) GaAs/GaAs photodiode ($d = 3\mu$m). The time scale is 50 ps/div. Both have a pulse width (FWHM) of $\simeq$45 ps and are quite similar other than a slightly more evident tail on the GaAs/Si response.

speed applications. At this short a pulse width, the rise time of the sampling head ($\simeq$25 ps) is affecting the measurement and the actual pulse width is somewhat less. Using a sum of the squares rule for deconvolving, the actual pulse width would be estimated at 37 ps (i.e. $\sqrt{45^2 - 25^2}$) although a more accurate determination of the oscilloscope sampling function is necessary for a reliable deconvolution.

This impulse response signal was also measured with a microwave spectrum analyzer (HP8565A) to display the frequency response of the detector which is shown in Fig. 3.7. Except for a slight dip at 3 GHz, the response is flat (within 3dB) out to >4 GHz. The same 3 GHz dip was observed with several other detectors having much different impulse responses (including a commercial photodiode with a rated corner frequency of 7 GHz) and is thus believed to be an artifact of the measurement set-up.

In summary, we have reported on the fabrication and measurement of GaAs *p-i-n* photodiodes grown on Si substrates by MBE. This includes the first high speed measurements of such devices. The detectors display a pulse width of 45ps (FWHM) and a -3dB corner frequency of >4 GHz. Within the resolution of these measurements, the high speed performance was virtually indistinguishable from that of a similar GaAs photodiode grown on a GaAs substrate. It is expected that the high frequency response can be improved upon simply by reduction of the diode dimensions.

On the negative side, although good devices were achieved, yield was not particularly high. Furthermore, several growths showed considerably slower response with pulse widths of 120 ps (FWHM). Also, reverse bias breakdown voltages were often less than 10 V and even as low as 2-3 V. Finally, there may be lifetime problems due to the stress and defects in the epitaxial GaAs. This was indicated by

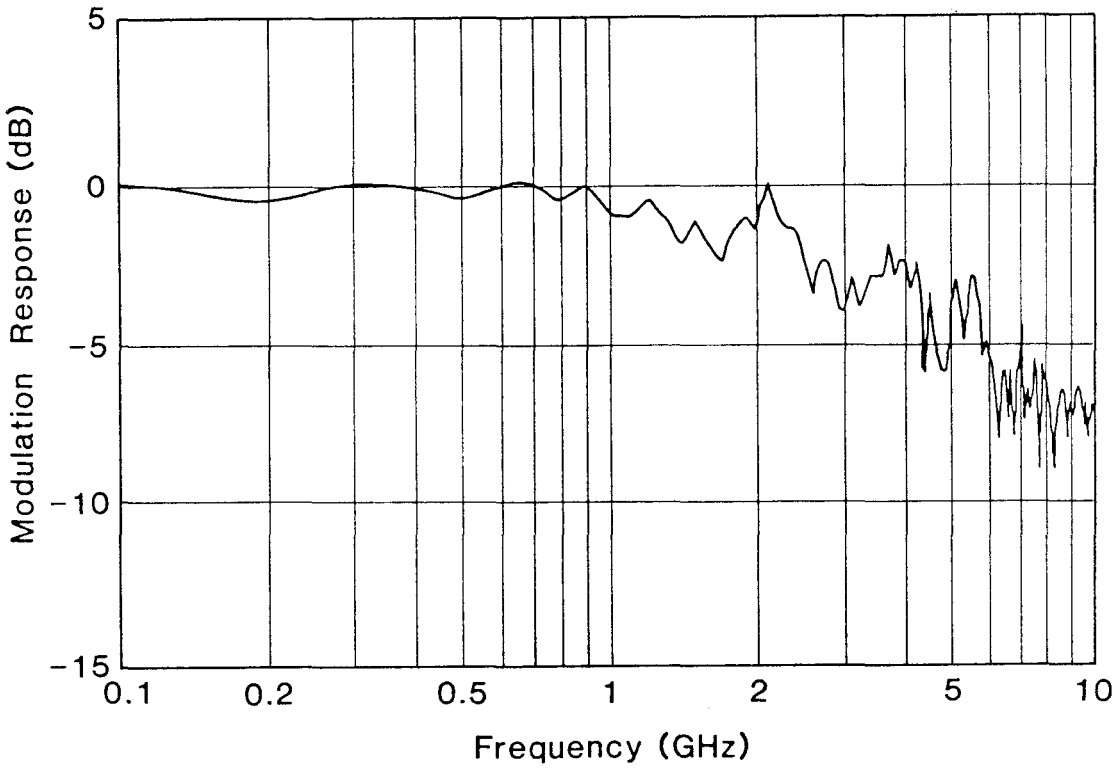


Fig. 3.7 Measured frequency response of a GaAs/Si photodiode ($d = 2$ μm). Actual data points are at multiples of 100 MHz corresponding to the repetition rate of the modelocked laser source.

difficulty in obtaining fast detectors from the same "good" wafer six months after the initial results.

References

[1] R. Houdré and H. Morkoç, to appear in CRC Critical Review.

[2] H. Z. Chen, A. Ghaffari, H. Wang, H. Morkoç and A. Yariv, Appl. Phys. Lett. **51**, 1320 (1987).

[3] H. Z. Chen, A. Ghaffari, H. Wang, H. Morkoç and A. Yariv, Opt. Lett. **12**, 812 (1987).

[4] H. Z. Chen, J. Paslaski, A. Yariv and H. Morkoç, Appl. Phys. Lett. **52**, 605 (1988).

[5] H. Z. Chen, A. Ghaffari, H. Wang, H. Morkoç and A. Yariv, Appl. Phys. Lett. **51**, 1320 (1987).

[6] H. Z. Chen, A. Ghaffari, H. Wang, H. Morkoç and A. Yariv, Opt. Lett. **12**, 812 (1987).

[7] K. Lau and A. Yariv, in *Semiconductors and Semimetals*, edited by W. T. Tsang (Academic, Orlando, FL, 1985), Vol. 22.

[8] Ortel Corp., Alhambra, CA.

[9] S. M. Sze, *Physics of Semiconductor Devices*, (Wiley, New York, 1981).

[10] G. W. Turner, G. M. Metze, V. Diadiuk, B-Y. Tsaur, and H. Q. Le, IEDM *Tech. Dig.* 1985, pp. 468-70.

[11] N. Chand, J. Allam, J. M. Gibson, F. Capasso, F. Beltram, A. T. Macrander, A. L. Hutchinson, L. C. Hopkins, C. G. Bethea, B. F. Levine, and A. Y. Cho, J. Vac. Sci. Technol. **B5**, 822 (1987).

[12] P. D. Hodson, R. R. Bradley, J. R. Riffat, T. B. Joyce, R. H. Wallis, Electron. Lett. **23**, 1094 (1987).

[13] N. Chand, R. Fischer, A. M. Sergent, D. V. Lang, S. J. Pearton, and A. Y. Cho, Appl. Phys. Lett. **51**, 1013 (1987).

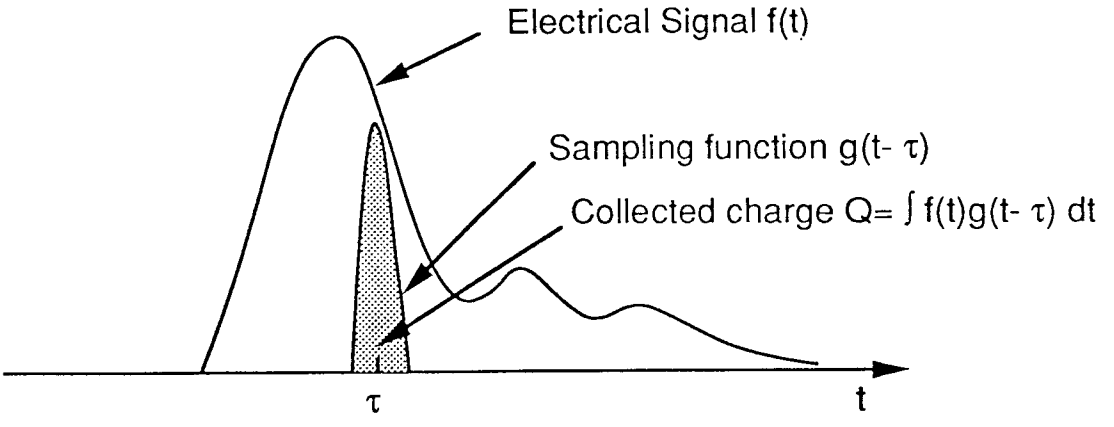
[14] Ortel Corporation, Alhambra, CA.

Chapter 4
Differential Sampling: Analysis

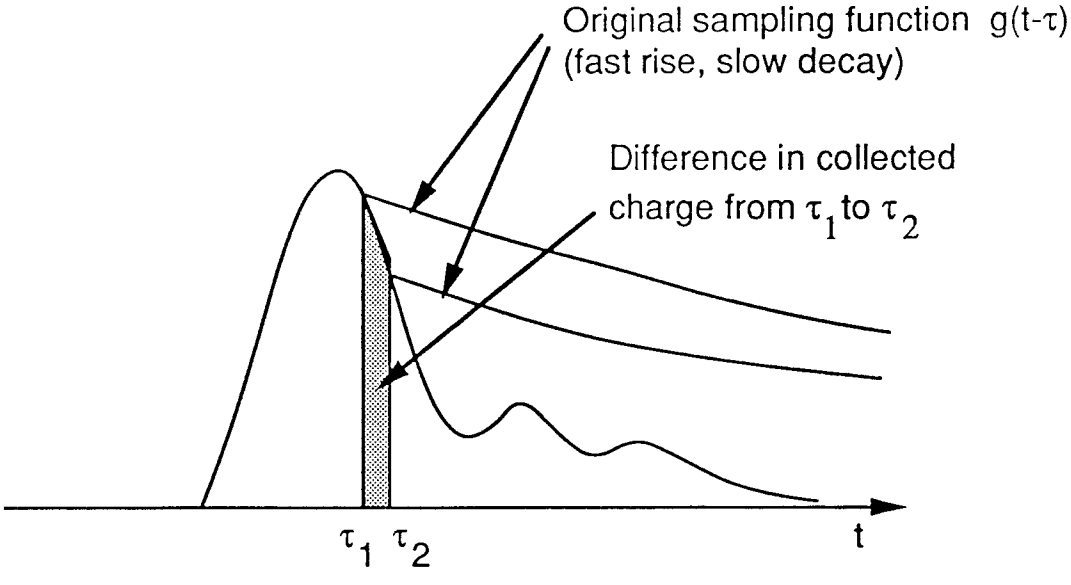
4.1 Introduction

For years, the high speed measurement needs of the electronics community have been served by commercial sampling oscilloscopes which offer resolutions of about 35 ps. Over the last decade, the development of high-speed electronics and optoelectronics has outgrown this capability as device response times have dropped to tens or even a few picoseconds. These advances have consequently necessitated the complementary development of a measurement technology to characterize them. Two very successful techniques which have been developed are electro-optic sampling and photoconductive sampling. Both of these take advantage of recent modelocked laser technology which offers optical pulses of typically a few ps down to < 10 fs. In addition, the use of optical triggering enables an unprecedented reduction in jitter which is crucial to sampling techniques due to the repetitive nature of such measurements. While the best temporal resolution has been achieved with the electro-optic technique, photoconductive sampling offers better sensitivity as well as being more adaptable to a variety of material systems.

In this chapter a new approach to photoconductive sampling is presented called differential sampling. The basic idea is illustrated by a comparison with the conventional approach, as shown schematically in Fig. 4.1. The conventional technique most often uses an optically activated switch to sample the voltage on a transmission line during the time the switch is closed. Consequently, the resolution is limited by how fast the switch can be closed *and* opened. While some of the first work used optical pulses to turn the switch both on and off [1], more recent work has typically relied on modifying material properties to enhance recombination of



Conventional Sampling



Differential Sampling

Fig. 4.1 Schematic comparison of conventional sampling approach (top) and differential sampling (bottom).

photogenerated carriers and thus turn off the conductivity of the switch faster. These include surface recombination [2], doping with deep level impurities [3], [4], [5], ion bombardment [6]-[11], and the use of amorphous material [12], [13]. The ion bombardment approach has been the most successful with typical lifetimes of 2-10 ps and even subpicosecond for the special case of heavily bombarded silicon-on-sapphire [14].

The differential scheme uses the difference between two conventional sampling measurements as its result, which gives the charge collected between the times (τ_1, τ_2) when the two switches are turned on. Some matching of the turn-off of the switches is required; but for good resolution, only a fast turn-on of the switch is now needed. The ultimate limits in resolution and sensitivity will be shown to be comparable to results obtained by ion bombardment. Furthermore, since the resolution of this new approach is independent of carrier lifetime, many of the drawbacks of bombardment can be avoided. Finally, a trade-off between resolution and sensitivity, which is not possible with a fixed sampling window, is easily implemented with this scheme.

4.2 Scheme and resolution

The uncertainties in the heuristic picture of the differential approach (i.e., actual resolution, effect of unmatched turn-off decays, etc.) are resolved by describing the scheme in terms of a new effective sampling function. We start with a quantitative description of the conventional sampling with a fast photoconductor [15]. A typical configuration using microstrip transmission line is shown in Fig. 4.2 along with the equivalent circuit. The usual model for a photoconductor is simply a time varying conductance in parallel with a parasitic capacitance [15]. The variation of the conductivity follows that of the carrier number in the photoconductor gap and will be denoted as $G(t)$. The light source is assumed to be a train of ultrashort

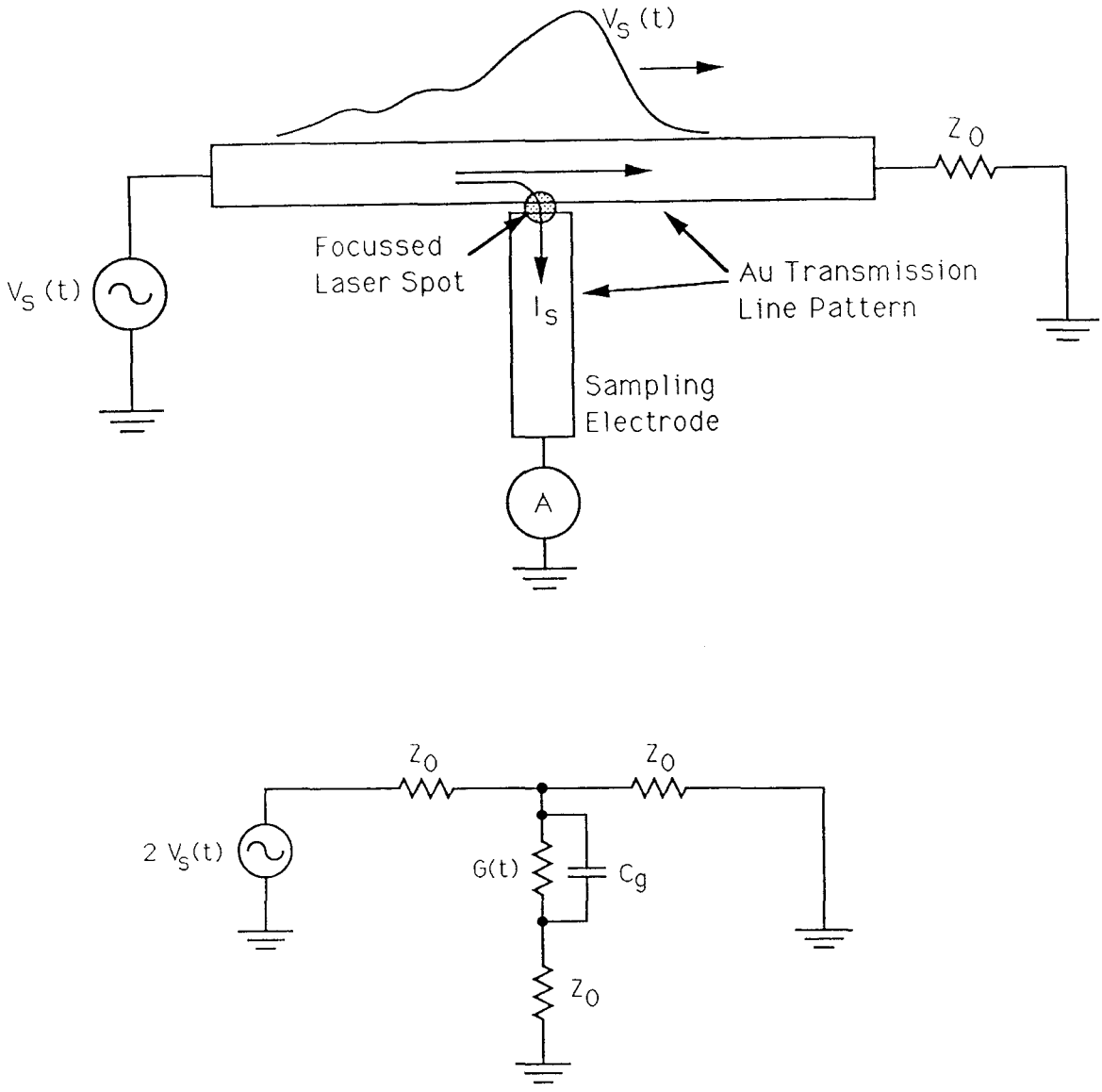


Fig. 4.2 (Top) Typical microstrip transmission line circuit for photoconductive sampling. The signal is represented by the voltage source V_s. The substrate material is a high resistivity semiconductor. Carriers created by the focussed light spot allow current, I_s, to flow to the sampling electrode until recombination turns off this switch.
(Bottom) Circuit equivalent of the photoconductive switch including the gap capacitance C_g. G(t) represents the time varying conductance of the photoconductive gap due to illumination.

pulses which instantaneously create free carriers when absorbed at the photoconductor gap, and $G(t)$ reflects the decay of these carriers. The measured signal is the total charge per pulse, Q_{meas}, — or dc current, $I_{meas} = Q_{meas}/T_{rep}$, for a repetitive source — flowing onto the sampling electrode. This is simply given by

$$\begin{aligned} Q_{meas}(\tau) = V_g \circ G &\equiv \int_{-\infty}^{\infty} dt\, V_g(t) G(t-\tau) \\ &= \int_{-\infty}^{\infty} dt\, V_g(t+\tau) G(t), \end{aligned} \tag{4.2.1}$$

where τ is a variable delay between the signal and the photoconductor turn-on, and $V_g(t)$ is the voltage across the photoconductor gap which is given in terms of the original signal, $V_s(t)$, as

$$V_g(t) \equiv V_s * f_{RC} = \int_{-\infty}^{\infty} dt'\, V_s(t') f_{RC}(t-t'). \tag{4.2.2}$$

The operations of correlation and convolution are denoted by $\circ$ and $*$ respectively. The function f_{RC} is just the transient associated with charging the gap capacitance through the transmission line impedance. Combining these two equations and changing the order of integration gives

$$Q_{meas}(\tau) = \int_{-\infty}^{\infty} dt'\, V_s(t') \int_{-\infty}^{\infty} dt\, G(t+t'-\tau) f_{RC}(t), \tag{4.2.3}$$

which is interpreted as the sampling of V_s by the function

$$f_{samp}(t') \equiv \int_{-\infty}^{\infty} dt\, G(t+t') f_{RC}(t) = G \circ f_{RC}. \tag{4.2.4}$$

The conventional approach to sampling relies on f_{samp} being sufficiently short temporally to be considered a delta function, in which case Q_{meas} is a direct representation of the signal V_s. The sampling function, $G \circ f_{RC}$, typically has a fast, circuit limited rise time (~few ps) and a much slower decay time (~100 ps - 3 ns) reflecting the carrier recombination for intrinsic semiconductors. Consequently, the

carrier lifetime limits the resoluton and much effort has been spent in finding ways to reduce it.

As an alternative we reconsider the use of a photoconductor with a very long carrier recombination time. Since f_{samp} still has a fast rise time, we expect the sampling result $f_{meas}(\tau)$ can be approximated by the integral of $V_s(t)$ from τ to infinity and a derivative operation should recover V_s. Differentiating Eqn. (4.2.3) and using (4.2.4) yields:

$$\frac{dQ_{meas}(\tau)}{d\tau} = \int_{-\infty}^{\infty} dt\, V_s(t) \left[\frac{df_{samp}(t-\tau)}{d\tau}\right] \tag{4.2.5}$$

which is just equivalent to sampling with the function $-df_{samp}(t)/dt$. In practice, a finite difference is often used which leads to:

$$Q_{meas}(\tau) - \alpha Q_{meas}(\tau + \Delta\tau) = \int_{-\infty}^{\infty} dt\, V_s(t)[f_{samp}(t-\tau) - \alpha f_{samp}(t-(\tau+\Delta\tau))] \tag{4.2.6}$$

where α is a constant to be discussed below. The result is then equivalent to sampling with a new effective sampling function defined as:

$$f_{\text{eff}}(t) = f_{samp}(t) - \alpha f_{samp}(t - \Delta\tau). \tag{4.2.7}$$

Assuming f_{samp} has a fast rise time and a slow decay, then for short $\Delta\tau$, f_{eff} will consist of a sharp "spike" followed by a long negative tail of equal-area (if $\alpha = 1$). The initial "spike" is the desired sampling feature and in the limit of very short $\Delta\tau$, the ultimate temporal resolution will be limited by the details of the leading edge of f_{samp}. The negative tail on the other hand will lend a *long-time* limit to the resolution since signals much longer than the tail will generate no net sampled signal. It is helpful to consider that for an exponential decay of f_{samp}, this tail is equivalent to high-pass filtering or ac coupling the signal with a single pole filter prior to a true measurement. While this effect may be undesirable, it is not too serious since

for decay times of ~100 ps, commercial instruments could be used to measure the long time features. It is also somewhat surprising that the reduction of this effect now favors a longer recombination time although this will adversely affect the noise performance. Finally, for the special case where the decay of f_{samp} is a single exponential with time constant τ_{rec}, the negative tail can be eliminated altogether by setting the parameter α in Eqn. (4.2.6) to $\alpha = \exp(-\Delta\tau/\tau_{rec})$. The function f_{eff} for this case is plotted in Fig. (4.3) for various values of the delay difference $\Delta\tau$. The original function, f_{samp}, used is the correlation of two exponentials—a gap charging time of 2 ps and carrier lifetime of 150 ps were chosen as representative of actual InP:Fe photoconductors presented in the next chapter. It should be noted that if the optical pulse width can not be neglected, then $G(t)$ and consequently the sampling functions f_{samp} and f_{eff} should be convolved with the optical pulse shape.

As can be seen, the sampling function is virtually independent of the recombination time and can be chosen with an arbitrary width down to the circuit rise time limit of 2 ps. The validity of this value may be questionable due to deficiencies in the simple electrical model of the photoconductive gap as a lumped element capacitance. Ground-plane reflections when using microstrip are well documented [3],[4] in high speed correlation measurements of fast photoconductors and have typical round trip times of 6-8 ps. Furthermore, calculations of gap capacitance are based on dc field distributions. This is only valid for use in a transmission line circuit if the gap is normal to the direction of wave propagation. In the case of a "side tapped" configuration, exact calculation becomes a complicated problem requiring knowledge of higher order and radiation modes of the transmission line for proper matching of fields. Qualitatively, however, one would expect artifacts corresponding to the transit time of signals across the width of the sampling transmission line.

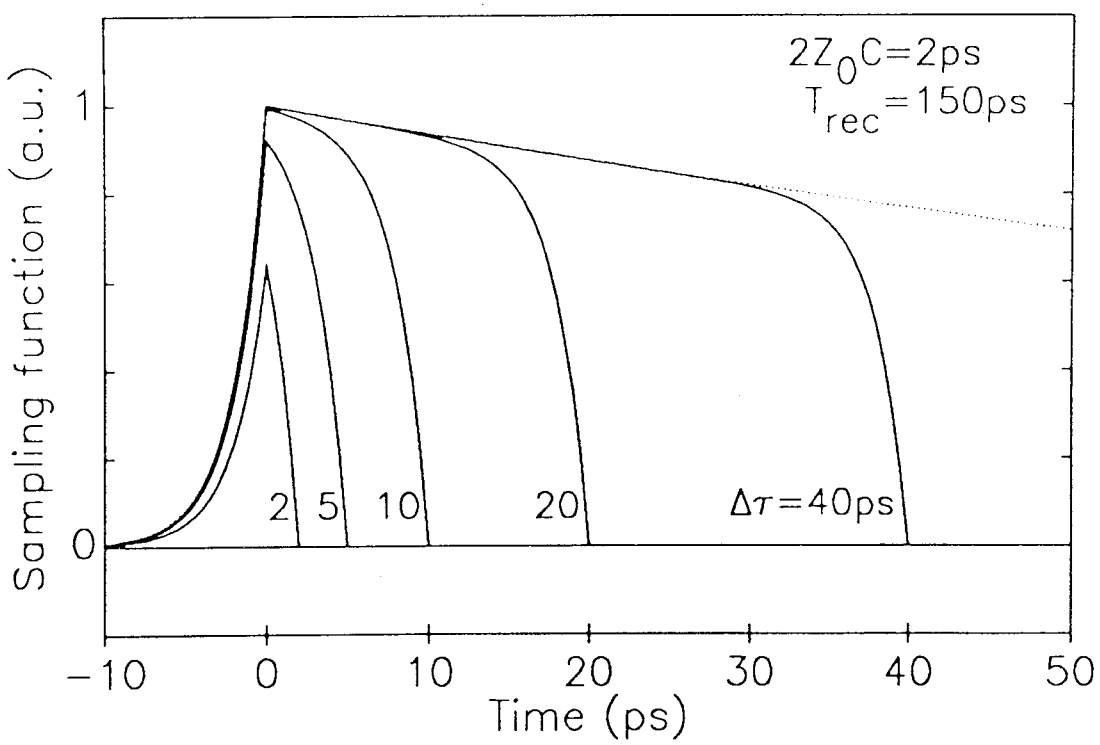


Fig. 4.3 Effective sampling functions for various values of the turn-on delay $\Delta\tau$. The gap charging and carrier decay times for a single photoconductor are 2 ps and 150 ps respectively.

Again, for typical 50 Ω microstrip lines on 300 μm substrates, this round trip time is about 6 ps. Although reduction of the transmission line dimensions would give some improvement, it is practically limited by difficulties in handling semiconductor substrates which are lapped extremely thin. A much better approach would be the use of coplanar waveguides. In addition to small circuit dimensions — limited by lithography — this geometry also offers the capacitance-free "sliding contact" scheme [16]. In essence, the capacitance charging limitation changes to a transit time of the electrical signal across the width of the optical spot size which is shorting the transmission line. This time can quite easily be 100 fs and so subpicosecond sampling seems feasible with this geometry.

The derivation leading to the effective sampling function suggests that the differential sampling scheme is similar to a deconvolution. This is easily checked in the frequency domain by Fourier transforming the left side of Eqn. (4.2.6):

$$\begin{aligned}\int_{-\infty}^{\infty} d\tau\, e^{-j\omega\tau}[Q_{meas}(\tau) - \alpha Q_{meas}(\tau+\Delta\tau)] \\ &= \tilde{Q}_{meas}(\omega)[1 - exp((j\omega - 1/\tau_{rec})\Delta\tau)] \\ &\simeq \tilde{Q}_{meas}(\omega)[1 - (1 + (j\omega - 1/\tau_{rec})\Delta\tau + \ldots)] \\ &\simeq \tilde{Q}_{meas}(\omega)[(-j\omega + 1/\tau_{rec})\Delta\tau + \ldots] \qquad (4.2.8)\end{aligned}$$

where the value $\alpha = \exp(-\Delta\tau/\tau_{rec})$ was used. Thus at low frequencies ($< 1/\Delta\tau$) the differential operation is equivalent to removing a pole at $\omega = -j/\tau_{rec}$. This is the conjugate of the pole expected for a deconvolution and the difference operation is instead a decorrelation which is what it should be. At higher frequencies, higher order Taylor terms are significant and the decorrelation is not perfect. This is not surprising though since the effective sampling functions of Fig. 4.3 are not completely devoid of the longtime decay behavior. Finally, this picture suggests that noise may be a problem since spectral features which lie below the noise level

can not be retrieved by multiplication with the factor $(-j\omega + 1/\tau_{rec})$. This issue is considered in more detail in the next section.

4.3 Sensitivity and noise

In addition to the temporal resolution, an important aspect of any sampling system is its noise performance. Ordinarily, such a differential scheme would be expected to suffer in this regard since the measurement result is only a small fraction of the signal which is actually measured. Nevertheless, an analysis of the noise limits shows this technique remains quite competitive with techniques which rely on high defect densities to reduce the carrier lifetime since these are usually accompanied by a significant reduction in mobility.

The following quantities are defined for use in the noise analysis:

$G_{ph}(0)$ = maximum photoconductivity at $t = 0^+$

$\overline{G_{ph}}$ = dc average photoconductivity

G_D = dark conductivity

τ_{rec} = carrier lifetime (assumed exponential decay)

T_s = effective sampling window width

L = length of photoconductor gap

V_{sig} = voltage of measured signal

P = average optical power absorbed

$1/T_{rep}$ = sampling repetition rate

Δf = bandwidth of dc sampling current measurement

λ = wavelength of optical pulse source

μ = appropriate mobility of photoconductor (usually electron)

h = Planck's constant

k = Boltzmann's constant

The fundamental noise limits in most photoconductor applications are (1) shot

noise on the average current and (2) Johnson noise from the average conductivity of each photoconductor. For a single photoconductor, these are given by [17]:

$$\overline{i_{SN}^2} = 4e(\overline{G_{ph}} + G_D)V_{sig}\frac{V_{sig}\mu\tau_{rec}}{L}\Delta f, \tag{4.3.1}$$

$$\overline{i_{JN}^2} = 4kT(\overline{G_{ph}} + G_D)\Delta f. \tag{4.3.2}$$

Comparing the two, it is found that the shot noise only dominates if

$$V_{sig}^2 > \frac{kT}{e}\frac{L^2}{\mu\tau_{rec}}, \qquad \text{for} \quad \overline{i_{SN}^2} > \overline{i_{JN}^2}. \tag{4.3.3}$$

The right side of this inequality is typically a few hundred mV which is much greater than expected minimum sensitivity levels ($< 10\ \mu$V) so the analysis can be restricted to Johnson noise.

The net sampling current collected is just the charge per pulse times the pulse repetition rate which is

$$\overline{i_{sig}} = V_{sig}G_{ph}(0)T_s/T_{rep}. \tag{4.3.4}$$

The peak photoconductivity is expressed in terms of the more easily measured *average* photoconductivity as:

$$G_{ph}(0) = \overline{G_{ph}}\,T_{rep}/\tau_{rec}. \tag{4.3.5}$$

Using this, the average net sampling current is:

$$\overline{i_{sig}} = V_{sig}\overline{G_{ph}}T_s/\tau_{rec}, \tag{4.3.6}$$

and thus the signal-to-noise ratio (SNR) is:

$$SNR = \frac{\overline{i_{sig}}^2}{\overline{i_{JN}^2}} = V_{sig}^2\frac{\overline{G_{ph}}^2(T_s/\tau_{rec})^2}{(2)4kT(\overline{G_{ph}} + G_D)\Delta f}. \tag{4.3.7}$$

This expression has been written to be applicable to both the differential technique as well as conventional sampling in which case the sampling window width $T_s \simeq \tau_{rec}$.

The factor of two in parentheses is valid for differential measurements since two signals with uncorrelated noise are subtracted leading to twice the noise power of a single photoconductor. Consistent with analyses of electro-optic sampling [18], the minimum detectable voltage V_{min}, is taken to be that which results in $SNR = 1$ and is thus

$$V_{min} = \left[\frac{(2)4kT(1 + G_D/\overline{G_{ph}})}{\overline{G_{ph}}(T_s/\tau_{rec})^2} \Delta f \right]^{1/2}. \tag{4.3.8}$$

This naturally breaks into two regimes depending on the relative amplitudes of $\overline{G_{ph}}$ and G_D. These are given by

$$V_{min} = \frac{L}{T_s} \sqrt{\frac{(2)4\frac{kT}{e}hc}{\lambda P \mu/\tau_{rec}} \Delta f} \qquad \overline{G_{ph}} \gg G_D, \tag{4.3.9}$$

and

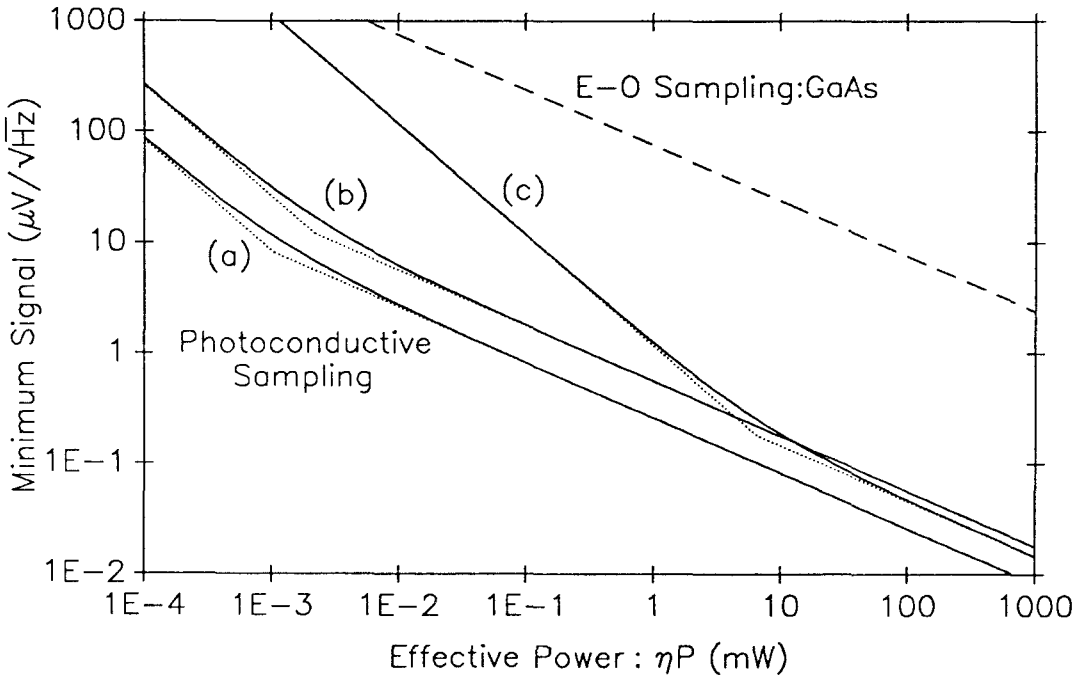
$$V_{min} = \frac{hcL^2}{e\lambda P \mu T_s} \sqrt{(2)4kTG_D \Delta f} \qquad \overline{G_{ph}} \ll G_D, \tag{4.3.10}$$

where we have used

$$\overline{G_{ph}} = \frac{e\lambda P \mu \tau_{rec}}{hcL^2}. \tag{4.3.11}$$

A plot of V_{min} versus optical power is shown in Fig. 4.4 for some typical values of high resistivity photoconductors. Also shown in this plot is a similar plot for electro-optic sampling illustrating the clear superiority of photoconductive techniques with regard to detection limits.

It is obvious that at any power level, it is desired to have $\overline{G_{ph}} > G_D$. Assuming this is attainable, Eqn. (4.3.9) says that for the same geometry and sampling resolution, the minimum detectable voltage is characterized by the material parameter μ/τ_{rec}. Techniques which utilize large defect densities to increase recombination rates also tend to reduce the mobility due to increased scattering off the same defects. Often, the ratio μ/τ_{rec}, remains almost constant over as much as a 100× decrease in lifetime [7],[13]. In special cases (i.e., InP:Fe bombarded with He^+



Curve	μ $cm^2/V\text{-}s$	τ_{rec} ps	G_D MΩ	L μm	T_s ps
a	1000	100	10	10	10
b	2000	150	10	25	10
c	200	10	0.5	25	10

Fig. 4.4 Theoretical minimum detectable voltage versus average optical power for various typical photoconductors. Curves (b) and (c) are typical of ordinary semi-insulating InP:Fe and ion bombarded InP respectively, while curve (a) is based on parameter values which allow easy scaling. Also shown is a curve for electro-optic sampling in GaAs which is one of the best materials in this regard [18].

ions), the lifetime can be reduced at a faster rate than the mobility [19] resulting in superior sensitivity. Eventually however, the mobility suffers a drastic decrease at a threshold bombardment dose and thus the lifetime reduction is limited.

In addition to decreased mobility, ion bombardment dopes the bombarded material and can significantly lower the dark resistance unless the material is carefully pre-doped with a compensating impurity. This can easily lead to a condition of $G_D > \overline{G_{ph}}$, and thus a comparison based on μ/τ_{rec} is no longer valid. A comparison in this case is easily seen by considering the change in the asymptotes of Fig. 4.4 with ion bombardment. As just discussed, the high power asymptote will stay about the same or slightly move down as μ/τ_{rec} increases somewhat. The low power asymptote has twice the slope of the high power regime and their intersection occurs at a power, P_{eq}, of

$$P_{eq} = \frac{hc\,L^2 G_D}{\lambda q\,\mu\tau_{rec}}. \tag{4.3.12}$$

For typical values with ion bombardment of τ_{rec} <10 ps, $\mu \sim 200$ cm^2/V-s and G_D of 0.1 - 1 MΩ, the transition point P_{eq} will increase 3-4 orders of magnitude from the intrinsic case to be at 1-10 mW which is typical of the available power from modelocked dye lasers. Thus, the use of ion bombardment is expected to degrade the minimum detectable voltage due to increased dark conductance noise, especially at low optical powers.

In all cases, V_{min} scales as $1/T_s$ which is just the trade-off of resolution for sensitivity which is characteristic of any photoconductive sampling scheme. However, its implementation with the differential approach is rather easy requiring only the adjustment of a delay between the triggering of two sampling windows. Conversely, such an adjustability would be very awkward with the conventional approach since a separately bombarded photoconductor would be needed for each desired resolution, and the lifetime reduction would have to be well characterized.

In addition to the fundamental noise limitations discussed here, any actual measurement will be plagued by additional sources of "excess" noise. Presumably, these sources can be eliminated or circumvented by various measurement and device design techniques, so that the fundamental noise is relevant to the ultimate sensitivity. However, with photoconductive sampling, the dominant noise source is often due to a photovoltaic response of the photoconductor which detects amplitude noise of the modelocked laser [20]. This effect is not well understood but is probably due to an imbalance of the Schottky responses at the two contacts of the photoconductor. An accurate analysis of the minimum detectable signal including this noise source would require a more detailed description of the photovoltaic response particularly with regard to doping levels, mobility, etc., which might be affected by ion bombardment or deep level traps. Roughly speaking, though, it is expected that the response only varies significantly with doping due to variations in the Schottky barrier width which directly affects collection efficiency in a planar structure. Reduced mobility and lifetime are not expected to affect the photovoltaic noise level except at the very short lifetimes where recombination prevents complete sweep-out from the Schottky barrier. Assuming then that the photovoltaic noise is not significantly changed, the best minimum detectable signal would be achieved by maximizing the signal — Eqn. (4.3.4) — which favors higher mobility and hence the differential scheme. A more extensive investigation of the photovoltaic response is needed before any conclusive comparisons can truly be made. In any case, this problem can be somewhat reduced by stabilization of the modelocked laser source and by using synchronous detection at higher frequencies since most lasers exhibit a noise spectrum which decreases at higher frequencies.

References

[1] D. H. Auston, Appl. Phys. Lett. **26**, 101 (1975).

[2] R. H. Moyer, P. Agmon, T. L. Koch, and A. Yariv, Appl. Phys. Lett. **39**, 266 (1981).

[3] R. A. Lawton and J. R. Andrews, IEEE Trans. Instr. Meas. **25**, 56 (1976).

[4] F. J. Leonberger and P. F. Moulton, Appl. Phys. Lett. **35**, 712 (1979).

[5] P. R. Smith, D. H. Auston, A. M. Johnson, and W. M. Augustyniak, Appl. Phys. Lett. **38**, 47 (1981).

[6] A. G. Foyt, F. J. Leonberger, and R. C. Williamson, Appl. Phys. Lett. **40**, 447 (1982).

[7] P. R. Smith, D. H. Auston, A. M. Johnson, and W. M. Augustyniak, Appl. Phys. Lett. **38**, 47 (1981).

[8] D. H. Auston and P. R. Smith, Appl. Phys. Lett. **41**, 599 (1982).

[9] P. M. Downey, B. Schwartz, Appl. Phys. Lett. **44**, 207 (1984).

[10] P. M. Downey, J. E. Bowers, C. A. Burrus, F. Mitschke, and L. F. Mollenauer, Appl. Phys. Lett. **49**, 430 (1986).

[11] R. Loepfe, A. Schaelin, H. Melchior, and M. Blaser, Appl. Phys. Lett. **52**, 2130 (1988).

[12] D. H. Auston, P. Lavallard, N. Sol, and D. Kaplan, Appl. Phys. Lett. **36**, 66 (1980).

[13] A. M. Johnson, D. H. Auston, P. R. Smith, J. C. Bean, J. P. Harbison, and A. C. Adams, Phys. Rev. B **23**, 6816 (1981).

[14] M. B. Ketchen, D. Grischkowsky, T. C. Chen, C. C. Chi, I. N. Duling, N. J. Halas, Appl. Phys. Lett. **48**, 751 (1986).

[15] D. H. Auston, IEEE Jour. Quant. Elec. **QE-19**, 639 (1983).

[16] D. R. Grischkowsky, M. B. Ketchen, C. C. Chi, I. N. Duling, III, N. J. Halas, J. M. Halbout, and P. G. May, IEEE Jour. Quant. Elec. **QE-24**, 221 (1988).

[17] A. Yariv, *Optical Electronics, Third Ed.*, Holt, Rinehart and Winston, New York (1985).

[18] B. H. Kolner, and D. M. Bloom, IEEE Jour. Quant. Elec. **QE-22**, 79 (1986).

[19] P. M. Downey and B. Tell, J. Appl. Phys. **56**, 2672 (1984).

[20] D. H. Auston, "Ultrashort Laser Pulses and Applications," Springer-Verlag, New York, 1988.

Chapter 5
Differential Sampling: Experiments

5.1 Introduction

In the last chapter, a differential measurement scheme was outlined for performing high speed sampling with a resolution much less than the turn-off time of the photoconductive switches used. The idea of differentiating or subtracting two delayed, decay functions with fast leading edges has been previously used to improve the performance of photoconductors in applications as fast photodetectors [1], [2] and as pulse generators [3]. In this chapter three possible implementations of the differential sampling approach are presented along with results demonstrating the success of the technique.

5.2 Experimental Set-up

Differencing schemes

In order to implement the difference operation of the last chapter, the following three approaches were tried: 1) numerically shift and subtract the result of a conventional sampling result, 2) dither the delay, τ, while synchronously detecting at the dithering frequency, and 3) simultaneously sample with two photoconductors having a relative delay in their turn-on times and subtract the results in real time. The most significant point of comparison of these alternatives is their sensitivity to excess noise sources, in particular the fluctuations in average power of the modelocked laser source.

Modeling the laser power fluctuations as:

$$P_{laser}(t) = P(1 + \xi(t)), \tag{5.2.1}$$

the result of an ordinary sampling measurement will vary as:

$$V_{meas}(\tau,t) = \left[\int dt'\, V_{sig}(t') f_{samp}(t'-\tau)\right](1+\xi(t))^2$$
$$\simeq \left[\int dt'\, V_{sig}(t') f_{samp}(t'-\tau)\right](1+2\xi(t)) \qquad (5.2.2)$$

assuming that the laser pulses are used to trigger both the measured signal and the photoconductive sampler. Assuming the fluctuations $\xi(t)$ are characterized by a spectral density function, $S_\xi(f)$, then the relative noise of the sampling measurement is given by

$$\frac{\text{RMS noise}}{\text{Signal}} = 2\sqrt{\overline{\xi^2}} = 2\sqrt{S_\xi(f)\,\Delta f}. \qquad (5.2.3)$$

A plot of the spectral density, $S_\xi(f)$, is shown in Fig. 5.1 and was taken by measuring the mean square current variation of a photodiode detecting the output of the modelocked dye laser. A two phase lock-in amplifier was used to measure this variation in a narrow bandwidth around a center frequency selected through its internal oscillator.

These fluctuations are especially serious to the first two schemes mentioned. In a typical scan of the delay, τ, points separated by $\Delta\tau$ are measured $\sim$ several seconds apart in real time, and hence the numerical shift and subtract scheme will be sensitive to laser noise around 1 Hz or less. Similarly, the second scheme will also be sensitive to laser fluctuations, though at the dithering frequency of τ. This frequency is limited to a maximum of $\sim$ 100 Hz since it typically requires mechanical movement of a mirror position. From Fig. 5.1, it is seen that this will lead to some improvement over the numerical technique but the improvement is limited. In addition, the mechanical dithering of the mirror caused some modulation in beam pointing which resulted in an extraneous signal due to scanning of the beam on the detector's active area. This scheme also requires some post-processing of the result

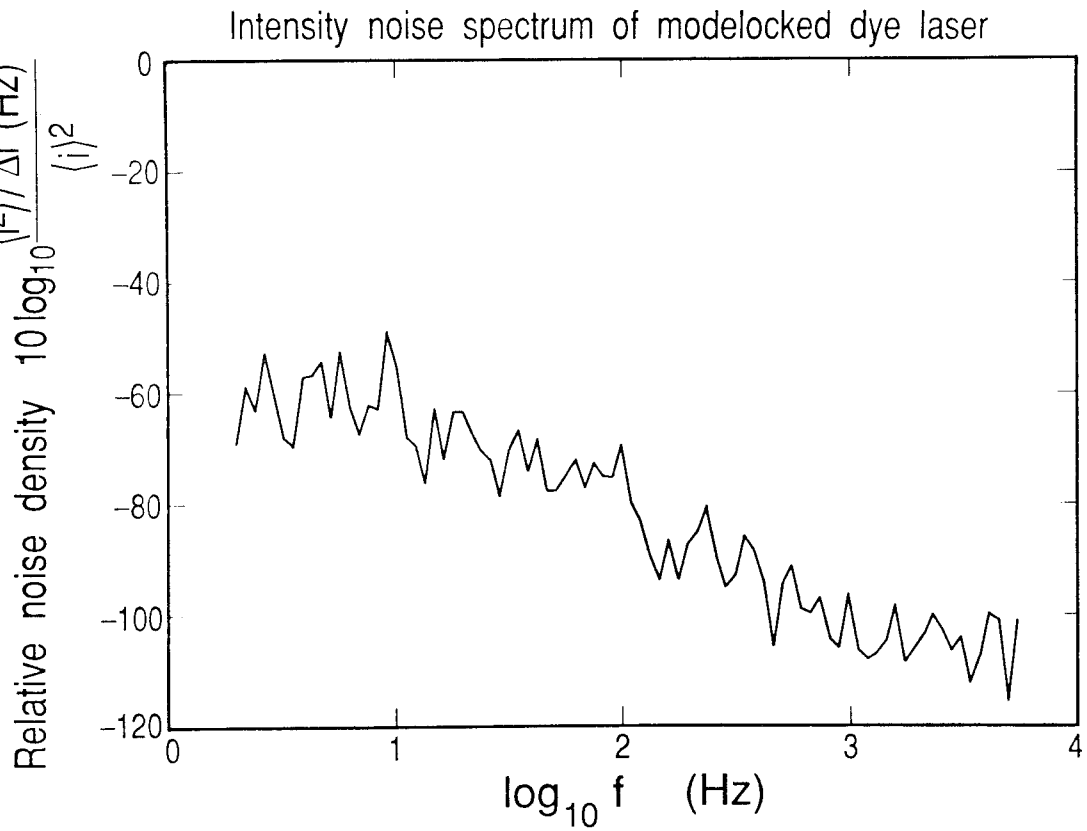


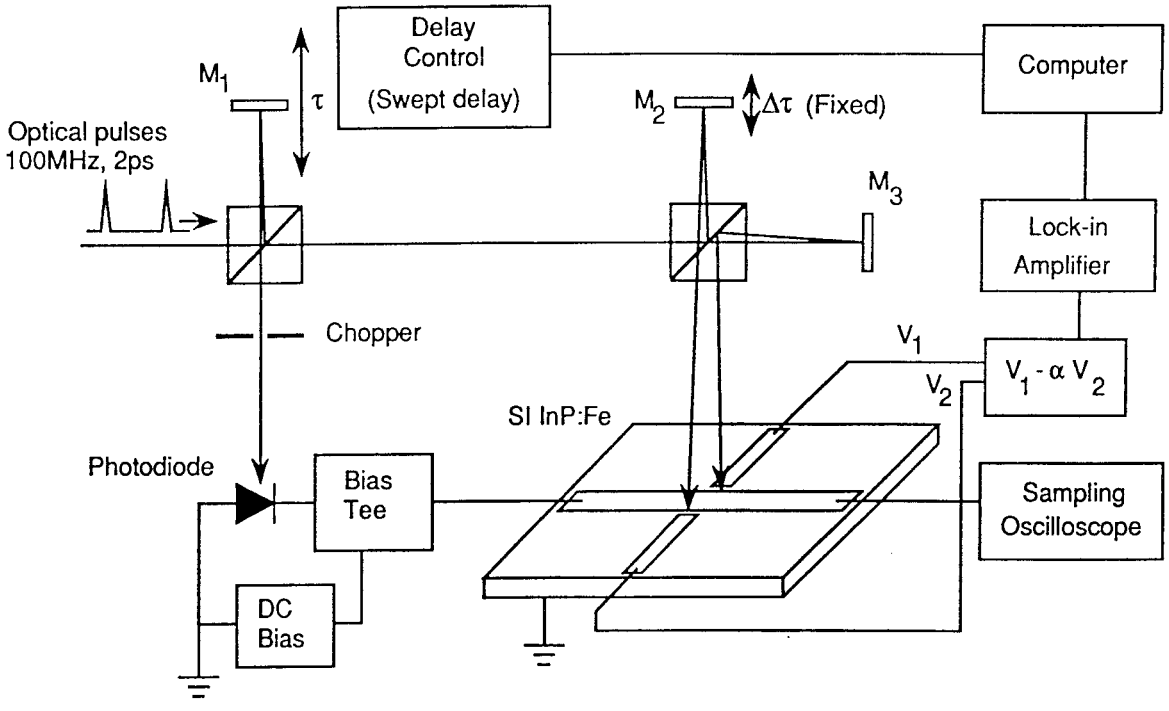
Fig. 5.1 Measured spectral density of the modelocked laser intensity noise, normalized by the average intensity.

to cancel the exponential tails as is done by the factor α in the finite difference schemes.

Finally, the double gap scheme is optimal in reducing the excess noise problems. Since the two signals are measured at the same time, fluctuations due to laser noise will be correlated and thus will tend to cancel when the two signals are subtracted. In essence, the *relative* noise of the difference will be the same as that of the original two signals. In contrast, the absolute noise in the numerical subtraction scheme is increased by $\sqrt{2}$ while the difference signal is only some fraction of the original signal; thus the relative noise can increase quite significantly and limits the practical resolution. A disadvantage of the double gap scheme is that it requires a certain degree of matching of the two photoconductors such that their carrier decay tails cancel well.

Set-up and Photoconductor Characterization

The experimental set-up used for sampling is shown in Fig. 5.2, illustrating the double gap scheme in measuring the impulse response of a photodiode. A synchronously pumped modelocked dye laser provides a train of optical pulses (3-5 ps pulsewidth, λ =600nm, 100 MHz rep. rate) for triggering of the devices. Part of the beam illuminates a high speed photodiode to generate the test signal which is coupled onto the center microstrip transmission line of the sampler. The other end of this line is either fed to a sampling oscilloscope (50Ω input) for monitoring or terminated in 50Ω to prevent reflections. The rest of the optical pulse train is then split again to illuminate the two sampling photoconductors on either side of the center transmission line. The average currents from these sampling electrodes are input to matched transimpedance amplifiers and the outputs are then subtracted with an adjustable factor, α, in a differential amplifier. This output is then detected with a lock-in amplifier synchronized to a chopper in the beam illuminating the



Experimental Set-up

Fig. 5.2 Experimental set-up for differential sampling measurement of a photodiode.

photodiode.

The delay between "turn-on" of the two photoconductors — $\Delta\tau$ — is adjusted by micrometer movement of the mirror M_2, and remains fixed during a sampling scan. The zero point delay is accurately established by searching for interference fringes when the two beams are overlapped. Since the coherence length of the pulses is actually much less than the pulse width — $\sim$ 300 fs versus $\sim$ 3 ps — this enables extremely accurate determination of the coincidence point and subsequently the sampling window width. The time variable, τ, is then swept by stepper motor control of the position of mirror M_1. Sweep speeds are kept slow enough that the lock-in time constant does not significantly degrade the resolution. The same set-up is used in the numerical shift and dithering schemes except only one photoconductor is illuminated, and the position of mirror M_1 is dithered ($\sim \pm 0.6$mm) in addition to the slow scan of τ.

The actual mounted sampler is shown in Fig. 5.3 along with the equivalent circuit of a single pair of photoconductive gaps. Multiple sampling gaps were used to conserve substrate material since success of any particular gap was not guaranteed and each sample required $\sim$ 13mm$\times$13mm square of substrate. All lines are designed for 50Ω impedance using the empirical formula [4]

$$\frac{w}{h} = 1.25\left[\frac{5.97}{\exp(Z_0\sqrt{\epsilon_r + 1.41}/87)} - \frac{t}{h}\right], \tag{5.2.4}$$

where Z_0 is the characteristic impedance, w is the width of the line, h is the substrate thickness, t is the metallization thickness, and ϵ_r is the relative dielectric constant. For the 25μm gaps used, the capacitance is estimated to be 0.03 pf [5] and the gap charging time of $2Z_0C_g$ is then about 3 ps. The AuGe/Au metallization is patterned by a conventional lift-off process and was annealed at 340 °C for 5 minutes. The processed wafer is cleaved somewhat oversize, glued to the copper mounting block, and then lapped flush with the edges to ensure continuity of the transmission line

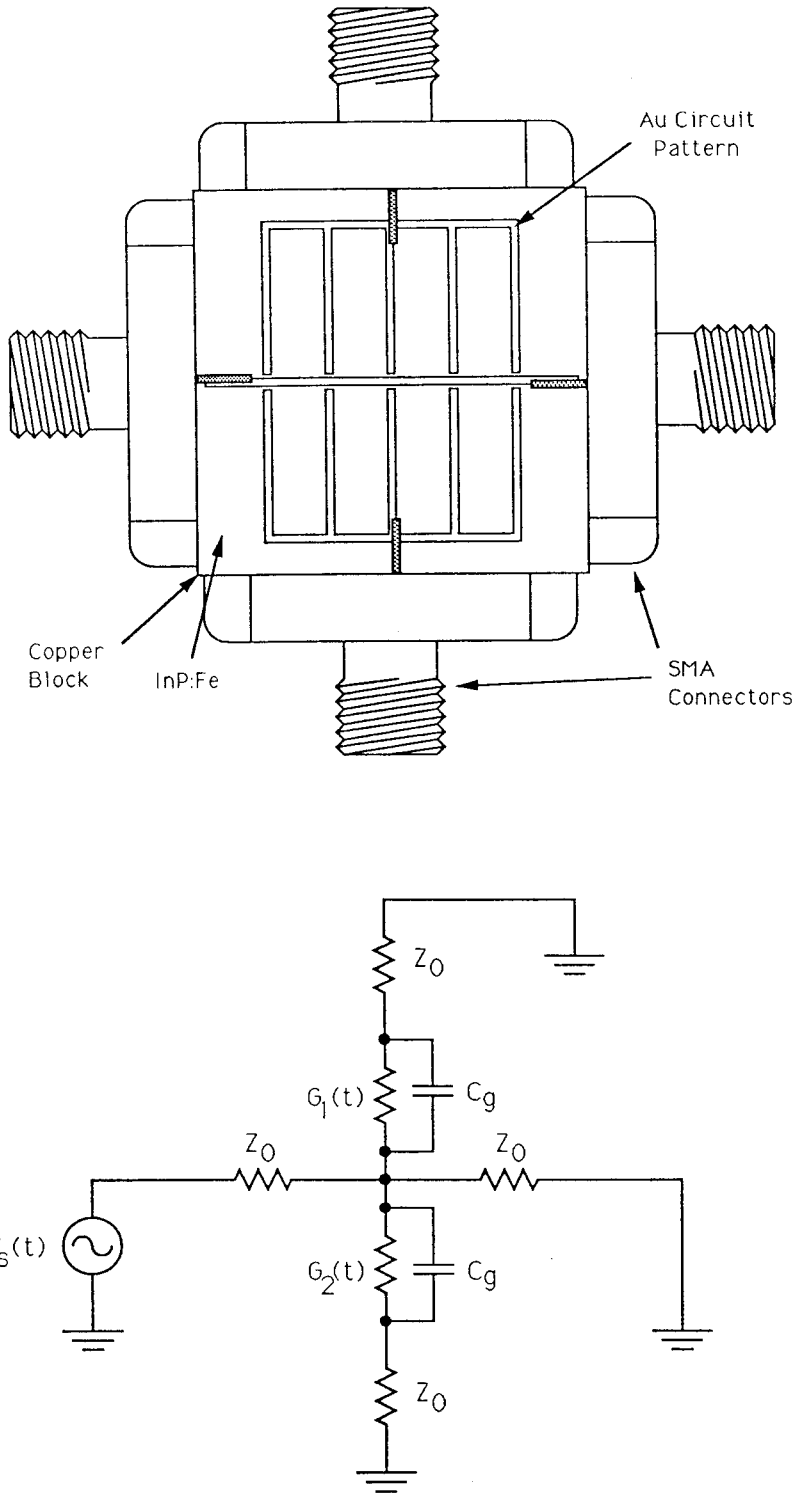


Fig. 5.3 Mounted sampling device (top) and equivalent circuit of a pair of photoconductors (bottom).

impedance at the connectors; thus avoiding potential reflections which can occur if the wafer does not butt against the connector.

Several substrates were tried including semi-insulting GaAs:Cr, InP:Fe, high resistivity undoped GaAs as well as some MBE grown layer structures on semi-insulating GaAs. The InP:Fe is preferred due to a lower surface recombination velocity which, if too large, can lead to a non-exponential photoconductive decay [6]. This is especially true when the illumination is far above the band gap energy such that absorption is strong and carriers are created very near the surface. In addition to this, the InP typically gave a stronger signal ($\sim 4\times$) under the same conditions and showed less photovoltaic effects. Both of these effects are believed to be due to somewhat better contacts with the InP, although good contacts were difficult with all semi-insulating materials.

The dc, open circuit photovoltaic response was as high as 100 mV with only 0.2 mW of optical power. By moving the beam focus position across the gap, the polarity could be reversed indicating that opposing Schottky barriers at the contact edges are the source of this signal. With careful positioning of the focus spot, the photovoltaic signal could be nulled to just a few per cent of the peak value, which helped reduce the noise in sampling measurements.

5.3 Measurement Results

The result of a single gap sampling measurement of the photodiode pulse response is shown in Fig. 5.4. Increasing τ corresponds to later turn-on of the sampling gap with respect to the photodiode signal. The left side of the peak reflects the conductivity decay of the photoconductor ($\sim$150ps) while the right side is predominantly due to the photodiode response. A significant amount of structure in the photodiode response is indicated, though it has been smoothed out by the effective integration of this measurement. The results of the numerical shift and subtract

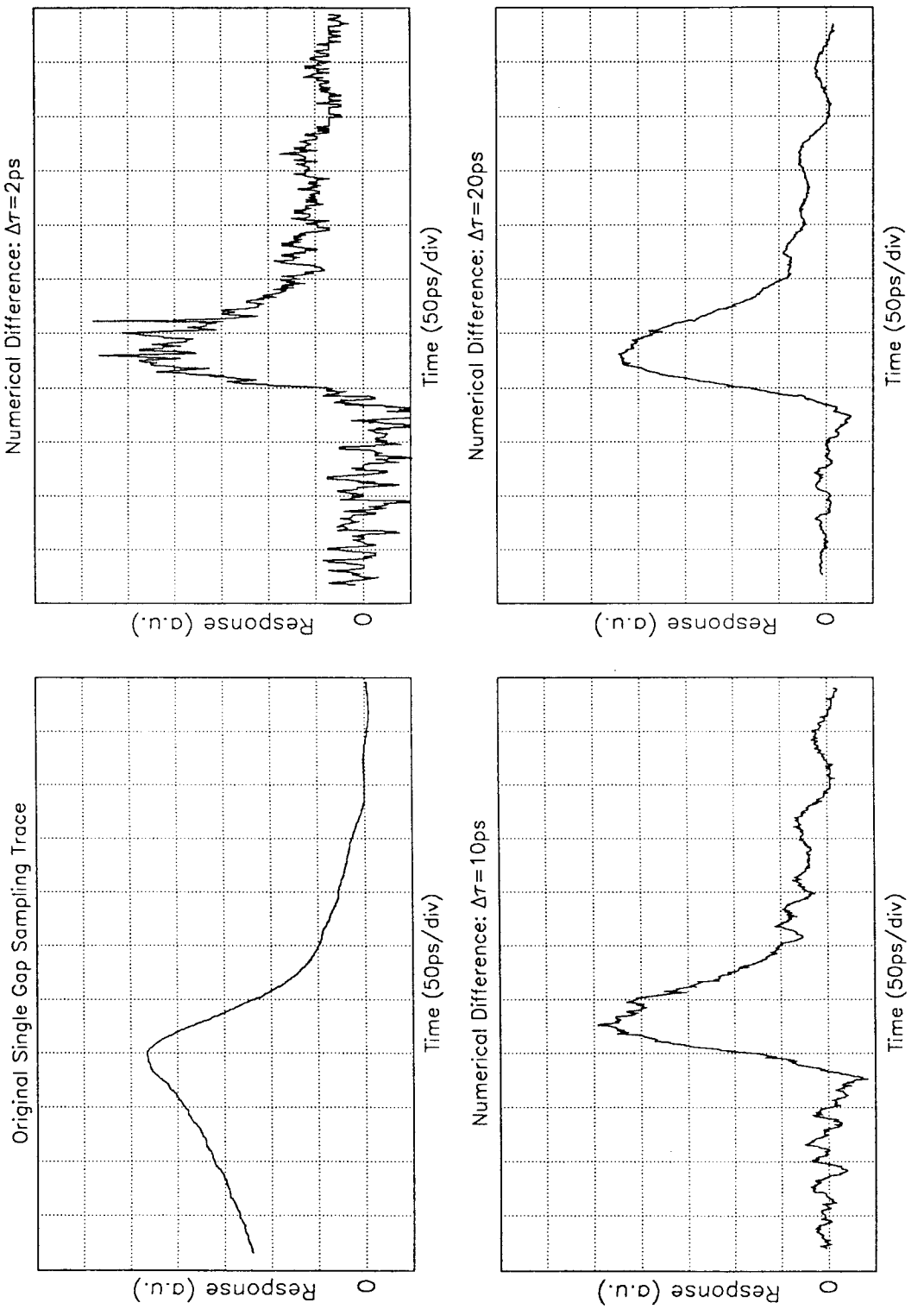
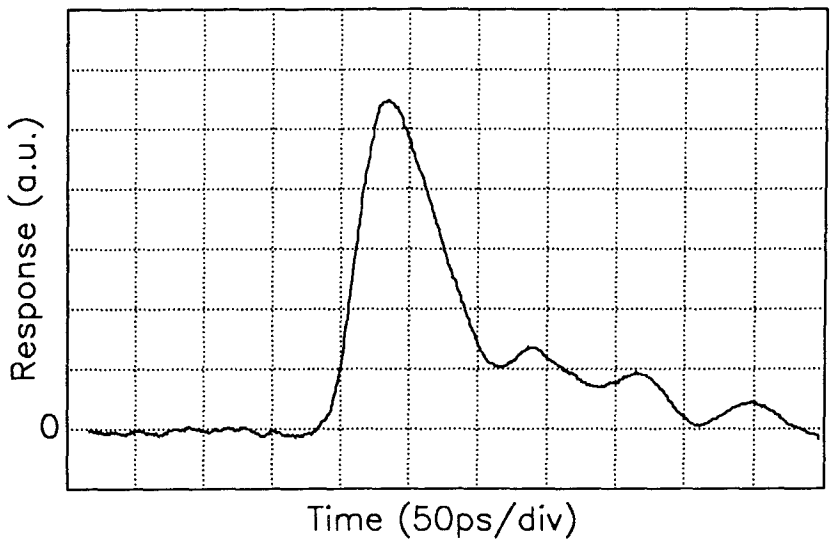


Fig. 5.4 Single gap sampling measurement of the photodiode response (upper left) and the results obtained by numerically differentiating this trace for various values of $\Delta\tau$.

scheme are also shown in Fig. 5.4 for various values of the shift $\Delta\tau$. The factor α is chosen so as to minimize the signal in the region to the left of the peak of the original since this is where the impulse response should be zero. Also, each difference signal has been appropriately rescaled for easier comparison. The signal recovery is surprisingly good, though the noise level is rather high as just discussed, especially with shorter sampling windows. The situation here is somewhat improved since the signal is shorter than the photoconductive decay. For a longer signal, the noise could be 2 to 3 times larger. Conversely, shorter signals would have lower noise and this may be a legitimate technique for measuring shorter pulses.

In contrast, the result of the double gap measurement is shown in Fig. 5.5. The factor α is adjusted to null the signal to the left of the peak and must compensate differences in the overall response of the two gaps as well as the matching factor $\exp(-\Delta\tau/\tau_{rec})$. Good cancellation also required testing of several gap pairs as well varying of the light spot position on each gap in order to get good matching of the decay times of the two gaps. The time window, $\Delta\tau$, was 10 ps for this measurement; however such resolution is unfortunately not verified here due to the lack of suitably fast features on the test signal. With such a slow test signal, the measured signal shape can be verified with a commercial sampling oscilloscope measurement of the same signal. As shown in Fig. 5.5, the agreement between the two is good, thus establishing the fidelity of the technique. The improvement in noise over the numerical approach is quite apparent and the trailing edge features are now well defined. These oscillations are due to ringing of the resonant circuit formed by the diode capacitance and bond wire inductance of the mounted photodiode. Repeatability of the double gap measurements was superior to that of the sampling oscilloscope, particularly with regard to these trailing edge features. Sampling oscilloscopes typically introduce "ringing" behavior of their own into measurement



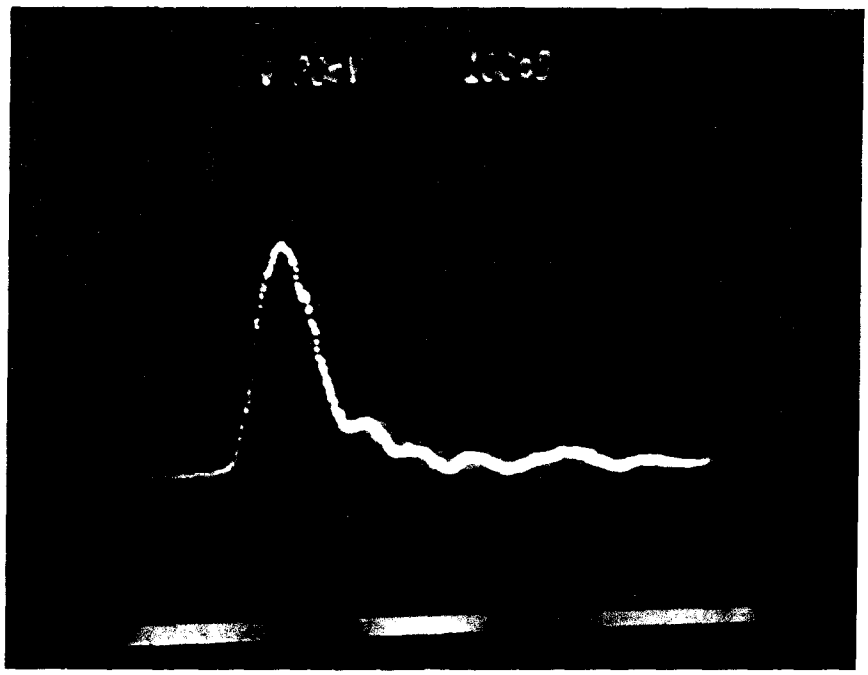


Fig. 5.5 Differential sampling result of photodiode using double gap technique (top), and commercial sampling oscilloscope measurement for comparison.

results (i.e., their sampling function has some oscillating tail feature) as well as suffering from jitter problems at such short time scales.

The sampling oscilloscope also calibrates the vertical sensitivity with a peak value of 60 mV. The corresponding peak sampling current measured by the lock-in is 11 pA, giving an overall sensitivity — $\overline{i_{sig}}/V_{sig}$ — of 180pA/V. Accounting for a factor of 1/2 due to the chopping of the photodiode beam, this is equivalent to a peak photoconductivity of $(2.8\text{M}\Omega)^{-1}$, while the expected value is $(100\text{k}\Omega)^{-1}$ for the average optical power of 5μW. On the other hand, the dc resistance does not show this discrepancy with measured values of 5-10MΩ and an expected value of 6.5MΩ. This is attributed to poor contacts which block the dc current when a pulsed bias is applied.

When the photodiode signal is zero, a noise level of 0.065 $\text{pA}/\sqrt{\text{Hz}}$ is measured from the sampling output. Together with the above sensitivity, the minimum detectable signal is then 360 $\mu\text{V}/\sqrt{\text{Hz}}$. Of the total noise, 0.055 $\text{pA}/\sqrt{\text{Hz}}$ can be attributed to the 10 MΩ feedback resistors of the transimpedance amplifiers, which indicates that significant improvement could be had by using a larger feedback resistance. However, if contacts were improved, then the expected gap resistance would dominate the Johnson noise and a higher feedback resistance would be unnecessary. In this case, the improved sensitivity afforded by the higher effective peak conductivity would reduce the minimum detectable signal level toward its theoretical limit of 16 $\mu\text{V}/\sqrt{\text{Hz}}$.

The optical power of 5 μW used in these measurements is quite low and better sensitivity should be obtained with higher powers. However, this was not found to be the case with these photoconductors and the best results were obtained with lower optical powers. At higher powers (from 5 μW to 500 μW), noise increased significantly while the signal level rose only as approximately the square root of

the optical power instead of linearly as would be expected. Such a square root dependence of photocurrent on optical power and can be explained by bimolecular recombination of the excess carriers [7]. However, even at the highest powers used and assuming a 0.1 μm absorption depth, the carrier densities are $< 2 \times 10^{17} \mathrm{cm}^{-3}$. This is well below the level at which bimolecular processes would dominate, especially in semi-insulating material with its high density of deep impurity levels. Furthermore, the single gap measurements provide a fairly good measurement of the photoconductor carrier recombination which is represented by the portion of the curve to the left of the peak in Fig. (5.4). If nonlinear recombination processes were significant, then with a 100$\times$ variation in optical power, some change would certainly be visible in the shape of this leading edge. Such measurements showed no such change and it is concluded that the carrier recombination is quite linear over this range of optical power. The sublinear current-light characteristic can be qualitatively explained by poor contacts which lead to space charge effects and blocking of the dc part of transient current pulses. While this unexplained behavior is disturbing, it is not critical to the accuracy of sampling measurements. Much more important is the linearity of the sampling with respect to the voltage signals being measured. By varying the optical power to the photodiode, the peak signal voltage was varied from 2 mV to 150 mV. The measured sensitivity over this range was found to be constant to better than 2% which was the measurement accuracy.

While good performance has been demonstrated, it is expected that improved performance could be obtained, especially with respect to sensitivity and general facility in obtaining success. The most likely avenue of success is improved contacts which seem to be related to most of the problems encountered. Improved contacts to semi-insulating semiconductors can definitely be obtained with improved metallization and annealing procedures; though the extent of improvement is uncertain

and it will certainly be difficult to match the performance of contacts made to highly doped layers. In light of this, it seems warranted to sacrifice some dark resistivity and use doped epitaxial layers to obtain good contacts. Such layers will also have longer decay times (~1 ns) which will also deteriorate the noise limit somewhat, but will reduce the matching requirements.

References

[1] W. Margulis, U. Österberg, B. Stoltz, A. S. L. Gomes and W. Sibbett, Optics Comm. **54**, 171 (1985).

[2] W. Margulis and R. Persson, Rev. Sci. Instrum. **56**, 1586 (1985).

[3] Y. Hori, J. Paslaski, M. Yi, and A. Yariv, Appl. Phys. Lett. **46**, 749 (1985).

[4] H. R. Kaupp, IEEE Trans. Elec. Computers **EC-16**, 185 (1967).

[5] M. Maede, IEEE Trans. Microwave Theory Tech. **MTT-20**, 390 (1972).

[6] K. K. Li, J. R. Whinnery, A. Dienes, in "Picosecond Optoelectronic Devices," (C. H. Lee, ed.) Academic Press, New York: 1984.

[7] R. Loepfe, A. Schaelin, H. Melchior, and M. Blaser, Appl. Phys. Lett. **52**, 2130 (1988).

Chapter 6
Retrieval of a Transient Impulse Response from its Autocorrelation

6.1 Introduction

The development of picosecond and femtosecond modelocked lasers over the last 10 years has been accompanied by a variety of measurement techniques which utilize these sources [1]. These include pump-and-probe techniques [2], photoconductive sampling [3], electro-optic sampling [4], [5] and second harmonic generation (or nonlinear frequency summing) to measure the ultra-fast optical pulses themselves [6]. Most of these can be classified as sampling schemes in that typically a short sampling function (or "window") is mixed with the signal to be measured and the total integrated result is measured with a slow detector. When the relative delay between the two mixed signals is varied, the cross-correlation of the two is readily obtained. In the limit that the sampling function can be considered a delta function, the measured signal is retrieved exactly. Frequently, however, circumstances dictate that a signal be sampled by itself, and consequently the obtained result is the auto-correlation of the measured signal. This can be due to the difficulties of synchronizing the measured signal with a sampling window from a separate source; or most commonly, because a suitably fast sampling signal is not available. This is the case when measuring the sampling function itself: since the temporal resolution of such a scheme is limited by the sampling window, it is chosen to be the shortest possible.

In general, it is not possible to retrieve the original function from its autocorrelation: there simply is not a unique solution. This is readily apparent in the frequency domain since the Fourier transform of the autocorrelation is just the squared magnitude of the transform of the original function. Thus, the phase of

the original transform can be arbitrarily chosen and there is an infinite set of original functions for any single autocorrelation. Consequently, the analysis of such measurements is usually limited to qualitative deduction of the pulse shape and an approximation of the pulse width. This approximation is usually based on the calculated ratio of the respective widths — full width at half maximum (FWHM) — of a chosen function and its autocorrelation. The choice of a functional form may be motivated by a calculated solution of the measured phenomena, but for the most part it is arbitrary.

In this chapter, we pursue the possibility of extracting more than just a pulse width estimate from the autocorrelation. By imposing additional known constraints on the original unknown function, the infinite set of phase functions is hopefully reduced to a tractable set or even a unique solution. This is a topic which has received considerable attention in the literature and is commonly referred to as the phase retrieval problem. In particular, applications to image formation and recovery have been a very active topic for nearly thirty years [7]. Other fields of application include x-ray diffraction [8] and optical coherence theory [9] to name a few. The analytic approach used here — the logarithmic Hilbert transform — has been extensively investigated for such applications but with only limited success. Due to the symmetric nature of problems in spatial coordinates, a unique solution is usually not determined by this method as will be explained later. As such, more recent work in these fields tends to be concerned with numerical algorithms which alternately enforce various constraints and approach a solution iteratively [10], [11]. On the other hand, the application of phase retrieval techniques to autocorrelation measurements of temporal transients appears to have been overlooked in the literature. Investigation shows that these problems are actually very well suited to the use of the logarithmic Hilbert transform due to some additional constraints which

are unique to many problems in the time domain.

In this chapter we present results on the retrieval of an impulse response from its autocorrelation. An analysis of the problem is given first, reviewing the logarithmic Hilbert transform and discussing the aspects of this problem which enable its successful application. Second, a numerical algorithm implementing this technique is presented along with results on some test functions.

6.2 Analysis

Logarithmic Hilbert transform

The logarithmic Hilbert transform has been extensively studied in the literature and so just a brief review is presented to clarify the results to be discussed here. In particular, we draw on many of the results from Burge et al. [12] and the reader is referred to that work for a more detailed discussion of much of the background.

The Fourier transform relations which are used are given as:

$$\tilde{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t}\, dt \tag{6.2.1}$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \tilde{f}(\omega) e^{j\omega t}\, d\omega \tag{6.2.2}$$

and $\tilde{f}(\omega)$ will be written as

$$\tilde{f}(\omega) \equiv |\tilde{f}(\omega)| e^{j\phi(\omega)}, \tag{6.2.3}$$

where ϕ is a real valued function of ω. Also, the cross-correlation $h(\tau)$ between two functions $f(t)$ and $g(t)$ is written as

$$h(\tau) \equiv f \circ g \equiv \int_{-\infty}^{\infty} f(t+\tau) g^*(t)\, dt. \tag{6.2.4}$$

It is easily shown in most introductory texts that the Fourier transform $\tilde{h}(\omega)$ of the cross-correlation function is $\tilde{f}(\omega)\tilde{g}^*(\omega)$; however, if ω is allowed to be complex, the correct expression is

$$\widetilde{[f \circ g]}(\omega) = \tilde{f}(\omega)\tilde{g}^*(\omega^*). \tag{6.2.5}$$

In any case, for real ω, the Fourier transform of the autocorrelation of a function $f(t)$ is

$$\widetilde{[f \circ f]}(\omega) = |\tilde{f}(\omega)|^2, \tag{6.2.6}$$

from which it is apparent that the phase information — $\phi(\omega)$ — has been lost.

In order to recover $\phi(\omega)$, some additional constraints must be supplied at this point. We start with two:

(1) the function $f(t)$ is to be restricted to a finite interval $a < t < b$,
i.e., $f(t) = 0$ for $t < a$ and $t > b$ where a and b are not yet specified;

(2) the function $f(t)$ is real valued.

The constraint that $f(t)$ be zero before $t = a$ is just the causality condition, assuming the signal to be a response initiated at some time $t < a$. The additional condition that $f(t)$ also vanish for $t > b$ is first motivated in that it allows the application of many results from the theory of "entire" or "integral functions." However, it would seem to exclude many common functional forms which typically characterize transient phenomena and theoretically persist forever (e.g., exponential decay). Nevertheless, in considering an actual measurement situation, this assumption is found to be quite justified. While one might expect an infinite duration signal, any measurement is practically limited to some finite time interval. Furthermore, if the signal is decaying, then it will be indistinguishable from zero after some finite time due to limitations of instrument resolution or noise. Of course these arguments apply only to the measured autocorrelation and do not necessarily mean that the function retrieved from it must be of finite duration also. For now we will just assume that such finite-time solutions exist and that they are the solutions we are interested in. Thus we resign ourselves to finding finite duration approximations to the possibly infinite autocorrelation and signal being retrieved.

The general approach at this point is to apply an argument similar to that used

to derive the Kramers-Kroenig relations which relate the real and imaginary parts of a causal frequency response function by a pair of Hilbert transforms. Seeking a relation between amplitude and phase, the complex logarithm of the frequency response is used here instead, and the result is referred to as the logarithmic Hilbert transform.

We start by considering the integral

$$\oint \frac{\log \tilde{f}(\omega')}{\omega' - \omega}\, d\omega' \tag{6.2.7}$$

over the closed contour shown in Fig. 6.1, comprised of the real axis and the semicircle, C_∞, in the lower half of the complex ω'-plane. The function $\log \tilde{f}(\omega')$ is singular at the poles and zeroes of $\tilde{f}(\omega')$. Causality ensures that all poles are in the upper half plane but dictates nothing about the zeroes. For now, it is assumed that all zeroes are also in the upper half plane and return to this point later. Application of the Cauchy integral formula will yield the desired result provided the contribution from C_∞ vanishes as the radius goes to infinity. Unfortunately, this is not so straightforward. Usually in problems like this, one shows that the magnitude of the kernel of the integral decays faster than $1/R$ everywhere along the semicircular contour of radius R. Then as $R \to \infty$, the path integral vanishes. This is clearly not the case for the integral considered here. For a finite energy signal $f(t)$, the spectral energy density, $|\tilde{f}(\omega)|^2$, must be integrable over all frequencies; hence, as $|\omega| \to \infty$, $|\tilde{f}(\omega)|$ vanishes and $\log \tilde{f}(\omega)$ diverges. This problem is addressed by Burge et al. [12] who also note that several authors seemingly neglect it altogether. They consider several solutions to this problem, most notably the so called modified logarithmic transform. If the lower limit in the Fourier transform definition is effectively $a = 0$, then the convergence problems are alleviated by using an additional factor of $1/\omega'$

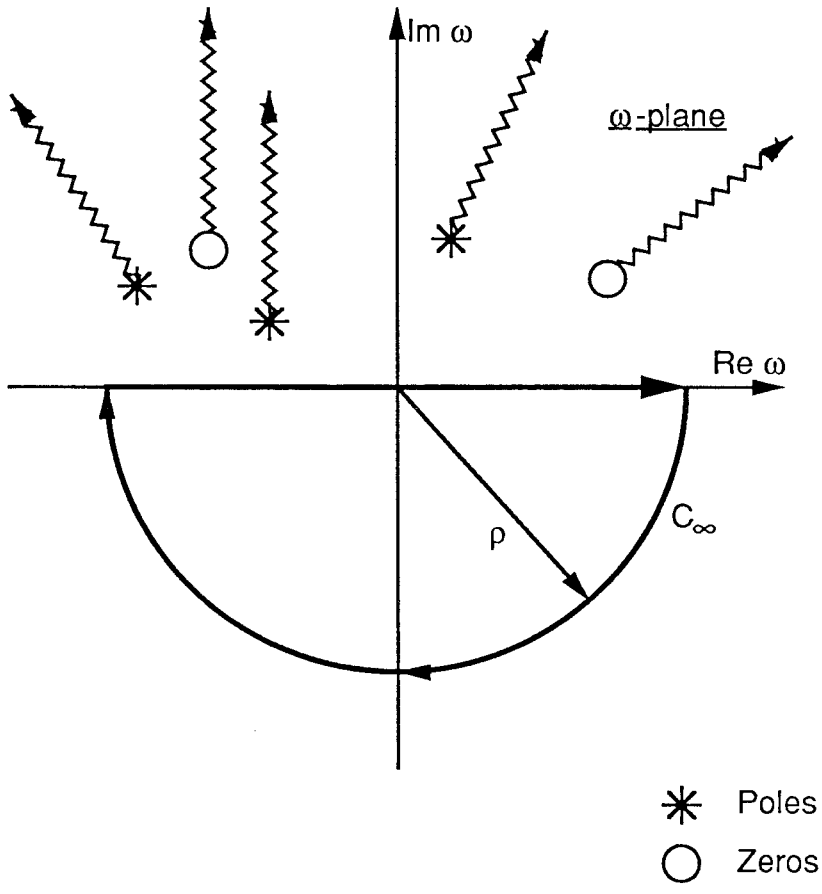


Fig. 6.1 Contour for evaluation of the path integral Eqn. (6.2.7) in the complex ω' plane. All poles of $\tilde{f}(\omega')$ are in the upper half plane and all zeroes are assumed to be there also. Hence all branch cuts of $\log \tilde{f}$ can be restricted to the upper half plane.

to give the modified integral

$$\oint \frac{\log \tilde{f}(\omega')}{\omega'(\omega' - \omega)} d\omega' \tag{6.2.8}$$

evaluated over the same contour as before. Applying the Cauchy integral formula and taking the real and imaginary parts of the result gives the following relations:

$$\phi(\omega) = \frac{\omega}{\pi} \int_{-\infty}^{\infty} \frac{\log |\tilde{f}(\omega')|}{\omega'(\omega' - \omega)} d\omega' + \phi(0), \tag{6.2.9}$$

$$\log |\tilde{f}(\omega)| = -\frac{\omega}{\pi} \int_{-\infty}^{\infty} \frac{\phi(\omega')}{\omega'(\omega' - \omega)} d\omega' + \log |\tilde{f}(0)|. \tag{6.2.10}$$

The first equation is the desired relation giving the phase function $\phi(\omega)$ in terms of the known amplitude $|\tilde{f}(\omega)|$.

This relation has been used successfully to do phase retrieval [13] and would be adequate for the problem here. However, the assumption of a real $f(t)$ allows Eqn. (6.2.9) to be simplified to one which requires less computation when implemented numerically. Expanding the denominator of Eqn. (6.2.9) we can rewrite it as

$$\phi(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\log |\tilde{f}(\omega')|}{\omega' - \omega} d\omega' - \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\log |\tilde{f}(\omega')|}{\omega'} d\omega' + \phi(0)$$

Now if $f(t)$ is real, then $\phi(0) = 0$ and $|\tilde{f}(-\omega)| = |\tilde{f}(\omega)|$. Taking the Cauchy principle value of the integral, the last two terms are then zero since $1/\omega'$ is an odd function of ω'. Our final result is then

$$\phi_m(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\log |\tilde{f}(\omega')|}{\omega' - \omega} d\omega', \qquad \text{for real } f(t). \tag{6.2.11}$$

This is just the original result we would have expected above but had to reject due to the divergence of $\log |\tilde{f}(\omega)|$ at infinity (the subscript m is explained below).

It is worth noting that an attempt to reduce Eqn. (6.2.10) to a complementary relation similar to Eqn. (6.2.11) leaves an arbitrary constant given by $\log |\tilde{f}(0)|$ plus

an integral. The integral does not vanish since $\phi(\omega')$ is an odd function, but it is independent of ω. Such an arbitrary constant is actually to be expected since all of the restrictions which led to the transforms would allow multiplication of $f(t)$ or $\tilde{f}(\omega)$ by a real constant. Thus $\log|\tilde{f}(\omega)|$ is indeterminate to at least an arbitrary added constant. There are also some similar arbitrary terms to $\phi(\omega)$ which have been subtly eliminated by various assumptions of the preceding argument and now bear explicit mention. First, an arbitrary added constant is uniquely determined by the reality condition which requires that $\phi(0) = 0$. Next, an arbitrary linear term is expected in the phase since the addition of a term $-\omega$T just shifts $f(t)$ by a time T which does not alter the autocorrelation nor does it violate the causality condition for T > 0. However, the validity of using the integrand of Eqn. (6.2.7) to alleviate the convergence problems along C_∞ depends on the effective lower limit of $f(t)$ being $t = 0$. Thus, in using Eqn. (6.2.9), the position of $f(t)$ has been implicitly established (and the arbitrary linear phase term fixed) such that the first nonzero behavior occurs at $t = 0$. This is verified in practice with the numerical signal recovery algorithm discussed later in the chapter.

Zeroes in the Lower Half Plane

We now return to the assumption that all zeroes of $\tilde{f}(\omega)$ are in the upper half of the complex ω-plane. The importance of zeroes is due to branch cuts of the logarithm function which originate at zeroes as well as poles of the argument. If we allow for zeroes in the lower half plane, then the integral path must be modified as indicated in Fig. 6.2. Evaluation then yields the modified phase result:

$$\phi(\omega) = \phi_m(\omega) + 2\sum_n \arg(\omega - \omega_n),$$

where the ω_n are the zeroes in the lower half plane. It is seen that all the terms in the sum will be positive, hence the term $\phi_m(\omega)$ is often called the minimal phase.

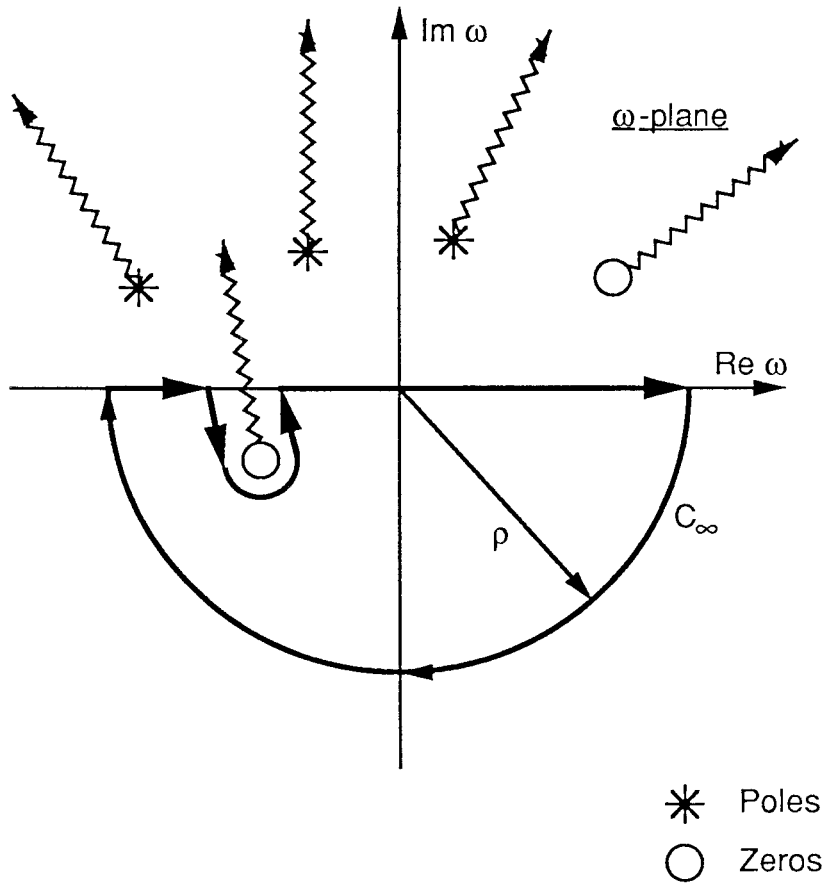


Fig. 6.2 Modified contour to accommodate branch cuts of $\log \tilde{f}(\omega')$ originating at zeroes of $\tilde{f}(\omega')$ in the lower half plane.

Others also refer to it as the Hilbert or canonical phase. All quantities associated with the minimal phase solution will be denoted by a subscript of m. With this result, the general retrieved transform can be written

$$\tilde{f}(\omega) = |\tilde{f}(\omega)|\, e^{j\phi_m(\omega)} \prod_n \frac{\omega - \omega_n}{\omega - \omega_n^*}. \tag{6.2.12}$$

The product term is commonly referred to as a Blaschke product and may be interpreted as the general meromorphic function having unity magnitude along the real axis. Of course this still represents an infinite set of solutions since the zeroes can be arbitrarily chosen and thus the problem of finding a unique solution reduces to specifying those zeroes of $\tilde{f}(\omega)$ which lie in the lower half plane.

The allowed combinations of lower half plane zeroes can be restricted by some general considerations. If we still maintain the restrictions of finite domain and integrability of $f(t)$, then $\tilde{f}(\omega)$ can have no poles in the finite ω-plane. Thus terms in the Blaschke product are limited to those for which ω_n^* is a zero of $\tilde{f}_m(\omega) \equiv |\tilde{f}(\omega)|\, e^{j\phi_m(\omega)}$. This can be viewed as limiting the possible solutions to those obtained by "flipping" various combinations of upper half-plane zeroes of $\tilde{f}_m$ across the real axis into the lower half plane. Next, the reality of $f(t)$ implies the following symmetry condition:

$$\tilde{f}(-\omega^*) = \tilde{f}^*(\omega). \tag{6.2.13}$$

Therefore, if a zero at ω_n is flipped, its mirror image (across the imaginary axis) must also be flipped to maintain this symmetry. Finally, it has been shown that the location of zeroes far from the origin depends primarily on the leading/trailing edge behavior of $f(t)$ while the general shape of $f(t)$ is governed by zeroes near the origin [8], [14]. Consequently, only a finite number of zeroes are relevant since the zeroes of $\tilde{f}_m$ have a finite density [15]. This point is further illustrated by considering the

effect of a Blaschke factor in the time domain, which is just the convolution of $f_m(t)$ with the transform of the Blaschke factor. For a zero $\omega_n = \mu_n + i\nu_n = |\omega_n|e^{i\alpha_n}$ in the lower half plane, and maintaining the symmetry condition of (6.2.13), this transform is

$$f_B(t) = \delta(t) + 4|\omega_n|\frac{\nu_n}{\mu_n}e^{\nu_n t}\cos(\mu_n t - \alpha_n), \tag{6.2.14a}$$

or if $\mu_n = 0$, then

$$f_B(t) = \delta(t) - 2|\omega_n|e^{\nu_n t}. \tag{6.2.14b}$$

If ω_n is far from the origin, then the decay and/or oscillations of these functions are very fast. Consequently, they are either similar to a delta function or give no net contribution when integrated against a "slowly varying" function. The exception to this will be near the termination points of the function $f_m(t)$. Consequently any unusual behavior at the edges of retrieved functions should be considered suspect — i.e., as possibly due to improper zero location far from the origin.

While these considerations reduce the allowed set of solutions, there is still generally a finite number of relevant zeroes of $\tilde{f}_m$ which could be flipped in various combinations giving a set of equally valid solutions. To further reduce this set, additional information must be supplied either from theoretically motivated constraints or additional measurements. In the latter category, schemes have been proposed for image reconstruction using an additional measurement with defocusing [16] or the introduction of a reference wave, similar to holography[9]. On the theoretical side, various constraints relevant to image fields have been imposed in attempts to find unique solutions. These have usually not been conducive to an analytic approach and are usually implemented in an iterative scheme where the various constraints are alternately satisfied in a repeated cycle [17]-[20].

Of course the simplest case would be if there were grounds to assert that there are no zeroes in the lower half plane and then the unique solution would be just the

minimal phase ϕ_m. While general conditions which would enable this assertion have been discussed in the literature, they were usually not applicable to the problems being considered. In fact, it has even been stated that the minimal phase is not very significant and that the behavior of the final solution would be dominated by the distribution of zeroes [14]. However, it will be shown that for a large class of temporal impulse response phenomena, there are in fact no zeroes in the lower half plane and thus the minimal phase is the desired solution. Following this, the relevant differences between these problems and imaging problems will be briefly discussed to illustrate why this approach works here.

Rather than try to define the entire set of such impulse functions, we will start with a few specific but useful theorems and show what a broad range of systems they can be successfully applied to. It is in no way implied that the set of functions so defined includes all possible cases with no lower half plane zeroes. Thus, there may be further examples for which the minimal phase is valid but which do not satisfy the criteria presented here.

We start with a theorem on integral functions given by Pólya [21] :

If $f(t)$ is continuous, positive and differentiable, except at a finite number of points, and if

$$\alpha \leq -\frac{f'(t)}{f(t)} \leq \beta \qquad (a < t < b),$$

then all the zeroes of $\tilde{f}(\omega)$ lie in the strip $\alpha \leq \mathrm{Im}\,\omega \leq \beta$.

The parameters a,b are the effective limits of integration in Eqn. (1) — i.e., $f(t) = 0$ for $t < a$ and $t > b$. As a special case of this theorem, another theorem — also due to Pólya — is that

If $f(t)$ is positive and non-increasing, then all the zeroes of $\tilde{f}(\omega)$ lie in the half-plane $\mathrm{Im}\,\omega \geq 0$.

While these theorems seem quite powerful in that they include all monotonic decay

behavior, direct application would require the pulse response to have an infinitely fast rise time which is unphysical.

We continue with some results from the theory of passive, linear networks. Although specifically stated in terms of electrical circuits, the results can of course be generalized to any analogous linear system. Simply stated, the transfer function of a simple ladder network has no zeroes in the lower half plane [22]. A simple ladder network is any circuit which can be drawn as in Fig. 6.3 and precludes the presence of node-bridging as well as certain transformer configurations. In other words, zeroes in the lower half plane can only occur if there is more than one path in the circuit from input to output such that the contributions from those paths might cancel at some complex frequency ω. In addition, the impedance (or admittance) between two points in *any* circuit has no zeroes in the lower half plane. While the response of many circuits is included in the theorems of Pólya, an important extension to include a variety of oscillatory behavior is added by these circuit theorems.

Finally, it is obvious that if $\tilde{f}(\omega)$ is the product of some functions $\tilde{f}_n(\omega)$ then the zeroes of $\tilde{f}(\omega)$ are just the collected zeroes of the functions $\tilde{f}_n(\omega)$. Actually, some zeroes could be cancelled by a pole of another function in the product, but no new zeroes can be created. Thus if all the functions $\tilde{f}_n(\omega)$ in the product can be shown to have no zeroes in the lower half plane, then it follows that the function $\tilde{f}(\omega)$ also has no zeroes in the lower half plane. In the time domain, this product decomposition means that $f(t)$ is the multiple convolution of functions $f_n(t)$, each of which satisfies some condition guaranteeing no zeroes in the lower half plane of ω. We also note that it is very important to distinguish convolution from correlation here (the two are loosely interchanged surprisingly often), for according to Eqn. (6.2.5) correlation flips the zeroes of the second function's transform into the lower half plane.

As a specific example of how these relations are applied in a general sense, we

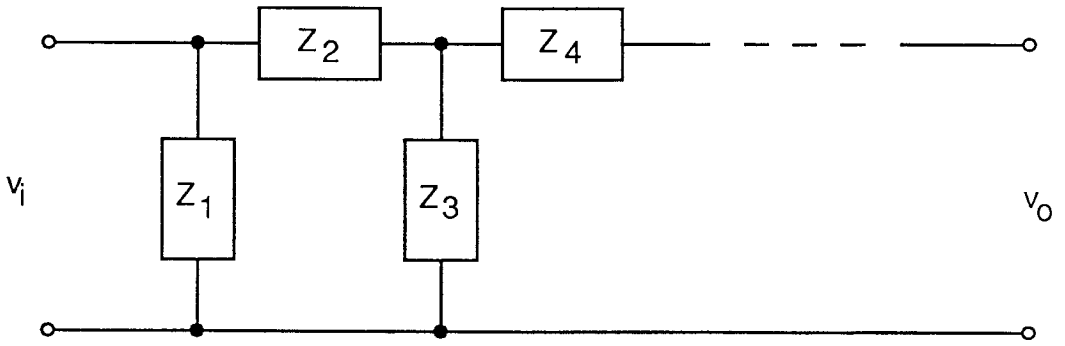


Fig. 6.3 Generalized "ladder circuit" configuration, where Z's are arbitrary linear, passive impedances.

briefly consider the case of the pulse response of photodiodes or photoconductors. Such photodetectors can generally be modeled as a time varying current source driving an external circuit including the load impedance. The current source turns on almost instantaneously when carriers are created and then decays with the removal of excess carriers either through recombination, diffusion, sweep-out, etc.. While a variety of functional forms are possible, most are monotonically decreasing and so satisfy the Pólya criteria. The external circuit is most simply a parallel combination of capacitance and the transmission line impedance, but may include bond wire inductance, parasitic contact resistance and capacitance, etc.. The ladder circuit configuration would still be satisfied in all but the most unusual cases or if intentional non-ladder circuits are included such as a phase shifter. Thus we can assert that the pulse response of the photodetector, which is the convolution of the current source and circuit response, has a transform with no lower-half-plane zeroes and the logarithmic Hilbert transform can be used for retrieval.

While it may seem irrelevant to consider photodiodes since they can not be cross correlated as photoconductors can, we have really been investigating phase retrieval in general and the "deautocorrelation" objective is only a specific application. Frequently the spectral response of devices is measured in amplitude only (so called "scalar" measurements) and hence the unknown phase and impulse response could be derived from such measurements using phase retrieval. The only restriction on the devices to be measured is that it somehow be assertable that there are no lower half plane zeroes. This should be very straightforward with most simple, passive devices. Unfortunately, active devices will probably not meet this criteria since they will typically display exponentially rising features if they can exhibit gain.

Finally, we'll briefly contrast the temporal problem considered here with the

more often studied spatial problems for which the logarithmic Hilbert transform is not so successful. First, the spatial problems usually deal with a truly complex field and so the condition that the final result be real is not valid. This leaves an additional undetermined constant in Eqn. 6.2.9 and also precludes the use of the simpler integral Eqn. 6.2.11. More significant though are the constraints on the location of zeroes. The theorems of Pólya and the circuit theorems suggest that upper half plane zeroes are associated with decay in the time domain which in turn implies that there be a distinction between forward and backward directions. In temporal problems such a distinction is determined by causality and entropy. However, in spatial problems there is usually a symmetry between forward and backward and thus it is not possible to assert that a quantity will be decaying in one direction or another. Hence the zeroes are equally likely to be in either half plane and are typically distributed in both. Hence the logarithmic Hilbert transform is not usually sufficient for phase retrieval in such problems and further constraints and techniques are required.

6.3 Numerical Algorithm and Results

Having established the validity of Eqn. (6.2.11) for phase retrieval of many transient phenomena, we now present results of a numerical algorithm implementing this relation. In addition to demonstrating successful signal reconstruction, we will also discuss several practical problems as well as possible pitfalls associated with numerically implementing the theory of Section II.

Algorithm and Test Functions

The overall signal recovery algorithm is schematically diagrammed in Fig. 6.4. The input is the autocorrelation which is transformed using a complex fast-Fourier-transform (FFT) to yield $|\tilde{f}(\omega)|^2$ and subsequently $\log|\tilde{f}(\omega)|$. According to Eqn. (6.2.11),

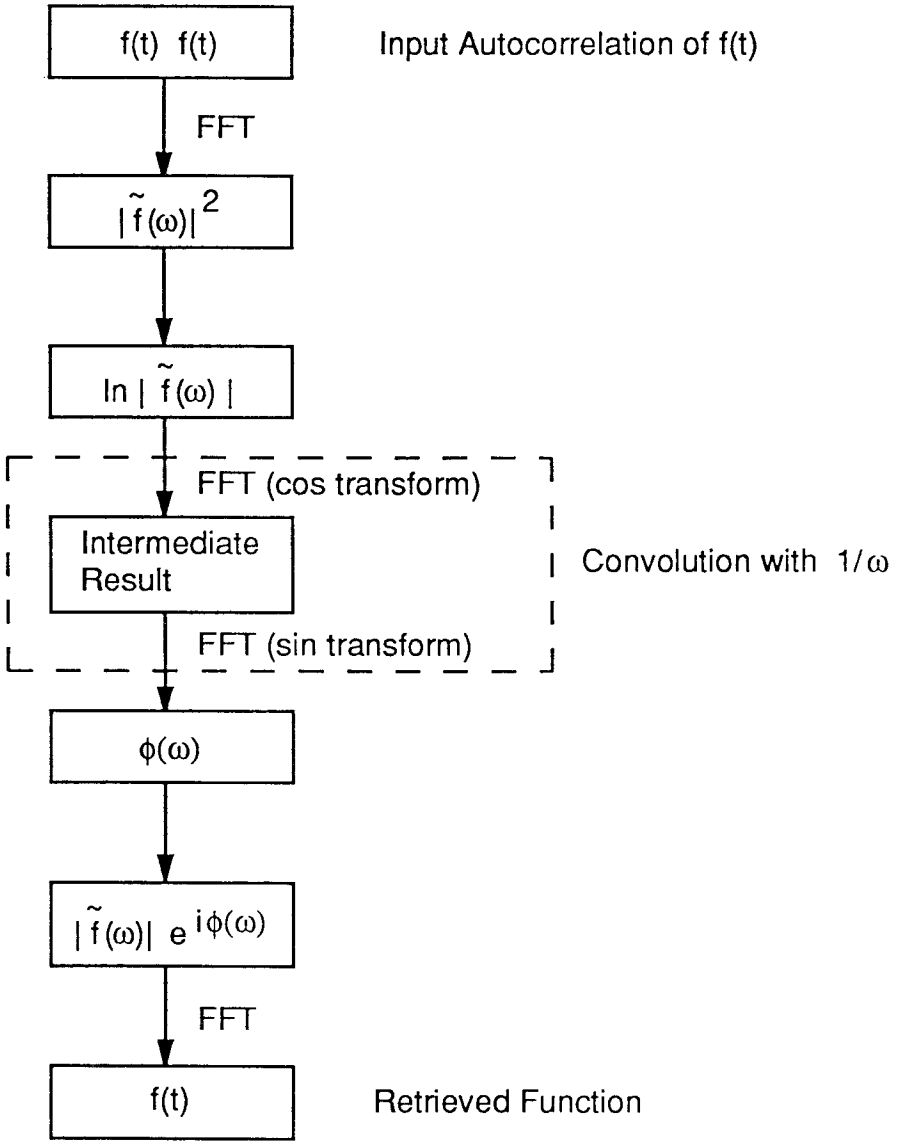


Fig. 6.4 Block diagram of the numerical algorithm for retrieving impulse response from its autocorrelation.

this must then be convolved with $1/\omega$ to yield the retrieved phase. Transforming this, the convolution becomes a multiplication with the Fourier transform of $1/\omega$ which is $-j\pi\text{sgn}(t)$ where

$$\text{sgn}(t) = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1 & t < 0. \end{cases}$$

Using this together with the even symmetry of $|\tilde{f}(\omega)|$ and then inverse transforming, the desired convolution becomes:

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\log|\tilde{f}(\omega')|}{\omega' - \omega} d\omega' = 4\int_0^{\infty} d\xi\, \sin(2\pi\xi\omega)\left[\int_0^{\infty} d\nu\, \log|\tilde{f}(2\pi\nu)|\cos(2\pi\nu\xi)\right]. \tag{6.3.1}$$

Evaluating the convolution in the Fourier space has been shown to give superior results compared to numerically evaluating the integral directly[13]. This is due to the inaccuracy of representing and integrating $1/(\omega' - \omega)$ near the singular point $\omega' = \omega$ using discrete points. This double transform then yields the phase $\phi_m(\omega)$ and inverse transforming $|\tilde{f}(\omega)|e^{j\phi_m(\omega)}$ with a final FFT step gives the retrieved impulse response $f(t)$.

In order to demonstrate the success of this scheme, we present a comparison of the retrieval of two similar functions. In addition to the final results, these examples are used to illustrate some practical problems which can degrade the results if not properly handled. The two sample functions used — see Fig. 6.5a — are $f_A(t) = te^{-t/\tau_A}$ and $f_B(t) = e^{-t/\tau_2} - e^{-t/\tau_1}$ with $\tau_2 = 5\tau_1$. These are impulse responses corresponding to a double pole frequency response and f_A is the limiting case of a second order pole (i.e., when $\tau_1 \to \tau_2$). The pulsewidth τ parameters have been selected such that the autocorrelations have the same width (FWHM) as shown in Fig. 6.5b. These curves have been rescaled to a peak value of 1 to facilitate comparison of shapes. The autocorrelations were derived analytically and

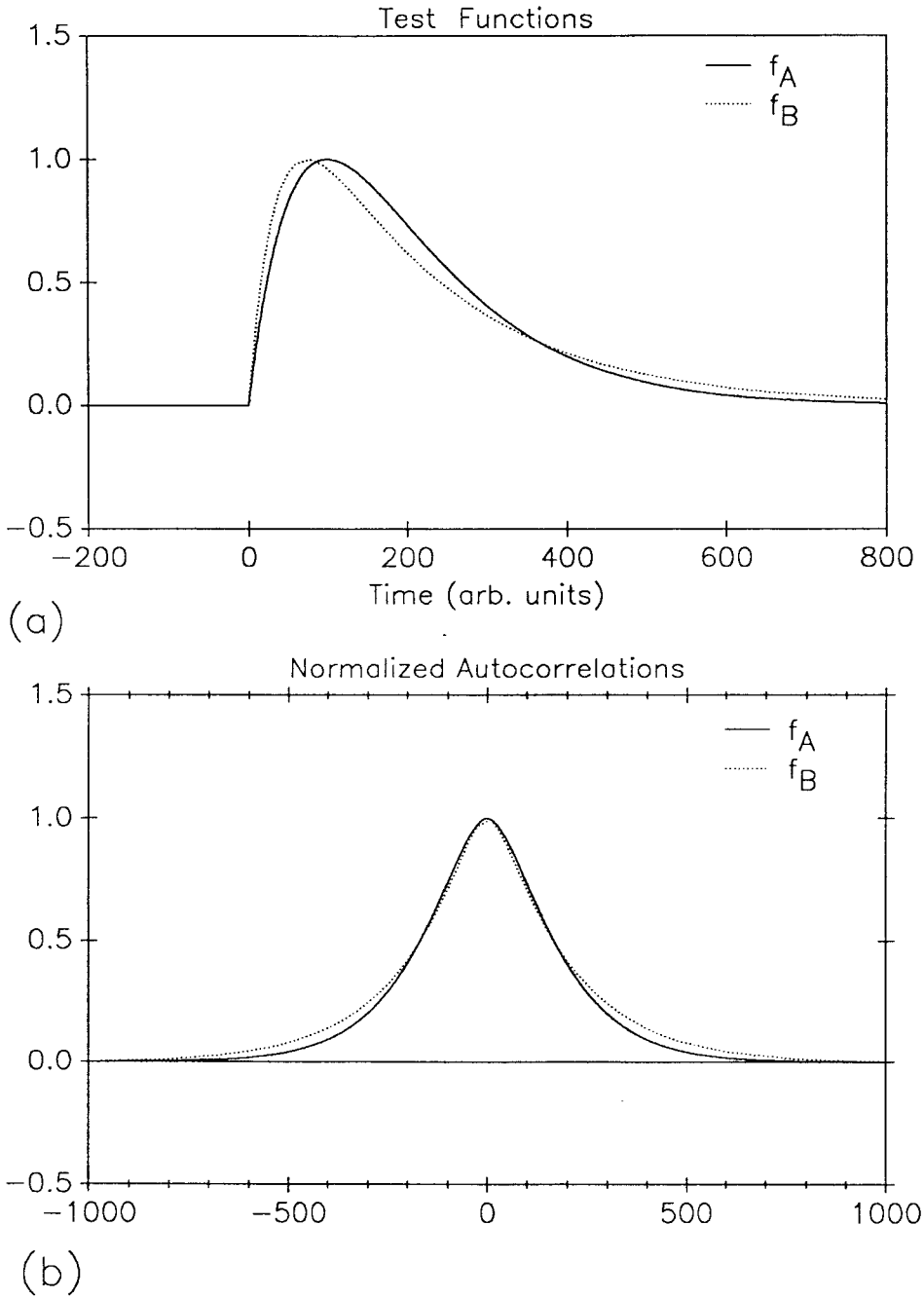


Fig. 6.5 (a) Original test functions f_A, f_B used and (b) their autocorrelations, normalized to the same peak value. The two autocorrelations have the same FWHM, but their detailed shapes differ as do the original functions.

then rounded to 10^{-3} of the peak value (i.e., ± 0.0005) to simulate measurement precision limits in an actual experiment.

Termination and Accuracy of the Autocorrelation

Before discussing the algorithm itself, we first consider effects due to the condition of the autocorrelation which is presumably the input data of the algorithm. In particular, the manner in which the autocorrelation is terminated and its accuracy have significant effects on the quality of the results obtained. The termination effects are important since in typical measurements where only a pulse width is sought, the autocorrelation is often not evaluated to where it is essentially zero. It will be shown that this is not adequate for the signal recovery algorithm and good results are only obtained if the autocorrelation is completely evaluated to where it is indistinguishable from zero (within the accuracy of the measurement). Second, the use of a rounded-off autocorrelation not only shows the effects of measurement accuracy limitations, but also indicates the significance of noise.

It is difficult to establish the effect of a poor termination in general due to the nonlinear nature of the problem. However, with a simple analysis we can at least predict some features which are to be expected particularly at the leading and trailing edges of the retrieved signal. If an arbitrary signal is assumed to be nonzero only over a finite interval given by $0 < t < t_T$, then its autocorrelation is zero for $|\tau| > t_T$ and the approach to zero is governed by the leading and trailing edge behavior of the original function. Specifically, if the leading and trailing edges are assumed to behave as:

$$f(t) \sim \begin{cases} t^{n_L}, & t > 0, \\ (t_T - t)^{n_T}, & t < t_T, \end{cases}$$

then the autocorrelation $h(\tau)$ goes to zero as:

$$h(\tau) \sim (t_T - \tau)^{n_L + n_T + 1} \qquad \text{as } \tau \to t_T, \text{ and } n_L + n_T \neq -1.$$

Now if the autocorrelation is abruptly terminated, then the original function should have a delta function behavior at one edge and an abrupt termination ($n = 0$) at the other. Even if the autocorrelation approaches zero linearly, the edges of the retrieved function will tend to terminate abruptly. In these cases the larger termination value, or the delta function will typically occur at the leading edge so that zeroes of the transform will remain in the upper half plane. This is explicitly brought out in theorems by Cartwright relating the asymptotic location of such zeroes to the leading and trailing edge behavior [23], [24].

The effect of an abrupt termination of the autocorrelation is shown in Fig. 6.6. The autocorrelation used (Fig. 6.6a) is that of f_A and is terminated at just 5% of its peak value. The resulting retrieved function is shown in Fig. 6.6b and is seen to be a poor representation of the original f_A. The result shows the predicted spike feature on the leading edge as well as an extra bump in the tail corresponding to the time at which the autocorrelation was terminated ($\sim$ 470). In addition there seems to be a slow oscillation as well as some sharp discontinuities in the result.

The leading edge spike is also typically generated as a result of noise (simulated by round-off) in the autocorrelation. Applying the algorithm to the complete autocorrelation of f_B, results in the retrieved function shown in Fig. 6.7. Other than the leading edge spike and noise, the result matches quite well with the original f_B.

A Bode plot of the frequency spectrum in this case — Fig. 6.8— shows an apparent noise floor which is actually due to the round-off of the autocorrelation as verified by comparing the "noise" level at different values of round-off precision. This noise floor leads to two features in the retrieved signal. First, the result is itself noisy which is not surprising, and second, there is a large spike at the leading edge of the signal. This spike can be interpreted as the result of the algorithm suppressing all signals for $t < 0$, including noise. In the limit of a truly *white* noise source,

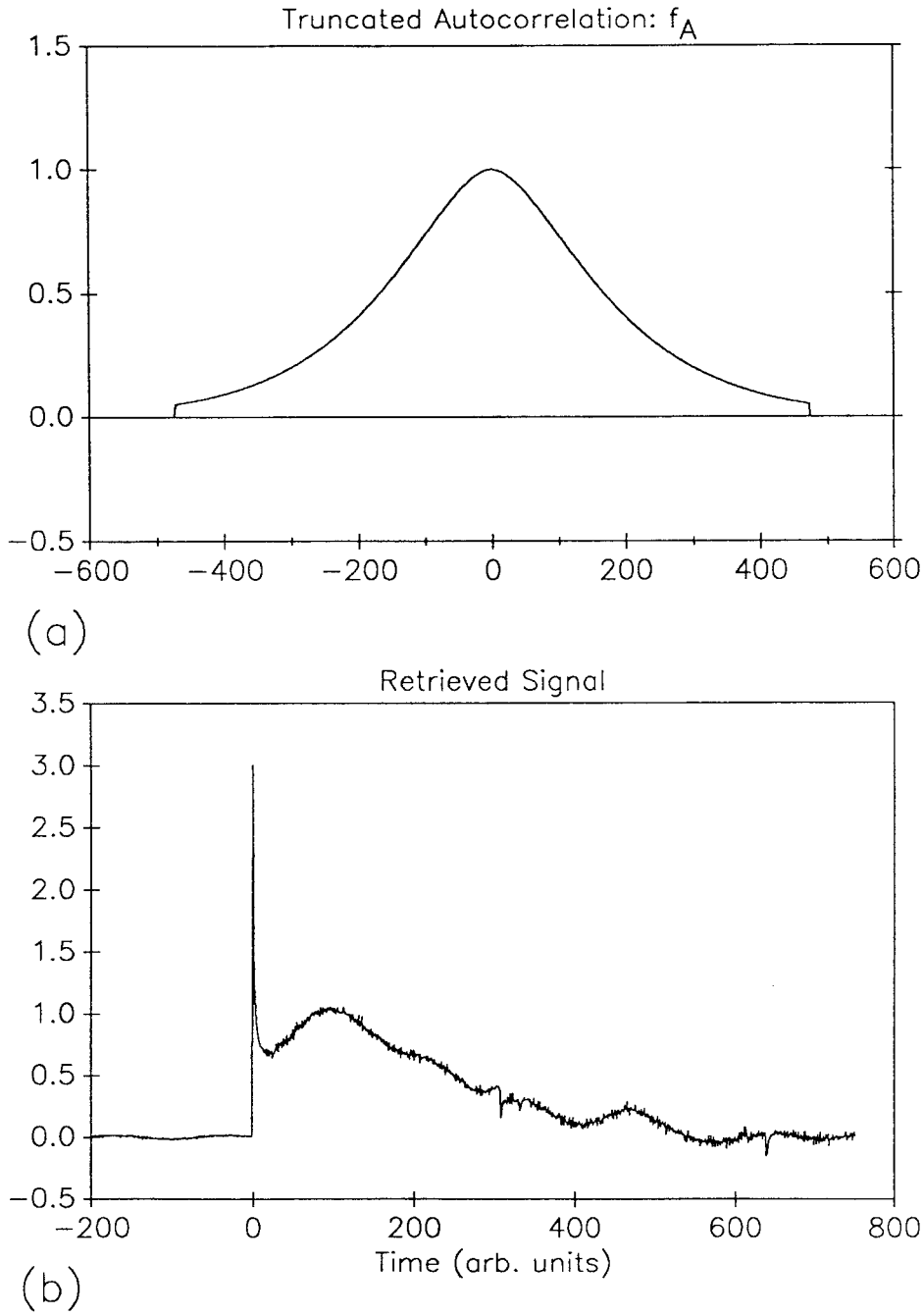


Fig. 6.6 (a) Autocorrelation of f_A truncated at 5% of its peak value. (b) Resulting retrieved function using phase retrieval algorithm. Large leading edge spike and poor overall shape can be attributed to truncation of autocorrelation.

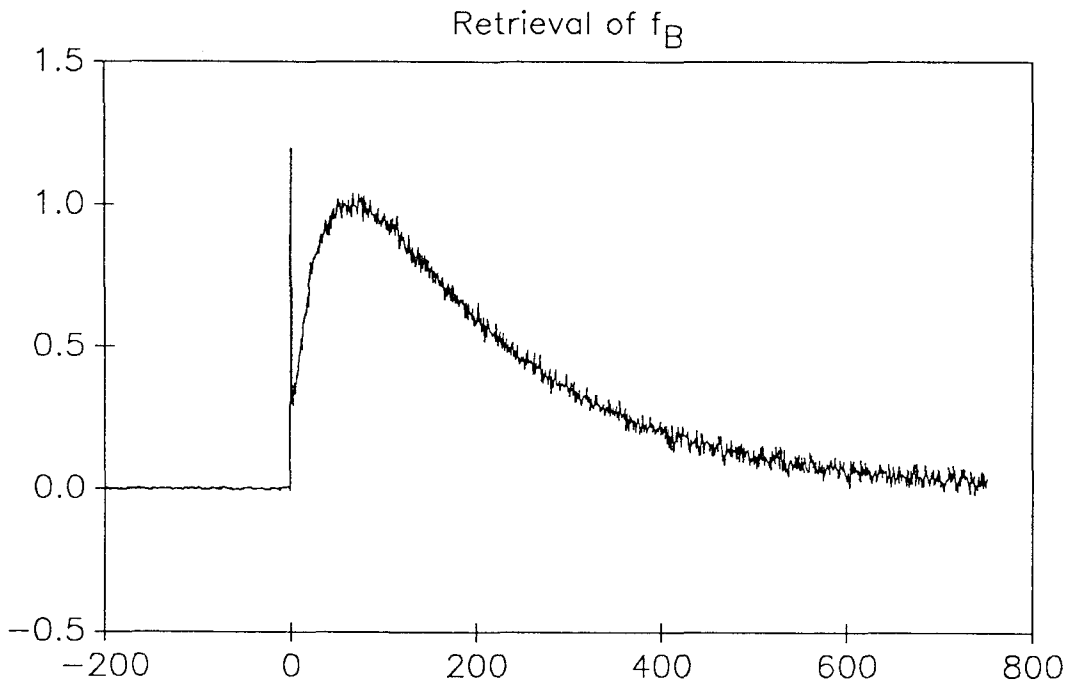


Fig. 6.7 Pulse response retrieval using the full autocorrelation of Fig. 6.5. The initial spike and noisiness are due to round-off of the autocorrelation.

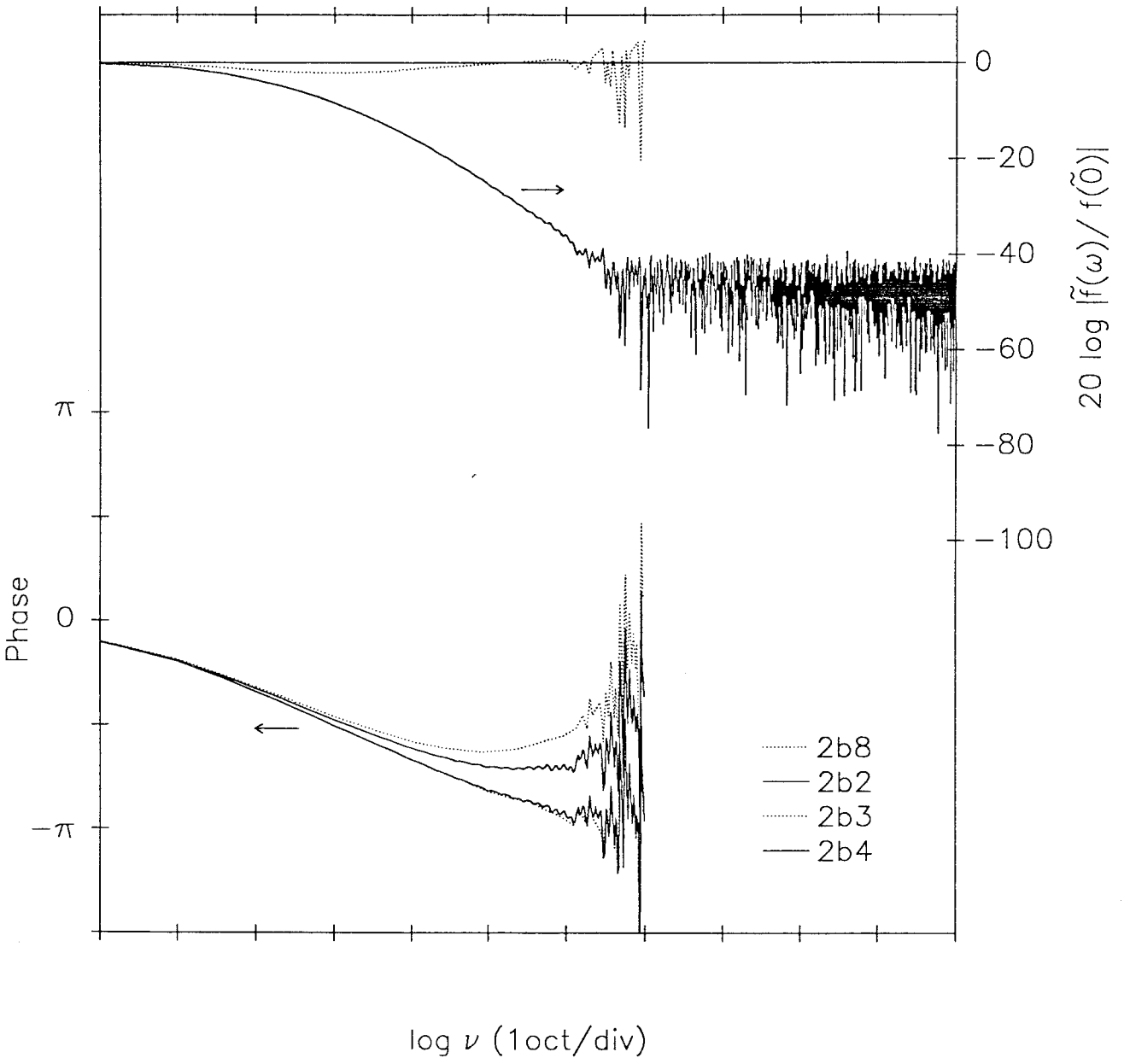


Fig. 6.8 (Top) Frequency spectrum of f_B (solid line) and "localized" spectrum after normalizing out high frequency behavior. Apparent noise level at -40 dB is due to round-off of the autocorrelation. (Bottom) Retrieved phase using four different approaches: (B8) the full spectrum is truncated at the onset of noise and only those frequencies are used in the double FFT algorithm of Eqn. (6.3.1); (B2) the entire spectrum is used in the double FFT algorithm; (B3) use normalization scheme and double FFT on localized spectrum truncated at $\nu = 2^7$; (B4) same as B3 except set $\log |\tilde{f}_{local}| = 0$ for $\nu > 2^7$ instead of truncating. There is very little difference between the results of these last two.

the spectrum would be flat and the minimum phase solution is $\phi_m(\omega) = 0$, which results in a delta function at $t = 0$. When only a finite portion of the frequency components are used and the spectrum varies around a flat average, then a finite spike results with additional noisiness for $t > 0$. This explanation is validated by the reduction of the spike and noisiness when the frequency spectrum is truncated prior to inverse transforming in the final step.

Divergence of Integrand and FFT Problems

In deriving the minimal phase integral relation, Eqn. (6.2.11), the issue of the divergence of $\log|\tilde{f}(\omega)|$ for large $|\omega|$ was shown to be irrelevant if $f(t)$ is real. However, in numerically evaluating the integral, it does become a problem since we are typically restricted to a finite frequency domain. The integral relation for ϕ_m implies that the phase at a particular frequency is a function of the behavior of $|\tilde{f}(\omega)|$ over all frequencies. However, the evaluation of this integral must be restricted to a finite domain due to computation limits or the temporal resolution of the autocorrelation which gives an upper frequency limit when the FFT is evaluated. This is somewhat alleviated by the local nature of the factor $1/(\omega - \omega')$ which gives higher weighting to the amplitude behavior near the frequency of interest. Nevertheless, the slow decay of this factor does not make it strongly local and far removed behavior can be significant particularly if it is diverging as $\log|\tilde{f}(\omega)|$ does. This effect will be especially noticeable at high frequencies approaching the point at which the spectrum is truncated. The use of FFT's to evaluate the modified convolution expression of Eqn. (6.3.13) further compounds this problem. The discrete nature of the FFT means that Fourier series and not transforms are actually evaluated, and the result of evaluating Eqn. (6.3.13) using FFT's is the convolution of $1/\omega$ with the frequency spectrum infinitely repeated with a period equal to twice the maximum frequency of the spectrum. This repeated function is not only symmetric

about $\omega = 0$, but also about $\pm\omega_{max}$. Thus the convolution with the odd function $1/\omega$ will return to zero at $\pm\omega_{max}$ which is clearly not correct.

This effect is illustrated in Fig. 6.8 where the retrieved phase is shown for f_B, and should asymptotically approach $-\pi$ at large frequencies. The middle curve is the result of applying the algorithm to the complete spectrum and shows the effect of the "flat" noise spectrum which tends to pull the phase toward zero at frequencies near where the noise becomes dominant. The upper curve shows the result when the spectrum is truncated before the phase is retrieved and clearly shows the FFT effect whereby the phase returns to zero at the maximum frequency of the spectrum.

An approach to reduce these problems is to factor out the asymptotic high/low frequency behaviors leaving only a localized function to be convolved using the FFT algorithm of Eqn. (6.3.1). Assuming the high frequency behavior to be asymptotic to $(\omega/\omega_0)^{-m}$, we factor the spectrum as follows:

$$|\tilde{f}(\omega)| = [1 + (\omega/\omega_0)^2]^{-m/2} |\tilde{f}(0)| |\tilde{f}_{local}(\omega)|, \tag{6.3.2}$$

where $\tilde{f}_{local}(\omega)$ is now a localized function which goes to zero at high frequencies. The desired phase is then the sum of the minimal phases of each of the three factors in Eqn. (6.3.2) and the constant factor gives zero. The causal transform corresponding to the asymptotic factor is given analytically as

$$\tilde{f}_{asym}(\omega) = \frac{1}{(1 + j\omega/\omega_0)^m} \tag{6.3.3}$$

and so its phase is simply

$$\phi_{asym}(\omega) = -m \tan^{-1}(\omega/\omega_0). \tag{6.3.4}$$

The phase of the remaining term, $\tilde{f}_{local}$, is now evaluated by the numerical integration algorithm and the high frequency problems are now greatly reduced since

$\log|\tilde{f}_{local}|$ goes to zero at high frequencies. The term $\log|\tilde{f}_{local}|$ for f_B is shown in Fig. 6.8 along with the frequency spectrum illustrating its suitability for use in the numerical convolution algorithm. Also, the retrieved phase using this approach is shown and clearly illustrates the improvement over the nonfactored approach.

The parameters ω_0 and m in the asymptotic factor of Eqn. (6.3.3) are found by best-fitting a line to the high frequency data in the Bode plot — before the onset of noise — and finding its intercept with the $|\tilde{f}(0)|$ value. Although the high frequency phase tends to $-m\pi/2$ and thus an accurate value for m would seem important, the retrieved signal was found to be fairly insensitive to the actual value used. For f_B, a variation in m from 1.6 to 2.4 produced only a slight advance ($m < 2$) or delay ($m > 2$) in the position of the pulse with very little change in its shape. In fact the nonfactored results corresponding to the poor phase curves of Fig. 6.8 were actually quite good with regard to retrieved pulse shape, and were just advanced in time. This can be attributed to the small amplitude of the spectrum at regions where the phase is poorly evaluated.

Finally, we consider using the normalization scheme to extrapolate the spectrum to high frequencies where the true spectrum has been lost in the noise. It is not suggested that the actual spectrum can be retrieved from the noise oblivion. In principle, the spectrum can take on virtually any shape below the noise level and such features must be regarded as lost. However, for the purposes of smoothly interpolating between the wider spaced time points that a truncated spectrum forces, it seems preferable to assume that the lost portion of the spectrum decays continuously along the asymptotic, pre-noise, behavior rather than just disregarding it altogether. This is supported by results from integral function theory: for temporal functions of finite amplitude and domain, the frequency spectrum is asymptotically *limited* by a finite power of ω at high frequencies [12].

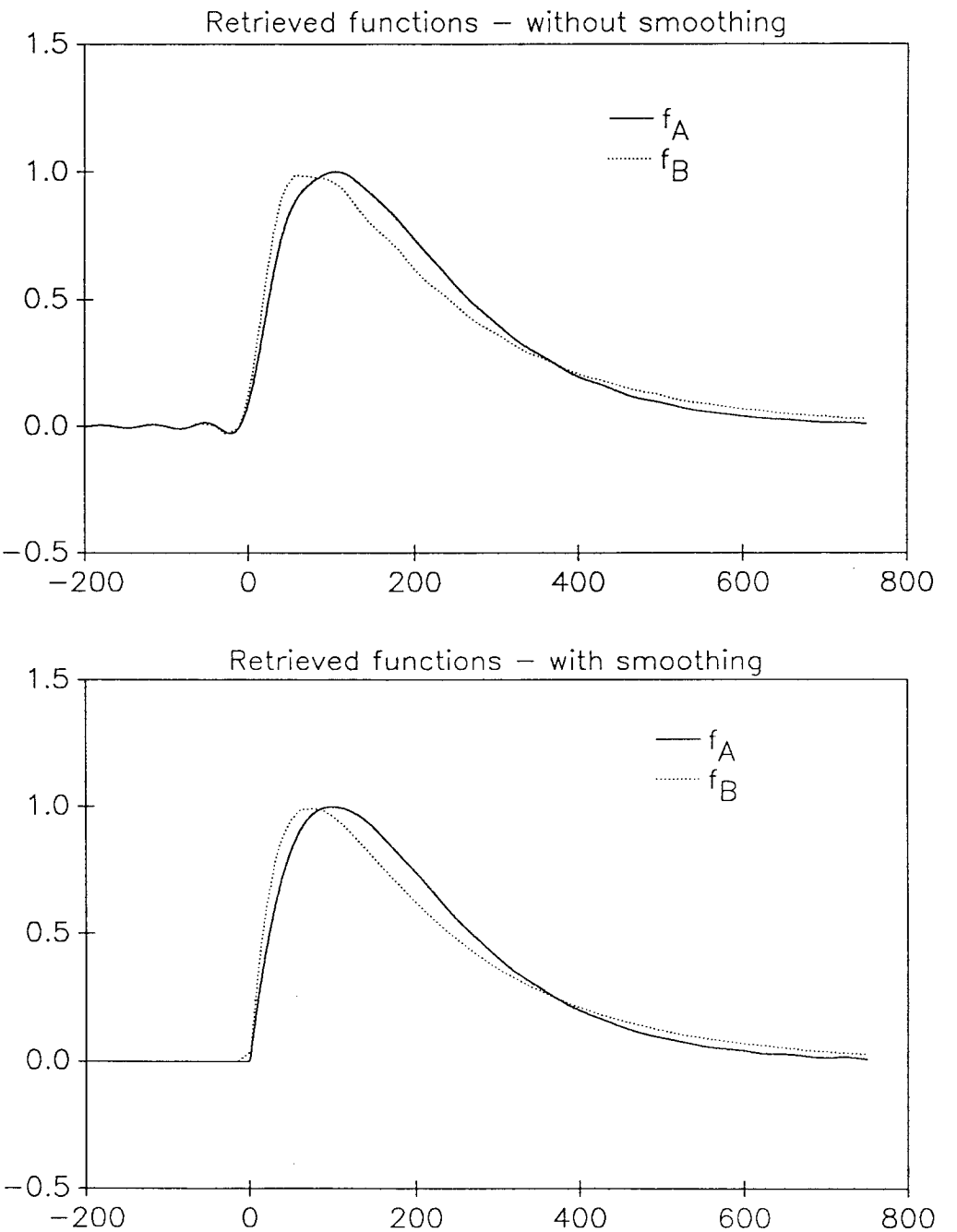


Fig. 6.9 Comparison of two interpolation approaches when high frequency portion of spectrum is truncated due to noise: (Top) Set $\tilde{f}(\omega) = 0$ for frequencies $\nu > 2^6$ and (Bottom) Use asymptotic normalization and set $\tilde{f}(\omega) = \tilde{f}_{asym}(\omega)$ for $\nu > 2^6$.

A comparison of these two interpolation approaches is shown in Fig. 6.9, which also demonstrates that the retrieval quality is sufficient to distinguish these two similar pulse shapes. It is seen that when the spectrum above the cut-off point is taken as zero, then the result has small undulations; while the result using the asymptotic decay as a high frequency extrapolation is quite smooth. The excellent agreement of the latter with the original functions should not be overrated since the test functions used are particularly well suited to this scheme. The undulations of the former approach are actually quite small if compared directly against the original function and a conclusive comparison of the two results is somewhat a matter of aesthetics. Of course there is a danger that the undulations could be interpreted as real; but the converse argument can be made that the smoothed result may imply the smoothness to be true beyond the actual resolution of the measurement. In short, the chosen conditions for arriving at any of the results should be carefully stated to avoid potential misinterpretation. It is also noted that the differences in the two approaches would be more pronounced if the known spectrum is limited to a lower value as would be the case with a higher noise level.

References

[1] See for example: *Ultrashort Light Pulses*, S. L. Shapiro, ed., New York: Springer-Verlag, 1988; *Ultrashort Laser Pulses*, W. Kaiser, ed., New York: Springer-Verlag, 1988.

[2] C. L. Tang and D. J. Erskine, Phys. Rev. Lett., **51**, 840 (1983).

[3] D. H. Auston, A. M. Johnson, P. R. Smith, and J. C. Bean, Appl. Phys. Lett., **37**, 371 (1980).

[4] J. A. Valdmanis, G. Mourou, and C. W. Gabel, Appl. Phys. Lett., **41**, 211 (1982).

[5] B. H. Kolner, D. M. Bloom, P. S. Cross, Electron. Lett., **19**, 574 (1983).

[6] E. P. Ippen and C. V. Shank, in *Ultrashort light pulses*, S. L. Shapiro, ed., New York: Springer-Verlag, 1977, pp. 83-122.

[7] A. Walther, Opt. Acta, **10**, 41 (1962).

[8] G. I. King, Acta Cryst., **A31**, 130 (1975).

[9] D. Kohler and L. Mandel, Jour. Opt. Soc. Amer. **63**, 126 (1973).

[10] See for example: "OSA Topical Meeting on Signal Recovery and Synthesis with Incomplete Information and Partial Constraints," J. Opt. Soc. Am. **73**, 1412 (1983); and "OSA Topical Meeting on Signal Recovery and Synthesis II," J. Opt. Soc. Am. A **4**, 105 (1987).

[11] J. R. Fienup, Appl. Optics **21**, 2758 (1982).

[12] R. E. Burge, M. A. Fiddy, A. H. Greenaway, and G. Ross, Proc. R. Soc. Lond. A. **350**, 191 (1976).

[13] N. Nakajima, Optics Lett. **11**, 600 (1986).

[14] H. M. Nussenzveig, Jour. Math. Phys. **8**, 561 (1967)

[15] E. C. Titchmarsh, Proc. Lond. Math. Soc. **25**, 283 (1925).

[16] B. J. Hoenders, Jour. Math. Phys. **16**, 1719 (1975).

[17] R. W. Gerchberg and W. O. Saxton, Optik **35**, 237 (1972).

[18] J. R. Fienup, Opt. Lett. **3**, 27 (1978).

[19] K. Chalasinska-Macukow and H. H. Arsenault, J. Opt. Soc. Am. A **2**, 46 (1985).

[20] R. H. T. Bates and W. R. Fright, J. Opt. Soc. Am. **73**, 358 (1983).

[21] G. Pólya, Math. Zeitschrift **2**, 352 (1918).

[22] E. S. Kuh, D. O. Pederson, *Principles of Circuit Synthesis.* New York: McGraw-Hill, 1959.

[23] M. L. Cartwright, Quart. J. Math. (Oxford) **1**, 38 (1930).

[24] M. L. Cartwright, Quart. J. Math. (Oxford) **2**, 113 (1931).