General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

First Quarterly Report
for
Contract No.: NAS5 - 11542

(1 June, 1968 - 31 August, 1968)

Prepared by

V. E. Suomi and T. H. Vonder Haar
Space Science and Engineering Center
The University of Wisconsin

For

Goddard Space Flight Center
Greenbelt, Maryland

Contracting Officer: Mr. Newchy Mignone
Technical Monitor : Mr. William Bandeen

Summary

Work in (a) Balloon Radio Altimetry (b) IR Imaging and Sounding and (c) Radio Occultation Techniques has been started. All of these efforts are directed to the use of satellite sensors and systems for inferring atmospheric structure. Some detailed results are given in Appendixes. Research will continue in each of these areas. In addition, the application of global satellite radiation data and measurements from geosynchronous satellites to meteorological problems has been proposed in an extension to the contract.

1. Introduction

This Type II Quarterly Report covers the period 1 June - 31 August, 1968. During this time, work in both contract research areas was accomplished. The main areas are: (a) Development of a Miniature Radio Altimetry Sensor for Super-Pressure Balloons and (b) Investigation of Infrared Imaging from Synchronous Satellites.

An extensive description of the first task is included in Appendix I. In the second area, preliminary studies of porposed instrumental characteristics and techniques are underway. In addition, related research on radio occultation techniques for inferring atmospheric density profiles was supported by this contract.

2. Discussion

a) Balloon Radio Altimeter

In order to study cross-isobaric flow, kinetic energy budgets and the deepening or filling of pressure systems from constant density balloon . a simultaneous measure of pressure and geometric altitudes is required. We a e developing a light-weight radio altimeter with an accuracy of ± 30 feet to meet these needs.

Appendix I gives details of the system which will be eventually interfaced with the IRLS system for satellite interrogation.

b) IR Imaging Studies

Our investigations of infrared imaging (and sounding) from geosynchronous altitude covers a wide range of possible sensors and techniques. At the present

time, prime emphasis is placed on the evaluation of proposed instrumentation and techniques for a sensor system having both imaging (IR and visible) and sounding (for vertical temperature profile) capabilities. Very little of this effort has advanced beyond the preliminary stage. One supporting project, A Survey of Noise Generation in Solid State Infrared Detectors, is attached as Appendix II.

Feasibility studies of multiple - array IR sensors have not advanced beyond the planning stage, and no large effort will be made in this area without further study and discussion.

c) Radio Occultation Techniques

During the reporting period, Bruce Lusignan of Stanford University was a visiting scientist at Space Science and Engineering Center. He worked on ray tracing as an inversion method for refractivity assuming initial profiles of refractivity vs. height. A more extensive account of his work will be included in a subsequent report.

In early September the SSEC was host to the Meteorological Satellite Radio Occultation Technique Working Group (headed by Dr. William Nordberg, GSFC). As a result of that meeting, some additional work in this area will be supported by the present contract and will be directed by Dr. Douglas Sargent.

3. New Technology

No "New Technology" was developed during the period of this report.

4. Program for Next Reporting Interval

Work will continue in:

a) Balloon Radio Altimetry (IRLS interface)

b) IR imaging and sounding (study of optical system and observation bands for ISIR (Imaging and Sounding Infrared Radiometer))

c) Radio Occultation Techniques (ray path calculations under realistic atmospheric conditions)

d) areas proposed in recent contract extension (use of global radiation data; application of geosynchronous satellite data)

5. Conclusions and Reccomendations

All contract work areas are receiving attention and short-term goals are now welldefined. Aside from the balloon altimeter project which has very specific objectives, all work areas remain flexible enough to respond to new scientific input. The Radio Occultation Working Group results provide a good example of this and we shall continue to participate in similar sessions that are relevant to contract tasks.

APPENDIX I

BALLOON-BORNE RADIO ALTIMETER

Progress Report #1
Under Contract NAS5-11542

Nadav Levanon
Project Coordinator

Space Science and Engineering Center

The University of Wisconsin

September 1968

Contents

1. Introduction

The radiosonde, an instrument package elevated to high altitudes in the atmosphere by a buoyant balloon, is the most wide spread means of obtaining vertical profiles of atmospheric parameters. These parameters are: Temperature, pressure, relative humidity, and wind velocity, whose changes with altitude are sought by the meteorologist. Sensors for the first three parameters are mounted in the balloon package, and their readings are transmitted to the ground. The wind is studied by tracking the balloon with optical or radio theodolites. The only parameter that is not measured directly is the altitude. However, we can allow one parameter not to be measured, because for thin layers of the atmosphere below 90 km, the following relation exists:[1]

$$\frac{P}{P_b} = EXP\left[-\frac{g_0 M h 2}{R(T+T_b)}\right] \qquad (1.1)$$

where:
- P - the pressure at the top of the layer
- P_b - the pressure at the base of the layer
- T - the temperature at the top of the layer
- T_b - the temperature at the base of the layer
- g_0 - acceleration due to gravity, at sea surface
- M - molecular weight of the air
- R - universal gas constant
- h - the thickness of the layer (in geopotential meters)

From 1.1 we see that once the pressure and temperature at a certain altitude (usually the surface)are known, the height measurement is not needed to get the vertical profile, as long as the temperature and pressure are measured at close intervals.

[1] U.S. Standard Atmosphere, 1962.

However, at high altitudes, these parameters are hard to measure accurately. For example, the barometer readings are in steps of 1 millibar, which are 0.1% of the total pressure at the surface, but 10% at an altitude of 30 km. The temperature readings are also less accurate at high altitude due to stronger solar radiation. A direct measurement of altitude at this range, could improve the accuracy of the vertical profile.

The meteorological community is devoting much effort to several programs of Global Weather Observations. Playing major roles in these programs are superpressure balloons, carrying their sensors at constant density altitude and floating with the winds for several months. They will float most of the time over the ocean where no ground stations exist to give the necessary reference reading of surface pressure and temperature. For these balloons a radio altimeter and a simple pressure measurement can supply such a reference.

The simultaneous reading of pressure and geometric altitude is most useful when we wish to study an unbalanced atmosphere. For example, the balloon in a balanced atmosphere will follow the wind along isobar lines. Therefore, constant pressure readings will be accompanied by constant altitude readings, but when the geometric altitude changes separately from the pressure altitude, that means that the wind crosses isobars, and a weather disturbance is growing or filling.

The above are just few examples of the great need for a radio altimeter on board the balloon package. However, conventional radio altimeters are too heavy for balloons. It seems that just

using modern circuit techniques like integrated circuitry would not be enough in reducing weight, size and price; and that, in addition, a more simple system is needed. It should be noted that the balloon offers one big advantage, and that is, its low rate of altitude changes compared to rockets, for example. This leaves plenty of time for averaging which can replace strong signals, or good noise figure. This advantage is fully exploited in the proposed balloon-borne radio altimeter.

The preliminary specifications outlined for this altimeter call for a radio altimeter that should cover the altitude range between 20,000 to 100,000 feet over the ocean, with an accuracy of $\pm$ 30 feet, whose weight should not exceed one pound.

This research is carried under NAS5-11542, with the final goal of interfacing the altimeter with the Interrogation, Recording, and Location Subsystem (IRLS).

2. Principle of Operation

The proposed radio altimeter uses a single superregenerative stage[1] serving as both the receiver and the transmitter. This method was used for beacons in the early days of radar. This superregenerative stage is used in a feedback system in such a way that the period between transmitted pulses is a measure of altitude. This section describes this operation in more detail.

The superregenerative stage, when triggered by the quench pulses, produces r.f. pulses whose envelope area depends on the r.f. signal existing at the input to the stage at the beginning of each quench pulse.

For no signal, the noise is the only input. When there is a signal, and when the period of the quenching pulses is equal to the delay of the return pulse, the superregenerative stage detects its own pulses. (Fig. 2.1) As we change the period (or repetition rate) of the quench pulses, we get a maximum output from the superregenerative detector (Fig. 2.2) near the repetition rate:

$$F_q = \frac{c}{2h} \qquad (2.1)$$

where: F_q - the repetition rate of the quench pulses
c - velocity of light
h - the altitude

Varying the repetition rate around this value is equivalent to scanning the returned pulse. As we will see later, the returned pulse is close to an imperfect integral of the transmitted pulse, and, therefore, has a defined peak at the end of the transmitted pulse.

The rest of the circuitry is a servo loop designed to lock on that peak. This is an example of a "single-dimensional sinusoidal perturbation adaptive system."[2]

---

[1] G. O. Hall: Superregenerative Receivers, in "Microwave Receivers" Radiation Lab. Series, Ch. 20.

[2] V. W. Eveleigh: "Adaptive Control and Optimization Techniques," McGraw Hill, 1967, Ch. 9.

The principle of this system (Fig. 2.3) is that sinusoidal perturbations serve to take the derivative of the curve. This means that at the peak we will get zero signal, while off the peak there will be some slope, either positive or negative. This type of signal can be used for an error signal in a servo loop. The block diagram of the system is shown in Fig. 2.4.

The Voltage Controlled Oscillator (VCO) determines the repetition rate of the quench pulses. This rate is linearly related to the voltage at the output of the integrator and is sinusoidly perturbed by the 200 H_z signal. In other words the quench frequency is frequency modulated by 200 H_z; the deviation of this FM is very small. The curve of the output of the superregenerative detector as function of the quench frequency serves as an FM detector for the 200 H_z signal. When the quench frequency is too high, the detected 200 H_z signal is out of phase, compared to the modulating 200 H_z signal; when f_q is too low, the 200 H_z signal is in phase; and when it is at the right frequency, we get only the second harmonics of the 200 H_z.

The output of the detector is passed through a narrow band amplifier centered around 200 H_z and then multiplied by the modulating signal. The output of this multiplier contains a positive d.c. voltage when the two signals are in phase, and a negative d.c. voltage when they are out of phase. The integrator output is affected by this d.c. voltage, and locks the VCO to the frequency which gives the peak at the output of the superregenerative detector.

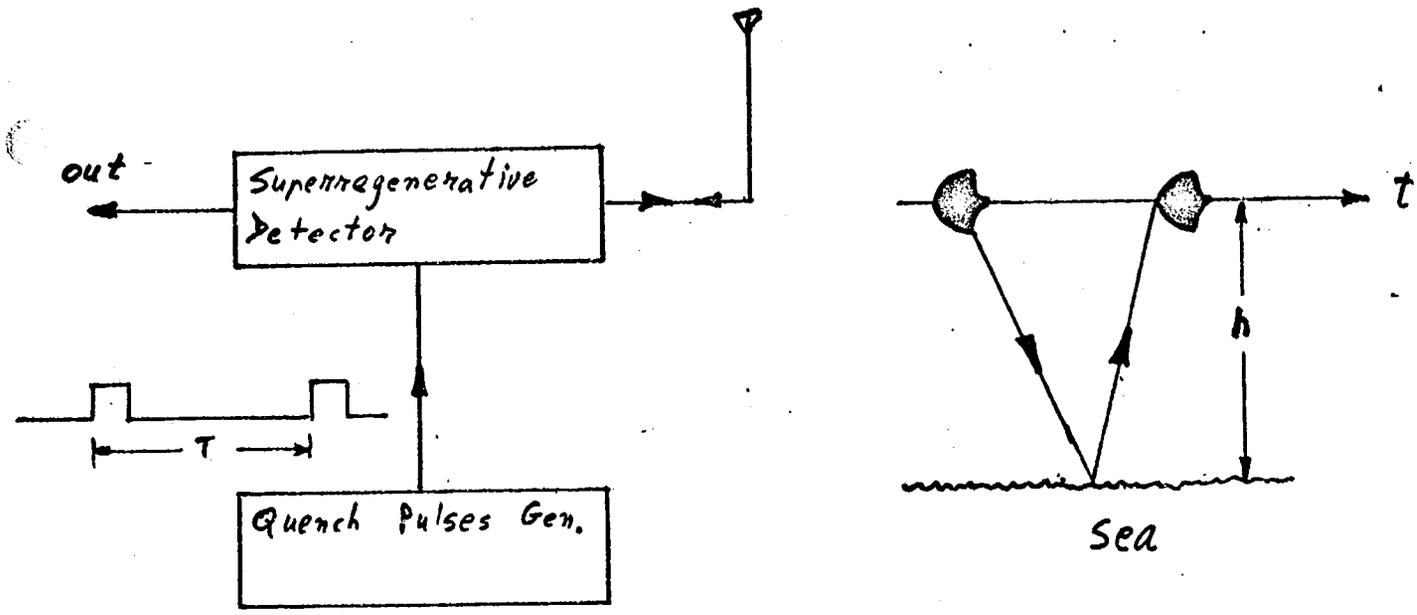


Fig. 2.1 The superregenerative detector receives its own pulses after they are returned from the sea surface.

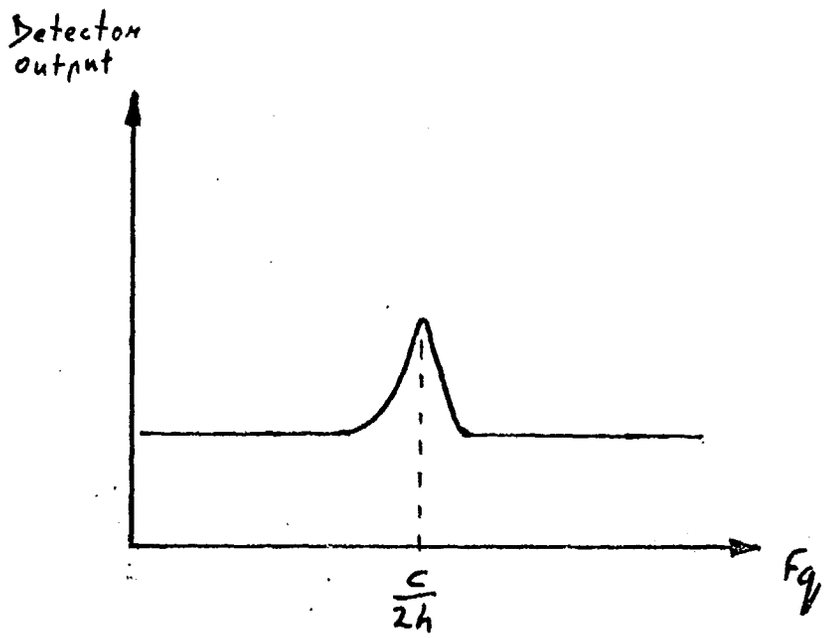


Fig.2.2 The output of the superregenerative detector peaks when the period of the quench pulses equals the delay of the returned pulse.

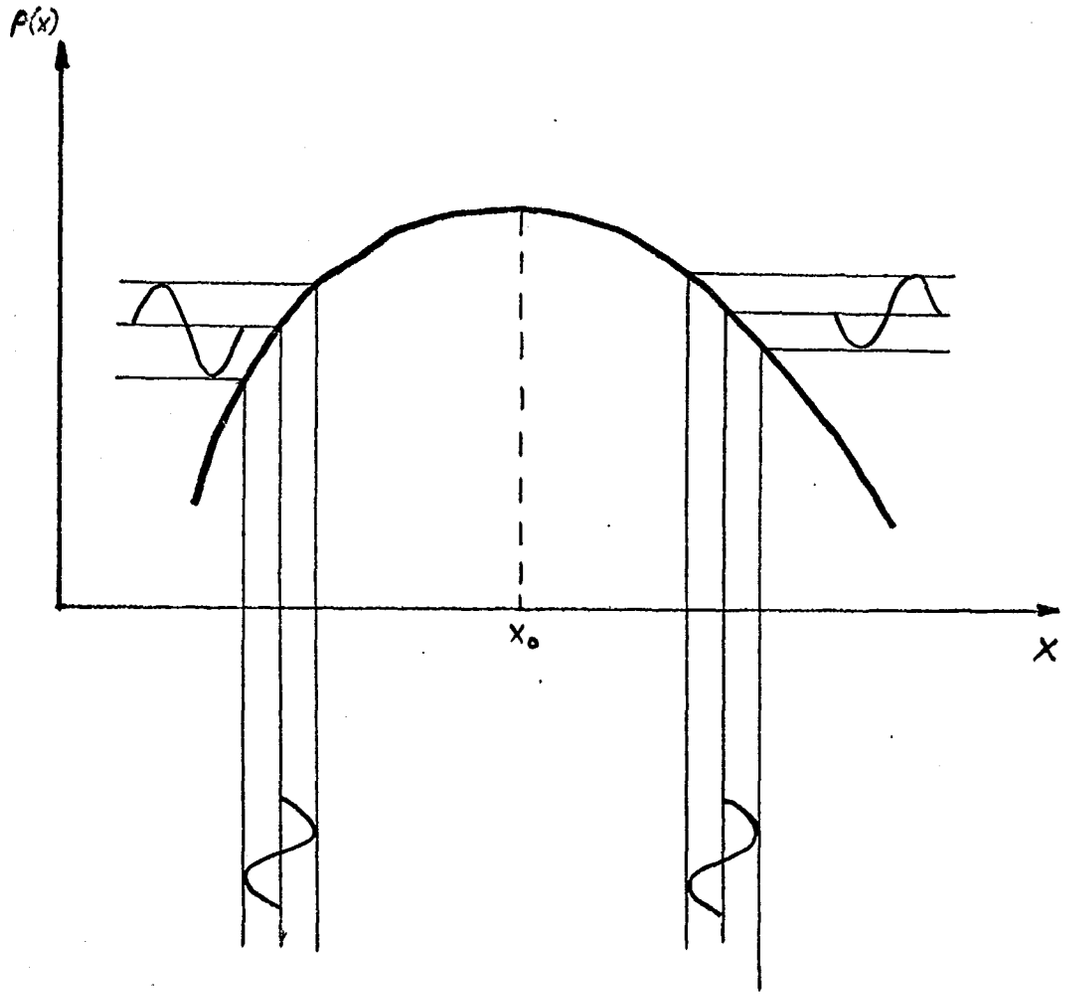


Fig. 2.3 The effect of sinusoidal parameter perturbation upon a single-dimensional even-function curve.

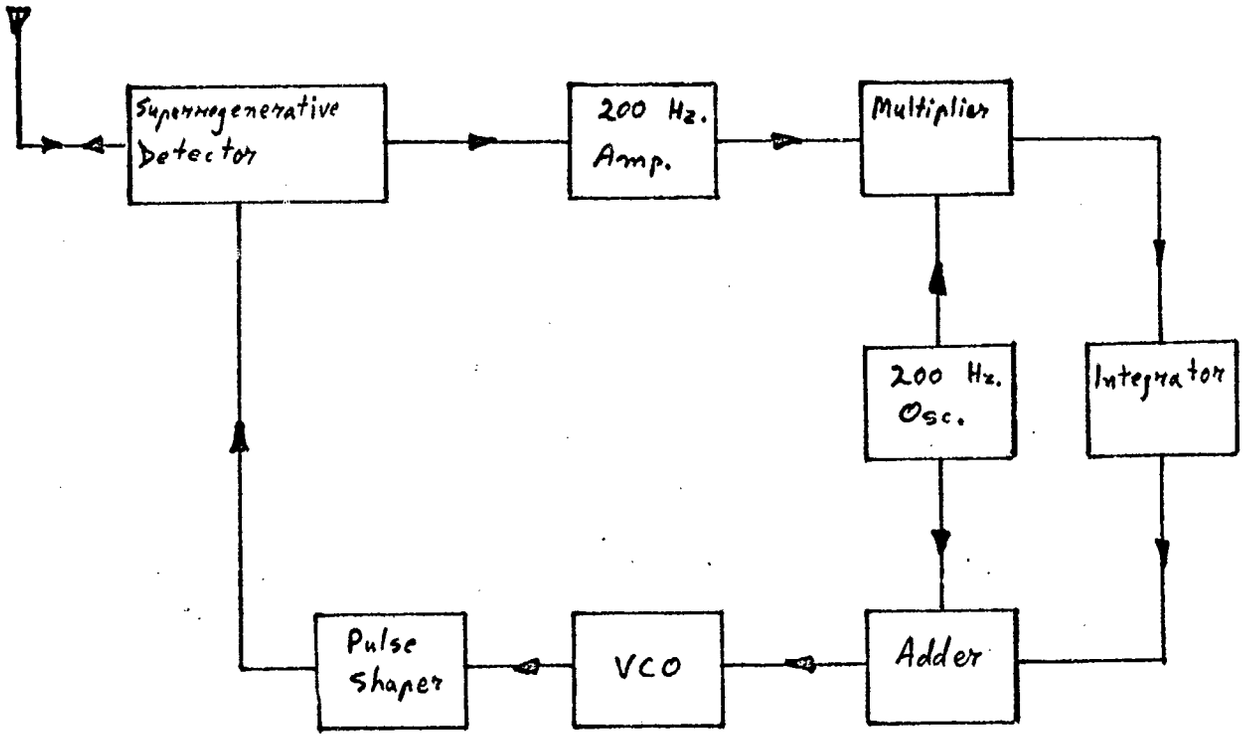


Fig. 2.4 Block diagram of the altimeter

3. The Superregenerative Detector

The fundamental action in a superregenerative circuit is centered around the growth of oscillations in an oscillator. The variation in grid voltage is shown in Fig. 3.1. Between positive pulses, the grid is at the negative potential - E which biases the tube beyond E_{co}, the cut off voltage. During the positive pulses the grid potential is raised to a point where the circuit oscillates, as indicated by projection into the shaded oscillation region. Thus, oscillation grows during the A periods and decaysduring the B periods.

There are two modes of operation defined for superregeneration, a linear mode (Fig. 3.1.b) and a logarithmic mode (Fig. 3.1.c). The linear mode results when the positive quench period A is so short that the oscillations do not have time to build to full saturation amplitude. The logarithmic mode occurs when the A period is sufficiently long to allow oscillations to build to full amplitude before the end of the A period. In our case the circuit operates also as the transmitter, and we need as much output power as possible. Therefore, we would like to operate in the logarithmic mode.

Before saturation is reached, the envelope of the oscillations will rise as,

$$e_c = V_o e^{\frac{Rt}{2L}} \qquad (3.1)$$

where: V_o - the input voltage when period A begins
R - the total negative resistance of the circuit
L - the circuit inductance

When there is no input signal, the input is the noise voltage, V_n. When a signal voltage, V_s, is present, the voltage at the start of the oscillations will be $V_s + V_n$ and Eq. (3.1) becomes,

$$e_c = (V_s + V_n) e^{\frac{Rt}{2L}} \qquad (3.2)$$

As can be seen in Fig. 3.2, a higher initial voltage would cause a shorter rise time period before saturation, and as a result a bigger pulse area. This additional area is added with each additional quench pulse, and the total change appears at the output of the detector. The change in average voltage of the detected output ΔV, assuming linear detection, is given by,[1]

$$\Delta V = F_q \, E_{max} \frac{2L}{R} \ln \frac{V_s + V_n}{V_n} \qquad (3.3)$$

This change in output divided by the change in the input gives the gain of the stage. Assume, for example, that the signal is equal to the noise; then the change in the input is equal to the noise which is a few microvolts. The change in the output, however, will be:

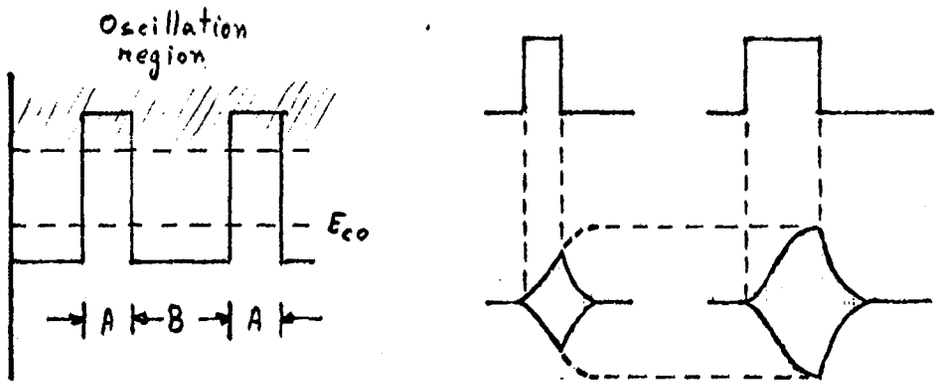
$$\Delta V = F_q \, E_{max} \frac{2L}{R} \ln 2 \qquad (3.4)$$

which is of the order of volts. That means that a gain of a million can be achieved in this single stage.

It is for both the high receiver gain and the doubling of the single stage as a transmitter and receiver, that the superregenerative stage has been chosen for a small, light, and accurate balloon-borne radio altimeter.

---

[1]G. O. Hall, Superregenerative receivers in "Microwave Receivers," Radiation Lab. Series, Ch. 20.



(a) (b) (c)

Quench voltage Modes of oscillation

Fig. 3.1 Quench voltage and relation to modes of oscillation

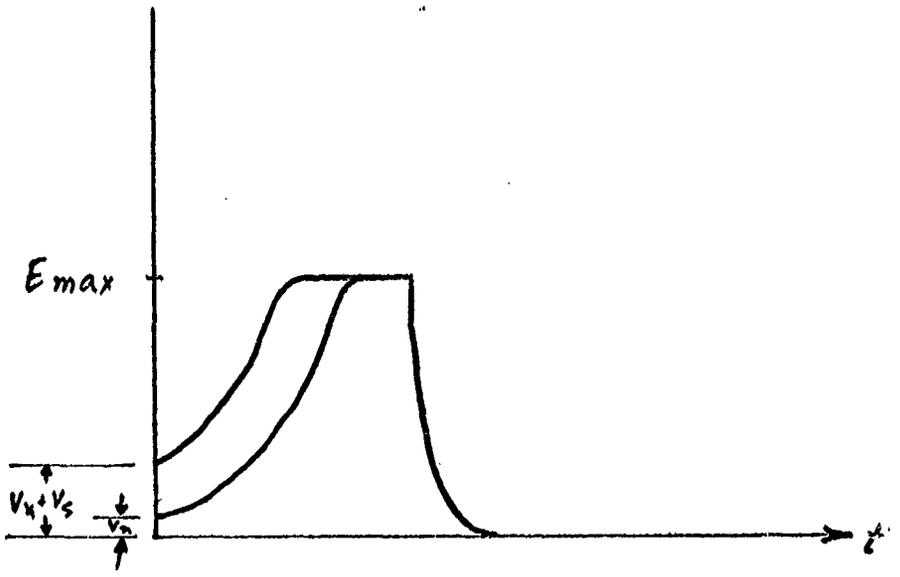


Fig. 3.2 The change in pulse area in the presence of signal

4. The Returned Pulse

The ocean surface plays the role of an imperfect integrator on the returned pulse. Let us assume, for a moment, that the three following conditions exist:

1. The average scattering cross section of the ocean is independent of the incident angle.
2. The altimeter antenna is isotropic.
3. The additional attenuation due to a longer pass of a side reflection compared to a downward reflection can be neglected.

With these three assumptions, we will follow the pulse on its way to the ocean surface and back. The first reflection occurs when the front end of the pulse reaches the point B (Fig. 4.1) just below the altimeter. Side returns like that from point C will start later, because of the additional distance, and will reach the altimeter even later because the returning pulse also has the additional distance to traverse. While the front edge of the pulse is being reflected from point C, latter portions of the pulse reach point B, and start to be reflected.

In other words, the pulse can be seen as traveling on the ocean surface. When the trailing edge is being reflected from point B, the whole pulse is reflected to the altimeter, and the intensity of the return reaches a maximum. With the above three assumptions, this level will remain constant as the pulse on the ground travels away from the altimeter sub-point.

Measurements have demonstrated[1-2] that, with occasional exceptions, a radar return from the ground is largely due to area scatter. For this reason, no phase relations were introduced, and when we speak of the returned pulse, we are referring to the mean pulse envelope.

---

[1]R. K. Moore and C. S. Williams, Jr.: Radar Terrain Return at Near-Vertical Incidence. Proceeding of the IRE, Feb. 1957. pp. 228-238.

[2]A. R. Edison, R. K. Moore and B. D. Warner: Radar Terrain Return Measured at Near-Vertical Incidence. IRE Transaction on Antenna's and Propagation. May, 1960. pp. 246-254.

As we mentioned above, the maximum of the returned pulse occurs at the end of the transmitted pulse (if we ignore the delay.) If we now remove the assumption that the antenna is isotropic, and, instead, assume that it has constant gain for $|\Theta| < \Theta_0$, and zero gain outside, we can calculate the minimum Θ_0 that still allows the peak to occur at the end of the transmitted pulse. From Fig. 4.1 we see that the return from point C has additional delay, τ , which is, if we assume a flat surface,

$$\tau = \frac{2h}{c}(\sec\Theta_0 - 1) \qquad (4.1)$$

where: h - the altitude
c - the velocity of light

If the length of the transmitted pulse is shorter than τ , then the peak of the return pulse occurs at the end of the transmitted pulse, which does not change with altitude. This is called "pulse-length-limited illumination". If the length of the transmitted pulse is larger than τ , the peak will occur at τ which is dependent on altitude. This is called "beam-width-limited illumination."

Our system locks on the peak of the returned pulse, and we would prefer the additional delay to be constant, and, therefore, use pulse-length-limited illumination. The condition for such illumination is that:

$$\sec\Theta_0 \geq \frac{c\,\tau_0}{2h_{min}} + 1 \qquad (4.2)$$

where: τ_0 - the length of the pulse
h_{min} - the minimum altitude

When we include the dependence of the average scattering cross section of the ocean on the angle of incidence, the additional attenuation due to additional distance, and the fact that the antenna pattern is not constant as function of θ, the returned signal can be expressed as the convolution of the transmitted pulse form in power units, with a function which includes effects of antenna pattern, ground properties and distance.

Moore, et. al.,[1] showed that for a flat earth and a system which is independent of azimuth orientation, the mean received pulse envelope in power units is given by:

$$P_R(d) = \frac{\lambda^2}{8(2\pi)^2} \int_h^{\frac{cd}{2}} \frac{1}{r^3} P_T\left(d - \frac{2r}{c}\right) G^2(\theta)\, \sigma_0(\theta)\, dr \qquad (4.3)$$

where: $r(=h\sec\theta)$ - the distance from the altimeter to the points of reflection

h - the altitude

θ - the angle from the vertical

$d(>\frac{2h}{c})$ - the time from the leading edge of the transmitted pulse

λ - the wavelength of transmitted energy

$G(\theta)$ - the antenna gain

$\sigma_0(\theta)$ - the average scattering cross section of the ocean.

$P_T(\ )$ - the envelope of the transmitted pulse in power units

$P_R(\ )$ - the envelope of the received pulse in power units

Moore, et. al.[1] suggests the inclusion of both the antenna pattern and the ground scattering pattern in one equation,

$$G^2(\theta)\sigma_0(\theta) = G^2\sigma_0 \cos^6\theta = G^2\sigma_0\left(\frac{h}{r}\right)^6 \qquad (4.4)$$

which is a good approximation in many cases.

Inserting (4) in (3) we get:

$$P_R(d) \approx \frac{\lambda^2 G^2 \sigma_0 h^6}{8(2\pi)^2} \int_h^{\frac{cd}{2}} \frac{1}{r^9} P_T\left(d - \frac{2r}{c}\right) dr \qquad (4.5)$$

Further approximations lead to:

$$P_R\left(t + \frac{2h}{c}\right) \approx \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} \int_0^t P_T(t-\tau)\left(1 - \frac{gc}{2h}\tau\right) d\tau \qquad (4.6)$$

where t is the time from the reception of the leading edge.

In Fig. 4.2 a typical superregenerative pulse is used as the transmitted pulse, P_T. When this envelope is used in Eq. (4.6), the envelope of the received pulse is as shown in Fig. 4.3. The envelope can be described by three lines.

I. $\quad 0 < t < t_1 = \frac{1}{\alpha} \ln \frac{P_{max}}{P_0}$

$$P_R\left(t + \frac{2h}{c}\right) = \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} \frac{P_0}{\alpha}\left[e^{\alpha t} - 1 - \frac{gc}{2h}\left(\frac{e^{\alpha t}}{\alpha} - t - \frac{1}{\alpha}\right)\right] \qquad (4.7)$$

$$P_1 = \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} \frac{P_{max}}{\alpha}\left(1 - \frac{gc}{2h\alpha}\right) \qquad (4.8)$$

II. $t_1 < t < T$

$$P_R\left(t+\frac{2h}{c}\right) \approx P_1 + \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} P_{max}(t-t_1)\left[1 - \frac{gc}{4h}(t-t_1) - \frac{gc}{2h\alpha}\right] \qquad (4.9)$$

$$P_2 \approx P_1 + \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} P_{max}(T-t_1)\left[1 - \frac{gc}{2h}\left(\frac{1}{\alpha} + \frac{T-T_1}{2}\right)\right] \qquad (4.10)$$

III. $T < t \ll \frac{2h}{c}$

$$P_R\left(t+\frac{2h}{c}\right) \approx P_2 - \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} P_{max}\left\{(t-T)\frac{gc}{2h}\left(\frac{1}{\alpha}+\frac{1}{B}+T-t_1\right) - \right.$$

$$\left. - \frac{1}{B}\left[\left(1+\frac{gc}{2h\beta}\right)\left(1-e^{-\beta(t-T)}\right)\right]\right\} \qquad (4.11)$$

For pulses of length in the order of $1\mu s$, the convolution integral can be approximated by a regular integral, and, this will give a simpler expression for the peak of the received pulse P_2.

$$P_2 = \frac{\lambda^2 G^2 \sigma_0 c}{16(2\pi)^2 h^3} P_{max}\left[T - \frac{1}{\alpha}\left(\ln\frac{P_{max}}{P_0} - 1\right)\right] \qquad (4.12)$$

For the following typical values:

$P_{max} = 1$ watt

$P_0 = 2.5\ 10^{-12}$ Watt $(12\mu V, 50\Omega)$

$\alpha = 5\ 10^{7}$ sec^{-1}

$\beta = 2.5\ 10^{8}$ sec^{-1}

$T = 1.5\ 10^{-6}$ sec

$\sigma_0 = 0.032$

$h = 2 \cdot 10^{4}$ m (60000 Ft.)

$G = 40$ (16 db)

$\lambda = 1$ m

We will get $P_2 = P_0 = 2.5\ 10^{-12}$ Watt

In the above discussion we have calculated the mean received pulse envelope. The pulses themselves fluctuate around this mean. However, in our system more than ten thousand pulses are averaged for each measurement. Because of this large number, we can accept the mean as a very good approximation of the returned pulse.

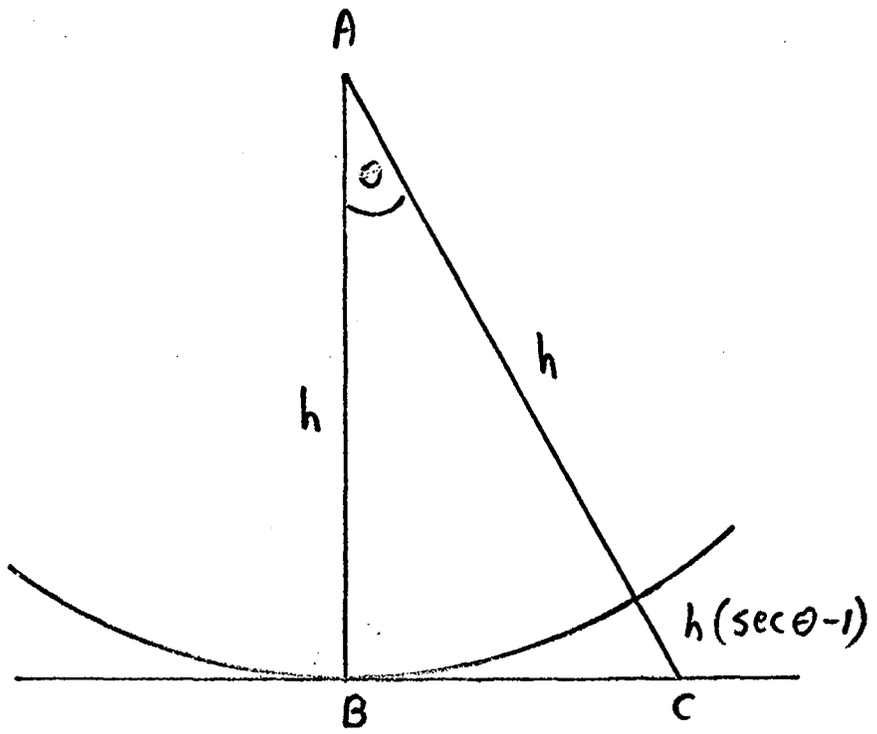


Fig. 4.1 The geometry of reflection

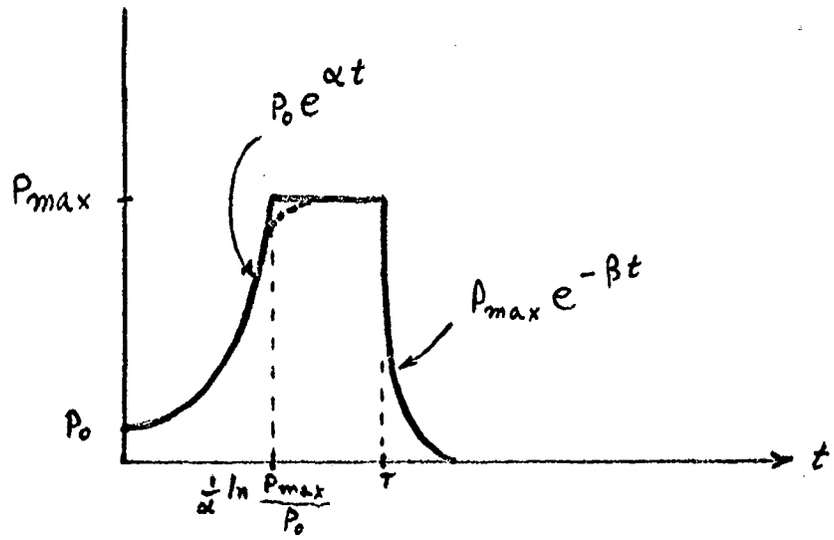


Fig. 4.2 The transmitted pulse

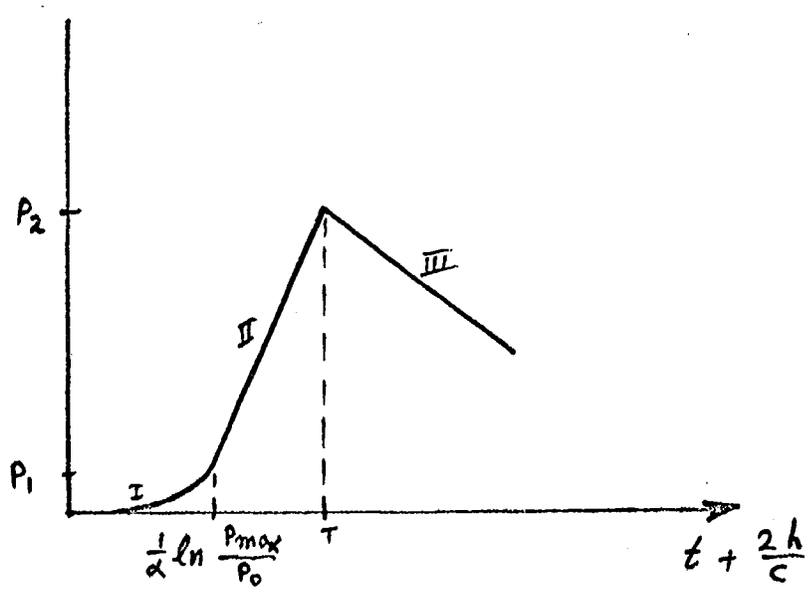


Fig. 4.3 The received pulse

5. Linear Approximation of the Altimeter

Looking at the block diagram (Fig. 2.4) of the altimeter, we note that the signal in one part of the loop has the form of voltage intensity, and in the other part, the form of frequency. There is a linear relation between the two, which is the linear characteristic of the VCO.

We can, therefore, replace frequency by voltage whenever it is necessary, and get an equivalent block diagram (Fig. 5.1). Here $\rho(x)$ describes the dependence of the detector output on the input voltage to the VCO. It also includes the delay, the intensity, and the shape of the returned pulse. In order to get our linear approximation we will assume $\rho(x)$ to be a parabola with a maximum at x=o,

$$\rho(x) = a_0 - a_1 x^2 \tag{5.1}$$

$$\rho'(x) = -2a_1 x = x\rho''(x) = x\rho'' \tag{5.2}$$

$$\rho''(x) = -2a_1 \tag{5.3}$$

When the loop is closed we have

$$x(t) = x_1 + c_1 \sin w_1 t \tag{5.4}$$

expanding $y_1(t)$ in a Taylor series provides (see Fig. 2.3), for small c_1

$$y_1(t) = \rho(x_1) + \rho'(x_1) c_1 \sin w_1 t + \rho''(x_1) 0.5 (c_1 \sin w_1 t)^2 + \ldots \tag{5.5}$$

Due to our choice of parabola, (5.5) becomes

$$y_1(t) = \rho(x_1) + x\rho'' c_1 \sin w_1 t + 0.5\rho''(c_1 \sin w_1 t)^2 \tag{5.6}$$

which means that the signal component at w_1 is proportional to the distance from x = o, where the curve has its maximum, and hence describes a linear relation.

The amplifier includes the gain of the superregenerative detector, and the attenuation of the path to the ocean and back. Although we showed earlier that the detector has logarithmic gain, we will assume linearity here.

The band pass amplifier is narrowband and centered around $w = w_1$. It serves to reject the d.c. and harmonic components of w_1. It also plays a major role in improving the signal-to-noise ratio. If A is the amplification of the amplifiers and if we assume that no phase shift is introduced in this stage at the frequency w_1, then

$$Y_2(t) = X P'' c_1 A \sin w_1 t \qquad (5.7)$$

(Ignoring the phase shift is justified because any phase shift at the frequency w_1, introduced by the amplifiers can be compensated by a similar phase shift introduced between the oscillator and the multiplier.)

At the output of the multiplier we get

$$Y_3(t) = X P'' c_1 A \sin w_1 t \sin w_1 t \qquad (5.8)$$

or

$$Y_3(t) = X P'' c_1 A \, 0.5 (1 - \cos 2 w_1 t) \qquad (5.9)$$

The low pass amplifier attenuates the second harmonic. Assuming its d.c. amplification to be included in A we get

$$Y_4 = X \, 0.5 A c_1 P'' \qquad (5.10)$$

The equivalent block diagram (Fig. 5.1) can now be transformed from its band pass form to a simplified low pass form.

Before we proceed in simplifying the block diagram, we must also consider the noise . Note that the receiver includes a multiplier after a narrow pass amplifier. White noise after a band pass network is described by

$$n(t) = n_1(t)\cos w_1 t - n_2(t)\sin w_1 t \qquad (5.11)$$

where both $n_1(t)$ and $n_2(t)$ are independent Gaussian variables, and each term includes half the mean noise power.

This noise is now multiplied by $\sin w_1 t$ and we get, due to noise only

$$Y_3 = [n_1(t)\cos w_1 t - n_2(t)\sin w_1 t]\sin w_1 t$$

$$(5.12)$$

$$Y_4 = 0.5[n_1(t)\sin 2w_1 t - n_2(t) + n_2(t)\cos 2w_1 t] .$$

The terms including $\sin 2w_1 t$ do not pass the following low pass amplifier, and we note that only $n_2(t)$ contributes to the d.c. signals.

This is the 3 db improvement in signal-to-noise ratio achieved whenever there is a synchronous detector, and we will include this effect by taking only half the noise power in the input.

The simplified equivalent block diagram of the altimeter is given in Fig. 5.2. In this block diagram ρ'' is taken as a positive constant, although, for a parabola with a maximum, the second derivative is a negative constant. This is corrected by introducing negative feedback instead of the positive feedback in Fig. 5.1.

A is the total gain and it includes the gain or attenuation of:

1. The r.f. path
2. The superregenerative circuit
3. The band pass amplifier
4. The integrator $\left(= \frac{1}{RC}\right)$
5. The low pass amplifier
6. The multiplier $(=\frac{1}{2})$

F(s) is the total response of the networks in the loop, and it includes:

1. The band pass amplifier response, transferred to low pass.
2. The low pass amplifier response.

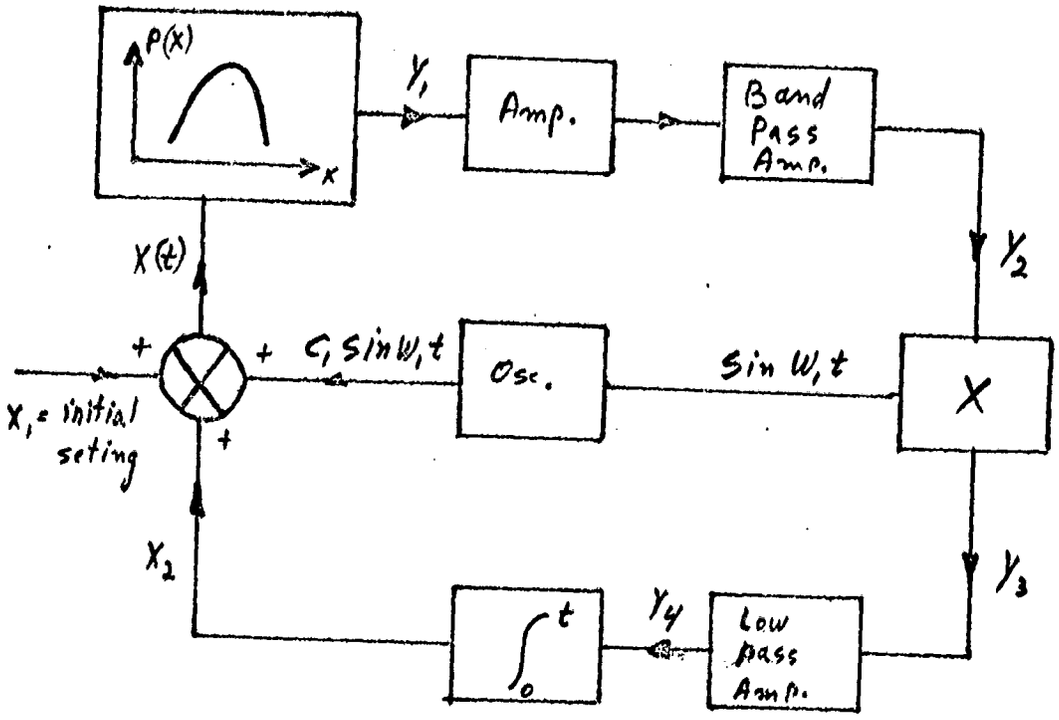


Fig. 5.1 The equivalent block diagram of the altimeter

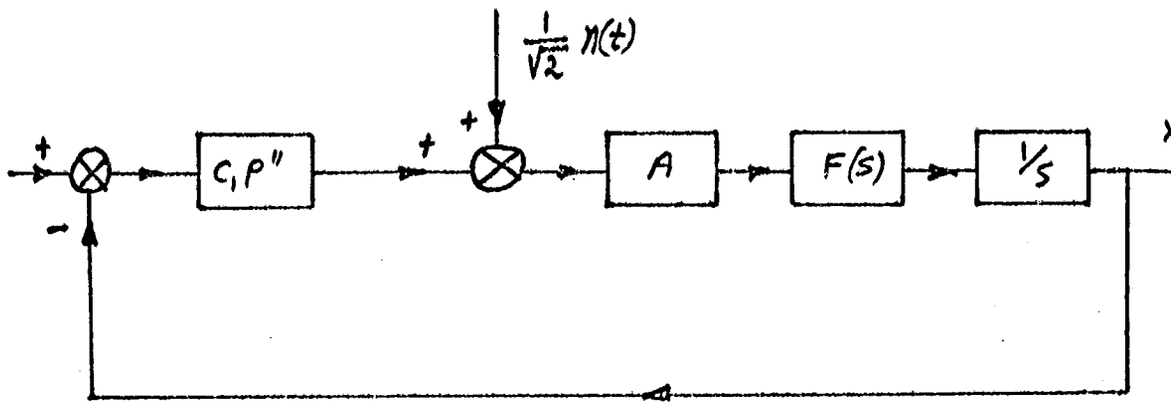


Fig. 5.2 A simplified equivalent block diagram of the altimeter

6. Noise and Mean-Square Error

The primary noise source of the altimeter is the superregenerative stage, and the noise bandwidth is the effective bandwidth of the loop. As we will show, the equivalent noise source temperature of the superregenerative stage is high and depends on the r.f. bandwidth of the stage. The reason for this is the sampling effect of the quenching pulses.

Let us assume that the oscillation frequency is F_o . The bandwidth of the r.f. stage is ΔF_o and the repetition rate, which is also the sampling rate, is F_q . Assume that an interferring signal exists at $f_o + f_m$ (Fig. 6.1). This interfering signal can be considered as a modulating frequency, F_m. As long as this modulating frequency is smaller than half the sampling frequency, the reconstructed frequency after the sampling, is unchanged (Fig. 6.2). When $F_m > \frac{1}{2} F_q$, the reconstructed frequency decreases. The dependence of the frequency at the output of the detector, on the modulating frequency is summarized in Fig. 6.3.

If f_b and f_a are the upper and lower limits of the pass band that follows the superregenerative stage, the output will be influenced not only by modulating signals between these two limits $F_a < F_m < F_b$, but also by higher modulating signals: $F_q - F_b < F_m < F_q - F_a$, $F_q + F_a < F_m < F_q + F_b$, $2F_q - F_b < F_m < 2F_q - F_a$, etc. The number of contributing regions, m, is limited by the r.f. bandwidth, as can be seen from Fig. 6.3, and equals

$$m = \frac{2 \Delta F_o}{F_q} \tag{6.1}$$

and the one-sided noise spectral density N_o is therefore

$$N_o = K T^o \frac{2 \Delta F_o}{F_q} \tag{6.2}$$

where: K = Boltzmann's constant
T° = Noise-source temperature, $^\circ K$

The simplified block diagram (Fig. 6.4) of the altimeter is similar to the linearized model of a phase lock loop.[1] Following Viterbi's treatment, we note that the spectral density of the output $S_x(w)$ is related to that of the noise, by

$$S_x(w) = \left| \frac{A \; F(w)/iw}{1 + c_1 \rho'' A \; F(w)/iw} \right|^2 S_n(w) \qquad (6.3)$$

while the closed loop transfer function is

$$H(s) = \frac{c_1 \rho'' A \; F(s)/s}{1 + c_1 \rho'' A \; F(s)/s} \qquad (6.4)$$

Now we let the noise be white with one-sided spectral density N_o. However due to the synchronous detection only half of this power enters the loop so we have an effective one-sided spectral density of $\frac{N_o}{2}$ Watts/H_z.

The spectral density of the output is:

$$S_x(w) = \frac{N_o}{4(c_1 \rho'')^2} \left| H(iw) \right|^2 \qquad (6.5)$$

---

[1] A. J. Viterbi, "Principles of Coherent Communication," McGraw-Hill, 1966, Chapters 1-4.

Then the variance of the output error due to noise is:

$$\sigma_x^2 = \frac{N_0}{4\pi (c_1 \rho'')^2} \int_0^\infty |H(iw)|^2 dw \tag{6.6}$$

We define the loop-noise bandwidth as:

$$B_L = \frac{1}{2\pi} \int_0^\infty |H(iw)|^2 dw \tag{6.7}$$

so that the mean square error due to noise can be written as:

$$\sigma_x^2 = \frac{N_0 \, B_L}{2 (c_1 \rho'')^2} \tag{6.8}$$

Let us call the total gain of the loop K such that

$$K = c_1 \rho'' A \tag{6.9}$$

Then, for different types of filters, the following are the loop-noise bandwidth:

Table 6.1

$F(s)$	B_L
1	$\frac{K}{4}$
$\frac{1}{s\tau+1}$	$\frac{K}{4}$
$\frac{1}{(s\tau+1)^2}$	$\frac{K}{2(2-K\tau)}$
$\frac{\tau s+1}{\alpha\tau s+1}$	$\frac{K(K\tau+1)}{4(\alpha K\tau+1)}$

With no filter we had

$$B_L = \frac{K}{4} = \frac{C_1 P'' A}{4} \tag{6.10}$$

and inserting (6.10) in (6.8) we see that the mean square error with no filter is

$$\sigma_X^2 = \frac{N_0 A}{8 C_1 P''} \tag{6.11}$$

As was indicated before, the noise entering the loop was a result of a narrow band process and was, therefore, normal. Our system, thus far, is linear. Therefore, the steady state probability density of the output X due to the noise is also normal. Fokker-Planck equations are used in Appendix A to show that if, when X reaches a value X_ℓ we say that the loop unlocks, then the mean time to unlock is given by

$$T = \frac{8}{N_0 A^2} \int_{X=0}^{X_\ell} e^{-\frac{\alpha X^2}{2}} \left[\int_{y=x}^{X_\ell} e^{\frac{\alpha y^2}{2}} dy \right] dx \tag{6.12}$$

where $$\alpha = \frac{8 P'' C_1}{N_0 A} = \frac{1}{\sigma_X^2}$$

We note that for small X_ℓ (6.12) will reduce to

$$T = \frac{4 X_\ell^2}{N_0 A^2} \tag{6.13}$$

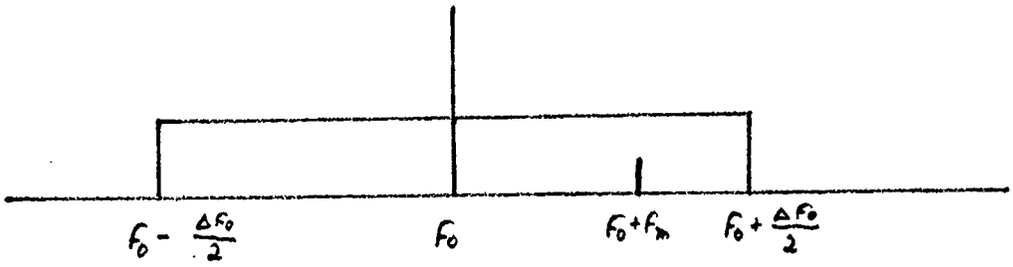


Fig. 6.1 The R.F. bandwidth

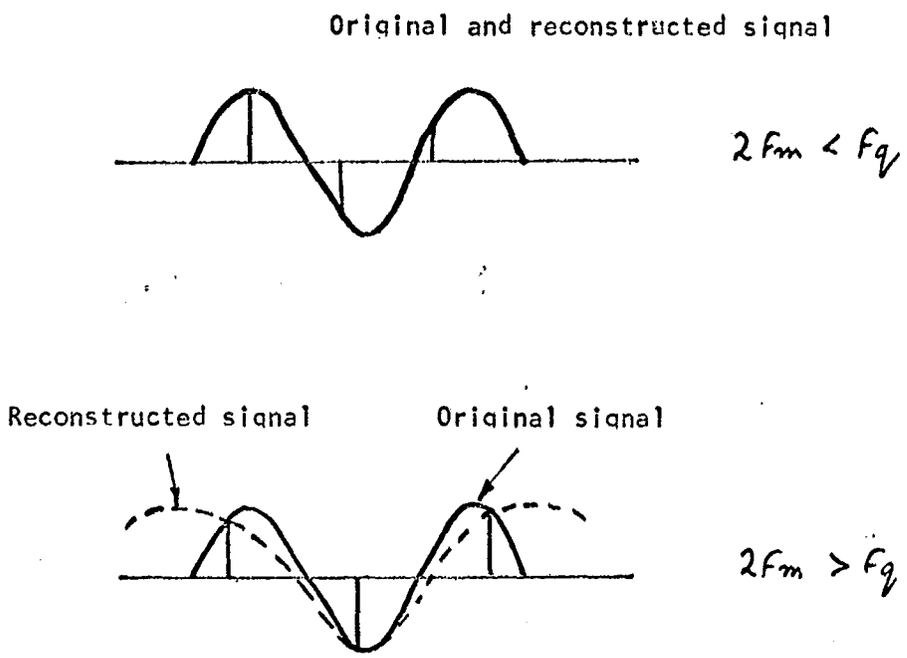


Fig. 6.2 The effect of low sampling rate on the reconstructed signal

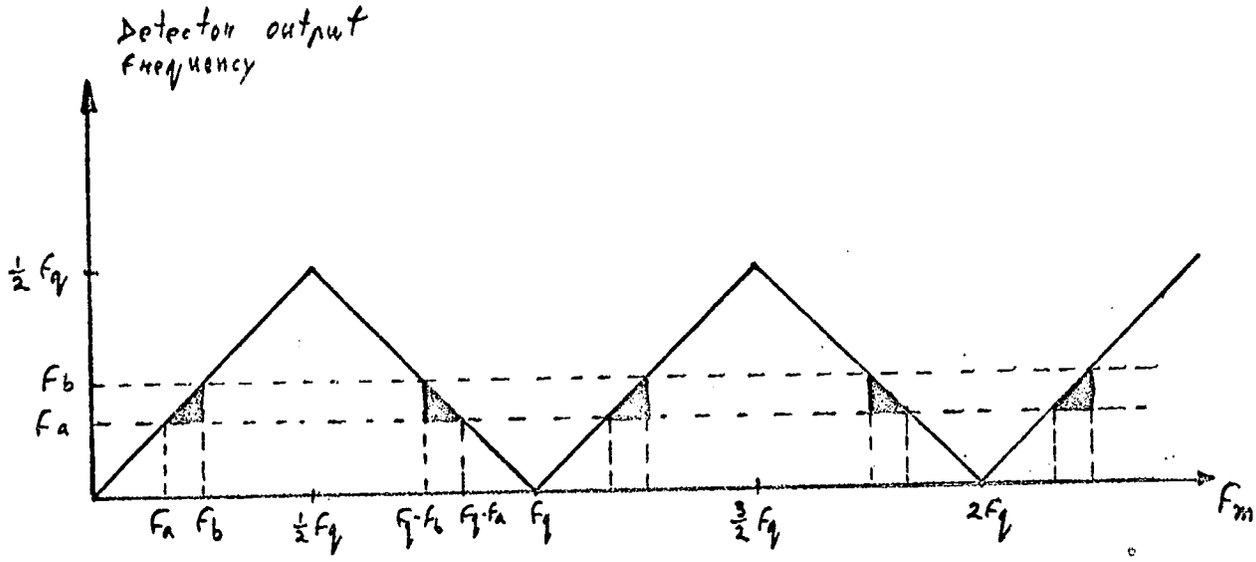


Fig. 6.3 Detector output frequency as function of modulating frequency

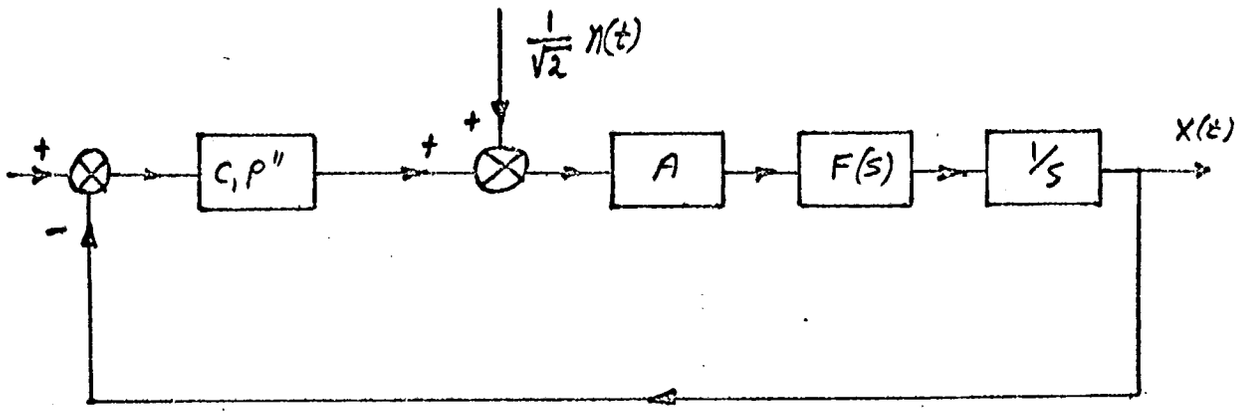


Fig. 6.4 A simplified equivalent block diagram of the altimeter

7. Sweeping Voltage, Steady State Error and Overshoot

In the previous section we introduced the linear approximation by choosing the dependence of the detector output $\rho(x)$ on the VCO input x, as a parabola. This curve is determined mainly by the shape of the return pulse which covers only a small portion of x (Fig. 6.1), or, in other words, the duration of the returned pulse is short compared to the total delay range in which the altimeter operates.

When x is such that $\rho(x)$ is zero, no signal exists in the loop, x would be a result of integration of noise, and the mean time to lock would be very long. To overcome this problem a sweeping voltage is introduced. This is a constant voltage connected to the input of the integrator and causing f_q (or x) to drift toward higher values (Fig. 7.1). If locking was lost and f_q has reached its maximum value (equivalent to the lowest operational altitude of the altimeter) the integrator's capacitor is shorted, and a new cycle of sweeping starts. This sweeping voltage causes a steady state error in the system. From the redrawn block diagram (Fig. 7.2) we get:

$$\frac{X(s)}{V(s)} = \frac{1}{s + K F(s)} \tag{7.1}$$

The sweeping voltage is constant:

$$V(s) = \frac{V}{s} \tag{7.2}$$

so

$$X(s) = \frac{V}{s\left[s + K F(s)\right]} \tag{7.3}$$

then

$$\lim_{t \to \infty} X(t) = \lim_{s \to 0} s X(s) = \frac{V}{K \lim_{s \to 0} F(s)} \tag{7.4}$$

For most filters of interest:

$$\lim_{s \to 0} F(s) = 1 \tag{7.5}$$

and we get for the steady state error

$$E_1 = \lim_{t \to \infty} X(t) = \frac{V}{K} \tag{7.6}$$

As we saw in the previous section, the mean square error due to noise is:

$$E_n^2 = \sigma_x^2 = \frac{N_0 B_L}{2(c_1 \rho'')^2} \tag{7.7}$$

Where B_L, the loop-noise bandwidth, is related to the total gain K. With no filter $B_L = \frac{K}{4}$, and we get

$$E_n^2 = \frac{N_0 K}{8(c_1 \rho'')^2} \tag{7.8}$$

In order to lower the noise error we have to reduce the gain K, but this will raise the steady state error E_1, because the square of the steady state error, is given from (7.6) as

$$E_1^2 = \frac{V^2}{K^2} \tag{7.9}$$

The total error is, therefore:

$$E^2 = E_n^2 + E_1^2 = \frac{N_0 K}{8(c_1 \rho'')^2} + \frac{V^2}{K^2} \tag{7.10}$$

With a filter we might be able to narrow the loop-noise bandwidth and to reduce the error due to noise without raising the steady

state error. However, an overshoot effect limits these improvements.

The way in which overshoot can cause the loss of locking is shown in Fig. 7.3. We recall that the returned pulse is limited, and, therefore $P(x)$ exists only between $-X_\ell$ to $+X_\ell$ (when the peak is at zero). Without a sweeping voltage, the overshoot could never exceed X_ℓ, and, as a result of the damping, will usually be lower. However, strong sweeping voltage might cause the overshoot to exceed X_ℓ and the loop to unlock.

The analysis of the overshoot will assume that the loop of Fig. 7.2 has initial conditions

$$\begin{aligned} X(0) &= -X_\ell \\ \dot{X}(0) &= 0 \\ v(t) &= V\,U(t) \end{aligned} \tag{7.11}$$

where U(t) is the step function. It will also assume the lag network which appears in Fig. 7.4, as the filter in the loop.

From Appendix B we get the locking behavior

$$X(t) = \frac{V}{K} + \frac{1}{W} e^{-\frac{a}{2}\left(\frac{K}{b}+1\right)t} \left\{ \left[V - \frac{a}{2}\left(\frac{K}{b}+1\right)\left(X_\ell + \frac{V}{K}\right)\right] \sin Wt - \left(X_\ell + \frac{V}{K}\right) W \cos Wt \right\} \tag{7.12}$$

$$\text{where} \quad W = \sqrt{aK - \frac{a^2}{4}\left(\frac{K}{b}+1\right)^2}$$

From table 6.1 we see that the loop noise bandwidth with this lag network in the loop, is given by

$$B_L = \frac{K}{4} \frac{\frac{K}{b} + 1}{\frac{b}{a}\frac{K}{b} + 1} \tag{7.13}$$

In order to get a significant reduction in B_L, compared to $\frac{K}{4}$, we would like two conditions to exist:

1. $\frac{K}{b} > 0.1$

2. $\frac{b}{a}$ large

For values of k/b large, compared to 1, the reduction in bandwidth approaches b/a. In Fig. 7.5, the locking behaviour of the loop is drawn for several ratios of b/a. From this figure we see that b/a cannot be as large as we wish, because the overshoot might exceed X_ℓ, and the system will not lock.

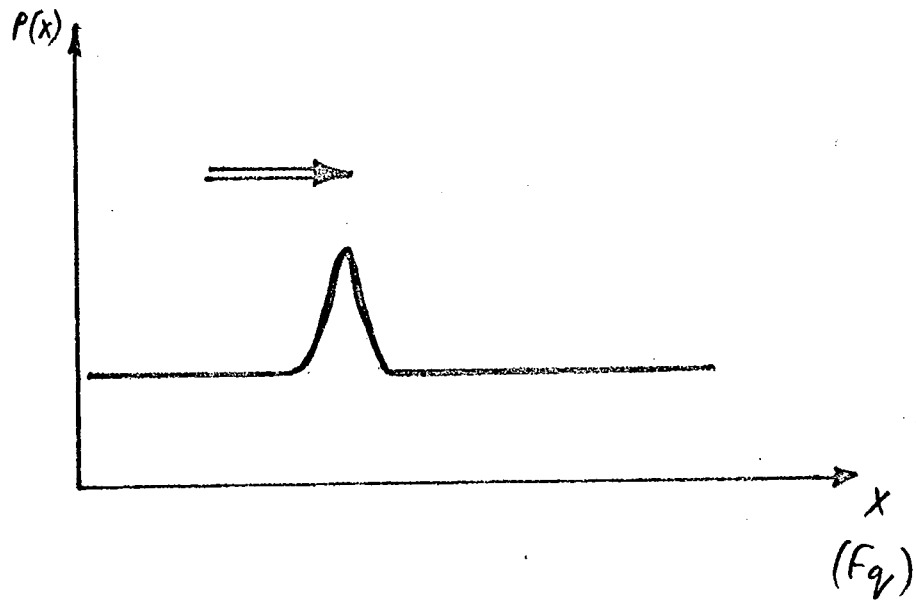


Fig. 7.1 The direction of sweeping

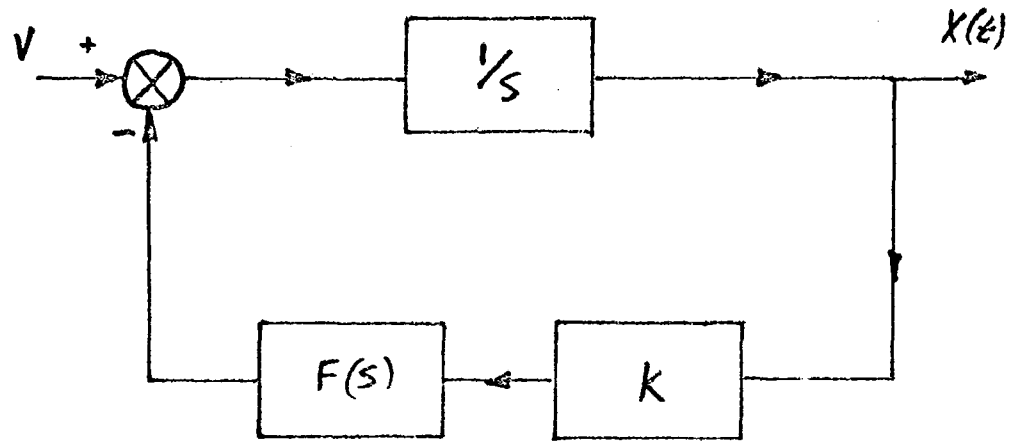


Fig. 7.2 The feedback loop redrawn with the sweeping voltage as an input

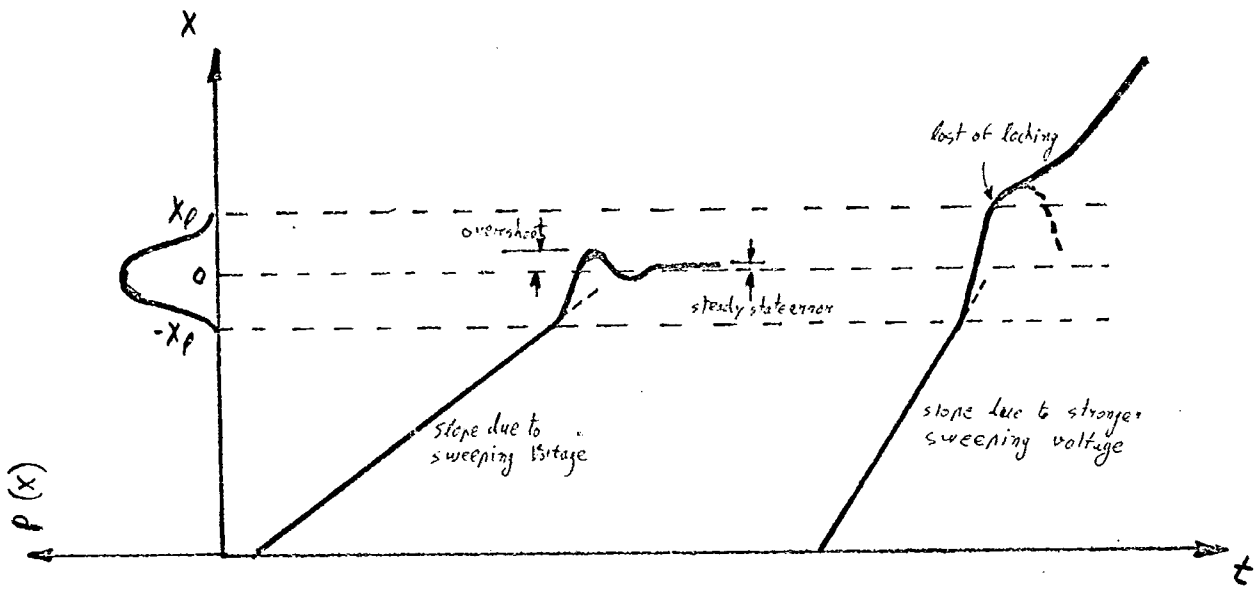
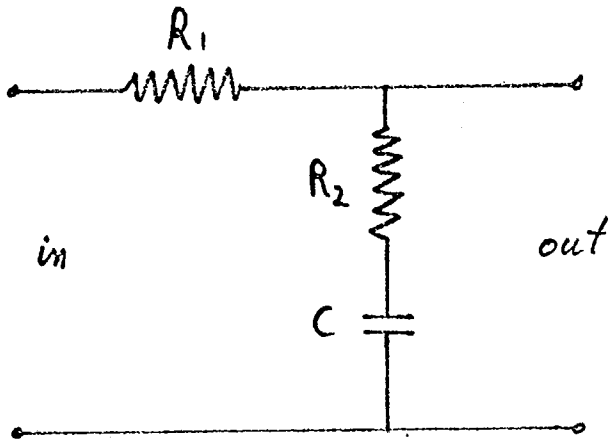


Fig. 7.3 Lost of locking due to overshoot



$$a = \frac{1}{(R_1 + R_2)C}$$

$$b = \frac{1}{R_2 C}$$

$$F(s) = \frac{a}{b}\,\frac{s+b}{s+a}$$

Fig. 7.4 The lag network in the feedback loop

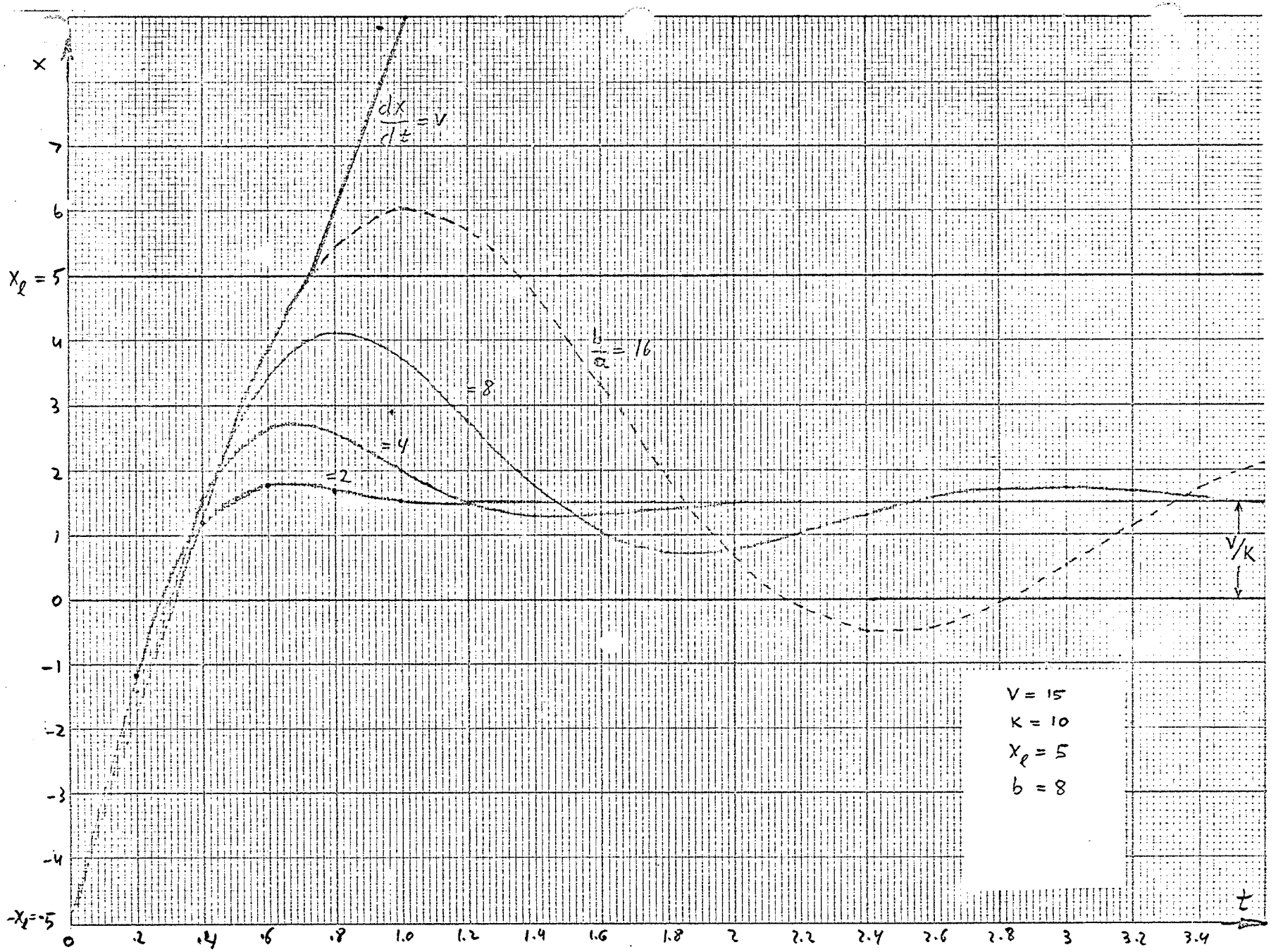


Fig. 7.5 Locking behavior of the loop

8. Antenna Design

The two alternatives for an antenna were either coating portions of the balloon as a reflector, or hanging an antenna below the altimeter. The first alternative would have interferred with the communication link between the balloon and the interrogation satellite. We, therefore, had to choose a hanging antenna. Because the antenna has to receive its own signal after it is reflected from the ocean, no antenna with circular polarization could be used.

The final choice was a rhombic antenna [1,2] which is light, very simple to construct, and capable of achieving a gain of 40. The use of the rhombic is usually limited to the HF band, but there is no reason why it should not fit for UHF use, if careful attention is given to the accuracy of the dimensions.

The rhombic antenna (Fig. 8.1) is a non-uniform transmission line, terminated by a matched impedance so that no reflection occurs. With no reflection each leg has two lobes (in any plane) tangent to its direction (Fig. 8.2). In order to have lobes 1 - 4 pointing down, the following relation should exist,

$$\frac{\pi \ell}{\lambda} \cos^2\phi = \tan\left[\frac{\pi \ell}{\lambda}(1-\sin\phi)\right] \qquad (8.1)$$

For $\ell = 3\lambda$, we get $\phi = 62°$

The correct termination for the above antenna (F=430 MH_z), was found experimentally by using the fact that the antenna is symmetric. Different resistors were connected at the output until the input impedance was equal to the output. This value was 330 ohms.

---

[1] F. E. Terman, "Radio Engineers' Handbook," McGraw-Hill, 1943, pp. 804-806.

[2] L. Thourel, "The Antenna," Chapman R. Hall, 1960, pp. 144-166.

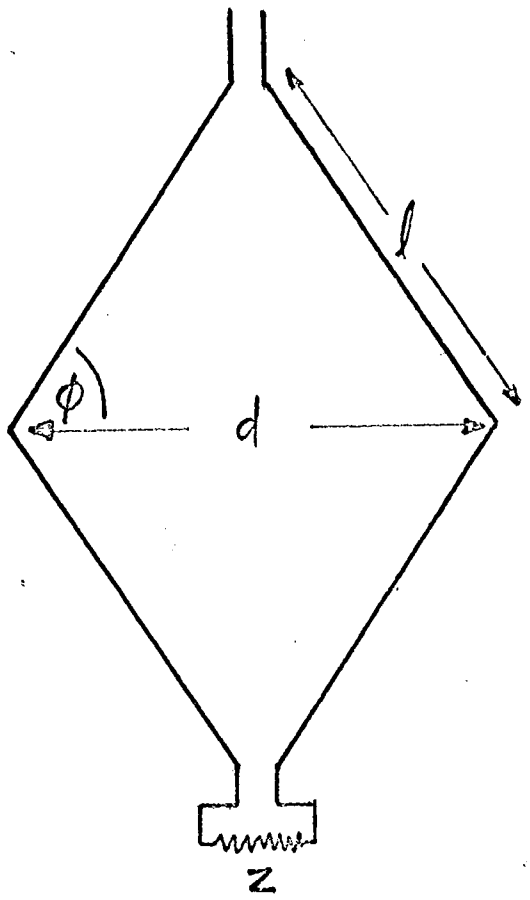


Fig. 8.1 The rhombic antenna

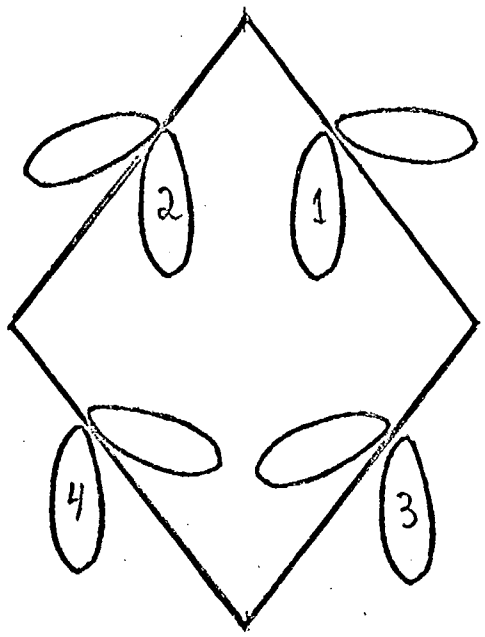


Fig. 8.2 The lobes of the rhombic legs

9. Circuitry and Testing

The superregenerative stage utilizes a nuvistor triode for the 430 MH_z version, and a pencil tube with integral resonator (the type used in radiosonde) for the 1680 MH_z version. The narrow band amplifier is an operational amplifier with a twin T network in the feedback loop. The multiplier is a ring demodulator. The integrator and oscillator utilize operational amplifiers. The VCO is an astable multivibrator in which the charging resistors are returned to a variable input. The pulse shaper includes a monostable stage that determines the duration of the quench pulse, and a Schmitt trigger that determines it's amplitude. Aside from the Schmitt trigger, the rest of the active stages are integrated circuits.

The laboratory test system appears in Fig. 9.1. The quenching pulses were used to trigger a delayed pulse generator. These delayed pulses modulated a UHF signal generator whose output simulated the returned pulses. The output of the signal generator was connected to the input (and also output) of the superregenerative detector. This point was also connected to a detector in front of an oscilloscope. This scope was used to monitor the shape of the envelope of the transmitted pulses. The counter was used to count the repetition rate. The chart recorder measured the input voltage to the VCO. These two equipments were used to measure steady state error and noise error. The chart recorder was also used to measure the loop bandwidth. To do this the delay was modulated by a VLF sine wave from the function generator. The dual chart recorder recorded both the sinusoidal signal at the output of the function generator, and the resulted sinusoidal signal at the output of the integrator. By comparing these two amplitudes at several frequencies it was possible to draw the frequency response of the loop.

The amplitude of the simulated returned pulse was controlled by the output attenuator of the signal generator. It should be noted that changes in the returned pulse amplitude affected mainly the signal to noise ratio, but also the gain of the loop, as the gain

includes in it the second derivative of the envelope of the returned pulse.

Without any sweeping voltage, the altimeter starts to lock when the peak of the returned pulses is as low as 2.5μV. With proper sweeping voltage (the entire range is scanned in ten seconds,) good locking (with mean time to unlock in the order of hours) is achieved with 10μV input. However, the steady state error is about .8% and the noise error .1%. With an input of 20μV the total error is below the specified .03%.

A flight test was conducted on June 5, 1968, over Lake Michigan, with an aircraft provided by the National Center for Atmospheric Research, Boulder, Colorado. The rhombic antenna could not be used on an aircraft, and it was replaced by a slotted dipole antenna. The gain of the slotted dipole was about 1.8 compared to 40 which is the expected gain from the rhombic. The maximum altitude at which the test was conducted was 8000 feet. From Eq. 4.17, the returned pulse is related to G^2/h^3. That means that this experiment is equivalent to testing the altimeter with the rhombic antenna, at an altitude of 70,000 feet.

The purpose of this flight test was to check whether the altimeter would lock to the returned signal, and how it would follow changes in the altitude of the aircraft. Excellent locking was achieved up to the maximum altitude of the flight--8000 feet. For the 30 minute period over the lake, locking was not lost even once, unless by external disturbance by the operator. Readings of altitude were performed by reading the period of the quench pulses with a portable oscilloscope. They were compared to the pressure altimeter of the aircraft. With these rather limited facilities, readings on the tested altimeter were within $\pm$ 0.1 percent from the readings on the aircraft altimeter.

The height of the waves in Lake Michigan at the time of the above test was 3 feet, according to the Coast Guard. On route back to Madison the altimeter was still locking over land, at an altitude of 8000 feet. However, locking was lost from time to time.

The specifications of the prototype used in this flight were:

Frequency	430 MH_Z
Peak Output Power	0.8 Watts
Total Input Power	1.8 Watts
Weight (not including batteries and antenna)	20 ozs.

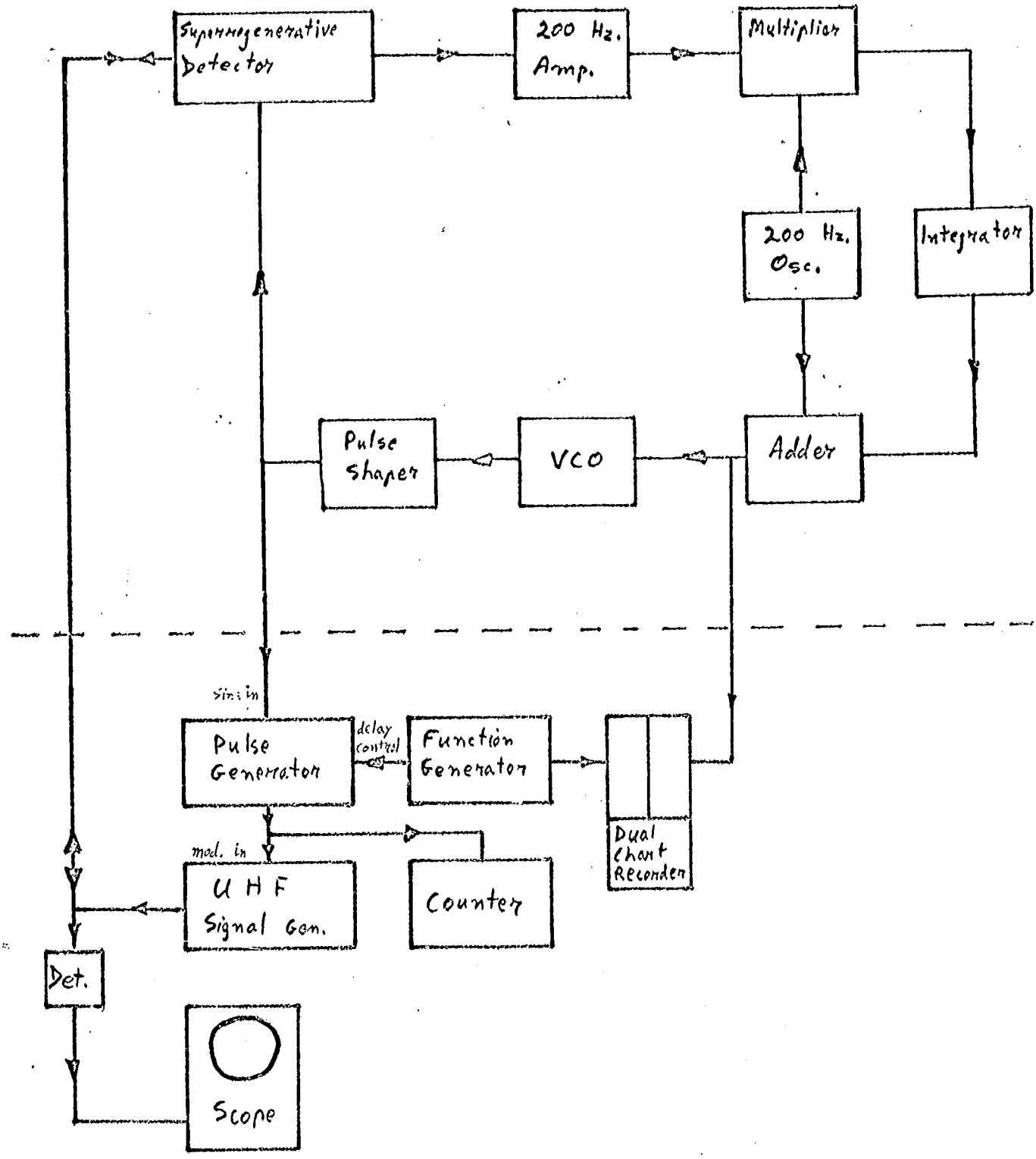


Fig. 9.1 Block diagram of the altimeter and the test system

10. Miscellaneous Problems

a. Additional Error Due to End Effects

The altitude information from the altimeter is extracted by counting the quench pulses or the r.f. pulses. If the counting period is not a multiple of the period of the perturbing sinusoidal signal, an error due to end effect will result.

Let us assume the perturbing sine wave is perfect, and that the VCO is linear. In Fig. 10.1 we have:

Δf_q - the maximum deviation of f_q

f_q - the correct repetition rate

T_o - the period of the perturbing signal

T - the counting period

α - the start point of the counting

n - the number of full periods of T_o in T

We define the additional portion of a period T_o in T as β(in degrees), and get:

$$\beta = 360\left(\frac{T}{T_o} - n\right) \tag{10.1}$$

This additional period can occur when f_q is high or low, and an error in the reading of the average f_q will occur.

This error E_F is given by

$$E_F = \frac{\Delta f_q}{n + \frac{\beta}{360}} \int_{\alpha}^{\alpha+\beta} \sin\theta \, d\theta = \frac{\Delta f_q}{n + \frac{\beta}{360}} \, 2 \sin\left(\alpha + \frac{\beta}{2}\right) \sin\frac{\beta}{2} \tag{10.2}$$

Note from (10.1) that

$$n + \frac{\beta}{360} = \frac{T}{T_o} \tag{10.3}$$

The maximum error occurs when

$$\sin\frac{\beta}{2} = 1 \qquad (10.4)$$

$$\sin\left(\alpha + \frac{\beta}{2}\right) = 1$$

and it is given by

$$\max E_F = \frac{2T_0}{T}\,\Delta F_q \qquad (10.5)$$

Finally we are interested in the relative error

$$\frac{\max E_F}{F_q} = \frac{2T_0}{T}\,\frac{\Delta F_q}{F_q} \qquad (10.6)$$

Practical numbers for our flight model are:

$$\frac{\Delta F_q}{F_q} = 2\ 10^{-2}$$

$$T_0 = \frac{1}{200}\ \text{sec}$$

$$T = 1\ \text{sec}$$

This would yield a maximum error of 0.02%, which is by itself close to the specified limit.

It is possible to get rid of this error by using the same basic oscillator for both T and T_0. This is possible when the counting is done in the balloon package, but impossible when the counting is done on the ground. To minimize the error it is required to have a small deviation ΔF_q ; however, the equivalent term in voltage C_1, is linearily related to the signal to noise ratio, and should be large.

b. Built-in Sweeping Voltage

We mentioned earlier that with no signal, $\rho(x)$ is a horizontal curve, whose derivatives are zero. Therefore, it would appear that there is no input to the integrator except for noise. This is not the case, because as F_q (or x) rise, there are more r.f. pulses, and

the superregenerative detector output rises in accordance. The true output of the detector, as a function of f_q when no return signal is present, is shown in Fig. 10.2. The slope of this curve is fixed and it therefore serves as a built-in sweeping voltage for the loop, with the only difference that it is inserted before the amplifier rather than after it. In practice this sweeping voltage is too strong and it had to be partly compensated.

c. Sub-harmonic Operation and Ambiguity

There is no reason why the altimeter cannot transmit a second pulse before the previous one is received. We shall call this sub-harmonic operation (Fig. 10.3). This mode of operation can happen by itself, at any altitude above twice the minimum altitude of operation. This immediately indicates that the minimum range of altitude required to get full coverage, is equal to the minimum altitude. If the minimum altitude of operation is 4000 feet, then the range of the altimeter should be from 4000 feet to 8000 feet. Above 8000 feet the altimeter will switch to a sub-harmonic mode. From Fig. 10.4 it is clear that the first ambiguity problem starts at 12000 feet which is 3 x 4000 feet, or, 2 x 6000 feet, and the altimeter can lock on both. The pattern of switching modes as the balloon ascends, in the case of no loss of locking except at the end of range, is marked in the Figure. However, this pattern is not guaranteed since loss of locking might occur. In any case, the ambiguity is at least the lowest altitude (in our example 4000 feet), and it can be resolved by pressure reading or by the history of the ascending.

Sub-harmonic operation is recommended even if the range of operation is limited as in the case of the super-pressure balloons. In this case, however, it is possible to assure that only one mode will cover this limited range. Operation with sub-harmonics, means higher quench frequency, f_q , and therefore, (see Appendix C) better signal to noise ratio.

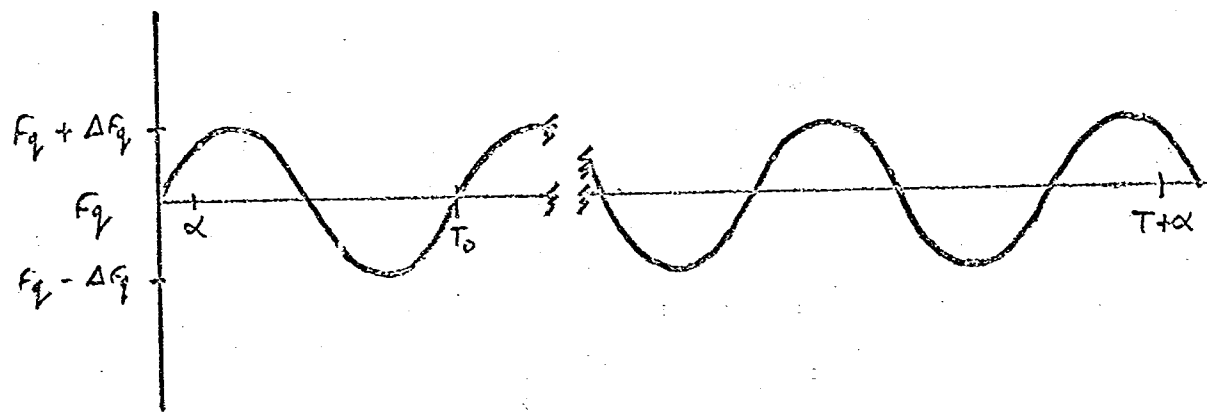


Fig. 10.1 The sinusoidal perturbation as a cause to error

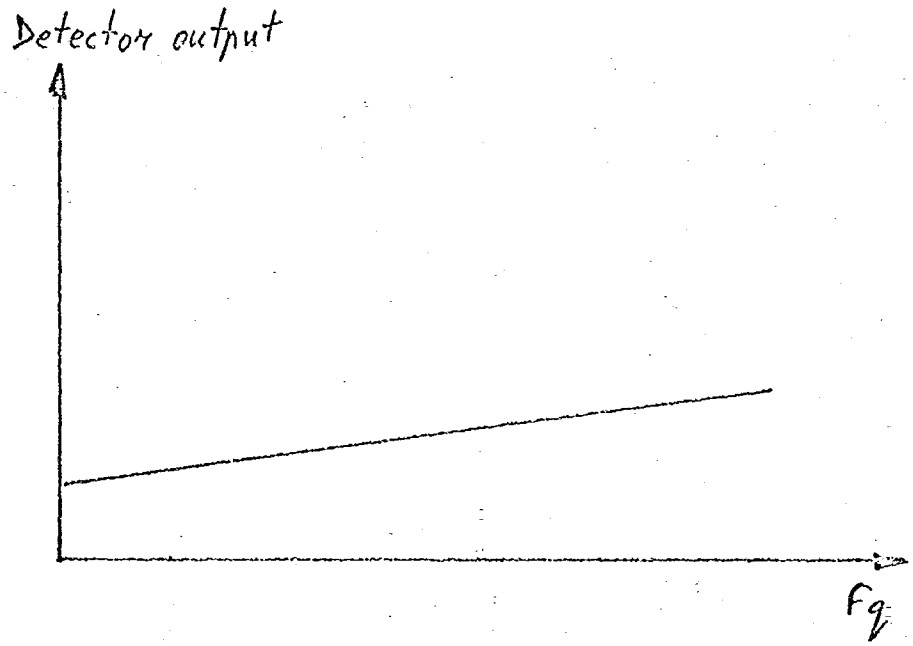


Fig. 10.2 The detector output as function of f_q , in the absence of the returned pulse

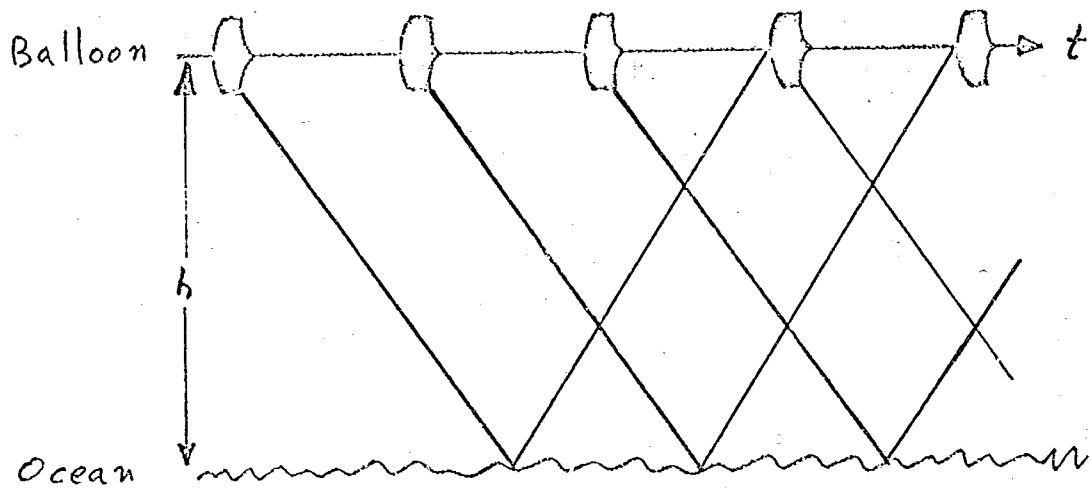


Fig. 10.3 Sub-harmonic operation

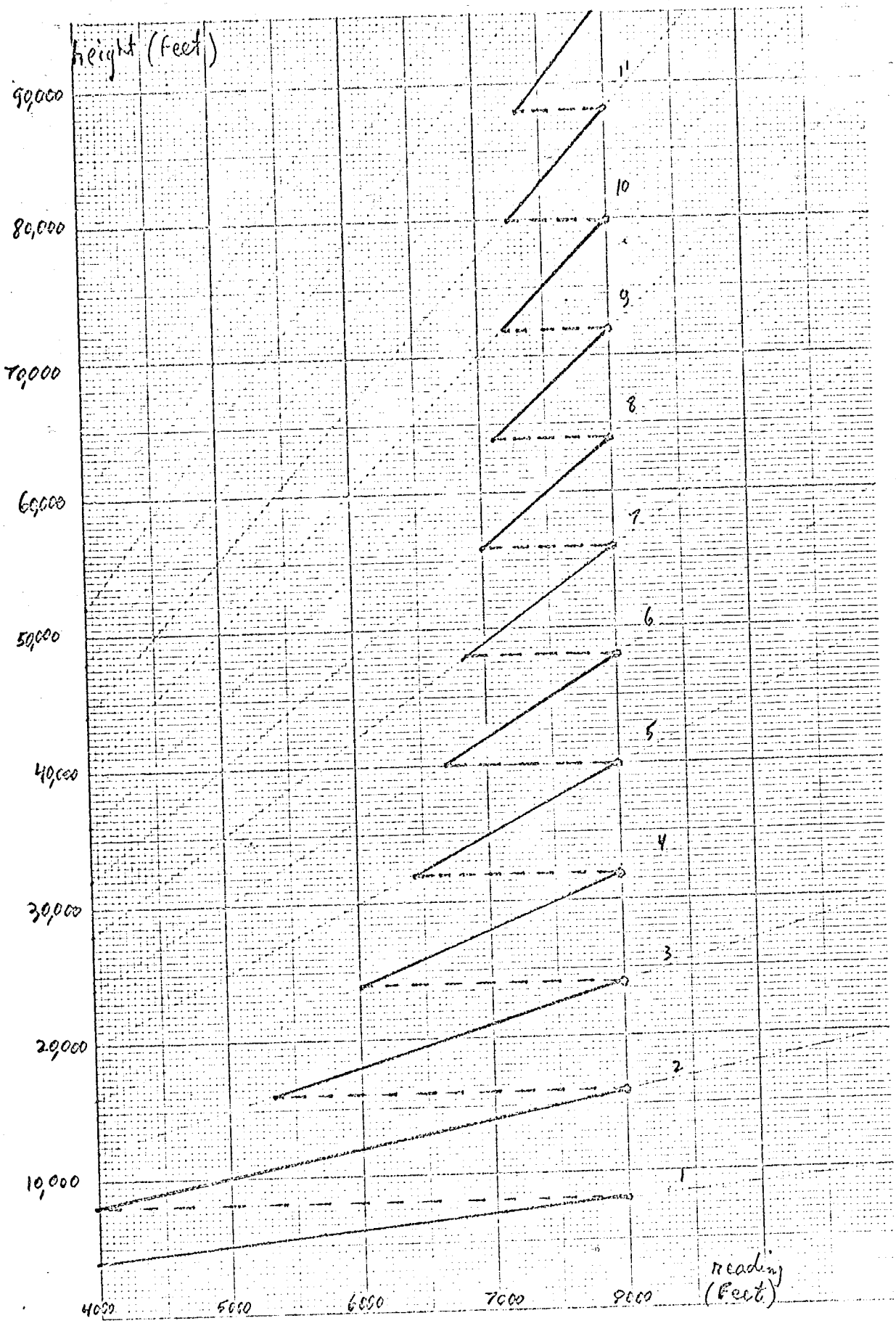


Fig. 10.4 Ambiguity pattern

11. Future Work

A second prototype is now being completed for a balloon flight test. The electrical specifications are the same, but the weight will be reduced to 7 ozs., and the package will pass environmental tests. This flight will include the rhombic antenna whose weight is 5 ozs.

In this test the altimeter will be flown with a regular radiosonde transmitter. It will use the original 100 v and 6v Magnesium Cuprous Chloride batteries, in addition to two 12v batteries of the same type. A ground station will receive the radiosonde signal on 1680 MH_z, and the altimeter pulses at 430 MH_z. The temperature and pressure reading from the radiosonde signal, and the repetition rate of the altimeter pulses, will be recorded simultaneously. The altitude as calculated from integration of temperature will then be compared to the altitude calculated from the repetition rate. The purpose of this test will be, mainly, to check the operation of the altimeter under true environmental condition; and the maximum operational altitude with the present power.

The accuracy of the altimeter will be checked in further flights, where radar tracking will be available. Additional flights will be conducted with the 1680 MH_z model, once it is completed.

On the theoretical side an effort will be made to analyze the true non-linear loop. This will include replacing the parabola by the mean returned pulse and introducing the logarithmic gain of the detector. Conclusions drawn from this analysis might cause changes in circuitry.

Appendix A

The Mean Time To Unlock

The loop in Fig. 6.4 is very similar to a phase lock loop treated extensivly by Viterbi[1]. We will follow his analysis rather closely. When there is no filter in the loop, and no change in the input, the differential equation of x(t) is:

$$\frac{d\,X(t)}{dt} + C_1 \rho'' A\, X(t) = \frac{A}{\sqrt{2}}\, n(t) \qquad (A.1)$$

we will assume that

$$P(x, t=0) = \delta(x - x_0) \qquad (A.2)$$

x(t) is a Markoff process and it therfore obeys the Fokker - Planck equation

$$\frac{\partial P(X,t)}{\partial t} = -\frac{\partial}{\partial X}\left[A_1(X)\, P(X,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial X^2}\left[A_2(X)\, P(X,t)\right] \qquad (A.3)$$

where :

$$A_1(X) = \lim_{\Delta t \to 0} \frac{E\left[\Delta X \mid X\right]}{\Delta t} \qquad (A.4)$$

$$A_2(X) = \lim_{\Delta t \to 0} \frac{E\left[(\Delta X)^2 \mid X\right]}{\Delta t} \qquad (A.5)$$

---

1. A.J. Viterbi , "Princples of Coherent Communication", McGraw - Hill, 1966,Ch. 1 - 4

We will integrate (A.3) over the infinitesimal interval from t to $t+\Delta t$, and get

$$\Delta X = X(t+\Delta t) - X(t) = -\rho'' c_1 A\, X(t)\Delta t + \frac{A}{\sqrt{2}} \int_t^{t+\Delta t} \eta(u)\, du \qquad (A.6)$$

from (A.4) and (A.6) we get

$$A_1 = -\rho'' c_1 A\, X(t) \qquad (A.7)$$

and from (A.5) and (A.6) we get

$$A_2 = \lim_{\Delta t \to 0} \frac{A^2}{2\Delta t} \int_t^{t+\Delta t} \int_t^{t+\Delta t} E\left[\eta(u)\,\eta(v)\right] du\, dv \qquad (A.8)$$

$$= \lim_{\Delta t \to 0} \frac{A^2 N_0}{4\Delta t} \int_t^{t+\Delta t} \int_t^{t+\Delta t} \delta(u-v)\, du\, dv = \frac{A^2 N_0}{4} \qquad (A.9)$$

Inserting (A.7) and (A.9) in (A.3) we get

$$\frac{\partial P(X,t)}{\partial t} = \frac{\partial}{\partial X}\left[\rho'' c_1 A\, X(t)\, P(X,t)\right] + \frac{A^2 N_0}{8} \frac{\partial^2 P(X,t)}{\partial X^2} \qquad (A.10)$$

We previously assumed

$$P(X,0) = \delta(X - X_0) \qquad (A.11)$$

we will now choose $X_0 = 0$. We will agree that unlocking occurs when

$$X = \pm X_\ell \qquad (A.12)$$

and in accordance will note that

$$P(X_\ell, t) = P(-X_\ell, t) = 0 \tag{A.13}$$

We will define the probability that X did not reach X_ℓ at time t , as

$$\Psi(t) = \int_{-X_\ell}^{X_\ell} P(X, t)\, dX \tag{A.14}$$

This is a single valued function (monotonic decreasing with t); its density is given by $-\frac{d\Psi}{dt}$. The expected time to reach X_ℓ for the first time is, therefore, given by

$$T = -\int_0^\infty t \frac{d\Psi(t)}{dt}\, dt = -\left[t\, \Psi(t)\right]_0^\infty + \int_0^\infty \Psi(t)\, dt \tag{A.15}$$

If we accept that $\Psi(t)$ approaches zero faster then 1/t , then the first term of (A.15) is zero, and we get

$$T = \int_{t=0}^{\infty} \int_{x=-X_\ell}^{+X_\ell} P(X, t)\, dx\, dt \tag{A.16}$$

We will integrate (A.10) from t = 0 to infinity, and get

$$P(X,\infty) - P(X,0) = \frac{d}{dX}\left[\rho'' c_1 A\, X(t)\, G(X)\right] + \frac{N_c A^2}{8} \frac{d^2 G(X)}{dX^2} \tag{A.17}$$

where

$$G(X) = \int_0^\infty P(X, t)\, dt \tag{A.18}$$

Clearly $P(x,\infty) = 0$; and, since X is assumed to be initially zero, $P(x,o) = \delta(x)$. (A.17) becomes

$$-\delta(x) = \frac{d}{dx}\left[\rho'' c_1 A\, x(t) Q(x)\right] + \frac{N_o A^2}{8}\,\frac{d^2 Q(x)}{dx^2} \qquad (A.19)$$

with the boundary conditions

$$Q(X_\ell) = \int_0^\infty P(X_\ell, t)\,dt = 0$$

$$Q(-X_\ell) = \int_0^\infty P(-X_\ell, t)\,dt = 0 \qquad (A.20)$$

Taking the integral of both sides of (A.19) from $X = -X_\ell$ to $X = X_\ell$ will yield

$$C - U(x) = \rho'' c_1 A\, X\, Q(x) + \frac{N_o A^2}{8}\,\frac{dQ(x)}{dx} \qquad (A.21)$$

where U(x) is the unit step function, and C is a constant to be evaluated from the boundary conditions. Let us define:

$$\gamma = \frac{N_o A^2}{8} \qquad (A.22)$$

$$\alpha = \frac{8\rho'' c_1}{N_o A} \qquad (A.23)$$

then (A.21) can be written as

$$Q'(x) + \alpha X Q(x) = \frac{1}{\gamma}\left[C - U(x)\right] \qquad (A.24)$$

Multiplying both sides by a function g(x) we get

$$g(x)\,Q' + \alpha x\, g(x)\, Q = \frac{1}{\gamma}\left[c - U(x)\right] g(x) \qquad (A.25)$$

let us have

$$g(x)\,Q' + \alpha x\, g(x)\, Q = \left[g(x)\,Q\right]' = g(x)\,Q' + g'(x)\,Q \qquad (A.26)$$

In order for (A.26) to be correct we must have

$$\alpha x\, g(x) = g'(x) \qquad (A.27)$$

or

$$\frac{g'(x)}{g(x)} = \alpha x$$

$$\therefore \quad \ln g(x) = \frac{\alpha x^2}{2}$$

$$\therefore \quad g(x) = e^{\frac{\alpha x^2}{2}} \qquad (A.28)$$

Using (A.28) and (A.26) in (A.25) gives

$$\left(e^{\frac{\alpha x^2}{2}} Q\right)' = \frac{1}{\gamma}\left[c - U(x)\right] e^{\frac{\alpha x^2}{2}} \qquad (A.29)$$

or

$$e^{\frac{\alpha x^2}{2}} Q = \frac{1}{\gamma}\int_{-x_0}^{x}\left[c - U(t)\right] e^{\frac{\alpha t^2}{2}}\, dt + D \qquad (A.30)$$

or

$$Q(x) = \frac{1}{\gamma} e^{-\frac{\alpha x^2}{2}} \int_{-x_\ell}^{x} \left[C - U(t)\right] e^{\frac{\alpha t^2}{2}} dt + D \qquad (A.31)$$

The boundary conditions give

$$Q(x = -x_\ell) = 0 \Rightarrow D = 0$$

$$Q(x = x_\ell) = 0 \Rightarrow C = \frac{1}{2} \qquad (A.32)$$

$$\therefore \; Q(x) = \frac{1}{\gamma} e^{-\frac{\alpha x^2}{2}} \int_{-x_\ell}^{x} \left[\frac{1}{2} - U(t)\right] e^{\frac{\alpha t^2}{2}} dt \qquad (A.33)$$

The mean time to unlock is given by (A.16) and (A.18) as

$$T = \int_{-x_\ell}^{x_\ell} Q(x)\, dx \qquad (A.34)$$

Inserting (A.33) in (A.34) we get

$$T = \frac{1}{\gamma} \int_{-x_\ell}^{x_\ell} e^{-\frac{\alpha x^2}{2}} \left[\int_{-x_\ell}^{x} \left[\frac{1}{2} - U(t)\right] e^{\frac{\alpha t^2}{2}} dt \right] dx$$

$$= \frac{1}{\gamma} \int_{-x_\ell}^{0} e^{-\frac{\alpha x^2}{2}} \left[\int_{-x_\ell}^{x} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt \right] dx$$

$$+ \frac{1}{\gamma} \int_{0}^{x_\ell} e^{-\frac{\alpha x^2}{2}} \left[\int_{-x_\ell}^{0} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt + \int_{0}^{x} \left(-\frac{1}{2}\right) e^{\frac{\alpha t^2}{2}} dt \right] dx$$

$$= \frac{1}{\gamma}\int_0^{x_\ell} e^{-\frac{\alpha x^2}{2}}\left[\int_{-x_\ell}^{-x} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt\right] dx$$

$$+ \frac{1}{\gamma}\int_0^{x_\ell} e^{-\frac{\alpha x^2}{2}}\left[\int_0^{x_\ell} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt - \int_0^{x} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt\right] dx$$

$$= \frac{1}{\gamma}\int_0^{x_\ell} e^{-\frac{\alpha x^2}{2}}\left[\int_x^{x_\ell} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt\right] dx$$

$$+ \frac{1}{\gamma}\int_0^{x_\ell} e^{-\frac{\alpha x^2}{2}}\left[\int_x^{x_\ell} \frac{1}{2} e^{\frac{\alpha t^2}{2}} dt\right] dx$$

$$\therefore \quad T = \frac{1}{\gamma}\int_0^{x_\ell} e^{-\frac{\alpha x^2}{2}}\left[\int_x^{x_\ell} e^{\frac{\alpha t^2}{2}} dt\right] dx \qquad (A.35)$$

Appendix B

Overshoot Analysis with a Lag Network

From Fig. 7.4 we have

$$F(s) = \frac{a}{b} \frac{s+b}{s+a} \tag{B.1}$$

From Eq. 7.1 we have

$$\frac{X(s)}{V(s)} = \frac{1}{s + K F(s)} \tag{B.2}$$

(B.2) in (B.1) gives

$$s\,X(s) + \frac{Ka}{b} \frac{s+b}{s+a} X(s) = V(s) \tag{B.3}$$

or

$$\left[s^2 + \left(\frac{K}{b}+1\right)as + Ka\right] X(s) = V(s)(s+a) \tag{3.4}$$

and in differential equation form

$$\frac{d^2x}{dt^2} + a\left(\frac{K}{b}+1\right)\frac{dx}{dt} + aKX = V(t)\,F^{-1}(s+a) \tag{B.5}$$

Returning to Laplace transform with initial conditions

$$s^2 X(s) - s\,X(0) - \dot{X}(0) + a\left(\frac{k}{b}+1\right) s\,X(s) - a\left(\frac{k}{b}+1\right) X(0) + a\,k\,X(s) = V(s)(s+a) \qquad (B.6)$$

The initial conditions are:

$$\begin{aligned} X(0) &= -X_\ell \\ \dot{X}(0) &= 0 \\ V(t) &= V\,U(t) \end{aligned} \qquad (B.7)$$

Inserting (B.7) in (B.6) we get

$$X(s) = \frac{\left(\frac{V}{s} - X_\ell\right)(s+a) - \frac{ak}{b} X_\ell}{s^2 + a\left(\frac{k}{b}+1\right)s + a\,k} \qquad (B.8)$$

or

$$X(s) = \frac{\left(\frac{V}{s} - X_\ell\right)(s+a) - \frac{ak}{b} X_\ell}{\left[s + \frac{a}{2}\left(\frac{k}{b}+1\right) - j\omega\right]\left[s + \frac{a}{2}\left(\frac{k}{b}+1\right) + j\omega\right]} \qquad (B.9)$$

where

$$\omega = \sqrt{a\,k - \frac{a^2}{4}\left(\frac{k}{b}+1\right)^2} \qquad (B.10)$$

Let us call:

$$\begin{aligned} \frac{a}{2}\left(\frac{k}{b}+1\right) - j\omega &= A \\ \frac{a}{2}\left(\frac{k}{b}+1\right) + j\omega &= B \end{aligned} \qquad (B.11)$$

then

$$X(s) = \frac{-sX_\ell}{(s+A)(s+B)} + \frac{V - aX_\ell\left(\frac{k}{b}+1\right)}{(s+A)(s+B)} + \frac{Va}{s(s+A)(s+B)} \qquad (B.12)$$

and in the time domain

$$X(t) = \frac{V - aX_\ell\left(\frac{k}{b}+1\right)}{B-A}\left(e^{-At} - e^{-Bt}\right)$$

$$- \frac{X_\ell}{B-A}\left(Be^{-Bt} - Ae^{-At}\right) \qquad (B.13)$$

$$+ Va\left[\frac{1}{B-A}\left(\frac{1}{B}e^{-Bt} - \frac{1}{A}e^{-At}\right) + \frac{1}{AB}\right]$$

We note that

$$B - A = 2j\omega \qquad (B.14)$$

$$BA = aK$$

so that

$$x(t) = \frac{V}{K} + \frac{1}{2j\omega}\left\{\left[V - \frac{aX_\ell}{2}\left(\frac{k}{b}+1\right) - jX_\ell\omega - \frac{Va}{A}\right]e^{-At}\right.$$

$$\left. - \left[V - \frac{aX_\ell}{2}\left(\frac{k}{b}+1\right) + jX_\ell\omega - \frac{Va}{B}\right]e^{-Bt}\right\} \qquad (B.15)$$

Note that

$$\frac{1}{A} = \frac{\frac{a}{2}\left(\frac{k}{b}+1\right) + j\omega}{aK} \qquad (B.16)$$

$$\frac{1}{B} = \frac{\frac{a}{2}\left(\frac{k}{b}+1\right) - j\omega}{aK}$$

From (B.11) , (B.15) and (B.16) we get

$$X(t) = \frac{V}{K} + \frac{1}{\omega} e^{-\frac{a}{2}\left(\frac{k}{b}+1\right)t} \left\{ \left[V - \frac{a}{2}\left(\frac{k}{b}+1\right)\left(X_0 + \frac{V}{K}\right)\right] \sin\omega t - \left(X_0 + \frac{V}{K}\right)\omega \cos\omega t \right\} \qquad (B.17)$$

Appendix C

Signal to Noise Ratio

If we will not include in equation (6.8) the $\frac{1}{2}$ due to synchronous detection we will get

$$\sigma_x^2 = \frac{N_0 \; B_L}{(c_1 \rho'')^2} \qquad (C.1)$$

We note that this is the inverse of the signal to noise ratio at the output, or

$$\left(\frac{S}{N}\right)_0 = \frac{1}{\sigma_x^2} \qquad (C.2)$$

In Fig. C.1 the various processes which the signal passes, and their effect on the signal to noise ratio are shown.

The first term is the $\frac{1}{F}$ change in signal to noise ratio due to the noise figure, F, of the r.f. stage. The second is the increase in noise due to sampling. The $\sqrt{n}$ improvement is due to averaging of n pulses.[1]

The term, 2, is due to the synchronous detection, and the last term is the improvement in signal to noise ratio due to the change from the r.f. bandwidth Δf_0 to the loop bandwidth B_L.

We note that the number of pulses averaged, n, is given by

$$n = \frac{f_q}{B_L} \qquad (C.3)$$

where: f_q - the repetition rate

B_L - the loop effective bandwidth

We thus get from Fig. C.1 that the signal to noise ratio at

---

[1] D.K Barton "Radar System Analysis," Prentice - Hall, 1964, pp. 25-29.

the output, $\left(\frac{S}{N}\right)_o$, is related to the signal to noise ratio at the input, $\left(\frac{S}{N}\right)_i$, as follows:

$$\left(\frac{S}{N}\right)_o = \frac{1}{F}\left(\frac{f_q}{B_L}\right)^{3/2}\left(\frac{S}{N}\right)_i \qquad (C.4)$$

Using (C.2) we get

$$\left(\frac{S}{N}\right)_i = F\,\sigma_x^{-2}\left(\frac{f_q}{B_L}\right)^{-\frac{3}{2}} \qquad (C.5)$$

The noise power at the input is given by

$$N_i = KT\,\Delta f_o \qquad (C.6)$$

where: K = Boltzmann's Constant

T = Noise-source temperature ^oK

Inserting (c.6) in (c.5) we get for the signal power input

$$S_i = KT\,\Delta f_o\;F\,\sigma_x^{-2}\left(\frac{f_q}{B_L}\right)^{-\frac{3}{2}} \qquad (C.7)$$

It will be interesting to put numbers in this equation. We recall that σ_x is the mean square error due to noise. If the total error allowed is .03% a sensible choice of σ_x will be .01% or 10^{-4}, thus, we have for the balloon flight model

$$KT = 5\;10^{-21}\ \text{watts/Hz}$$

$$\Delta f_o = 5\;10^{6}\ \text{Hz}$$

$$F = 8$$

$$\sigma_x = 10^{-4}$$

$$f_q = 10^{5}\ \text{Hz}$$

$$B_L = 1\ \text{Hz}$$

and the resulted required signal power at the input is

$$S_i = 5\ 10^{-13} \text{ watts}$$

The available power, S, from a voltage source E with an internal resistor R is given by

$$S = \frac{E^2}{4R} \qquad (C.8)$$

For R = 50 and $S = S_i$ we will get a typical value for the input voltage

$$E = 10\,\mu V$$

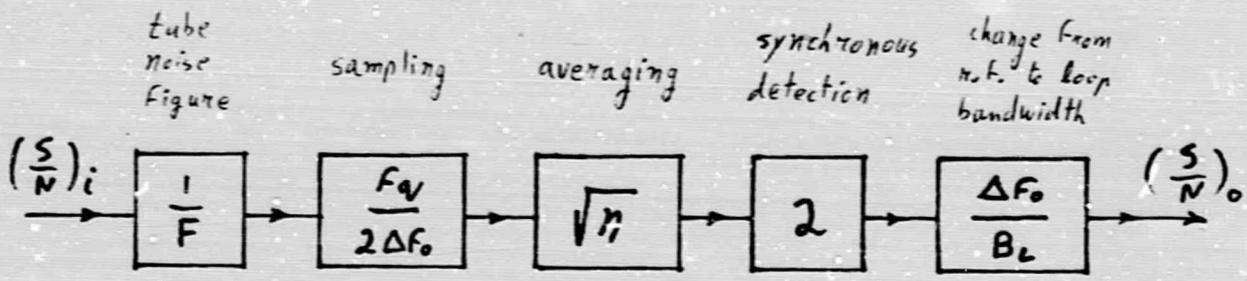


Fig. C. 1 Signal-to-noise-ratio budget

APPENDIX II

A SURVEY OF NOISE GENERATION IN SOLID STATE INFRARED DETECTORS

Progress Report # 1
Under Contract NAS5 - 11542

Robert Krauss
Project Specialist

Space Science and Engineering Center
The University of Wisconsin

CONTENTS

Robert Krauss
October 2, 1968

A Survey of Noise Generation in Solid State Infrared Detectors

OPTICAL PROPERTIES

Many dielectric materials transmit in the visible and absorb in the infrared. Most semiconductors are opaque in the visible but transparent in a large part of the infrared. Figure I shows the relation between absorption and wavelength in a typical semiconductor. Where $\lambda < \lambda_o$, incident photons have enough energy to excite electrons across the forbidden energy band. This intrinsic absorption at short wavelengths is so efficient that even very thin material is completely opaque.

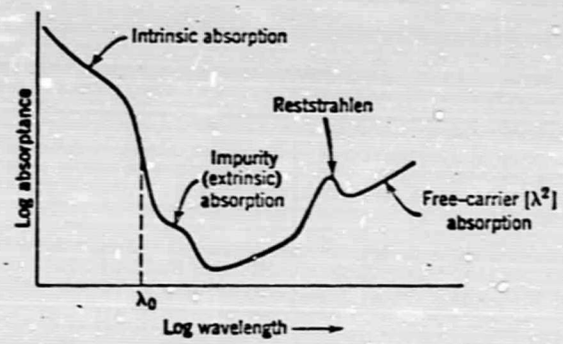


Figure 1. Spectral absorption in a semiconductor.

The absorption edge, λ_o, is defined by

$$\lambda_o = \frac{hc_o}{E}$$

where E is the width of the forbidden energy gap. At wavelengths greater than λ_o, the atoms in a valence crystal are nearly neutral and give rise to very little dipole moment and abstract little energy from the beam.
The only other absorption possible for $\lambda > \lambda_o$ is due to the presence of absorption bands in the forbidden gap due to impurities and phonons. These can be excited thermally and optically, and are therefore one reason for keeping detectors cooled. Since the width E of the forbidden band is somewhat temperature dependent, λ_o may exhibit a temperature dependent shift.

The primary bonding in semiconductors is covalent. In cases where ionic bonding exhists, one often observes reststrahlen, which are optical resonances related to anomalous dispersion. Finally, at longer wavelengths, one finds free carrier absorption, which is proportional to λ^2 at shorter wavelengths and reaches a temperature dependent constant at long wavelengths. In impure semiconductors absorption bands appear due to excitation of electrons from donor centers to the conduction bands or from the valence band to acceptor centers. These centers are generally near the band edges, so the excitation energy is small (see figure 2).

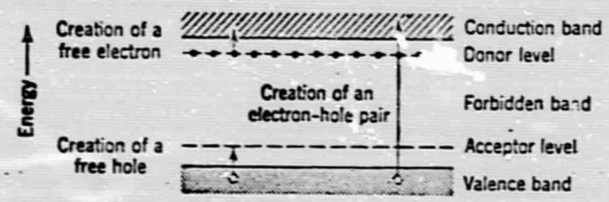


Figure 2. Energy bands and excitation processes in an extrinsic semiconductor.

Another type of absorption, which may produce bands considerably more intense than free carrier absorption at wavelengths $\lambda > \lambda_o$, results from transitions within an energy band - the transitions between "heavy" and "light" holes, as in p-type germanium.

Another, often valuable, optical property of semiconductors is their high index of refraction. Figure 3 on the next page shows a comparison of refractive index for various optical materials.

ELECTRICAL STRUCTURE AT THERMAL EQUILIBRIUM

Semiconductors are often classified as intrinsic or extrinsic. Intrinsic absorption causes excitation of electrons across the forbidden band. Extrinsic absorption causes transitions of electrons between impurity states and the conduction or valence band. The distance between energy levels in a band is of the order of 10^{-14} ev. Thus the levels form practically a continuum. The forbidden band widths of some pure semiconductors are given in table I.

One can further divide extrinsic semiconductors into "n-type" and "p-type". The n-type semiconductors have donor atoms near the conduction band which control the conductivity by producing negative charge carriers. In p-type semiconductors, acceptor atoms get electrons from the valence band, creating positive hole carriers. Intrinsic conduction produces electron-hole pairs, where both electrons and holes contribute to electrical conduction. Figure 2 shows all three processes.

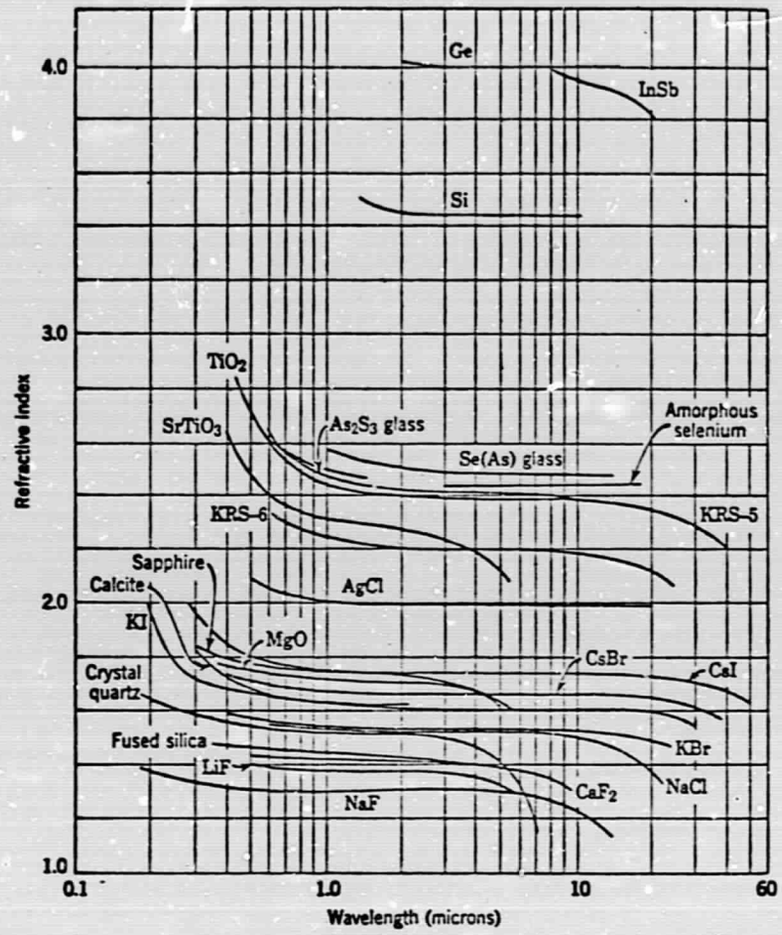


Figure 3. The refractive indices of selected optical materials. [After W. L. Wolfe and S. S. Ballard, *Proc. Inst. Radio Engrs.* 47, 1540 (1959).]

Table I. Width of forbidden band in pure semiconductors at room temperature

Semiconductor	Band Gap (ev)	Semiconductor	Band Gap (ev)
Ge	0.67	GaAs	1.45
Se	2	GaSb	0.80
Si	1.12	InAs	0.35
Sn (gray)	0.08	InP	1.25
Te	0.33	InSb	0.18
AlSb	1.6	Mg_2Sn	0.33
CdS	2.4	PbS	0.37
ZnS	3.6	PbSe	0.27
CdSe	1.8	PbTe	0.30
CdTe	1.5	HgTe	0.02

The influence of impurities on the electrical properties of semiconductors is remarkable. One impurity atom per million is able to completely dominate the lattice electrical properties. This is due to the fact that the free carrier concentration depends exponentially on the ratio of Fermi energy to temperature, so that only a very small fraction of lattice sites is ionized at any given temperature.

In order to obtain quantitative relationships of the electrical properties of semiconducting materials, one must determine the free carrier concentration at equilibrium. The probability that an energy level E is occupied is

$$f_e(E) = \frac{1}{e^{\frac{E-E_f}{kT}} + 1}$$

where E_f is the Fermi level, at which the probability of occupation is 0.5, and where energies are measured from the bottom of the conduction band. The density of states $g_e(E)$, the number of available energy levels per unit volume per unit energy is

$$g_e(E) = 4\pi \left(\frac{2m_e}{h^2}\right)^{3/2} E^{1/2}$$

where m_e is the effective electron mass. The density of electrons is then

$$n(E) = g(E)f(E) = 4\pi \left(\frac{2m_e}{h^2}\right)^{3/2} \frac{E^{1/2}}{e^{\frac{E-E_f}{kT}} + 1}$$

and the number of electrons in the conduction band is

$$n = \int_0^\infty n(E)\,dE = 4\pi \left(\frac{2m_e}{h^2}\right)^{3/2} \int_0^\infty \frac{E^{1/2}\,dE}{e^{\frac{E-E_f}{kT}} + 1}$$

$$(1) \qquad n = N_c e^{E_f/kT} \qquad\qquad N_c \equiv 2\left(\frac{2\pi m_e kT}{h^2}\right)^{3/2}$$

where N_c is the effective electron volume density at the bottom of the conduction band.

For the valence band, we do the same calculation using $f_h(E) = 1 - f_e(E)$

$$f_h(E) = \frac{1}{e^{\frac{E_f-E}{kT}} + 1}$$

$$g_h(E) = 4\pi \left(\frac{2m_h}{h^2}\right)^{3/2} (-E_i - E)^{1/2}$$

where m_h is the hole effective mass and E_i is the intrinsic excitation energy i.e. forbidden band width. The density of free holes is then

$$p = \int_{-\infty}^{-E_i} g_h(E)\, f_h(E)\, dE = 4\pi \left(\frac{2m_h}{h}\right)^{3/2} \int_{-\infty}^{-E_i} \frac{(-E-E_i)^{1/2}\, dE}{e^{\frac{E_f - E}{kT}} + 1}$$

(2) $$p = N_v e^{\frac{-E_f - E_i}{kT}} \qquad N_v \equiv 2\left(\frac{2\pi m_h kT}{h^2}\right)^{3/2}$$

and N_v is called the effective hole density at the top of the valence band.

Using equations (1) and (2), one can find the Fermi energy

$$\frac{E_f}{kT} = -\frac{E_i}{2kT} - \frac{3}{4} \ln \frac{m_e}{m_h}$$

$$np = N_c N_v e^{-E_i/kT}$$

and since no assumptions were made about intrinsic or extrinsic properties in deriving equations (1) and (2), we arrive at the very useful relationship for any semiconductor that, if there is a large free electron concentration, the free hole concentration is reduced, and vice versa.

In intrinsic semiconductors, $n = p \equiv n_i$, and is defined as intrinsic concentration. We then have from above

$$n_i^2 = np = N_c N_v e^{-E_i/kT}$$

Next, consider the effect of an impurity semiconductor with N_d donors per cm^3. All the donors lie at energy E_d below the conduction band, taken as E=0, but sufficiently close so that all the conduction electrons come from the donor levels and have a density n given by the density of states and the occupation probability at energy E_d. Using the subscript e for electrons or unionized donors and h for holes or ionized donors, and with a factor of $\frac{1}{2}$ for spin degeneracy

$$(N_d)_e = N_d f(-E_d) = \frac{N_d}{1 + \frac{1}{2} e^{(-E_d/kT + E_f/kT)}}$$

$$(N_d)_h = N_d - (N_d)_e = \frac{N_d}{1 + 2e^{(-E_d/kT + E_f/kT}}$$

which is the same as the concentration of free electron carriers in this case, that is,

$$n = N_c e^{E_f/kT} = \frac{N_d}{1 + 2e^{(E_d/kT + E_f/kT)}}$$

At sufficiently low temperatures $n \ll N_d$ and by eliminating E_f/kT, we get

$$n^2 = \frac{1}{2}(N_d - n)N_c e^{-E_d/kT}$$

(3) $$n = \frac{1}{\sqrt{2}}(N_d N_c)^{1/2} e^{-E_d/2kT}$$

Similarly for acceptor concentration N_a and energy level E_a

(4) $$p = \frac{1}{\sqrt{2}}(N_a N_v)^{1/2} e^{\left(-\frac{E_a - E_i}{2kT}\right)}$$

Thus, in terms of the charge carriers, with mobility defined as $\mu = \frac{E}{v}$ the ratio of electric field magnitude to the carrier speed in the direction of the field, the conductivity for a semiconductor is

(5) $$\sigma = nq\mu_e + pq\mu_h$$

NON-EQUILIBRIUM PROCESSES

At thermal equilibrium, the concentration of electrons and holes is seen from the preceeding to be completely determined by the temperature, width of the forbidden band. and the carrier effective masses. It is possible to upset this equilibrium and increase the concentrations by means of carrier injection. This can be done by applying forward bias or allowing photon absorption. If the carrier injection is halted, the concentrations begin to decrease, eventually reaching the equilibrium value for that temperature. The process is known as recombination. The major recombination process in most semiconductors is not a direct action between electrons and holes, but rather occurs through the medium of lattice imperfections called recombination centers. They have energy levels deep within the forbidden band and capture minority carriers. The increase in charge causes majority carriers to migrate to the center and neutralize it. Another process of importance in some semiconductors is Auger recombination in which an electron and a hole recombine and give energy to an electron which pushes it into a higher energy state.

We can define a lifetime τ which is the average length of time spent by carriers in the free state. In phenomena taking place at potential barriers, for which diffusion is basic, the minority carrier lifetime is of importance. In conduction phenomena one must consider the time during

which either carrier is free.

We call the carrier density in the absence of radiation n_0 and the increase Δn_0. If the radiation is removed, the excess carrier concentration decreases in an exponential manner due to recombination as it becomes harder and harder to find available recombination sites, so that

$$n - n_0 = \Delta n_0 e^{-t/\tau}$$

$$\frac{dn}{dt} = -\frac{\Delta n_0}{\tau} e^{-t/\tau} = \frac{n_0}{\tau} - \frac{n}{\tau}$$

The first term on the right is the generation rate; the second term is the recombination rate. At equilibrium $dn/dt = 0$ and $n = n_0$.

Table II gives parameters for various semiconductors. The time derived above is for indirect recombination. There also exist direct recombination and surface recombination effects which add up to shorten the relaxation time.

Table II. Parameters of selected intrinsic semiconductors at room temperature

Material	n_i (cm^{-3})	μ_e (cm^2/volt sec)	μ_h (cm^2/volt sec)	τ (μsec)
Si	1.5×10^{10}	1700	350	1000
Ge	2.4×10^{13}	3800	1800	1000
PbS	2.9×10^{15}	800	500	100
PbSe	2×10^{17}	1200	600	2
PbTe	6×10^{16}	2000	800	30
InAs	2×10^{15}	40,000	600	5
InSb	1.8×10^{16}	60,000	600	0.2
Te	1×10^{16}	1700	500	30

SOURCES OF NOISE IN DETECTORS

Noise appearing in an infrared detector system may be categorized in two ways, namely noise inherent in the detector, and noise from the radiation background. In the detector, noise is generated by fluctuations in the concentrations and motion of current carriers.

The first form of noise to be considered is due to random motions of the charge carriers and is called thermal noise or Johnson noise. It appears in any resistive material, even a metal. At frequencies below 10^{13} Hz (10 THz) one may assume the classical $\frac{1}{2}kT$ per degree of freedom and use the derivation of noise voltage due to Nyquist.

Consider two resistors R_1 and R_2 connected together, both at temperature T. They are noninductive and noncapacitive. Each transfers noise power to the other in equal amounts since they remain at the same temperature. Putting an ideal filter of bandwidth Δf between the two resistors changes the amount of power transmitted, but not the thermodynamic equilibrium. Since R_1, R_2, Δf are all arbitrary, the power transferred in any bandwidth must be independent of resistance.

If the frequency dependence of the open circuit noise voltages is $V_{N1}(f)$ and $V_{N2}(f)$, the mean square noise voltages in bandwidth Δf are $[\overline{V_{N1}(f)\Delta f}]^2$ and $[\overline{V_{N2}(f)\Delta f}]^2$. Now put a transmission line between the resistors. The energy per mode of vibration is kT and the number of modes per unit bandwidth is $2\ell/v$ where v is the propagation speed. At equilibrium, the energy stored in the line in a bandwidth Δf will be

$$W = \frac{2\ell}{v} kT\Delta f$$

Half the energy is carried in each direction so that the power transferred into each resistor is

$$P = \frac{1}{2}\frac{W}{t} = \frac{1}{2}\left(\frac{v}{\ell}\right)\left(\frac{2\ell}{v} kT\Delta f\right) = kT\Delta f$$

The average power is also given by

$$P = [\overline{V_N(f)\Delta f}]^2 \frac{R_2}{(R_1+R_2)^2} = \frac{[\overline{V_N(f)\Delta f}]^2}{4R} = kT\Delta f$$

where $R_1 = R_2$. We note that the noise power is dependent only on the bandwidth and not on frequency, so we can specify V_N as the noise voltage in a given frequency band Δf.

$$\overline{V_N^2} \equiv [\overline{V_N(f)\Delta f}]^2 = 4RkT\Delta f$$

(6) $$V_N = (4RkT\Delta f)^{1/2}$$

(7) $$I_N = \left(\frac{4kT\Delta f}{R}\right)^{1/2}$$

The value of kT at room temperature (295°K) is 4.07×10^{-21} watt sec so for a 1 Hz bandwidth, the noise power is always 4.07×10^{-21} watts, which in a 1 ohm resistor amounts to 1.28×10^{-10} volts.

In a network involving more than one resistor, each at a different temperature, the noise voltage is considered as arising from each independently, so for R_1 and R_2 in series

$$V_N = 4k\Delta f\left[R_1T_1 + R_2T_2\right]$$

One can derive an equivalent noise temperature T_N for the whole circuit.

$$V_N = \left[4kT_N(R_1+R_2)\Delta f\right]^{1/2}$$

$$T_N = T_1\frac{R_1}{R_1+R_2} + T_2\frac{R_2}{R_1+R_2}$$

Similarly for parallel resistors

$$I_N = \left[4k\left(\frac{T_1}{R_1} + \frac{T_2}{R_2}\right)\Delta f\right]^{1/2}$$

$$I_N = 4kT_N\left(\frac{1}{R_1} + \frac{1}{R_2}\right)\Delta f$$

$$T_N = T_1\frac{1/R_1}{1/R_1 + 1/R_2} + T_2\frac{1/R_2}{1/R_1 + 1/R_2}$$

At frequencies above 10 THz, one should use the Planck theory for the energy per degree of freedom. Then

$$P_N = \frac{hf\Delta f}{e^{hf/kT}-1}$$

$$V_N = \left(\frac{4Rhf\Delta f}{e^{hf/kT}-1}\right)^{1/2}$$

The presence of capacitors and inductors does not change the thermal noise level in a circuit, since this noise is connected only with energy dissipation.

Another source of noise in semiconductors is current noise, sometimes called flicker noise, modulation noise, or 1/f noise. Some authors also use the term current noise to refer to all types of noise which depend on bias current, including generation-recombination noise. We shall use it only to refer to 1/f noise. The power spectrum is

$$\overline{I_N}^2 = K \frac{I^\alpha \Delta f}{f^\beta} \tag{8}$$

where K is a constant, $\alpha \approx 2$, and $\beta \approx 1$.

Much is known about the properties of this noise, but conflicting information has made a determination of its source a problem. Several sources displaying the observed power spectrum have been suggested. They are all associated with potential barriers, whether they exist at intergranular contacts, rectifying electrodes, semiconductor surfaces, dislocations, point contacts, or p-n junctions. The dependence of the noise voltage upon bias current suggests that current noise is the result of fluctuations in the conductivity of the material, which modulate the bias current, possible due to carrier density fluctuations.

It is impossible to represent the 1/f power dependence with a model having a single time constant. This has been attributed to a distribution of diffusion activation energies which lead to ion migration and barrier height modulation. The 1/f noise has been observed down to a frequency of 2×10^{-4} Hz, necessitating lifetimes of the order of hours. This is hard to reconcile with the magnitude of other relaxation times in solids. Experiments have led to disagreement on whether the noise has association with majority carriers, minority carriers, or both.

The temperature dependence of current noise has been shown to be slight, and indicates no definite trends. On the other hand, the magnitude of the noise voltage has been shown to depend on the dimensions of the material. If we assume $\alpha = 2$, then

$$\overline{V_N}^2 = \overline{I_N}^2 R^2 = K \frac{I^2 R^2}{f^\beta} \Delta f = \frac{C_1}{\ell A} \frac{I^2 R^2}{f^\beta} \Delta f = \frac{C_1 \rho^2 I^2 \ell \Delta f}{f^\beta A^3} \tag{9}$$

If $\alpha > 2$ the effect is even more dependent on dimensions. One may state however, that current or 1/f noise is still not very well understood, even though its properties have been extremely well catalogued.

The third basic type of noise in semiconductors is generation-recombination or gr noise. It is characterized by a power spectrum which is constant at low frequencies and decreases rapidly beyond a characteristic chopping frequency $f_1 = 1/\tau$ related to the inverse of the carrier lifetime. Generation-recombination noise is due to statistical fluctuation in the concentration of carriers. This fluctuation is inherent in any semiconductor due to the statistical processes involved in radiation absorption and carrier generation. One can thus expect gr noise even in a constant current circuit.

Since carrier current effects rather than diffusion are involved, it is necessary to separate extrinsic and intrinsic semiconductors, since the number of carriers is different in each case. We shall omit the derivation and quote only the results for the power spectrum and noise voltages, where $\bar{N}$ is the mean total number of electron carriers, τ is the mean recombination lifetime, and b is the mobility ratio of electrons to holes. The mean total number of holes is denoted by $\bar{P}$.

(10) $$\overline{I^2} = \frac{4I^2\tau}{\bar{N}\left(1+(2\pi f)^2\tau^2\right)} = \frac{4I^2\tau}{\bar{N}\left(1+\omega^2\tau^2\right)} \quad \text{(EXTRINSIC)}$$

(11) $$\overline{I^2} = 4I^2\left[\frac{(b+1)^2\bar{N}\bar{P}}{(b\bar{N}+\bar{P})^2(\bar{N}+\bar{P})}\right]\frac{\tau}{1+\omega^2\tau^2} \quad \text{(NEAR INTRINSIC)}$$

(12) $$\overline{I^2} = \frac{2I^2}{\bar{N}}\frac{\tau}{1+\omega^2\tau^2} \quad \text{(INTRINSIC)}$$

One must be careful which equation he uses. While the photoconduction effects in a semiconductor such as HgCdTe are intrinsic in that the electrons are photo-excited from the valence to the conduction band and the impurities are all temperature ionized, the donors or acceptors do provide excess carriers, which must recombine in the valence band. Thus the electrical properties may be those of an extrinsic or near-intrinsic semiconductor, and the two-carrier formula should be used.

For convenience, the functional dependence of the three types of noise is summarized in the table on the following page, where $\gamma/\bar{N}$ is the term in brackets above specifying the carrier concentration mobility ratio dependence.

TABLE III

Summary of Noise Sources

TYPE OF NOISE	FUNCTIONAL DEPENDENCE
Thermal, Nyquist or Johnson	$V_{rms} = (4RkT\Delta f)^{1/2}$ $I_{rms} = \left(\frac{4kT\Delta f}{R}\right)^{1/2}$
Current, 1/f, modulation, or flicker	$V_{rms} = \left(K\frac{I^{\alpha}}{f^{\beta}}R^2\Delta f\right)^{1/2}$ $I_{rms} = \left(K\frac{I^{\alpha}}{f^{\beta}}\Delta f\right)^{1/2}$ $\left[\begin{matrix}\alpha \approx 2 \\ \beta \approx 1\end{matrix}\right]$
Generation-recombination or gr	$V_{rms} = \left(\frac{2I^2R^2\gamma\tau\Delta f}{\bar{N}(1+\omega^2\tau^2)}\right)^{1/2}$ $I_{rms} = \left(\frac{2I^2\gamma\tau\Delta f}{\bar{N}(1+\omega^2\tau^2)}\right)^{1/2}$ $\gamma \approx 1$ intrinsic $\gamma \approx 2$ extrinsic

The spectrum of noise current is thus given by the relation

$$\overline{I^2} = \left[\frac{K_1 I^{\alpha}}{f^{\beta}} + \frac{K_2 I^2}{1+(f/f_1)^2} + \frac{4kT}{R}\right]\Delta f \tag{13}$$

as shown in figure 4, where K_1, K_2, f_1, f_c, α, β are constants and cross-over frequencies for different types of noise.

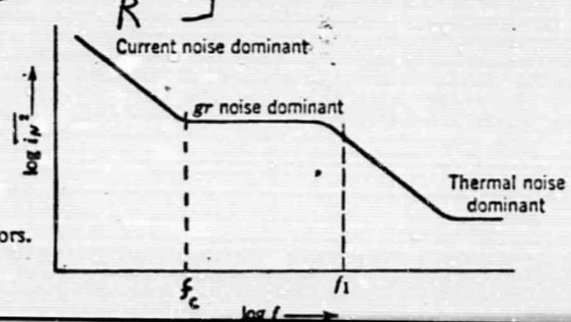


Figure 4. Spectrum of noise current in semiconductors.

We note three distinct regions in figure 4. At high frequencies only thermal noise exists. At intermediate frequencies near f_1, generation recombination noise becomes important. Below the crossover frequency f_c, current noise is predominant. In some semiconductors f_1 and f_c are reversed so one passes directly from predominant current noise to predominant thermal noise. Occasionally, thermal noise is dominant at all frequencies.

SOURCES OF NOISE IN RADIATION BACKGROUND

Light falling on a semiconductor can have several effects. If more carriers are made available, we get a change in the conductivity (photoconductivity). If the light falls on a p-n junction, the existing electric field separates the charges and produces a potential difference (photovoltaic effect). A third effect makes use of a magnetic field to separate the charge carriers (photoelectromagnetic effect). The first two are the most common of which use has been made in infrared detectors, so we shall only consider those.

We recall the excitation energy for light of wavelength λ being absorbed in a semiconductor.

$$\Delta E = hf = \frac{hc_0}{\lambda}$$

At equilibrium in a radiation field, the number of carriers equals the product of generation rate and carrier lifetime. From equation (5),

$$\begin{aligned}\sigma &= (n+\Delta n)\, q\mu_e + (p+\Delta p)\, q\mu_h \\ &= q\mu_h \left[bn + p + \Delta n\,(b+1)\right]\end{aligned}$$

where b is the mobility ratio. The change in conductivity is

$$\frac{\Delta\sigma}{\sigma} = \frac{\Delta n\,(b+1)}{bn+p}$$

In order to detect the change in σ, one supplies a bias voltage and a load resistor equal to the resistance of the semiconductor to assure maximum signal voltage per unit radiant flux. The radiation is then modulated or "chopped" in some fashion to produce an alternating current. The resulting variation in voltage across the load resistor can then be amplified.

In a photovoltaic detector, the voltage is produced directly by the action of photons without the need for a bias supply or a load resistor. Consider the energy level diagram of a p-n junction (figure 5).

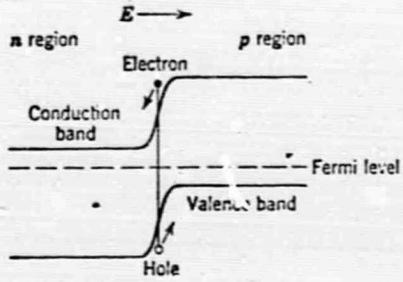


Figure 5. Carrier excitation at a *p-n* junction.

A photon of wavelength sufficiently short to cause intrinsic excitation, when absorbed at the barrier, produces a free hole-electron pair. The positive charges build up in the p-region and the negative charges in the n-region, causing a difference in potential between the faces of the semiconductor, which is detectable with a potentiometer.

One can also supply a bias to the junction in the reverse direction. In the absence of radiation, the "dark" condition, one has a characteristic junction rectification curve as in figure 6. Applying reverse bias puts one at point c. Radiation of the proper wavelength increases leakage current over the top of the potential barrier, and the resulting increase in reverse current is detected by measuring the voltage across a load resistor.

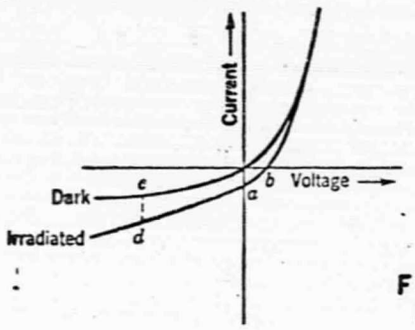


Figure 6. Rectification characteristic of a photodiode.

We have already discussed the generation of noise in the detector. It is evident that since photons come from a radiating source, their emission is somewhat random, being limited by Bose-Einstein statistics about the mean of the Planck function at their emission temperature. Thus even a perfect or near-perfect detector willbe photon noise limited in the ideal case.

The average rate at which photons are emitted in a spectral interval $d\nu$ from a blackbody is

$$\bar{N} = \frac{M(\nu, T)}{h\nu} d\nu = \frac{2\pi \nu^2 / c_0^2 \, d\nu}{e^{h\nu/kT} - 1} \tag{14}$$

where $M(\nu, T)$ is the Planck function. The noise spectrum of a fluctuation $P(t)$ in power in the frequency interval between f and f+df is

$$\overline{P(f)}^2 = 2\bar{N} \left| \int_0^\infty P(t) e^{i\omega t} dt \right|^2 .$$

Assuming that the photon emission is governed by Bose-Einstein statistics, $P(t)$ is given by the Bose distribution function and

$$\overline{P(f)}^2 = \frac{4A\pi h^2}{c_0^2} \int_0^\infty \frac{\nu^4 e^{h\nu/kT} d\nu}{\left(e^{h\nu/kT} - 1\right)^2}$$

$$\overline{P(f)}^2 = 8Ak\sigma T^5$$

where A is the source area. If the source is a grey body, the right hand side should be multiplied by the effective emissivity ϵ .

The fluctuation in number per unit bandwidth at frequency f is

$$P_N(f) = \frac{4\pi A}{c_0^2} \int_0^\infty \frac{\nu^2 e^{h\nu/kT} d\nu}{\left(e^{h\nu/kT} - 1\right)^2} = 4.17 A T^3 \times 10^{11}$$

We assume the detector is sensitive to all possible wavelengths. If it has emissivity ϵ , the net mean square power over bandwidth Δf will be

$$\overline{P^2} = 8A\epsilon\sigma k\left(T_1^5 + T_2^5\right)\Delta f \tag{15}$$

where T_1 is the detector temperature and T_2 is the temperature of the surroundings.

FIGURES OF MERIT

There are four questions to be asked for any infrared detector system:

1. What is the minimum intensity of radiant power falling on the detector which will give rise to a signal-to-noise ratio of 1?
2. What signal will be obtained per unit radiant power falling the detector?
3. How does the signal very with the wavelength of the incident radiation?
4. What is the modulation frequency response of the detector?

Question one is evidently very dependent on the spectral distribution of the radiation source, noise mechanisms of emission and detection, bandwidth, detector frequency dependence, temperature, area, intensity of incident radiation, and so on. Thus the conditions under which the figure of merit is measured must be defined as precisely as possible.

One of the most widely used figures of merit is the noise equivalent power, P_N, or NEP. The NEP is defined as the rms value of the sinusoidally modulated radiant power falling on a detector which will give rise to an rms signal voltage equal to the rms noise voltage from the detector. The reference temperature is either the laboratory ambient 22°C (295°K) or the cooled operating temperature. The reference area is usually 1 cm^2 and the radiant power and field of view should be specified. The temperature of the blackbody source, modulation frequency, and bandwidth at that frequency are specified with the NEP value as NEP(500°K, 900 Hz, 1 Hz). If A is the detector area and V_N/V_S is the noise to signal ratio, and $\mathcal{H}_{rms}$ is the incident rms irradiance, then

$$\text{NEP} \equiv \mathcal{H}_{rms} A \left(\frac{V_N}{V_S}\right) \frac{1}{\Delta f^{1/2}} \qquad (\text{Watt Hz}^{-\frac{1}{2}})$$

A second figure of merit is the detectivity, defined as the reciprocal of the noise equivalent power.

$$D \equiv \frac{1}{\text{NEP}} \qquad (\text{Hz}^{\frac{1}{2}}\ \text{watt}^{-1})$$

Another common figure of merit is the noise equivalent input, NEI, which is the irradiance required to give a signal-to-noise ratio of 1.

$$\text{NEI} = \frac{\text{NEP}}{A} = \frac{1}{AD} \qquad (\text{watt cm}^{-2}\ \text{Hz}^{-\frac{1}{2}})$$

One can define noise equivalent radiance, NER, as the noise equivalent irradiance per unit solid angle.

$$\text{NER} = \frac{\text{NEI}}{\Omega} = \frac{\text{NEP}}{\Omega A} = \frac{1}{AD\Omega} \qquad (\text{watt Hz}^{-\frac{1}{2}}\ \text{cm}^{-2}\ \text{ster}^{-1})$$

An area independent figure of merit is defined by multiplying NEI by the square root of the area. The reciprocal of this quantity is known as D^*, and has units of cm $Hz^{\frac{1}{2}}$ $watt^{-1}$. It is also called detectivity. The only way to distinguish D^* from D is to look for the extra dimension of length in D^*. Usually $D^*(T,f,\Delta f) = D^*(T,f)$ is the quantity quoted, being given with the temperature and the modulation frequency. A bandwidth of 1 is always assumed.

The current noise limited detectors whose NEP is proportional to the square root of the area, allow one to define a quantity S which is both area and frequency independent. S removes the frequency dependence by multiplying $NEP/A^{1/2}$ by the square root of the measuring frequency. Then

$$S = \frac{NEP}{A^{1/2}}(\Delta f)^{1/2} = \frac{(\Delta f)^{1/2}}{D^*} \quad (\text{watt cm}^{-1}) \tag{19}$$

Having characterized the signal-to-noise performance for given input, we consider next the second question, characterizing the signal voltage per unit radiant power. Since we are not interested in noise, we need not specify the bandwidth. Also, unless we are dealing in the neighborhood of the detector response time, the frequancy may also be omitted. We must state the radiator and detector temperatures and the detector area, however. Then we can define the responsivity $\mathcal{R}$ as the rms signal voltage per uni power falling on the detector.

$$\mathcal{R} \equiv \frac{V_s}{P} = \frac{V_s}{\mathcal{H} A} \quad (\text{volt watt}^{-1}) \tag{20}$$

The third question is relatively easy to answer. Photon detectors essentially measure the rate at which quanta are absorbed, hence the incident photons must have more than a minimum energy before being absorbed and contributing to the detector output. Because the number of photons per second per watt is directly proportional to wavelength, and their energy if absorbed is unimportant, the response of a photon detector for equal amounts of radiant power per unit wavelength interval decreases as the wavelength decreases below the minimum energy cutoff (see figure 7), while thermal detectors integrate power over the whole spectral band.

A meaningful display would be a plot of some spectral figure of merit versus wavelength, or relative responsivity per unit radiant power versus wavelength.

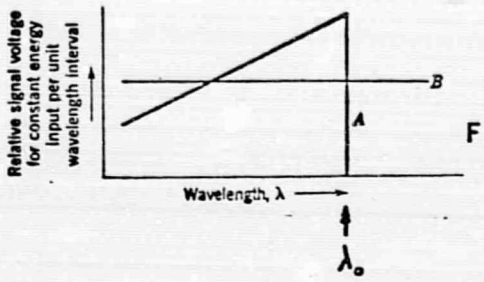


Figure 7. Comparison of idealized spectral responses of photon and thermal detectors.

In real detectors, the long wavelength limit shows a dropoff to zero over a short finite interval rather than as shown for the ideal case in figure 7. The percentage of maximum response should be given at the chosen λ_o point, since there is no agreed upon standard as to where λ_o is defined.

Finally, the fourth question concerning frequency response must be answered. What is the dependence of responsivity and detectivity upon the chopping rate? It is difficult to get a detector which has both figures of merit large over a large frequency range. Because of carrier relaxation or response time, the instantaneous carrier density cannot follow the rise and fall of sufficiently rapid modulation frequancies. Because the rise and decay times follow an exponential law, the frequency response of signal voltage for unit radiant intensity is similar to that of a low pass filter, that is,

$$(21) \qquad \mathcal{R}(f) = \frac{\mathcal{R}_o}{(1+4\pi^2 f^2 \tau^2)^{1/2}}$$

where $\mathcal{R}_o$ is the responsivity at frequency $f=0$, and $1/\tau$ is the time constant (see figure 8).

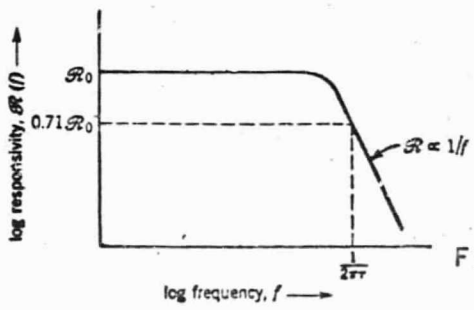


Figure 8. Frequency dependence of responsivity.

In semiconductors characterized by a simple recombination mechanism, the value of τ corresponds closely to the carrier lifetime. Thus photoconductors have response times dictated by the majority carrier lifetime, whereas photovoltaic detectors, which depend on carrier diffusion, have response times dictated by the minority carrier lifetime.

For white noise limited detectors, the frequency dependence of the detectivity and responsivity have the same form, since the noise, by definition, is frequency independent. However, for current noise limited detectors, the noise voltage has a $(1/f)^{\frac{1}{2}}$ dependence and D^* has the form

$$D^*(f) = \frac{kf^{1/2}}{(1+4\pi^2 f^2 \tau^2)^{1/2}} \tag{22}$$

where k is a proportionality constant. This is shown in figure 9. Any value of detectivity with current noise limited detectors should include a notation of the frequency at which it was measured.

The importance of specifying all the conditions under which the measured values of detector performance are determined must again be stressed. In some detectors, for example, the response depends on the wavelength of the exciting radiation. In many, it depends on the radiation intensity and the detector temperature. In another example, the spectral detectivity exhibits multiple time constant behavior depending on the chopping frequency of the incident radiation. In most cases the source intensity is important. The field of view and background temperature influence the performance of photon noise limited detectors. Even the ambient humidity has been known to be a factor.

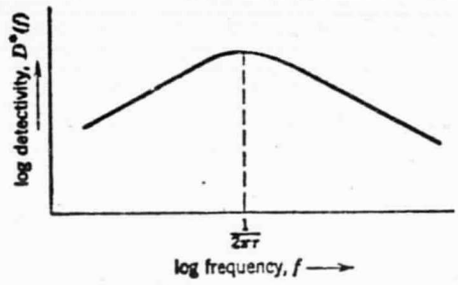


Figure 9. Dependence of D^* upon frequency for $1/f$ noise limited detectors.

NOISE LIMITED DETECTORS

As mentioned previously, photon emission is a statistical process. Thus there is inherent noise in any radiation. Secondly, since detectors are normally limited to a certain spectral region, the radiation noise from the source will only be encountered in the region of sensivity of the detector rather than over the whole band of frequencies. We shall introduce a quantum

efficiency $\eta(\nu)$, which is defined as the ratio of excited charge carriers per incident quantum.

The determination of D^* is relatively straightforward if we assume monochromatic signal radiation on a blackbody background. The average rate of arrival at a detector area A of photons of frequency ν from a background of temperature T_2 is

$$\bar{N}_t A = \frac{M(\nu, T_2)}{h\nu} A \, d\nu$$

where $\bar{N}_t$ is the average rate of arrival per unit area and $M(\nu, T_2)$ is the PLanck radiation function. The mean square deviation $\bar{N}_t^2$ in the rate of arrival per unit area, of photons, governed by Bose-Einstein statistics, is given by

$$\bar{N}_t^2 = \bar{N}_t \frac{e^{h\nu/kT_2}}{e^{h\nu/kT_2} - 1}$$

Setting $\nu_0 = hc_0/\lambda_0$ where ν_0 is the cutoff energy at which the quantum efficiency is set to zero, we determine $\bar{N}_t^2(f)$, the frequency dependence of mean square fluctuations in the rate of current carriers due to the arrival of photons in the spectral range of the detector.

$$\bar{N}_t^2(f) = A \int_{\nu_0}^{\infty} \eta(\nu) \bar{N}_t^2 \, d\nu = \frac{A}{h} \int_{\nu_0}^{\infty} \eta(\nu) \frac{M(\nu, T_2)}{\nu} \frac{e^{h\nu/kT} d\nu}{e^{h\nu/kT} - 1}$$

Introducing

$$M(\nu, T) = \frac{2\pi h \nu^3 / c_0^2}{e^{h\nu/kT} - 1}$$

gives

$$(23) \qquad \bar{N}_t^2(f) = \frac{2\pi A}{c_0^2} \int_{\nu_0}^{\infty} \eta(\nu) \frac{\nu^2 e^{h\nu/kT_2}}{\left[e^{h\nu/kT_2} - 1\right]^2} d\nu$$

Next we determine the intensity of monochromatic radiation from the signal source required to give an average rate of carrier generation equal to the rms fluctuation in the arrival of photons from the background in a bandwidth Δf. The rate of generation N_s of carriers excited by photons from the monochromatic source of power P_s incident on the detector is

given by

$$N_s = \frac{\eta(\nu_s) P_s}{h\nu_s}$$

where $\eta(\nu_s)$ is the quantum efficiency for photons of frequency ν_s .

The rms fluctuation in bandwidth Δf is given by

(24) $$N_s = \left[\overline{N_t^2}(f) \cdot 2\Delta f\right]^{1/2}$$

so that

(25) $$P_s = \frac{h\nu_s (4\pi A \Delta f)^{1/2}}{c_0 \eta(\nu_s)} \left[\int_{\nu_0}^{\infty} \eta(\nu) \frac{\nu^2 e^{h\nu/kT_2}}{\left[e^{h\nu/kT} - 1\right]^2} d\nu \right]^{1/2}$$

which is the monochromatic power from the signal source needed to give rise to a signal from the detector equal to the signal from background photon noise in the spectral interval to which the detector responds and in the bandwidth Δf.

(26) $$D_\lambda^* = \frac{A^{1/2}}{P_{s\,min}} = \frac{c\,\eta(\nu_s)}{2\pi^2 h\nu_s \left[\int_{\nu_0}^{\infty} \frac{\eta(\nu)\nu^2 e^{h\nu/kT}}{\left[e^{h\nu/kT} - 1\right]^2} \right]^{1/2}}$$

Figure 10 shows the wavelength dependence of D_λ^* assuming unit quantum efficiency, while figure 11 shows D_λ^* for actual detectors operating at 77°k.

We note from the figure that detectors having short cutoff wavelengths see less of the photon noise from the background than those having cutoff wavelengths near the minimum. Thus, as the wavelength is decreased, D_λ^* increases. As one passes the minimum and goes out to longer wavelengths, the number of photons per watt of signal power more than compensates for the noise increase, and D_λ^* again rises, provided such low energy photons can generate charge carriers in the detector.

One must distinguish in photon detectors between the various photon detecting mechanisms. The above equation is good for photovoltaic detectors, where only background fluctuations contribute to the noise. On the other hand, the photoconductive detectors depend upon change in charge carrier concentration, which introduces other sources of noise. It can be shown that at equilibrium in a photoconductor, the total noise power due to photon and gr noise is greater than twice the photon noise alone. Thus the D_λ^* values in this section and in figure 10 must be divided by $\sqrt{2}$ when applied to photoconductors.

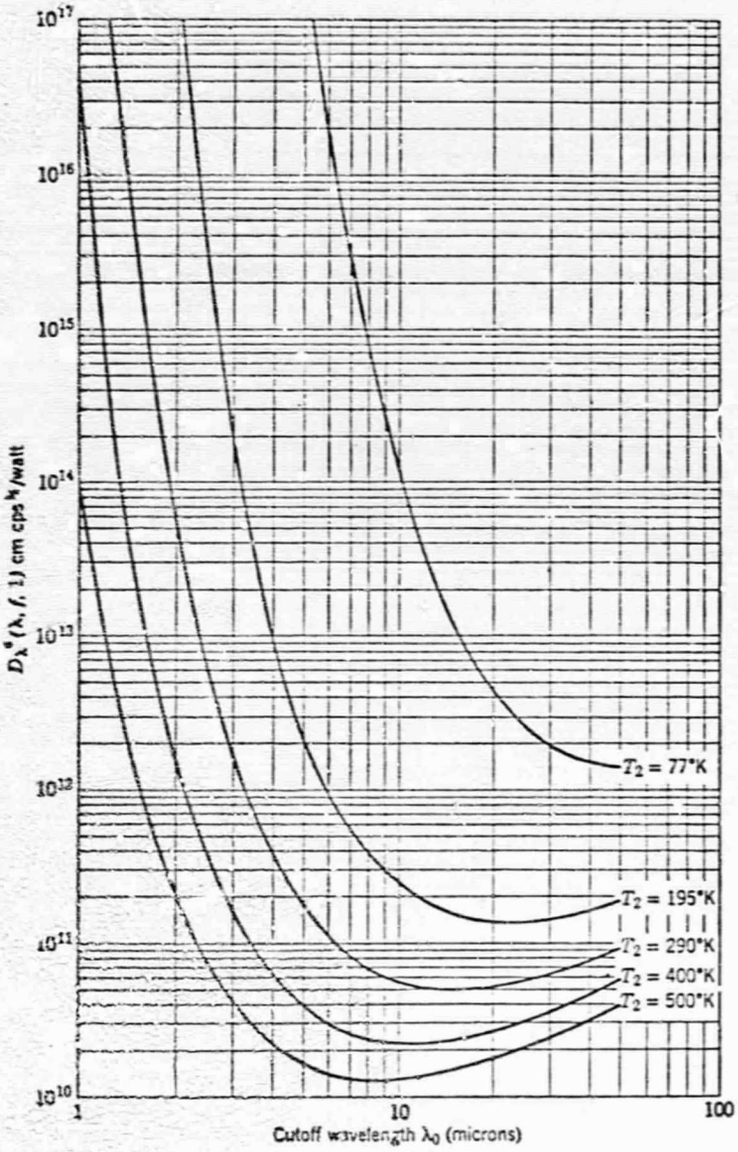


Figure 10. Photon noise limited D^* at peak wavelength, assumed to be cutoff wavelength, for background temperatures of 77°K, 195°K, 290°K, 400°K, and 500°K. Viewing angle of 2π steradians and unit quantum efficiency.

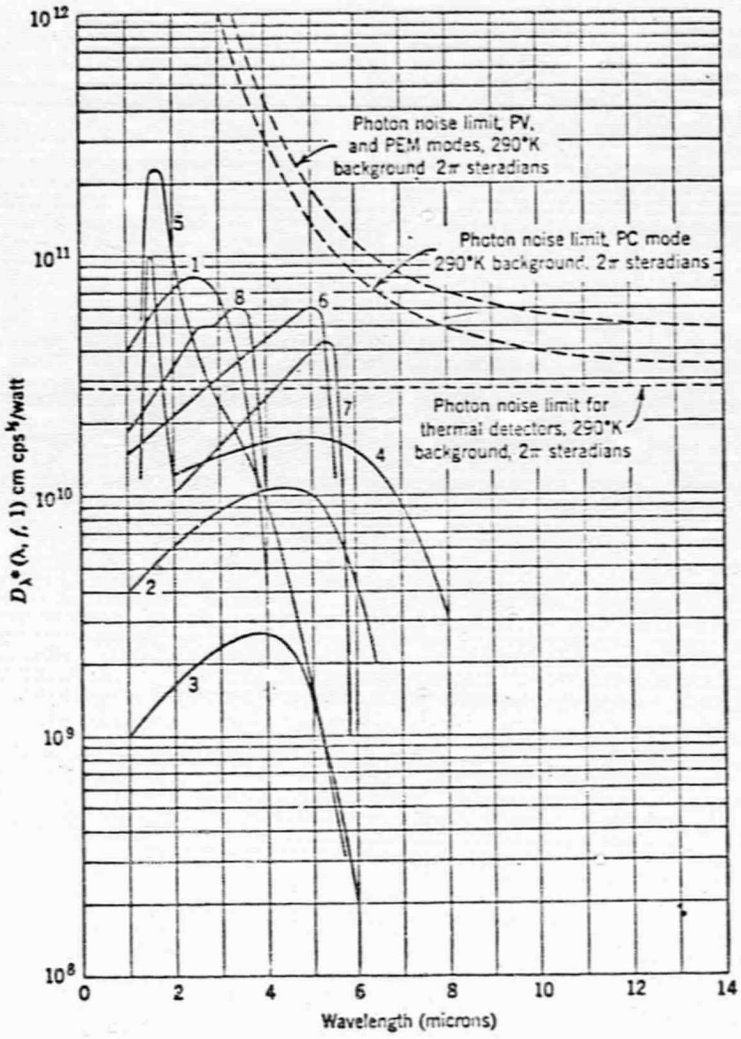


Figure 11. Spectral D_λ^* of detectors operating at 77°K. 1. PbS, PC (90 cps). 2. PbSe, PC (90 cps). 3. PbTe, PC (90 cps). 4. Ge:Au, PC (900 cps). 5. Ge:Au,Sb, PC (90 cps). 6. InSb, PC (900 cps) (60° field of view). 7. InSb, PV (900 cps). 8. Te, PC (900 cps).

There is yet another problem. In photovoltaic detectors, photons arriving from the back hemisphere will be absorbed before reaching the junction, and therefore will have no effect. In photoconductors however, photons arriving from any direction will produce charge carriers, so we have another $\sqrt{2}$ to divide by as the noise power is doubled unless the detector is sufficiently cooled to reduce the radiation from other directions to negligible quantities.

Up to now we have considered monochromatic noise limited detectors. Next we must consider the signal as arising from a blackbody, (1), instead of being monochromatic. We wish to determine the rate of absorption of photons emitted from a blackbody at temperature T_3, having wavelengths in the spectral sensitivity range of the detector. This is given by

$$N_s' = \int_{\nu_0}^{\infty} \eta(\nu) \frac{M(\nu, T_3)}{h\nu} d\nu.$$

Where P_s' is the total radiation from the blackbody, we express the photon absorption rate from the signal as

$$N_s' = \frac{1}{h} \int_{\nu_0}^{\infty} \frac{\eta(\nu) \frac{M(\nu, T_3)}{\nu}}{\sigma T_3^4} d\nu \, P_s'$$

We can express N_s' in terms of N_s

$$N_s' = G N_s \frac{P_s'}{P_s}$$

where

$$G = \frac{\nu_s}{\sigma T_3^4 \eta(\nu_s)} \int_{\nu_0}^{\infty} \frac{\eta(\nu) M(\nu, T_3)}{\nu} d\nu. \tag{27}$$

The rate of absorption of signal photons can be equated to the rms value of photon noise in the bandwidth Δf.

$$P_{s\,min}' = \frac{P_{s\,min}}{G}$$

$$D^* = D_\lambda^* G \tag{28}$$

Figure 12 shows G for several different blackbody sources, while figure 13 plots D^*. Note that D^* is much more level in the 3-7 micron

region. The detector is less susceptible to disturbance from background radiation, but is less able to see the photons from the signal source since as λ_o decreases, the detector responds to a smaller and smaller fraction of the photons. At wavelengths shorter than 3 microns, the reduction in the number of photons from the source more than equals the reduction in photon noise from the background and $\bar{J}^*$ drops as λ_o decreases.

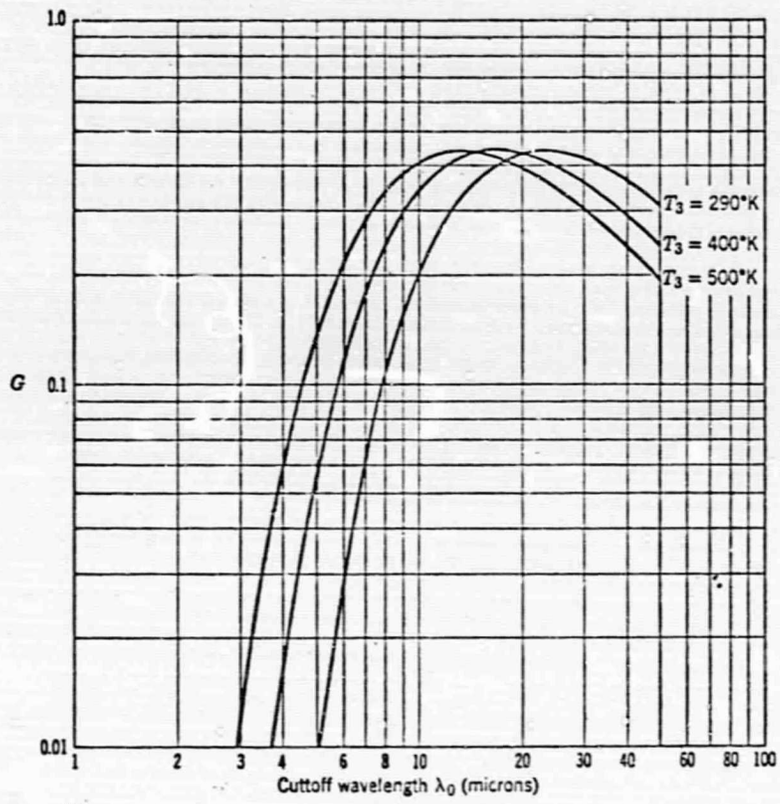


Figure 12. Dependence of the function G upon detector cutoff wavelength λ_0 for black body source temperatures of 290°K, 400°K, and 500°K and viewing angle of 2π steradians.

Note the surprising crossover in figure 13. Photon noise limited detectors with cutoff wavelengths above 18 microns can detect a 290°K source against a 290°K background noise more readily than a 500°K source against the same background. Since the limiting noise is the same in both cases, the explanation must lie in the relative number of photons per watt from the signal source. The 290°K source must emit more photons per second per watt above 18 microns than the 500°K source. Thus the detectivity, or ratio of absorbed signal photons per second to the total source power, must be higher for the 290°K source in this wavelength region.

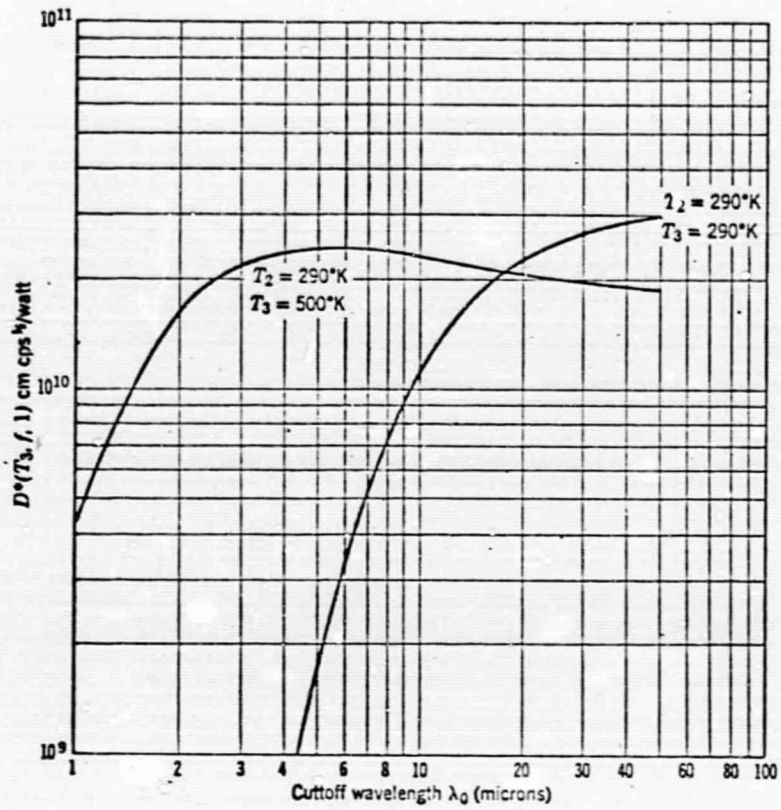


Figure 13. Photon noise limited D^* for 290°K and 500°K source temperatures as a function of detector cutoff wavelength λ_0 for a background temperature of 290°K, viewing angle of 2π steradians, and unit quantum efficiency.

FILTERING AND SHIELDING

If we are looking at a monochromatic source against a background, a good way to improve the performance of photon noise limited detectors is to use a narrow band optical filter on the same wavelength. In this way we make use of the monochromatic detectivity, whichis always higher than the blackbody detectivity. Obviously the filter must be cooled or it will radiate at the same temperature as the background. This filtering can yield improvements of several orders of magnitude for short wavelengths, as can be seen in figures 10 and 12. Even when looking at a black or grey source, one can often reduce background noise by using a cold filter

restricting the view to the region of maximum source radiation. Just as in matching the amplifier impedance to the detector, the ultimate performance of photon noise limited detectors is obtained by getting a proper match between spectral emission from the source and spectral response of the detector for any given background temperature.

Another way of avoiding background noise is by viewing as little background as possible. Viewing the target through a cooled aperture which is barely large enough can make an order of magnitude difference or more in the detectivity. Figures 14 and 15 emphasize this fact.

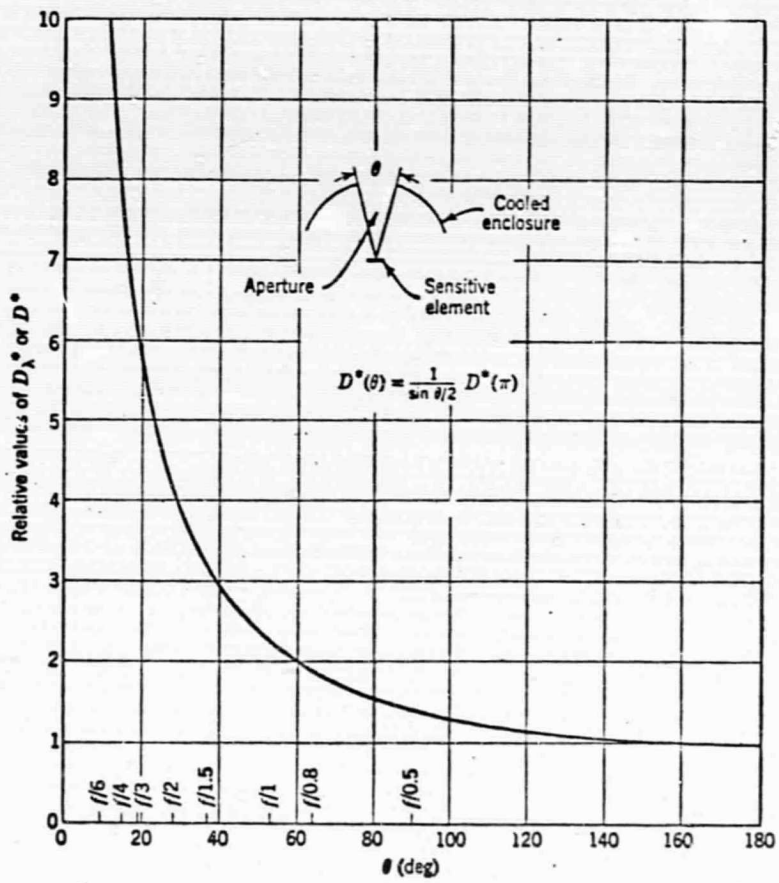


Figure 14. Relative increase in photon noise limited D_A^* and D^* achieved by using cooled aperture in front of Lambertian detector.

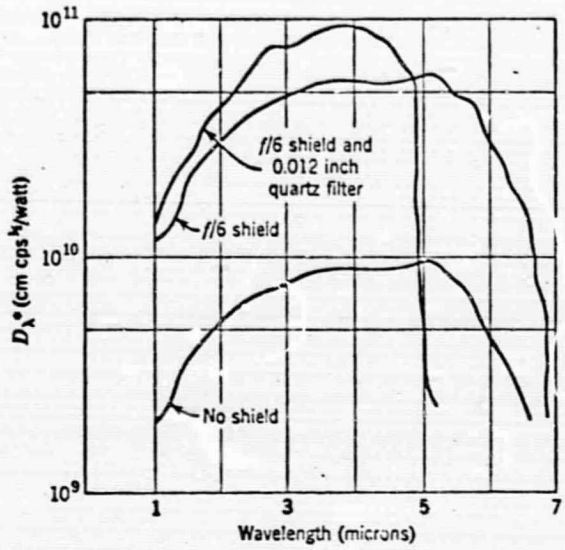


Figure 15. Improvement in lead selenide photoconductor obtained through use of cooled filter. (Courtesy Santa Barbara Research Center and *Missiles and Rockets.*)

PHOTON NOISE LIMIT WHEN VIEWING THE EARTH THROUGH THE ATMOSPHERE

Radiation from the earth at about 290°K peaks at 10 microns. Variations in average temperature from point to point are of the order of a fraction of a degree Kelvin. The performance of a thermal mapping system is determined, therefore, by the ability of the detector to discriminate scene variations of this size. Figure 13 shows the appropriate $D^*(290°K, f, 1)$ when there is no atmospheric absorption.

Where absorption is great, we will see noise but no signal. Where absorption is small, as in the 8-14 micron band, we see both signal and noise. The rate of absorption of signal photons is

$$N_{t1} = 0 \qquad (\lambda < 8\mu)$$

$$N_{t2} = \int_{\nu_1 = \frac{hc}{14\mu}}^{\nu_2 = \frac{hc}{8\mu}} \eta(\nu) \frac{M(\nu, T_2)}{h\nu} d\nu \qquad \left(8\mu < \lambda < 14\mu, \ \nu = \frac{hc}{\lambda}\right)$$

$$N_{t3} = \int_{\nu_3=\infty}^{\nu_2=\frac{hc}{14\mu}} \eta(\nu)\,\frac{M(\nu,T_2)}{h\nu}\,d\nu \qquad \lambda > 14\mu$$

$$N_t = G_t N_s \frac{P_t}{P_s}$$

where

(29) $$G_{t1} = 0 \qquad (\lambda < 8\mu)$$

(30) $$G_{t2} = \frac{\nu_s}{\sigma T_2^4 \eta(\nu_s)} \int_{\nu_1}^{\nu_2} \eta(\nu)\,\frac{M(\nu,T_2)}{\nu}\,d\nu \qquad (8\mu < \lambda < 14\mu)$$

(31) $$G_{t3} = \frac{\nu_s}{\sigma T_2^2 \eta(\nu_s)} \int_{\nu_3}^{\nu_2} \eta(\nu)\,\frac{M(\nu,T_2)}{\nu}\,d\nu \qquad (\lambda > 14\mu)$$

Using $D^* = D^*_\lambda(\lambda, f, 1)$, we get the curve shown below in figure 15.

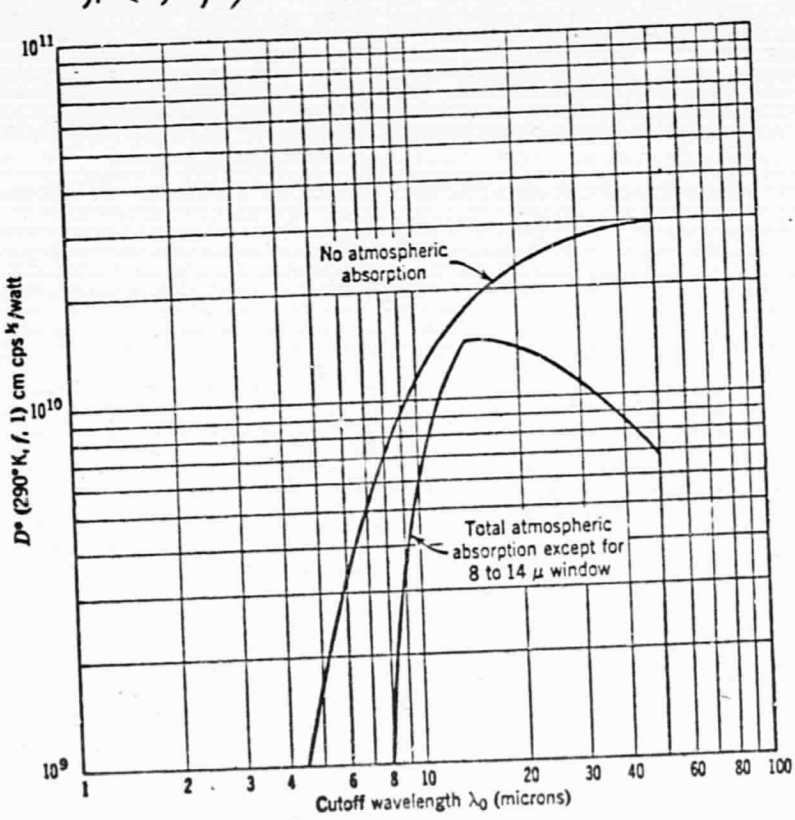


Figure 16. Photon noise limited $D^*(290°K, f, 1)$ for 290°K background temperature as a function of detector cutoff wavelength λ_0. Illustrating the effects of atmospheric absorption. Viewing angle of 2π steradians and unit quantum efficiency.

Detectors having long wavelength limits much greater or less than 14 microns will not perform well as photon noise limited detectors in the 8-14 micron atmospheric window. Below 8 μ there is only noise and no signal. Above 14 μ we have the full signal, but keep integrating noise, which gradually degrades D^*.

SIGNAL FLUCTUATION LIMIT

It is possible for a detector to be limited by the signal fluctuations rather than the noise from the background. The emission of photons can be considered an almost random process, limited by Bose-Einstein statistics. During a time interval t, let $\bar{N}$ be the average number of signal photons absorbed by the detector. The probability $P(\alpha, \bar{N})$, of detecting α photons in equal time is given by the Poisson distribution function

$$P(\alpha, \bar{N}) = \frac{\bar{N}^{\alpha} e^{-\bar{N}}}{\alpha!}$$

If we require 99% probability of detecting at least one photon, there is a 1% chance of finding no photons.

$$P(0, \bar{N}) = 0.01 = e^{-\bar{N}}$$

$$\bar{N} = \ln 100 = 4.61$$

The minimum detectable power is $P_{s_{min}}$,

$$(32) \qquad P_{s\,min} = \frac{4.61}{t} \frac{hc_0}{\lambda} = 9.22 \frac{hc_0}{\lambda} \Delta f$$

where λ is the photon wavelength and the bandwidth $\Delta f = 1/2t$. Note that the minimum detectable power is independent of area and depends linearly on bandwidth. Thus it is meaningless to use D^*, which is proportional to $(A\Delta f)^{1/2}$. Rather, one should compare P_s with the minimum detectable monochromatic power for photon noise from a radiating background. This can be obtained from D^*_λ.

One can see from figure 10 that $P_{s\,min}$ decreases very rapidly as one goes to short wavelengths in the visible and near infrared. Moreover, short observation times (large bandwidth) tend to cause signal fluctuations to dominate over photon noise from a 290°K background.

MEANS OF ACHIEVING PHOTON NOISE LIMITED PERFORMANCE

The approach to photon noise limited performance in photon detectors is through reduction of all internal noise sources. The quantum efficiency should approach unity. Antireflection films can be used to increase the absorptance, especially in materials with a high index of refraction. Current and thermal noise can usually be made small. Elimination of current noise is a problem in technology involving, for example, methods of preparing ohmic contacts to crystals. Thermal noise, arising from interchange of energy between the detector and its surroundings, can be suppressed as follows.

From table III on page 12, the thermal noise voltage is

$$V_{N,th} = (4kTR\Delta f)^{1/2}$$

and the generation-recombination noise voltage is

$$V_{N,gr} = I_N R = \frac{2I\tau^{1/2}}{N^{1/2}} R(\Delta f)^{1/2}$$

where I is the bias current, N the total number of charge carriers, and τ the recombination lifetime. The ratio of the two voltages in terms of area A, mobility μ, and carrier concentration n, is

$$\frac{V_{N,th}}{V_{N,gr}} = \frac{(qkT)^{1/2} n A \mu^{1/2}}{I \tau^{1/2}}$$

Therefore, to reduce thermal noise to drop below gr noise, we require temperatures which are low, as well as low mobility and low cross sectional area. This indicates slightly p-type material. One cannot increase bias current because of Joulean heating.

The manner in which gr noise can be suppressed below photon noise can be accomplished in two ways: by reducing lattice temperature to excite fewer carriers, and by increasing the lifetime associated with phonon excitation until it exceeds that due to photon excitation. Both ways are used, with the latter of extreme importance.

We shall assume the material is sufficiently pure and free from crystal defects so that the lifetime due to imperfection recombination is long. Then one only has to worry about keeping the Auger radiative recombination lifetime longer than the gr lifetime by varying temperature and purity.

The following example illustrates the approach. We shall construct an intrinsic detector operating in the 8-14 μ atmospheric window with 14.5 μ cutoff wavelength. We wish a 1 μsec response time and operation with a minimum amount of cooling.

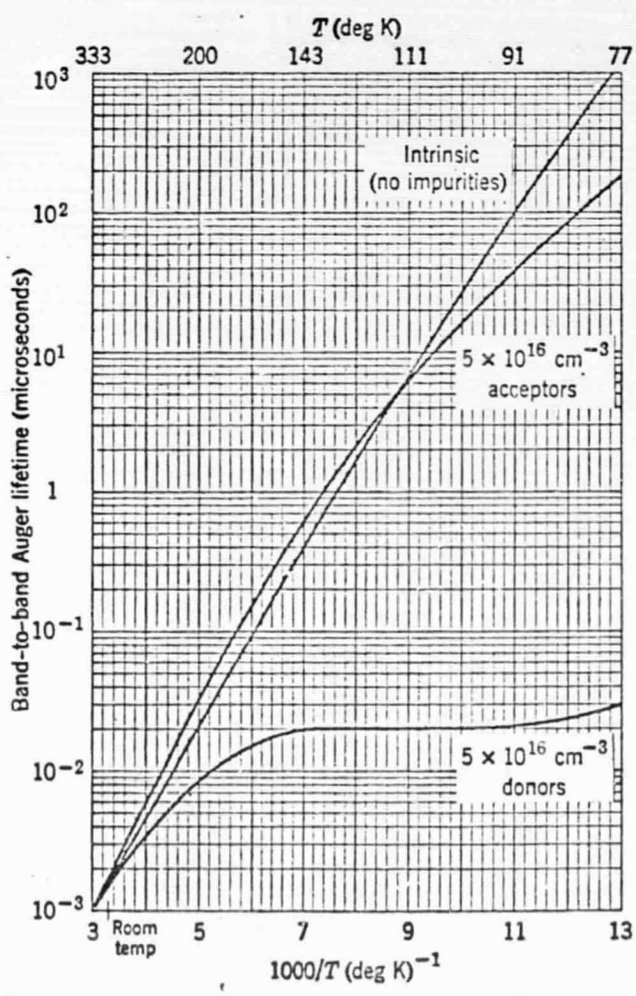


Figure 17. Dependence of Auger lifetime upon temperature for assumed semiconductor. (After Blakemore[25])

The energy gap required for 14.5 micron response is 0.086 ev. A look at narrow gap semiconductors indicates we can expect the effective masses of electrons and holes to be respectively, .05 and .20 electron masses. The dielectric constant will be around 20. One can then construct the plots of lifetime and temperature shown in figures 17 and 18.

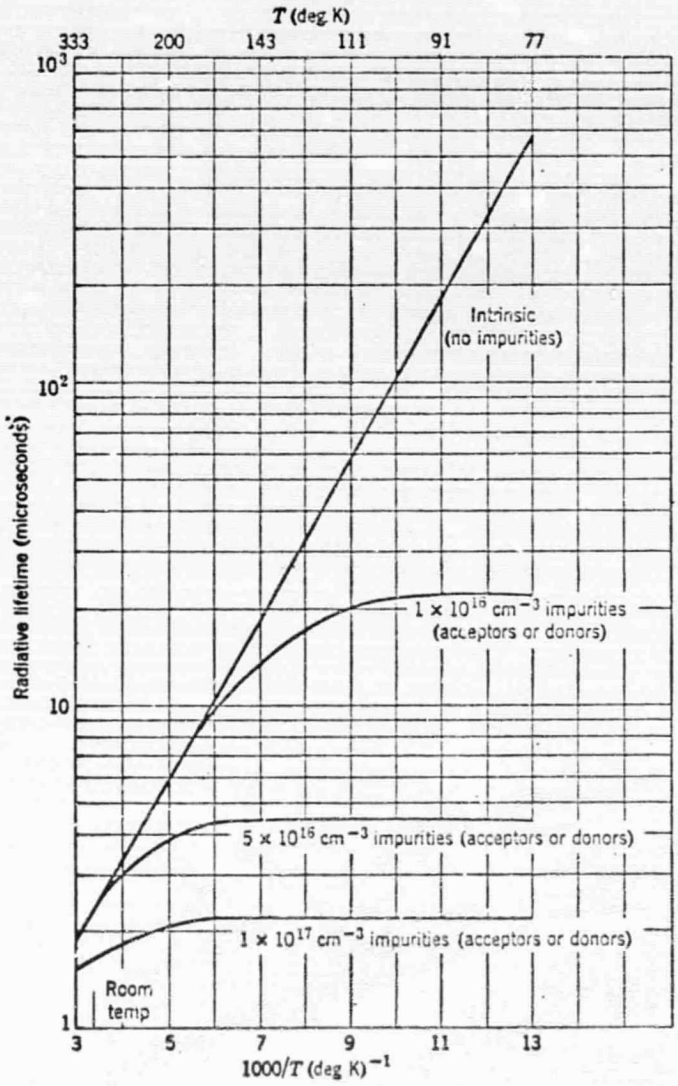


Figure 18. Dependence of radiative lifetime upon temperature for assumed semiconductor. (After Blakemore.[33])

It can be seen that for intrinsic material, the room temperature radiative lifetime is 2 μsec whereas the Auger lifetime is .001 μsec. Since the Auger lifetime has greater slope, one eventually reaches a low enough temperature where the radiative lifetime dominates. For intrinsic material, this occurs at 77°K where the radiative lifetime is 600 μsec and the Auger lifetime is 1000 μsec. This is too long a response time. We see, however, that with 5×10^{16} acceptors/cm^3 (p-type material), the radiative lifetime is dominant at 110°K with a value of 5 μsec. Thus, in practice, one has to choose the material with the proper energy gap and perform calculations based upon known values of effective mass and index of refraction and dielectric constant. So armed, one can choose the optimum conditions to satisfy the operational requirements.

One final note regarding choice of materials. Medium resistance (100 ohms - 1 megohm) offers use of the most simple amplifier techniques, but lower resistance takes advantage of the $\sqrt{TR}$ noise dependence. Very high resistances have associated problems with microphonics and electrical pickup, and because of the RC time constant, have reduced frequency response. One can expect resistance of a semiconductor to increase as temperature is lowered. Thus, in most applications where low noise is desired, one is justified in choosing as low a resistance as is compatible with operating temperature, detector resistivity, detector shape and size, and amplifier technology.