FSTC-HT-23-1114-70

US ARMY FOREIGN SCIENCE AND TECHNOLOGY CENTER

HANDBOOK OF RADAR ENGINEERING FUNDAMENTALS

by

V. Ya. Tsylov, et. al.

Country: USSR

This document is a rendition of the original foreign text without any analytical or editorial comment.

698

TECHNICAL TRANSLATION

FSTC-HT-23-1114-70

ENGLISH TITLE: HANDBOOK OF RADAR ENGINEERING FUNDAMENTALS

FOREIGN TITLE: SPRAVOCHNIK PO OSNOVAM RADIOLOKATSIONNOY TEKHNIKI

AUTHOR: V. Ya. Tsylov, et. al.

SOURCE: Military Publishing House of the Ministry of Defense of the USSR, 1967, Moscow

Translated for MID by Translation Consultants, Ltd

NOTICE

The contents of this publication have been translated as presented in the original text. No attempt has been made to verify the accuracy of any statement contained herein. This translation is published with a minimum of copy editing and graphics preparation in order to expedite the dissemination of information. Requests for additional copies of this document should be addressed to the Defense Documentation Center, Cameron Station, Alexandria, Virginia, ATTN: TSR-1.

This translation was accomplished from a xerox manuscript. The graphics were not reproducible. An attempt to obtain the original graphics yielded negative results. Thus, this document was published as is, in order to make it available on a timely basis.

BLANK PAGE

Table of Contents

Page

Page

Page

Page

Page

CHAPTER XIV. ELECTRICAL AND RADIO MEASUREMENTS

CHAPTER XV. CYBERNETICS AND DIGITAL COMPUTERS

CHAPTER XVI. FUNDAMENTALS OF GENERAL INFORMATION THEORY

CHAPTER XVII. RADAR SYSTEMS AND ANTIMISSILE DEFENSE CHARACTERISTICS

Page

Foreword

The rapid progress made in radar is associated with the considerable expension that has taken place in its field of application, as well as with the uninterrupted introduction of new achievements derived from science and engineering. The development of radar, and the complexity of the tasks radar can perform, makes it necessary to modernize the means used to realize radar information, and for the control of equipment and troops.

Officers engaged in the combat operation of radio frequency installations and in the training of personnel, are aware of the lack of literature taking a practical approach to their problems, literature that will increase their qualifications.

The purpose of this Handbook is to help officers to complete, and expand, their special knowledge.

The first eleven, and the fourteenth chapters in the Handbook contain information on the fundamentals of radar and on the installations that make up radar systems. The other chapters discuss the installations in the systems used for processing information, for control, and for communications. The examples cited in the Handbook are hypothetical in nature, and are presented for purposes of illustration only.

Chapter I was written by V. Ya. Tsylov, chapters II and VII by A. M. Pedak, Chapter II and the second part of Chapter IV by Candidate of Technical Sciences, Docent F. R. Kholyavko, the first part of Chapter IV by Candidate of Technical Sciences, Docent M. Z. Chashnik, Chapter V by Candidate of Technical Sciences, Docent K. S. Labets and Ye. S. Batenin, Chapter VI by Candidate of Technical Sciences I. K. Tregub, Chapter VIII by Candidate of Technical Sciences B. I. Kurilin, Chapter IX by Candidate of Technical Sciences, Docent A. S. Magdesiyev, Chapter X by Candidate of Technical Sciences, Docent G. M. Rakovchuk, Chapter XI by P. I. Baklashov, Chapter XII by Candidate of Technical Sciences, Docent V. G. Koryakov, Chapter XIII by Candidate of Technical Sciences, Docent L. L. Barvinskiy, Chapter XIV by Candiddate of Technical Sciences, Docent A. K. Krishtafovich, Chapter XV by Doctor of Technical Sciences Ye. N. Vavilov, and Candidate of Technical Sciences Yu. A. Buzunov, Chapter XVI by A. M. Mikhaylov, and Chapter XVII by A. S. Kucherov.

Where the significance of materials to be used was the same, the authors gave precedence to new questions, and for this reason the Handbook does not include certain types of information. Readers will be able to find such information in the literature recommended in each of the chapters.

Since the Handbook includes an extremely broad scope of questions a great many open source Soviet and foreign publications were used in its compilation, but a complete list would take up entirely too much space. We wish to express our deep appreciation to the authors whose works helped us, in one way or another, in our work of compiling this Handbook.

Please address testimonials and comments concerning the Handbook to Moscow, K-160, Military Publishing House.

The Authors

Letter Designations

1. Designation of physical magnitudes

a height of rectangular waveguide; value of measured magnitude

A number of correctly used symbols

b width of rectangular waveguide

B flux density; susceptance

c velocity of propagation of radio waves; velocity of light in a vacuum

C capacity; information reception rate; capacity of a communication channel; dynamic error coefficient

d diameter; distance

d_s spot diameter

D radar range, range to target; dynamic range; scale of range indicator scale; coupler directivity

D_B horizon range (refraction taken into consideration)

e electron charge; instantaneous value of electromotive force (emf)

E electric field strength; constant voltage

$E_{rec\ min}$ - real sensitivity of receiver to voltage

$\dot{E}$ complex amplitude of electric field strength

f high frequency

f_d limits of frequency deflection (deviation)

F low frequency

F_D Doppler frequency

F_p pulse repetition frequency

F_M modulation frequency

$F(\epsilon)$ directivity pattern

g real conductivity; antenna gain

G directive gain

$G(\omega)$ spectral density of noise power

h Plank's constant ($6.6 \cdot 10^{-36}$ joules/meter); height; altitude; transistor parameter

H target altitude; magnetic field strength; excess of signal over noise

I information content; current strength

I_a continuous anode current; pulse anode current

j_{beam} - cathode ray tube beam current density

k proportionality coefficient; gain factor; transmission coefficient; Boltzmann's constant ($1.38744 \cdot 10^{-23}$ joules/degree)

K_{tw} traveling wave ratio

K_{sw} standing wave ratio

K_f frequency shift(ing) coefficient

K_φ phase shift(ing) coefficient

$K(m)$	transfer function
l	length
L	inductance
L_s	length of scanning line on indicator screen
m	amplitude modulation factor; test number; integer
m_j	frequency modulation index
M	linear scanning scale; moment
n	index of refraction; integer
n_A	antenna rpm
n_p	pulse transformer transformation ratio
$n(t)$	number of elements rejected in time t
N	receiver noise figure; number of molecules (atoms) for the energy level; number of elements in a system
N_{turns}	number of turns in a winding
N_p	number of pulses
p	probability; reflection factor
P	probability; power
P_A	total power input to antenna
P_{rad}	power radiated by radar set (transmitter)
P_p	pulse power
P_0	input power
P_Σ	power radiated by antenna
$P_{rec\ min}$	real receiver sensitivity
$P'_{rec\ min}$	limit of receiver sensitivity
P_{fa}	false alarm probability
P_{cd}	probability of correct detection
P_{tn}	transmitter noise power
Q	quality factor; duty; report number
$Q(t)$	probability of failure
r, R	radius; resistance
R_A	antenna radiation resistance
R_g	DC generator resistance
R_e	radius of the earth
R_{load}	load resistance
R_{eff}	effective radius of the earth
R_{nt}	equivalent noise resistance of an electron tube
R_i	internal resistance of a tube
$R(\tau)$	correlation function
S	tube slope
S_A	capture area of antenna
S_g	geometric area (aperture) of the antenna
S_c	area of core section

$S_s(\omega)$	signal spectral density
t	time
t_r	recovery time
t_d	discreteness of range scale markers
t_{delay}	signal delay time
t_f	time to transmit one figure
T	period; absolute temperature
T_r	average recovery time
T_c	charge period
T_p	pulse recurrence period
T_m	modulation period
T_0	time to failure; period of natural oscillations
u	instantaneous value of voltage
U_{a2}	voltage across the second anode of a cathode ray tube
U_m	voltage amplitude
U_{per}	permissible voltage
U_{in}	incident wave voltage
U_{break}	breakdown voltage
v	rate of propagation of electromagnetic waves
v_s	scope beam sweep speed
v_r	radial speed of target
v_p	wave phase velocity
$v(\tau)$	probability of recovery in time τ
V	volume
V_c	communication channel capacity
V_s	signal volume
w	instanteneous value of energy
$w(x)$	probability density for the magnitude x
x	random error in coordinate measurement
X	reactance; true value of measured magnitude; coordinate (of an index, of a target)
Y	complex conductance; coordinate (of an index, of a target)
Y_e	established value for a regulated magnitude
Z	impedance
α	circuit decay coefficient; transistor current gain factor; phase change coefficient or wave number; target bearing; weight factor
β	azimuth; wave decay coefficient in a transmission line; current gain factor for a transistor
γ	propagation factor ($\gamma = \beta + j\alpha$)
γ_f	frequency tuning rate

δ	relative instrument error
δR	range resolution
δH	height resolution
δV	resolving volume
$\delta\beta$	bearing resolution
$\delta\varepsilon$	elevation resolution
Δ	measurement error
ΔB	increment of magnetic induction
ΔF	receiver pass band
ΔX	deflection of X coordinate
$\Delta\beta$	radar azimuth search sector
$\Delta\varepsilon$	radar elevation search sector
ε	target elevation; dielectric constant
η	efficiency
η_A	antenna efficiency
η_F	antenna feeder installation efficiency
λ	wavelength; element failure rate (in a system)
λ_w	wavelength in pipe
λ_{crit}	critical (limiting) wavelength in pipe
μ	permeability; tube gain
ν	signal-noise ratio
ν_o	signal-noise ratio for optimum signal processing
ν_c	classification factor
Π	power flow density; direction finding capability
ρ	wave (characteristic) impedance
σ	mean square error; mean square deviation from average service life
σ_{tar}	target effective echoing area
σ_R	mean square error in range measurement
σ_β	mean square error in azimuth measurement
τ	circuit time constant; pulse duration
τ_p	pulse length
τ_{pc}	pulse length after compression
τ_d	pulse delay time
τ_c	channel occupancy time
τ_s	signal length
τ_{co}	duration of pulse cut-off (decay)
τ_f	duration of pulse front
φ	phase angle; reflection factor phase
φ_{co}	coherent oscillator phase angle

φ_{lo}	local oscillator phase angle
$\varphi_{0.5}$	width of antenna radiation pattern in the horizontal plane at the 0.5 power level
Φ	magnetic flux
ψ	angle of refraction; dephasing
ω	angular frequency
ω_0	resonant frequency
ω_c	carrier frequency

2. Abbreviated subscripts

a	anode
A	antenna
b	base, beat
in	circuit (element) input
out	circuit (element) output
g	generator
per	permissible value
g	grid
c	cathode
ch	channel
co	collector
max	maximum value
min	minimum value
opt	optimum value
n	noise
thr	threshold value
s	signal
av	average
st	stable
est	established
sp	specific
f	feeder, filter
ph	phase
sh	shunt
sc	scale
em	emitter
el	element
eq	equivalent
eff	effective
m	amplitude

Chapter I

Radar Fundamentals

1.1 Radar and Its Forms

Radar is that field of radio engineering that uses reflections, re-radiations or self-radiations of electromagnetic waves, to detect various targets (objects), as well as to measure their coordinates and movement parameters.

A radar target is understood to mean any material object capable of detection, and its position and movement parameters measured by radar methods.

Radar can be subdivided into active, active with active response, semi-active, and passive.

Active radar operation means the irradiation of a target by electromagnetic energy radiated by a radar antenna, and the reception of the energy reflected from the target.

Active radar operation with active response differs from the previous definition in that a responder, a transceiver, is installed on the target and responds to signals from the radar.

Semi-active radar operation differs from active operation in that the target is irradiated by a single radar (a radar located on the ground, for example), but the reception and detection of the signals reflected from the target are through the medium of another object (a rocket, for example).

Passive radar use means the reception of energy radiated from the target.

From the radio engineering point of view, the problem of detecting some particular target reduces to detecting a signal radiated, or reradiated, from that target against a background of a different type, of noise.

Any target irradiated by a radar becomes a source of secondary radiation. The power of the secondary radiation depends on many factors, such as the field strength created by the radar near the target, target parameters (dimensions, shapes, and electrical properties), the position of the target relative to the radar, polarization of the primary field, and wavelength.

Passive radio detection is based on the phenomenon of the radiation of electromagnetic energy by any physical body, the temperature of which is above absolute zero. Since all targets satisfy this condition, they can be detected without preliminary irradiation.

Radar is based on the properties of radio waves to be propagated in a homogeneous medium on a straight line, and at constant velocity. These properties of radio waves are what make it possible to establish the direction to the target, as well as the length of the path over which they are propagated. Radar is therefore subdivided into radio range finding and radio direction finding.

Radio range finding is the determination of range to a target by measuring the length of the path over which radio waves are propagated to the target and return.

Radio direction finding is the determination of the direction to the target, that is, measuring the angular coordinates of the target.

1.2 Radio Range Finding Methods

Determination of the range to the target (D) involves the measurement of the delay in the reflected signal with respect to the main pulse.

The main pulse (signal) is the pulse (signal) of high-powered, high-frequency electromagnetic energy formed by the transmitter and radiated into space by the antenna.

The instant at which the main pulse is radiated is taken as the origin of the reading of the time of propagation of the radio wave.

The reflected signal (pulse) is the signal (pulse) of electromagnetic energy reflected from the target and received by the receiver.

The interval of time between the instant the main pulse is radiated and the instant the reflected signal is received is called the reflected signal delay time (t_d)

$$t_d = 2D/c. \tag{1.1}$$

From whence

$$D = ct_d/2, \tag{1.2}$$

where

D is the distance between the radar and the target;

c is the velocity of propagation of radio waves.

The real medium is not strictly homogeneous, so the path covered by the radio waves will not be a straight line, strictly speaking, nor will the rate at which the radio waves are propagated be strictly constant over the entire propagation path. However, the relationships cited are correct for the real medium as well if c is understood to be the average value for the rate of propagation of the radio waves over distance D ($c \approx 3 \cdot 10^8$ meters/second). (According to the last measurement, the speed of light in a vacuum is $c = 299792 \pm 0.4$ km/second).

There are various methods of radio range finding, and these depend on the method used to measure the time interval, t_d, such as pulse, frequency, phase, and frequency-pulse. Radar methods differ in the same way.

1.3 The Pulse Method in Radar

Let us consider the principle upon which the pulse radar functions by using the block schematic shown in Figure 1.1. The electronic range finder transmitter radiates superhigh frequency oscillations in the form of main pulses that repeat periodically. Reception of reflected pulses takes place in the time interval between main pulses. Pulses received are fed from the receiver output to the indicator, where the interval of time between the beginning of the radiation of the main pulse and the beginning of reception of the reflected signal (1.1) is measured and, as a result, the distance to the reflecting target is also established.

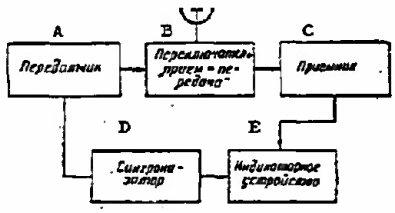


Figure 1.1. Block schematic of a pulse radar.
A - transmitter; B - T-R switch; C - receiver;
D - synchronizer; E - indicator.

Where linear sweep is used in the indicator the relationship between the beam deflection (l) and the measured range (D) can be established through the formula

$$l = v_s t_d = v_s\, 2D/c = SD, \qquad (1.3)$$

where

v_s is the constant speed of the sweep;

$S = 2v_s/c$ is the linear sweep scale.

Synchronization between the pulse transmitter and the indicator, that is, the instant the main pulse is radiated and the beginning of the indicator sweep must coincide precisely.

1.4 Pulse Compression Methods in Radar

Pulse energy must be increased if radar range is to be increased

$$W_p = P_p \tau_p, \qquad (1.4)$$

where

W_p is the pulse energy;

P_p is the pulse power;

τ_p is the pulse length.

As will be seen from the formula at (1.4), an increase in pulse length will increase the pulse energy. Increase in the pulse length degrades the radar range resolution, but if special modulation of the radiated pulses is used these pulses can be compressed by special processing in the receiver to a length providing the specified range resolution.

Intrapulse linear frequency modulation and intrapulse phase shift keying are such modulation methods.

The pulse-frequency method in radar

Figure 1.2 is a simplified block schematic of a radar using intrapulse linear frequency modulation.

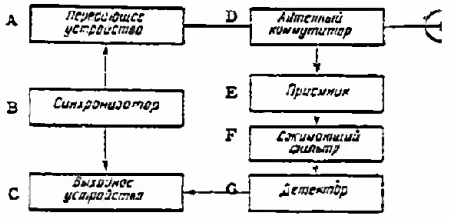


Figure 1.2. Block schematic of a radar using intrapulse linear frequency modulation.

A - transmitter; B - synchronizer; C - output; D - antenna switch; E - receiver; F - compressing filter; G - detector.

The transmitter forms long length, τ_p, radar pulses. The frequency inside the pulse changes in accordance with a linear law

$$f = f_0 - at, \tag{1.5}$$

where

a is the rate of change in the frequency.

Figure 1.3 a and b shows the shape of the radar pulse and the law governing the change in the frequency.

Signals reflected from the target are picked up by the radar receiver and are fed into a special compressing filter. The compressing filter is a delay line, the delay time in which is linearly dependent on the frequency (fig. 1.3c).

In this filter the high frequencies in a pulse, which arrive earlier, are delayed longer, while the low frequencies, which arrive later, are delayed for a shorter period of time. The result is that all the frequency components of the pulse are displaced in time to the end of the pulse; the pulse is compressed in time.

The degree of pulse compression is wholly dependent on the limits of the change in the frequencies in the pulse (the deviations in frequency)

$$f_d = a\tau_p.$$

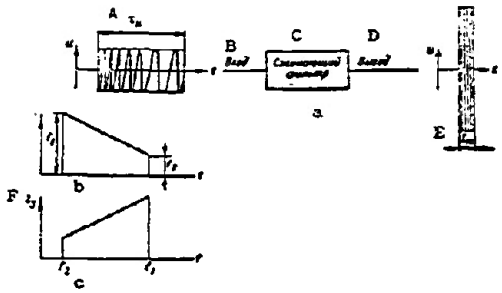


Figure 1.3. The principle involved in pulse compression.

a - shapes of radar pulses at the input and output of the compressing filter; b - law governing the change in the frequency of a radar pulse; c - dependence of filter delay time on the frequency.

A - τ_p; B - input; C - compressing filter; D - output; E - τ_{pc}; F - t_d.

Pulse length at the filter output equals

$$\tau_{pc} = 1/f_d. \tag{1.6}$$

The pulse compression ratio is

$$k = \tau_p/\tau_{pc} = f_d\tau_p. \tag{1.7}$$

Pulse power at the output of the compressing filter increases by a factor of k, or

$$P_{p\ out} = kP_{p\ in}. \tag{1.8}$$

For example, compressing a pulse with a length of 500 microseconds by a factor of 100 requires a frequency deviation in the pulse of f_d = 200 MHz, and a rate of change in frequency a = 400 MHz/second. But if pulse power at the input is $P_{p\ in}$ = 1 microvolt, the pulse power at the filter output will be $P_{p\ out}$ = 100 microvolt.

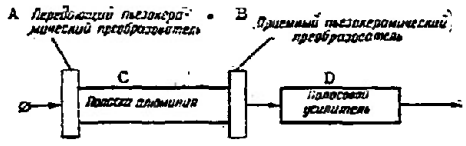


Figure 1.4. Functional schematic of a compressing filter.

A - transmitting piezo-ceramic converter; B - receiving piezo-ceramic converter; C - aluminum strip; D - band-pass amplifier.

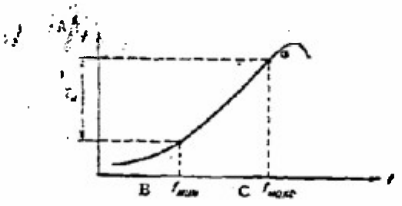


Figure 1.5. Dependence of t_d on frequency for a dispersion ultrasonic delay line.

A - t_d; B - f_{min}; C - f_{max}.

An optimum filter, consisting of an ultrasonic dispersion delay line with a correcting bandpass amplifier across the output (fig. 1.4), for example, can be used for the compressing filter.

The ultrasonic dispersion delay line consists of two piezo-ceramic converters of electrical oscillations into mechanical oscillations and an aluminum strip (fig. 1.4). The delay time of this ultrasonic delay line with respect to the frequency (fig. 1.5), and within the limits of the frequency from f_{min} to f_{max} changes linearly with the frequency.

Intrapulse phase shift keying method in radar

A simplified block schematic of a radar with intrapulse phase shift keying is shown in Figure 1.6. The transmitter forms the main pulses with constant frequency and long duration τ_p. These pulses can be split into equal segments, code intervals τ_c. Within the limits of each code interval is the initial phase of high-frequency oscillations.

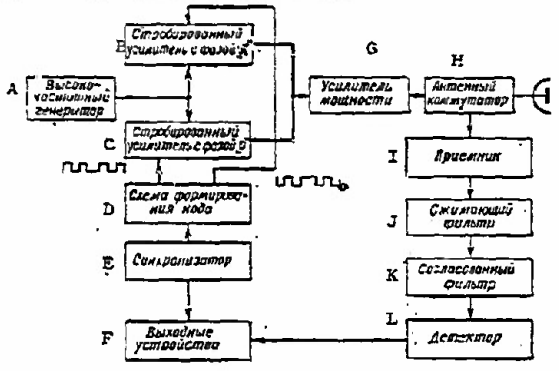


Figure 1.6. Block schematic of a radar with intrapulse phase shift keying.

A - HF oscillator; B - strobe amplifier with phase "π"; C - strobe amplifier with phase "0"; D - code forming circuit; E - synchronizer; F - output devices; G - power amplifier; H - antenna switch; I - receiver; J - compressing filter; K - matching filter; L - detector.

The length of the code interval, τ_c, is determined by the specified range resolution, and the code interval sequence is established by the code selected.

The Barker codes, in which the initial phases in adjacent code intervals equal π or 0, and the number of code intervals in the pulse can be 3, 4, 5, 7, 11, etc., can be used, for example.

Figure 1.7 shows a simplified schematic of the optimum processing of phase shift keyed pulses.

Figure 1.8a shows a signal consisting of seven code intervals (n = 7). The code intervals with initial phase 0 are conventionally designated by the sign "+1" while those with phase π are conventionally designated by the sign "-1."

The signal is processed in a device, the circuit of which is shown in Figure 1.7, in order to compress it. This device is connected to a delay line with six cells (each of which has a time constant τ_c) and seven take-offs. All the take-offs are connected to a summator through amplifiers with identical gain factors. Amplifiers with the "a" sign do not change the signal phase, but those with the "-a" sign change the phase by π. All voltages are summed in the summator, with phases taken into consideration. The oscillations are fed from the summator output into an optimum (matching) filter for pulse length τ_c.

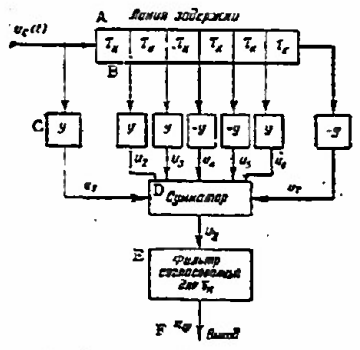


Figure 1.7. Schematic diagram of the optimum processing of phase shift keyed pulses.

A - delay line; B - τ_c; C - a; D - summator; E - matching filter for τ_c; F - output.

The shape of the voltage after the summator and optimum filter is shown in figures 1.8b and 1.8c.

Table 1 lists the voltages fed from the delay line take-offs to the summator, using the accepted designations, as well as the resultant voltages.

Table 1.1

A Напряжение сумматора	B Кодовые интервалы 1	2	3	4	5	6	7	8	9	10	11	12	13
u_1	−1	+1	−1	−1	+1	+1	+1	0	0	0	0	0	0
u_2	0	−1	+1	−1	−1	+1	+1	+1	0	0	0	0	0
u_3	0	0	−1	+1	−1	−1	+1	+1	+1	0	0	0	0
u_4	0	0	0	+1	−1	+1	+1	−1	−1	−1	0	0	0
u_5	0	0	0	0	+1	−1	+1	+1	−1	−1	−1	0	0
u_6	0	0	0	0	0	−1	+1	−1	−1	+1	+1	+1	0
u_7	0	0	0	0	0	0	+1	−1	+1	+1	−1	−1	−1
C Суммарное напряжение	−1	0	−1	0	−1	0	+7	0	−1	0	−1	0	−1

Key: A – voltage from summator; B – code intervals; C – summed voltage.

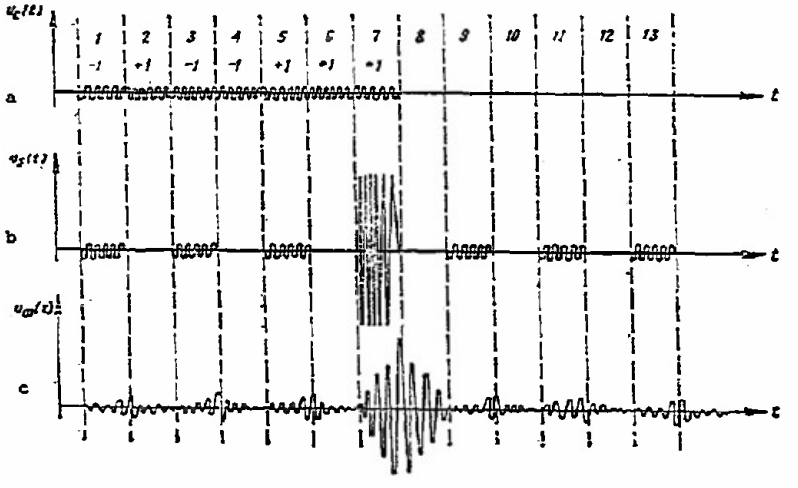


Figure 1.3. Conversion of a phase shift keyed pulse.

When phase shift keying is used the amplitude of the compressed pulse increases by a factor of n, while the power increases by a factor of n^2. The length of the compressed pulse is shorter than that of the main pulse by a factor of n.

1.5 The Frequency Method in Radar

The principle upon which the frequency radar operates is as follows. The simplified block schematic is shown in Figure 1.9.

The voltage of an outgoing frequency modulated signal, u_1 (fig. 1.10a) is fed into the detector from the high frequency generator. Simultaneously the receiving antenna supplies the reflected signal voltage u_2 (fig. 1.10b) to the detector input. If the distance from the reflecting target, and its effective area, remain unchanged with time there will be no additional modulation (amplitude and frequency) of the oscillations during reflection. Given these conditions, the reflected signal differs from the outgoing signal only in amplitude. The magnitude of the time delay in the reflected signal can be established through the formula at (1.1).

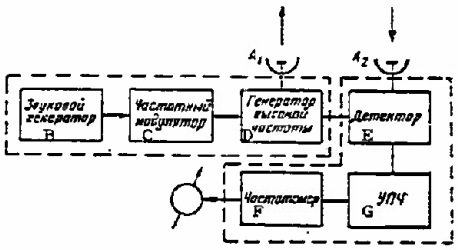


Figure 1.9. Block schematic of a frequency radio range finder.
A_1 - transmitting antenna; A_2 - receiving antenna;
B - audio oscillator; C - frequency modulator; D - high-frequency generator; E - detector; F - frequency meter; G - intermediate-frequency amplifier.

Beats occur when the outgoing and reflected signals are added (fig. 1.10c). The resultant signal turns out to be both frequency and amplitude modulated. The number of maximums in the envelope of the resultant oscillation in unit time will depend on the reflected signal delay time, that is, on the distance from the reflecting target.

Now, if by detecting the resultant signal the envelops is divided (fig. 1.10d) and after the necessary amplification is fed into a frequency meter, its readings will correspond to the distance being measured. The frequency meter can be calibrated in distance units.

The beat frequency in the case of linear frequency modulation equals

$$F_b = a \cdot 2D/c, \tag{1.9}$$

where

a is the rate of change in the frequency.

In the case of modulation in accordance with the law for a symmetrical triangular curve (fig. 1.11)

$$F_b = |f_1 - f_2| = 4f_d F_m / c \cdot D \tag{1.10}$$

where

f_d is the frequency deviation;

F_m is the modulation frequency.

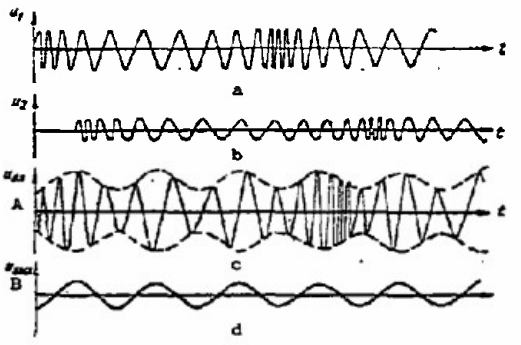


Figure 1.10. Voltage curves explaining the operation of the frequency radio range finder.

A - input; B - output.

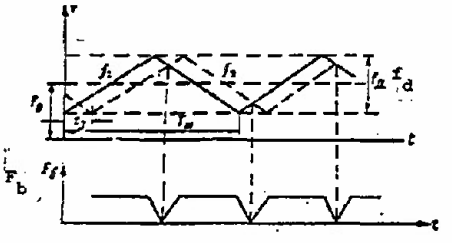


Figure 1.11. Dependence of the modulus of the difference in the frequencies of the outgoing and reflected signals on time in the case of modulation in accordance with the law for a symmetrical triangular curve.

The formula at (1.10) is correct for the condition $|f_1-f_2| > F_m$.

For the case of modulation in accordance with a harmonic law

$$F_b = |f_1-f_2| = |f_d \sin \Omega_M t_d/2 \sin \Omega_M(t-t_d/2)|, \qquad (1.11)$$

where

t_d is the delay time for the reflected signal.

As will be seen from the expression at (1.11), if there is harmonic modulation and unchanged distance, the beat frequency will change periodically. But the readings on the frequency meter (given adequate inertia in the meter) will correspond to the average beat frequency, as determined through the formula at (1.10).

The formula at (1.10) is approximate for cases of modulation with respect to a symmetrical triangular curve and with respect to the harmonic law, and yields discrete values that are multiples of the modulation frequency. The result is that the measured distance can differ from the true distance by a magnitude $\pm D_0$, and in the case of measurements of short distances even by as much as $2D_0$, where

$$D_0 = c/4f_d, \tag{1.12}$$

which establishes the accuracy of the frequency method of radio range finding.

The foregoing is correct in its entirety for the condition that the amplitude of the outgoing signal does not change with time. In point of fact, because of the resonant properties of the transmitting and receiving antennas and of the receiver input circuit, the frequency modulation is accompanied by amplitude modulation as well. This causes the accuracy of this particular method to deteriorate.

If the distance between the range finder and the target changes, the law for the change in the frequency of the reflected signal differs from the law for the change in the frequency of the outgoing signal because of the Doppler effect.

Now the beat frequency equals

$$F_b = |F_d \pm F_D|, \tag{1.13}$$

where

F_d is the beat frequency resulting from the delay in the reflected signal and established through the formula at (1.9);

F_D is the Doppler frequency.

The Doppler frequency is

$$F_D = f_c \cdot 2v_r/c = 2v_r/\lambda, \tag{1.14}$$

where

f_c is the radio range finder's carrier frequency;

v_r is the target's radial velocity, that is, the rate of displacement of the target in the direction toward the radio range finder;

λ is the wavelength.

Range to the target and the speed at which it is moving can be measured separately for the case of modulation in accordance with the law for the symmetrical triangular curve (fig. 1.12).

If $F_d > F_D$,

$$F_d = F_д = \frac{1}{2}(|F_1| + |F_2|), \tag{1.15}$$

$$F_D = F_з = \frac{1}{2}(|F_1| - |F_2|). \tag{1.16}$$

Thus, by measuring the difference between F_1 and F_2 (fig. 1.12), we can find the distance to the target, as well as its speed

$$D = cF_d/2a, \tag{1.17}$$

$$v_r = cF_d/2f_c. \tag{1.18}$$

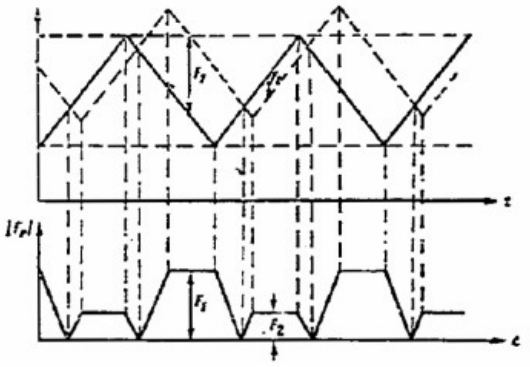


Figure 1.12. Influence of the Doppler effect in the case of modulation in accordance with the law for the symmetrical triangular curve.

1.6 The Phase Method in Radar

The simplified block schematic of the phase radio range finder is shown in Figure 1.13.

Its operating principle is as follows. Two voltages, one from the scale frequency generator,

$$u_1 = U_{m1}\sin(\omega_s t + \varphi_0), \tag{1.19}$$

where

ω_s is the scale frequency;

φ_0 is the initial phase,

and one from the output of the receiver,

$$u_2 = U_{m2}\sin[\omega_s(t-t_d) + \varphi_0 - \varphi_r - \varphi_{ref}], \tag{1.20}$$

where

φ_r is the delay in the phase of the scale oscillation in the radio range finder circuits;

φ_{ref} is the lag in the phase of the scale oscillation occurring during the reflection from the target;

t_d is the reflected signal delay time,

are fed into the phase meter, P.

The difference in the phases of voltages u_1 and u_2 equals

$$\varphi_{diff} = \omega_s t_d + \varphi_r + \varphi_{ref}. \tag{1.21}$$

The difference in phase in the range finder circuits, as well as the lag, during reflection, are constant and can be computed, or determined, experimentally. Then, measuring the difference in the phases of voltages u_1

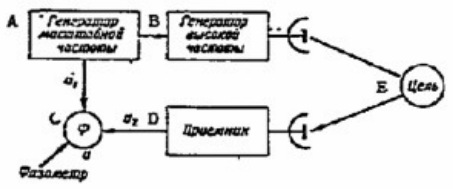


Figure 1.13. Simplified block schematic of a phase radio range finder.
A - scale frequency generator; B - high-frequency generator; C - phase meter; D - receiver; E - target.

and u_2, the distance to the target can be established as

$$D = c/2\omega_s\ (\varphi_{diff} - \varphi_r - \varphi_{ref}). \qquad (1.22)$$

If radio range finding is to be ensured by the phase method there must be reliable selection of the signal reflected from the target, the distance to which must be established.

The Doppler effect can usually be used to make the selection. Two transmitters, operating on frequencies ω_1 and ω_2, are usually used. Found at the receiver input (fig. 1.14) are the outgoing signals with frequencies ω_1 and ω_2, as well as the signals reflected from the target with frequencies $\omega_1+\Omega_1$ and $\omega_2+\Omega_2$, where $\Omega_1 = \omega_1 \cdot 2v_r/c$ and $\Omega_2 = \omega_2 \cdot 2v_r/c$ are the Doppler frequencies.

From the output of the receiver the voltage is fed into two bandpass filters, one of which passes the band of frequencies from $\Omega_{1\ min}$ to $\Omega_{1\ max}$, the other those from $\Omega_{2\ min}$ to $\Omega_{2\ max}$. Frequencies ω_1 and ω_2 should be selected such that the bands indicated do not overlap.

If ω_1 and ω_2 differ little from each other, and $\Omega_1 \approx \Omega_2$, then

$$D \approx c\varphi_{diff\ D}/2(\omega_1-\omega_2) \qquad (1.23)$$

where

$\varphi_{diff\ D} = (\omega_1-\omega_2)\cdot 2D/c$ is the difference in the phases of voltages u_1 and u_2 received by the receiver.

As wel see, the distance in the case under consideration can be established by measuring the difference in the phases of the oscillations of two Doppler frequencies.

The phase range finder considered has no range resolution, but does have target speed resolution. This latter is what makes target selection with respect to the speed at which it is moving possible, and, as a result, provides range measurement even when several targets are in the "field of

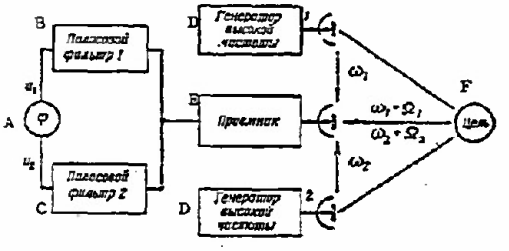


Figure 1.14. Simplified block schematic of a radio range finder using the Doppler effect.

A - phase meter; B - bandpass filter 1; C - bandpass filter 2; D - high-frequency generator; E - receiver; F - target.

vision" of the radio range finder. Narrow band filters with a variable resonant frequency should be envisaged in this case, rather than bandpass filters (fig. 1.14).

1.7 The Continuous Radiation with Phase Code Shift Keying Method

The radar transmitter generates continuous oscillations at constant frequency. The initial phase of these oscillations changes in accordance with a predetermined code.

A repetition period for the radar, T_p, corresponding to the specified range for the radar, can be selected. This period can be divided up into time intervals τ_c (fig. 1.15).

The interval τ_c is selected on the basis of the range resolution needed for the radar. Total intervals are

$$N = T_p/\tau_c. \tag{1.24}$$

The phase shift keying is done in such a way that in approximately half of the intervals the phase is equal to φ_0, and in the remainder of the intervals the phase is equal to $\varphi_0 + \pi$. The phase alternation is selected in accordance with a predetermined code (fig. 1.15).

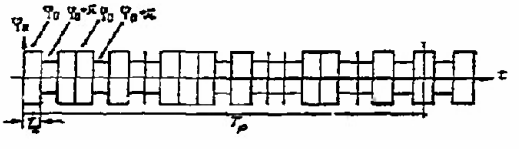


Figure 1.15. Example of the phase code shift keying of a continuous signal.

The phase code shift keyed signal is radiated into space, reflected from the target, and received by the receiver, but delayed by the time required to propagate the electromagnetic energy to the target and back. The received signal is fed into the mixer in the receiver, as is an identical phase code shift keyed reference signal, with the difference that this latter signal is time delayed by the delay line. By changing the delay in the reference signal the codes are made to coincide in time, the phase shift keying is equalized, and a narrow band, unmodulated signal, is obtained. The magnitude of the delay in the reference signal determines the range to the target. There are many advantages to this method as compared with the pulse method (while retaining all the positive aspects of the latter):

the level of the peak power is equal to the average power, so there is no need to impose rigid requirements with respect to the electrical strength of waveguides and high voltage insulators;

higher noise stability, because this method makes it possible to use narrow band filtering and is very sensitive to target speed;

three target coordinates (range, azimuth, or elevation, and radial velocity) can be measured.

1.8 The Continuous Radiation with Noise Modulation Method

Noise modulation of a high-frequency signal can be used to determine the range in the case of continuous radiation. Modulation can be amplitude or phase. Let us consider the operating principle of the radar with amplitude noise modulation by using the block schematic shown in Figure 1.16.

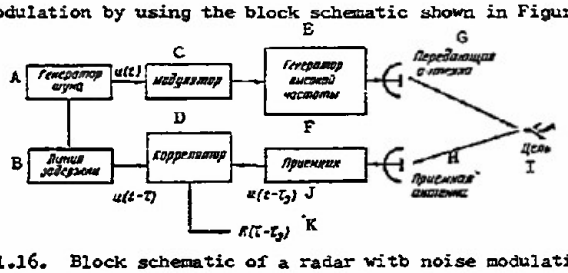


Figure 1.16. Block schematic of a radar with noise modulation.

A - noise generator; B - delay line; C - modulator; D - correlator; E - high-frequency generator; F - receiver; G - transmitting antenna; H - receiving antenna; I - target; J - $u(t-\tau_d)$; K - $R(\tau-\tau_d)$.

The generator produces the noise signal u(t) (fig. 1.17a). The oscillations produced by the high-frequency generator are amplitude modulated by this signal. The reflected signal is fed into the receiver. The original noise signal is separated at the detector output, but with time delay τ_d corresponding to the range to the target (fig. 1.17b). The signal is fed

from the detector output to the correlator, at the second input to which is fed the original noise signal u(t) delayed in the delay line by time τ (fig. 1.17c). The delay time in the line is variable.

The correlator performs an operation on both signals of the form

$$R(\tau-\tau_d) = 1/T_{ob} \int_0^{T_{ob}} u(t-\tau_d)u(t-\tau)dt, \qquad (1.25)$$

where

T_{ob} is time of observation.

In the theory of random processes the function $R(\tau-\tau_d)$ is called the autocorrelation function, and the circuit has therefore been named the correlator. Produced at the correlator is a voltage proportional to the value of the autocorrelation function of the noise $R(\Delta\tau)$ when $\Delta\tau = \tau-\tau_d$. The autocorrelation function of the noise is a maximum when $\Delta\tau = 0$.

Consequently, when the line delay time equals τ_d the voltage at the autocorrelator output will be a maximum (fig. 1.17d). If the entire distance is to be scanned the delay in the line should change from zero to a maximum, for this will correspond to the distance to the most distant target. The principal properties of a radar with noise modulation are the following:

the noise signal, as distinguished from a periodically regulated signal, makes it possible to obtain an unambiguous determination of the distance to the target;

the use of noise modulation, particularly in the case of phase modulation, increases the average power of the signal as compared with the pulse regime, just as in radars with other types of continuous radiation;

the radar does not have high pulse power, simplifying the design of the transmitter and the antenna feeder path;

the noise signal in the radar, particularly in the case of amplitude modulation, is similar to the internal noises found in receivers, and this can make possible more secret operation of the radar because, first of all, it is difficult to establish the very fact of radar operation, and, second, it is difficult to establish the radar's parameters.

1.9 Detection of Radar Signals

Radar detection of a target means the process of deciding on the presence, or absence, of a target in a particular area of space by receiving and processing radar signals.

Signals are always received against a background of noise. The principal types of natural noises found on the wavebands used in radar work are thermal and cosmic noises, and the noises that develop inside the receiver.

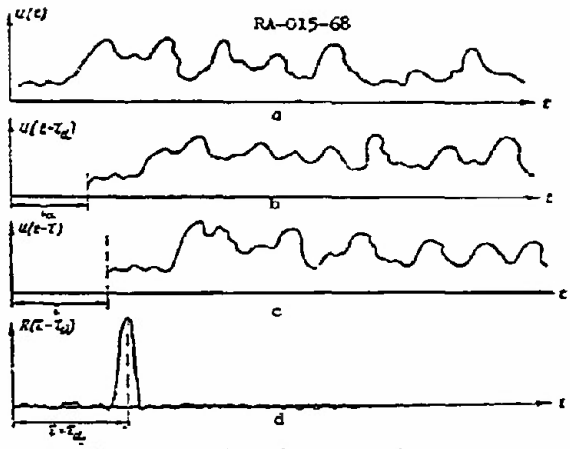


Figure 1.17. The explanation of the principle of operation of the correlator in a radar with noise modulation.

Noise causes signal distortion and errors in evaluating the situation. The decision as to the presence, or absence, of a target is made in the light of two conflicting conditions:

there is, in fact, a target;

there is, in fact, no target.

These conditions are unknowns in the development of the decision.

There are two types of decisions relevant to each of these conditions:

"there is a target";

"there is no target."

Four situations are possible during detection. Given the condition that there is, in fact, a target, the decision that there "is a target" is correct detection, and the decision that there "is no target" is a target miss.

Given the condition that there is, in fact, no target, the decision that there "is no target" is correct nondetection, while the decision that there "is a target" is a false alarm.

Target miss and false alarm are errors in target detection.

Since, in the general case, radar signals and noise are random functions of time, the making of any particular decision too is random in nature.

The possibilities that these situations will occur is taken as characterizing the probabilities of correct and mistaken decisions:

the probability of correct detection, P_{cd};

the probability of correct nondetection, P_{cnd};

the target miss probability, P_{mt};

the probability of false alarm, P_{fa}.

Correct detection and target miss (when a target is, in fact, present) form a complete group of incompatible events, so

$$P_{cd} + P_{mt} = 1. \tag{1.26}$$

In precisely the same way, false alarm and correct nondetection form a complete group of incompatible events when there is no target, and

$$P_{fa} + P_{cnd} = 1. \tag{1.27}$$

The formulas at (1.26) and (1.27) demonstrate that among the probabilities listed there are only two magnitudes that are independent. Ordinarily, the probability of correct detection and the probability of a false alarm are used as the two independent probabilities for purposes of characterizing the detection devices.

The device used to process radar information must meet contradictory requirements. If target miss is to be avoided, a decision as to its presence must be made, even when the target signal is badly distorted by interference and it is impossible to confirm with confidence the certainty that there is no target. The false alarm probability increases, of course. On the other hand, if the attempt is made to reduce the probability of a false alarm, the decision must be made as to target presence only when there is a clear excess of signal over noise. The probability of target miss is now increased.

This therefore leads to an educated compromise between the contradictory requirements, selecting the optimum method for processing the information. Certain criteria, in accordance with which we can compare various devices with each other, must be used in order to judge the quality with which the detection devices function. The optimum detection device is one which makes it possible to obtain the best (as compared with others) value for the selected criterion, all other conditions being equal.

The criterion most often used in radar is the Neyman-Pearson criterion, in accordance with which the optimum detection device should provide the maximum probability of correct detection, P_{cd}, for a specified value for the probability of false alarm, P_{fa}.

In the optimum receiver the detection of radar signals results from the establishment of a posteriori (that is, after the fact) probabilities of there being various types of information (for example, information that a target is, or is not present), and indications as to the quality of the decision made with respect to the information, the probability of which is greater than that of any others, or the establishment of the likelihood ratio characterizing the likelihood of a particular hypothesis with respect to the information transmitted.

The signal at the receiver input, u(t), is the sum of the random processes, given the condition that there is, in fact, a target

$$u(t) = u_s(t) + u_n(t), \tag{1.28}$$

where

$u_s(t)$ is the target signal;

$u_n(t)$ is the noise.

If there is no target

$$u(t) = u_n(t). \tag{1.29}$$

Like any random process, the signal $u(t)$ can be described in full by the probability-density function for the signal en elope $w(u)$ and phase $w(\varphi)$.

The shape of the probability distribution for the case of $u(t) = u_s(t) + u_n(t)$ and $u(t) = u_n(t)$ is shown in Figure 1.18.

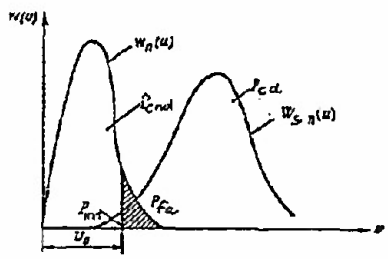


Figure 1.18. Probability-density functions for noise and signal-plus-noise envelopes.

Knowing the probability-density function $w(u)$, and assigning threshold values for u_0, we can compute the probabilities of correct and mistaken decisions, established through the following expressions

$$P_{cd} = \int_{U_0}^{\infty} w_{s+n}(u)du, \tag{1.30}$$

$$P_{cnd} = \int_{0}^{U_0} w_n(u)du, \tag{1.31}$$

$$P_{mt} = \int_{0}^{U_0} w_{s+n}(u)du, \tag{1.32}$$

$$P_{fa} = \int_{U_0}^{\infty} w_n(u)du, \tag{1.33}$$

where

$w_{s+n}(u)$ is the probability-density function for the signal-plus-noise envelope;

$w_n(u)$ is the probability-density function for the noise envelope;

U_0 is the threshold value for the voltage the excess in which is used in making the decision that a target is present.

The likelihood ratio can be established through the expression

$$\lambda = w_{s+n}(u)/w_n(u). \tag{1.34}$$

In accordance with the Neyman-Pearson criterion the decision that a target is present is made if the likelihood ratio exceeds the specified threshold value λ_{th}, that is, $\lambda > \lambda_{th}$.

The magnitude of λ_{th} is selected such that the false-alarm probability, P_{fa}, does not exceed a permissible value, $P_{fa\ per}$. In reality, the target signal used in existing radars operating in the space scan regime is a pulse train. The number of pulses in the train can be established through the formula

$$N_p = F_p \cdot \varphi_{0.5}/6n_A, \tag{1.35}$$

where

$\varphi_{0.5}$ is the width of the power pattern at the half-power level;

n_A is the antenna rotation rate (rpm).

There are serious mathematical difficulties involved in computing the probabilities of correct detection of the target and of false alarm. However, in the case of weak signals, and this case is one of great practical interest, these probabilities can be established through the following formulas

$$P_{cd} = \mathcal{P}_{\omega} = \frac{1}{2}\left[1 + \Phi\left(v\sqrt{\frac{N_p}{2}} - \frac{\ln \lambda_{th}}{v\sqrt{2N_p}}\right)\right], \tag{1.36}$$

$$P_{fa} = \mathcal{P}_{\alpha} = \frac{1}{2}\left[1 - \Phi\left(\frac{\ln \lambda_{th}}{v\sqrt{2N_p}}\right)\right], \tag{1.37}$$

where

$\Phi(x)$ is the probability integral;

$v = P_s/P_n$ is the signal/noise power ratio at the output of the linear section of the receiver;

λ_{th} is the threshold value for the likelihood ratio;

N_p is the number of target pulses processed.

The correct detection probability for a specified false-alarm probability is greater the greater the signal/noise ratio at the output of the linear section of the receiver, and the larger the number of pulses processed.

Figure 1.19 shows the dependencies of the probability of correct target detection, P_{cd}, for a specified false-alarm probability, P_{fa}, on the magnitude of the signal/noise ratio, v, and the number of pulses processed, N_p. The curves for these dependencies are called the detection curves (the receiver

performance curves) and provide a visual representation of the observability of radar signals. These curves can be used to find the threshold value, v_{th}, at which specified probabilities P_{cd} and P_{fa} will be ensured for optimum signal processing

$$v_{th} = x(P_{cd}, P_{fa})/\sqrt{N_p}\ , \qquad (1.38$$

where

$x(P_{cd}, P_{fa})$ is the abscissa of a point on the curve corresponding to the specified probabilities P_{cd} and P_{fa}.

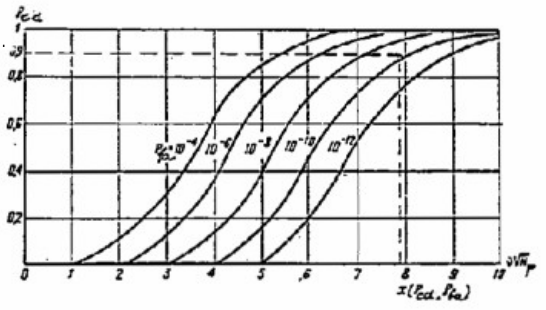


Figure 1.19. Detection curve in the case of a weak signal of constant amplitude

P_{cd} is usually given from 0.5 to 0.9, and P_{fa} from 10^{-10} to 10^{-6}.

The receiver detecting the signal should be able to evaluate the likelihood ratio $\lambda = w_{s+n}(u)/w_n(u)$. The greater the likelihood ratio, the greater the probability that a signal is present. Computing the likelihood ratio for a signal with completely known parameters, one can conclude that the receiver should form the integral

$$q = \frac{2}{w_n}\int_0^{T_0} u(t)\,u'(t)\,dt \qquad (1.39)$$

(where

w_n is the energy of the noise, computed as equal to its spectral density;

$u'(t)$ is the completely known signal;

$u(t)$ is the signal at the receiver input;

T_0 is time of observation),

and compare it with some level (threshold). At the same time, this operation will, with equal effect, compute the likelihood ratio and compare it with the threshold value. The expression at (1.39) is called the correlation integral. A receiver carrying out this operation with the incoming signal is an optimum receiver and has the structure shown in Figure 1.20.

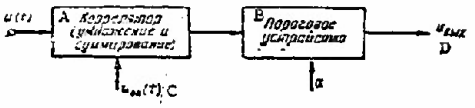


Figure 1.20. Block schematic of a correlation receiver.
A - correlator (multiplication and summing); B - threshold device; C - $u_{ref}(t)$; D - u_{out}.

The correlator is a device used to multiply the input signal u(t) by the reference signal $u_{ref}(t)$, the latter a copy of the signal with the completely known parameters, u'(t), coinciding in time with the incoming signal, and integrated (summed) in time.

The computed correlation integral is compared with the threshold, α, in the threshold device, and if the integral is in excess of the threshold, the decision is that a target is present.

This type of receiver is called a correlation receiver.

A filter, matched with the signal (or an optimum filter), as well as a correlator, can be used to process signals in the optimum receiver. The filter is said to be optimum when its frequency curve, $K_{opt}(\omega)$, is a complexly conjugate spectrum of the signal $S(\omega)$

$$K_{opt}(\omega) = ce^{-j\omega t_0}\dot{S}^*(\omega). \tag{1.40}$$

If the signal spectrum is

$$\dot{S}(\omega) = S(\omega)e^{j\varphi_c(\omega)}, \tag{1.41}$$

the filter curve will be

$$\dot{K}_{opt}(\omega) = K_{opt}(\omega)e^{j\varphi_{opt}(\omega)} = cS(\omega)e^{-j\varphi_c(\omega)}e^{j\omega t_0}, \tag{1.42}$$

where

$K_{opt}(\omega)$ is the filter's amplitude-frequency curve;
$\varphi_{opt}(\omega)$ is the filter's phase-frequency curve;
$S(\omega)$ is the signal's amplitude-frequency spectrum;
$\varphi_s(\omega)$ is the signal's phase-frequency spectrum;
τ_0 is the time delay in the filter.

As will be seen from the expression at (1.42), the amplitude-frequency curve for the optimum filter, $K_{opt}(\omega)$, is proportional to the signal's amplitude-frequency spectrum, $S(\omega)$, that is, the shape of the filter's frequency curve coincides with the shape of the signal's frequency spectrum.

The best example of an optimum filter will pass the spectral components most pronounced in the signal spectrum. The weak components in the signal spectrum are suppressed by the filter. The components of the noise spectrum

uniformly distributed over a broad band of frequencies are suppressed along with the weak components.

The phase-frequency curve, $\varphi_{opt}(\omega)$, for the optimum filter is proportional to the phase-frequency curve for the signal, $\varphi_s(\omega)$, with reversed sign, or

$$\varphi_{opt}(\omega) = -\varphi_s(\omega) + \omega t_0. \tag{1.43}$$

This signifies that the phase displacements of the spectral components of the signal are compensated for by the filter, and that there is a period in time t_0 (filter delay) when all the spectral components are in phase and add arithmetically. The signal voltage peak occurs at the filter output at this moment in time, but this phenomenon does not take place for noise (fig. 1.21).

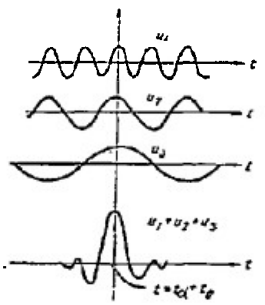


Figure 1.21. Superposition of the maxima for the harmonic components of the useful signal at the output of the filter in the case of an optimum phase-frequency curve.

The optimum filter provides the best signal/noise ratio at the filter output when compared with any other types of filters

$$\nu_0 = 2w_s/G, \tag{1.44}$$

where

w_s is signal energy;

G is the spectral density of the noise.

The signal/noise ratio provided by the optimum filter is established by the signal energy at the filter input and the spectral density of the noise, and does not depend on the shape of the signal. Signal shape only establishes the filter structure.

The only filters that can be used to process signals are those matched to the signals in passband width.

Figure 1.22 shows the dependence of ν/ν_0 on the product $\Delta F\tau_p$ (ν is the signal/noise ratio at the non-optimum filter output; ν_0 is the signal/noise ratio at the optimum filter output; τ_p is the pulse length; ΔF is the filter passband).

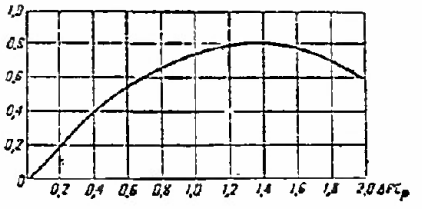


Figure 1.22. Dependence of the signal/noise ratio on the product $\Delta F\tau_p$.

As will be seen from Figure 1.22, there is an optimum value for the product, $(\Delta F\tau_p)_{opt} = 1.37$, for which the ratio v/v_0 is maximum, or, putting it another way, there is an optimum passband

$$\Delta F_{opt} = 1.37/\tau_p, \tag{1.45}$$

at which $v/v_0 = 0.825$.

The loss in the signal/noise ratio in this filter, as compared with that taking place in the optimum filter, is only 17%, or degraded by a factor of 1.2. The optimum filter provides a single pulse, as well as pulse trains, for processing. The engineering of the filter for a single pulse is simpler than is the case for the filter that handles pulse trains.

The solution to the detection problem during optimum processing of pulse trains reduces to the following operations:

optimum filtration of each pulse in the train;

amplitude detection;

synchronous integration of video signals;

testing of the summed signal at the threshold.

The first two operations are usually performed in the receiver, the others by the radar's output devices. The synchronous integration operation is carried out by a device for summing the signals corresponding to the same range in the different repetition periods. Needed for carrying out this operation is a device that can remember the signals for the repetition period (a delay line, for example). The block schematic of a synchronous integrator for four pulses in a square pulse train, as well as the curves explaining the synchronous integration process, are shown in Figure 1.23.

The use of optimum signal processing reduces, in sum, to reducing threshold power. By threshold power (P_{th}) of radar signals is understood to mean the minimum signal power at the receiver input which provides specified probabilities of correct target detection and false alarm.

The magnitude of threshold power depends on the specified values for the probabilities of correct target detection and false alarm, the radar signal parameters, time of observation, and the type of processing given the radar signals.

For the case of optimum filtering of single high-frequency square pulses

$$P_{th} = \nu_{th} \cdot NkT_0/\tau_p \tag{1.46}$$

where

N is the receiver noise figures;

k is the Boltzmann's constant, $k = 1.37 \cdot 10^{-23}$ joules/degree;

T_0 is the Kelvin temperature, $T_0 = 300°K$;

ν_{th} is the signal/noise ratio, established using the detection curve (fig. 1.19) and the specified probabilities for P_{cd} and P_{fa}.

Threshold power must be stepped up when filtering is non-optimum so the same probabilities of correct detection and false alarm exist as pertain when using optimum filtering.

The signal/noise ratio for non-optimum filtering is lower than it is for optimum filtering by a factor of ν_0/ν, in which $\nu_0 = 2w_s/G$ is the signal/noise ratio at the output of the optimum filter, and ν is the signal/noise ratio at the output of the linear section of the receiver in the case of non-optimum filtering.

Consequently, the threshold power in the case of non-optimum filtering is equal to

$$P'_{th} = \nu_0/\nu \cdot \nu_{th} \cdot NkT_0/\tau_p = \nu_c \cdot NkT_0/\tau_p = \nu_c P_n. \tag{1.47}$$

The factor ν_c is called the classification factor and signifies the factor by which the power of minimum received signals must exceed the power of the receiver's internal noises in order to have the signals from the target detected with specified probabilities of correct detection and false alarm.

For example, a receiver with optimum passband will have a ratio $\nu_0/\nu = 1.2$ and $\nu_c = 1.2\nu_{th}$.

The threshold power is the real sensitivity of the receiver, $P_{rec\ min}$. Therefore, taking $1/\tau_p = \Delta F$, we can write the formula at (1.47) in the form

$$P_{rec\ min} = \nu_c P'_{rec\ min},$$

where

$P'_{rec\ min} = NkT_0\Delta F = P_n$ is the limit of sensitivity for the receiver.

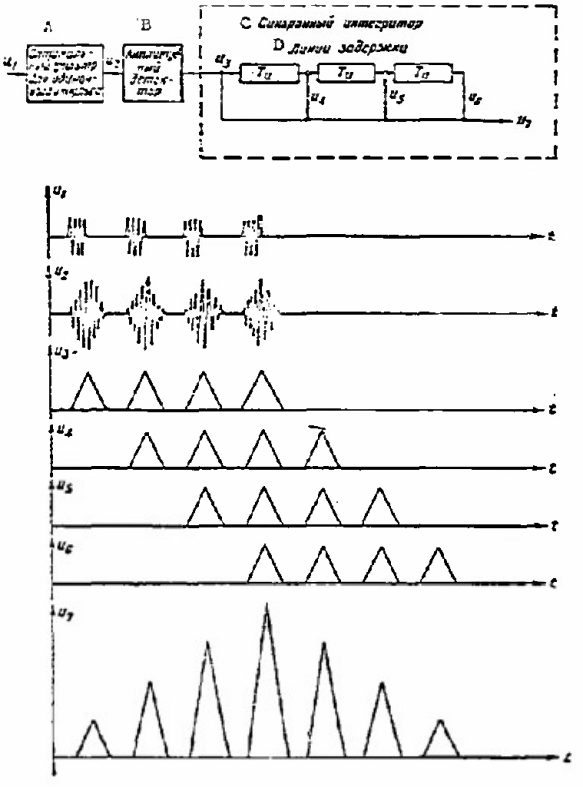


Figure 1.23. Optimum processing of signals.

A - optimum filter for single pulse; B - amplitude detector; C - synchronous integrator; D - delay lines.

1.10 Radar Range

Target detection is an operation that must separate signals reflected from the target from the background of receiver internal noise and fluctuating noise. The special feature of this operation is the statistical nature of its results, occasioned by the random nature of the change in the noise voltage and the fluctuations in the magnitude of the effective cross section. Separation of the useful signal can be accomplished with a definite degree of authenticity and which, above all else, can be characterized by the target detection probability.

In establishing the radar range we must take into consideration the probable nature of the detection of target signals against a background of noise.

With this as a point of departure, we will call the maximum radar range that limiting distance to a target at which detection of target signals against a background of noise will be ensured with a specified probability.

In free space, and given the condition that maximum radiation from the radar antenna is aimed at the target and that atmospheric absorption of radio waves is not taken into consideration, maximum radar range (D_{max}) can be established through the equation

$$D_{max} = \sqrt[4]{\frac{P_{trans} G_{max} S_a \sigma_t}{16\pi^2 P'_{rec\ min} v_c}}, \qquad (1.48)$$

where

P_{trans} is the power radiated by the transmitter;
$P'_{rec\ min}$ is the limiting sensitivity of the receiver;
G_{max} is the maximum antenna directivity;
S_A is the capture area of the antenna;
σ_t is the effective cross section;
v_c is the classification factor.

Example. Determine the range at which a radar will detect an aircraft with an effective cross section $\sigma_t = 5\ m^2$ with a probability of correct detection of $P_{cd} = 0.9$ and a false-alarm probability of $P_{fa} = 10^{-8}$. Radar characteristics are $P_{trans} = 845$ kw; $P_n = 10^{-12}$ watt; $S_A = 30/\pi\ m^2$; $G = 12{,}000$; $\lambda = 10$ cm; $F_p = 450$ hertz; $\tau_p = 2$ microseconds; $n_A = 6$ rpm; $\varphi_{0.5} = 2°$; and $\Delta F = 1.37/\tau_p$.

Solution. 1. Number of pulses in the train

$$N_p = F_p \varphi_{0.5}/6n_A = 450 \cdot 2/6 \cdot 6 = 25 \text{ pulses}$$

2. We find that $\vartheta_{th}\sqrt{N_p} \approx 8$ from the curves in Figure 1.19 for given $P_{cd} = 0.9$ and $P_{fa} = 10^{-8}$, so

$$v_{th} = 8/\sqrt{N_p} = 8/5 = 1.6.$$

3. Since filtering is not optimum, but separation of the signal is by a receiver with a passband of ΔF_{opt}, from the curve in Figure 1.22 $v_0/v \approx 1.2$ and the classification factor is

$$v_c = v_0/v \cdot v_{th} = 1.2 \cdot 1.6 \approx 1.9.$$

4. The maximum range at which the target can be detected is

$$D_{max} = \sqrt[4]{\frac{P_{trans} G_{max} S_A \sigma_t}{16\pi^2 P'_{rec\ min} \nu_t}} =$$

$$= \sqrt[4]{\frac{845 \cdot 10^3 \cdot 1.2 \cdot 10^4 \cdot 30 \cdot 5}{16\pi^2 \pi 10^{-12} \cdot 1.9}} = 2 \cdot 10^5 \text{meters} = 200 \text{ km}.$$

1.11 Radio Direction Finding Methods

The task of radio direction finding reduces to establishing the direction from which radio waves radiated, or reradiated, by the target, are arriving. In other words, the task is to establish the azimuth, or elevation, of the target (the target bearing).

Target bearing (α) is understood to mean the angle between a base line and the direction to the target. The base line most often used is:

in the horizontal plane - the north point (sometimes the south);

in the vertical plane - the plane of the horizon.

Radio direction finding can use the following methods: amplitude methods; phase methods; and amplitude-phase methods.

Amplitude methods of radio direction finding

The amplitude methods of radio direction finding are based on the use of the directional properties of antennas (fig. 1.24).

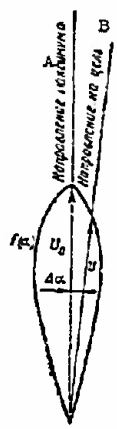


Figure 1.24. Direction finding by the maximum method.
A - direction of the maximum; B - direction to the target.

When a movable pencil-beam antenna is used the amplitude of the signal reflected from the target at the receiver output (fig. 1.25) depends on the direction of the reception pattern with respect to the target (the target bearing). If the antenna reception pattern has its maximum directed at the

target the amplitude of the reflected signal will be maximum at the receiver output, and in all other cases the signal amplitude will be less than the maximum, and even equal to zero when the reception pattern is turned away from the target.

The dependence of the voltage across the receiver output on the target bearing is called the direction finding characteristic $f(\alpha)$.

When the directional properties of the transmitting antenna, $f_{A1}(\alpha)$, and of the receiving antenna, $f_{A2}(\alpha)$, are used for direction finding, the direction finding characteristic is

$$f(\alpha) = f_{A1}(\alpha) f_{A2}(\alpha), \tag{1.49}$$

where

$f_A(\alpha)$ is the antenna directivity pattern.

If a radar has one antenna used for transmitting, as well as receiving, the direction finding characteristic is

$$f(\alpha) = f_A^2(\alpha). \tag{1.50}$$

The amplitude methods used for direction finding are the maximum method, the minimum method, the equisignal sector method, and the comparison method.

The maximum-signal method. In this method one establishes the direction to the target by the direction of the maximum for the direction finding response curve (fig. 1.24).

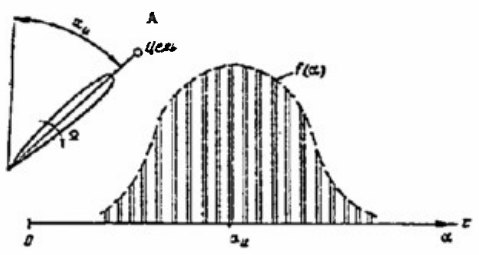


Figure 1.25. Signal at the receiver output when the maximum-signal method is used for direction finding (the direction finding response curve).

A - target.

When the amplitude of the signal from the target at the receiver output is a maximum, it is taken that the maximum for the direction finding response curve coincides with the direction to the target, and is considered to be the target bearing.

Let the radar antenna be rotating in the horizontal plane. A signal from a target, appearing on the screen of an amplitude indicator, will change its amplitude continuously (fig. 1.25).

When the signal reaches its maximum the operator (or the automatic device) reads the value of the target azimuth off the azimuth instrument indicating the angular position of the antenna.

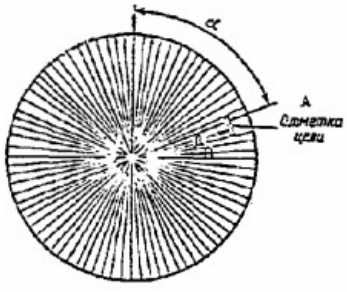


Figure 1.26. View of the screen of a circular scan indicator when finding the direction to a target by using the maximum-signal method.

A - target pip; B - range.

If a circular scanning scope is used the range scanning line of the scope will rotate in synchronism with the antenna. The target azimuth can be established by the angular position of the pip on the screen of the scope (fig. 1.26).

A scope with a range-azimuth (fig. 1.27) is usually used with sector scan radars. The scanning line on the scope in this case moves along the axis of the azimuth in synchronism with the beam switching.

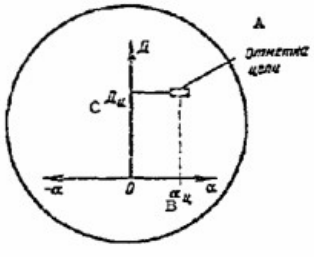


Figure 1.27. Screen of a range-azimuth scope.

A - pip; B - α_t; C - D_t.

Target azimuth is established from the position of the bright marker on the indicator screen relative to the zero azimuth line.

The principal advantages of the maximum-signal method are the simplicity with which bearing finding can be accomplished, and the possibility of taking bearings under the most favorable of signal/noise ratio conditions. The shortcoming in this particular method is the relatively low degree of accuracy in taking bearings.

The bearing error is

$$\Delta\alpha = \sqrt{\frac{2m}{|f''(0)|}} \qquad (1.51)$$

where

$m = U_0 - U/U_0$ is the relative error in establishing the maximum for the target signal;

U_0 is the true value of the signal maximum;

U is the measured signal value;

$f''(0)$ is the second derivative of $f(\alpha)$ when $\alpha = 0$;

$f(\alpha)$ is the direction finding response curve;

α is the angle between the cardinal direction and the direction to the target (target bearing).

When the signal classification is good the mean square value of the relative erros is $\sigma_m = 0.05$ to 0.15, and the mean square error in the bearing is equal to

$$\sigma_\alpha = (0{,}15 \div 0{,}25)\,\theta_{0,5},$$

where

$\theta_{0.5}$ is the width of the direction finding response curve at the 0.5 power level.

<u>The minimum-signal method.</u> The minimum-signal method, as applied to direction finding, means reading the bearing at the moment in time when the direction of the minimum for the direction finding response curve coincides with the direction to the target (fig. 1.28). A direction finding response curve with a sharply defined minimum can be obtained by cutting in two antennas that are opposite in phase.

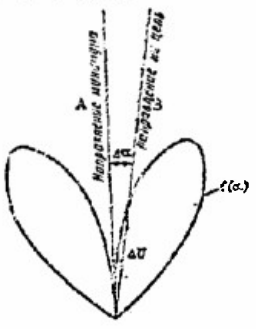


Figure 1.28. Direction finding by the minimum-signal method.
A - direction of the minimum; B - direction to the target.

The principal advantage of the minimum-signal method is much more accurate direction finding as compared with the maximum-signal method because the steepness of the direction finding response curve in the region of the

minimum signal is much more pronounced than in the region of the maximum signal.

This method is widely used in navigation because comparatively simple equipment can be used for direction finding, while providing acceptable accuracy.

The mean square error in the bearing in this case is

$$\sigma_\alpha = \frac{K}{\frac{2U_0}{U_n} f'(0)} \qquad (1.51a)$$

where

K is the proportionality factor (K = 0.2 to 0.5);

U_n is the mean square value of the noise voltage;

$f'(0)$ is the first derivative of the direction finding response curve at zero;

U_0 is the voltage across the receiver input when the direction of the maximum in the reception pattern coincides with the direction to the target.

The equisignal zone method. Realization of this method requires the availability of two reception patterns positioned in space as shown in Figure 1.29.

When the equisignal zone method is used for direction finding the bearing is read at that moment in time when the equisignal direction coincides with the direction to the target; that is, when the amplitudes of the signals from the target corresponding to each of the reception patterns are equal. In the simplest of cases direction finding by the equisignal method can be carried out as follows.

Let the antenna array consist of two identical antennas with the directions of their maxima in the reception patterns spaced at some angle $2\alpha_0$ (fig. 1.29). The signals reflected from the target are either received in turn by the antennas if there is but one receiver (one channel) in use, or by each of the antennas independently if there are two receiving channels.

The operator, rotating the antenna array in the plane of direction finding, attempts to arrive at that position in which the signals picked up by each of the antennas will be equal to each other. It is at this moment in time that the operator reads the bearing to the target off the instrument used to indicate the angular position of the antenna array.

The equisignal direction can be formed in one of several ways; the two antenna method (in the manner pointed out above); the three antenna method; the dephasing method; and the defocusing method.

The conical scan method is the one most often used in single-channel systems to obtain the equisignal direction. What is involved in this method is the shifting of the antenna exciter out of the focus of the parabolic

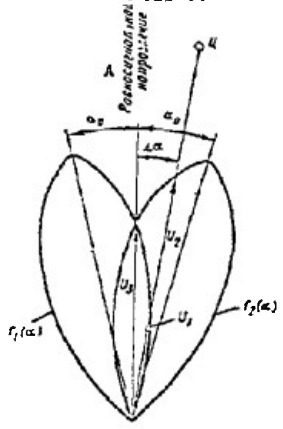


Figure 1.29. Bearing finding by the equisignal zone method.
A - equisignal direction.

mirror of the antenna and rotating it around the foxal axis of the mirror, thus causing the reception pattern to describe the figure of a cone in space, occupying the right, upper, left, and lower positions, successively.

The right and left positions are used to establish target azimuth, while the upper and lower positions are used to establish the elevation.

The principal advantages of the equisignal method of direction finding are greater accuracy than is the case when the maximum-signal method is used, and the capacity to establish the side to which the target is displaced from the equisignal direction, thus providing for automatic tracking in direction.

The accuracy of the equisignal method can be established by the accuracy in establishing the moment in time when the amplitudes of the signals corresponding to the two intersecting direction finding response curves are equal to each other.

The direction finding error in this case can be established through the formula

$$\Delta\alpha = \frac{m}{\left|\frac{f'(\alpha_0)}{f(\alpha)}\right|} = \frac{m}{\Pi}, \qquad (1.52)$$

where

m is the amplitude modulation factor, the result of the displacement of the direction finding response curves

$$m = \frac{U_1 - U_2}{2U_3};$$

U_1 and U_2 are the target signal voltages for the first and second direction finding response curves;

U_3 is the target signal voltage in the equisignal direction (see fig. 1.29);

$\Pi = \left| \frac{f'(\alpha_0)}{f(\alpha)} \right| \frac{1}{\text{rad}}$ has come to be called the direction finding capability (sensitivity).

The mean square error in the bearing is

$$\sigma_\alpha = \sigma_m / \Pi \text{ rad.}$$

When signal classification is good, $\sigma_m = 0.02$.

As will be seen from the formula at (1.52), the accuracy will be greater the steeper the direction finding response curve in the equisignal direction. And the steepness of the direction finding response curve in the equisignal direction will be greater the larger the angle α_0, by which the reception pattern is displaced from the equisignal direction.

Angle α_0 cannot be increased very much because if it is there will be a severe reduction in the magnitude of the reflected signal in the equisignal direction. There is an optimum angle of displacement, $\alpha_{0\ opt}$, at which the necessary signal magnitude and minimum bearing error can be ensured.

When receiving and transmitting antennas are used

$$2\alpha_{0\ opt} = 0.85\theta_{0.5}. \tag{1.53}$$

As a practical matter, $2\alpha_0$ will be somewhat smaller. Ordinarily $2\alpha_0 = 0.6\theta_{0.5}$, because detection range is reduced for optimum α_0.

The signal-comparison method. The direction to the target when the signal-comparison method is used for direction finding can be judged by the magnitude of the relationship between the amplitudes of two incoming signals corresponding to two intersecting direction finding response curves (fig. 1.30).

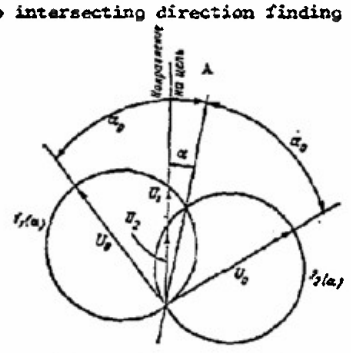


Figure 1.30. Direction finding by the signal-comparison method. A - direction to the target.

Let the direction finding response curves be identical, and let the angle between the direction to the target and the equisignal direction be equal to α. Then

$$U_1 = U_0 f(\alpha_0 - \alpha),$$
$$U_2 = U_0 f(\alpha_0 + \alpha). \tag{1.54}$$

And

$$\frac{U_1}{U_2} = \frac{f(\alpha_0 - \alpha)}{f(\alpha_0 + \alpha)}. \qquad (1.55)$$

By measuring this relationship we can establish the target bearing for known direction finding response curves. This method can be used with a single channel, as well as with a multichannel set. If the multichannel variant is used this method makes it possible to do what is, in principle, monopulse direction finding.

A characteristic feature of the signal-comparison method is that direction finding can be done with fixed direction finding response curves.

The advantage of the single channel arrangement for making the comparison is the simplicity of the equipment used.

The principal shortcoming of the single channel arrangement is the presence of additional direction finding errors, because a change in the target signal intensity during the switching period will yield a false mismatch voltage. Noise, the amplitude of which changes with switching frequency, can completely upset system operation.

The multichannel (monopulse) arrangement for making the comparison lacks these shortcomings. Any change in the strength of the incoming signal is reflected in equal measure on both of the signals being compared and does not show up in system operation. Therefore, so far as monopulse direction finders are concerned, there is no way that an effective angular noise from the same point in space as that occupied by the target proper can be created. In principle, the multichannel (monopulse) arrangement can take bearings on several targets individually, something that is impossible in single channel systems.

Direction finding using phase methods

Phase methods of radio direction finding establish the target bearing by the difference in the phases of the voltages across two receiving antennas, A and B (fig. 1.31), separated in space. The base between the antennas is equal to d.

The difference in the phases of the incoming oscillations equals

$$\varphi = 2\pi/\lambda\ (D_1 - D_2) = 2\pi/\lambda\ \Delta D, \qquad (1.56)$$

where

$\Delta D = D_1 - D_2$ is the propagation difference;

D_1 is the distance between points A and T;

D_2 is the distance between points B and T;

λ is the radar wavelength.

Since, ordinarily, $D_1 \gg d$, and $D_2 \gg d$, then

$$\Delta D = d \sin \alpha,$$

$$\varphi = 2\pi/\lambda\ d \sin \alpha, \qquad (1.57)$$

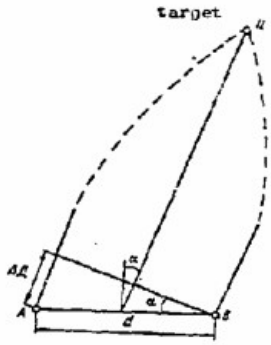


Figure 1.31. Direction finding by the phase method.

where

α is the target bearing, equal to the angle between the normal to the base, d, and the direction to the target.

What follows from (1.57) is that

$$\alpha = \arcsin \frac{\varphi}{\frac{2\pi}{\lambda} d}. \tag{1.58}$$

The accuracy of direction finding using the phase method is, in the main, dependent on the accuracy with which the difference in the phases of the incoming signals is measured, and on the magnitude of the base. The mean square error in the bearing is

$$\sigma_\alpha = \frac{\sigma_\varphi}{\frac{2\pi}{\lambda} d \cos\alpha}. \tag{1.59}$$

In turn

$$\sigma_\varphi = \frac{1}{\sqrt{\frac{P_s}{P_n}}} \tag{1.60}$$

where

P_s is signal power;

P_n is noise power.

The phase method of direction finding, in principle, permits monopulse direction finding with a multichannel direction finder.

Direction finding using amplitude-phase methods

These methods are based on the use of the amplitude, as well as the phase, relationships between the voltages across two receiving antennas separated in space.

An example of the amplitude-phase method of radio direction finding is the phase-antiphase method, otherwise known as the sum and difference method. In essence, the method is as follows. The antenna array consists

of two identical antennas spaced distance d apart. The directions of the maxima in the reception patterns for both antennas are perpendicular to base d. In the case of a single channel direction finder, when the two receiving antennas are cut in in-phase with each other, the summed voltage U_s (fig. 1.32) across the receiver input will be

$$U_s = 2U_0 f(\alpha) \cos\left(\frac{\pi d}{\lambda} \sin\alpha\right) \tag{1.61}$$

where

U_0 is the voltage across the receiver input when the direction of the maximum in the reception pattern coincides with the direction to the target;

$f(\alpha)$ is the direction finding response curve for the antenna;

d is the base between the antennas.

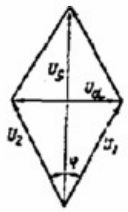


Figure 1.32. Establishing the sum and difference voltages.

If the antennas are cut in so their phases are opposite, the voltage across the receiver input will be

$$U_d = 2U_0 f(\alpha) \sin\left(\frac{\pi d}{\lambda} \sin\alpha\right). \tag{1.62}$$

If we measure the relationship

$$\frac{U_d}{U_s} = \operatorname{tg}\left(\frac{\pi d}{\lambda} \sin\alpha\right), \tag{1.63}$$

we can establish the target bearing.

This method can also be used for monopulse direction finding in the case of a multichannel arrangement in the direction finder.

The block schematic of one of the variants of a dual-channel monopulse radio direction finder for realization of this method of direction finding is shown in Figure 1.33. The magnitude of the deflection of the pip on the screen of the cathode ray tube caused by the intermediate frequency pulses depends on the direction to the target.

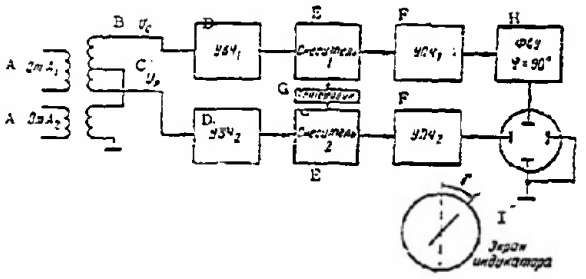


Figure 1.33. Block schematic of a dual-channel monopulse radio direction finder.

A - from; B - U_s; C - U_d; D - high frequency amplifier; E - mixer; F - intermediate frequency amplifier; G - local oscillator; H - photoelectric reader; I - indicator screen.

If the amplification in both channels, and the X and Y tube sensitivities are identical

$$\operatorname{tg}\gamma = \operatorname{tg}\left(\frac{\pi d}{\lambda}\sin\alpha\right), \tag{1.64}$$

where

γ is the angle between the mark on the indicator screen and the Y axis.

Consequently,

$$\gamma = \frac{\pi d}{\lambda}\sin\alpha, \tag{1.65}$$

from whence

$$\alpha = \arcsin\frac{\gamma}{\frac{\pi d}{\lambda}}. \tag{1.66}$$

The accuracy of this method of direction finding can be established by the accuracy with which angle γ is read. The mean square error in direction finding is equal to

$$\sigma_\alpha = \frac{\sigma_\gamma}{\frac{\pi d}{\lambda}\cos\alpha}. \tag{1.67}$$

When the signal/noise ratio is high σ_γ = 1 to 3°.

1.12 The Partial Patterns Method

In the case of the partial patterns method a specified sector is observed by overlapping it with several receiving antennas, each of which is connected to a separate receiver channel. The signals are fed from the receiver outputs into a special analyzer (fig. 1.34).

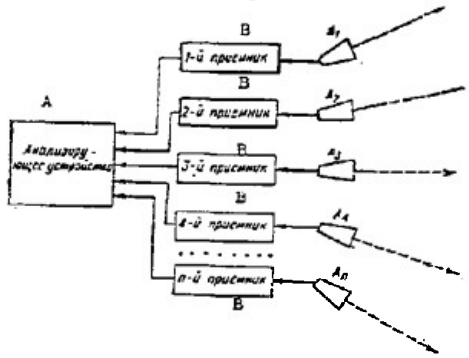


Figure 1.34. Block schematic of a radar with target direction finding by the partial patterns method.

A - analyzer; B - receivers as numbered.

The antenna reception patterns partially overlap (fig. 1.35). A coarse approximation of the direction to the target can be the number of the receiving channel, the input signal to which has the greatest amplitude. The precise bearing can be established by using one of the direction finding methods already considered that utilize intersecting reception patterns.

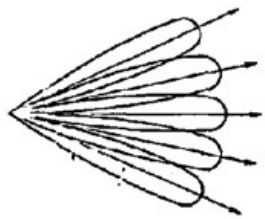


Figure 1.35. Partial reception patterns.

1.13 The Monopulse Method

In the monopulse radar each pulse reflected from the target carries all the information on target position, not only the angular coordinates, but range as well. The information is separated by simultaneously comparing the amplitudes and phases of the reflected signals received by several antennas.

Monopulse radars are used for automatically tracking by angular coordinates, for the most part, but surveillance radars can also be used. Automatic tracking in one plane requires two channels and two antennas, with four required for tracking in azimuth and elevation.

Monopulse radars are more complicated than single-channel radars, but do provide more accurate coordinate determination. This is because the low-frequency amplitude fluctuations in the reflected signals have no effect on the functioning of these systems.

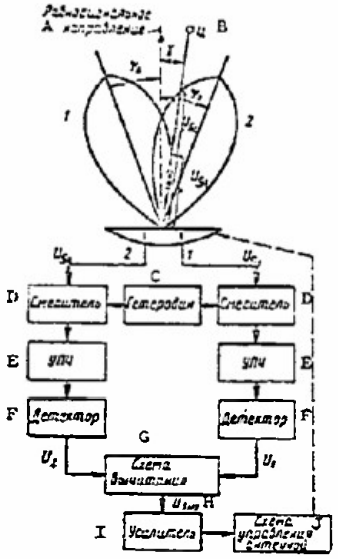


Figure 1.36. Block schematic of an amplitude difference radar.

A - equisignal direction; B - target; C - mixer;
D - local oscillator; E - intermediate frequency amplifier;
F - detector; G - subtract circuit; H - U_{out}; I - amplifier;
J - antenna control circuit.

Let us consider the principle of operation of the monopulse radar by using a very simple amplitude difference radar (fig. 1.36) in which the amplitudes of signals received by two channels in the radar (for direction finding in one plane) are compared in order to establish the direction to the target.

The antenna directivity patterns form the equisignal direction. The signals received by each of the antennas is amplified by separate receiver, detected, and then their difference is found.

The signal received by antenna 1 is in the following form at the receiver input

$$u_{c1} = kF(\psi_0 + \gamma)\cos(\omega t + \varphi), \tag{1.68}$$

where

k is the proportionality factor;

$F(\beta)$ is the directivity pattern for the antennas:

$$\beta = \psi_0 + \gamma,$$

ψ_0 is the angle of deflection of the maximum for the directivity pattern from the equisignal direction;

γ is the angle of deflection of the equisignal direction from the direction to the target (the angle of mismatch);

ω is the signal frequency;

σ is the phase.

The signal at the output from the second antenna is

$$u_2 = kF(\psi_0 - \gamma)\cos(\omega t + \sigma). \tag{1.69}$$

After transformation, intermediate frequency amplification, and linear detection, the signals in the radar channels at the input to the subtractor will equal, respectively,

$$\begin{aligned} u_1 &= kK_1F(\psi_0 + \gamma), \\ u_2 &= kK_2F(\psi_0 - \gamma), \end{aligned} \tag{1.70}$$

where

K_1, K_2 are the signal transfer functions in the channels.

The signal at the subtract circuit output is equal to

$$u_{out} = k\left\{(K_1 - K_2)F(\psi_0) - (K_1 + K_2)\left[\frac{dF(\beta)}{d\beta}\right]_{\beta=\psi_0}\gamma\right\} \tag{1.71}$$

for small mismatch angles.

If the channel transfer functions are equal to $K_1 = K_2 = K'$, the signal at the output will equal

$$u_{out} = 2K'k\left[\frac{dF(\beta)}{d\beta}\right]_{\beta=\psi_0}\gamma. \tag{1.72}$$

As will be seen from the formula at (1.72), the signal at the subtract circuit output is directly proportional to the mismatch angle. This signal is fed into the antenna control circuit that rotates the antenna so as to continuously match the equisignal direction with the direction to the target, that is, to reduce the mismatch signal to zero.

Figure 1.37 shows the direction finding curves for the system. The shortcoming in the system is the dependence of the zero value of the direction finding curve on the stability and the equality of the signal transfer functions in the individual channels on each other. There is no such shortcoming in the amplitude sum-difference radar (fig. 1.38).

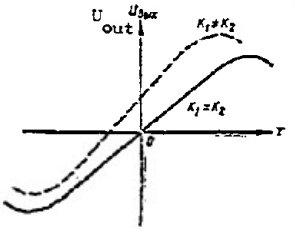


Figure 1.37.

Direction finding curve for an amplitude difference radar.

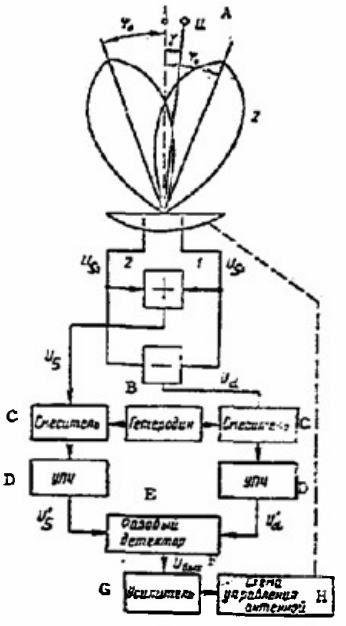


Figure 1.38. Block schematic of a sum-difference radar.

A - target; B - mixer; C - local oscillator; D - intermediate frequency amplifier; E - phase detector; F - U_{out}; G - amplifier; H - antenna control circuit.

The signal at the output of this system is equal to

$$u_{out} = 2K_1K_2K_{0a}F(\psi_0)\left[\frac{dF(\beta)}{d\beta}\right]_{\beta=\psi_0}\gamma\cdot\cos(\varphi_1-\varphi_2). \tag{1.73}$$

for small mismatch angles.

In this system the zero direction does not depend on the amplitude and phase characteristics of the channels ($u_{out} = 0$ when $\gamma = 0$). Change in the amplitude (K_1 and K_2) and phase (φ_1 and φ_2) characteristics of the channels will only result in some change in the steepness of the direction finding response curve.

We can have phase difference and phase sum-difference radars, as well as amplitude monopulse radars. They differ from those described above only in the fact that the phase, rather than the amplitudes, are compared.

1.14. Methods Used to Measure the Height at Which a Target is Flying

Target height can be established with respect to known slant range to it, and by the target elevation. At the same time, such factors as atmospheric refraction and the curvature of the earth's surface must be taken into consideration.

Atmospheric refraction can be taken into consideration by replacing the actual radius of the earth, R_e, by the so-called effective radius of the earth, R_{eff}. In the case of normal atmospheric refraction

$$R_{eff} = 4/3\ R_e = 4/3 \cdot 6370 = 8500 \text{ km} \tag{1.74}$$

For the case of $H \ll R_e$ and $D \ll R_e$

$$H = h_A + D \sin \varepsilon + D^2/2R_{eff}, \tag{1.75}$$

where

H is the height at which the target is flying;

h_A is the height of the radar antenna;

D is the slant range to the target;

ε is target elevation;

R_{eff} is the effective radius of the earth.

Ordinarily, $h_A \ll H$, therefore

$$H \approx D \sin \varepsilon + D^2/2R_{eff}. \tag{1.76}$$

The centimeter band and measuring the height at which a target is flying

The maximum, comparison, and V-beam methods are available for measuring target elevation when the centimeter waveband is used.

The method of the maximum. A beam, narrow in the vertical plane, and wide in the horizontal plane, continuously scans a specified elevation sector when the maximum method is used to determine height. Reflected signals are fed from the receiver output into the elevation-position indicator (fig. 1.39). In this indicator the range and elevation angle sweep is the result of feeding a sawtooth range voltage to the X plate and a voltage proportional to the angle of elevation to the Y plate.

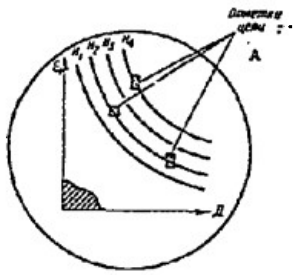


Figure 1.39. Screen of the elevation-position indicator.

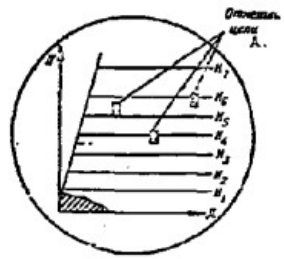


Figure 1.40. Screen of the height-range indicator.

A - target pips.

A series of curves of constant heights, computed through the formula at (1.76), are plotted on the indicator screen and are used to find target height.

A height-range indicator (fig. 1.40) can be used instead of the elevation-position indicator. Height sweep is formed by delivering a voltage in the form

$$U_y = kl \sin \epsilon, \tag{1.77}$$

where

k is the proportionality coefficient,

to the Y plate.

The method of the maximum can also be used to find the elevation angle by using partial directivity patterns.

The comparison method was described above.

The V-beam method. The V-beam method is, in essence, a radar antenna array consisting of two antennas forming two flat beams. One of the beams is vertical, the other tilted (fig. 1.41). The angle between the planes of the beams is usually equal to 45°. Both antennas are installed on the same house and rotate in the horizontal plane. The direction of rotation is such that the vertical beam is displaced before the tilted beam is, with the result that the target is illuminated twice per revolution of the antenna array. The angle α_t by which the antenna array is turned from the time the target is illuminated by the vertical beam to the time it is illuminated by the tilted beam depends on the height and range. The relationship between target height and this angle is

$$H \approx \frac{D \sin \alpha_t}{\sqrt{1 + \sin^2 \alpha_t}} \tag{1.78}$$

Accordingly, target height can be found by measuring the slant range and the difference in the azimuths of α_t.

The meter band radar and measuring target height. Target angles of elevation are measured by comparing the emfs induced in antennas at different heights above the surface of the earth (fig. 1.42). Each antenna has its own directivity pattern in the vertical plane. That of the lower antenna has one lobe, that of the upper two lobes, for example. Equisignal directions are formed in space by the intersection of the lobes, and these directions can be used to find the target's angle of elevation (fig. 1.42b), particularly when a goniometer is used.

A goniometer is a device with two mutually perpendicular stator coils and one rotor coil. The latter can be rotated in the magnetic fields of the stator coils (fig. 1.43). The output terminals of both antennas are connected to the stator coils. The rotor coil is connected to the receiver. The emf induced in the rotor coil will depend on the summed magnetic field created

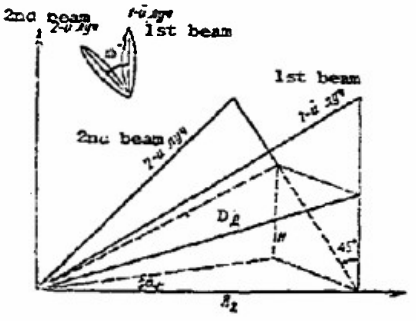


Figure 1.41. The V-beam method used to find target height.

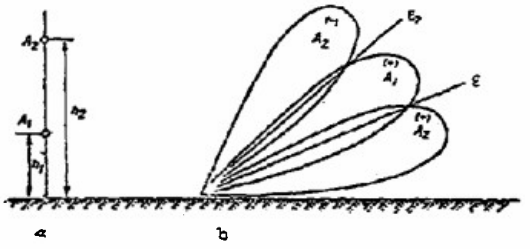


Figure 1.42. Measuring target angle of elevation by using two antennas placed at different heights.

a – antenna positions; b – directivity pattern positions.

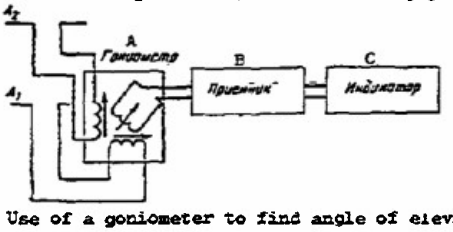


Figure 1.43. Use of a goniometer to find angle of elevation.

A – goniometer; B – receiver; C – indicator.

by the stator coils, while the magnetic field will depend on the currents flowing in the stator coils.

The current flowing in the first coil will depend on the signal received by antenna A_1

$$I_1 = k\varphi_1(\varepsilon), \tag{1.79}$$

where

ε is target angle of elevation;

$\varphi_1(\varepsilon)$ is the equation for the directivity pattern of antenna A_1;

k is the proportionality coefficient.

The current flowing in the second coil is

$$I_2 = k\varphi_2(\varepsilon). \tag{1.80}$$

If the rotor coil is turned so the resultant emf across its output is zero, the angle of rotation of coil γ will fix the target angle of elevation ε completely

$$-\operatorname{tg}\gamma = \frac{\varphi_1(\varepsilon)}{\varphi_2(\varepsilon)}. \tag{1.81}$$

We can construct the curve for $\gamma = \Phi(\varepsilon)$, for use in finding the angle of elevation ε, or we can graduate the goniometer scale to read the angle of elevation directly.

1.15. Moving Target Selection Systems

The moving target selection system is a set of special equipments able to separate signals from moving targets against a background of reflections from fixed and slowly moving targets. These targets include local objects, water surfaces, hydrometeorological objects (clouds, rain, hail, snow), and others. Interfering reflections from all of the above targets are called passive noises.

Passive noise can exceed the level of receiver noise in intensity by 30 to 80 db, resulting in overloading the receiver and loss of useful signal. The useful signal can be lost even when there is no overloading, the result of the masking effect of the noise.

Useful signal and passive noise, alike the result of the phenomenon of secondary emission of electromagnetic energy from the radar transmitter have much in common, so far as their characteristics are concerned. The basic difference, and the one on which moving target selection is based, is the difference in the frequencies of the reflected signals, the result of the different radial components of the speed at which the target is moving and the sources of the passive noise.

A transmitter radiates oscillations

$$u_{trans}(t) = U_m \cos(\omega_g t + \varphi_0). \tag{1.82}$$

Oscillations reflected from a fixed target at distance D_0 are in the form

$$u_{fix}(t) = U_m \cos[\omega_g t - 2D_0/s\ \omega_g + \varphi_0]. \tag{1.83}$$

Oscillations reflected from a target at distance D_0, moving in a straight line and uniformly with radial speed v_r, have the form

$$u_{tar}(t) = U_m \cos[\omega_g t - 2D_0/s\ \omega_g \pm 2v_r/s\ \omega_g t + \varphi_0], \tag{1.84}$$

where

$2D_0/s\ \omega_g = \varphi_{D_0}$ is the constant lag resulting from the distance to the reflecting target;

$2v_r/s\ \omega_g = \Omega_D$ is the Doppler frequency resulting from the radial movement of the target.

The appearance of the Doppler frequency in the moving target signal is used for moving target selection.

All that is necessary to detect the signals from moving targets against the background of the reflections from fixed targets when using a continuously radiating radar is to feed transmitter oscillations into the receiver simultaneously with the reflections from the target. Beating of the oscillations from the transmitter and the reflected signal establishes the fact that a moving target is present and establish its radial speed. However, a system such as this will not establish the distance to the target directly.

Moving target selection can also be obtained by the use of pulse radars, and here the distance to the target can be established through the use of conventional pulse methods.

Coherent pulse moving target selection systems

The place of the moving target selection system in pulse radars can be seen from Figure 1.44. Pulse moving target selection systems are called coherent pulse systems.

In these systems the oscillations from a special, so-called coherent oscillator are used as the reference oscillations and the reflected signals are compared with them. These oscillations are closely synchronized in phase with the transmitter oscillations, so closely that the difference in phases between the two oscillations is constant during each period in the pulse train. This type of oscillation is called coherent. The most widely used circuitry in the centimeter waveband is that in which the phasing of the coherent oscillator and the combining of the coherent and reflected oscillations takes place on an intermediate frequency.

We can review the operation of coherent pulse moving target selection systems by using the schematic shown in Figure 1.44 as an example.

The antenna radiates the pulses of high frequency energy from the transmitter into space.

Oscillations from the transmitter, converted into pulses at the intermediate frequency (phasing pulses) by the voltage from the local oscillator, are fed into the coherent oscillator to phase it. During phasing the frequency and phase of the coherent oscillator oscillations are equal to the frequency and phase of the phasing pulse oscillations.

When the phasing process is concluded the phase of the coherent oscillator oscillations is closely coupled with the phase of the transmitter oscillations, although it is not equal to it. The result is coherency in transmitter and coherent oscillator oscillations.

The phased voltage is applied to the phase detector from the coherent oscillator. The phase detector also takes the reflected signal, converted into a signal at intermediate frequency by the voltage from this same local oscillator.

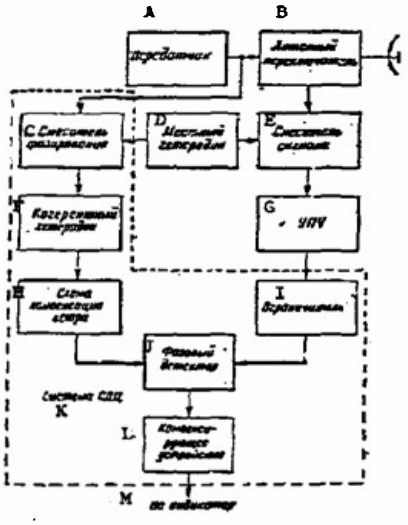


Figure 1.44. Simplified block schematic for moving target selection.

A – transmitter; B – antenna switch; C – phasing mixer; D – local oscillator; E – signal mixer; F – coherent oscillator; G – intermediate frequency amplifier; H – wind compensation circuit; I – limiter; J – phase detector; K – moving target selection system; L – compensator; M – to indicator.

An amplitude limiter is installed ahead of the phase detector, and its purpose is to eliminate parasitic modulation of signals from fixed targets caused by antenna rotation as space is canned.

The reflected signal voltage after the mixer in the nth period of repetition can be represented in the form

$$u_s = U_{m\,s}\sin[(\omega_g + \Omega_D - \omega_{lo})t + \Omega_D(n-1)T_p - \varphi_{D_0} - \varphi_g + \varphi_{lo}], \qquad (1.85)$$

where

ω_g is the generator frequency;

Ω_D is the Doppler frequency;

ω_{lo} is the local oscillator frequency;

T_p is the pulse repetition period;

φ_g is the initial phase of generator oscillations;

$n = 1, 2, 3...$;

φ_{lo} is the initial phase of local oscillator oscillations.

The coherent oscillator voltage is

$$u_{co} = U_{m\,co} \sin\,[(\omega_g - \omega_{1o})t - \varphi_{co}]. \tag{1.86}$$

The voltage across the phase detector output is

$$u_{pd} = U_{m\,pd} \cos\,[(\omega_g + \Omega_D - \omega_{1o})t - (\omega_g - \omega_{1o})t + \Omega_D$$

$$(n-1)T_p - \varphi_{D_o} - \varphi_g + \varphi_{1o} + \varphi_{co}] = U_{m\,pd} \cos[\Omega_D t +$$

$$+ \Omega_D(n-1)T_p - \varphi_{D_o} + \varphi_{1o} - \varphi_g + \varphi_{co}], \tag{1.87}$$

where

$U_{m\,pd} = 1/2\, U_{m\,s} U_{m\,co}$.

If generator, local oscillator, and coherent oscillator frequencies are stable the fixed target ($\Omega = 0$) video signals at the phase detector output are constant in amplitude from period to period in the trains. When the target is a moving one ($\Omega \neq 0$) the video pulses at the phase detector output will have amplitudes that will change from period to period in accordance with the law $\cos \Omega_D t$.

The envelope of the video pulses will change in accordance with a harmonic law with frequency F_b, equal to the Doppler frequency, only when

$$F_D < F_p/2,$$

where

F_p is the pulse repetition frequency.

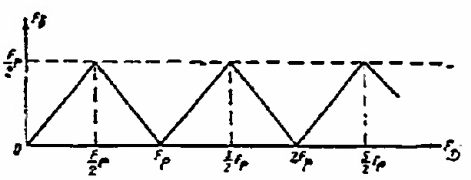


Figure 1.45. Dependence of the video pulse envelope frequency on the Doppler frequency.

If this condition is not satisfied there will be a stroboscopic effect when frequency F_b of the video pulse envelope changes with the Doppler frequency in accordance with a sawtooth law (fig. 1.45).

Moving target selection compensators

Signals flow from the phase detector to a compensator.

The compensator makes the alternate period subtraction of the video signals from the phase detector output whereas the video signals with fixed

amplitude are suppressed and the pulses, the amplitudes of which change from period to period, are isolated.

The compensator can be a delay line or a storage tube.

A compensator in the form of an ultrasonic delay line is shown in Figure 1.46.

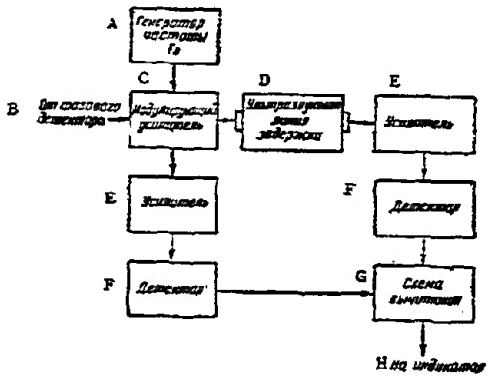


Figure 1.46. Block schematic of a compensator in the form of an ultrasonic delay line.

A - f_o frequency generator; B - from phase detector; C - modulating amplifier; D - ultrasonic delay line; E - amplifier; F - detector; G - subtraction circuit; H - to indicator.

Signals pass over two channels so the alternate-period subtraction can be done in the compensator, over a direct channel, and over a channel with a delay of T_p, and are fed into the subtraction circuit where the difference signals are isolated.

Pulses are delayed for the period of repetition in the channel with the ultrasonic delay line.

The video pulses are converted into radio pulses with an auxiliary frequency, f_o (10 to 12 mHz) so the transmission through the ultrasonic delay line will be undistorted. The signal at the output from the direct channel is in the form

$$u_{dir} = U_m \cos [\Omega_D t + \Omega_D (n-1) T_p - \varphi]. \quad (1.88)$$

The signal at the output of the channel with the delay is

$$u_{delay} = U_m \cos[\Omega_D t + \Omega_D n T_p - \varphi]. \quad (1.89)$$

The signal at the output of the compensator (after subtraction) is

$$u_{com} = u_{dir} - u_{delay} = 2U_m \sin(1/2\,\Omega_D T_p) \sin[\Omega_D t + \Omega_D T_p (n - 1/2) - \varphi]. \quad (1.90)$$

As will be seen from this formula, the difference signal equals zero for signals from fixed targets ($\Omega_D = 0$), but differs from zero for signals from moving targets.

If good compensation is to be provided signals from fixed targets there must be strict equality between the summed delay times for the signals, T_{delay}, and the repetition period, T_p. This greatly complicates the compensator circuit.

More modern are compensators with memory tubes combining the functions of memory and compensator.

The memory tube compensator maintains the equality $T_{delay} = T_p$ automatically at all times, so strict constancy of T_p is not required.

Target "blind" speeds

The frequency curve for a compensator in the case of single subtraction

$$K(F) = 2\left|\sin(\pi F_D T_p)\right| \quad (1.91)$$

is in the form depicted in Figure 1.47.

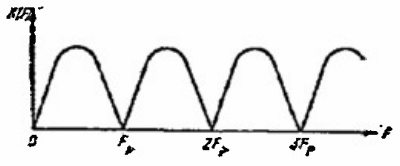


Figure 1.47. Frequency curve for a compensator.

The Doppler frequencies, F_D, for which the product is

$$F_D T_p = n, \quad (1.92)$$

and where n = 1, 2, 3, ..., that is, Doppler frequencies, multiples of the pulse repetition, yield a signal equal to zero at the compensator output. The radial target speeds causing these Doppler frequencies, that is,

$$v_{r\ blind} = n \cdot \lambda F_p/2, \quad (1.93)$$

are called "blind" speeds. The reason is that in time T_p the target covers distances equal to multiples of $\lambda/2$.

Consequently, targets moving at "blind" speeds cannot be detected.

A variable repetition frequency is used to avoid "blind" speeds in radars with moving target selection. While the speed will be blind on one repetition frequency, it will differ from the "blind" speed on another, and the target can be detected.

The repetition frequency can be changed smoothly, or in bounds.

Change in radar frequency can, in principle, be used to cope with "blind" speeds, but this is a more complicated method.

"Wind compensation" devices

Certain types of passive interference sources (rain clouds, chaff clouds) move about with the wind as a single target, that is, they have a regular speed component. At the same time, the video pulses turn out to be modulated by the frequency of the Doppler beats, and uncompensated interference residuals will be present at the compensator output. The movement of the source of interference can be compensated for by removing the uncompensated residuals. The frequency of the oscillations of the coherent oscillator can be changed by a special, so-called "wind compensation" device, so as to change the frequency of the signal reflected from the moving source of interference, and this will eliminate the residuals. The result is to make the phase of the interference constant from period to period with respect to the coherent oscillations, and interference compensation quality will be improved.

1.16 Externally Coherent Pulse Moving Target Selection Systems

The use of internally coherent moving target selection is difficult because of the expansion in the band of Doppler frequencies of the passive interference. Two factors are involved in this expansion: the increase in the ceiling of the radar detection zones; and the shortening of the wave length on which the radars operate. In the centimeter band the band of Doppler frequencies of the passive interference can reach a magnitude such that it is difficult to suppress the interference in an internally coherent moving target selection system.

In this case externally coherent moving target selection systems are used, and these are based on the same principles as the internally coherent moving target selection systems, except that the coherent oscillator is phased with the passive interference signal, rather than with the main signal, or the passive interference itself is used as the reference voltage. In order to prevent useful signals from targets flying in an interference cloud being suppressed by this moving target selection system, the signal used for phasing the coherent oscillator (or used as the reference voltage) is delayed for a period of time qual to, or slightly longer, than the length of the radar pulse. Compensation for wind influence in these systems is automatic. However, the moving target can be selected in these systems only if a passive interference cloud is present. If passive interference is not suppressed in the radar receiving pattern the phasing of the coherent oscillator can only be done by target signals, and this will lead to

suppression of useful signals. Moreover, externally coherent systems will not suppress the leading edge of the interference cloud. Broad band compensation for wind effect can be realized in dual-frequency radars without the shortcomings inherent in the externally coherent method.

1.17 The Dual-Frequency Method of Moving Target Selection

A radar is said to be a dual-frequency radar when it radiates radio frequency pulses, and receives reflected signals, on two different frequencies at the same time. In the dual-frequency radar the same transformation takes place in front of the phase detector, just as it does in the single-frequency radar, but in the two sub-channels in the receiver that correspond to the two different frequencies of the main and reflected signals.

In this case the reflected signal frequency is

$$f_{I\ ref} = f_I + F_{D1}, \tag{1.94}$$

where

$F_{D1} = 2v_r/\lambda_1$;

$$f_{II\ ref} = f_{II} + F_{D2}, \tag{1.95}$$

where

$F_{D2} = 2v_r/\lambda_2$.

f_I and f_{II} in the formulas at (1.94) and (1.95) are the carrier frequencies for the transmitters in the dual-frequency radar.

Two signals, with frequencies as follows, are fed into the phase detector after the frequency conversion.

$$f_{I\ tf} = f_{tf} \pm F_{D1};\ f_{II\ tf} = f_{tf} \pm F_{D2}. \tag{1.96}$$

The echo signals are added to each other geometrically, and not to the coherent voltages, in the phase detector of the dual-frequency radar. The result is heating, detected by the amplitude detector. Video pulses, the envelope of which will change with difference in the Doppler frequency, are formed at the detector output.

$$F_b = F_{D1} - F_{D2} = 2v_r/\lambda_1 - 2v_r/\lambda_2 = (f_I - f_{II})2v_r/s. \tag{1.97}$$

If, by way of example, the frequencies for transmitters f_I and f_{II} are equal to 1550 and 1500 mHz, respectively, in the case of the single-frequency radar the Doppler frequency bands for the passive interference moved by the wind at speeds of from 0 to 50 meters/second would cover the sections of the frequency curve from zero to F_{D1} = 517 hertz, and F_{D2} = 500 hertz, respectively, and in the case of the dual-frequency radar $F_b = F_{D1} - F_{D2}$ = 517 - - 500 = 17 hertz.

Naturally enough, it is quite easy to suppress passive interference in so narrow a band of Doppler frequencies.

Chapter II

Radar Tactical and Engineering Data

Radar tactical and engineering data are data characterizing the expected efficiency of the radar to carry out concrete tactical missions in combat.

Radar can be designated in accordance with its tactical purpose as:

warning;

early warning;

detection and guidance;

acquisition;

rocket guidance, and others.

Single purpose radars can establish different numbers of coordinates and target characteristics.

The most important radar tactical and engineering data are:

limiting range;

limiting angular coordinates;

period of scan;

resolution;

accuracy in establishing coordinates;

information capability;

interference immunity;

standards for servicing in combat (setting up time, time required to energize, striking time, and the like);

climatic conditions for combat utilization.

2.1 Maximum Operating Range

Radar range is based on maximum, R_{max}, and minimum, R_{min}, operating range.

The maximum operating range (detection, for example) is specified by tactical requirements. It can be evaluated, approximately, through the basic radar equation.

Maximum operating range depends on many of the radar's engineering characteristics, on the conditions under which radio waves are propagated, and on target characteristics, which, in the real conditions prevailing when the radar is used in combat, are all subject to random changes.

Accordingly, evaluation of maximum operating range is a probability.

The value of maximum operating range, R_{max}, is usually indicated with respect to a specific target (a fighter plane, for example) and with a specified probability.

Preliminary tests can be used when developing new radars to establish R_{max} and to confirm the probability of realizing that R_{max}. It can usually

be shown that R_{max} is some value for n% (90%, for example) of all the tests conducted.

This determination of R_{max} means that this magnitude can be guaranteed in n% of all cases of target detection. In 100 - n% of the cases the target cannot be detected at ranges equal to R_{max}, but will be detected at somewhat lesser distances.

The operating range of a long-range detection radar is usually limited to the range of direct radar visibility, R_v, and this can be established through the formula

$$R_v = 4{,}1(\sqrt{h_A} + \sqrt{H}), \tag{2.1}$$

where

R_v is in kilometers;

h_A is the radar antenna height, in meters;

H is target height, in meters.

2.2 Minimum Detection Range

Minimum radar detection range, R_{min}, depends on the limits within which the antenna array can function with respect to the angle of elevation ε. It differs for different heights, and is established by the magnitude of the dead funnel.

For a cosecant radiation pattern

$$R_{min} = H/\sin \varepsilon_{max}, \tag{2.2}$$

where

ε_{max} is the elevation limit.

In ground radars with small elevation angles the real value of R_{min} can be established by the flares on the indicator screen at the sweep origin by the reflections from local objects.

If the antenna array imposes no limitations, the minimum operating range for the radar can be established by the pulse duration, τ_p, by the antenna switch restoration time, t_r, and the indicator resolution, δR_i

$$R_{min} \approx c/2\ (\tau_p + t_r) + \delta R_i. \tag{2.3}$$

2.3 Radar Azimuth Limits

Most of the existing radars operate in the circular scan mode, that is, the azimuth, β, is established from 0 to 360°.

Some radars use sector scan, dictated by tactical considerations.

2.4 Radar Elevation Operating Limits

Radar operating limits in the vertical plane are determined by the elevation sector scan, ε_{min}, and a maximum elevation, ε_{max}.

The minimum elevation is limited by the conditions for forming the radiation pattern, with the effect of the ground, and design and operational nature, taken into consideration. ϵ_{min} is usually equal to several tens of degrees for a radar in the centimeter band, and from 1 to 3° for a radar in the decimeter and meter bands. The maximum elevation for modern detection radars is within limits ranging from 30 to 45°.

A reduction in the magnitude of ϵ_{min} increases the operating range against low-flying targets, and is obtained by raising the antenna array above the ground surface.

An increase in ϵ_{max} increases the radar's ability to detect high targets. Engineering-wise, it is arrived at by complicating the design of the antenna arrays, by increasing the number of operating channels in the radar, and by increasing the power of the radio frequency transmitters.

2.5 The Radar Detection Zone

The radar detection zone is the space within the limits of which the radar can detect a target with a specified probability, and can measure target coordinates with the required accuracy.

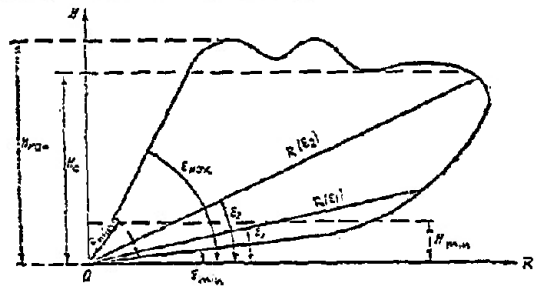


Figure 2.1. The radar detection zone. Detection range depends on target height. H_c is height of the ceiling for the radar. It is possible to find targets at heights of $H > H_c$ in the scan zone. H_{min} is the target height at which a low-flying target can be detected at a specified range.

The detection zone (fig. 2.1) in the vertical plane can be represented graphically in coordinates of the height, H, and the range, R. Operating range differs for different elevations, ϵ. The shape of the detection zone depends on the purpose for which the radar is intended, and the engineering decisions made in the radar design.

2.6 Radar Height Ceiling

The radar height ceiling, H_c, is the maximum height at which a target can be detected at any range within the limits of the detection zone.

Targets can be detected at heights above H_c at certain ranges within the limits of the scan zone.

2.7 Scan Period

The radar scan period, T_{scan}, is the interval of time required to illuminate all the points in the space that is the radar's scan zone.

Continuity and required definition in the reflected signal indications can be provided, given the condition that

$$T_{scan} \geqslant N_{p\ min} \Delta\beta\ \Delta\varepsilon / F_p \varphi_{0.5} \theta_{0.5} \qquad (2.4)$$

where

T_{scan} is in seconds;

$N_{p\ min}$ is the minimum number of pulses reflected from the target needed to detect the target with the specified probability ($N_{p\ min}$=3 to 25);

F_p is the pulse repetition frequency;

$\Delta\beta$ is the radar's sector scan in the horizontal plane;

$\Delta\varepsilon$ is the sector scan in the vertical plane;

$\varphi_{0.5}\theta_{0.5}$ are the widths of the antenna radiation patterns in the horizontal and vertical planes at the 0.5 power level, respectively.

In detection and acquisition radars, $\Delta\varepsilon = \theta_{0.5}$ and

$$T_{scan} \geqslant N_{p\ min} \Delta\beta / F_p \varphi_{0.5} \qquad (2.5)$$

The scan period determines the interval between two successive measurements of target coordinates. Given modern target speeds, the smoothness with which the pips are moved on the scopes, and the accuracy with which the target trajectories are reproduced are greater the shorter the scan period.

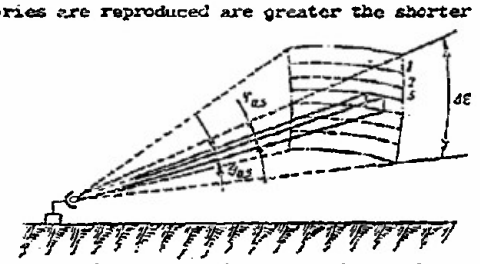


Figure 2.2. Scan in an elevation sector by one beam.

One way to reduce the scan period is to increase the aperture angle in the radar antenna radiation pattern. However, narrow radiation patterns are necessary to improve the resolution and increase accuracy in establishing coordinates. The contradictions can be reconciled by a compromise selection of the radiation pattern and the method used to scan space.

Circular scanning radars have a scan period equal to the time for one complete antenna revolution

$$T_{scan} = 60/n_A, \tag{2.6}$$

where

n_A is antenna rpm.

A radar can have sector scan in azimuth, as well as in elevation. Elevation sector scan can be by one beam (fig. 2.2), or by several beams (fig. 2.3), each of which has its own zone. In the latter case, the scan period is shortened.

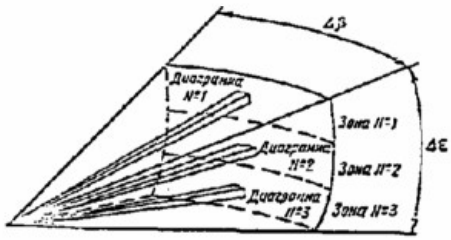


Figure 2.3. Elevation sector scan by several beams. Patterns 1, 2, 3; zones 1, 2, 3.

Given screw scan (fig. 2.4), one can make a circular scan in a zone. The scan period for the zone equals the time required for the antenna to make the several revolutions needed to scan the zone

$$T_{scan\ zone} = 60/n_A \cdot \Delta\varepsilon_z/\theta_{0.5}, \tag{2.7}$$

where

$\Delta\varepsilon_z$ is the zone sector.

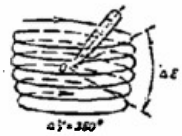


Figure 2.4. Screw scan.

The full scan period, assuming zones are identical, and that the scan mode does not change from zone to zone, equals

$$T_{scan} = 60/n_A \cdot \Delta\varepsilon_z/\theta_{0.5}\, n_z k_r, \tag{2.8}$$

where

n_z is the number of zones;

k_r is a factor that takes into consideration the time needed to reverse antenna movement; it depends on the number of zones, and on the sequence and the method used to shift from zone to zone; k_r = 1 to 3.

2.8 Range Resolution

Radar resolution is the capacity of the radar to make separate observations and measurements of the coordinates of two targets close together.

Radar range resolution, δR, is the minimum distance in range between two targets with the same angular coordinates at which it is still possible to make separate observations and range measurements for each target

$$\delta R = c\tau/2 + \delta R_s, \tag{2.9}$$

where

c is the electromagnetic energy propagation rate;

τ is pulse length;

δR_s is scope resolution (see Chapter IX).

Resolution in conventional pulse radars is higher (δR is less) the shorter the pulse length and the higher the scope resolution. It is primarily dependent on the length of the main pulse, τ_p.

The first term in the right side of the equation at (2.9) is determined by the potential resolution, δR_p, of the radar

$$\delta R_p = c\tau/2. \tag{2.10}$$

Resolution in a radar with intrapulse frequency modulation and pulse compression during the processing of the reflected signal can be determined by the duration of the "compressed" pulse, τ_{cp},

$$\delta R_p = c\tau_{cp}/2. \tag{2.11}$$

Resolution in a radar with phase-code shift keying can be determined by the duration of the code interval, τ_{pc},

$$\delta R_p = c\tau_{pc}/2 \tag{2.12}$$

where

$\tau_{pc} = \tau_p/n$;

τ_p is the length of the main pulse;

n is the number of code intervals.

2.9 Azimuth Resolution

Radar azimuth resolution, $\delta\beta$, is the difference in azimuths between targets at the same range, and at the same altitude, but as close to each other as they can be and still have the radar make separate measurements of the azimuth of each target

$$\delta\beta = \varphi_{0.5} + \delta\beta_p \qquad (2.13)$$

where

$\varphi_{0.5}$ is the width of the radiation pattern at half power in the horizontal plane;

$\delta\beta_p$ is the azimuth resolution for the scope (see Chapter IX).

For circular scan scopes, for example,

$$\delta\beta = \varphi_{0.5} + 57.3\ d_s/r_D, \qquad (2.14)$$

where

r_D is the distance of the marker from the center of the scope;

d_s is the diameter of the bright spot on the scope.

2.10 Elevation Resolution

Radar elevation resolution, $\delta\epsilon$, is the difference in the elevations of two adjacent targets at the same range and at the same azimuth at which it is still possible to make separate measurements of elevation of each of the targets.

Radar resolution $\delta\epsilon$ with beam nodding in the vertical plane equals

$$\delta\epsilon = \theta_{0.5} + \delta\epsilon_s \qquad (2.15)$$

where

$\theta_{0.5}$ is the width of the radar antenna radiation pattern in the vertical plane;

$\delta\epsilon_s$ is the scope elevation resolution.

What follows from (2.13) and (2.15) is that the resolution of the angular coordinates depends primarily on the width of the antenna radiation pattern in the corresponding plane. The first summands in these equations determine the potential radar resolution of angular coordinates.

2.11 Height Resolution

A radar with high elevation resolution, and particularly one using the partial patterns method, or the beam nodding in the vertical plane method, to determine height, has height resolution.

Radar height resolution is the minimum difference in height between two targets at the same distance, and at the same azimuth, at which the height of each of the targets can still be measured separately (fig. 2.5).

For the radar with beam nodding in the vertical plane

$$\delta H \approx \theta_{0.5} R/\cos\epsilon + \delta H_s, \qquad (2.16)$$

where

R is the range to the target;

ϵ is the present elevation;

δH_s is the height scope resolution.

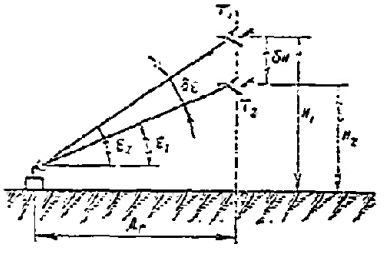


Figure 2.5. Radar height resolution.

2.12 Radar Resolution Volume

Radar resolution (pulse) volume is the volume of the part of space in the radar's scan zone delimited by the distances equal to the radar's resolution in range and angular coordinates (fig. 2.6). It equals

$$\delta V = D^2 \delta D \delta\beta \delta\varepsilon \qquad (2.17)$$

or

$$\delta V = 4\pi D^2/G \cdot c\tau/2, \qquad (2.18)$$

where

G is the antenna directive gain.

Targets within the limits of the radar resolution volume are detected as a single target.

The shorter the main pulse, and the sharper the radiation pattern of the antenna array, the smaller the pulse volume, and, consequently, the better the radar's resolution.

A radar with a smaller pulse volume is less prone to the effects of passive interference.

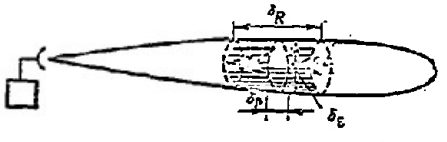


Figure 2.6. Resolution (pulse) volume of a radar.

Resolution in the case of automatic tracking. Radar resolution in the case of automatic tracking is understood to mean the minimum difference in target coordinates at which positive automatic tracking of each of the targets is still possible.

2.13 Accuracy in the Determination of Target Coordinates

The magnitude of the errors found in coordinate measurements is a criterion of the accuracy in determining target coordinates.

Errors can be external and instrumental, according to source, and can be coarse, systematic, and random, according to their behavior patterns.

Coarse errors, or misses, occur as a result of errors in calculations and are readily eliminated. A well-drilled radar team will not permit them to occur.

Systematic errors in radar operation remain constant, or change in accordance with a known law. These errors can be established in advance and taken into consideration. Personnel operating the radar should know the sources of possible systematic errors and should always eliminate their causes, as well as know how to evaluate permissible errors and introduce the necessary correction factors in measurement results.

Systematic errors are usually instrumental errors. They depend on the accuracy with which the radar is tuned and adjusted, the condition of the equipment, errors in locking and orienting the radar on the terrain, the degree of training given the operators, and the like.

Systematic errors can also be external errors, and include:

errors caused by the ground, and local objects;

errors affecting wave propagation, the result of concrete climatic conditions;

errors that depend on the nature and maneuvers of the target.

Random errors are inevitable errors, the result of the random nature of all the processes in the radar assemblies and units, in the propagation of radio waves, in the reflections from the target, in the observations made on the screen, and the like.

Random errors can be computed through the statistical theory for measuring radar signal parameters, which is based on the fact that the presence of various noises (interference) in the equipment for receiving and processing the signals cause the appearance of random errors in measurement. Random errors in operating radars are established by preliminary tests. It is the random errors that establish the accuracy with which target coordinates are measured.

Mean square σ, average (probable), and maximum errors are used to evaluate random errors.

Often used in practice as well is the error characterizing the definite probability that such error will appear.

Mean square error

The mean square error is

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}, \qquad (2.19)$$

where

$x_i = a_i - X$ is the random error in the ith measurement;

a_i is the result of the ith measurement;

X is the true value of the coordinate being measured;

n is the number of measurements.

The mean square error can be computed with sufficient accuracy even when the number of measurements is comparatively small ($n \geq 10$).

If the mean square errors $\sigma_1, \sigma_2, \ldots, \sigma_n$, depending on various independent sources, are known, the resultant mean square error is

$$\sigma_\Sigma = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2} \tag{2.20}$$

Average error

The average eeror is equal to the average arithmetical error, derived from the absolute values of the random errors in a series of measurements

$$x_{av} = \pm \frac{|x_1| + |x_2| + \ldots + |x_n|}{n} \tag{2.21}$$

The normal law for the distribution of random errors

Random errors have a normal distribution law

$$w(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2}$$

where

$w(x)$ is the probability density of random errors;

x is the random error;

σ is the mean square error.

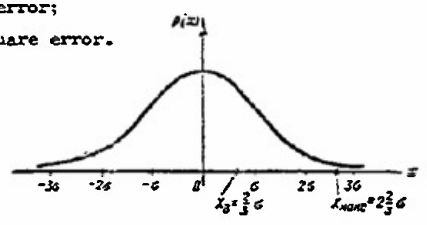


Figure 2.7. The normal law for the distribution of probability of random errors.

Figure 2.7 depicts in graphical form the normal law for the distribution of random errors. When measurements are more precise the measurement errors are smaller, but in any case

$$\int_{-\infty}^{+\infty} w(x)\,dx = 1,$$

so the probability density of errors is greater the closer to $x \approx 0$. There is an equal probability of the errors being positive and negative.

The probability of a definite error, x_k, can be established through the equation

$$P(x_k) = 2\int_0^{x_k} w(x)\,dx.$$

The probability of the mean square error is $P(\sigma) = 0.683$, for example. What this means is that of all measurements made the error will not be in excess of σ in 68.3% of the cases, whereas in 31.7% of the measurements the error will be larger than the mean square error.

Probable error

The probable, x_{prob}, or average error is that value of the error with respect to which the equal probability of the random error is greater, the smaller it is. The probability of an average error is $P(x_{prob}) = 0.5$. Consequently, 50% of the measurements have an error less than x_{prob}, and 50% of the measurements have an error greater than x_{prob}.

The probable error is in the following ratio in terms of the mean square error

$$x_{prob} = 2/3\ \sigma, \tag{2.22}$$

Maximum error

The maximum error is the largest random error possible under predetermined measurement conditions. It is taken as equal to

$$x_{max} = 4x_{prob}. \tag{2.23}$$

Consequently,

$$x_{max} \approx 3\sigma. \tag{2.24}$$

The probability of a maximum error occurring in a specified series of measurements is $P(x_{max}) = 0.993$. Consequently, only 0.7% of the measurements have an error greater than x_{max}.

Errors of given probability

Statistical errors, based on special statistical measurements, are often used to evaluate the accuracy with which target coordinates are determined by radars. In such cases a great many measurements of target coordinates are made and after the results have been processed mathematically it can be established that in n% of the measurements of the total measurements made the error, x_n, will not exceed some specific value.

Errors for 85 or 95% of the measurements are most often given. These errors are, naturally, in a definite ratio to the mean square error

$$x_{0.85} = 1.44\sigma, \tag{2.25}$$

$$x_{0.95} = 2\sigma. \tag{2.26}$$

It is convenient to use errors of given probability in day-to-day practical work. Mean square, and other errors are usually used in theoretical research, and for engineering computations.

2.14 Accuracy in Range Measurement

The modern theory of the measurement of radar signal parameters is what makes it possible to evaluate the errors in the measurement of target coordinates with an accuracy adequate for practical purposes.

The accuracy with which range is measured depends on the accuracy with which the delay in the reflected signal is measured, on errors because of less than optimum processing of the signals, on the presence of unanticipated delays in the signal in the transmission, reception, and indication channels, and on random errors in the measurement of range in the scopes used.

The errors in the scopes are the result of instability in the scale markers, and of errors in reading. The latter is a compound of the error in establishing the center of the marker and errors in interpolation (see Chapter IX).

The potential accuracy in the radar range measurement can be characterized by the mean square error, σ_{pot}, equal to

$$\sigma_{pot} = c\tau_p/2\sqrt{2\pi\nu_{v\ one}} \tag{2.27}$$

where

$\nu_{v\ one}$ is the value of the visibility factor for one pulse.

The mean square error in measuring the range, σ_R, is, naturally, larger than the potential mean square error, σ_{pot}. It is equal to

$$\sigma_R = \gamma_R\sigma_{pot}, \tag{2.28}$$

where

γ_R is the accuracy impairment factor for a real radar.

γ_R = 1.5 to 15, depending on the radar.

The factor γ_R can be represented in the form

$$\gamma_R = \sqrt{1 + \sigma^2_{prop}/\sigma^2_{pot} + \Sigma\sigma^2_i/\sigma^2_{pot}}\ , \tag{2.29}$$

where

σ_{prop} is the error caused by the curvature in the trajectory over which the radio waves are propagated;

σ_i is the error in the ith device in the radar.

The calculation of the error in σ_R can be made with sufficient accuracy even if it is taken that

$$\gamma_R = \sqrt{1 + \sigma^2_{ind}/\sigma^2_{pot}}, \tag{2.30}$$

where

σ_{ind} is the mean square error in the indicator.

2.15 Accuracy in Azimuth Measurement

Systematic errors in measuring the azimuth can occur when the radar antenna array is not oriented accurately and because of non-correspondence between the position of the antenna and the electrical scale of the azimuth scale.

Random errors in measuring target azimuth result from instability in the operation of the antenna rotation system, instability in the circuit that forms the azimuth markers, and from errors in reading.

The mean square error in measuring the azimuth is

$$\sigma_\beta = \sqrt{3}/\pi \cdot \varphi_{0.5}/\sqrt{2\nu_{v\ one}} \cdot \gamma_\beta, \tag{2.31}$$

where

γ_β is the accuracy impairment factor in determining the azimuth for a real radar; it is established through a formula similar to that used to establish γ_R;

$\varphi_{0.5}$ and σ_β are in degrees.

2.16 Accuracy in Elevation Measurement

The accuracy with which elevation is measured can be established by virtually the same factors as apply to the accuracy in measuring azimuth.

The mean square error in measuring the elevation can be assessed through a formula similar to the one at (2.31)

$$\sigma_\epsilon = \sqrt{3}/\pi \cdot \theta_{0.5}/\sqrt{2\nu_{v\ one}} \cdot \gamma_\epsilon, \tag{2.32}$$

where

$\theta_{0.5}$ and σ_ϵ are in degrees;

γ_ϵ is the accuracy impairment factor in the determination of the elevation by a real radar.

2.17 Accuracy in Height Measurement

In accordance with the formula at (1.76), the error in measuring the height is a combination of the error in measuring the range and the error in measuring the elevation, and can be established through the equation

$$\sigma_H = \sigma_R\left(\sin\epsilon + \frac{R}{R_e}\right) + \sigma_\epsilon R\cos\epsilon, \tag{2.33}$$

where

σ_H, σ_R are in meters;

σ_ϵ is in radians.

2.18 Information Capability of a Radar

Radars provide information on the coordinates of all targets within the scanning zone of the radar, on the characteristics of the targets, and on their accessories. Information from the radar can be evaluated qualitatively and quantitatively.

The quality of the information is the volume of data on the targets, on data accuracy and resolution.

Information can be assessed quantitatively by various criteria.

The information content I (in bits), can be established as follows, for example

$$I = nH(X) \text{ bits} \tag{2.34}$$

where

$n = V_{scan}/\delta V$ is the number of elements in the information;
V_{scan} is the volume of the scan zone for the radar;
δV is the resolution volume;
$H(X)$ is the entropy (see Chapter XVI);
X is the system of events.

Information here is understood to mean the possible number of resolution volumes, δV, in the radar's scan zone volume, V_{scan}.

The most widely disseminated system of events, X, in radar consists of two events, of equal probability: "there is a target," in the resolution volume, and "there is no target" in the resolution volume, or

$$X = \begin{pmatrix} x_1 & x_2 \\ p_1 & p_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \\ 0{,}5 & 0{,}5 \end{pmatrix}$$

where

p_1 is the probability of event x_1;
p_2 is the probability of event x_2.

In this case the entropy is

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i = (0{,}5 \log_2 0{,}5 + 0{,}5 \log_2 0{,}5) = 1 \text{ bit/element},$$

and

$$I = n \tag{2.35}$$

that is, the information content is equal to the number of elements, its components.

The technical informational capacity (the information content) of a radar can be evaluated through the formula

$$I_{tech} = (R_{max} - R_{min})\Delta\theta \; \Delta\epsilon/\delta R \; \delta\theta \; \delta\epsilon, \tag{2.36}$$

where

R_{max}, R_{min} are the operating limits for the radar in range;

$\Delta\beta$, $\Delta\varepsilon$ are the scanning sectors for the radar in azimuth and elevation, respectively;

δR, $\delta\beta$, $\delta\varepsilon$ are the resolutions.

The rate (technical) information can be obtained from the radar is

$$C_{tech} = I_{tech}/T_{scan}, \qquad (2.37)$$

where

T_{scan} is the scan period.

The practice often is to evaluate the information capacity of a radar by the number of targets and the number of pieces of data on each target in fact received in unit time, that is, the real rate at which information is received. This latter factor is sometimes called the tactical information capacity and is established in the following manner

$$C = 60mn/\Delta t \text{ locations per minute} \qquad (2.38)$$

where

m is the number of targets being worked;

n is the number of pieces of data on the target;

Δt is the capacity for information presentation.

2.19 Radar Immunity to Jamming

A radar's immunity to jamming means its capacity to retain tactical and engineering characteristics in the face of the effects of various types of radio frequency interference.

The effect of the jamming appears in a reduction in the signal/noise energy ratio at the input to the radar receiver. The result is that the target can be detected with a given probability at shorter ranges, or cannot be detected at all.

The quantitative criteria for evaluating the immunity of a radar to jamming can be quite different.

It is customary to evaluate the immunity of the radar to jamming by its operating range in the face of the jamming.

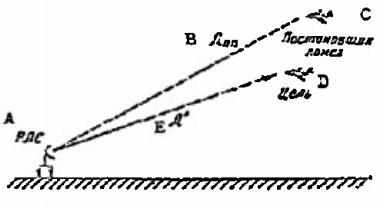


Figure 2.8. Schematic diagram of the location of the jammer and the target covered by the jammer.

A - radar; B - R_{jt}; C - jammer; D - target; E - R'.

The radar operating range, R', in the presence of noise jamming produced by the jamming transmitter does not coincide with the range to the victim (fig. 2.8), but equals

$$R' = R\sqrt[4]{16\pi^2 R_{jt}^2 P'_{lim\ min}/G_d g_{jam} g(\varphi_{jt}\theta_{jt})\Delta F\lambda^2}, \qquad (2.39)$$

where

R is the radar range when jamming is absent;

R_{jt} is the distance from the radar to the jamming transmitter;

$P'_{lim\ min}\Delta F$ are sensitivity limit and radar receiver bandwidth;

$g(\varphi_{jt}, \theta_{jt})$ is the radar antenna gain in the direction to the jamming transmitter;

G_d is the jammer transmitter noise power density, watts/hertz;

g_{jam} is the radar antenna gain for the jamming.

If the jamming transmitter is aligned with the victim

$$R' = R^2\sqrt{16\pi^2 P'_{lim\ min}/G_d g_{jam} g_{max}\Delta F\lambda^2}, \qquad (2.40)$$

where

g_{max} is the maximum gain for the radar antenna.

When jamming occurs, the range of the radar not immune to jamming can be reduced to the point where it can be considered as completely jammed, for all practical purposes. This is why a variety of measures are undertaken to improve the immunity to jamming of radars (see Chapter XI).

The literature on the subject points out that a radar can be made immune to jamming by:

operating the radar over a wide band and by rapid frequency changing;

multichannel radar construction;

high energy potential for the radar;

reducing the level of the side lobes in the antenna pattern;

by changing the pulse repetition frequency;

by controlling the polarization of the radiated signal;

by expanding the dynamic range of the receiver-indicator channel;

special types of modulation of the radiated oscillations;

special methods of processing the incoming signals;

the use of various devices and antijam circuits in the receiving and signal processing channel.

2.30 Climatic Conditions

The radar should be able to function in all types of climatic conditions. The permissible limits for temperature, humidity, pressure, and wind velocity in which the radar can retain its capacity to perform are usually given.

Temperature, humidity, and atmospheric pressure have a direct influence on the ability of the various units and devices in the radar to perform their functions.

2.21 Radar Engineering Data

Radar engineering data are the values of the magnitudes contained in the radar equation, as well as a number of other engineering characteristics of the principal devices in the radar. Principal among them are:

the operating wavelength, λ, or the wave band;

radiated power, P_{rad};

receiver sensitivity, $P_{rec\ min}$;

antenna gain, g_A;

widths of the radiation pattern in the horizontal, $\varphi_{0.5}$, and vertical, $\theta_{0.5}$, planes;

pulse repetition frequency, F_p;

antenna rotation rate, n_A;

the power required by the radar;

pulse length, τ_p;

receiver passband, ΔF.

2.22 Wavelength

The dimensions of the antenna array for the required values for the width of the antenna radiation pattern, and the directive gain, depend on the radar operating wavelength selected.

Taken into consideration when selecting the wavelength are the possibilities of obtaining the necessary power from the transmitter, and providing the required receiver sensitivity. Taken into consideration as well as the absorbtion and scattering features of weather conditions (clouds, rain, snow) and of the atmosphere (oxygen and water vapor).

The equation at (1.48) can be written in the form

$$R_{max} = \sqrt[4]{P_p S_A^2 \sigma_t / 4\pi\lambda^2 P_{rec\ min}} \qquad (2.41)$$

The range of a specific radar, and the magnitudes of P_p, S_A, $P_{rec\ min}$, which are given, increases with a decrease in the wavelength because $R \sim \sqrt{1/\lambda}$.*

This dependence of R on λ can be explained by the improvement in the directional properties of an antenna with specified geometric dimensions with reduction in λ.

---

* It is taken that σ_t = constant, regardless of λ.

Usually measured when tuning a radar is the frequency of the radiated oscillations, f, which can be found for the wavelength through

$$\lambda = \frac{3 \cdot 10^4}{f}, \tag{2.42}$$

where

λ is in centimeters;

f is in mHz.

Greater accuracy in determination of coordinates and resolution can be realized more readily at the shorter wavelengths in the radar range.

2.23 Transmitter Power

The power radiated by a radar, P_{rad}, is practically always characterized by the transmitter pulse power, P_p.

These powers are associated by the relationship

$$P_{rad} = \eta_f P_p, \tag{2.43}$$

where

η_f is the efficiency of the antenna feeder system.

Transmitter pulse power is understood to mean the average power delivered to the feeder system by the transmitter while the pulse lasts.

Pulse power and average transmitter power over the pulse sequence period, P_{av}, are associated by the relationship

$$P_p = P_{av} / \tau_p F_p, \tag{2.44}$$

where

τ_p is the pulse length, in seconds;

F_p is the pulse sequence frequency in Hz.

From (2.44)

$$P_{av} = \tau_p F_p P_p. \tag{2.45}$$

Transmitter energy is

$$W_p = P_p \tau_p = P_{av} T_p. \tag{2.46}$$

Radar range is $R \sim \sqrt[4]{W_p}$, that is, it is primarily determined by transmitter energy.

An increase in range is obtained by increasing the transmitter pulse power for a given pulse length, and conversely for a specified pulse power the radar range will increase with lengthening of the pulse. It is precisely this latter variant that can be realized in the radar using intrapulse frequency modulation and intrapulse phase shift keying.

2.24 Receiver Sensitivity

Real receiver sensitivity, $P_{rec\ min}$, is the minimum input signal power at which it is still possible to receive and detect reflected signals with a given probability.

Signal detection takes place against the background of the noise inherent in the receiver, so

$$P_{rec\ min} = \nu_v P_{n\ O}, \tag{2.47}$$

where

$P_{n\ O}$ is the noise power in the receiver passband at temperature T_O (see Chapter VIII);

ν_v is the visibility factor (see the formula at 1.47).

In the equation at (2.41), the receiver sensitivity is expressed in watts. The higher the receiver sensitivity (the lower $P_{rec\ min}$) the greater the radar range because $R \sim 1/\sqrt[4]{P_{rec\ min}}$.

Receiver sensitivity is often expressed in decibels

$$P_{rec\ min}\ [db] = 10 \log P_{rl}/P_{rec\ min}. \tag{2.48}$$

The reading level is taken as $P_{rl} = 10^{-5}$ watt, or $P_{rl} = 10^{-3}$ watt = = 1 milliwatt. For example, $P_{rec\ min} = 10^{-14}$ watt corresponds to 90 db with respect to $P_{rl} = 10^{-5}$ watt or 110 db with respect to P_{rl} = 1 milliwatt. In the watter case it is said that the sensitivity is expressed in decibel-milliwatts.

2.25 Antenna Gain

The antenna gain characterizes the directional properties of the antenna array (see Chapter IV).

The antenna gain, g_A, is, for all practical purposes, equal to the antenna's directive gain, G, because the radar antenna efficiency is extremely high ($\eta_A \approx 1$). This is why the antenna's directive gain is often used instead of the gain. The directive gain is associated with the capture area of the antenna, S_A, by the relationship

$$G = 4\pi S_A/\lambda^2. \tag{2.49}$$

With this relationship in mind, the equation at (2.41) can be written in the form

$$R_{max} = P_p G^2 G_t \lambda^2 / 64\pi^3 P_{rec\ min}. \tag{2.50}$$

There is a significant dependence of radar range for a specified working wavelength, that is, when λ = constant, on the directive gain because $R \sim \sqrt[4]{G}$. Today, antenna size is often increased in order to increase radar range.

2.26 Pulse Repetition Frequency

The maximum repetition frequency, F_p, of the main pulses should satisfy the following condition in order to have unambiguous determination of targets at specified ranges

$$F_{p\ max} \leqslant c/2R_{max}k_a, \tag{2.51}$$

where

c is the rate of propagation of radio waves;

k_a is the assurance factor, equal to 1.15 to 1.25.

Consequently

$$F_{p\ max} \leqslant 1.3 \cdot 10^5/R_{max}, \tag{2.52}$$

where

R_{max} is in kilometers.

At the same time, the pulse repetition frequency should be such that for a specified space scan rate the number of pulses, N_p, illuminating the target will be adequate to detect the target with a specified probability. With these considerations in mind, it is customary to provide

$$F_{p\ min} \geqslant N_{p\ min}\Delta\theta\ \Delta\epsilon/T_{scan}\varphi_{0.5}\theta_{0.5}. \tag{2.53}$$

For surveillance radars, and with the propagation time for electromagnetic energy to and from the target taken into consideration,

$$F_{p\ min} \geqslant \frac{N_p}{\dfrac{\varphi_{0.5}}{6n_A} - \dfrac{R_{max}}{1.5 \cdot 10^5}}, \tag{2.54}$$

where

n_A is antenna rpm;

$\varphi_{0.5}$ is in degrees;

R_{max} is in kilometers.

There is no need to observe the condition at (2.51) if the radar is fitted out to eliminate ambiguity in target determination. In such case the pulse repetition frequency can be three to five times higher.

The pulse repetition frequency establishes the number of pulses in the train for a given antenna radiation pattern width, and, as a result, the effect on the visibility factor for the receiver-indicator channel; that is, on the observability of the signal on the scopes.

2.27 Antenna Rotation Rate

With the conditions stipulated by the equation at (2.6) in mind, the relationship at (2.5) can be written in the form

$$n_A \leqslant 60F_p\varphi_{0.5}/N_{p\ min}\Delta\beta. \qquad (2.55)$$

Antenna rotations, n_A, should not exceed a predetermined number for specified values of $\varphi_{0.5}$, F_p, and scanning sector $\Delta\beta$, otherwise it can be difficult to observe the reflected signals because of the low $N_{p\ min}$.

For radars with circular scan, when $\Delta\beta = 360°$,

$$n_A \leqslant F_p\varphi_{0.5}/6N_{p\ min}. \qquad (2.56)$$

Other of the engineering characteristics of radars will be reviewed in subsequent chapters.

Chapter III

Radio Wave Propagation

3.1 The Spectrum of Electromagnetic Oscillations

Electromagnetic oscillations cover the wave band from 10^{-11} to $3 \cdot 10^{10}$ cm. Included are radio waves, infrared rays, visible light, ultraviolet rays, X-rays, and gamma rays. In the spectrum of electromagnetic oscillations, radio waves occupy the band of wavelengths from 100 km to 0.3 mm, or frequencies from 3 kHz to 10^6 mHz.

The USSR divides radio waves into the bands listed in Table 3.1.

Table 3.1

Classifications of radio wave bands adopted in the USSR

Name of band	Ultralong waves	Long waves	Medium waves	Short waves	Ultrashort waves
Wavelengths, meters	100000 - 10000	10000 - 1000	1000 - 100	100 - 10	10 - 0.0003
Frequency, mHz	$3 \cdot 10^{-3}$ - $3 \cdot 10^{-2}$	$3 \cdot 10^{-2}$ - $3 \cdot 10^{-1}$	$3 \cdot 10^{-1}$ - 3	3 - $3 \cdot 10$	$3 \cdot 10$ - 10^6

The ultrashort wave band is divided into the sub-bands indicated in Table 3.2.

The United States and England have adopted the classifications and designations for radio wave bands listed in Table 3.3.

Table 3.2

Classifications of the ultrashort wave band adopted in the USSR

Name of sub-band	Meter waves	Decimeter waves	Centimeter waves	Millimeter waves
Wavelengths, meters	10-1	1 - 0.1	0.1 - 0.01	0.01 - 0.0003
Frequency, mHz	$3 \cdot 10$ - $3 \cdot 10^2$	$3 \cdot 10^2$ - $3 \cdot 10^3$	$3 \cdot 10^3$ - $3 \cdot 10^4$	$3 \cdot 10^4$ - 10^6

Table 3.3

Classifications of ultrashort wave bands adopted in the United States and England

Designations adopted		Frequency band, gigahertz	λ, cm
in the United States	in England		
P	P	0,225—0,390	133,5—76,90
Lp — Lz	L	0,390—1,550	76,90—19,30
Sa — Sh	S	1,550—3,90	19,3—7,69
C	C	3,90—6,20	7,69—4,84
Xd — Xs	X	6,20—10,90	4,84—2,75
Kp — Kz	J	10,90—17,25	2,75—1,74
Kt—Kl	K	17,25—33,0	1,74—0,91
Ka — Qc	Q	33,0—46,00	0,91—0,65
Va — Vc	V	46,0—56,0	0,65—0,54

3.2 Electromagnetic Field Characteristics

The speed of free propagation of radio waves in an unbounded medium with a dielectric constant ε and a permeability μ can be established through the formula

$$v = \frac{1}{\sqrt{\varepsilon\mu}}. \tag{3.1}$$

This speed is equal to the speed of light, c, in free space (in a vacuum). For free space

$$\left.\begin{aligned} \varepsilon_0 &= \frac{1}{36\pi 10^9} \quad [\text{farad/meter}] \\ \text{and} \qquad \mu_0 &= 4\pi 10^{-7} \quad [\text{farad/meter}] \end{aligned}\right\} \tag{3.2}$$

The electromagnetic field can be characterized by the intensities of the electrical (E, volts/meter) and magnetic (H, amp/meter) fields.

The energies moved by the electromagnetic wave are concentrated in part (W_E) in the electrical, and in part (W_H) in the magnetic fields. The volume densities of these energies are

$$W_E = \frac{\varepsilon E^2}{2} \quad [\text{joules/m}^3] \tag{3.3}$$

$$W_H = \frac{\mu H^2}{2} \quad [\text{joules/m}^3] \tag{3.4}$$

The intensity at which the energy is moved can be characterized by the power flow density (Π, watts/m^2), that is, by the flow of energy over unit surface per second. The direction in which the energy is moved can be determined by the Umov-Poynting vector

$$\overline{\Pi} = [\overline{E}\overline{H}]. \tag{3.5}$$

In the electromagnetic field created by an isotropic radiator (that is, by a radiator, the field power flow density of which is the same in all directions) the magnitude of the power flow density, averaged over the period of oscillations, equals

$$\Pi_{av} = \frac{P_\Sigma}{4\pi R^2}, \tag{3.6}$$

where

P_Σ is the power radiated by the source;

R is the distance.

The wave front is the surface, all points on which have the same phase, that is, the surface at each point on which radio waves radiated by the antenna in different directions arrive at the same time. In the case of a point source the wave front is spherical, and in the case of a linear source the wave front is cylindrical. A small section of the wave front is a plane perpendicular to the direction of propagation at long distances from the source. The wave is called a plane wave in this case.

The amplitude of the field in a plane wave propagated without attenuation does not change with distance. In the spherical wave it changes in proportion to 1/R, and in the cylindrical wave it changes in proportion to $1/\sqrt{R}$.

The ratio of field amplitudes in a plane wave propagated in any medium is constant and is called the wave impedance of the medium

$$\rho = \frac{E}{H} = \sqrt{\frac{\mu}{\varepsilon}}. \tag{3.7}$$

The wave impedance of free space equals

$$\rho_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 120\pi \approx 377 \text{ ohms} \tag{3.8}$$

The average power flow density in free space is

$$\Pi_{av} = \frac{EH}{2} = \frac{E^2}{240\pi} = 60\pi H^2, \tag{3.9}$$

where

E and H are the amplitude values.

The rate at which the wave is dephased (φ) with distance

$$\alpha = \frac{d\varphi}{dR} = \frac{2\pi}{\lambda} \tag{3.10}$$

is called the phase factor, or phase constant.

In an absorbing medium the amplitude of the plane wave field diminishes

$$E = E_0 10^{-\beta D_f/20}, \tag{3.11}$$

or

$$E = E_0 e^{-\beta' D_f}, \tag{3.12}$$

where

β' is the attenuation constant in nepers/kilometer;

$\beta = \beta'/0.115$ is the attenuation constant in db/kilometer;

E_0 is the amplitude of the field at the input to the absorbing medium;

E is the amplitude of the field at distance D_f from the input to the absorbing medium.

Region profoundly effecting the propagation of radio waves

The propagation of radio waves from their source to the point of observation is within a definite region, in the main. This region is an ellipsoid encompassing several first Fresnel zones (fig. 3.1). The Fresnel zones are sections of the wave front of a size such that the phase of a wave moving from the source through the initial zone differs from the phase of a wave moving through the edge of this zone by 180°.

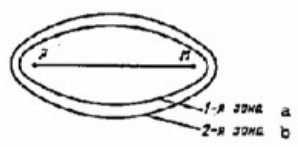


Figure 3.1. Region profoundly effecting the propagation of radio waves.

a - 1st zone; b - 2nd zone.

The first Fresnel zone is a circle, the center of which lies on the line between the source and the point of observation, and is the origin of the zone. The second, and subsequent, zones are annular belts, each of which encompasses the preceding zone. The phases of the fields at the boundaries of neighboring zones differ by 180°.

3.3 The Influence of the Ground on the Propagation of Radio Waves

The ground effects the propagation of radio waves because some of the energy from the radiation source reaches the ground and is reflected from

it. This results in two waves, the direct wave and the reflected wave, appearing at the observation site (fig. 3.2). Some of the energy is absorbed by the earth.

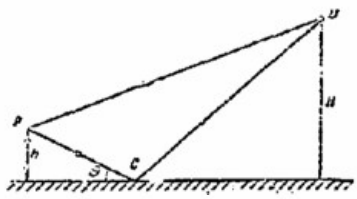


Figure 3.2. Interference of the direct and reflected waves.

Reflection coefficients

The complex amplitude of the field of the reflected wave ($\dot{E}_{ref}$) is determined by the amplitude of the field of the incident wave ($\dot{E}_{inc}$), by multiplying the latter by some factor, called the reflection coefficient ($\dot{R}_{ref}$). In accordance with the determination

$$\dot{R}_{ref} = \dot{E}_{ref}/\dot{E}_{inc} = \bar{R}_{ref} \; e^{j\beta_{ref}} \tag{3.13}$$

where

β_{ref} is the change in the phase of the wave upon reflection, or the phase of the reflection coefficient.

The modulus and the phase of the reflection coefficient depend on the polarization of the incident wave and the electrical parameters of the soil, on the dielectric constant, and on the conductivity σ.

For horizontal and vertical polarizations, the reflection coefficients can be established through the formulas

$$\dot{R}_h = \frac{\sin\Theta - \sqrt{\varepsilon' - \cos^2\Theta}}{\sin\Theta + \sqrt{\varepsilon' - \cos^2\Theta}} = |\dot{R}_h| e^{j\beta_h}, \tag{3.14}$$

$$\dot{R}_v = \frac{\varepsilon'\sin\Theta - \sqrt{\varepsilon' - \cos^2\Theta}}{\varepsilon'\sin\Theta + \sqrt{\varepsilon' - \cos^2\Theta}} = |\dot{R}_v| e^{j\beta_v}, \tag{3.15}$$

where

ξ is the angle of slip of the radio beam (fig. 3.2);

$\varepsilon' = \frac{\varepsilon}{\varepsilon_0} = \varepsilon_r - j\eta$ is the relative dielectric constant for the soil;

$\eta = 60\lambda\sigma$ is the imaginary part of ε';

ε_r is the real part of ε'.

The moduli and phases of the reflection coefficients are found by using the curves (figs. 3.3 - 3.8) plotted through the use of the formulas at (3.14) and (3.15). A parameter common to all curves is the magnitude η, the value of which is plotted on the sloping scales individually for the vertical and horizontal polarizations.

The electrical parameters of selected soils and water are listed in Table 3.4.

Table 3.4

Electrical parameters of selected soils and of water

In the meter wave band

Soil, water	ε_r		σ, mho/m	
	from	to	from	to
Sea water	80	-	0.66	6.6
Wet ground	5	20	10^{-3}	10^{-2}
Dry ground	2	6	10^{-4}	$4 \cdot 10^{-3}$
Fresh water	80	-	10^{-3}	$5 \cdot 10^{-3}$

In the centimeter wave band

Soil, water	λ, cm	ε_r	σ, mho-m
Sea water, 20-25°C	10	69	6.5
Dry, sandy	9	2	0.03
Wet sandy	9	24	0.6
Fresh water, +20°C	10	79	2.06
	3	64	18.4

Example. Determine the modulus and phase of the reflection coefficient when polarization is vertical and $\Theta = 1°$. A radar with $\lambda = 3$ meters is located on soil that can be categorized as wet ground.

From Table 3.4 we use $\varepsilon_r = 10$; $\sigma = 10^{-2}$ mho/m. We determine the imaginary part of the dielectric constant to be

$$\eta = 60\lambda\sigma = 60 \cdot 3 \cdot 10^{-2} = 1.8.$$

We find the solid line for $\eta = 1$ on the curve (fig. 3.5) along the left sloping scale of values. Let us now draw a line parallel to this line and corresponding to $\eta = 2$ to its point of intersection with the vertical line for $\Theta = 1°$. Let us now take the point of intersection to the right to the axis of the value for R_v where we find $R_v = 0.88$. The point of intersection of the line for $\eta = 2$ lies below the base of the curve on the curve for phases of the reflection coefficient (fig. 3.6). We will therefore take $\beta_v = 180°$ as an approximation.

The interference coefficient (the ground coefficient)

The field at any point above the ground can be obtained as a result of the superimposition (interference) of the field of the incident wave and the field of the wave reflected from the ground. The complex amplitude of

the resultant field ($\bar{E}_{res}$) is determined by the amplitude of the field of the incident wave by multiplying the latter by a coefficient called the interference coefficient, or the "ground coefficient" [$\dot{\phi}(\Theta)$]. The interference coefficient in the case of horizontal [$\dot{\phi}_h(\Theta)$] and vertical [$\dot{\phi}_v(\Theta)$] polarizations can be established through the formula

$$\Phi_{h,v}(\Theta)=\sqrt{1+|\dot{R}_{h,v}|^2+2|\dot{R}_{h,v}|\cos\left(\frac{4\pi h_A}{\lambda}\sin\Theta+\vartheta_{h,v}\right)}, \qquad (3.16)$$

where

h_A is the height of the source of radiation above the surface of the ground.

When the soil can be considered to be a good conductor ($\eta \gg \varepsilon_r$), the ground coefficients equal

$$\Phi_h(\Theta)=2\sin\left(\frac{2\pi h_A}{\lambda}\sin\Theta\right), \qquad (3.17)$$

$$\Phi_v(\Theta)=2\cos\left(\frac{2\pi h_A}{\lambda}\sin\Theta\right). \qquad (3.18)$$

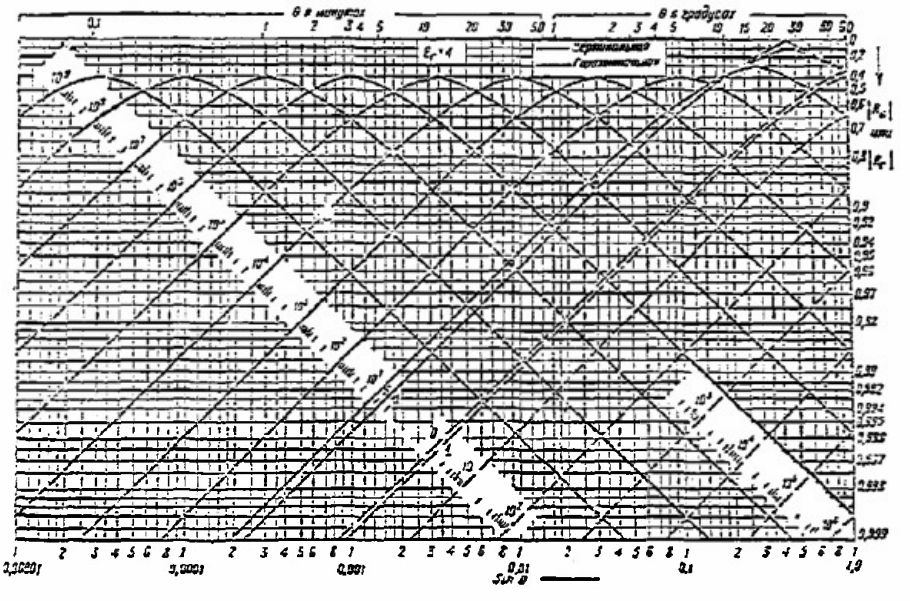


Figure 3.3. Graphs for the moduli of the reflection coefficients as functions of the angle of slip for various values for the parameter $\eta = 60\lambda\sigma$ when $\varepsilon_r = 4$.

——— vertical
- - - - horizontal

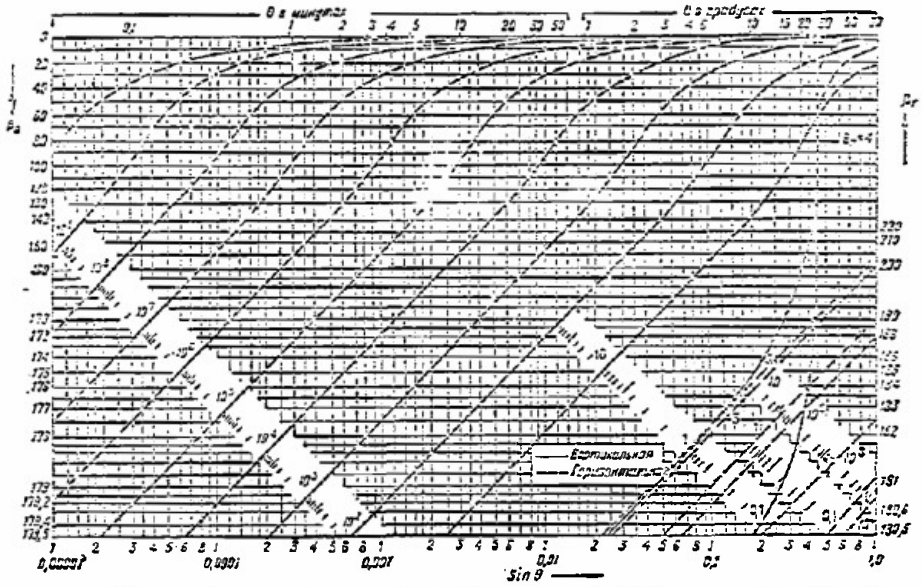


Figure 3.4. Graphs for the arguments for the reflection coefficient as functions of the angle of slip for various values for the parameter η when $\epsilon_r = 4$.

——— vertical
- - - - horizontal

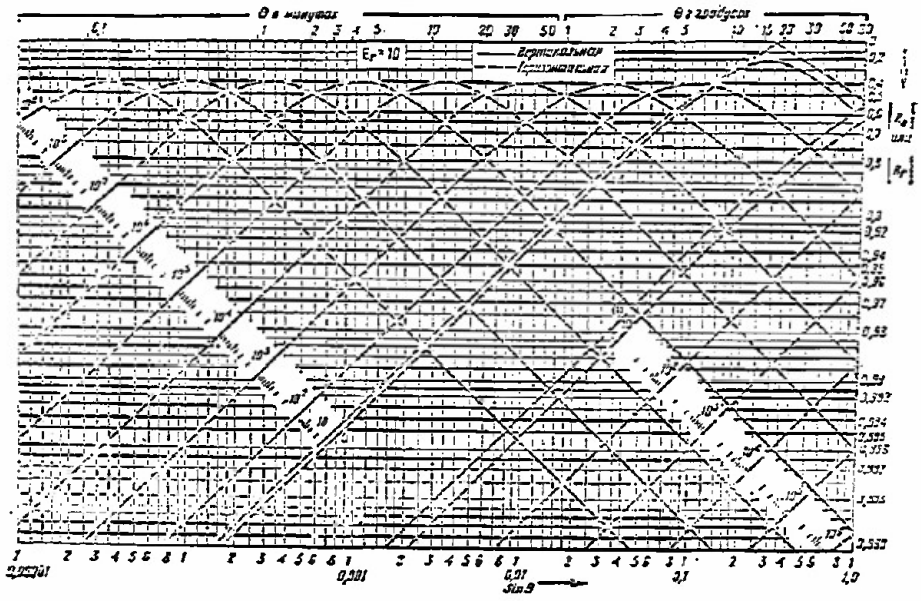


Figure 3.5. Graphs for the moduli of the reflection coefficients as functions of the angle of slip for various values for the parameter η when $\epsilon_r = 10$.

——— vertical
- - - - horizontal

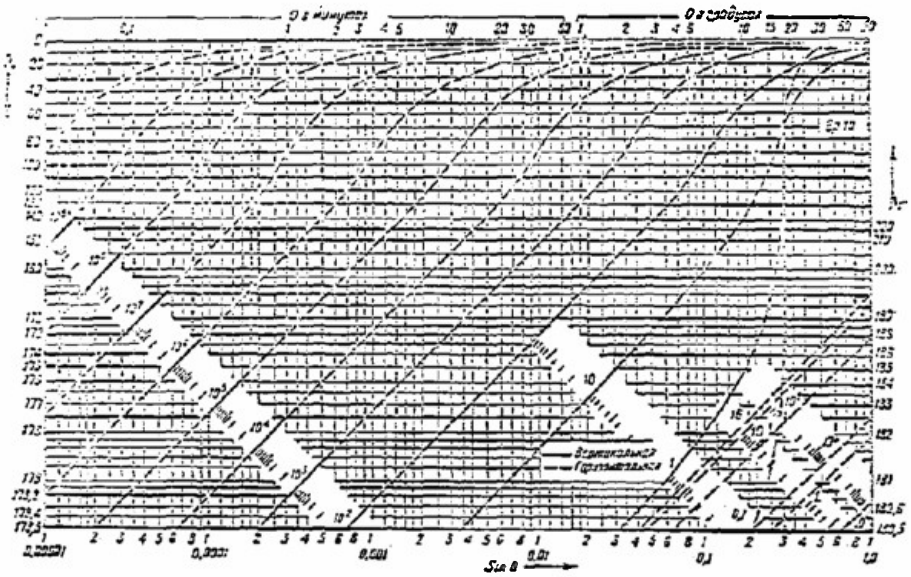


Figure 3.6. Graphs for the arguments for the reflection coefficients as functions of the angle of slip for various values for the parameter η when $\epsilon_r = 10$.

——— vertical
- - - - horizontal

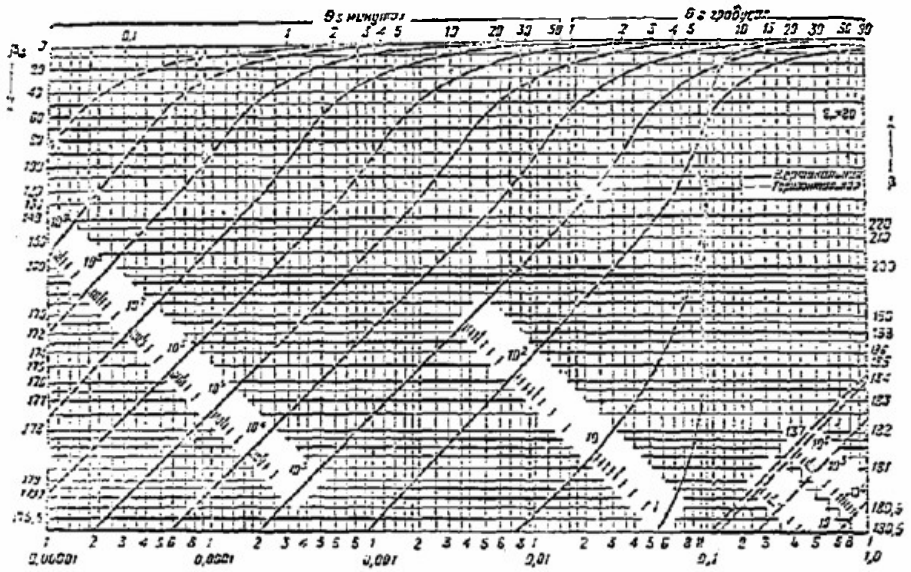


Figure 3.7. Graphs for the arguments for the reflection coefficients as functions of the angle of slip for various values for the parameter η when $\epsilon_r = 80$.

——— vertical
- - - - horizontal

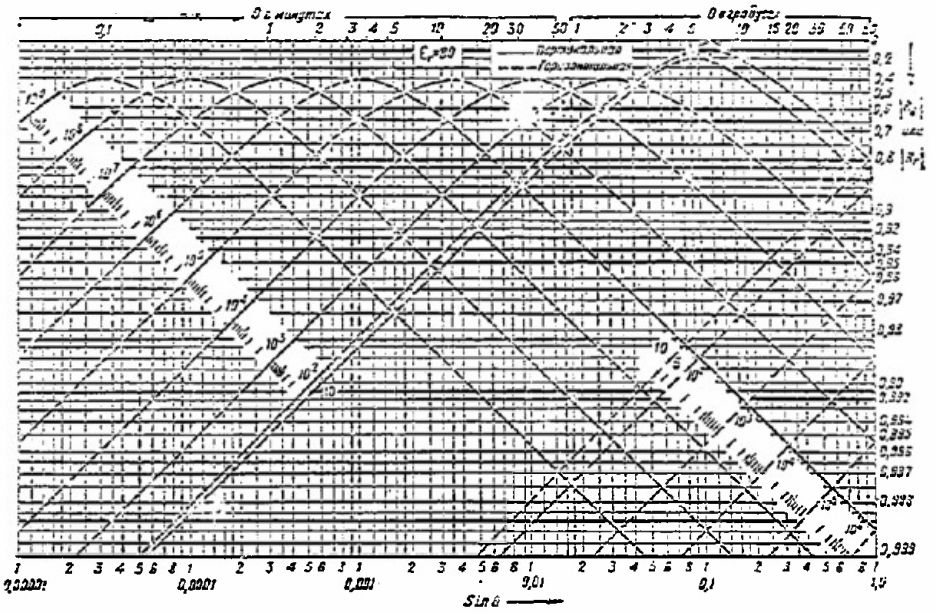


Figure 3.8. Graphs for the moduli of the reflection coefficients as functions of the angle of slip for various values for the parameter η when $\epsilon_r = 80$.

——— vertical
- - - - horizontal

Region profoundly effecting the reflection

Since there is a region in space that profoundly effects the propagation of radio waves, there is also a region on the earth's surface that profoundly effects reflection. This region is a section of the surface of the earth of the ellipsoid encompassing at least the first Fresnel zone (fig. 3.9). Therefore, the region profoundly effecting the reflection is an ellipse that at least contains the first Fresnel zone.

When the location of air targets is such that the antenna height (h_A) is considerably less than target height, the position and the dimensions of this ellipse can be established through the following approximate relationships:

distance from the antenna to the center of the ellipse

$$X_{01} \approx 12 \frac{h_A^2}{\lambda}; \tag{3.19}$$

length of the major axis of the ellipse

$$a_1 \approx 11{,}3 \frac{h_A^2}{\lambda}; \tag{3.20}$$

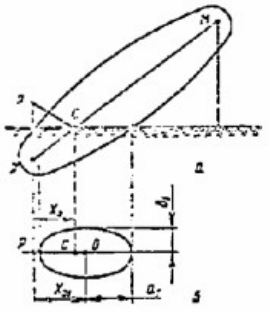


Figure 3.9. Region profoundly effecting the reflection of radio waves from the earth's surface.

a - in the vertical plane; b - in the horizontal plane.

length of the minor axis of the ellipse

$$b_1 \approx 2{,}82 h_A. \tag{3.21}$$

The nature of the reflection should be specular in the region profoundly effecting the reflection. This means that the permissible irregularities, Δh_i, in the reflecting surface should meet the condition

$$\Delta h_i < \frac{\lambda}{16 \sin \Theta} \tag{3.22}$$

The radius of the area around the radar in the case of specular reflection is

$$R_{area} = X_{01} + a_1 \approx 23{,}3 \frac{h_A^2}{\lambda} \tag{3.23}$$

Example. Determine the dimensions and permissible irregularity in the area around an antenna with height h_A = 6 meters. The radar is operating in the circular scan mode on a wavelength of λ = 2 meters.

Solution

$$X_{01} \approx 12 \frac{h_A^2}{\lambda} = 12 \frac{6^2}{2} = 216 \text{ meters,}$$

$$a_1 \approx 11{,}3 \frac{h_A^2}{\lambda} = 11{,}3 \frac{6^2}{2} = 202 \text{ meters,}$$

$$b_1 = 2{,}82 h_A = 2{,}82 \cdot 6 = 16{,}8 \text{ meters.}$$

The radius of the area around the antenna is

$$R_{area} = X_{01} + a_1 = 216 + 202 = 418 \text{ meters.}$$

The sizes of the permissible irregularities are determined at the origin of the area ($\Theta = \Theta_o$) and in its center ($\Theta = \Theta_c$).

In this case

$$\sin\Theta_c = \frac{h_A}{\sqrt{(X_{01}-a_1)^2 + h_A^2}} = \frac{6}{15{,}2} \approx 0{,}4,$$

$$\sin\Theta_c \approx \operatorname{tg}\Theta_c = \frac{h_A}{X_{01}} = \frac{\lambda}{12h_A}.$$

The permissible irregularities at the origin of the area are

$$\Delta h_{i0} < \frac{\lambda}{16\sin\Theta_0} = \frac{2}{16\cdot 0{,}4} \approx 0{,}3 \text{ meter.}$$

The permissible irregularities in the center of the area are

$$\Delta h_{ic} < \frac{\lambda}{16\sin\Theta_c} = \frac{3h_A}{4} \approx 4{,}5 \text{ meters.}$$

Reduced heights. Line of sight.

So-called reduced heights (h'_A and H') must be used in the calculations that utilize the interference formulas, for these take into consideration the sphericity of the earth. They are computed through the following formulas

$$h'_A = h_A - \frac{D^2}{2R_e}\left(\frac{h_A}{h_A + H}\right)^2, \tag{3.24}$$

$$H' = H - \frac{D^2}{2R_e}\left(\frac{H}{h_A + H}\right)^2, \tag{3.25}$$

where

D is the distance between the observation and radiation points;

R_c = 6370 kilometers, and is the radius of the earth.

The sphericity of the earth is what causes a shadow region below the line of the horizon (fig. 3.10), as a result of which location is only possible from the boundary of that region when line of sight to the target is established. Line of sight (D'_s) is established through the formula

$$D'_s = \sqrt{2R_e}\left(\sqrt{h_A} + \sqrt{H}\right). \tag{3.26}$$

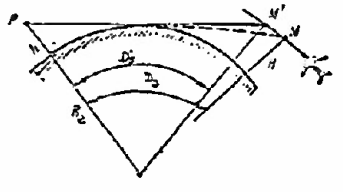


Figure 3.10. Target line of sight. D'_s - geometric; D_s - with refraction of radio waves taken into consideration.

3.4 The Influence of the Troposphere on the Propagation of Radio Waves

The troposphere is the lower boundary of the atmosphere, extending to heights of 10 to 12 kilometers. It contains air gases, water vapors, and hydrometeors. The electrical parameters of the troposphere are determined by its meteorological condition and atmospheric processes, and these depend on how the troposphere is heated by the earth, and on weather conditions. The result is that the following effects occur in different ways in the troposphere:

refraction of radio waves;

scattering of the radio waves by discontinuities;

absorption of radio wave energy;

attenuation of radio waves by hydrometeors.

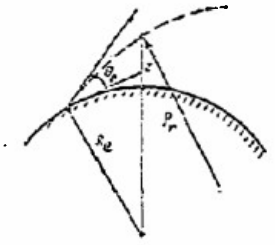


Figure 3.11. Refraction of radio waves in the troposphere.

Refraction is the curvature of the path over which radio waves are propagated.

Tropospheric refraction is the result of a smooth change in the index of refraction ($n = \sqrt{\epsilon'}$) of the troposphere with height (z). The degree of curvature of the radio beam (fig. 3.11) can be evaluated by the radius of curvature

$$\rho_r = -\frac{1}{\frac{dn}{dz}\cos\theta_0} = kR_e, \qquad (3.27)$$

where

Θ_0 is the elevation of the initial direction of the radio beam;

k is the curvature coefficient for the radio beam;

dn/dz is the rate of change (gradient) in the index of refraction with height.

The troposphere is said to be normal when its state is such that the absolute temperature (T) and pressure of the water vapor (e) decline with height in accordance with a linear law

$$T^\circ = 288 - 0.0065z,$$
$$e\ [\text{millibars}] = 10 - 0.0035z.$$

In the normal troposphere the gradient of the index of refraction equals

$$\frac{dn}{dz} = -4\cdot 10^{-8}\ \frac{1}{м}.$$

Refraction is called standard in this case. Th least radius of curvature in the beam will be during radiation along the horizon ($\Theta_0 = 0$). In the standard refraction case $\rho \approx 4R_e$ ($k = 4$).

The different cases of refraction in the troposphere are shown in Figure 3.12.

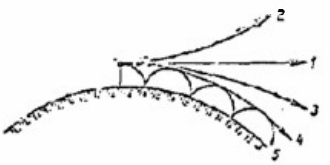


Figure 3.12. Different cases of refraction in the troposphere when $\Theta_0 = 0$.
1 - no refraction ($k = \infty$); 2 - negative refraction ($k < 0$); 3 - standard refraction ($k = 4$); 4 - increased refraction ($k = 1$); 5 - superrefraction ($k < 1$).

The radio beam bends downward with an increase in the index of refraction with height ($dn/dz > 0$) and negative refraction ($k < 0$) occurs. A decrease, or an increase, in the refraction occurs with less, or more, of a change in the index of refraction with height with respect to standard refraction.

The case when the coefficient of curvature is $k < 1$ is called superrefraction, or an atmospheric duct. The radio beam is then propagated in the layer near the ground and is reflected from the surface of the earth many times.

Negative refraction is usually encountered in the winter time, during snow storms, and in the polar regions over the sea. It reduces line of sight.

The atmospheric duct is formed by the layer of warm air found above a layer of cooler, moist air. It is usually encountered over the sea and during the morning hours in the summer time over land in windless eather.

Positive refraction ($k > 0$) causes the radio beam to enter the shadow region (fig. 3.10), increasing line of sight. The effect of refraction on line of sight can be taken into consideration by replacing the radius of the earth by a radius (R_e) equivalent to it and established through the relationship

$$1/R_e = 1/R_e - 1/\rho_r. \qquad (3.28)$$

In the standard refraction case

$$R_e = 4/3\ R_e \approx 8500 \text{ km}.$$

And the line of sight can be established through the formula at (2.1).

Scattering of radio waves by the discontinuities in the troposphere is caused by the continuous, and disordered, movements of the air masses with

different electrical parameters. Scattered by the discontinuities, the radio waves also enter the shadow region (fig. 3.13). The amplitude of the scattered field in the shadow region is much greater than that of the diffraction field. This phenomenon is what makes long-range tropospheric communications possible over distances up to 600 to 700 kilometers. These communications involve the use of special stations with powerful transmitters and pencil-beam antennas aimed at each other.

Absorption of radio waves in the troposphere occurs in the oxygen and in the water vapor, and depends on the frequency (fig. 3.14). The absorption lines of the other gases are beyond the radio wave band. Radio wave scattering in the gases in the troposphere is slight.

Attenuation of radio waves by hydrometeors is the result of scattering and absorption of their energy. Figures 3.15 and 3.16 show the attenuation factors in rain and in fog with respect to their intensity. Attenuation in dry snow and in hail is slight. If the snow is wet, attenuation will be the same as that occurring in rain of the same intensity.

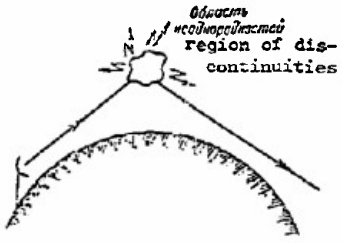


Figure 3.13. Scattering of radio waves by discontinuities in the troposphere.

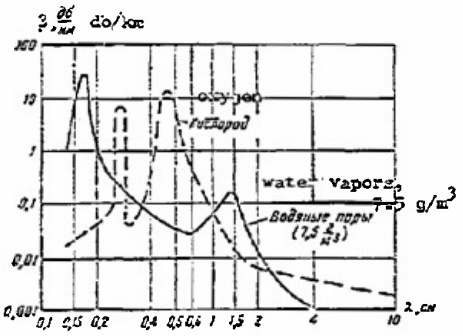


Figure 3.14. Curves of the dependence of the attenuation factor for radio waves on the wavelength for oxygen and water vapors.

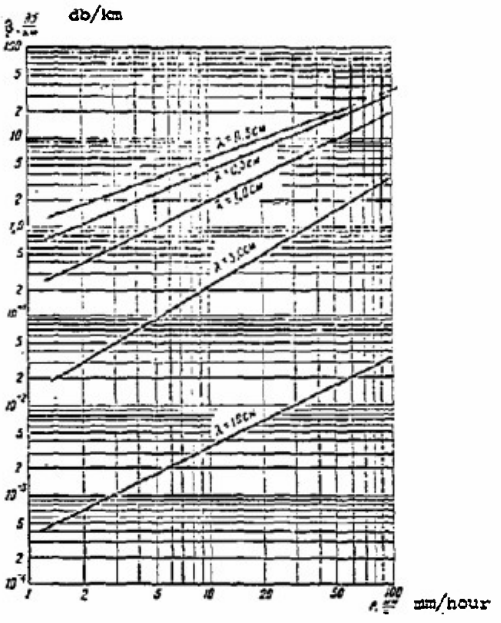


Figure 3.15. Attenuation factor as a function of the rain intensity for various wavelengths.

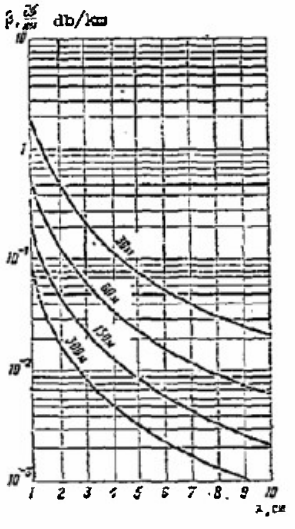


Figure 3.16. Attenuation factor in fog of different intensities in the centimeter wave band (figures on the curves are line of sight in fog).

The radar range (R) can be established through

$$R = R_0 10^{-0.5\beta R},$$

where

R_0 is the range when there is no absorption;

β is the attenuation factor in db/km,

with absorption of radio waves taken into consideration.

3.5 The Influence of the Ionosphere on the Propagation of Radio Waves

Extending above the troposphere at a height above 12 km and rising to 60 to 70 km is the stratosphere. Propagation of radio waves in this region takes place as it does in a vacuum. The upper layer of the earth's gas envelope, which is located above the 60 to 70 km level, is called the ionosphere. The ionosphere, according to present-day concepts, extends several thousands of kilometers from the earth.

The ionosphere consists of ionized gases with an admixture of neutral atoms and molecules.

The source of the ionization is solar radiation, particularly ultraviolet, X-ray and gamma radiation. The energy of these radiations diminishes with penetration into the depth of the atmosphere. Hence, the electron concentration (N) changes with height from slight in the lower layers of the ionosphere to some maximum (N_{max}) at some definite height (z_{max}). Rarification of the air causes a further reduction at even greater heights. This simplified model of the structure of the ionosphere (fig. 3.17) is called a simple layer.

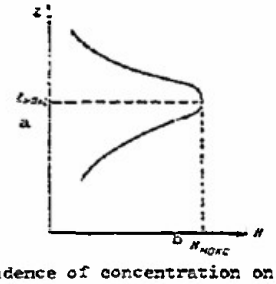


Figure 3.17. Dependence of concentration on height in a simple layer. a - z_{max}; b - N_{max}.

The electrical parameters of the ionosphere have the following values

$$\varepsilon = 1 - \frac{\omega_0^2}{\omega^2}, \tag{3.29}$$

$$\sigma = \varepsilon_0 \nu \frac{\omega_0^2}{\omega^2}, \tag{3.30}$$

where

ω_0 is the plasma frequency, that is, the natural frequency of free oscillations of the electrons in an ionized medium ($\omega_0 = 2\pi\sqrt{80.8N}$);

N is the electron concentration in electrons/m^3;

ν is the number of electron collisions in unit time.

The propagation of radio waves in a simple ionized layer

The index of refraction in the ionosphere

$$n = \sqrt{1 - \frac{80.8N}{f^2}} \tag{3.31}$$

(where f is the frequency of electromagnetic oscillations) can be less than 1, zero, and imaginary. Propagation of radio waves is impossible when its values are zero and imaginary. With N given any value there is, in the layer, a frequency at which radio waves will be reflected from the given layer. The highest of these frequencies

$$f_{crit} = \sqrt{80.8N_{max}}, \tag{3.32}$$

at which the radio wave striking the simple layer vertically can still be reflected from it is called the critical frequency. Radio waves with frequencies higher than f_{crit} pass on through the ionosphere.

If the radio wave strikes the ionosphere at an angle, Θ_0, the highest frequency

$$f_{max} = f_{crit}/\sin\Theta_0 \tag{3.33}$$

is called the maximum frequency.

An averaging of many years of data on the values of customary electron concentrations yields critical and maximum frequencies between 16 and 48 mHz. Radio waves at frequencies below these values will not pass through the ionosphere.

The actual mechanism involved in the formation of the ionosphere is more complicated than the mechanism involved in the formation of the simple layer. The ionosphere is made up of individual layers, known as the D, E, F_1, and F_2 layers (fig. 3.18). The electron concentrations in the layers, and the heights of the layers, change with time of day and time of year, as well as with the degree of solar activity. There is rapid recombination of the ionized atoms in the denser layers at night, the low (D) layer disappears, and the electron concentration in the rest of the layers decreases. The atmosphere becomes more dense in the winter time, the maximum for the concentration increases, and drops lower down. The F_1 layer disappears in the winter time.

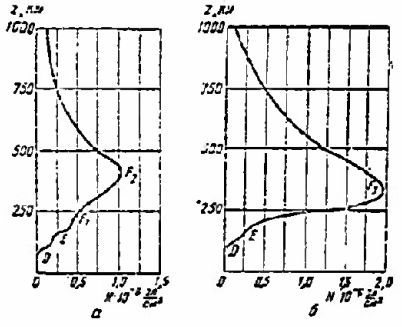


Figure 3.18. Average distribution of electron concentrations by height in the summer (a) and winter (b). Horizontal: electrons/cm^3.

The effect of the ionosphere on the propagation of waves in different bands

Ultralong waves and long waves are propagated between the ionosphere and the earth as if by a waveguide. The earth's surface guides the waves in this band within the earth envelope, causing them to propagate in the form of a so-called ground wave.

Long waves are propagated as ground, as well as sky, waves. The radio waves in the ground wave attenuate relatively quickly. The radio waves in the sky wave penetrate the D and E layers, where absorption is intensive, the waves are reflected, and they return to earth. These waves cover comparatively short distances in the daytime because of this absorption, but at night, when the D layer disappears, and when the concentrations of electrons in the E layer diminish, the range of transmission of long waves is much greater than during the day.

Short waves propagated by a sky wave are reflected repeatedly from the ionosphereand from the earth. For this reason the range of radio communications on these waves is very great.

The N. I. Kabanov effect

The Kabanov effect occurs in the short wave band, and simply stated says that radio waves reaching the earth's surface after having been reflected from the ionosphere can be scattered by the earth. Some of the scattered radiation returns to the radiation source, where it can be recorded. The return scattered signals can be received at ranges varying from several hundred to several thousand kilometers.

If the reflecting surface creates a direct reflection there will be no return scattered signal.

Ultrashort waves pass through the ionosphere, except for the long wave portion of this band. Absorption and scattering of their energies occurs, as does refraction, and change in propagation rate.

3.6 The Effect of the Ionized Regions of Atomic Explosions on the Propagation of Radio Waves (L5)

The ionized regions of atomic explosions have the properties of the ionosphere. The dimensions of these regions can be in the tens of kilometers, and the densities of the electron concentrations in them can have magnitudes close to those found in the ionosphere.

These regions can have the following effect on radar operation:

cause a bending in the path of the radio waves (fig. 3.19), resulting in an error in determining target height

$$\Delta H = h_{reg}\left(\frac{\cos\varphi}{\sqrt{\varepsilon' - \sin^2\varphi}} - 1\right), \qquad (3.34)$$

where

h_{reg} is the vertical dimension of the region;

φ is the angle of incidence of the beam at the region boundary;

cause a change to take place in the group signal velocity ($v_{group} = c\sqrt{\varepsilon'}$), in the ionized region, leading to errors in determining the range to the target, and equaling

$$\Delta R = 2L_{reg}\,(1 - 1/\sqrt{\varepsilon'}) \ , \qquad (3.35)$$

where

L_{reg} is the distance over which the ionized region extends;

cause some of the field energy to be absorbed, scattered, and reflected in the ionized region, and thus causing a reduction in radar operating range.

The scattered radiation from the ionized region can be picked up by the radar in the same way that scattered radiation from any target is picked up. This is the basis for the possibility of determining the coordinates of atomic explosions and for arriving at an approximate determination of their characteristics.

3.7 The Effective Reflecting Surface of a Target

The effective reflecting surface of a target is the area of nondirectional radiation that determines the magnitude of the reradiated power and creates the same flow of energy at the reception site as would the real target.

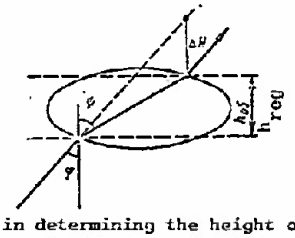


Figure 3.19. Error in determining the height of a target beyond the ionized region of an atomic explosion.

The effective reflecting surface depends on target size, shape, material, and radar wavelength. The directivity pattern of the secondary radiation from the target is badly cut up because of the complex configuration of the target. As the target moves the nature of this dissection changes constantly, the result of the random changes in target position and changes in the direction of illumination. It is this that causes fluctuations in the power of the signal reflected by the target. Hence, the effective reflecting surface of a target is a variable. The probability that σ_t will at least be equal to the selected magnitude (x) can be established using the graph shown in Figure 3.20.

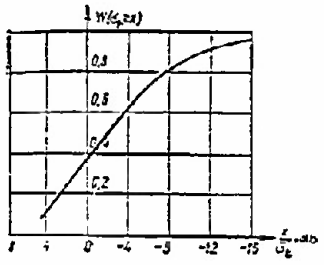


Figure 3.20. Graph of the probability that the effective reflecting surface of a target, σ_t, will at least be equal to the selected design magnitude, x.

Table 3.5 lists the average values for the effective reflecting surface ($\sigma_{t\ o}$) for selected targets.

Table 3.5

Average values for effective reflecting surface for selected targets

Type target	$\sigma_{t\ o}$, m^2
Fighter plane	3 to 5
Long-range bomber	20
Ballistic missile warhead	0.2
Transport aircraft	50
Man	0.8

3.8 Radar Coverage Pattern

The coverage pattern is that region of space at the boundary of which a target can be detected with specified probability. The boundary of the coverage pattern can be established through the radar equation

$$R = \sqrt[4]{P_p G^2_{max} \sigma_t \lambda^2 / 64\pi^3 P'_{rec\ min} v_v}\ 10^{-0.053R} \cdot F(\varphi, \epsilon), \qquad (3.36)$$

where $F(\varphi, \epsilon)$ is the standardized antenna pattern (see Chapter IV).

The random nature of the position of the boundary of the coverage zone is caused by the fuzzing up of the signal by noise and by fluctuations in the effective reflecting surface of the target. Hence, target detection is random in nature; that is, it can, or cannot, occur. This is taken into consideration in the equation at (3.36) by the visibility factor, ϑ_v. The necessary values of the latter providing for specified probabilities of correct target detection, (P_{cd}), and false alarm, (P_{fa}), at predetermined ranges, establish the boundary of the coverage pattern.

Radar coverage pattern in the vertical plane

The boundary of the coverage pattern is established by the resultant antenna pattern. Detection and target acquisition radars have separate antenna patterns in the horizontal and vertical planes, so the coverage patterns too are constructed in these planes. The coverage pattern in the vertical plane is taken with respect to the maximum in the antenna pattern in the horizontal plane, so its equation has the form

$$R(\epsilon) = R_{max}F(\epsilon), \tag{3.37}$$

where

$F(\epsilon)$ is the radar antenna pattern in the vertical plane.

Meter and decimeter band radars, in which the antenna pattern is formed by part of the energy reflected from the earth, have the equation at (3.37) in the form

$$R(\epsilon) = R_0F(\epsilon)\xi(\epsilon), \tag{3.38}$$

where

R_0 is the maximum radar operating range in free space;

$F(\epsilon)$ is the standardized antenna pattern in free space;

$\xi(\epsilon)$ is the multiplier for the earth (see the expressions at 3.18 - 3.18).

Figure 3.21 is an example of the coverage pattern for a radar, the antenna pattern of which is formed as a result of the energy reflected from the earth. The shape of the coverage pattern can change with the profile of the position and the nature of the soil. It is desirable to set up in each position with coverage patterns for characteristic azimuths.

The coverage pattern in the vertical plane (fig. 3.21a) is usually constructed in the rectangular system of coordinates, in terms of the height (H) and the slant range (R). Plotted on the coordinate grid are the elevation lines, the lines of reduced heights (H'), and the equal altitude curves with the sphericity of the earth taken into consideration. The reductions (ΔH) in the latter below the lines of reduced heights can be established through (3.25)

$$\Delta H \simeq R^2/2R_e \ . \tag{3.39}$$

The equal altitude curves are used in the evaluation of the combat capabilities of a radar against targets flying at different altitudes. The coverage patterns in the vertical plane can be dissimilar in different directions, depending on the nature of the radar position. With this in mind, the combat capabilities of the radar can be evaluated by the set of coverage patterns taken for characteristic azimuths.

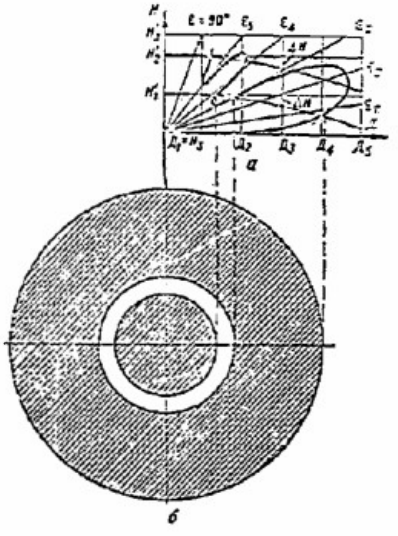


Figure 3.21. Example of coverage pattern in the vertical (a) and horizontal (b) planes for a radar on a level, homogeneous site.

Radar coverage pattern in the horizontal plane

The coverage pattern in the horizontal plane is a horizontal section at a definite height in the coverage patterns in the vertical plane. When the site is level and homogeneous the coverage pattern in the horizontal plane is a circle (fig. 3.21b) with internal, unexamined annular bands. They form at the site of the nulls in the antenna pattern.

When the position in which the radar is located is uneven, the coverage pattern in the horizontal plane is established from the set of coverage patterns in the vertical plane for characteristic azimuths, and with the screening effect of local objects at the given height taken into consideration.

Maximum operating range, range at a specific height, and antenna pattern for the site are made more precise by using data obtained under real radar

operating conditions. Overflights can be used when needed to make these refinements.

3.9 Propagation of Electromagnetic Waves in the Infrared Spectrum

Infrared rays cover the band from 0.3 mm to 0.75 microns, that is, the band between the millimeter waves and visible light. Practically all bodies with temperatures above absolute zero radiate infrared rays.

When propagated into the atmosphere the beam-type flux from the infrared spectrum is attenuated because of scattering and absorption. Scattering is caused by refraction, reflection, and diffraction of the beam-type flux by the discontinuities. The attenuation factor due to scattering can be established through the relationship

$$\beta = \beta_0 \left(\frac{0.55}{\lambda}\right)^4,$$

where

λ is the wavelength in microns;

β_0 is the attenuation factor for visible light (λ = 0.55 micron).

β_0 values, depending on atmospheric conditions, are listed in Table 3.6.

Table 3.6

Attenuation factors for visible light energy in accordance with atmospheric conditions

Atmospheric conditions	β'_0, neper/km	β_0, db/km
Very heavy fog	>48	>416
Heavy fog	48	416
Moderate fog	12	104
Light fog	4.75	41
Heavy smoke	1.2	10.4
Light smoke	0.6	5.2
Clear	0.24	2.1
Very clear	0.12	1.04
Exceptionally clear	0.04	0.35

Some of the radiation energy is converted into other types of energy when the beam-type flux is absorbed. Absorption in the molecules and gases in the atmosphere is resonant in nature.

Ozone in the infrared region of the spectrum has absorption lines in the 4.63-4.95, 8.3-10.6, and 12.1-16.4 micron bands.

CO_2 has the broadest absorption in the 4-4.8 and 12.9-17.1 micron bands.

Water vapor has heavy, broad absorption in the 0.926-0.978, 1.095-1.165, 1.319-1.498, 1.762-1.977, and 2.32-2.845 micron bands. Width of the bands, and intensity of absorption by water vapors and CO_2 are greatest at the earth's surface. They diminish with increase in height above sea level.

Chapter IV

Antenna Feeder Systems

Transmission Lines

Transmission lines are systems designed to transmit the energy of electromagnetic waves.

The following types of transmission lines are currently in use:

two-wire open-wire and shielded;

coaxial;

rectangular and round waveguides;

strip waveguides.

The process of transmitting electromagnetic energy along a line is always wave-like in nature in the transmission line, regardless of the type of line used.

4.1 Voltages and Currents on a Transmission Line

An incident wave is propagated from the generator to the load. On a long lossy line the amplitude of the incident wave diminishes approaching the load (fig. 4.1). If the load is unmatched some of the power is reflected and there is a reflected wave (fig. 4.2).

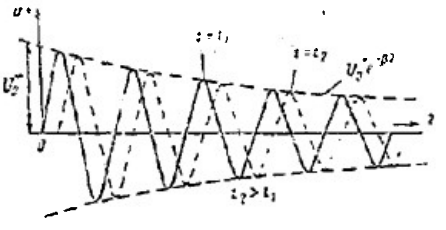


Figure 4.1. Graph of the voltage of an incident wave on a line.

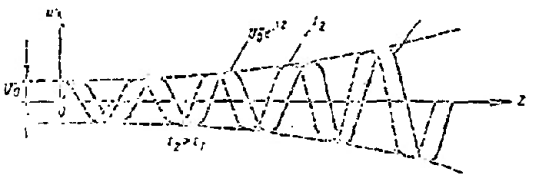


Figure 4.2. Graph of the voltage of a reflected wave on a line.

The expressions for the instantaneous values of the voltages of an incident, u^+, and a reflected, u^-, wave on a transmission line are

$$\left.\begin{aligned} u^+ &= U_0^+ e^{-\alpha z}\cos(\omega t - \alpha z), \\ u^- &= U_0^- e^{\alpha z}\cos(\omega t + \alpha z + \psi), \end{aligned}\right\} \qquad (4.1a)$$

where

U_0^+ and U_0^- are the amplitudes of the incident and reflected waves at the origin of the line (when $z = 0$);

z is a coordinate of a point on the line;

ω is the cyclic frequency of the oscillation;

$\alpha = 2\pi/\lambda$ is the phase factor; it characterizes the difference in phase in the oscillations of the traveling wave at points spaced unit length from each other;

β is the attenuation factor;

ψ is an angle taken into consideration the phase displacement at the load and the delay between the direct and the return propagation of the waves.

For current waves

$$\left.\begin{aligned} i^+ &= \frac{U_0^+}{\rho} e^{-\beta z} \cos(\omega t - \alpha z), \\ i^- &= -\frac{U_0^-}{\rho} e^{\beta z} \cos(\omega t + \alpha z + \psi), \end{aligned}\right\} \tag{4.1b}$$

where

ρ is the characteristic impedance.

The resultant voltage and current on the line are the application (superposition) of the instantaneous values of the incident and reflected waves of voltage and current

$$u = u^+ + u^-, \tag{4.2a}$$

$$i = i^+ + i^-. \tag{4.2b}$$

The latter equation, when transformed with (4.1a) and (4.1b) taken into consideration can be written

$$i = \frac{1}{\rho}(u^+ - u^-). \tag{4.3}$$

The relationships describing the propagation of the transverse components of the electrical and magnetic fields on any transmission line, including waveguides, are similar in structure.

In complex form, the resultant voltage and current on the line are described as follows

$$u = \dot{U}(z) e^{j\omega t}, \quad i = \dot{I}(z) e^{j\omega t}. \tag{4.4}$$

where

$\dot{U}(z)$ and $\dot{I}(z)$ are the resultant complex amplitudes;

$$\begin{aligned} \dot{U}(z) &= \dot{U}^+(z) + \dot{U}^-(z), \\ \dot{I}(z) &= \dot{I}^+(z) + \dot{I}^-(z) = \frac{1}{\rho}[\dot{U}^+(z) - \dot{U}^-(z)]. \end{aligned} \tag{4.5}$$

Reflection Coefficient

The complex reflection coefficient is

$$p = V + jW = p_0 e^{j2\alpha z} e^{2\beta z} \tag{4.6}$$

and characterizes the ratio of the amplitudes of incident and reflected waves and the phase displacement between these oscillations at the point under consideration. In the formula at (4.6), $p_0 = \dot{U}_0/U_0^{\uparrow}$ is the complex coefficient of reflection through the section z_0. The value of the modulus (p_0) is within the limits

$$0 \leqslant |p_0| \leqslant 1, \tag{4.7}$$

because the amplitude of the reflected wave cannot be greater than that of the incident wave.

The mulitplier $e^{2\beta z}$ characterizes the change in amplitudes in the case of direct and the return passage from the section under consideration to the load. The multiplier $e^{j2\alpha z}$ characterizes the change in phase compared with the phase of the reflection coefficient through the section z_0.

4.2 Conditions on the Lossless Transmission Line

On short sections of a transmission line, one with a length of several wavelengths, and for purposes of describing the conditions on the line, losses can be ignored, the attenuation factor can be taken as $\beta \approx 0$, and the same relationships that exist on the lossless line can be used to describe conditions. The condition will depend on the ratio of the incident and reflected wave amplitudes, that is, on the modulus of the reflection coefficient $|p|$. The positions of maxima and minima are established by the phase of the reflection coefficient.

Traveling Wave Conditions

If the load is matched to the transmission line and completely absorbs the power incident to it, there will be no reflected wave ($p = 0$) and there will be traveling waves on the line.

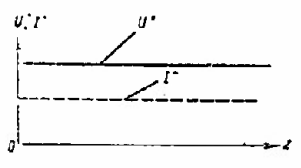


Figure 4.3. Distribution of amplitudes on a line in the case of traveling waves.

As will be seen from (4.5), the amplitudes of the voltages (currents) along the line remain constant (fig. 4.3). Only the phase of the oscillation along the line changes.

Standing Wave Conditions

If the load does not scatter the actual power and completely reflects the incident wave, $|p| = 1$, that is, the amplitude of the reflected wave will equal the amplitude of the incident wave. At the points where the phases of the voltages of the incident and reflected waves coincide the reflection coefficient, p, is a real magnitude. Since $|p| = 1$, the amplitude of the voltage at these points equals

$$U_{max} = 2U^{+}. \tag{4.8}$$

The points indicated are spaced $\Delta z = \lambda/2$ apart.

In accordance with (4.5), in these sections the current amplitude is

$$I_{min} = 0. \tag{4.9}$$

The same is true for points where the phases of the voltages of the incident and reflected waves are opposite, and the resultant amplitude of the voltage equals zero

$$U_{min} = 0, \tag{4.10}$$

while the amplitude of the current has a maximum

$$I_{max} = 2I^{+} = U_{max}/\rho. \tag{4.11}$$

Thus, the resultant voltage and the current in the case of complete reflection (when $|p| = 1$) are represented by the pattern of standing waves with nodes and loops (fig. 4.4).

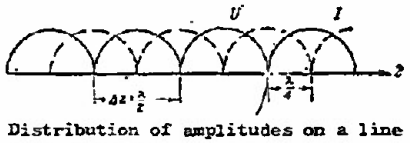


Figure 4.4. Distribution of amplitudes on a line in the standing wave condition.

Mixed Condition

If the load at the end of the line absorbs some of the power, but some is reflected, the value of the modulus of the reflection coefficient is in the limits

$$0 < |p| < 1. \tag{4.12}$$

The condition obtained (fig. 4.5) is somewhere between the traveling and standing wave condition. The sections of maxima and minima are called characteristic sections. Across these sections where the amplitude of the voltage is maximum the amplitude of the current is minimum, and vice versa. The distance between two adjacent maxima (minima) is $\Delta z = \lambda/2$.

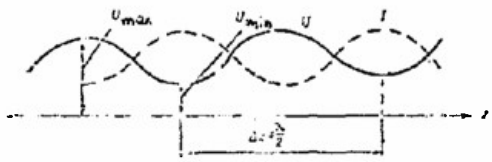


Figure 4.5. Distribution of amplitudes on a line when mixed waves prevail.

Characteristics of the Condition on a Transmission Line

It has been accepted that the condition will be characterized by the value of the traveling wave ratio, K_{tw}. The traveling wave ratio is numerically equal to the ratio of the minimum amplitude of the voltage (or current) to the maximum,

$$K_{tw} = U_{min}/U_{max} = I_{min}/I_{max}. \tag{4.13}$$

The magnitude the inverse of K_{tw} is called the standing wave ratio, K_{sw}.

$$K_{sw} = 1/K_{tw} = U_{max}/U_{min} = I_{max}/I_{min}. \tag{4.14}$$

The association between K_{sw} and the magnitude of $|p|$ stems from the relationship at (4.5)

$$K_{sw} = 1 + |p|/1 - |p|. \tag{4.15}$$

For traveling wave conditions $|p| = 0$, $K_{tw} = 1$, $K_{sw} = 1$. For standing wave conditions $|p| = 1$, $K_{tw} = 0$, $K_{sw} = \infty$. In the mixed wave condition $0 < |p| < 1$, $0 < K_{tw} < 1$, $1 < K_{sw} < \infty$.

Input Impedance of a Line Segment

The input impedance, $Z(z)$ of a line of section z is the impedance of a line of length $l - z$, loaded by impedance, Z_{load}

$$Z(z) = \frac{U(z)}{I(z)} = \rho \frac{1+p}{1-p}. \tag{4.16}$$

The reduced (standardized) impedance is

$$Z'(z) = \frac{Z(z)}{\rho} = \frac{1+p}{1-p} \tag{4.17}$$

$Z(z)$ is a complex magnitude for an arbitrarily selected section.

The magnitude $p = |p|$ in sections of voltage maximum (current minimum), so

$$Z'_{max} = 1 + |p|/1 - |p| = K_{sw}, \tag{4.18}$$

that is, it is a real magnitude.

Similarly, across the sections of voltage minimum (current maximum) the magnitude $p = -|p|$, and across these sections

$$Z'_{min} = 1 - |p|/1 + |p| = K_{tw}. \tag{4.19}$$

Consequently, the input impedance of the line is pure resistance across all characteristic sections and has a complex nature, as in the other sections.

4.3 The Circle Diagram for Long Lines (the Wolpert Diagram)

The Wolpert circle diagram (fig. 4.6) is a pattern of curves of constant values for the reduced pure, R', and reactive, X', resistances constructed on the plane of the complex reflection coefficient

$$p = W + jV = \frac{Z' - 1}{Z' + 1} \tag{4.20}$$

The curves for $R' = \text{const} = f(W, V)$ and $X' = \text{const} = F(W, V)$ are families of circles.

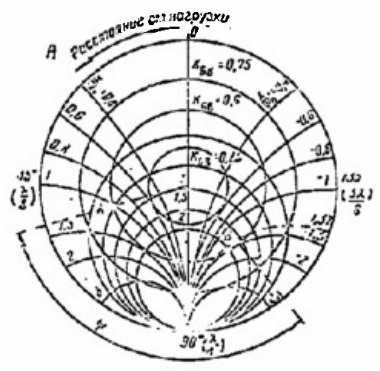


Figure 4.6. Circle diagram for long lines.
A - distance from load.

The curves for the constant value

$$K_{sw} = \text{const and } Z' = z/\lambda = \text{const}$$

are plotted on this same diagram.

These two families, together with the circles for $R' = \text{const}$ and $X' = \text{const}$, yield the circle diagram (fig. 4.6).

The diagram can be used to establish the impedance of various sections of the line for a given line condition and, conversely, to determine what conditions are for a known load impedance. This is why the diagram, constructed using relative magnitudes $R' = R/\rho$, $X' = X/\rho$, $z' = z/\lambda$, as well as the dimensionless magnitude K_{sw}, is universal and can be used for computations on transmission lines of any type.

In this diagram the circles for $K_{sw} = 1$ (or $p = 0$), that is, the match condition corresponds to the central point. The circle corresponding to $K_{sw} = \infty$, that is, standing wave conditions, is the outside circle. The intermediate values of K_{sw} are not plotted to avoid complicating the diagram. The value of K_{sw} can be taken from the curves for $R' = \text{const}$ where they intersect the vertical axis (the axis of real magnitudes). The upper part of the axis of real magnitudes corresponds to the sections of voltage minima (current maxima). On this line $X' = 0$ and $R' = K_{tw} < 1$. The lower part of the axis of real magnitudes corresponds to sections of voltage maxima (current minima). On this line $X' = 0$ and $R' = K_{sw} > 1$. To the left are the arcs of the circles when $-X' = \text{const}$, and to the right the arcs of the circles when $X' = \text{const}$.

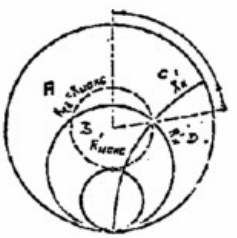


Figure 4.7. Use of the circle diagram to determine the magnitude of K_{sw} on a line for a specified load impedance.
$A - K_{sw} = R_{max}$; $B - R'_{max}$; $C - X'_{load} + jX'_{load}$.

<u>Example</u>. It is given that $Z'_{load} = R'_{load} + jX'_{load}$. Find K_{sw} and the position of the closest minimum with respect to the load.

The solution sequence will be explained through the schematic shown in Figure 4.7.

1. Find, on the diagram, the curves for R'_{load} and X'_{load}. The points at which they intersect establish the position of the impedance, Z'_{load} on the plane of complex p.

2. A concentric circle passing through this point establishes the value of K_{sw}. The value $K_{sw} = R'_{max}$ can be read at the site of the intersection of this circle with the line of maxima.

3. The reduced distance from the load to the closest minimum corresponds to the angle, read counterclockwise, between the radial line passing through the point Z'_{load} and the line of the minima. The figures for the values of the reduced length $z' = z/\lambda$ are plotted on the outside circle in the diagram (fig. 4.6).

4.4 Matching Load and Line

If the line load does not equal the characteristic impedance, there will be a reflected wave on the line in addition to the incident wave, as has already been pointed out. The larger $|\rho|$, the greater the amplitude of the reflected wave and the closer conditions on the line to the standing wave conditions. The reflected wave causes an increase in losses, and the appearance of over-voltages reduces the electrical strength of the line and intensifies the dependence of the input impedance on the frequency.

A matching device (fig. 4.8) is installed near the load in order to eliminate the reflected wave.

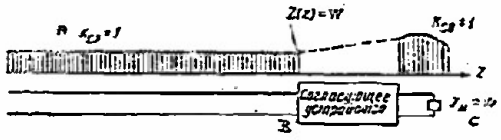


Figure 4.8. Schematic of the installation of a matching device in a transmission line.
A – K_{sw}; B – matching device; C – Z_{load}.

Matching device parameters are selected such that the device will transform load impedance into characteristic impedance; that is, the input impedance of the device will be equal to the characteristic impedance

$$Z_{in} = \rho, \text{ that is } Z' = 1 \text{ or } Y_{in} = 1/\rho, \text{ that is } Y'_{in} = 1. \qquad (4.21)$$

Now pure traveling wave conditions will pertain on the line to the left of the matching device.

The matching device ought not introduce loss, so should be a reactive sink.

Quarter-wave transformers and reactive stubs are most often used as matching devices.

4.5 Two-Wire and Coaxial Lines

Two-wire (symmetrical) lines can be open, or shielded. The construction of open (unshielded) feeders is shown in Figure 4.9, that of shielded in Figure 4.10.

At the present time open feeders are used primarily for feeding symmetrically-coupled long wave and medium wave antennas. The open feeder is held by rigid insulators, or is fixed in place by guys and insulators.

The coaxial (unsymmetrical) feeder can be rigid or flexible. In the rigid coaxial feeder (fig. 4.11) the inner wire is supported by dielectric spacers, or by metal insulators. The latter are short-circuited segments of the

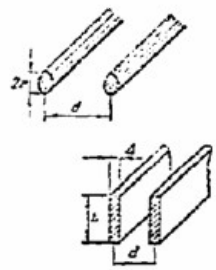


Figure 4.9. Construction of symmetrical two-wire unshielded lines.

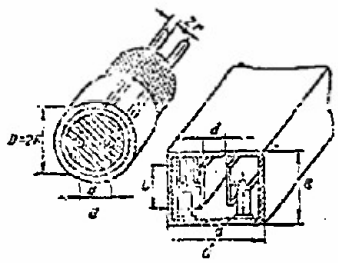


Figure 4.10. Construction of symmetrical two-wire shielded lines. a - with a circular shield; b - with strip conductors.

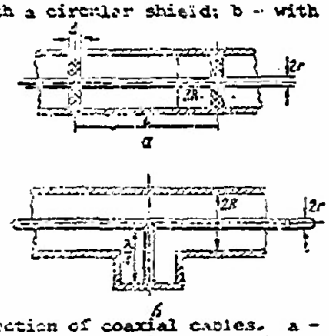


Figure 4.11. Construction of coaxial cables. a - with spacers as insulators; b - with metal insulators.

feeder with electrical length $l_{ins} = \lambda/4$. The impedance of the supporting segment (of the insulator) can be computed through the formula

$$Z_{in\ ins} = 4\rho^2/\beta\lambda, \tag{4.22}$$

where

β is the attenuation factor for the insulator feeder.

Since $\rho \gg \beta\lambda$, $Z_{in\ ins} \gg \rho$, and the insulator will not, for all practical purposes, shunt the main feeder.

The flexible coaxial feeder consists of a flexible outer metallic braid and an inner wire supported by insulators. The insulation most often used is a solid filling of a polyethylene type dielectric.

Two-Wire and Coaxial Line Parameters

These lines have the following main parameters:

characteristic impedance ρ;

line wavelength λ_{line};

velocity rate $\xi = \lambda/\lambda_{line}$;

permissible voltage U_{perm};

attenuation factor β.

The two-wire cable has a

characteristic impedance of

$$\rho \approx \frac{276}{\sqrt{\epsilon_r}} \lg \frac{d}{r}, \tag{4.23}$$

where

d is the distance between the wires;

r is the radius of the wires;

ϵ_r is the relative dielectric constant for the dielectric envelope;

for an aerial cable $\epsilon_r = 1$;

a permissible voltage for an aerial cable of

$$U_{perm} = U_{limit}/k, \tag{4.24}$$

where

$$U_{limit} = 4.6E_{limit}r[cm] \log d/r,\ E_{limit} = 30\ kv/cm; \tag{4.25}$$

$k = 2$ to 3, and is the assurance factor with respect to electrical strength;

an attenuation factor for typical aerial cables in the meter band of

$$\beta = (0.01 \text{ to } 0.04)\ db/meter. \tag{4.26}$$

The coaxial cable has a

characteristic impedance of

$$\rho = \frac{138}{\sqrt{\epsilon_r}} \lg \frac{R}{r}, \tag{4.27}$$

a permissible voltage for an aerial cable of

$$U_{perm} = U_{limit}/k, \tag{4.28}$$*

where

$$U_{limit} = 2.3\ E_{limit}\, r[cm] \log R/r.$$

The wavelength for a cable with a dielectric filler can be computed through the formula

$$\lambda_{line} = \lambda/\sqrt{\varepsilon_r}, \tag{4.29}$$

where

$\lambda = 300/f[mHz]$ is the wavelength in free space, in meters.

Correspondingly, the velocity rate is

$$\xi = \sqrt{\varepsilon_r}. \tag{4.30}$$

For a cable with dielectric spacers

$$\xi = 1.01 \text{ to } 1.05. \tag{4.31}$$

Table 4.1

Electrical characteristics of flexible coaxial cables

Type cable	Characteristic impedance, ohms	Attenuation when f = 100 mHz	Outside diameter, mm	Working voltage, kv
RK-29	50	0.113	9.8 ± 0.5	1.5
RK-19	52	0.2	4.2 ± 0.2	1.0
RK-6	52	0.05	12.4 ± 0.6	4.5
RK-47	52	0.08	10.3 ± 0.5	1
RK-49	72	0.13	6.9 ± 0.3	1
RK-3	75	0.07	13.0 ± 0.7	5.5
RK-1	77	0.11	7.3 ± 0.4	3.0
RK-2	92	0.09	9.6 ± 0.5	4.5

4.6 Waveguides

A waveguide is a hollow metal tube used for the transmission of electromagnetic waves. Any shape of tube cross-section can be used as a waveguide. From the standpoint of construction, the rectangular waveguides are more convenient to use. Round waveguides are most often used when axial symmetry of the waveguide is required, with rotating junctions, for example. Any waveguide

* k = (6 to 7) for a line with metal insulators; k = (15 to 20) for a line with dielectric spacers.

is a unique filter of the upper frequencies and these can propagate oscillations along the waveguide if their wavelengths are shorter than some critical wavelength, λ_{crit}

$$\lambda < \lambda_{crit}. \tag{4.32}$$

The value of λ_{crit} is determined by the transverse dimensions and type of oscillation propagated in the waveguide. The smaller the transverse dimensions, and the more complex the pattern of the oscillation field, the shorter the critical wavelength.

Ordinarily λ_{crit} has the same order as that of the transverse dimension of the waveguide, so it is desirable to use waveguides to transmit waves in the decimeter, centimeter, and millimeter bands.

What is established in any waveguide is a field pattern such that on the walls of the waveguide the tangent component of the magnetic field equals zero.

Transverse electrical oscillations with a longitudinal component H_z of the magnetic field, called type TE waves (or H waves), as well as transverse magnetic waves with a longitudinal component E_z of the electrical field, can be propagated along a hollow tube. The latter are called type TM waves (or E waves).

The wave that will be propagated in a waveguide alone, without the mixing in of other waves, is called the fundamental wave for the waveguide, and the others are called higher types of waves.

For the fundamental wave

$$\lambda_{crit\ fund} > \lambda_{crit\ high}. \tag{4.33}$$

The waveguide should ordinarily operate on the fundamental wave without the mixing in of higher waves. For this purpose the dimensions of the waveguide are selected such that they satisfy the inequality

$$\lambda_{crit\ fund} > \lambda > \lambda_{crit\ high}. \tag{4.34}$$

The wavelength λ_w in the waveguide for the selected oscillation equals

$$\lambda_w = \lambda/\sqrt{1 - (\lambda/\lambda_{crit})^2}. \tag{4.35}$$

The wavelength, λ_w, increases with increase in λ. When $\lambda \to \lambda_{crit}$, the magnitude $\lambda_w \to \infty$, indicating the curtailment of the propagation along the waveguide of the type of oscillation under consideration. The phase velocity, v_p, of a wave in a waveguide characterizing the movement of the field pattern along the waveguide, equals

$$v_p = \lambda_w f = c/\sqrt{1 - (\lambda/\lambda_{crit})^2}. \qquad (4.36)$$

It will always be greater than the speed of light, that is, $v_p > c$.

At the same time, the energy transference rate is

$$v_e = c\sqrt{1 - (\lambda/\lambda_{crit})^2} \qquad (4.37)$$

and always is less than the speed of light, that is, $v_e < c$.

If the condition $\lambda < \lambda_{crit}$ is satisfied for the selected type of oscillation, traveling, standing, and mixed wave conditions can exist in the waveguide, depending on the nature of the load.

4.7 Rectangular Waveguide

E and H waves can exist in a waveguide (fig. 4.12). The critical wavelength for the E and H waves can be established through the relationship

$$\lambda_{crit} = \frac{2}{\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}} \qquad (4.38)$$

For E waves: m = 1, 2, 3, ...; n = 1, 2, 3...

For H waves: m = 0, 1, 2, 3, ...; n = 0, 1, 2, 3 ...

Values m = 0 and n = 0 at the same time are impossible for H waves. The indices m and n indicate the number of standing half-waves along sides a and b, respectively. For the case when the dimension b > a, the highest λ_{crit} can be obtained for H_{01} waves (m = 0; n = 1), which is therefore the fundamental wave for a rectangular waveguide.

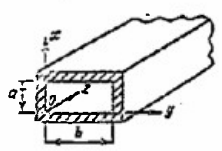


Figure 4.12. Rectangular waveguide.

Fundamental (H_{01}) wave in the rectangular waveguide. Figure 4.13 shows the structure of the electromagnetic field of the wave for the matching case. For this wave $\lambda_{crit} = 2$ b.

The electrical strength of the waveguide operating on wave H_{01} can be established through the expression

$$P_{limit} = E^2_{limit}/480\pi\ ab\sqrt{1 - (\lambda/2b)^2}. \qquad (4.39)$$

For air $E_{limit} = 30$ kv/cm.

Ordinarily

$$P_{perm} = (1/3 \text{ to } 1/5)\ P_{limit}. \qquad (4.40)$$

In order to preclude the possibility of the appearance of higher types of waves, the operating wave is selected from the condition

$$1{,}8\,b > \lambda > \begin{cases} b \\ 2a \end{cases}. \tag{4.41}$$

The characteristic impedance with respect to the voltage is

$$\rho_0 = U^2_{inc}/2P_{inc} = \frac{240\pi\frac{a}{b}}{\sqrt{1-\left(\frac{\lambda}{2b}\right)^2}}. \tag{4.42}$$

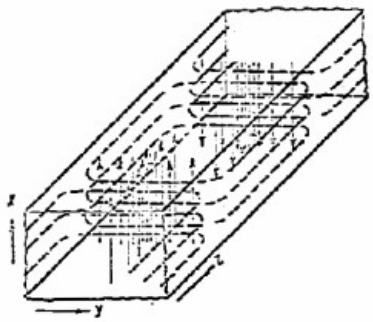


Figure 4.13. Structure of the field in a rectangular waveguide operating on a fundamental wave, H_{01}.

Basic data on rectangular waveguides are listed in Table 4.2.

Table 4.2

Rectangular waveguides

A Внутренние размеры, мм	B Полоса пропускания, см	C Толщина стенок, мм	D Рабочая мощность, Мвт	E Максимальное затухание, $\frac{дб}{м}$
145×292	33—46,8	3,18	45,5	0,003
96×196	20,5—31,2	3,18	18,9	0,003
65×130	13,65—20,7	2,03	8,3	0,012
45×90	9,1—13,83	2,03	3,9	0,017
34×72	7,6—11,55	2,03	2,4	0,024
34×43	5,01— 7,62	1,63	1,04	0,044
15×35	3,67— 5,58	1,63	0,54	0,072
10×23	2,5— 3,03	1,27	0,23	0,127

Key: A - inside dimensions, mm; B - passband, cm; C - wall thickness, mm; D - operating power, mw; E - maximum attenuation, db/meter.

4.8 Round Waveguides

Both E and H waves can exist in round waveguides. The critical wavelength is

$$\lambda_{crit} = 2\pi a/U \qquad (4.43)$$

where

a is the waveguide radius;

$U = \begin{cases} U_{mn} & \text{for E waves} \\ U'_{mn} & \text{for H waves} \end{cases}$

U_{mn} is the nth root of the Bessel function of the first kind of the mth order;

U'_{mn} is the nth root of an arbitary Bessel function of the first kind of the mth order.

Possible values are $m = 0, 1, 2, 3...$; $n = 1, 2, 3...$

The index m indicates the periodicity of the field of the wave (the number of "standing waves" around the perimeter of the circumference of the waveguide). The case of $m = 0$ corresponds to a wave with axial symmetry.

The index n indicates the number of independent coaxial regions, in each of which the energy flows as in an independent waveguide, without flowing into a neighboring region.

The fundamental wave in the round waveguide is the H_{11} wave (fig. 4.14). For this wave

$$\lambda_{crit} = 3.41\ a \qquad (4.44)$$

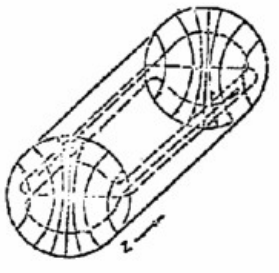


Figure 4.14. H_{11} wave in a round waveguide.

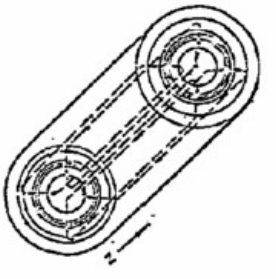


Figure 4.15. H_{01} wave in a round waveguide.

The H_{11} wave can be used in installations in which it is necessary to turn the plane of polarization, as well as in various waveguide transitions.

A symmetrical wave of the H_{01} type (fig. 4.15) can be used in addition to the H_{11} wave. This wave has little attenuation, subsiding with increase in frequency, so it can be used in long-range transmission waveguide lines.

The symmetrical E_{01} wave (fig. 4.16) can be used in rotating junctions (fig. 4.17). In these junctions the contact with the outer walls is through a quarter-wave coaxial stub, 4, which provides the electrical contact at the rotation site. The quarter-wave dividing sleeve, 6, suppresses the radiation of super-high frequency energy. The purpose of ring 3 is to suppress the H_{11} wave, which can appear in the junction because $\lambda_{crit\ H_{11}} > \lambda_{crit\ E_{01}}$.

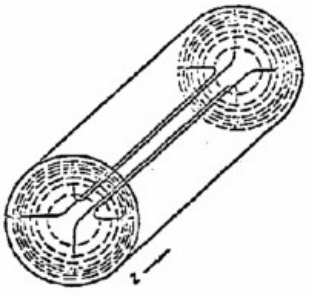


Figure 4.16. E_{01} wave in a round waveguide.

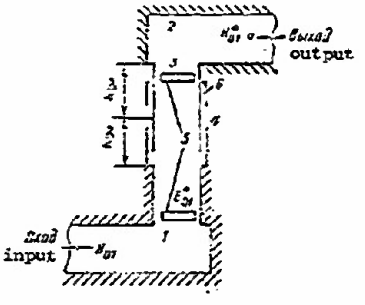


Figure 4.17. Rotating waveguide junction.

The H_{11} and E_{01} waves are also used in cutoff attenuators (fig. 4.18). The dimensions of the waveguide for the cutoff attenuator are selected such that the oscillator wavelength satisfies the condition $\lambda > \lambda_{crit}$. Given this selection of the wavelength, the propagation factor is $\gamma = \beta$, and is real, so the super-high frequency oscillations attenuate along the waveguide.

The dimensions of the cross section of the attenuator are usually selected small and the oscillator wavelength turns out to be a good deal longer than the critical $\lambda \gg \lambda_{crit}$. Now the attenuation factor can be computed through the approximate formula

$$\beta\ [\text{nep/m}] \cong 2\pi/\lambda_{crit}[\text{m}]. \qquad (4.45)$$

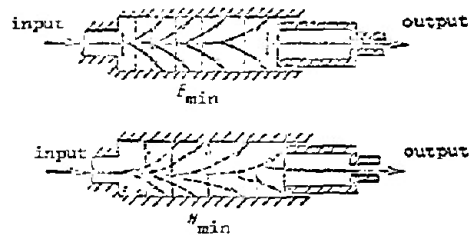


Figure 4.18. E and H type attenuators.

The magnitude of the attenuation derived from typical attenuators ranges within the limits of β = (1 to 3) nep/cm, or β = (10 to 30) db/cm.

4.9 Directional Couplers

Directional couplers are used to transmit some of the electromagnetic energy from one waveguide to another. Because of the directional properties, the incident and reflected waves in the main waveguide in turn set up traveling waves in different directions in the auxiliary waveguide. If there is a wave in one direction in the main waveguide, there will only be a wave in one direction in the case of ideal directivity in the auxiliary waveguide. And in real couplers there is a small amplitude wave in an undesirable direction. The relationship between the powers of the waves propagated on the auxiliary waveguide in desirable and undesirable directions characterizes the directivity of the coupler

$$D\ [db] = 10 \log P^{+}_{aux}/P^{-}_{aux}. \qquad (4.46)$$

In the case of the ideal coupler $P^{-}_{aux} = 0$ and $D = \infty$. Ordinarily the magnitude is D = 20 to 40 db.

The coupling between the main and the auxiliary waveguides is characterized by the iterative attenuation

$$C\ [db] = 10 \log P^{+}_{main}/P^{+}_{aux}. \qquad (4.47)$$

The magnitude of C can take values from C = 0 (full transmission of energy from the main waveguide to the auxiliary) to C = 50 to 70 db, when part of the main power is transmitted by the auxiliary waveguide (10^{-5} to 10^{-7}).

If an absorbing matched load is inserted as a directional coupler in one arm of the auxiliary waveguide, and an indicator is inserted in the other arm, an assembly such as this will only react to one wave, the incident wave, for example. If this directional coupler is then turned so the output

flange is coupled to the oscillator, we obtain an arrangement that will react only to the reflected wave.

The following varieties of couplers are to be distinguished.

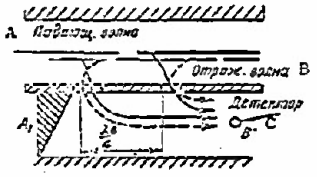


Figure 4.19. Directional coupler with a quarter-wave separation of the coupling holes.

A - incident wave; B - reflected wave; C - detector.

Couplers with quarter-wave separation of the coupling elements. The simplest coupler of this type is the two-hole coupler with the coupling holes in the narrow wall of a rectangular waveguide functioning on the main wave (fig. 4.19). Here directivity is achieved because waves moving through the hole add in phase in the required direction, but are opposite in phase, and mutually compensatory in the case of the opposite (undesirable) direction because of the difference in electrical paths, $\Delta\partial = 2 \cdot \lambda_w/4 = \lambda_w = 2$.

Couplers with several $\lambda_w/4$ spaced holes are used to reduce the iterative attenuation and to increase the directivity in the frequency range. A coupler such as this can be considered as a stage made up of several two-hole couplers.

Quarter-wave stubs, coupling the wide walls (fig. 4.20), are sometime used instead of round holes.

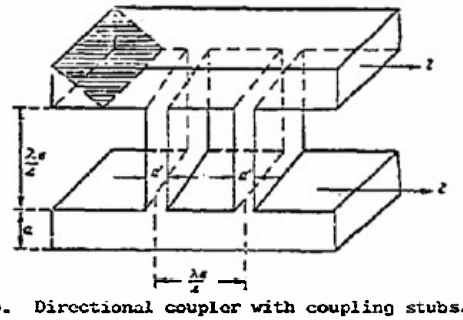


Figure 4.20. Directional coupler with coupling stubs.

Couplers with mixed electromagnetic coupling. Examples of these couplers are those with round and +-shaped holes in the wide wall of a rectangular waveguide (fig. 4.21 a and b). Directivity in these couplings is arrived at

because of the directional properties of the coupling element proper. Specifically, in the coupler shown in Figure 4.21a, the directional properties of the round hole are established by the simultaneous excitation in the hole of a symmetrical E and an unsymmetrical H wave of a circular waveguide. Because of the coupling of those waves the resultant characteristic radiation from the hole into the auxiliary waveguide has directional properties.

In the coupler shown in Figure 4.21b the directivity of the ⊸-shaped hole is established by excitation in the slots of the cross that has time and space shifts of 90° and the resultant characteristic of the radiation from the main waveguide into the auxiliary waveguide too proves to be directional.

Couplers with distributed coupling. This coupler is designed in the form of two waveguides connected through a long slot in a common wall, or in the form of two ribbon lines located quite close to each other. Each section of the slot (line) can be considered an elementary exciter. If a traveling wave is propagated in the main waveguide the phases of the elementary exciters will change along the waveguide in accordance with the law for the change in phase for the particular line. Because of the functions performed by these exciters, a traveling wave in a predetermined direction can also be excited in the auxiliary waveguide.

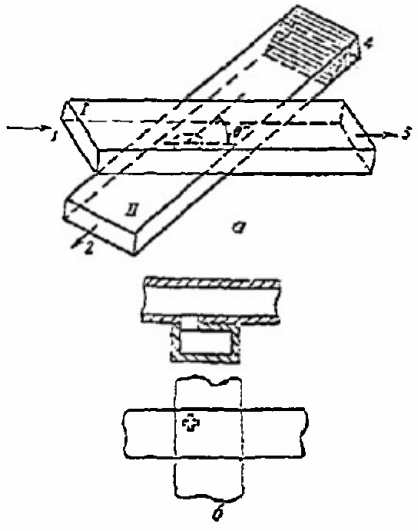


Figure 4.21. Directional coupler with mixed coupling. a - with a round coupling hole; b - with a ⊸-shaped coupling hole.

The change in the direction of propagation of the wave in the main waveguide causes a change in the direction of the wave in the auxiliary waveguide.

Consequently, this system is a directional coupler.

4.10 Waveguide Bridges

Waveguide bridges are used to decouple oscillators on different frequencies and operating on a common load, for measuring mismatches, and as elements in antenna switches and balanced mixers. The bridges most often used are T-bridges and slot bridges.

The T-bridge (fig. 4.22a) is a combination of two waveguide triplets. If matched loads are connected to arms 3 and 4 of the bridge, and then fed from the arm 2 side, there will be no energy flow in arm 1, because of bridge symmetry. If the feed is from the arm 1 the energy will flow to arms 3 and 4, but not to arm 2. Arms 1 and 2 are therefore decoupled when arms 3 and 4 carry symmetrical loads.

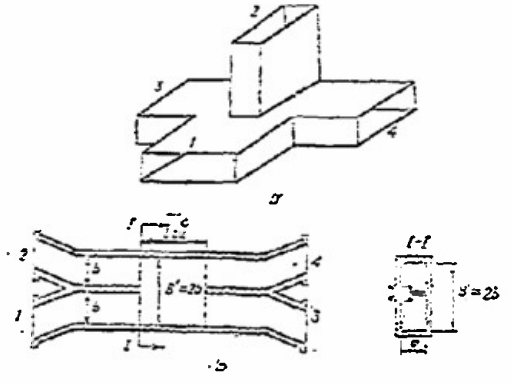


Figure 4.22. Waveguide bridges.
a - waveguide T-bridge; b - slot bridge; c - hertz.

If the bridge is fed through arm 2, a mismatch in either of arms 3 and 4 will result in super-high frequency energy flowing in arm 1. This phenomenon is what makes it possible to use the bridge as a mismatch indicator. The disadvantages of the T-bridge are the need for an additional matching device, and the rigid manufacturing specifications with respect to accuracy.

The slot bridge. The most widely used design is that in the form of two waveguides with a common narrow wall, part of which is cut out (fig. 4.22b). A capacitive stub in the central section compensates for the reflected waves occurring at the site of the junction between the narrow and wide sections.

The slot bridge can be considered a directional coupler with distributed coupling and an iterative attenuation of C = 3 db. If the bridge is fed from the arm 1 side, and if matched loads are connected to arms 3 and 4, no energy will flow in arm 2.

The length of the slot, l_{slot}, in the bridge is selected for the condition

$$l_{slot} \approx \frac{\lambda}{4} \cdot \frac{1}{\sqrt{1-\left(\frac{\lambda}{\lambda_b}\right)^2} - \sqrt{1-\left(\frac{\lambda}{\lambda_b}\right)^2}}. \tag{4.48}$$

In this case power fed to arm 1 will divide equally between arms 3 and 4. At the same time, the phase of the oscillations in the arms will differ by 90°. No signal will pass through arm 2. If one of the loads on arms 3 or 4 is mismatched there will be a reflected wave, part of which will travel on arm 2. Accordingly, the slot bridge is similar to the T-bridge, so far as the functions it performs are concerned. Simplicity of design and good band coverage are advantages of the slot bridge.

4.11 Strip Lines

Strip lines are usually used in super-high frequency circuits in receivers and are made using printed-wire methods. The advantages of these lines are their good engineering features and small size. They can be readily coupled to coaxial lines and waveguides. Disadvantages are low electrical strength and quite heavy losses.

Unsymmetrical (fig. 4.23a) and symmetrical (fig. 4.23b) strip lines can be used.

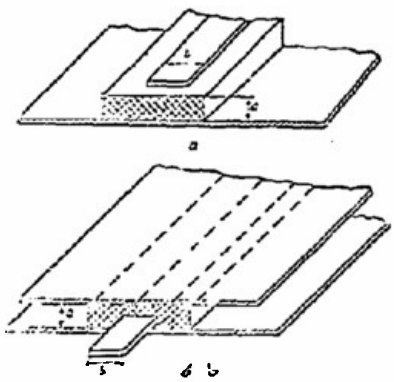


Figure 4.23. Strip lines. a - unsymmetrical strip line; b - symmetrical strip line.

The characteristic impedance of the unsymmetrical line is

$$\rho \approx \frac{120\pi}{\sqrt{\varepsilon_r}} \cdot \frac{d}{b}. \tag{4.49}$$

For the symmetrical line

$$\rho \approx \frac{15\pi^2}{\ln 2 + \frac{\pi}{2}\frac{b}{d}}. \tag{4.50}$$

Strip lines can be used to build directional couplers, bridges, filters, attenuators, and other devices. Design principles are those applicable to waveguides. Strip lines are used primarily in radar receiver plumbing and in small super-high frequency gear.

4.12 The Nonreciprocal Elements in Transmission Lines

Devices in which anisotropic materials, the properties of which differ for waves from different directions, have been developed for use as the dielectric in present-day transmission lines.

The anisotropic material most often used is a magnetized ferrite.

A ferrite is a magneto-dielectric with a crystalline structure which, in its external appearance, resembles a ceramic. An anisotropic magnetized ferrite makes it appear that waves with a different direction of rotation of the plane of polarization are propagated in the ferrite at different phase velocities and are absorbed differently. Specifically, if the exciting field is selected in accordance with the relationship

$$H_0\ [a/m] \approx 28.6\ f[mHz], \qquad (4.51)$$

where

f is the frequency of oscillations of the wave,

then for a wave from one of the directions there is intensive absorption, occasioned by ferromagnetic resonance in the ferrite.

By changing the exciting field we can change the ratio of the propagation parameters for waves moving out, and returning.

The following effects are most often used when ferrite transmission lines are employed:

nonreciprocal (that is, different for direct and return waves) rotation of the plane of polarization;

nonreciprocal phase time delay;

nonreciprocal absorption.

The Nonreciprocal Rotator

Devices providing for the nonreciprocal rotation of the plane of polarization are called nonreciprocal rotators. Most widely used are rotators based on the use of the Faraday effect (fig. 4.24a). A rotator of this type is a circular waveguide stub working on an H_{11} wave, in the center of which is a longitudinally magnetized ferrite rod.

A change in the slope of the plane of polarization is made by changing the excitation current in the coil.

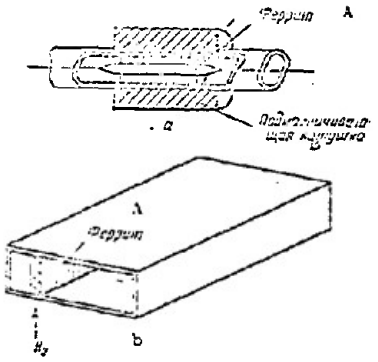


Figure 4.24. Ferrite waveguide elements. a - nonreciprocal rotator; b - nonreciprocal phase shifter (when $H \ll H_0$).
A - ferrite; B - excitation coil.

The Nonreciprocal Phase Shifter and Absorber

The device providing the nonreciprocal phase time delay of the waves is called the nonreciprocal (unidirectional) phase shifter, or nonreciprocal phase switcher.

The most convenient phase shifter is in the form of a stub of rectangular waveguide with a longitudinally magnetized ferrite plate (fig. 4.24b). The difference in the phase time delays for waves moving in opposite directions can be established by the magnitude of the exciting field, and the intensity of the H magnetic field is taken much lower than resonance ($H \ll H_0$), so absorption will be approximately identical, and small, for direct, as well as return, waves.

Devices that absorb direct and return waves differently are called nonreciprocal attenuators (reflected wave insulators).

The simplest attenuator can be obtained if, in the device reviewed above (fig. 4.24b) in the form of a stub of rectangular waveguide with a transversely magnetized ferrite, the value of the exciting field is increased to some magnitude $H \approx H_0$ that will correspond to the condition of ferromagnetic resonance and of intensive absorption for one of the waves (the reflected wave, for example). At the same time the direct wave will move with slight attenuation.

The Circulator

The circulator is a piece of plumbing in which, because nonreciprocal elements are used, a wave moving to one of the arms will be propagated inside the circulator over a path different from that over which the wave leaving this arm is propagated. Figure 4.25a is a diagrammatic layout of a circulator with four feed arms.

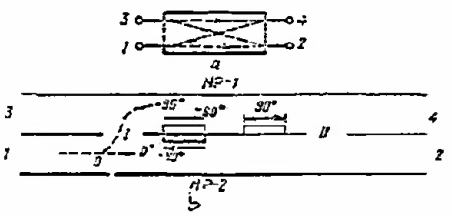


Figure 4.25. The circulator. a - schematic diagram of circulator operation; b - circulator in which the effect of a nonreciprocal phase shift is used.

If the feed is from the arm 1 side, the signal is fed into arm 2. At the same time, if the circulator is fed from the arm 2 side the signal is no longer fed into arm 2, but to arm 3 (or 4), etc. Thus, signal passage in the circulator is in accordance with the pattern 1-2-3-4-1...

The circulator can be used as an antenna transfer switch if the transmitter is connected to arm 1, the antenna to arm 2, the receiver to arm 3, and the absorbing matched load to arm 4.

The circulator can be used as a unidirectional absorber-matching device if, as before, the oscillator and the load are connected to arms 1 and 2, and the matched loads are connected to arms 3 and 4. Now the reflected wave is fed from arm 2 to arm 3, then to arm 4 where it is completely absorbed. Thus, only the incident wave remains on arm 1 and the oscillator feeding the load through the circulator is always working on a matched load.

Circulators using the effect of a nonreciprocal phase shift (fig. 4.25b) are the ones most widely used. The principal parts of the circulator are the slot bridges, I and II, the nonreciprocal phase shifters, NP-1 and NP-2 with $\Delta\varphi_n = -90°$, and the reciprocal phase shifter with $\Delta\varphi_0 = -90°$.

Let us consider the passage of a signal fed into arm 1. After slot bridge I the signal power is halved, and, because of the properties of the bridge, the phase of the signal in the lower section will lead that of the signal in the upper section by 90°. The nonreciprocal and reciprocal phase shifters in the upper part in turn introduce an additional 180° shift, at the same time that the lower nonreciprocal phase shifter provides no phase shift because the ferrite in it is located anti-symmetrically with respect to the ferrite in the upper phase shifter. This, the "upper" and "lower" signals at the input to bridge II will have a 270° shift, as a result of which signals fed into arm 4 are balanced and all the energy flows in arm 2.

If the circulator is fed through arm 2 the signal will halve after passing through bridge II and the "upper" and "lower" signals will be obtained with identical 90° additional phase shifts because of the upper reciprocal and lower nonreciprocal phase shifters. The upper nonreciprocal phase shifter introduces no additional shift. Following reasoning similar to that

in the foregoing, one can be persuaded that in this case the signal is only fed into arm 3. If the device is supplied from the arm 3 side the signal is fed into arm 4. Thus, the device has all the properties of the circulator. Circulators of a type similar to this are widely used in modern radars.

Antenna Devices

Devices used for the radiation and reception of electromagnetic waves are called antennas.

Antennas have the property of reciprocity. Antenna parameters in the reception mode are established by its parameters in the transmission mode.

4.13 Transmitting Antenna Parameters

The radiation pattern shows the dependence of the power (Π) density, or of the amplitude of the antenna field strength (E), on the direction in space when the distance to the points of observation is constant, that is

$$E(\varphi, \Theta) = E_{max} F(\varphi, \Theta), \tag{4.52}$$

$$\Pi(\varphi, \Theta) = \Pi_{max} F^2(\varphi, \Theta), \tag{4.53}$$

where

E_{max} and Π_{max} are the maximum values for the respective magnitudes;

φ, Θ, are the azimuth and meridional angles.

The pattern is depicted in the principal planes by a plane curve in a polar (fig. 4.26a) or in a rectangular (fig. 4.26b) system of coordinates.

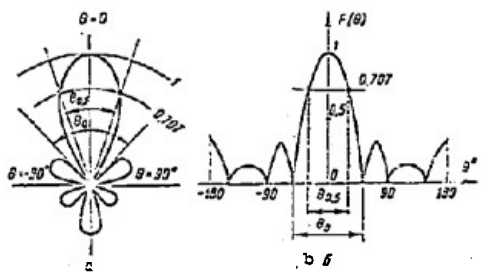


Figure 4.26. Radiation patterns in a polar (a) and in a rectangular (b) system of coordinates.

Patterns are classified according to their shapes.

Figure 4.27a is that of a cosecant pattern, fixed in the vertical plane through the equation

$$F(\varepsilon) = \sin \varepsilon_w \operatorname{cosec} \varepsilon \text{ when } \varepsilon_w \leqslant \varepsilon \leqslant \varepsilon_{max}, \tag{4.54}$$

where

ε_w and ε_{max} are the minimum and maximum elevation angles in the equation for the cosecant pattern.

The cosecant pattern can be used in ground radars for observation purposes and in aircraft radars. The level of the signal reflected from targets at various distances from the radar within its operating range remains unchanged at the receiver input in the case of the cosecant pattern.

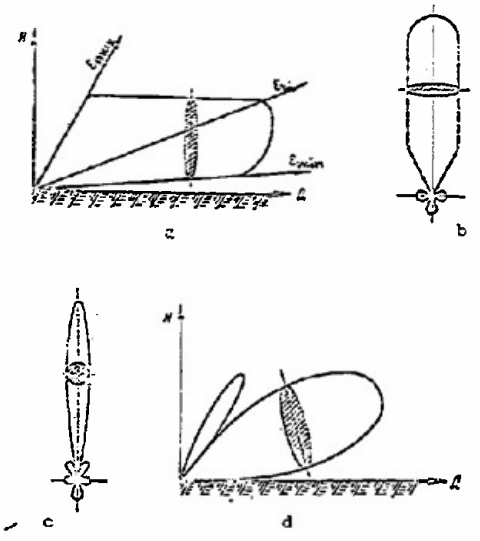


Figure 4.27. Radiation patterns. a - cosecant; b - spade; c - pencil; d - fan.

The spade-shaped pattern, used in radio altimeters, is shown in Figure 4.27b.

The pattern, the major lobe of which is approximately symmetrical with respect to the direction of its maximum, is called pencil-shaped (fig. 4.27c).

The pattern expanded in one principal plane and compressed in the other principal plane is called fan-shaped (fig. 4.27d).

The directive gain of an antenna (G) is that number indicating the gain in the power density, or in radiated power (P_Σ), provided at the point of observation by a directional antenna as compared with a nondirectional antenna (fig. 4.28)

$$G(\varphi, \theta) = \frac{\Pi_\Sigma(\varphi, \theta)}{\Pi_0} \quad \text{when } P_{\Sigma 0} = P_\Sigma, \tag{4.55}$$

$$G(\varphi, \theta) = \frac{P_\Sigma}{P_{\Sigma 0}}, \quad \text{when } E_\Sigma = E_0. \tag{4.56}$$

In the formulas at (4.55) and (4.56) the index Σ denotes a magnitude equated to the field of a directional antenna, while the index 0 denotes that to the field of a nondirectional antenna.

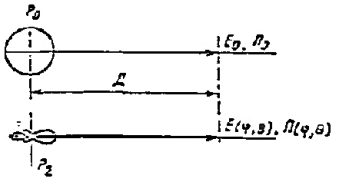


Figure 4.28. Determination of directive gain of an antenna.

The dependence of the directive gain on the direction in space can be established through the pattern by the relationship

$$G(\varphi, \theta) = G_{max} F^2(\varphi, \theta). \tag{4.57}$$

The maximum value of the directive gain can be stablished through the formula

$$G_{max} = E^2_{max} D^2 / 60 P_\Sigma, \tag{4.58}$$

where

D is the distance between the antenna and the point of observation.

The radar resistance (R_Σ) is a factor associating the power radiated by the antenna with the square of the current (I) at a particular point on the antenna

$$P_\Sigma = \frac{I R_\Sigma}{2}$$

The antenna efficiency, η_A, is the ratio of the power radiated by the antenna to the power fed to it (P_A)

$$\eta_A = \frac{P_\Sigma}{P_A} = \frac{P_\Sigma}{P_\Sigma + P_{loss}}, \tag{4.59}$$

where

P_{loss} is the power loss.

The antenna gain, g_A, is a number indicating the actual gain in power density, or in radiated power, provided by a directional antenna, that is, the gain with losses taken into consideration

$$g_A(\varphi, \theta) = G(\varphi, \theta)\, \eta_A = g_{max} F^2(\varphi, \theta). \tag{4.60}$$

The input impedance (Z_A) of a transmitting antenna is the impedance with which the antenna feeder is loaded.

The resistance (R_A) in the input impedance includes the radiation resistance equated to the current flowing in the antenna input ($R_{\Sigma A}$) and the resistance of the losses (R_{loss})

$$R_A = R_{\Sigma A} + R_{loss}.$$

The antenna radiation height is the theoretical length of some equivalent linear array antenna with uniformly distributed current along a conductor and a "current area" (S_I), equal to the "current area" of the real antenna (fig. 4.29)

$$h_{rad} = S_I/I_A,$$

where

I_A is the current at the antenna terminals.

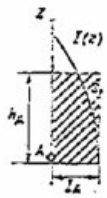


Figure 4.29. Determination of antenna radiation height.

Antenna parameters are associated with h_{rad} by the dependencies

$$h_{rad} = \lambda/\pi\ f_{max}, \tag{4.61}$$

where

$f_{max} = \sqrt{R_{\Sigma A} G_{max}/120}$ is the maximum value of an unnormalized pattern;

$$E_{max} = 60\pi I_A/\lambda D \cdot h_{rad}. \tag{4.61a}$$

4.14 The Receiving Antenna and Its Parameters

The receiving antenna can be represented as a generator of an emf with an internal impedance Z_i. The equivalent circuit of a receiving antenna loaded with impedance Z_{load} is shown in Figure 4.30.

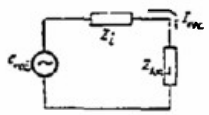


Figure 4.30. Equivalent circuit for a receiving antenna.

The receiving pattern of a receiving antenna can be established as the dependence of the emf induced on the angle of incidence of the arriving wave.

The effective height of the receiving antenna can be established as the dependence of the emf induced on the angle of incidence of the arriving wave.

The effective height of the receiving antenna is the coefficient linking the antenna emf (e_{rec}) with the intensity of the electrif field for the direction of maximum reception (given the condition that the receiving antenna is oriented in accordance with field polarization)

$$e_{rec\ max} = Eh_{rad}. \tag{4.62}$$

The directive gain of a receiving antenna is a number indicating the factor by which the power of the incoming signal arriving from the direction of maximum reception is greater for a directional antenna than it is for an isotropic radiator.

The receiving pattern, the effective height, and the directive gain of an antenna in the receiving mode are the same as the corresponding parameters of the same antenna operating as a transmitter. The internal impedance of a receiving antenna is equal to the input impedance of the same antenna in the transmitting mode. These are the properties that are the antenna reciprocity properties.

The power, P_{rec}, delivered by the receiving antenna to the matched load is established through the formula

$$P_{rec} = \Pi S_A,$$

where

Π is the power density in an incident plane wave;

S_A is the antenna capture area.

The antenna capture area is linked with the other parameters by the dependency

$$S_A = \lambda^2/4\pi \cdot g_{max} F^2(\varphi,\Theta). \tag{4.63}$$

In the event of a mismatch of the load, the magnitude of $S_{A\ mis}$ is rediced in proportion to the degree of mismatch

$$S_{A\ mis} = S_A(1 - |p_{ref}|^2), \tag{4.64}$$

where

p_{ref} is the voltage reflection coefficient.

The effective surfaces of antennas are the radiating surfaces with size S_g equals

$$S_A = S_g K_{uf}, \tag{4.65}$$

where

K_{uf} is the surface utilization factor.

4.15 Systems of Radiators

The resultant system pattern (F_{rs}) for identical radiators can be established by the multiplication rule. It equals the pattern of one radiator in the system (F_1) multiplied by the system factor (F_s)

$$F_{rs} = F_1 F_s. \tag{4.66}$$

The system of radiators is said to be linear if the radiators are positioned in line.

The factor for a discrete equal amplitude linear system of identical radiators (fig. 4.31) with a linear increase in the initial phases of the fields can be established through the formula

$$F_{S.x}(p_x) = \frac{\sin\frac{N_x p_x}{2}}{N_x \sin\frac{p_x}{2}}, \tag{4.67}$$

where

N_x is the number of elements in the system;

$p_x = 2\pi/\lambda\ d_x \sin\Theta - \psi_x$ is the generalized coordinate;

d_x is the distance between radiators in the system;

ψ_x is the difference in the initial phases of the fields of adjacent radiators.

The factor for a similar continuous system of radiators is in the following form

$$F_{S.x}(\xi_x) = \frac{\sin\xi_x}{\xi_x}, \tag{4.68}$$

where

$\xi_x = L_x/2(2\pi/\lambda\ \sin\Theta - \varphi'_x)$ is the generalized coordinate;

L_x is the system length;

φ'_x is the phase displacement per unit length of the system.

The width of the major lobe of the factor for linear, equal amplitude systems can be established through the following approximate formulas:

for a discrete system when $N_x \geqslant 5$

$$\Theta_{0.5} \approx 2\arcsin\frac{0.45\lambda}{N_x d_x}; \tag{4.69}$$

for a continuous system

$$\Theta_{0.5} \approx 51\frac{\lambda}{L_x}. \tag{4.70}$$

In linear systems in which the cosine amplitude distribution diminishes toward the edge the width of the major lobe of the system factor is approximately 1.3 times that found in a system with uniform amplitude distribution.

A flat, rectangular lattice of radiators (fig. 4.32) can be taken as a linear system consisting of N_y parallel rows, each of which is a linear system of N_x elements. In accordance with the multiplication rule, the factor for an equal amplitude system of radiators located in a rectangular plane can be established as

for a discrete system

$$F_S(p_x, p_y) = \frac{\sin\frac{N_x p_x}{2}}{N_x \sin\frac{p_x}{2}} \cdot \frac{\sin\frac{N_y p_y}{2}}{N_y \sin\frac{p_y}{2}}; \tag{4.71}$$

for a continuous system

$$F_S(\xi_x, \xi_y) = \frac{\sin\xi_x}{\xi_x} \cdot \frac{\sin\xi_y}{\xi_y}. \tag{4.72}$$

In the formulas at (4.71) and (4.72) the generalized coordinates p_y and ξ_y can be found similarly through the formulas at (4.67) and (4.68).

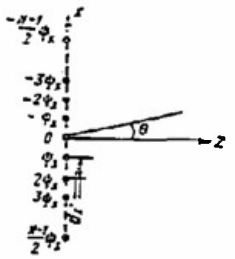


Figure 4.31. A linear straight-line-phase system of radiators.

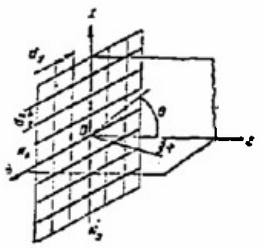


Figure 4.32. A two-dimensional lattice of radiators.

The gain for a system of radiators continuously distributed on a plane can be computed through the formula

$$g_{max} = 4\pi/\lambda^2 S_{geo} K_{uf} \eta_A. \tag{4.73}$$

In the equal amplitude system $K_{uf} = 1$, and in a system with a cosine distribution on one of the sides of the rectangular plane it equals 0.81.

The position of the main maxima of the factors for systems of radiators positioned discretely and continuously are respectively determined on the basis of the conditions that

$$\sin \alpha_{max} = \frac{\varphi\lambda}{2\pi d}, \tag{4.74}$$

$$\sin \alpha_{max} = \frac{\varphi\lambda}{2\pi} \tag{4.75}$$

where

α_{max} is an angle includes by the direction of the maximum radiation and the perpendicular to the axis of the radiators.

4.16 Electrical Control of the Radiation Pattern

The position of the main maximum of the radiation by the system can be controlled by changing the distance between the elements, d, by changing the wavelength, λ, and by changing the phase displacements φ and φ'.

Nonmechanical tilting of the radiation pattern is based on this and is usually done by controlling the phase displacements (phase control), or by changing frequency (frequency control). Both types of control can be exercised mechanically, or electronically.

Phase control is exercised by using various types of mechanical and electrical phase shifters, with the phase shifters connected into the circuit of each of the elements in the lattice. The increase in phase from element to element takes place in accordance with a linear law. Control reduces to smoothly changing the phase in each element from the highest value to zero, and then to the highest value with opposite sign, causing the main maximum to shift from one extreme position to the other.

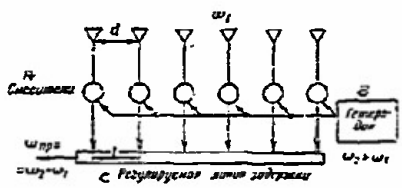


Figure 4.33. Schematic diagram of phase control of the switching of the radiation pattern by frequency conversion.
A - mixers; B - local oscillator; C - receiver; D - adjustable delay line.

Used as mechanical phase shifters are changeable length stubs, controlled phase transformers, rotatable type phase shifters, and others. The electrical phase shifters usually used are waveguides with magnetized ferrites, semiconductor diodes with a controlled capacitance for the junction, and controlled inductances with saturation. It is also possible to use waveguides with an electrified ferroelectric material, or an electron plasma.

One such phase control circuit, called a circuit with frequency conversion, is shown in Figure 4.33. Here the direction of the principal maximum is determined from the condition

$$\sin\Theta_{max} = \frac{\varphi_1}{\varphi_0} - \frac{2\pi}{\varphi_0}, \qquad (4.76)$$

where

$\varphi_1 = 2\pi/\lambda_{line} \cdot l$ is the dephasing of successively radiating elements over section l of a delay line between the taps of two adjacent channels;

λ_{line} is the line wavelength;

$\sigma_0 = 2\pi/\lambda \cdot d$ is the dephasing of successively radiating elements along the array between two adjacent radiators;

d is the distance between these radiators;

n is a factor selected in accordance with the magnitude of the phase displacement and the size of the sector of radiation pattern switching.

Frequency control is illustrated by the series circuit shown in Figure 4.34. The position of the principal maximum for the radiation from the array is, in this case as well, determined from the condition at (4.76), which is in the form

$$\sin \Theta_{max} = \lambda/\lambda_{line} \cdot l/d - n\,\lambda/d. \qquad (4.77)$$

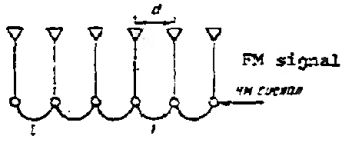


Figure 4.34. Frequency control series circuit for switching a radiation pattern.

Frequency control is accomplished as a result of the dephasing of successively radiating elements from element to element, and is, in fact, the result of change in frequency. This control can be realized through the use of frequency modulation.

Switching of frequency and phase can be done from pulse to pulse, as well as inside a pulse.

4.17 Dipole Antennas. The Symmetrical Dipole

The radiation pattern of a symmetrical dipole (fig. 4.35) in the plane of the electrical vector can be expressed through the following dependency

$$f(\Theta) = \frac{\cos\left(\frac{2\pi}{\lambda} l \sin\Theta\right) - \cos\left(\frac{2\pi}{\lambda} l\right)}{\cos\Theta}. \qquad (4.78)$$

The dipole radiates nondirectionally (fig. 4.36b) in the magnetic plane. The radiation pattern of the most widely used half-wave symmetrical dipole ($2l = \lambda/2$) can be established through the expression

$$F(\Theta) = f(\Theta) = \frac{\cos\left(\frac{\pi}{2}\sin\Theta\right)}{\cos\Theta} \qquad (4.79)$$

and is in the form shown in Figure 4.36a. The amplitude of the field

associated with this dipole can be computed through the formula

$$E(\theta) = 60\, I_{loop}/D \cdot F(\theta), \tag{4.80}$$

where

I_{loop} is the loop current;
D is the distance.

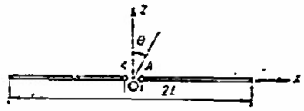


Figure 4.35. Symmetrical dipole.

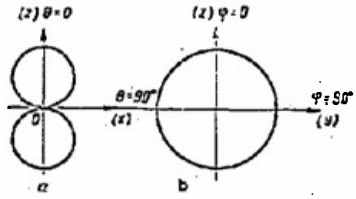


Figure 4.36. Radiation patterns of a half-wave symmetrical dipole in the electrical (a) and magnetic (b) planes.

The active portion of the input impedance of a half-wave symmetrical dipole equals $R_{\Sigma A} = 73.1$ ohms. The effective height of this dipole equals λ/π, as per the formula at (4.61a), and the directive gain is equal to 1.64.

The reactive portion of the input impedance of a half-wave thin symmetrica dipole is inductive in nature and equals 42.5 ohms. Consequently, the length of this dipole is longer than the resonant length. The dipole arm must be shortened to a length equal to the following in order to tune to resonance

$$\Delta = \frac{42{,}5}{2\pi\rho_A}, \tag{4.81}$$

where

ρ_A is the equivalent characteristic impedance of the dipole

$$\rho_A = 120 \ln \frac{\lambda}{1{,}78 r}; \tag{4.82}$$

r is the radius of the dipole conductor.

Dipole arrays. The radiation pattern of an array can be established through the multiplication rule at (4.66). The total radiation resistance and the effective height of a cophasal array consisting of an adequate number of half-wave dipoles, using a parasitic reflector, can be computed, approximately, through these formulas

$$R_{\Sigma\ array}[\text{ohms}] = 135 N_x N_y, \tag{4.83}$$

$$h_{eff\ array} = 2N_x N_y\ \lambda/\pi, \tag{4.84}$$

where

N_x and N_y are the numbers of dipoles along the OX and OY axes.

The array's directive gain can be computed through the formula at (4.73). At the same time the capture area of the array equals

$$S_A \approx 30\pi \cdot h^2_{eff\ array}/R_{\Sigma\ array}. \tag{4.85}$$

4.18 Director Antenna

A director antenna is the name given to a linear array (fig. 4.31) consisting of a directly-fed antenna and parasitic elements, the maximum radiation from which is directed to one side along its axis.

Suppression of the radiation to the opposite side of the array axis is achieved because the elements are spaced $\lambda/4$ apart (fig. 4.37) and have a phase difference of $\psi = 90°$.

In the director antenna the supply is delivered to the directly-fed antenna only. The other elements reradiate the directly-fed antenna field and are called parasitic elements. The initial phase of the field associated with the parasitic directors lags the phase associated with the field of the directly-fed antenna, so the directors are shorter than the directly-fed antenna and their reactance is capacitive.

The field reradiated by the reflector leads the field associated with the directly-fed antenna. The reactance of the reflector is inductive, and its length is longer than that of the directly-fed antenna.

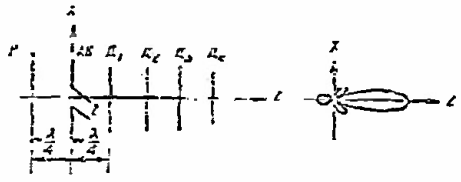


Figure 4.37. Director antenna ("wave duct").

A director antenna with more than one director is usually called a "wave duct." The average value of the distance in the "wave duct" is between 0.25 and 0.35 λ. The directive gain of a wave duct with length L equals

$$G_{max} \approx 5(1 + L/\lambda). \tag{4.86}$$

The directive gain of a wave duct is usually 20 to 30, and the width of the radiation pattern is 30 to 40°. It can be computed, approximately, through the formula

$$\Theta_{0.5} = B\sqrt{\lambda/L}, \tag{4.87}$$

where

B is a factor with a value of 65 to 70.

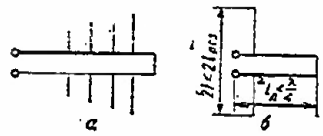
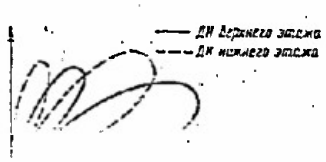


Figure 4.38. Wideband multiunit device with reactance compensation (a) and the principle behind this compensation (b).
1 - resonant; 2 - line.

The wave duct is a comparatively narrow-band system. In order to widen its bandpass, a loop dipole, sometimes made with jumpers and bosses, is used, as is a special device in the form of an array of directly-fed, shortened antennas (fig. 4.38), as the directly-fed antenna. The band voerage by this device is improved because the compensation from the capacitance of each of its dipoles is opposite in sign to the reactivity of the short-circuited line. The band coverage by the wave channel can also be improved by using broadband reflectors (fig. 4.39).

A linear array of "wave ducts" has greater radiation directivity than a simple wave duct. The directive gain of this system is higher at optimum distance and the magnitude of this distance is taken as equal to 1 to 1.2 λ.

Figure 4.39. Two types of multiple-tuned dipole reflectors.



Two-stacked, director-type antennas are used with detection radars to form the radiation pattern, which looks like the cosecant pattern, and to provide a way in which to measure target elevation. This close similarity to the cosecant shaped radiation pattern is the result of the distribution of the power radiated between the upper ($P_{\Sigma u}$) and lower ($P_{\Sigma l}$) antenna stacks, given the condition that

$$h_u/h_l = \sqrt{P_{\Sigma u}/P_{\Sigma l}} = \sqrt{m_u/m_l}, \tag{4.88}$$

where

$h_{u(l)}$ is the height of the upper (lower) stack;

$m_{u(l)} = P_{\Sigma u(l)}/P_{\Sigma}$ is the power distribution factor.

The distance between the stacks can be computed for the condition that the mutual overlapping of the direction of zero radiation by one stack correspond to the maximum of the radiation from the other stack (fig. 4.4). At the same time, h_u = (1.5 to 2) h_l.

If the stages are fed about 90° apart in phase, the radiation pattern of the two-stacked antenna in the vertical plane is in the form

$$f(\varepsilon) = F_{wd}(\varepsilon)\sqrt{m_u \sin^2\left(\frac{2\pi}{\lambda}h_u \sin\varepsilon\right) + m_l \sin^2\left(\frac{2\pi}{\lambda}h_l \sin\varepsilon\right)}, \tag{4.89}$$

where

ε is the elevation;

$F_{wd}(\varepsilon)$ is the wave duct radiation pattern in the vertical plane.

This wave duct is advantageous because it has no deep nulls. A Y-transformer (fig. 4.41) is used to distribute the power between the stacks in the specified ratio. It simultaneously provides a match between the feeder coming from the generator and the feeder branches going to the stacks. The following relationships exist in the Y-transformer circuitry

$$\left.\begin{aligned} m_u &= \sin^2\left(\frac{2\pi}{\lambda}l_u\right) = \cos^2\left(\frac{2\pi}{\lambda}l_l\right); \\ m_l &= \sin^2\left(\frac{2\pi}{\lambda}l_l\right); \quad m_u + m_l = 1; \\ l_u &= l_l + \frac{\lambda}{4}. \end{aligned}\right\} \tag{4.90}$$

The radiation pattern of a two-stacked antenna provides a means for making an approximation of target elevation if this pattern is controlled by a special device. Used for the latter purpose is a stack supply transfer switch that supplies the stacks in phase and out of phase, as well as a goniometer.

A goniometer is a Y-transformer with sliding contacts at points 1 and 2. The condition of $l_u = l_l + \lambda/4$ must not be upset when these contacts are moved. The resultant elevation radiation pattern changes with the positions of brushes 1 and 2, so ε can be determined by the change as shown on a suitably scaled goniometer.

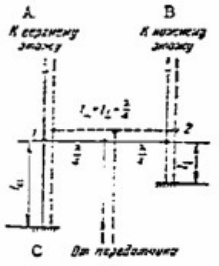


Figure 4.41. Schematic diagram of a Y-transformer and a goniometer. A - to upper stack; B - to lower stack; C - from transmitter.

4.19 Mirror Antennas

Antennas in which the field in the aperture is formed as a result of the reflection of electromagnetic waves from a metal surface with a special profile are called mirror antennas. This antenna, in addition to the mirror, also has an exciter. The type of antenna is determined by the shape of the mirror. The field created by a mirror antenna in space, as well as its field distribution, are determined by the amplitude and phase distribution of the field in the aperture. The aperture is a flat, continuous system of exciters, so its feed pattern can be determined in accordance with the rules for flat continuous systems. The magnitudes characterizing the feed pattern of various apertures with respect to the amplitude distribution are listed in Table 4.3.

Table 4.3

Principal characteristics of radiating apertures

Shape of aperture	Amplitude distribution	K_{ip}	RP width at 0.5P, degrees	First side lobe level
Rectangular a x b	Uniform with respect to a	1	51 λ/a	0.21
	Uniform with respect to b		51 λ/b	0.21
Rectangular a x b	Uniform with respect to a	0.81	51 λ/a	0.21
	Cosinusoidal with respect to b		67 λ/b	0.07
Round, radius r_0	Uniform	1	60 $\lambda/2r_0$	0.2

The gain of a mirror antenna can be established through the formula at (4.73). The efficiency of this type of antenna depends not only on the energy loss during the energy conversion process, but on the losses resulting from energy leakage from the mirror as well. The gain of a mirror antenna has its highest value in the case of the optimum product $K_{ip}\eta_A$. This value

is usually equal to 0.3 to 0.4. This can be provided by that feed pattern in which the power density incident at the edge of the mirror is reduced to 0.1 its value at the center.

Single radiators, continuous and discrete systems of sources, can be used as mirror antenna exciters.

4.20 The Parabolic Antenna

A mirror antenna in which the profile of the mirror is a parabola is called a parabolic antenna.

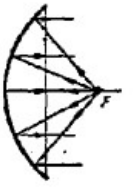


Figure 4.42. Parabolic antenna operating principle.

In the parabolic antenna, the length of the optical path covered by all beams from the focal point to the mirror, and after reflection to the surface of the aperture, is identical (fig. 4.42). Thus, the parabolic antenna can convert the spherical front of a wave from the source, which is at the focal point, onto a flat front in the aperture.

Moreover, after reflection from the mirror the beams diverging from the focal point are parallel. This property is what makes for the highly directive radiation from parabolic antennas.

The mirror, a paraboloid of rotation, forms a pencil-beam pattern. Truncated paraboloids are used to form fan-beam patterns. The truncation can be symmetrical, as well as unsymmetrical, with respect to the focal plane (fig. 4.43). This latter is what makes it possible to considerably reduce the screening effect of the exciter and the reaction of the mirror on the exciter by positioning the exciter outside the zone of the most intensive mirror field.

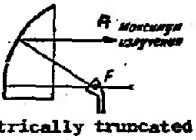


Figure 4.43. Unsymmetrically truncated parabolic antenna with radiation maximum displaced in order to reduce the screening effect of the exciter and mirror reaction on the exciter.

A - radiation maximum.

The width of the radiation patterns of parabolic antennas can be computed approximately:

in the electrical plane

$$\theta^{0}_{0.5} \approx 80 \frac{\lambda}{2R_{0E}}. \qquad (4.91)$$

in the magnetic plane

$$\varphi^0_{0.5} \simeq 70 \frac{\lambda}{2R_{0H}}, \qquad (4.92)$$

where

$R_{OE(H)}$ is the radius of the paerture in the corresponding plane.

The parabolic antenna will form a fan-beam pattern if a linear system of discrete sources (fig. 4.44) is used as the exciter. The displacement, ΔX, of the source from the focal point (fig. 4.44a) leads to a displacement of the maximum in the radiation pattern by the angle $\Delta\theta$. These magnitudes are associated by the relationship

$$\Delta X = 2f \operatorname{tg} \frac{\Delta\theta}{2}, \qquad (4.93)$$

where

f is the parabola's focal length.

By using a ruler of discrete exciters fed from a common channel (fig. 4.44b) we can form a fan-beam pattern because the fields of the individual sources are added. If each of the exciters if ed independently (fig. 4.44c) we can form a fan of partial (independent) patterns.

When the exciter is moved out of the focal point defocusing occurs, and the result is an increase in the pattern width and a drop in the directive gain (fig. 4.45). This is a shortcoming of parabolic antennas. It can be eliminated in part in spherical and in spherical-parabolic antennas.

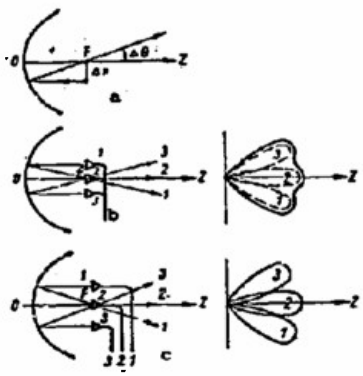


Figure 4.44. Principles involved in the deflection of the maximum in the radiation of a parabolic antenna (a), and their use as the basis for the formation of a fan-beam (b) and of partial (c) patterns.

A shallow, spherical mirror will function in almost the same way as does a parabolic mirror, if the exciter is located at a point half the radius

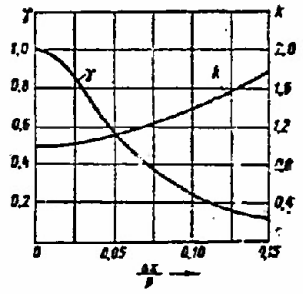


Figure 4.45. The coefficient of reduction in the directive gain (γ) and the coefficient of expansion of the main lobe (k) in the radiation pattern of a parabolic antenna in terms of the displacement of the exciter from the focal point.

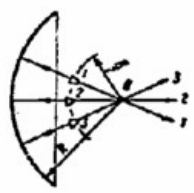


Figure 4.46. Principle of operation of the spherical mirror.

Figure 4.47. Spherical-parabolic mirror.

of the sphere, and if the direction of its maximum radiation is oriented along the radius. The beam reflected from the mirror passes through the center of the sphere. If the exciter is displaced by some angle, the beam will be deflected by that same angle. This is the basis on which the fan-beam pattern (fig. 4.46) can be constructed.

The shape of the spherical mirror can be made parabolic to improve its focal properties in one of the planes. The spherical-parabolic antenna can thus be formed (fig. 4.47).

4.21 Mirror Antennas with a Cosecant Pattern

The methods discussed above in connection with the formation of fan-beam patterns can be used to form cosecant radiation patterns. Then too, that method which changes the shape of the parabolic mirror, that is, using

a mirror with a dual curvature (fig. 4.48), can also be used.

Figure 4.48. Principle of operation of a mirror with dual curvature.

The cosecant pattern has two distinctive sections (fig. 4.27). A so-called triangular fan is contained within elevation limits from ϵ_{min} to ϵ_{fan}. The cosecant fan is contained within the sector from ϵ_{fan} to ϵ_{max}. The triangular fan forms as in the conventional fan-beam pattern.

The number of displaced exciters needed to form this fan can be established through the formula

$$n = \epsilon_{fan} - \epsilon_{min}/\theta_{0.5}k_{av}. \tag{4.94}$$

Here the average value of the expansion factor (k_{av}) for the pattern as a result of the displacement of the exciter from the focal point can be taken according to the number (n) of exciters required. This is why the computation through the formula at (4.94) can be carried out by successive approximations.

The number of exciters needed to form the cosecant fan can be computed similarly.

A cosecant shaped pattern can be arrived at by a redistribution of the powers radiated in the fan and by changing the directive gain in various sections of the fan.

The intensity of the electromagnetic field at the maximum of the lobe in the antenna radiation pattern can be approximated through the formulas

in the near zone

$$\Pi_{av} = 3P_{\Sigma\ av}/S_A, \tag{4.95}$$

in the far zone

$$\Pi_{av} = P_{\Sigma\ av}G/4\pi D^2, \tag{4.96}$$

where

D is the distance.

The distance to the boundary between the near and far zones can be approximated from the condition that

$$D_{boundary} \geq \pi L_A^2/8\lambda, \tag{4.97}$$

where

L_A is the largest dimension of the antenna aperture.

4.22 Anti-T-R Boxes

Anti-T-R boxes are used with radars having a common antenna to switch the antenna to the transmitter when the latter is transmitting, and to the receiver when that unit is receiving. Anti-T-R boxes can be divided into switchgear and those built using non-reversible elements.

Resonant stubs, balanced bridges, and circulators (balanced anti-T-R boxes) are used in switchgear types. A typical schematic of an anti-T-R box using resonant stubs is shown in Figure 4.49. A gas discharger is the switch in this circuitry. Three stages of blocking in the receiving branch provide reliable protection for the receiver within the operating band of frequencies. Discharger P_2 is cut in through a transformer in the first blocking stage, reducing the "hot" resistance of the discharger by a factor of $(\rho_1/\rho_2)^2$, providing for heavy attenuation of the leakage power in the receiver.

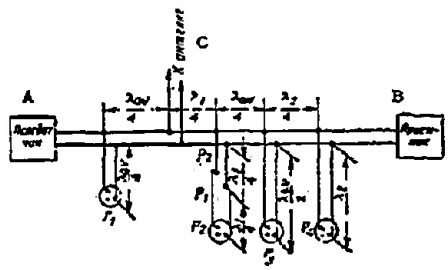


Figure 4.49. Typical schematic of the switchgear type radar antenna transfer switch.

A - transmitter; B - receiver; C - to antenna.

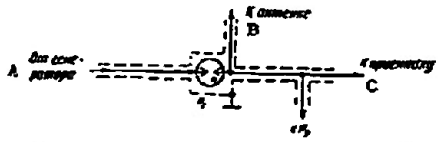


Figure 4.50. Schematic diagram of how a transfer gas discharger that switches the branch coming from the generator into an antenna is cut in.

A - from generator; B - to antenna; C - to receiver.

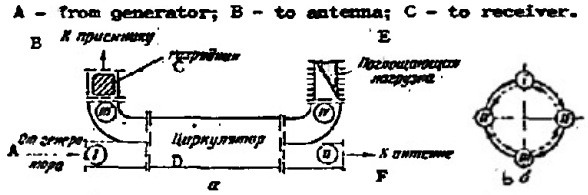


Figure 4.51. Schematic diagram of a balanced antenna anti-T-R box designed to use a circulator with a ferrite phase shifter (a) and the schematic diagram of the circulator in this transfer switch (b).

A - from generator; B - to receiver; C - discharger; D - circulator; E - absorbing load; F - to antenna.

In the meter band radars the transmitter circuit is blocked by transfer dischargers inserted in the circuit as shown in the diagram in Figure 4.50.

A typical balanced transfer switch circuit, using a circulator, is shown in Figure 4.51.

The switching devices are connected to the circulator arms as follows:

to arm I - generator;

to arm II - antenna;

to arm III - receiver;

to arm IV - the absorbing load.

Discharger P_2, the blocking branch for the receiver in the transmit mode, is cut into arm II. This blocking is needed to protect the receiver from the energy fed into this arm if the antenna is not matched with the feed line. The energy reflected from discharger P_2 is fed into the absorbing load.

Upon change in the direction of circulation the feedback energy from the transmitter is fed from arm I into the matched load (arm IV).

Chapter V

Electronic Devices

5.1 Classification of Electronic Devices

The principles of operation of electronic devices are based on the phenomena created by currents of electrons, on controlling the motion of these currents and converting them into elements of the devices themselves. With reference to the characteristics of the space in which the electron currents move, electronic devices are divided into vacuum, gas discharge, and semiconductor devices.

Using electronic devices it is possible to accomplish:

rectification (the conversion of ac current into dc);

generation (the conversion of dc current into ac);

amplification (conversion of low-power electrical oscillations into high-power electrical oscillations);

conversion of electrical energy into light energy or light energy into electrical energy;

conversion of frequency and shape of electrical oscillstions, and

stabilization and commutation of supply voltages and currents.

Vacuum, ion, and semiconductor electronic devices according to their construction, purpose, type of electrical discharges in the discharge gases, and also the character of energy conversion, are divided into a series of standard groups or classes, the most important of which are:

converting devices; in this group are found rectifiers, amplifiers, generators, frequency converters, switches, and other special devices designed to convert electrical currents and voltages;

photoelectric devices; this group includes vacuum and gas-filled photoelements, photo-multipliers, semiconductor photodiodes and phototransistors, electro-optic image converters, and also transmitting television electron ray tubes. Photoelectric devices are used to convert light signals into electrical signals;

electronic indicating devices; this group includes the cathode ray tubes in oscillographs, receiving television tubes (kinescopes), print-out indicator tubes (charactrons), and s series of other special devices. Electronic indicating devices are used to convert electrical signals into light signals that can be observed on luminescent screens.

5.2 Conventional Classification of Vacuum and Semiconductor Electronic Devices

Depending on the operating principles, purpose, and construction characteristics, different types of electronic devices are classified (marked) in accordance with a system of conventional nomenclature, stipulated by the All Union Standard GOST 5461-59. The classification of electronic devices consists of four elements of letters and numbers (Table 5.1).

Table 5.1

Conventional classification of electronic devices

Group	Classification
First element of the classification	
Long-wave and short wave generator tubes (frequencies up to 25 MHz)	GK
Ultrashortwave generator tubes (frequencies from 25 to 600 MHz)	GU
Pulse generating tubes	GI
Kenotrons	V
Voltage stabilizers (stabilitrons)	SG
Current stabilizers (current regulators)	ST
Pulse modulator tubes	GMI
Receiving-amplifying tubes and low-power kenotrons	A number indicating in round numbers the heater voltage in volts
Cathode ray tubes	A number indicating the diameter or diagonal of the screen in centimeters
High voltage pulse diodes	VI
Charge-storage tubes	LN
Gas-filled pulse thyratrons	TGI
Second element of the classification	
Diodes	D
Dual diodes	Kh
Triodes	S
Tetrodes	E
Low-frequency pentodes and ray tetrodes	P
Pentodes and ray tetrodes with sharp cut-off	Zh
Pentodes and ray tetrodes with remote cut-off	K
Frequency-converting tubes with two control grids	A
Triodes with one or two diodes	G
Pentodes with one or two diodes	B
Dual triodes	N

Triode-pentode	F
Triode-hexode and triode-heptode	I
Tuning indicators	Ye
Oscillograph tubes and kinescopes with electrostatic ray deflection	LO
Oscillograph tubes with electromagnetic ray deflection	LM
Kinescopes with electromagnetic ray deflection	LK
Low-power kenotrons	Ts
High-voltage kenotrons)	A number indicating the serial number of the type of device
Voltage stabilizers)	
Charge-storage tubes)	

Third element of the classification

Generating tubes of all ranges)	A number indicating the serial number of the type of device
Cathode ray tubes of all types)	
Receiving-amplifying tubes and low-power kenotrons)	

Fourth element of the classification

Receiving-amplifying tubes and low-power kenotrons:	
in a metal envelope	None
in a glass envelope	S
in a ceramic envelope	K
"acorn" type	Zh
miniature ("finger-tip"), 19 and 22.5 mm in diameter	P
subminiature, over 10 mm in diameter	G
subminiature, 10 mm in diameter	B
subminiature, 6 mm in diameter	A
with a retainer in the switch	L
with disc solders	D

Notes: 1. A nonexisting element in the conventional classification (besides the last one) is marked with a dash.

2. Besides the four elements in the classification for receiving-amplifying tubes, there may be additional letters: V - for vibration-resistance devices, Ye - for long-lived devices, I - for pulse devices.

A fraction appears as the fourth element of the classification for high-voltage kenotrons, gas-filled tube rectifiers, and thyratrons: the numerator gives the average and pulse current in amperes; the denominator gives the allowed reverse voltage in kilovolts.

Examples: 1. 6N1P-Ye: "6" - heater voltage 6.3 V; "N" - dual triode; "1" - series number of the type; "P" miniature configuration ("finger-tip" tube): "Ye" - long-lived device (5,000 hour life-time). 2. 45LM1V: "45" - screen diameter approximately 45 cm; "LM" - oscillograph tube with electromagnetic ray deflection; "1" - serial number of the type; "V" - two-layer screen, white glow, yellow afterglow (see the basic parameters of cathode ray tubes). 3. TGI1-700/25: "TGI" - gas-filled thyratron, pulse; "1" - serial number of the type; "700/25" - 700 A anode current in the pulse and 25 kv maximum anode voltage.

System of Marking Semiconductor Devices

1st element in the classification: a letter or a number specifying the original material.

G or 1 for germanium. Letter if $t_{max} < 60°C$; number of $t_{max} \geq 70°C$.

K or 2 for silicon. Letter if $t_{max} < 85°C$; number of $t_{max} > 120°C$.

A or 3 for gallium arsenide.

2nd element: a letter designating class or group of devices.

Diode - D;

transistors - T;

variable capacitors - V;

superhigh frequency diodes - A;

photo devices - F;

noncontrolled thyristors - N;

controlled thyristors - U;

tunnel diodes - I;

stabilitrons (Zener diodes) - S;

rectifier columns or blocks - Ts.

3d element: number specifying the purpose or electrical characteristics of the devices (Table 5.2).

4th element: letters (A, B, C, ...), determining the variety of a given development type.

Examples: Germanium parametric diode - 1A402A; silicon mixing detector - 2A101A; tunnel diodes of gallium arsenide - 3I301a to 3I301G; germanium transistor - 1T308A to 1T308G; silicon transistor - 2T301 to 2T301Zh.

Table 5.2

Third element of the classification for semiconductor devices

For Diodes

Rectifiers						UHF diodes							Noncontrolled thyristors		
1 низкой мощности	2 средней мощности	3 большой мощности	4 Универсальные	5 Импульсные	6 Варикапы	7 смесительные	8 видео	9 модуляторные	10 параметрические	11 переключательные	12 фотодиоды	13 фототранзисторы	1 низкой мощности	2 средней мощности	3 большой мощности
101–199	201–299	301–399	401–499	501–599	101–199	101–199	201–299	301–399	401–499	501–599	101–199	201–299	101–199	201–299	301–399

Controlled thyristors			Tunnel			Stabilitrons								
						Low power			Average power			High power		
1 низкой мощности	2 средней мощности	3 большой мощности	14 генераторные	15 усилительные	16 переключательные	0,1–9,9 в	10–99 в	100–199 в	0,1–9,9 в	10–99 в	100–199 в	0,1–9,9 в	10–99 в	100–199 в
101–199	201–299	301–399	101–199	201–299	301–399	101–199	201–299	301–399	401–499	501–599	601–699	701–799	801–899	901–999

For Transistors

Low power			Average power			High power		
17 низкочастотные	18 среднечастотные	19 высокочастотные	17 низкочастотные	18 среднечастотные	19 высокочастотные	17 низкочастотные	18 среднечастотные	19 высокочастотные
101–199	201–299	301–399	401–499	501–599	601–699	701–799	801–899	901–999

Key: 1 - low power; 2 - average power; 3 - high power; 4 - universal; 5 - pulse; 6 - variable capacitors; 7 - mixing; 8 - video; 9 - modulator; 10 - parametric; 11 - switching; 12 - photodiodes; 13 - photo transistors; 14 - generators; 15 - amplifiers; 16 - switching; 17 - low frequency; 18 - middle frequency; 19 - high frequency.

Before introducing the new classification system for semiconductor devices in 1964, all semiconductor diodes developed earlier had D as the first element of the classification, and transistors had the letter P. The second element for diodes and transistors was a number signifying the purpose or electrical properties of the device, and the third element was a letter (A, B, C...), which differentiated the different types of a given device.

Examples of the old nomenclature: diodes - D1A to D1Zh; D219A, ..., transistors - P15, P502 to P502V, etc.

5.3 Vacuum Tubes

The principle of operation of vacuum receiving-amplifying and generating (transmitting) tubes is based on controlling the density of an electron flow in the operating space of the tube by changing the voltage applied to the electrodes.

Amplifying activity of the tube is caused by the difference in degree to which the grid and anode voltages affect the anode current; grid voltage has a much greater effect on anode current. This allows weak signals, conducted to the controlling grid, to create large changes in anode current and to effect amplification or generation of electrical oscillations.

Tubes with control grids may be used for amplification or generation. However, special generating and modulating tubes are usually used in transmitting tube circuits. The characteristic peculiarity of these tubes in comparison with receiving-amplifying tubes is the increased size of the electrode system, the comparatively large values of supply voltages, currents, and dissipated power. Generating and modulating tubes in most cases are cooled by forced air or liquid.

The physical processes in generating and modulating tubes are practically no different from the physical processes taking place in receiving-amplifying tubes with control grids.

Fundamental Characteristics of Electron Tubes

The properties of electron tubes are divided into characteristics and parameters, which are given in special handbooks of electronic devices or which are indicated on rating plates on each electron tube. The most useful characteristics are the following: anode, anode-grid, grid-anode, and grid.

The anode characteristic is the relationship of anode current to the voltage on the anode with constant voltage on the remaining electrodes.

The anode-grid characteristic is the ratio of anode current to voltage on the first (controlling) grid with constant voltage on the other electrodes.

The grid-anode characteristic is the relationship of current through one of the grids to anode voltage with constant voltage on the other electrodes.

The grid characteristic is the relationship of current in one of the grids to voltage on one or another grid with constant voltate on the other electrodes.

Fundamental Parameters of Electron Tubes

1. The electrical values which determine the standard and extreme operating states are: heater voltage and current, voltage on the electrodes, allowable dissipated power on the electrodes, and other values.

2. Static or rated parameters:

slope of the characteristic (transconductance)

$$S = \Delta I_a/\Delta U_{g1} \text{ [mA/V] when } U_a = \text{const};$$

internal ac resistance of the tube

$$R_i = \Delta U_a/\Delta I_a \text{ [ohm] when } U_{g1} = \text{const};$$

gain

$$\mu = \Delta U_a/\Delta U_{gl} \text{ when } I_a = \text{const.}$$

ΔI_a, ΔU_{gl}, ΔU_a indicate small increments of anode current, grid and anode voltage.

These parameters are linked by the equation

$$SR_i = \mu. \tag{5.1}$$

3. Values which characterize the resistance of the tube to external influences are: allowable intervals of temperature, pressure, and humidity of the surrounding medium, allowable mechanical loads, and other values.

Tubes of a single type always vary somewhat in these parameters from one sample to another. Handbooks give average values of the parameters for tubes of the given type.

4. Interelectrode capacitance:

input C_{in}, capacitance between the control grid and other electrodes not working on ac voltages at the same frequency as that supplied to the control grid;

output C_{out}, capacitance between the anode and other electrodes which do not operate on ac voltage of the same frequency as that of the voltage supplied to the resistance load of the tube;

transfer capacitance C_{tr}, capacitance between anode and control grid.

Example. For a triode, $C_{in} = C_{gc}$, i.e., the capacitance between grid and cathode; $C_{out} = C_{ac}$, i.e., the capacitance between anode and cathode; $C_{tr} = C_{ag}$, i.e., the capacitance between anode and grid.

The value of "cold" capacitance is given in handbooks. In an operating tube the "warm" capacitance is increased by approximately 20 to 30%.

Standard values of the parameters for receiving-amplifying tubes are given in Table 5.3.

The most important electron tube parameters are the following:

Bandwidth coefficient of a tube

$$\gamma = S/C_{in} + C_{out} \text{ [mA/V·pF]}. \tag{5.2}$$

The bandwidth coefficient of the usual amplifying tubes is less than one, but in special tubes for wideband amplification the coefficient is equal to 1 to 3 mA/V·pF.

Table 5.3

Standard values of parameters for receiving-amplifying tubes

1 Группы приборов	2 Крутизна характеристики S [мА/В]	3 Внутреннее сопротивление R_i [кОм]	4 Коэффициент усиления μ	5 Входная емкость $C_{вх}$ [пФ]	6 Проходная емкость $C_{пр}$ [пФ]	7 Выходная емкость $C_{вых}$ [пФ]
8 Триоды для усиления мощности	3—10	0,3—10.	2—10	5—10	3—15	5—10
9 Триоды для усиления напряжения	2—20	10—50	20—100	1—5	0,5—3	1—5
10 Низкочастотные тетроды и пентоды	2—15	10—100	50—1000	3—15	0,1—1	3—10
11 Высокочастотные тетроды и пентоды	3—30	100—2500	1000—10000	3—10	0,1—0,003	2—10

Key: 1 - group; 2 - slope of characteristic, S [mA/V]; 3 - internal resistance, R_i [kΩ]; 4 - gain, μ; 5 - input capacitance, C_{in} [pF]; 6 - transfer capacitance, C_{tr} [pF]; 7 - output capacitance, C_{out} [pF]; 8 - triodes for power amplification; 9 - triodes for voltage amplification; 10 - low-frequency tetrodes and pentodes; 11 - high-frequency tetrodes and pentodes.

Equivalent noise resistance of a tube, $R_{n\,eq}$. The equivalent noise resistance of a tube is that resistance which, when the grid is connected into the circuit at room temperature, creates a noise current in the anode circuit equal to the noise current of the given tube.

For triodes

$$R_{n\,eq} = 2.5 \text{ to } 3/S \text{ [k}\Omega\text{]}. \tag{5.3}$$

For pentrodes and tetrodes

$$R_{n\,eq} \approx I_a/I_a+I_{g2}\,(3/S + 20I_{g2}/S^2) \text{ [k}\Omega\text{]}, \tag{5.4}$$

where

S[mA/V] is transconductance;

I_a, I_{g2} are anode current and current in the screen grid in milliamps.

Active input resistance of the tube at UHF.

$$R_{in} = a\lambda^2 \text{ [k}\Omega\text{]}, \tag{5.5}$$

where

$a = 0.1$ to 3 [kΩ/m^2];

λ is the wavelength in meters.

Maximum leakage resistance in the first grid circuit. The value of this resistance, R_{g1}, is given in handbooks and determines the allowable grid current for negative grid voltage. Usually R_{g1} = 0.1 to 1 MΩ.

5.4 Cathode Ray tubes

Cathode ray tubes is the name customarily given to a large group of electronic devices in which a focused electron ray (beam) controlled by electric signals is used.

Depending on their purpose they are classed as: cathode ray tubes for measurements of physical values and visual observations of regular electrical processes (oscillographic tubes); cathode ray tubes for receiving television images (kinescopes); projection electron ray tubes; and other special classes.

Dependong on focusing and deflection of the electron ray, cathode ray tubes are divided into the following basic types:

with electrostatic focusing and ray deflection;

with magnetic focusing and ray deflection;

with a combined system of focusing and ray deflection.

Cathode Ray Tubes with Electrostatic Control

The construction and supply circuit of a cathode ray tube with electrostatic focusing and deflection of the electron ray are shown in Figure 5.1. All electrodes of the gun are made cylindrical to receive a radially symmetrical electric field, assuring the best conditions for focusing the electron ray. To decrease the effect of the plate field on focusing, each plate is electrically connected to the second anode of the gun through a leakage resistor R ≈ 5 to 7 MΩ.

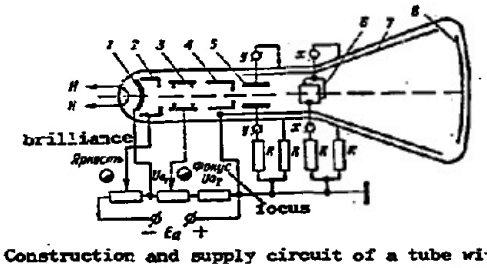


Figure 5.1. Construction and supply circuit of a tube with electrostatic focusing and electrostatic deflection of the electron ray.

1 - cathode; 2 - control electrode (modulator); 3 - first anode of the gun; 4 - second anode of the gun; 5 - vertical deflection plate; 6 - horizontal deflection plate; 7 - graphite cover; 8 - luminescent screen.

Cathode Ray Tubes with Magnetic Control

The deflection system of tubes with magnetic control in most cases consists of two pairs of deflecting windings positioned on the throat of the tube. The illuminated spot on the screen in this type of tube is positioned by varying the currents in the deflection windings. The construction of an electron-optic system and the supply circuit for a CRT with magnetic control are shown in Figure 5.2.

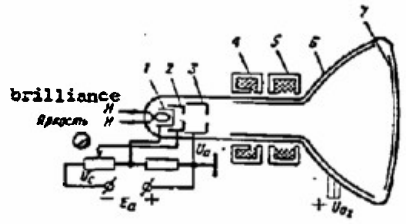


Figure 5.2. Construction and supply circuit for a tube with magnetic focusing and magnetic deflection of the electron ray.

1 - cathode; 2 - control electrode (modulator); 3 - anode of the gun (accelerating electrode); 4 - focusing winding; 5 - deflection winding; 6 - graphite coating - anode of the gun; 7 - luminescent screen.

Compared with tubes with electrostatic control, tubes with magnetic control have the following advantages:

they permit better focusing of the ray with greater current density in the beam;

magnetic deflection permits deflection of the ray at a greater angle, allowing a larger screen to be used with a relatively short tube;

simpler constructions and circuits can be used to obtain radially-circular and spiral scan;

the construction is simpler.

However, the impossibility of having a set of focusing and deflecting windings, the impossibility of expending energy for creating focusing and deflection magnetic fields, and also the limited possibilities of using magnetic tubes for studying rapidly changing electrical processes because of the inductance in the deflecting windings are all essential faults of oscillograph tubes with magnetic control in comparison with tubes with electrostatic control.

Modulation Characteristic of Cathode Ray Tubes

The modulation characteristic of a cathode ray tube is the relationship of the ray current to voltage on the control electrode (modulator) with constant voltages on the remaining electrodes.

Cutoff voltage is the negative voltage on the control electrode (modulator) at which the ray current is equal to zero.

Screen Brilliance

Screen brilliance is calculated by the formula

$$B = AJ_L(U - U_0)^n, \qquad (5.6)$$

where

J_L is the current density of the rsy;

U is potential difference between cathode and screen;

U_0 is minimum potential difference between cathode and screen at which the screen is illuminated;

A is the constant, depending on the physical properties of the phosphor;

$n = 1$ to 2 for the technical phosphors now in use.

Brilliance is almost always controlled by changing J_L by varying the voltage on the control electrode (figs. 5.1 and 5.2, "brilliance" control).

Fundamental Parameters of Cathode Ray Tubes

1. Electrical values which determine standard and extreme operating modes.

2. Diameter or diagonal of the screen. In conventional notation of tube types, this is indicated in centimeters by the first element.

3. Afterglow time (the time required after excitation is removed for the brilliance to decrease to one percentoof its initial value). Tubes are divided into five groups according to the afterglow time of the phosphors used in the screens:

very short afterglow (less than 10 μsec);

short afterglow (10 μsec to 0.01 sec);

average afterglow (0101 to 0.1 sec);

long afterglow (0.1 to 16 sec);

very long afterglow (greater than 16 sec).

4. Glow color. Glow color and afterglow time are indicated by the fourth element of the conventional designation in letters as follows:

A - dark blue glow, short afterglow;

B - white glow, short or sverage afterglow;

V - two-layer screen, white glow, yellow afterglow, long afterglow;

G - violet glow, very long afterglow;

D - light blue glow, green afterglow, long afterglow;

Ye - two-tone screen in shifting bands (first band has an orange glow, and the second band has a green glow), long afterglow;

Zh - light blue glow, very short afterglow;

I - green glow, aversge afterglow;

K - rosy (orange) glow, long afterglow;

M - light blue glow, short afterglow.

5. Tube sensitivity. Sensitivity of tubes with electrostatic control is defined as the amount the light spot on the screen shifts when a voltage equal to one volt is applied to a given pair of deflection plates. It is measured in mm/V.

For plane-parallel plates the sensitivity may be approximated by the formula

$$H' = \frac{1}{2U_{a2}} \cdot \frac{l_1}{d}\left(\frac{l_1}{2} + l_2\right) [\text{mm/V}], \qquad (5.7)$$

where

U_{a2} is voltage on the second anode of the gun, volts;

l_1 is length of the plate in mm;

d is distance between plates in mm;

l_2 is distance from the screen to the plate in mm.

Handbooks give tube sensitivity for each pair of deflection plates. Usually h' = 0.1 to 0.8 mm/V.

Sensitivity of tubes with magnetic control is defined as the distance the light spot on the screen shifts when one amp-turn is applied to the deflection winding, measured in mm/amp turn. The sensitivity of magnetic tubes is determined experimentally and is not given in handbooks, since it depends on the parameters of the deflection windings.

Dark-Trace Tubes (Skiatrons)

These tubes are a variety of cathode ray tubes in which the luminescent screen is replaced by a screen consisting of potassium chloride or sodium chloride crystals. Exciting such a screen with an electron beam produces not a light but a dark spot or a dark trace.

At temperatures close to 300°K the dark spot or trace on the screen may persist for a long time after the excitation is removed, even days or weeks. To clear the screen (erase the recording) in several seconds or minutes, it is heated or subjected to strong external illumination.

Readout Cathode Ray Tubes (Charactrons)

These tubes illuminate letters, numbers and other symbols on a luminescent screen using an electron ray. The cross section of the electron beam must assume the form of the symbol which it is desired to record; then the electron beam, falling on the luminescent screen, prints the required symbol on it.

Construction of a charactron is sketched in Figure 5.3. The significant difference in the construction of this tube in comparison with oscillograph tubes consists of the matrix (metal disc with holes), and the selection and compensation plates. The electron ray, after passing through the selected opening in the matrix, is returned to the axis of the tube by

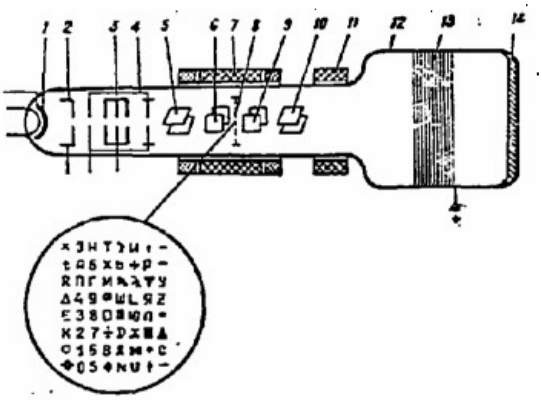


Figure 5.3. Schematic of a charactron.

1 - cathode; 2 - controlling electrode; 3 - first anode; 4 - second anode; 5 and 6 - deflection plates for a symbol selection system; 7 - focusing and correcting windings; 8 - matrix; 9 and 10 - compensating system plates; 11 - deflection winding of the address system; 12 - bulb; 13 - postdeflection acceleration system; 14 - luminescent screen.

the compensating plates and subsequently deflected to the desired point on the screen by the magnetic field of the address system, where it illuminates the symbol chosen on the matrix. The number and type of these symbols depend on the number and shape of the openings in the matrix. Industrially produced charactrons have a matrix with 63 different symbols. Response speed in charactrons reaches 20,000 symbols per second.

Symbolic Cathode Ray Tube (Typotron)

A typotron is a variation of a charactron with a memory for symbolic information on the screen. Formation and choice of the symbols on the matrix in a typotron is accomplished just as in a charactron. The fundamental difference in a typotron compared with a charactron is a special memory device, which permits retention of the inscribed symbol on the screen for a long time (symbolic cathode ray tubes are described in the book by I. Ye. Soloveychik, P. M. Anishchenko, Symbolic Presentation and Its Application in Contemporary Radioelectronic Systems, Sov. Radio Press, 1959.)

Charge Storage Tubes

Charge storage tubes are special cathode ray tubes used for recording, storing and reproducing electrical signals, recorded on a dielectric target.

Recording electrical signals on a dielectric target by an electron beam is based on using the phenomenon of secondary electron emission.

An essential property of the construction of charge storage tubes in comparison with oscillograph tubes is the dielectric target, deposited as a thin layer on a conducting base (signal plate) in place of the luminescent screen.

Charge storage tubes with a barrier grid or readout charge storage tubes (shown in fig. 5.4) enjoy wide use at the present time.

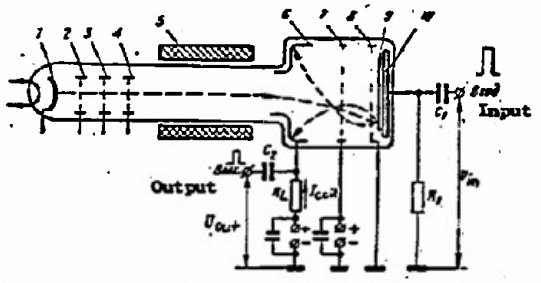


Figure 5.4. Schematic diagram of a charge storage tube with a barrier grid.

1 - cathode; 2 - control electrode; 3 - first anode of the gun; 4 - second anode of the gun; 5 - deflection winding; 6 - collector; 7 - screen grid; 8 - barrier grid; 9 - dielectric target; 10 - signal plate.

Focusing and deflection of the electron ray in charge storage tubes may be electrostatic or magnetic. Scanning the surface of the target by an electron beam may be effected by a raster or spiral.

The operating state of the readout charge storage tube is chosen so that the coefficient of secondary emission, $\sigma = I_2/I_1$ (where I_1 is the current of the primary electron beam, I_2 the current of the secondary electrons dislodged from the target), is larger than one. In this state the input electric signals subject to recording on the dielectric target are lead to the signal plate, and the load resistance R_L, from which output signals are removed on readout by the electron beam before the new signals are recorded, is connected into the circuit of the collector, barrier grid, or signal plate.

In the absence of input signals, the surface of the dielectric target under the influence of the electron beam acquires a surplus positive charge. This is accompanied by an increase in potential on the surface of the target relative to the barrier grid and by the formation of a braking electric field for secondary electrons in the "barrier grid-target" gap. Consequently the flow of secondary electrons from the target decreases.

The potential on the surface of the target increases under the influence of the electron beam until dynamic equilibrium is established, at which the flow of secondary electrons leaving the target becomes equal to the flow of

primary electrons in the beam. This amount of surface potential on the target in the dynamic equilibrium state is called the equilibrium potential and is designated U_{eq}.

Thus at dynamic equilibrium, the potentials of all elements on the target surface reach the value of U_{eq}, positive relative to the barrier grid (potentials of the barrier grid and signal plate are identical and equal to zero). Thereupon, current in the collector circuit, I_{col}, is constant and the output signal $U_{out} = 0$.

On application of input signals, the electric field in the "barrier-grid-target" gap changes. If the input signal is positive, the braking electric field for secondary electrons in the "barrier grid-target" space increases, collector current decreases, and a positive signal proportional to the amplitude of the input signal appears on the resistance R_L in the collector circuit (this output signal appearing on the load while input signal, U_{in}, is applied is called the recording signal, U_{rec}). While the input signal is applied, the potentials of the points on the surface of the target scanned by the electron beam decrease in comparison with U_{eq}; the positive input signal is recorded on the target (fig. 5.5a).

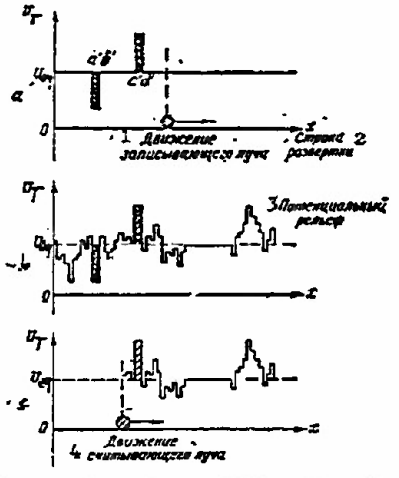


Figure 5.5. Potential changes on the surface of a dielectric target under the influence of an electron beam. a - distribution of potential on the surface of the target during recording of short pulse signals (U_{eq} - equilibrium potential; a'b' - recording of a positive input signal; c'd' - recording of a negative input signal); b - recording on the target of a complex signal; c - change in potential on the target during readout of recorded signals (potential relief) by an electron beam.

1 - recording ray motion; 2 - scanning line; 3 - potential relief; 4 - readout ray motion.

Thus input signals applied to the signal plate produce a change in potential distribution along the scan of the target by the electron beam; a so-called potential relief appears on the target (fig. 5.5b).

Reproduction or readout of the signals recorded on the screen is effected by the electron beam on its return motion over the parts of the target surface on which recording took place. If there are no input signals during readout, then in the readout process, the surface potentials of the target are again returned to the equilibrium value, U_{eq}, by the electron beam (fig. 5.5c). Meanwhile the potential relief modulates the secondary emission current of the collector; the number of secondary electrons trapped by the collector will decrease in relation to the potential distribution on the target along the scan. Consequently, signals corresponding to the previously recorded signals will appear on the output circuit of the charge storage tube on load resistor R_L (these output signals are called reading signals, U_{read}). Polarity of the reading signals is opposite to that of the recorded input signals. If the output signals are removed from the load in the barrier grid or signal plate circuits, their polarity corresponds to the previously recorded input signals/ but this method of removing output signals requires special measures for frequency separation of input and output signals.

During readout of the recorded signals the potential relief is simultaneously erased. Consequently, it is possible to record new input signals on the last scan of the target by the electron beam, and later to read them out.

A charge storage tube with barrier grid is usually used as a readout device. The principle of signal readout is illustrated in Figure 5.6., where $u_{in\ 1}$ is the input signal of positive polarity applied to the signal plate at a time t_1; $u_{1\ rec}$ is the output signal taken from load resistor R_L during recording of $u_{in\ 1}$; $u_{1\ read}$ is the output signal taken from resistor R_L at a time $t_2 = t_1 + T_s$ while reading out the previously recorded signal (T_s is target scan period); $u_{in\ 2}$ is the input signal of positive polarity applied to the signal plate at time $t_2 = t_1 + T_s$; $u_{2\ rec}$ is the output signal taken from R_L while recording $u_{in\ 2}$; $u_{out} = u_{1\ read} - u_{2\ rec}$ is the resultant differential output signal taken from R_L at $t_2 = t_1 + T_s$.

If the charge storage tube were an ideal readout device, then a periodically repeating signal, constant in amplitude and polarity, applied to its input would be read out completely, beginning from the second scan period.

In actuality, not only a useful signal, but also so-called residual and parasitic signals are created in the charge storage tube. Consequently, "zero" differential signal at the output is only obtained after n-fold readout of identical input signals.

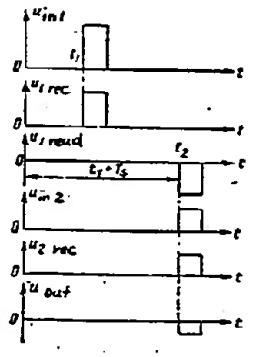


Figure 5.6. Diagram showing the signal readout principle of a charge storage tube.

Fundamental Parameters of Readout Charge Storage Tubes

Recharge coefficient η. The recharge coefficient (coefficient of the first residual) is defined as the ratio of amplitude of the second $U_{2\ read}$ and the first $U_{1\ read}$ readout signals for application to the tube input (to the signal plate) of repetitive sinusoidal voltages. It is measured in %, i.e.,

$$\eta = U_{2\ read}/U_{1\ read} \cdot 100\%.$$

Target suppression factor P_T. Target suppression factor (coefficient of seeding) is defined as the ratio of amplitude of the first output recording signal $U_{1\ rec}$ to the amplitude of residual output signal U_0 with repetitive sinusoidal signals applied to the input of the charge storage tube, i.e.,

$$P_T = U_{1\ rec}/U_0.$$

Dynamic range, D. Dynamic range is defined as the ratio of the sum of output signals $U_{1\ rec}$ and $U_{1\ read}$ to full range of intrinsic noise voltage in the charge storage tube $2U_{n\ max}$ with repetitive sinusoidal signals applied to the input of the tube, i.e.,

$$D = U_{1\ rec} + U_{1\ read}/2U_{n\ max}.$$

For the given conventional operating states of a charge storage tube and with the given operating amplitude of the input signals, η, P_T, D, and the output signal current I_c, depend on the constant component of collector current $I_{col\ 0}$ (i.e., on the ray current). The character of this dependence is shown in Figure 5.7. In every actual application of a readout charge storage tube, it is possible to choose the optimum readout mode by using the relationship of η, P_T, D, and I_c to ray current.

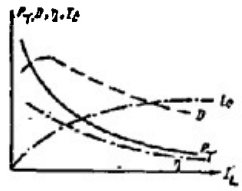


Figure 5.7. Relationship of the basic parameters of a readout charge storage tube to ray current with input signal parameter held constant (U_{in} = const).

5.5 Ion or Gas-Discharge Tubes

Ion or gas-discharge tubes include a large group of tubes based on the physical process in which an electric current passes through a gas-discharge gap.

In contrast to vacuum tubes, in ion tubes current is created by the movement of positive and negative ions, as well as of electrons. When a current is passed through an ion tube, the electric field of slowly moving positive ions compensates for the field of the negative space charge of the electrons; consequently the internal resistance of a gas-discharge tube is very small. This results in large discharge currents for small voltages applied to a gas-discharge tube.

Gas-filled ion tubes usually contain inert gases (helium, argon, neon, krypton, xenon) or mercury vapor. Some types of ion tubes are filled with hydrogen or a mixture of several inert gases.

The density of the gas or mercury vapor filling the tube significantly affects the conditions under which electric current passes through the tube. In gas-filled tubes, the gas pressure for the most part does not exceed tenths of mmHg. In some tubes it reaches tens of mmHg and even several atmospheres.

Ion tubes are divided into two basic groups according to the type of electric discharge:

semi-self-maintained discharge tubes;

self-maintained discharge tubes.

In semi-self-maintained discharge tubes, electron emission from the cathode, necessary to create a discharge current is established by heating the cathode with an external current source (thermoelectronic emission) or by the action on the surface of the cathode and the gas column by quants of radiant energy (photoelectronic emission). In self-maintained discharge tubes, electron emission from the cathode is accomplished by bombarding the surface of the cathode with positive ions incident on the cathode from the discharge gap.

A large variety of ion tubes of both types is used in radar technology: gas-filled tube rectifiers, thyratrons, voltage regulators, spark gaps, fluorescent lamps, decatrons, cyphertrons, and others.

Gas-filled tube rectifiers are gas-filled two-electrode tubes with heated cathode. They are used for rectifying currents at industrial frequency. The discharge is a semi-self-maintained arc.

Thyratrons are three- or four-electrode gas-filled devices. There are thyratrons with heated cathodes (using semi-self-maintained arc discharge) and with cold cathodes (using self-maintained glow discharge). Thyratron grids are designed only for controlling striking, since after discharge the voltage on the grids has practically no effect on anode current. Therefore, such tubes may only be used in certain switching circuits or rectifier circuits with smooth regulation of the rectified current.

Stabilitrons are two-electrode tubes with self-maintaining discharge, used for stabilizing voltages in dc circuits. There are stabilitrons with glow discharge (stabilization voltage U_{st} = 75 to 150 V) and corona discharge (U_{st} = 1000 to 10,000 V).

Decatrons are multi-electrode gas-filled tubes with glow discharge, designed to count electric pulses in a decimal system. Decatrons are classed as two-pulse and one-pulse, according to the method of transferring a charge from one discharge gap to another, and are classed as calculators or commutators according to the function they fulfill.

Cypherotrons are indicating tubes with glow discharge for displaying various data in the form of illuminated numbers.

Fundamental Characteristics and Parameters of Ion Tubes

Depending on the construction peculiarities and purpose of the ion tube, its properties are determined by various characteristics and parameters. Here only some of these parameters, illustrating the most common properties of ion tubes, will be discussed.

Glow potential, U_g. This is the voltage on the electrodes of a gas-discharge gap at which self-maintaining discharge occurs. It is determined from the empirical formula

$$U_g = U_{ion} \cdot pr/const + \ln(pr), \qquad (5.8)$$

where

U_{ion} is the ionization potential of the given gas;

p is gas pressure;

r is the distance between electrodes;

const is a constant for the given gas.

The function $U_g = f(pr)$ is shown in Figure 5.8. From the graph it can be seen that by choosing the product pr, and the composition of the gas, the required value of U_g can be established in the design or construction stage. U_g is the fundamental parameter of ion tubes.

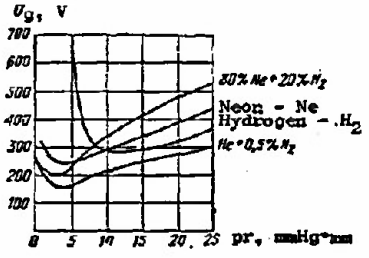


Figure 5.8. Dependence of glow voltage for certain gases and gas mixtures on the product, pr.

In an actual circuit depending on the value of discharge current and temperature of the surrounding medium, pressure, p, inside the tube does not remain constant, which leads to changes in U_g.

Arc voltage, U_{arc}, is the potential difference between anode and cathode produced on passage of the discharge current. U_{arc} is always less than U_g. For mercury devices $U_{arc} \approx 10$ to 20 V; for tubes filled with inert gases and hydrogen, U_{arc} may be as high as tens and hundreds of volts.

Discharge development time, t_{dev}, is the time required to establish current in the discharge gap after voltage is applied to its electrodes. This time varies from several nanoseconds to milliseconds. Consequently, ion devices in many applications should be considered inertial tubes.

The electrical parameters of gas-discharge tubes are maximum and mean current, operating and maximum permissible voltage, limiting frequencies, limiting resistances in the electrode circuits, stabilized voltages, and others.

5.6 Semiconductor Diodes

Semiconductor diodes are classed in two groups according to the construction of their principle working part (the area of contact between the electron and hole-conducting semiconductors, which is usually called electron-hole or p-n junction): junction and point-contact. The majority of diodes is made on a base of germanium or silicon by introducing into the monocrystal strictly controlled quantities of impurities of other elements, which change the character of conduction. Germanium or silicon doped with tri-valent elements (acceptor additives) has hole conductivity (p-type). If five-valent elements are added (donor additive), then the conductivity is is electron (n-type).

According to the preparation method of the p-n junction, planar semiconductor diodes are classed as alloyed, diffused, or "grown."

Current through the p-n junction at constant temperature depends on amplitude and plarity of the applied voltage

$$I = I_0\left(e^{\frac{qU}{kT}} - 1\right), \qquad (579)$$

where

I is the current flowing through the junction;

$q = 1.6 \cdot 10^{-19}$ k is the electron charge;

$k = 8.7 \cdot 10^{-5}$ eV/°K is the Boltzmann constant;

T is the junction temperature in °K;

$e \approx 2.73$ is the base of the natural logarithms;

I_0 is reverse current of the diode, obtained when a certain external voltage is applied, minus to the p region and plus to the n region (I_0 is also called the reverse saturation current);

U is applied coltage.

A graph of this equation for a diode is the volt-ampere characteristic (fig. 5.9).

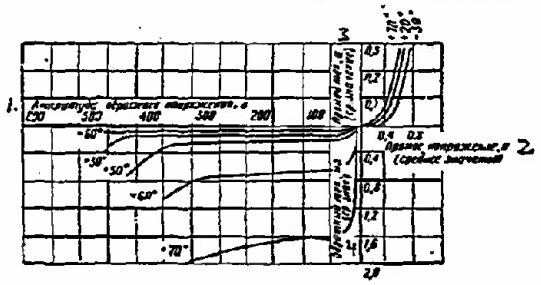


Figure 5.9. Volt-ampere characteristic of a semiconductor diode.

1 - reverse voltage, V; 2 - forward voltage, V (average); 3 - forward current, A (average); 4 - reverse current, mA (average).

As is evident from the graph, a diode is a nonlinear element in an electrical circuit.

Its forward and reverse resistances are very different. Current I_0 depends on temperature. For germanium diodes, reverse current doubles for each 9 to 12°C increase in temperature; for silicon it doubles for each 6 to 8.5°C.

At sufficiently large U_{rev} a sharp rise in reverse current is observed. An increase in reverse current is caused by thermal or electrical breakdown. If current is not limited under these conditions, the diode is destroyed.

Besides active resistance, a p-n junction (diode) has a capacitance, which depends on the original material (mgermanium or silicon), the concentration of additives (thickness of the p-n junction), junction area, and applied voltage. The relationship of capacitance to applied coltage is shown in Figure 5.10.

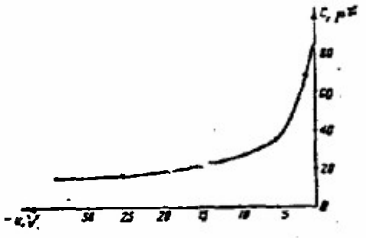


Figure 5.10. Graph of capacitance of a semiconductor diode as compared to applied voltage.

Diodes are classed in three groups according to their frequency characteristics:

low-frequency (rectifying or power diodes);

high-frequency (universal or pulse);

superhigh-frequency.

Within each group, semiconductor diodes are divided according to classification as shown in Figure 5.11.

The properties of the first two groups of diodes are characterized by the following fundamental parameters:

I_{rev} [μA] - reverse current at a certain reverse voltage: for silicon diodes I_{rev}, as a rule, is considerably less than for idential germanium diodes;

U_f [V] - forward voltage drop in the diode at nominal current; for germanium diodes U_f = (0.3 to 0.5) V; for silicon diodes U_f =(0.8 to 1.5) V;

C [pF] - diode capacitance with a certain reverse voltage applied;

f_{max} [MHz] - maximum frequency at which operation without lowering the the rectified current is possible;

U_{max} [V] - breakdown voltage at a given temperature;

$\Delta t°$ - operating temperature range; for germanium diodes $\Delta t°$ = (-60 to +70)°C, for silicon $\Delta t°$ = (-60 to +120)°C.

Parameters given in handbooks are average values, and variation from specimen to specimen may be considerable. Therefore, when diodes are to be connected in parallel or in series, they should be tested, and matching resistances should be connected.

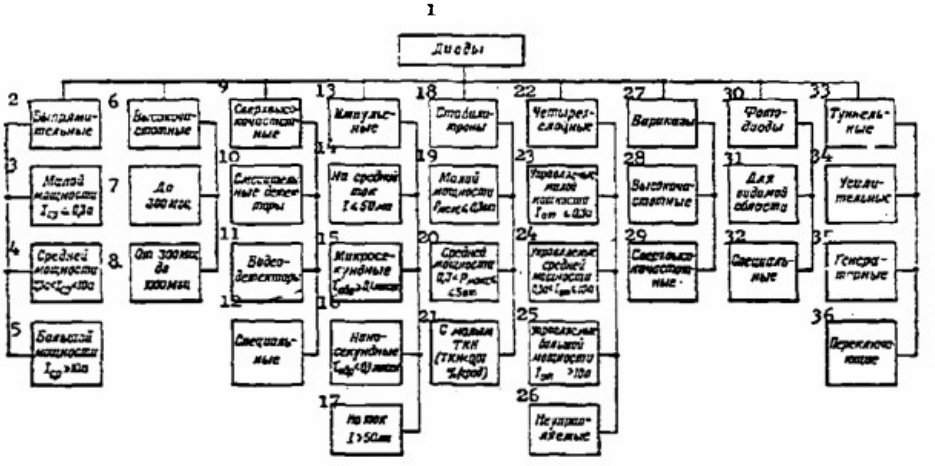


Figure 5.11. Classification of semiconductor diodes.

1 - diodes; 2 - rectifying; 3 - low power $I_{av} < 013$ A; 4 - average power 0.3 A $< I_{av} <$ 10 A; 5 - high power $I_{av} >$ 10 A; 6 - high frequency; 7 0 to 300 MHz; 8 - from 300 MHz to 1000 MHz; 9 - superhigh frequency; 10 - mixing detectors; 11 - video detectors; 12 - special; 13 - pulse; 14 - on average current $I \leqslant 50$ mA; 15 - microsecond $t_{rr} \geqslant 0.1$ μsec; 16 - nanosecond $t_{rr} < 0.1$ μsec; 17 - on current $I > 50$ mA; 18 - stabilitrons; 19 - low power $P_{max} \leqslant 0.3$ W; 20 - average power $0.3 < P_{max} \leqslant 5$ W; 21 - with low TC (TC < 0.10 %/deg); 22 - four layer; 23 - rectifying, low power $I \leqslant 0.3$ A; 24 - rectifying, average power 0.3 A $< I \leqslant$ 10 A; 25 - rectifying, high power $I >$ 10 A; 26 - noncontrolled; 27 - variable capacitors; 28 - high frequency; 29 - superhigh frequency; 30 - photodiodes; 31 - for visible range; 32 - special; 33 - tunnel; 34 - amplifying; 35 - generating; 36 - switching.

Fundamental parameters of superhigh frequency diodes are:

operating wave length, λ [cm];

conversion losses

$$L = 10 \lg \frac{\text{applied power at the signal frequency}}{\text{output power at an intermediate frequency}} \text{ [db]},$$

usually amount to 6.5 to 8.5 db;

current sensitivity $\beta = I_{rect}/P_{in}$ [A/W]; for present-day diodes it reaches 4 A/W and depends on the amount of constant positive bias, whose optimum value for silicon diodes is 0.1 to 0.2 V, and 0.3 to 0.4 V for germanium.

Relative temperature noise t_n is the ratio of noise power created by the diode to the termal noise power of an ohmic resistance, equal to the output impedance of the diode at T = 239°K, usually t_n = 2 to 3.

Special Types of Semiconducting Diodes

Reference diodes (stabilitrons). For a sufficiently large concentration of admistures ($N \approx 10^{19}$ atoms/cm^3) in each of the semiconductor regions, the thickness of the p-n junction is small (to $5 \cdot 10^{-6}$ cm). Then reverse types of electrical breakdown of the p-n junction may take place - avalanche and zener. Current at the breakdown may vary within wide limits with practically unchanged voltage ($I \approx kU^{23}$).

The volt-ampere characteristic of such a diode is shown in Figure 5.12.

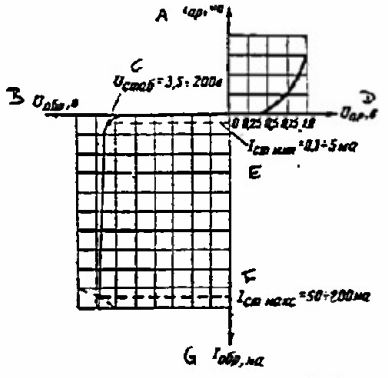


Figure 5.12. Volt-ampere characteristic of a silicon stabilitron.
A - I_f, mA; B - U_{rev}, V; c - U_{stab}; D - U_f, V;
E - $I_{st\ min}$; F - $I_{st\ max}$; G - I_{rev}, mA.

Stabilitrons are made with a silicon base. Their fundamental parameters are:

U_{st} [V] - stabilization voltage, it ranges from 3.5 to 200 V;

$R_d = dU/dI$ [ohm] - dynamic resistance;

$TC = dU_{st}/dt° \cdot 1/U_{st} \cdot 100\%$/degree is temperature coefficient of stabilization voltage, for the majority of diodes TC = 0.2 to 0.5 %/deg;

$I_{st\ min}$; $I_{st\ max}$ are minimum and maximum stabilization currents, usually $I_{st\ min} \approx$ (0.1 to 10)mA; $I_{st\ max} \approx$ (30 mA to 1 A).

Tunnel diodes. With very high concentrations of admixtures in each region of the crystal ($10^{10} - 10^{21}$ cm^{-3}), the thickness of the p-n junction is very small ($d \approx 10^{-6}$ cm). A significant field gradient ($E \geq 10^5$ V/cm) can exist on such a thin junction with only small external voltages. High gradient and small junction thickness allows current carriers to penetrate through the electrical barrier on the junction practically without energy

loss. This called the tunnel effect. The volt-ampere characteristic of such a diode is shown in Figure 5.13.

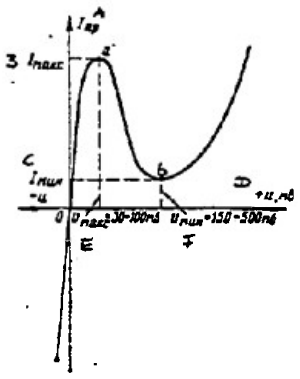


Figure 5.13. Volt-ampere characteristic of a tunnel diode.
A - I_c; B - I_{max}; C - I_{min}; D - +u, mV; E - U_{max}; F - U_{min}.

With a moderate increase in forward voltage (to 60 to 200 mV), the tunnel current of a diode begins to rise (0a part of the characteristic), and then decreases because of a decrease in the number of free energy levels in the p-region to which the electrons of the n-region may transfer by the tunnel effect (ab part). With a further increase in applied voltage, current through the diode rises again, but because of an increase in the usual diffusion current. With reverse biasing, current monotonically and rapidly builds up to the limiting value.

Tunnel current I_{peak} is strongly dependent on the area of the junction. For present-day diodes $I_{peak} \approx$ (0.1 to 50) mA. For germanium diodes, the ratio of currents I_{peak}/I_{valley} = 12 to 14, and for gallim arsenide diodes it reaches 40.

Because of the falling portion on the volt-ampere characteristic (where $R_i < 0$), they may be used for amplifying or generating oscillations (to frequencies of 10^{10} Hz) or for high-speed switching circuits (switching speed to 1 to 2 ns). Tunnel diodes may operate over a wide temperature range (4 to 600°K) and at high radiation levels. A fault of these diodes is the poor reproducibility of parameters in production. In addition, to assure their operation at SHF, the junction area must be made very small ($S \geqslant 50\ \mu^2$) because of large specific capacitance ($C \leqslant 5\ \mu F/cm^2$).

Parametric diodes (variable capacitors). Variation of diode capacitance with applied voltage may be used for parametric amplification and generation of electric oscillations, effecting frequency modulation, network tuning, and other purposes. At the present time special types of diodes, called parametric diodes, have been designed for low-noise amplification at SHF. They are made of silicon, germanium, and gallium arsenide; structurally they may be planar or point-contact. It is primarily important that diode capacitance C_0 permit network tuning at resonance, and that changes in total capacitance of the network ΔC with small amplitudes of the controlling voltage result in sufficiently high modulation coefficients ($m = \Delta C/C_0$).

The relationship of diode capacitance change to change in voltage, determined by the junction structure, is desirably close to linear. Under these conditions, the resistance loss R_s introduced into the resonance system by the diode should be minimum, so that greatest amplification and least noise level is achieved. Decrease in R_s is attained by increasing the donor concentration, and also by introducing special admixtures (gold, silver, nickel). At the present time, diodes are being used in which $C_0 \approx 0.5$ to 2 pF; $R_s \approx 1$ to 5 ohms; $m = 0.2$ to 0.6; $f_{max} \geq 10$ GHz, permitting 20 to 35 db amplification in the 3-cm range with noise coefficients of 1.5 to 2.5 db.

Photodiodes

Photodiodes are designed to convert light signals into electrical signals. Germanium photodiodes are the most widely used. They may operate in two ways:

a) in the absence of an external voltage source. In this mode the photodiode operates as a barrier-layer cell. Intrinsic emf reaches 60-90mV, greatest current goes up to 100 μA (with illumination of 7000 lux and R_L = = 1000 ohms). This mode is characterized by low noise levels, and it is feasibly used in detecting light fluxes and low frequency tracking of light pulses (F_i up to 100 kHz);

b) in the presence of an external reverse voltage. This is the so-called photodiode regime. Here the operating voltage drop on the load is equal to the difference in voltage drop with a dark current (reverse current of the diode I_{dark} = 5-50 μA) and current when the diode is illuminates ($I_{max} \leq 1$ mA). The operating voltages of the supply sources are chosen equal to 3-30 V. Integral sensitivity reaches $K = dI/d\Phi \approx 50$ mA/lu. Maximum spectral sensitivity corresponds to the wave length $\lambda = 1.5$ μ. Limiting repetition frequency of the light pulses is up to 100 kHz.

Switching diodes (thyristors). Switching diodes have a four-layer structure, p-n-p-n. The basis of the switching activity of this structure is the dependence of current gain of the "generating" transistors on the electrical state.

The majority of thyristors is made on a silicon base by alloying, diffusion, or a combination of both methods. A volt-ampere characteristic of such diodes has the form shown in Figure 5.14 in solid lines. As a consequence of the portion of the volt-ampere characteristic with negative resistance (part ab) they may be used in various circuits for automation, computer technology, and others, as switching elements. On-off time may be as short as 10 ns.

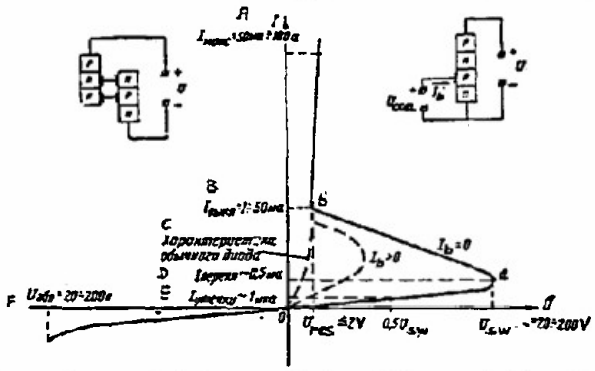


Figure 5.14. Volt-ampere characteristic curve of switching diodes. A - I_{max} = 50 mA-100 A; B - I_{off} = 1-50 mA; C - characteristic curve for a conventional diode; D - I_{sw} ~ 0.5 mA; E - I_{leak} ~ 1 μA; F - U_{rev} = 20-200 V.

Diodes are also made with controlling electrodes. Here a third output is made from one of the internal regions of the structure. By varying the voltage on this electrode, it is possible to vary the reset voltage, as shown in Figure 5.14 in dotted lines.

Numerical values for the parameters of contemporary switching diodes are given on graphs of volt-ampere characteristics.

5.7 Transistors

A transistor is a system of two adjacent electron-hole junctions in the monocrystal of a semiconductor.

At present only planar germanium and silicon transistors are manufactured. Their junctions are made by diffusion or alloying, and also by combining these two methods or by pulling from a melt. Accordingly, differentiations are made between alloyed, diffused, and "grown" transistors.

The different conductivity regions in a semiconductor monocrystal may be arranged p-n-p or n-p-n. The construction and conventional symbols for transistors in the principal circuits are shown in Figure 5.15.

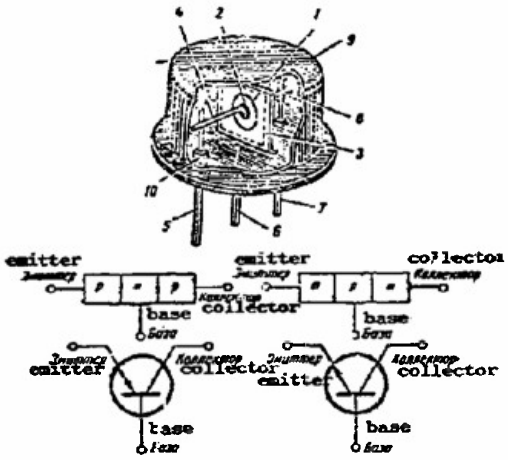


Figure 5.15. Construction and conventional symbols for transistors.

1 - semiconductor crystal; 2 - emitter; 3 - crystal holder (base output); 4 - internal emitter output; 5 - external emitter output; 6 - base output; 7 - external collector output; 8 - internal collector output; 9 - metal casing; 10 - glass insulator.

The left region of the crystal (fig. 5.15) is called the emitter, the center region is called the base, and the right is called the collector.

On amplifying electrical oscillations, bias voltage at the junction between emitter and base drops in the forward direction from the source E_e (tenths of a volt), and, at the collector, junction voltage drops from the source E_c (tens of volts) in the reverse direction, as shown in Figure 5.16.

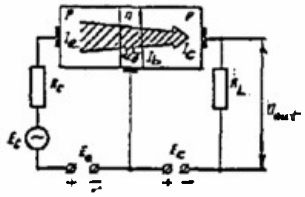


Figure 5.16. Schematic of a p-n-p transistor unblocked for amplifying signals.

Under the influence of U_{eb}, holes from the emitter are injected into the base, causing a current in the emitter I_e. In the base region these carriers will move to the collector because of diffusion. Arriving at the collector, the holes are attracted by the electric field created by E_c at

the junction from the base into the collector region, thus increasing current in the collector junction,

$$I_c = I_{co} + \alpha I_e,$$

where

I_{co} is the initial reverse current in the collector junction;
$\alpha = \Delta I_c / \Delta I_e \Big|_{U_{cb}} = \text{const}$ is the current amplification factor (α = = 0.9 to 0.999).

Because the base width is made very small ($w_b < 50\ \mu$), almost all the holes (90-99.9%) leaving the emitter reach the collector. The remaining holes recombine in the base with electrons. To establish electrical neutrality in the base, electrons enter it from the source E_e. This creates a small current in the base circuit $I_b = I_e - I_c$.

A change in emitter current (at the input of a low impedance circuit) causes a change in collector current approximately equal to it (at the output of a high-impedance circuit). This is the origin of the term "transistor" (transfer of resistor). This is the operating principle of a transistor, an element which may be used to amplify voltage and power.

The physical processes in n-p-n transistors are practically the same as those already discussed (it is only necessary to change polarity of the supply sources E_e and E_c to the reverse and observe the motion of electrons in the base). A transistor may be connected in reverse, where the collector takes the role of the emitter and the emitter becomes the collector. But since the area of the collector junction is approximately twice as large as the area of the emitter junction, while the impurity concentration in the emitter region is chosen considerably larger than in the collector region, current amplification factor ($\alpha_I \leqslant 0.5$) and allowable power dissipation for inverse transistor connection are small. Therefore, such a connection appears infeasible.

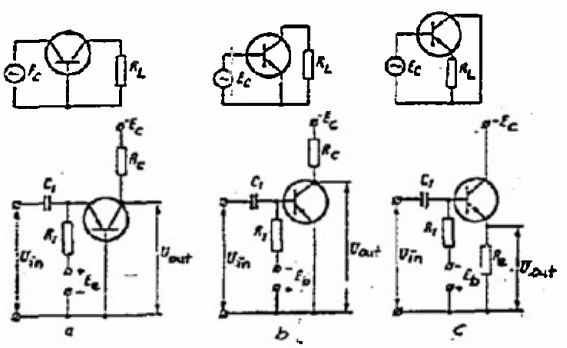


Figure 5.17. Three basic circuits for connecting transistors. a - common base; b - common emitter; c - common collector.

Depending on which electrode of the transistor is common to input and output current (voltage) sources, there are three basic configurations for connecting transistors: common base, common emitter, and common collector (fig. 5.17).

Characteristics and Systems of Characteristic Parameters for Transistors

The fundamental properties of transistors at low frequencies are determined by families of input and output static characteristics. In handbooks, input and output characteristics are usually given for only two transistor circuits: common base and common emitter.

From these characteristics it is possible to construct two more families: characteristics of voltage feedback and transfer characteristics (forward current transmission), but they are rarely used.

At the present time, the most widely used characteristics are those with I_{in} and U_{out} as the independent variables. Graphs of these characteristics take the form shown in Figures 5.18 and 5.19.

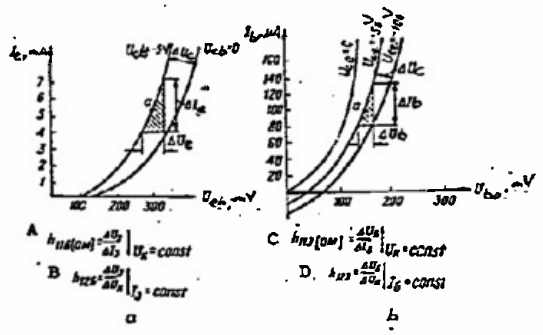


Figure 5.18. Characteristic input curves of transistors.
a - common base; b - common emitter.

A - $h_{11\,b}[\text{ohm}] = \Delta U_e/\Delta I_e \,|\, U_c = \text{const}$;

B - $h_{12\,b} = \Delta U_e/\Delta U_c \,|\, I_e = \text{const}$;

C - $h_{113}\,[\text{ohm}] = \Delta U_b/\Delta I_b \,|\, U_c = \text{const}$;

D - $h_{123} = \Delta U_b/\Delta U_c \,|\, I_b = \text{const}$.

Characteristics for common collector circuits differ only slightly from those for common emitter, and therefore are not used in practice.

For small changes in voltage and currents (small signals) the nonlinear portions of the characteristics may be considered linear and the transistor may be considered as an active linear four-terminal network for varying

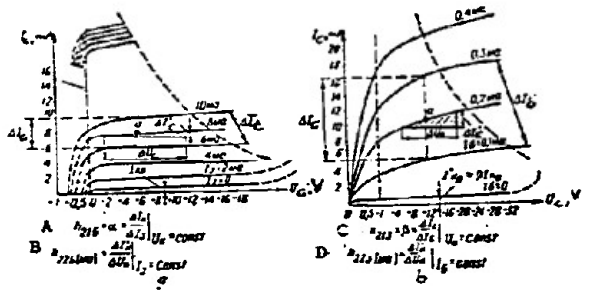


Figure 5.19. Output characteristics of transistors.
a - common base; b - common emitter.

A - $h_{21b} = \alpha = \Delta I_c/\Delta I_e \big|_{U_c = const}$;

B - h_{22b}[megohm] $= \Delta I_c'/\Delta U_c \big|_{I_e = const}$;

C - $h_{213} = \beta = \Delta I_c/\Delta I_b \big|_{U_c = const}$;

D - h_{223} [megohm] $= \Delta I_c/\Delta U_c \big|_{I_b = const}$

components of current and voltages (fig. 5.20). The association between currents and coltages for such a four-terminal network may oe expressed by one of six pairs of linear equations. The systems of characteristic transistor

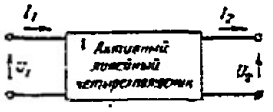


Figure 5.20. Representation of a transistor as an active linear four-terminal network for varying currents and voltages.

I - Active linear four-terminal network.

parameters are differentiated according to the chosen pair of equations. The following three systems are frequently used:

System of Z parameters. For this system

$$\dot{U}_1 = Z_{11}\dot{I}_1 + Z_{12}\dot{I}_2;$$
$$\dot{U}_2 = Z_{21}\dot{I}_1 + Z_{22}\dot{I}_2.$$

System of Y parameters. For this system

$$\dot{I}_1 = Y_{11}\dot{U}_1 + Y_{12}\dot{U}_2;$$
$$\dot{I}_2 = Y_{21}\dot{U}_1 + Y_{22}\dot{U}_2.$$

System of H parameters. For this system:

$$\dot{U}_1 = H_{11}\dot{I}_1 + H_{12}\dot{U}_2;$$
$$\dot{I}_2 = H_{21}\dot{I}_1 + H_{22}\dot{U}_2.$$

At low frequencies the Z, Y, and H parameters are purely active, and are defined as $Z_{11} = r_{11}$; $Z_{12} = r_{12}$; $Z_{21} = r_{21}$; $Z_{22} = r_{22}$; $Y_{11} = g_{11}$; $Y_{12} = g_{12}$; $Y_{21} = g_{21}$; $Y_{22} = g_{22}$; $H_{11} = h_{11}$; $H_{12} = h_{12}$; $H_{21} = h_{21}$; $H_{22} = h_{22}$.

Since with experimental measurements of characteristic parameters the greatest accuracy for small input resistance of the transistor is obtained with no load on the input, but with large output load (short-circuit on the output), most handbooks now give the h-parameters.

Numerical values of the h-parameters (Table 5.4) are also determined from the characteristics, as shown in Figures 5.18 and 5.19. Accordingly,

$$h_{11}[\text{ohm}] = \dot{U}_1/\dot{I}_1 \Big|_{\dot{U}_2=0} = \Delta U_{in}/\Delta I_{in} \Big|_{U_{out}} = \text{const}$$

is the input impedance of the alternating current transistor with the alternating current shorted;

$$h_{12} = \dot{U}_1/\dot{U}_2 \Big|_{\dot{I}_1=0} = \Delta U_{in}/\Delta U_{out} \Big|_{I_{in}} = \text{const}$$

is the open-circuit feedback voltage for alternating current across the input;

$$h_{21} = \dot{I}_2/\dot{I}_1 \Big|_{\dot{U}_2=0} = \Delta I_{out}/\Delta I_{in} \Big|_{U_{out}} = \text{const}$$

is the alternating short-circuited current amplification factor at the output; for circuits with a common base the current amplification factor is designated by α, and by β for the circuit with a common emitter;

$$h_{22}[\text{ohm}] = \dot{I}_2/\dot{U}_2 \Big|_{\dot{I}_1=0} = \Delta I_{out}/\Delta U_{out} \Big|_{I_{in}} = \text{const}$$

is the alternating current output open circuit admittance at the input.

Numerical values of the transistor parameters depend on the circuit, electrical mode, operating frequency, and temperature.

Table 5.4

Typical values of the h-parameters of low-frequency alloyed germanium transistors

1 Параметр	2 Схема с общей базой	3 Схема с общим эмиттером	4 Схема с общим коллектором
h_{11} [ом] 5	40	200	2000
h_{12}	$4 \cdot 10^{-4}$	$16 \cdot 10^{-4}$	1
h_{21}	$\alpha = 0{,}98$	$\beta = \frac{\alpha}{1-\alpha} = 49$	$(\beta+1) = \frac{1}{1-\alpha} = 50$
h_{22} [мкмо] 6	1	50	50
$g_{21}\left[\frac{ма}{в}\right]$ 7	25	25	25

Key: 1 - parameter; 2 - common base; 3 - common emitter; 4 - common collector; 5 - ohm; 6 - μmho; 7 - mA/V.

The advantages and inadequacies of various transistor circuits can be easily seen from Table 5.4. Handbooks, as a rule, usually give the h-parameters for a common base circuit. To determine h_{11}, h_{12}, h_{21}, and h_{22} in two different circuits, it is required to divide the corresponding parameters of the common base circuit by $(1-\alpha)$. The parameter $h_{12c} = 1$, and $h_{12e} = h_{11b}h_{22b}/1-\alpha - h_{12b}$.

If the Z or Y parameters are required, they can be found from the following

$$r_{11} = \frac{h_{11}h_{22} - h_{12}h_{21}}{h_{22}}, \quad g_{11} = \frac{1}{h_{11}};$$

$$r_{12} = \frac{h_{12}}{h_{22}}, \quad g_{12} = -\frac{h_{12}}{h_{11}};$$

$$r_{21} = -\frac{h_{21}}{h_{22}} \quad g_{21} = \frac{h_{21}}{h_{11}};$$

$$r_{22} = \frac{1}{h_{22}}, \quad g_{22} = \frac{h_{11}h_{22} - h_{12}h_{21}}{h_{11}}.$$

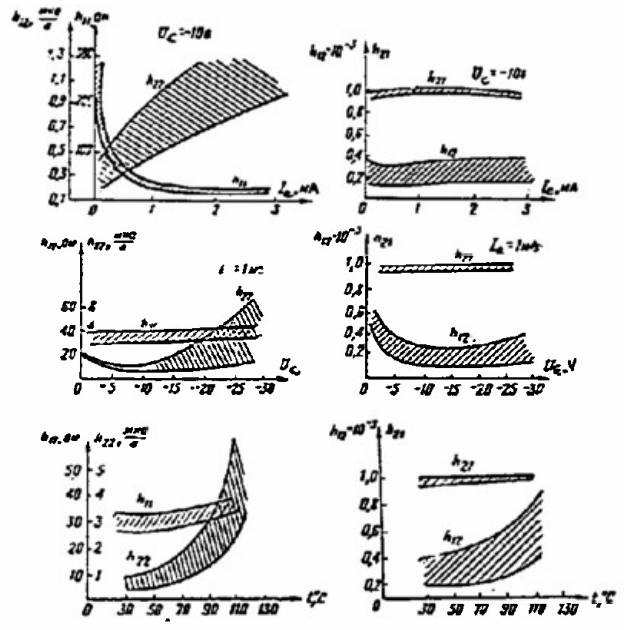


Figure 5.21. Relationship of h-parameters of low-power germanium low-frequency transistors to operating regime and temperature (cross hatching shows the regions of possible variation in parameters from one unit to another among devices of the same type).

The manner in which the h-parameters depend on state, temperature and other factors is quite complex. These relationships for low-power germanium alloyed triodes is shown in Figure 5.21.

In addition to the Z, Y, and H parameters, there is another system of intrinsic (physical) parameters whose numerical values do not depend on the circuit configuration. These parameters correspond to the T equivalent circuit (fig. 5.22).

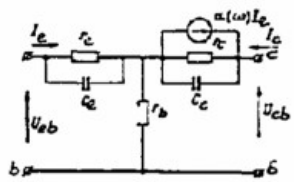


Figure 5.22. Simplest equivalent circuit of a transistor connected with a common base.

r_e - emitter junction resistance; C_e - emitter junction capacitance; r_b - base resistance; r_c - collector junction resistance; C_c - collector junction capacitance; I_e - equivalent current generator.

In this system of parameters are introduced:

the current amplification factor,

$$\alpha = -h_{21}; \tag{5.10}$$

resistance of the p-n junction,

$$r_e = 2h_{11} - 2\frac{h_{12}}{h_{22}}(1 + h_{21}); \tag{5.11}$$

base resistance

$$r_b = r'_b + r''_b = h_{12}/h_{22}/ \tag{5.12}$$

resistance of the collector p-n junction

$$r_c = 1/h_{22}. \tag{5.13}$$

5.8 Frequency and Noise Properties of Transistors

Frequency properties of transistors are basically determined by the time required for the carriers to cross the base region and by the effect of the junction capacitance. They are characterized by a parameter called the specific current amplification frequency, f_α [MHz]. f_α is the frequency at which the modulus of the current amplification factor α decreases by $\sqrt{2}$ in comparison with the value measured at low frequency. Since the capacitance and base width depend on applied voltages and currents, f_α of a transistor depends not only on its construction but also on the operating mode. These relationships are shown in Figure 5.23.

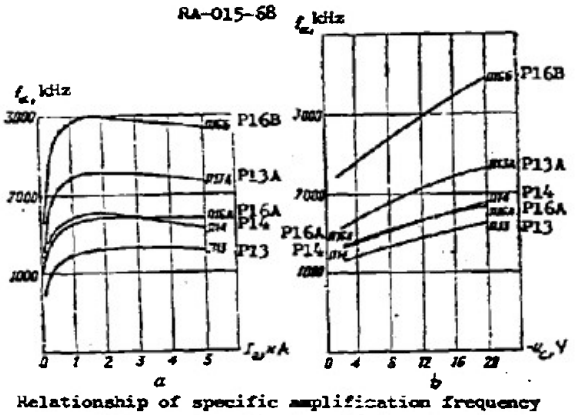


Figure 5.23. Relationship of specific amplification frequency of transistors P13, P14, and P16 on the operating mode. a - plotted against emitter current; b - plotted against collector voltage.

Transistors are presently classed according to their frequency characteristics as low-frequency ($f \leqslant 3$ MHz); average frequency (30 MHz $<$ $< f \leqslant 30$ MHz); high frequency (30 MHz $< f \leqslant 300$ MHz), and superhigh frequency ($f > 300$ MHz).

When using transistors it must be kept in mind that the frequency properties of a transistor in a common emitter circuit are considerably worse than in a common base circuit. The specific amplification frequency of a transistor in common emitter circuit f_β is approximately calculated by the formula

$$f_\beta \simeq 0{,}7 f_\alpha (1-\alpha) \text{ [MHz]}. \tag{5.14}$$

Noise properties of mass production transistors is worse than those of vacuum triodes. Noise depends on the parameters of the transistor (α, r_c, r_b, I_{c0}), its operating mode (I_e, U_c), operating frequency f, temperature, and impedance of the signal source R_g.

These relationships are complex in character. Change of noise factor F_n with changes in emitter current, temperature, R_g, and frequency are shown in Figures 5.24a and 5.24b.

The noise factor of transistors usually is within the limits 5-30 db and may vary considerably from one unit to another of the same transistor type. Noise of a transistor varies inversely with its α and r_c and directly with its r_b and I_{c0}.

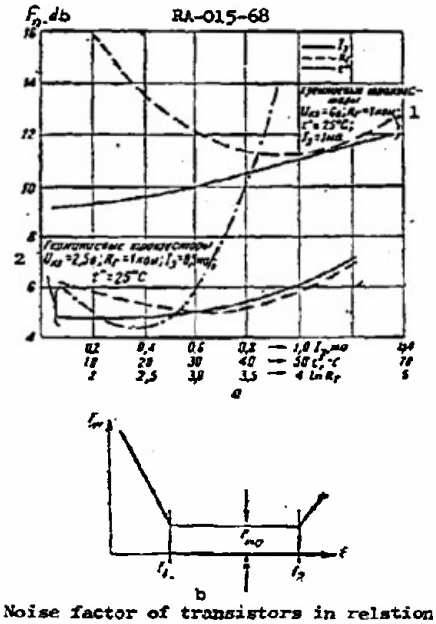


Figure 5.24. Noise factor of transistors in relation to (a) operating mode and temperature, and (b) frequency.

1 - Silicon transistors:
$U_{cc} = 6$ V; $R_g = 1$kΩ; $t° = 25°C$; $I_e = 1$ mA.

2 - Germanium transistors:
$U_{ce} = 2.5$ V; $R_g = 1$ kΩ; $I_e = 0.5$ mA; $t° = 25°C$.

5.9 Operating Modes and Uses of Transistors

It is possible to distinguish three regions on the family of output curves for a transistor (for example, for a common emitter circuit): active, collector current cut-off region, and saturation region (fig. 5.25).

In practice transistors are used in two modes; amplification and switching. In the amplification mode with small signals, the transistor operates only in the active region; with large signals, it operates in the collector current cut-off region and in the active region. In the switching mode, it operates in all three regions. In the active region, the emitter junction is biased in the forward direction, and the collector junction in the reverse direction. Consequently, output current varies proportionally with changes in input current,

$$I_c = \beta(I_{c0} + I_b).$$

The cut-off region (blocked state) is characterized by reverse voltage on both junctions. Under these conditions it is possible to consider that approximately constant reverse currents flow in the circuits of both electrodes of the transistor: $I_{e0} \approx I_{c0}$; $I_{b0} \approx I_{c0} + I_{e0}$.

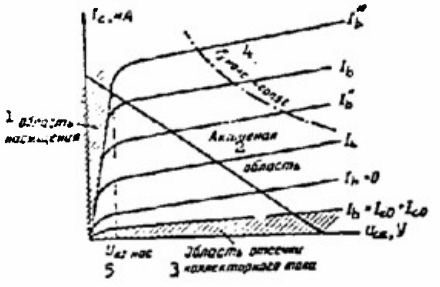


Figure 5.25. Possible transistor operating modes.

1 - saturation region; 2 - active region; 3 - collector current cut-off region; 4 - $P_{c\ max}$ = const; 5 - $U_{ce\ sat}$.

The semiconductor triode is saturated when both junctions are biased in the forward direction. Then further change in base current produces no change in collector current. Residual voltage on the collector, $U_{ce\ sat}$, is small, usually amounting to tenths of a volt. Accordingly, dissipated power on the collector p-n junction is also small, and collector current is determined by the value of load resistance and supply voltage E_c. This results in good switching properties for transistors in pulse circuits.

Inertial properties of transistors depend on the circuit, electrical mode, junction capacitances, and transfer time for the carriers to cross the base.

If the operating point is not in the saturation region, the time constant of the transition process at the transistor output in a common emitter circuit (τ_β) is determined by

$$\tau_\beta \approx \frac{\beta + 1}{5.15 f_\alpha} \text{ [sec]}. \qquad (5.15)$$

If the operating point is in the saturation region, then the processes in the semiconductor triode after the end of the base current pulse consist of two stages (fig. 5.26). During the first stage a decrease in concentration (dissipation) of accumulated holes in the base takes place. This process is characterized by a dissipation time $\tau_{diss} \approx \tau_\beta \ln S$, where $S = I_b/I_{b\ sat}$ is the degree of saturation; then the operating point shifts to the active region, in which collector current decreases exponentially with time constant τ_β. Thus, recovery time from saturation to the blocked state is greater than t_i, which has a negative effect on high-speed pulse circuit.

Transistors are exceptionally reliable elements of electronic circuits, but incorrect use reduces their reliability and is one of the reasons for failures in operation.

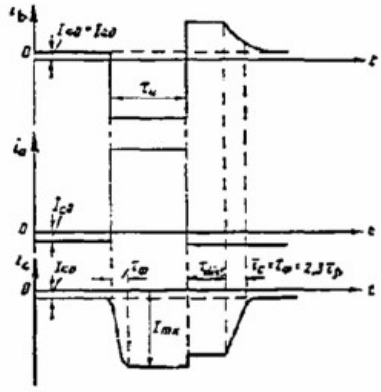


Figure 5.26. Diagrams of transistor currents for pulse operation.

The most frequently encountered examples of incorrect operation are the following:

1. Using devices at the maximum allowable conditions or even exceeding the operational limits. In all handbooks for each device are given maximum allowable currents, voltages, and power, taking into account the corresponding safety factor only on technological breakdown. These limits should not be exceeded in any static, dynamic, or non-steady state which might occur in circuits.

2. Disregard for the thermal state of the transistor. Since at increased temperatures the allowable voltage, current and power limits are reduced, almost all of the transistor parameters change. For example, reverse collector current on the average doubles with each 10°C rise in temperature. Therefore, special attention should be paid to cooling without regard to how far the dissipated power is from the allowable limit. Germanium transistors operate in a temperature range from -60 to +70°C, and silicon from -60 to +120°C. Transistors should only be soldered into a circuit in such a manner as to minimize heating of the junctions.

3. Using transistors in a common emitter circuit with large resistance in the base circuit. This resistance should be minimum. Avoiding the transistor output from the arrangement does not assume isolation of the base circuit in the presence of voltages on the other electrodes.

4. Securing the transistor by its leads. The transistor is resistant to mechanical stress only when it is mounted to the chassis by its casing. This also assures the best heat conduction and generally increases stability and reliability of operation.

Chapter VI

Pulse Engineering

6.1 Pulse Voltages and Currents

An electrical pulse is understood to mean the voltage (current) acting in a circuit for a short time period, short as compared with the transient processes taking place in that circuit.

Video pulses and radio pulses are different, in that video pulses are voltages (currents), the instantaneous values of which differ from zero, or from a constant level, for a short period of time (fig. 6.1a). Radio pulses are pulses of high-frequency sine voltage (current) (fig. 6.1b). The envelope of a radio pulse is a video pulse.

Pulses can be of different shapes, depending on the use for which intended. Most often used in radar work are square (fig. 6.1c), sharp (fig. 6.1d), sawtooth (fig. 6.1e), and trapezoidal (fig. 6.1f, g) pulses.

Video pulses can have positive and negative polarities.

The principal parameters of a single pulse are its amplitude $U_m(I_m)$, width τ_p, width of front τ_f, and cut-off τ_{co} (fig. 6.1a). Conventionally, the pulse width is measured at the level of 0.1 $U_m(I_m)$, the width of the front can be determined as the time required for the voltage (current) to rise from 0.1 $U_m(I_m)$ to 0.9 $U_m(I_m)$, and the cut-off width as the fall from 0.9 $U_m(I_m)$ to 0.1 $U_m(I_m)$. The reading level can sometimes be selected at other levels, but in any case is always stipulated.

The principal parameters of a periodic sequence of pulses are the repetition (recurrence) period, T_p (fig. 6.1a), the repetition (recurrence) frequency, $F_p = 1/T_p$, the pulse duty factor

$$Q = T_p/\tau_p = 1/F_p\tau_p \tag{6.1}$$

and its reciprocal, the duty cycle

$$K_d = 1/Q = \tau_p/T_p = F_p\tau_p. \tag{6.2}$$

Yet another parameter of the radio pulse is the carrier frequency, f_c.

6.2 Pulse Formation by Linear Electrical Circuits

Rheostate type capacitive circuits, RC, and rheostate type inductive circuits, RL, for differentiating and integrating pulses, impulsing circuits, and lines with distributed and lumped parameters, are all used to form pulses.

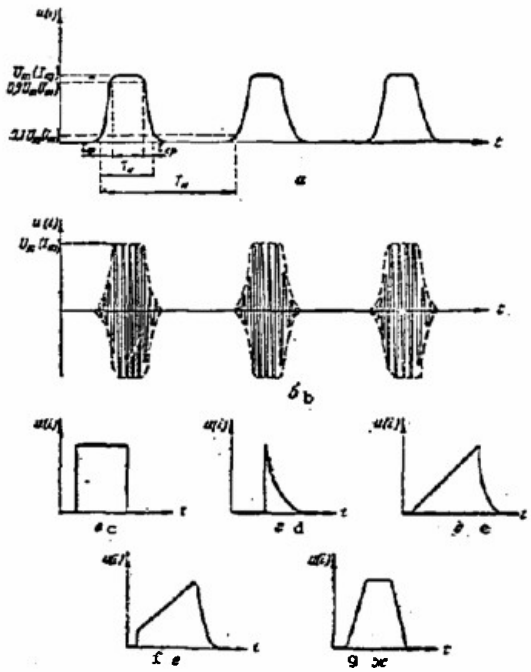


Figure 6.1. Video and radio pulses.

a - video pulses; b - radio pulses; c - square video pulse; d - sharp video pulse; e - sawtoothed video pulse; f and g - trapezoidal video pulses.

Differentiating circuits

A differentiating circuit is one in which the voltage across the output is proportional to the derivative of the input voltage

$$U_{out} = k \cdot dU_{in}/dt.$$

Capacitive (fig. 6.2a) and inductive (fig. 6.2b) circuits can be used for differentiation. The capacitive differentiating circuit is the one most widely used in practice. The smaller the time constant ($\tau = RC$ for the capacitive, $\tau = L/R$ for the inductive circuit) the more precise the differentiation, but the smaller the magnitude of the output signal. Hence, the magnituce of the time constant is selected in accordance with concrete engineering conditions.

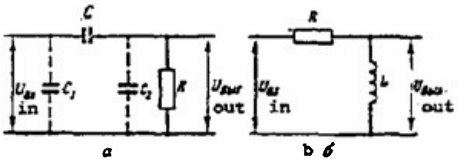


Figure 6.2. Differentiating circuits.
a - capacitive; b - inductive.

Differentiating circuits are used for the following operations.

1. Carry out the mathematical operation of differentiation. In the electronic computer this use of the differentiating circuit is for continuous action. A combination of a differentiating circuit and an amplifier (an operational differentiating amplifier) is used in this case to improve the quality of differentiation when the output signal has sufficient amplitude.

2. Formation of sharp pulses from the edges of the voltage. When positive square pulses (fig. 6.3a) are fed into the input of the differentiating circuit, sharp pulses (positive for the time of the pulse front, negative for the time of the pulse cut-off) are obtained at the output. Formation of sharp pulses is connected with the rapid charge (discharge) of the capacitance. The smaller the time constant, the shorter the pulse. When the pulse width is measured at the 0.1 U_m level

$$\tau_p = 2.3\ \tau = 2.3\ RC. \qquad (6.3)$$

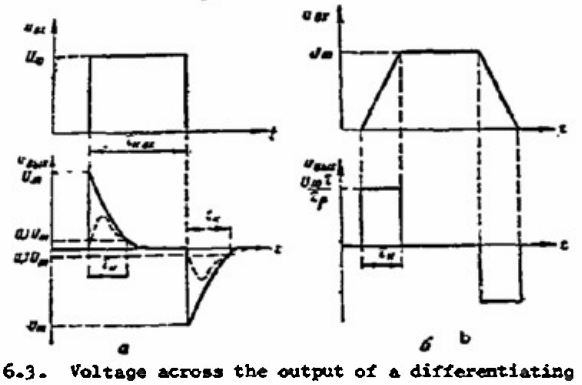


Figure 6.3. Voltage across the output of a differentiating circuit.
a - in the case of differentiation of a square pulse;
b - in the case of differentiation of a trapezoidal pulse.

The shape of the output pulse shown in Figure 6.3a, occurs only under ideal conditions; infinitely short width of the input signal front, zero generator resistance, no shunt capacitances.

A real pulse at the output of the differentiating circuit has less amplitude, is wider, and has a finite width of front (the dotted line in fig. 6.3a). Capacitance C is selected 4 to 5 times larger than C_1 and C_2 (fig. 6.2a) in order to reduce the effect of the shunt capacitances. Ordinarily, C_1 and C_2 are 10 to 15 pf, with the magnitude of $C \geqslant 50$ to 100 pf.

The obtaining of sharp pulses is sometimes called pulse contraction, and the circuit used a contracting circuit. Use of the contracting circuit makes it possible to obtain pulses with a width of $\tau_p \geqslant 0.5$ to 1 microseconds. Pulse contraction can be used to form short-duration pulses, for selecting pulses according to width, and thelike.

3. Formation of square pulses from trapezoidal. When a trapezoidal pulse is fed into the input of a differentiating circuit, there will be two square pulses at the output; positive during the time of the front, negative during the time of input pulse cut-off (fig. 6.3b). Thus, the differentiating circuit can be used to obtain the sensitizing pulse in a scope with raster scan.

Integrating circuits

An integrating circuit is one in which the voltage across the output is proportional to the integral of the input voltage

$$U_{out} = k \int_0^t U_{in} dt.$$

Used for the most part for the integration is the RC, capacitive, circuit (fig. 6.4a). The RL, inductive, circuit (fig. 6.4b) is practically not used.

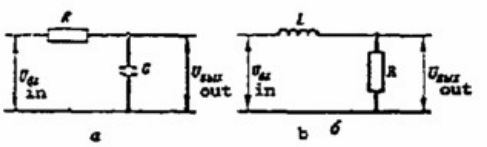


Figure 6.4. Integrating circuits.
a - capacitive; b - inductive.

The larger the integrating circuit's time constant, the better the quality of integration, but the smaller the magnitude of the output signal.

Principal uses of the integrating circuit are to carry out the mathematical operation of integration (as an element in an operational integrating amplifier), to make a sawtooth pulse out of a square pulse (fig. 6.5a) to form the time-base sweeps, for selecting pulse widths, and for obtaining the parabolic voltage from the sawtooth voltage (fig. 6.5b) to form the height scan in the indicator of an altimeter system.

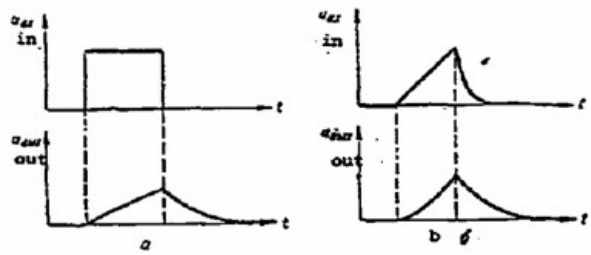


Figure 6.5. Voltage across the output of an integrating circuit.
a - integration of a square pulse; b - integration of a sawtooth pulse.

Systems with Impact Excitation Circuits

Circuit excitation caused by the current drop can be used in these systems. Systems with impact excitation circuits are used to obtain a series of sinusoidal oscillations, short sharp and square pulses, for pulse expansion, etc.

Obtaining series of sinusoidal oscillations. Positive feedback circuits are most widely used (fig. 6.6a). Upon initial operation tube T_1 is triggered, the place current in the tube flows through the coil and stores energy in the magnetic field of the coil. When tube T_2 receives an input pulse it is blocked and oscillations with frequency $f = 1/2\pi\sqrt{LC}$ appear in the circuit. These oscillations are repeated by the cathode follower, so energy is available in the circuit. When resistor R_3 is of the optimum magnitude the energy losses in the circuit are compensated for and the amplitude of the oscillations remains practically constant (fig. 6.6b). Connection point "a" is usually selected at the middle of the coil. Now the optimum magnitude of $R_{3\ opt}$ can be established through the formula

$$R_{3\ opt} = 1/4 \cdot 1/CR. \qquad (6.4)$$

This circuitry is widely used in calibrators to obtain the range markers.

Obtaining short sharp pulses. Here the critical condition is used, and the circuit can be shunted by resistance

$$R = \frac{1}{2}\sqrt{\frac{L}{C}}. \qquad (6.5)$$

Delivery of a negative pulse (fig. 6.7a) results in the formation of two sharp pulses (fig. 6.7b), similar in shape to the pulse from the differentiating circuit, at the circuit output. This is why this circuit is often called an inductive differentiating circuit. The principal advantage

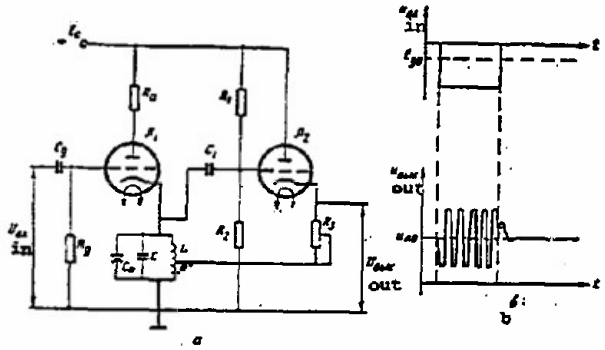


Figure 6.6. Arrangement with impact excitation circuit and positive feedback.

a – schematic; b – voltage time diagrams.

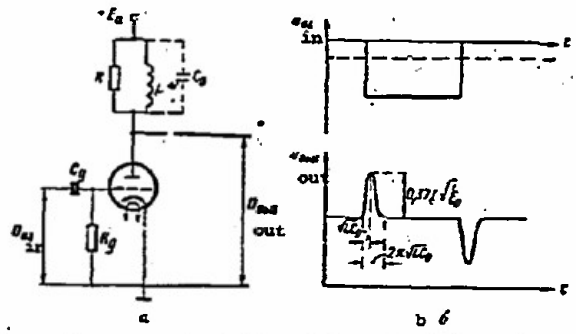


Figure 6.7. Stage with inductive differentiating circuit.

a – schematic; b – voltage time diagrams.

of the circuit is the possibility of obtaining a large amplitude pulse, considerably in excess of the source voltage E_a.

Forming Lines

Distributed-parameter lines, as well as artificial lines with lumped parameters, can be used to form short square pulses. The advantages of these devices include good shape and a highly stable pulse width, and the possibility of obtaining high power pulses.

The line can be used as an energy bank for purposes of obtaining powerful pulses. The line is charged by closing the key (fig. 6.8a). The charging time is

$$t_{charge} = 3(R + R_{load})lC, \qquad (6.6)$$

where

l is the line (cable) length, meters;

C is the line capacitance (capacitance per meter of length).

The line discharges to load $R_{load} = \rho = \sqrt{L/C}$ at the moment the key is closed (a thyratron is often used as the key). Waves of voltage and current are propagated on the line at velocity $v = c/\sqrt{\varepsilon}$ (c is the speed of light; ε is the dielectric constant of the cable dielectric). By time $T_d = l/v$ the voltage (current) wave reaches the end of the line and is reflected from the open end of the line to be propagated to its origin at the same velocity. When the reflected wave arrives at the origin of the line the line is completely discharged. A pulse with amplitude E/2 and width

$$\tau_p = 2T_d = 2\frac{l}{c}\sqrt{\varepsilon} = 6{,}6\cdot 10^{-9} l\sqrt{\varepsilon} \qquad \text{[microseconds]} \quad (6.7)$$

is formed at the load.

The pulse shape is dostorted (fig. 6.8b) when there is a line mismatch ($R_{load} > \rho$ or $R_{load} < \rho$).

The principal advantage of using a coaxial cable (RK-1, RK-3, RK-20, and others) for forming is the undistorted pulse shape (when the cable is matched to the load). Disadvantages are low characteristic impedance (ρ = 50 to 150 ohms) and long length (in order to obtain T_d = 1 microsecond a length of l = 200 to 300 meters is required). One of the conductors is twisted into a spiral (spiral lines RKZ-400, RKZ-401, and others) to reduce the length of the cable. These lines provide a delay of 1 microsecond per 10 cm of cable length at a characteristic impedance of about 1,000 ohms. The principal shortcoming of the spiral line is the extensive distortion of the pulse shape. Spiral lines usually have a delay of $T_d \leqslant 1$ microsecond.

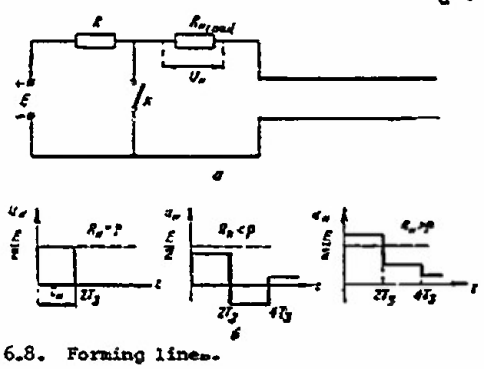


Figure 6.8. Forming lines.
a - schematic; b - voltage time delays.

Artificial chain lines (fig. 6.9a) are those most widely used because they are compact, and because different characteristic impedances (from tens of ohms to several kilohms) and delays (from thens of a microsecond to several tens of microseconds) can be obtained. Chain lines consist of series-connected cells of inductances and capacitances.

The characteristic impedance ρ_{line} and the delay time in one link can be established through the formulas

$$\rho_{line} = \sqrt{L_{cell}/C_{cell}}\ ; \tag{6.8}$$

$$T_{d\ 1} = \sqrt{L_{cell}/C_{cell}} \tag{6.9}$$

The width of the pulse that is formed is

$$\tau_p = 2T_d = 2n\sqrt{L_{cell}C_{cell}}\ , \tag{6.10}$$

where

n is the number of cells.

Increasing the number of cells results in a shortening of the front (fig. 6.9b). For a given front width

$$n = 0.27 \cdot \tau_p/\tau_f. \tag{6.11}$$

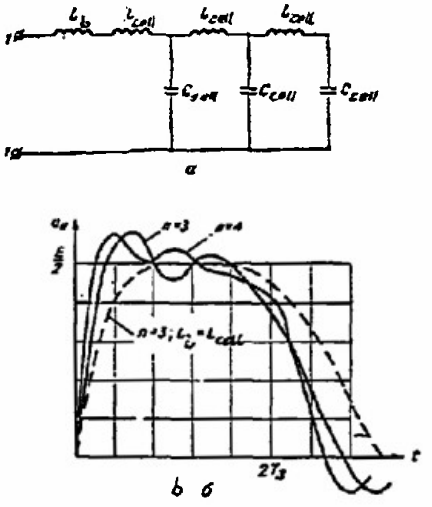


Figure 6.9. Artificial chain line.
a - schematic; b - voltage time diagrams.

However, an increase in the number of cells does not do away with the oscillations at the apex of the pulse. Correction is used to improve the shape of the apex, and that most often used is a booster inductance, L_b. When $L_b = L_{cell}$ there are virtually no oscillations at the apex, but the width of the front doubles (the dotted line in fig. 6.9b).

Forming lines are used in the pulse modulators of radars, in coding and decoding devices, for stabilization of pulse generators.

6.3 Nonlinear Electrical Circuits Used for Pulse Formation

The parameters R, L, and C depend on the magnitude of the voltage (current) in nonlinear forming circuits.

Clippers

A clipper is a device, the voltage across the output of which remains constant when the input voltage goes beyond the limits of a fixed level, the quantizer threshold. When the values of the input voltage are inside the quantizer threshold the voltage should be reproduced across the output without distortion. Limitation of the upper part of the input voltage results in top clipping, limitation of the lower results in bottom clipping, and when limitation is of the upper and lower parts in bilaterial clipping.

Clippers are used to obtain square pulses from a sine voltage, for amplitude and polarity selection, and for improving pulse shape. Elements in which the resistance very definitely depends on the regime are used for clipping. Diodes, multi-electrode tubes, and transistors are used as nonlinear elements such as this.

Diode clippers. The principles of operation of diode clippers are based on unidirectional conductivity of the diode. They are used in electrovacuum and semiconductor diodes. The advantages of the vacuum diode include a high value of back resistance and permissible voltage, and a low temperature ratio. Shortcomings include large size, less efficiency and reliability (than semiconductor diodes), dependence of the characteristic on the filament voltage, and high inter-electrode capacitance. Today semiconductor diodes are used in clippers, for the most part. There are series and parallel diode clipper circuits, depending on the method used to cut in the diode.

In the series diode clipper circuit (fig. 6.10) the diode is unblocked when the voltage is positive and the voltage at the output will repeat the input voltage, whereas, when the voltage is negative the diode will be blocked and the output voltage will be clipped. The clipping level can be changed by changing the magnitude and polarity of the bias source. The advantate of this arrangement is the clearly defined clipping level. The disadvantage is the capacitive coupling between the input and output because of the plate-cathode capacitance, causing distortion in the shape of the output voltage, particularly when the front is quite steep.

In the parallel circuit of the diode clipper (fig. 6.11), when the diode is unblocked the load resistance R_{load} is shunted by the low resistance of the diode and the output voltage is limited, while if the diode is blocked

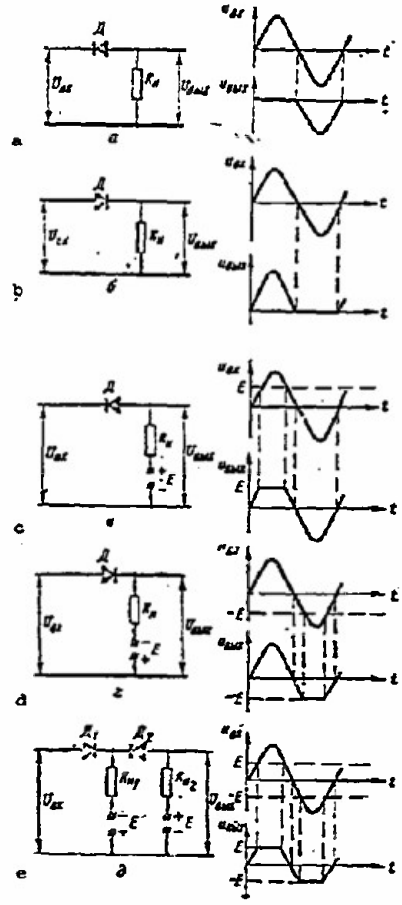


Figure 6.10. Clippers with series connected load.

a - top clipper to zero level; b - bottom clipper to zero level; c - top clipper to positive level; d - bottom clipper to negative level; e - bilateral level.

the voltage across the output repeats the shape of the input voltage. The condition $R_{load} \gg R_{clip} \gg R_b$ must be satisfied if the circuit is to function normally. The absence of capacitive coupling between the input and output is the advantage of this circuit, whereas the disadvantages are the need for a bias source with low internal resistance, less definition of the clipping, and attenuation of the output signal caused by the drop in voltage across R_{clip}.

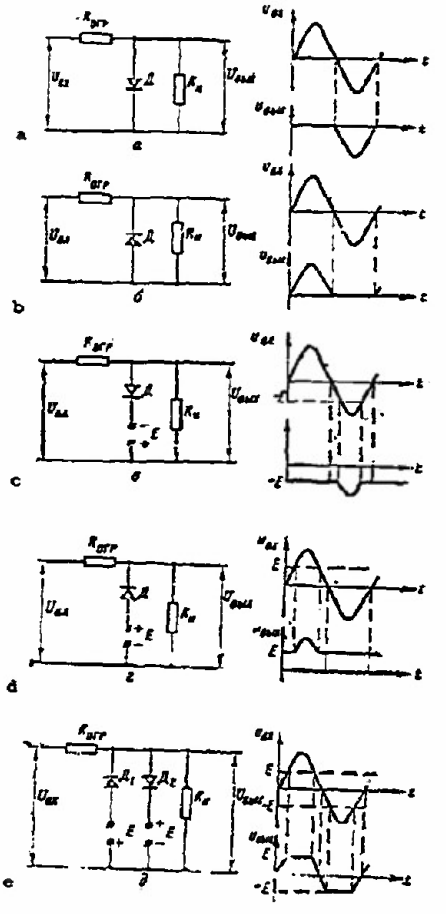


Figure 6.11. Clippers with parallel connected load.

a - top clipper to zero level; b - bottom clipper to zero level; c - top clipper to negative level; d - bottom clipper to positive level; e - bilateral clipper.

The advantages of all diode clipper circuits are simplicity, efficiency, and stability of the clipping level. Lack of amplification is a shortcoming.

Clipping Amplifiers

Multi-electrode tubes, or transistors, can be used to obtain clipping and amplification. Three types of clipping are possible, as a result of the flow of grid currents (grid-circuit clipping), as a result of the lower and upper bending of the plate-grid curve (plate-circuit clipping top and bottom).

Grid-circuit clipping. The circuit describes an amplifier with resistance R_{clip} inserted in its grid circuit (fig. 6.12a). When the voltage across the grid is positive the voltage across the amplifier input is clipped by the flow of grid current through this resistance similar to the parallel diode clipper (fig. 6.12b). The voltage clipped in the grid circuit is amplified and inverted in the plate circuit.

Bottom plate-circuit clipping is obtained in the amplifier (fig. 6.13a) by plate current cutoff. When the input voltage drops below E_{g0} the tube is blocked, the voltage across the plate reaches the magnitude E_a, and is clipped (fig. 6.13b).

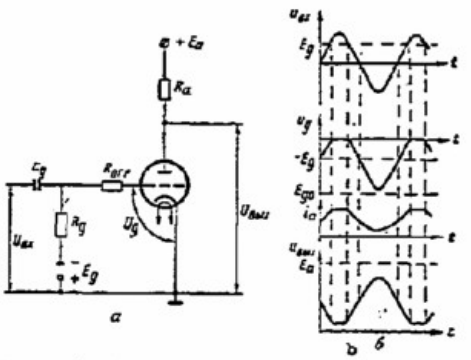


Figure 6.12. Grid-circuit clipper.
a - schematic; b - voltage and current time diagrams.

Top plate-circuit clipping. A pentode and a high-resistance plate load must be used to obtain this clipping. The clipping results from the transition of the pentode into the critical condition, and this establishes the minimum voltage across the plate (fig. 6.14).

A combination of bottom plate-circuit clipping with grid-circuit clipping, or top plate-circuit clipping, is used to obtain bilateral clipping. A transistor can be used for unilateral or bilateral clipping (fig. 6.15). Top clipping is by blocking the transistor, while bottom clipping results from the transition to the saturation condition (for transistors of the p-n-p type).

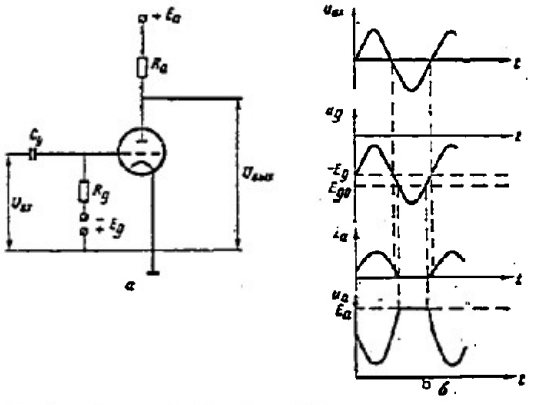


Figure 6.13. Bottom plate-circuit clipper.

a - schematic; b - voltage and current time diagrams.

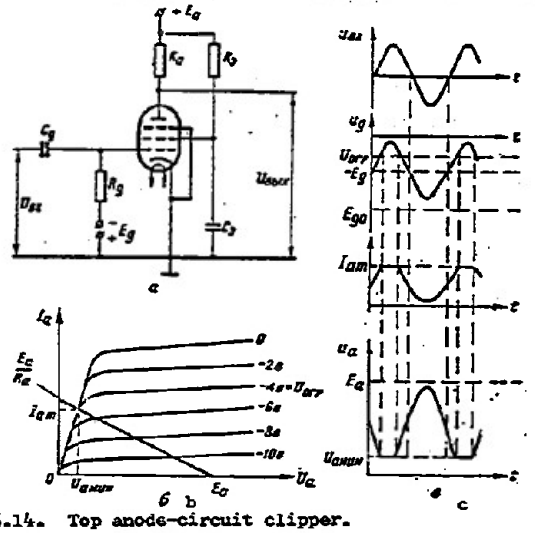


Figure 6.14. Top anode-circuit clipper.

a - schematic; b - voltage and current time diagrams.

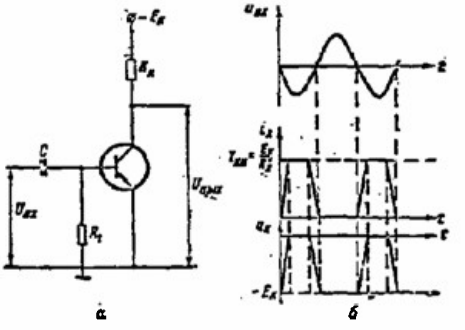


Figure 6.15. Bilateral transistor limiter.
a - schematic; b - voltages and current time diagrams.

Level Clampers

Separation circuits (RC circuits and transformers) will not pass the direct components (fig. 6.16a), and level clampers make d.c. restoration possible. A diode is usually used for this purpose (fig. 6.16b). In the interval between the pulses the condenser quickly discharges through the diode, eradicating the output voltage biasing and changing the zero level. A source of biasing voltage must be cut in in series with the diode to change the clamping level. Clampers are used in the output stages of the sweeps of scopes, for background restoration in television sets, and the like.

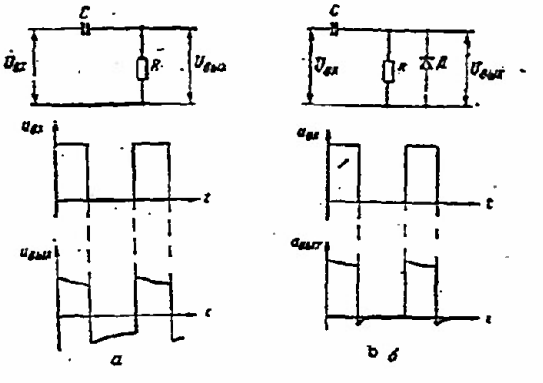


Figure 6.16. Level clampers. a - change in the initial level as pulses pass through an RC circuit; b - clamper circuit and output voltage time diagram.

Circuits with Nonlinear Magnetic Elements

The cores of nonlinear magnetic elements for the most part are manufactured from ferrites with a rectangular hysteresis loop (fig. 6.17). Circuits with nonlinear magnetic elements are used to form pulses (peak transformers), or as circuits with two stable states.

Peak transformers (fig. 6.18a) are used to convert the sine voltage into short pulses. The operating condition is selected such that the transformer is saturated for a good part of the period. The voltage is zero across the output. Short pulses (fig. 6.18b) appear at the output as the transformer shifts from one saturated state to the other. Advantages of the peak transformer are reliability, simplicity, compactness, great steepness of output pulse cut-off. Shortcomings are not very steep pulse edges and small amplitude output pulses.

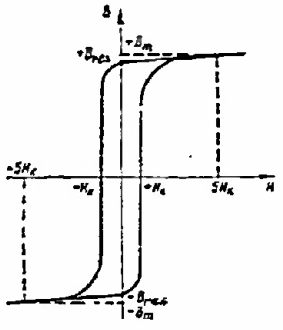


Figure 6.17. Hysteresis loop of a ferromagnetic material.

Circuits with two stable equilibrium states. The circuits are based on the presence of two states of residual magnetization of the core (toroid), $+B_{res}$ and $-B_{res}$ (fig. 6.17). The basis of the circuit is the toroid with several windings (fig. 6.19a).

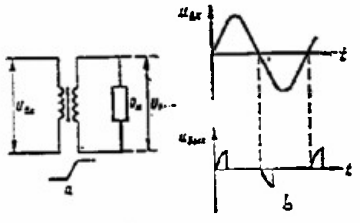


Figure 6.18. Peak transformer.
a - schematic; b - voltage time diagrams.

If the magnetic state of the toroid is established by the point $-B_{res}$, when a pulse current is fed to winding I the core undergoes a reversal of magnetism to $+B_m$. When the pulse current ceases the toroid enters the second

stable state, one that corresponds to the residual magnetization $+B_{res}$. Now, if a pulse current is fed to winding II the core undergoes a reversal of magnetism to $-B_m$, and after this pulse subsides the toroid returns to the original state of residual magnetization $-B_{res}$.

An emf pulse is induced at the output winding each time the toroid undergoes a reversal of magnetism (fig. 6.19b).

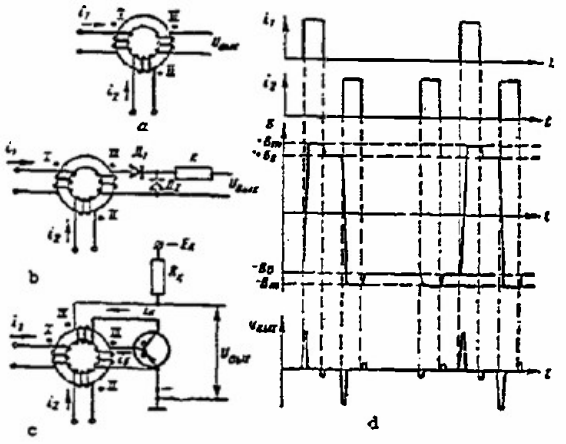


Figure 6.19. Ferrites used in circuits.

a - ferrite toroid; b - time diagrams of currents, voltages, and remanence; c - ferrite-diode element; d - ferrite-transistor element.

If the toroid had been in state $+B_{res}$ prior to feeding the pulse to winding I, upon delivery of the induction pulse the increase is to $+B_m$, and after the pulse has died out the toroid returns to the original state, $+B_{res}$. The toroid does not undergo a reversal of magnetism and the emf pulse induced in the output winding is small in amplitude. Nor does reversal of magnetism occur with the delivery of pulse current to winding II if prior to the delivery the magnetic state of the toroid had been established by the remanence, $-B_{res}$.

Diodes (ferrite-diode elements), or transistors (ferrite-transistor elements), are inserted as couplings between the cores to provide unidirectionality of information. Insertion of transistors also provides amplification of the input signal. Other elements (such as transfluxors, "small ladders," and the like), are made of ferrites, in addition to toroids. Ferrite circuits are widely used as switching, logical, and memory elements.

6.4 Square Pulse Generation

These generators use devices with a falling section on the volt-ampere curve. It is a known fact that on this section the device's equilibrium state is unstable, resulting in the voltage (current) jumps needed to form square pulses. Positive feedback, or devices with a falling section on their curves (thyratrons, tunnel diodes, secondary emission tubes, and the like), are used to obtain this curve.

Pulse generators can function in the following modes:

autooscillation, when generation occurs without outside excitation;

synchronization, when outside signals controlling the operation of the generator are fed to the generator functioning in the autooscillation mode;

start-stop (driven), when the generation is the result of the effect of the outside signal only.

Multivibrators as Generators

In generators of this type the positive feedback is created by using two-stage amplifiers in which the output from one stage is coupled to the input of the other.

Multivibrators operating in the autooscillation mode. The principal circuit in the multivibrator (fig. 6.20a) is a two-state resistance-coupled amplifier, in which the output from each of the amplifiers is capacitance-coupled to the input of the other through a separation capacitor. Generation involves sequential blocking of one tube and firing the other (a process known as inverting the circuit). After sequential inversion the capacitor connected to the plate of the triggered tube discharges through this tube and the leakage resistance, creating a negative voltage across the grid of the blocked tube, keeps this tube blocked. When the capacitor discharges the voltage across the grid of the blocked tube increases and when it reaches a magnitude equal to that of the blocking voltage the tube fires, the positive feedback circuit is restored and the next inversion, the blocked tube is triggered, the triggered tube is blocked, takes place in the circuit.

Square pulses are formed at the tube plates (fig. 6.20b). The width of the pulses, and their repetition frequency, can be established through the formulas

$$\tau_{p\,1} = R_1 C_2 \ln U_{m2}/|E_{g0}|; \qquad (6.12)$$

$$\tau_{p\,2} = R_2 C_1 \ln U_{m1}/|E_{g0}|; \qquad (6.13)$$

$$T_p = \tau_{p\,1} + \tau_{p\,2}. \qquad (6.14)$$

Pulse width and repetition frequency are regulated by changing the magnitudes of the leakage resistances (R_1 and R_2), or of the capacitance.

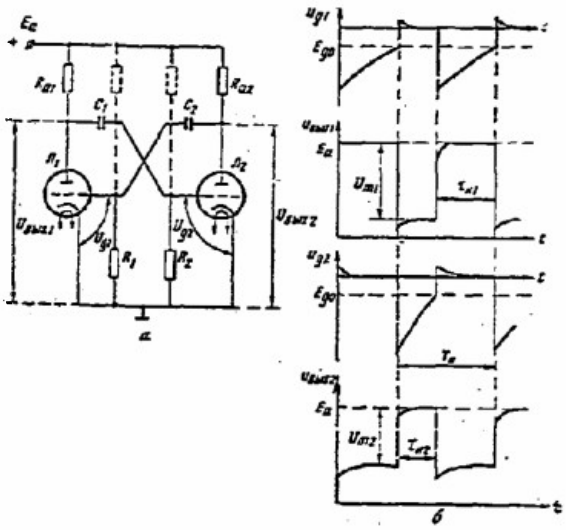


Figure 6.20. Vacuum tube multivibrator.
a - schematic diagram; b - voltage time diagrams.

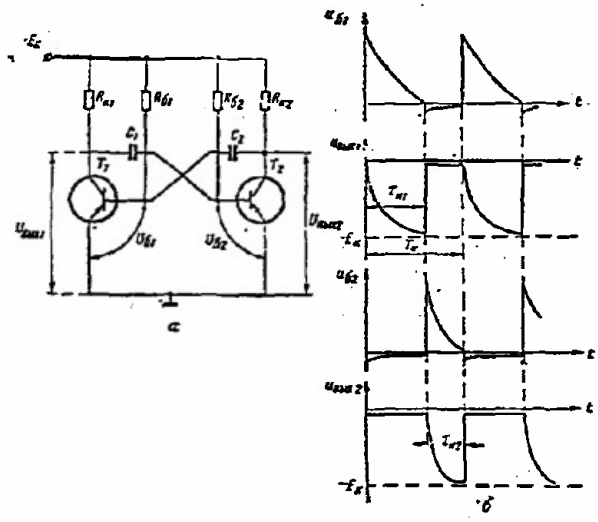


Figure 6.21. Transistor-type multivibrators.
a - schematic diagram; b - voltage time diagrams.

Minimum multivibrator pulse width is 0.5 to 2 microseconds, the oscillation frequency is from a fraction of a hertz to hundreds of thousands of hertz, and the pulse rate can reach 100. Frequency stability is low and is but a few percentage points in the best of cases.

A simple way in which to increase stability is to use a multivibrator with "positive" grids (the dotted line in fig. 6.20a shows the leakage resistances cut in in this circuit). Other stabilization methods will be reviewed in what follows.

The transistor-type multivibrator (fig. 6.21) is no different, in principle, than the tube type, but as a practical matter bias is always fed to the base in the case of the transistor generators to increase stability.

Pulse width, repetition frequency, and amplitude of pulses are calculated through the formulas

$$\tau_{p\,1} = R_{b1}C_2 \ln \frac{2 + I_{c0}R_{b1}/E_c}{1 + I_{c0}R_{b1}/E_c}\ ; \tag{6.15}$$

$$\tau_{p\,2} = R_{b2}C_1 \ln \frac{2 + I_{c0}R_{b2}/E_c}{1 + I_{c0}R_{b2}/E_c}\ ; \tag{6.16}$$

$$T_p = \tau_{p\,1} + \tau_{p\,2}; \tag{6.17}$$

$$U_{m1} \approx U_{m2} \approx E_c. \tag{6.18}$$

The dependence of the initial current, I_{c0}, on the temperature is in the main determined by the temperature stability of the pulse repetition period. In order to increase the stability it is desirable, insofar as this is possible, to increase E_c, reduce resistance R_b, and select transistors with small current values, I_{c0}. One way to increase stability is to insert a base of silicon diodes in the circuit. Instability of the repetition period can be reduced to a few percentage points by the use of these measures.

Triggered multivibrator.* Triggered multivibrators are understood to be multivibrators with a single stable equilibrium condition, producing a single pulse triggered by an outside signal. Triggered multivibrators are used as pulse expanders, strobe pulse generators, time delay devices, and the like.

There are many circuit designs for the triggered multivibrator. Specifically, we can make a triggered multivibrator by blanking one of the

* These circuits are also called single-pulse multivibrators, single-vibrators, trigger circuits, kipp relays.

multivibrator tubes (fig. 6.20a) by an outside biasing source. Most frequently used is the triggered multivibrator with cathode coupling (fig. 6.22a). The advantages of this circuit are stability, simplicity, absence of a negative bias source, good pulse shape at the plate of tube T_2. In its initial state tube T_1 is blocked by voltage U_c created by the current from the unblocked tube, T_2, at resistance R_c. Upon delivery of the trigger pulse tube T_1 is unblocked and T_2 is blocked. Condenser C, discharging through unblocked tube T_1, provides a negative voltage across resistance R, keeping tube T_2 blocked. As the condenser discharges this voltage increases and when it reaches the blocking voltage tube T_2 is unblocked and tube T_1 is blocked. The original condition of the circuit is restored (fig. 5.22b) after condenser C is charged from source E_a.

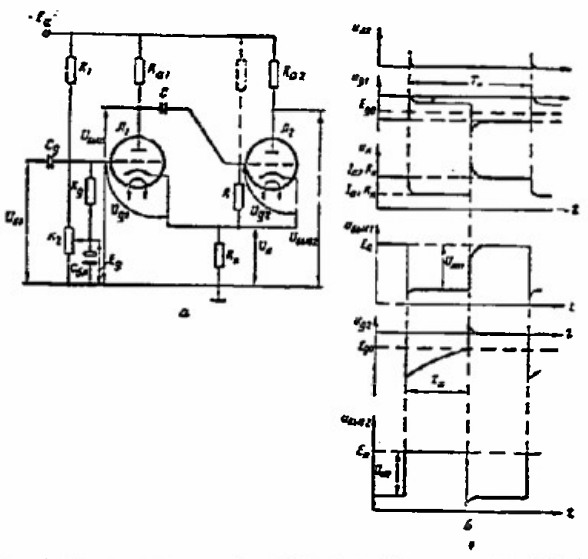


Figure 6.22. Triggered multivibrator with cathode coupling using electron tubes.

a - circuit; b - voltage time diagrams.

The pulse width can be calculated through the formula

$$\tau_p \approx RC\ln \cdot U_{m1}/|E_{g0}|. \qquad (6.19)$$

Width is adjusted by changing R and C, as well as by changing voltage E_g.

The circuit provides pulses from units in microseconds to fractions of a second. Maximum permissible pulse duty factor is 0.3 to 0.5. The use of certain types of circuits (pick-off diode, cathode follower) results in considerably increasing the duty factor. Stability of the pulse width is in

percentage. Connecting the resistance to source E_a (the dotted line in fig. 6.22a) results in a several-fold increase in the stability.

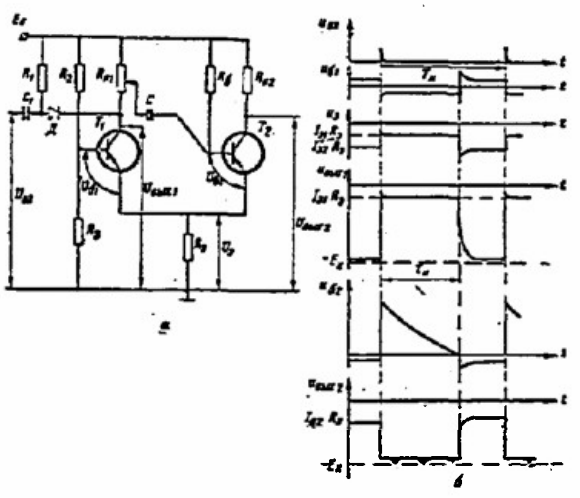


Figure 6.23. Triggered multivibrator with emitter coupling using transistors.

a - schematic; b - voltage time diagrams.

Circuits using transistors to make triggered multivibrators are similar to corresponding circuits using tubes. Figure 6.23 shows a triggered multivibrator circuit with emitter coupling.

The pulse width can be calculated, approximately, through the formula

$$\tau_p \approx R_b C \ln \frac{2 - R_e\left[\frac{1}{R_{c1}+R_e}+\frac{1}{R_{c2}+R_e}\right]+\frac{I_{co}R_b}{E_c}}{1-\frac{R_b}{R_{c1}+R_e}+\frac{I_{co}R_b}{E_c}} \qquad (6.20)$$

Triggers

Triggers are start-stop devices with two stable equilibrium conditions. Rheostat coupling between stages is used to obtain this mode. Triggers are divided into symmetrical triggers, in which both stages are identical, and unsymmetrical. Symmetrical triggers are used as frequency dividers for pulse repetition, for controlling various types of devices, as memory, counting, and switching cells in electronic computers, and the like. Unsymmetrical triggers are used mainly to form square voltages from sine voltages and as amplitude discriminators.

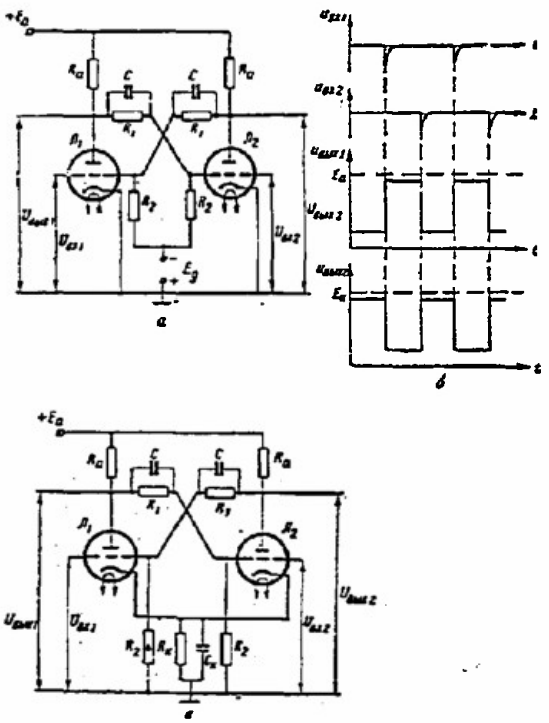


Figure 6.24. Vacuum tube symmetrical triggers.

a - trigger with external biasing; b - voltage time diagrams; c - trigger with automatic biasing.

Vacuum tube symmetrical triggers. The circuit most widely used is the one with an external biasing source (fig. 6.24a). The circuit has two stable equilibrium conditions:

tube T_1 is unblocked, T_2 is blocked;

tube T_1 is blocked, T_2 is unblocked.

A negative pulse, blocking the unblocked tube (or a positive pulse unblocking the blocked tube) must be fed into the input in order to shift the trigger from one stable state to the other. The voltage across the plate of the unblocked tube will increase and will flow through the coupling circuit to the grid of the blocked tube and unblock it. The circuit shifts to the second stable state, and this holds until the arrival of the next trigger pulse.

An automatic biasing circuit (fig. 6.24c) is often used, in addition to the circuit discussed. This circuit has the advantage of the absence of a biasing source, but also has the disadvantages that go with having two additional elements and a reduction in the amplitudes of the output pulses. The use of the external biasing source circuit is desirable in devices with a great many triggers.

The type of triggering reviewed, in which pulses of just one polarity are fed to the tube grid in turn, or in which pulses of alternating polarity are fed to the grid of a single tube, is called split triggering. Another type of triggering, common (or counting), is used when triggers are used in counters. Pulses of just one polarity are fed to the grid (or plate) of the trigger tube through a diode (fig. 6.25).

Quick-acting triggers can be characterized by maximum frequency of repetition of the trigger pulses, at which the trigger functions stably. Triode triggers can provide quick action up to 1 MHz. The action can be increased to 10 MHz and higher by some complication of the circuit (the use of pentodes, additional cathode followers, high-frequency correction, clamping diodes, and others).

Symmetrical transistor triggers. Transistor triggers are similar to tube-type triggers (fig. 6.26). The speed at which these triggers function is usually slower than that of the tube-type triggers, because of the lower (as compared with tubes) frequency properties of the transistor, as well as because of the delay in the beginning of the turnover at the time the unblocked transistor leaves the saturated mode. A variety of circuit arrangements can be used to increase the action, such as nonlinear feedback to eliminate the saturation mode, emitter repeaters, diode clamping, high frequency correction, and the like (fig. 6.27). The use of these measures can result in arriving at a speed of a few megahertz.

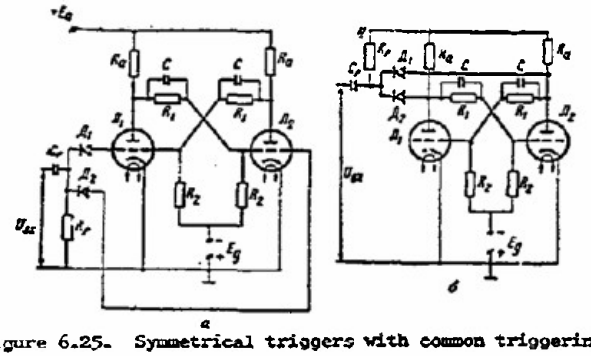


Figure 6.25. Symmetrical triggers with common triggering.
a - grid; b - plate.

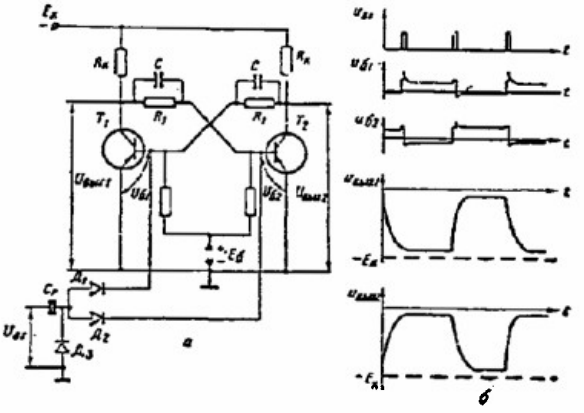


Figure 6.26. Symmetrical saturated transistor trigger.
a - schematic; b - voltage time diagrams.

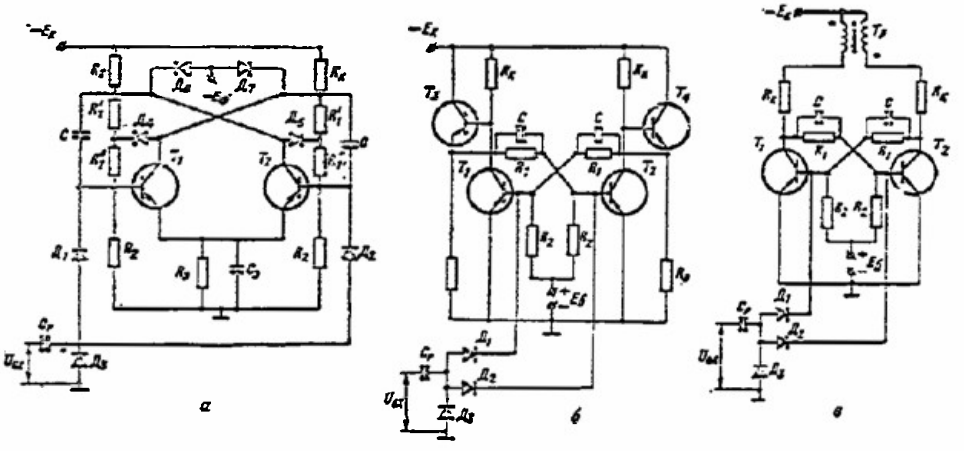


Figure 6.27. Different transistor trigger designs.
a - with diode clamping; b - with emitter followers;
c - with high-frequency correction.

Unsymmetrical trigger with cathode coupling (fig. 6.28). This circuit has a number of advantages, including a well-shaped output pulse, and a high degree of sensitivity. However, triggering will only occur with alternating polarity pulses so the unsymmetrical trigger with cathode coupling is

not used in counter circuits. When a voltage is delivered to the input the trigger flip-flops any time the input voltage passes through the operation level. The transistor circuit is similar to the tube circuit.

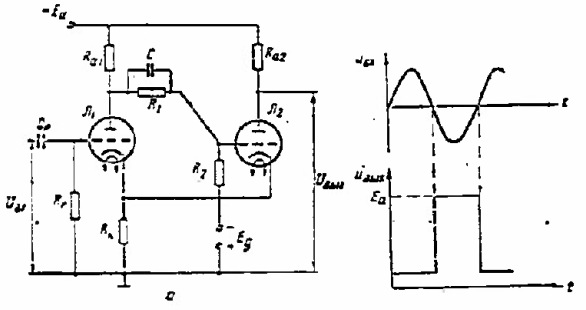


Figure 6.28. Unsymmetrical trigger.
a - schematic; b - voltage time diagrams.

Gas discharge triggers (thyratrons, cold cathode tubes) make it possible to switch heavy currents. The chief shortcoming of gas discharge triggers is their slow action, a few tens of kHz. They are used in automatic circuits, electronic switches, and counters. Figure 6.29 is the schematic of one design of a cold-cathode tube trigger. Initially one of the tubes is fired (unblocked), the other is blocked, and condenser C is charged. Upon delivery of a positive pulse the blocked tube fires, condenser C begins to discharge through this tube, the plate voltage of the tube which had fired drops sharply, and the tube is killed. The trigger shifts to the second stable condition.

Tunnel diode trigger (fig. 6.30a). The falling section of the diode curve is used in tunnel diode generators. The load line should intersect the diode curve at three points (fig. 6.30b), in order to obtain two stable states; points 1 and 3 are stable, point 2 is unstable. The circuit can be triggered by pulses of alternating polarity, and when the pulse is negative the transition is made from point 3 to point 1, when positive from point 1 to point 3. Advantages of this circuit include simplicity, high speed (up to several tens of megahertz), and high temperature stability. Shortcomings include lack of a unidirectional signal, making it difficult to match stages, and low voltage (tenths of a volt). The use of transistors will do away with these shortcomings. A speed of several MHz can be reached with combination circuits such as these. In addition to their use in triggers, tunnel diodes can be used in multivibrators for autooscillation and triggered modes.

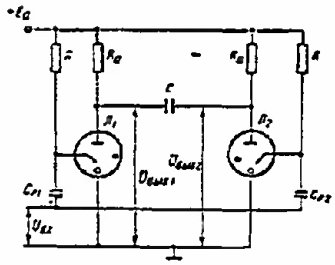


Figure 6.29. Cold-cathode tube trigger.

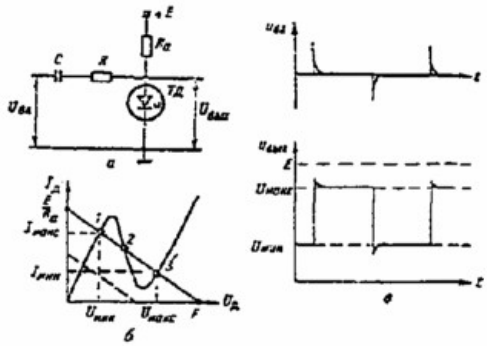


Figure 6.30. Tunnel diode trigger.

a - schematic; b - volt-ampere curve; c - voltage time diagrams.

Blocking Oscillators

A pulse transformer is used to form the positive feedback in these oscillators. Blocking oscillators are used to generate short pulses (from a few tens of nanoseconds to several tens of microseconds) with a relatively high pulse rate (from several tens to several hundreds of thousands). An important advantage of the blocking oscillator over the multivibrators is the ability to obtain high power in a pulse for low average power. The most widely used design of a blocking oscillator in the autooscillation mode is shown in Figure 6.31a.

The tube is blocked by negative voltage across the grid, created across resistance R by the discharge of condenser C, in the interval between pulse generations. The discharge causes the voltage across the grid to increase and when it reaches the magnitude of the trigger voltage the tube fires, the feedback circuit is closed and pulse generation occurs (fig. 6.31b). During

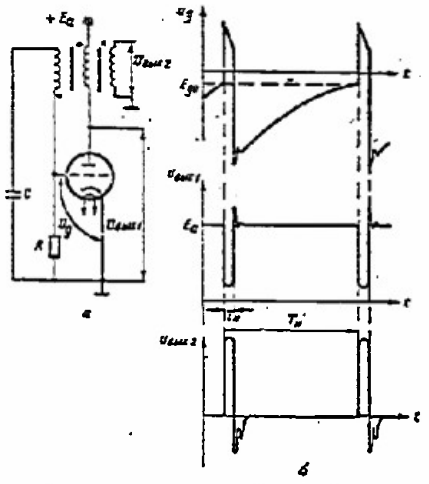


Figure 6.31. Electron tube blocking oscillator.
a - schematic; b - voltage time diagrams.

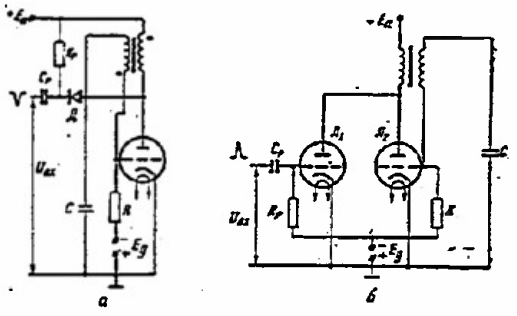


Figure 6.32. Start-stop blocking oscillator.
a - with triggering through the diode; b - with a triggering stage.

generation condenser C is charged by the grid current, the voltage across the grid decreases, and the tube is once again blocked at some predetermined time. The transformer windings must buck each other in order for generation to occur. If generation fails to occur the ends of one of the windings must be reversed. The voltage for the blocking oscillator can be picked off the plate, or the third (load) winding, so it is easy to obtain the required polarity and output pulse amplitude. Sometimes the output voltage can be taken off at the resistance connected to the plate or cathode circuit. Negative pulses can be taken off at the plate load, positive ones at the cathode load.

The repetition period can be adjusted by changing the magnitude of resistance R and capacitance C, the pulse width by changing capacitance C. Use of a "positive" grid circuit results in a several-fold increase in repetition period stability. The blocking oscillator can be blocked by an external source to obtain the triggered mode.

One of the most widely used methods of triggering is the delivery of the trigger pulses through an isolation diode (fig. 6.32a). Sometimes an additional decoupling stage (a cathode follower, or a tube connected in parallel with the blocking oscillator tube) can be used for triggering (fig. 6.32b).

Transistor blocking oscillators can generate pulses with widths of from a few tenths of a microsecond to several hundred microseconds with pulse rates from several units to several thousand. The circuits and the operating principles are similar to corresponding tube blocking oscillators. Most frequently used is the transistor blocking oscillator with a common emitter (fig. 6.33). The diode, D, is used to reduce the kickback occurring after pulse formation.

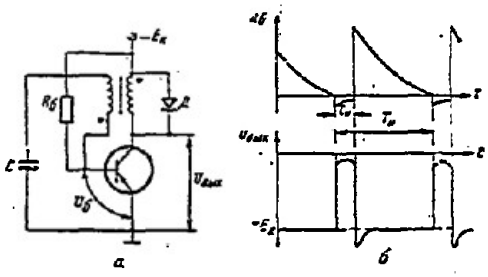


Figure 6.33. Transistor blocking oscillator.
a - schematic; b - voltage time diagrams.

Well-shaped pulses, a steep front, large load-carrying capacity, are all advantages of this circuit. Instability in pulse width is a few tenths of a percent, and the repetition period is 10 to 20%.

A several-fold increase in the stability of the repetition period can be obtained by using a circuit with an emitter capacitance (fig. 6.34). The shortcoming in this circuit is the high pulse peak cutoff. Artificial chain lines are used to increase pulse width stability.

Stabilization of pulse time parameters. Pulse excitation circuits (fig. 6.35a) are often used to increase the stability of the repetition period. The circuit oscillationa are added to the voltage across condenser C, with the result that the time at which the tube is unblocked can be stabilized (fig. 6.35b). The use of the circuits increases the stability of the pulse repetition period by a factor of 3 to 5. Circuits for increasing the stability of multivibrator-type oscillators can be used similarly.

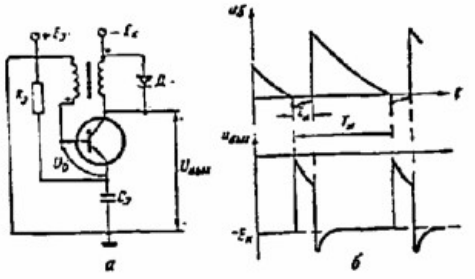


Figure 6.34. Blocking oscillator with emitter capacitance.

a - schematic; b - voltage time diagrams.

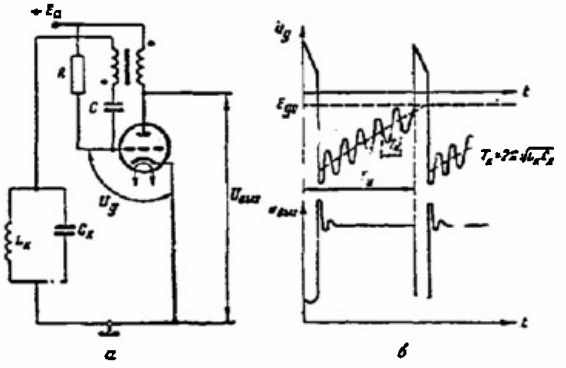


Figure 6.35. Blocking oscillator stabilized by a pulse excitation circuit.

a - schematic; b - voltage time diagrams.

An artificial chain line (fig. 6.36) can be used instead of capacitance C to stabilize the blocking oscillator pulse width. In this circuit the pulse width can be calculated by the artificial line parameters

$$\tau_p = 2n\sqrt{L_c C_c}. \qquad (6.21)$$

The number of line cells is usually 3 to 5. The circuit provides very good pulse stability but does have a number of shortcomings, including a reduction in pulse power by a factor of 1.5 to 2, deterioration in pulse shape, and an increase in the front length. The circuit is therefore only used when the requirements with respect to pulse width stability are rigid. The artificial line is seldom used to stabilize multivibrator type generators.

Generator synchronization is one way to improve the stability of pulse time parameters. External periodic voltages act on a generator operating in the automatic oscillation mode when synchronization is used. Synchronization involves the matching (multiplicity) of the repetition frequency of the

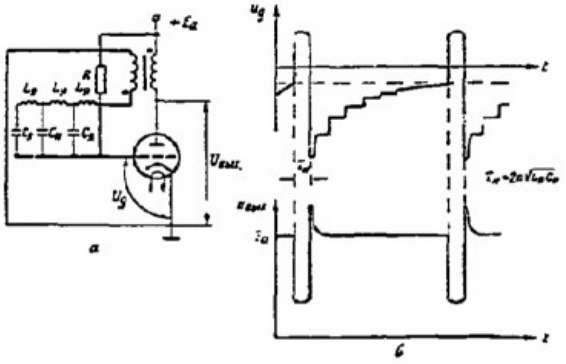


Figure 6.36. Blocking oscillator stabilized by an artificial chain line.

a - schematic; b - voltage time diagrams.

pulses from the generator with the frequency of the synchronizing voltage. A sine voltage and short pulses are most often used as the synchronizing voltage.

Figure 6.37a shows a blocking oscillator in the synchronization mode. The voltage across the grid is the sum of the voltage at the condenser and the synchronizing voltage. Generation occurs when this summed voltage reaches the blocking voltage value (fig. 6.37b). Other types of generators can be synchronized in a similar way.

The condition for synchronization is $T_0 > T_p$; the ratio $n = T_p/T_{syn}$ is called the synchronization multipleness. The higher this latter is, the more stable synchronization will be. Synchronization is widely used for time matching of the operation of pulsers, frequency dividers, and for the stabilization of generators.

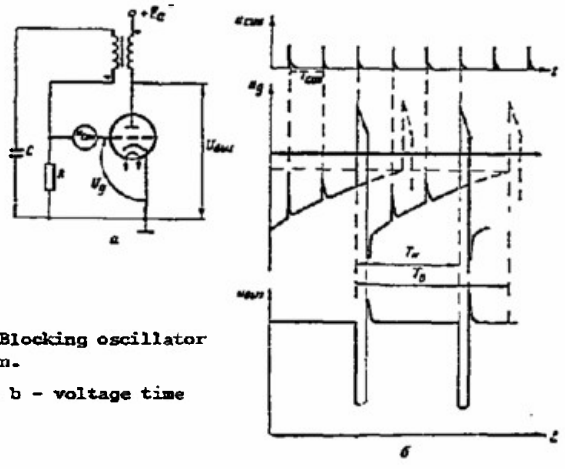


Figure 6.37. Blocking oscillator synchronization.

a - schematic; b - voltage time diagrams.

6.5 Generation of Sawtooth Voltage and Currents

The sawtooth voltage is used to obtain the time-base sweep in cathode ray tubes with electrostatic deflection, a time delay, and for other purposes. Sawtooth current is used primarily to obtain the time-base sweep in tubes with electromagnetic deflection (scopes, television devices, and the like).

Sawtooth Voltage Generators

The basis for obtaining a sawtooth voltage is the charging (discharging) of a condenser. In the simplest circuit (fig. 6.38a) condenser C is charged through R_a while the tube is blocked, but discharges quickly through the tube when the tube is unblocked. A sawtooth voltage is generated across the condenser (fig. 6.38b).

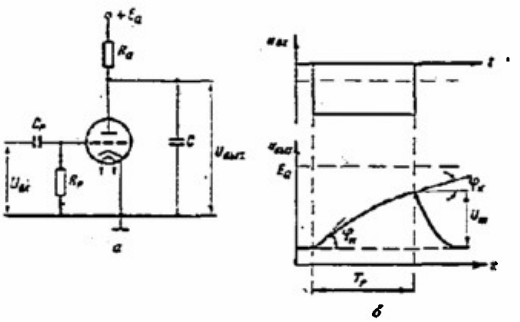


Figure 6.38. A simple sawtooth voltage generator.
a - schematic; b - voltage time diagrams.

The most important parameters of the sawtooth voltage are the duration of the operating rate, T_o, the average velocity, $v_{av} = U_m/T_o$, voltage use factor, $\xi = U_m/E_a$, and the coefficient of nonlinearity ϵ, which characterizes the relative change in the rate of rise (fall) at the beginning and at the end of the sawtooth voltage

$$\varepsilon = \frac{v_b - v_e}{v_b} = \frac{\left|\frac{du}{dt}\right|_b - \left|\frac{du}{dt}\right|_e}{\left|\frac{du}{dt}\right|_b} = \frac{\operatorname{tg}\varphi_b - \operatorname{tg}\varphi_e}{\operatorname{tg}\varphi_b}. \qquad (6.22)$$

The coefficient of nonlinearity should not exceed 0.1 (10%) in the case of the television sweep, 0.02 to 0.03 (2 to 3%) in the case of radar scopes, and 0.001 (0.1%) in the case of accurate time delay. In the simple generator considered $\epsilon = \xi$. Consequently, the nonlinearity is very high in the case of the practically desirable use of the source of supply ($\xi \geqslant 70$ to 80%). A stage such as this can be used in devices in which the linearity requirement is low.

The principal method used to obtain a linearly changing voltage is that of charging (discharging) a condenser with direct current. Current stabilizers (pentodes, transistors), as well as feedbacks, are used for current stabilization.

Sawtooth voltage generator with current stabilizer. Figure 6.39a shows a circuit with a current stabilizing pentode. When tube T_1 is blocked condenser C will discharge what is, for all practical purposes, direct current through the pentode because the pentode's plate curves have but a slight slope. A linearly falling voltage (fig. 6.39b) is generated at the output. When tube T_1 is unblocked condenser C will charge from source E_a. The minimum coefficient of nonlinearity is about 1%. The coefficient of nonlinearity can be reduced by using feedback to the pentode and a supplemental source of supply. The circuit can be used in sweep generators in certain types of oscillographs in combination with a start-stop multivibrator.

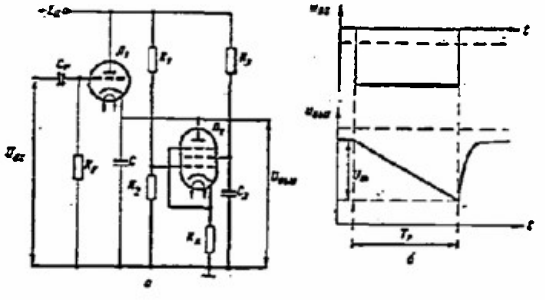


Figure 6.39. Sawtooth voltage generator with current stabilizing pentode.

a - schematic; b - voltage time diagrams.

Figure 6.40a is that of a similar circuit with a current stabilizing transistor. Initially both transistors are unblocked and condenser C is charged to practically the source voltage. Upon the arrival of a trigger pulse transistor T_1 is blocked, and condenser C discharges with the collector current of transistor T_2. Since the dependence of the collector current on the collector voltage is slight, the condenser will discharge in accordance with a linear law (fig. 6.40b). The minimum coefficient of nonlinearity for this circuit too is close to 1%.

There is a significant increase in the coefficient of nonlinearity when a load is connected to the condenser, not only in the case of the tube-type circuit, but in the case of the transistorized circuit as well. Consequently, a low-ohmic load must be connected through a transformer stage (such as a

cathode follower, for example). Adjustment of the rate of the linear voltage is made by changing the capacitance C, as well as with resistance R_c (in the tube design) and R_e (in the transistor design).

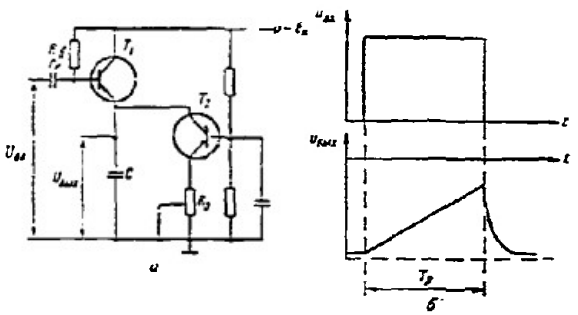


Figure 6.40. Sawtooth voltage generator with current stabilizing transistor.

a - schematic; b - voltage time diagrams.

Sawtooth voltage generator with positive feedback. Condenser C will begin to charge when switching tube T_1 (fig. 6.41a) is blocked. A compensating voltage, fed into the charging circuit from the cathode follower output, is used for current stabilization. Minimum coefficient of nonlinearity is about 1%. Condenser C and resistance R are used to adjust the rate of rise in the voltage. One advantage of the circuit is good load capacity (the voltage drops with the output of the cathode follower). The circuit is widely used in the range sweep channel of radar scopes.

The circuit is frequently combined with a trigger circuit (fig. 6.42). The role of the switching tube in the circuit is taken by the grid-cathode section of tube T_2. Initially T_2 is unblocked and T_1 is blocked. Feeding a trigger pulse unblocks T_1, causing an increase in the voltage across resistance R_{c1}, and T_2 is blocked. Condenser C is charged through the cathode follower, as shown in Figure 6.41. When the voltage across the grid-cathode section of tube T_2 reaches the level of the blocking voltage this tube is unblocked and T_1 is blocked. When the condensers are recharged the circuit returns to its initial state. Change in R and C results in a simultaneous change in velocity and duration of the operating rate, but amplitude U_m will not change, and this is an advantage of the circuit.

The transistor design of the main direcuit is similar to that of the tube design (fig. 6.43). It is recommended that a source E_e = 1.5 to 2 volts be inserted in the emitter circuit to prevent blocking the emitter follower during the circuit restoration period, and thus reduce the time required to recharge condenser C_{main}. The minimum coefficient of nonlinearity equals 1 to 2%.

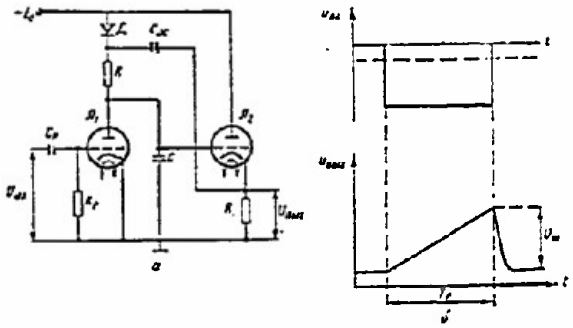


Figure 6.41. Sawtooth voltage generator with positive feedback.

a - schematic; b - voltage time diagrams.

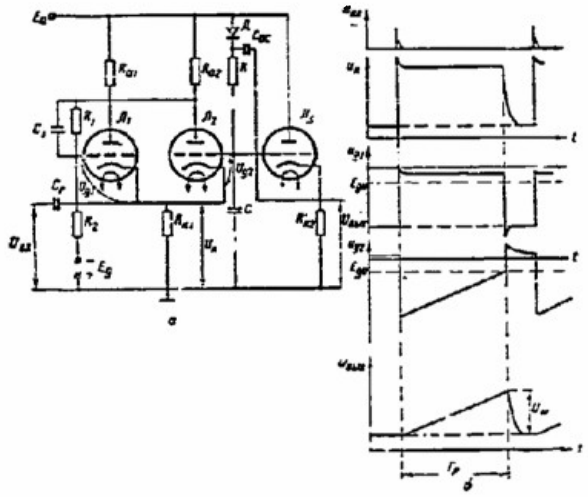


Figure 6.42. Trigger circuit combined with a sawtooth voltage generator.

a - schematic; b - voltage time diagrams.

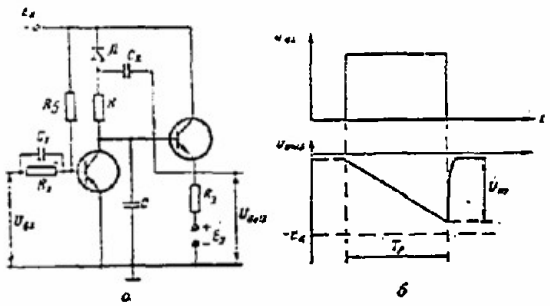


Figure 6.43. Sawtooth voltage generator with transistorized positive feedback.

a - schematic; b - voltage time diagrams.

Sawtooth voltage generator with negative feedback (fig. 6.44a). The tube is initially blocked at the third grid by negative voltage from the condenser, obtained by using a clamping circuit (diode D). When a positive pulse is supplied the tube is unblocked and condenser C discharges through resistance R and the tube. The discharge current and linearization of the output voltage (fig. 6.44b) are stabilized by the feedback from the tube plate into the discharge circuit. The chief advantage of this circuit is the high linearity (a coefficient of nonlinearity on the order of hundreds of a percent can be obtained). This is why the circuit is often used in devices for precise variable delay. Shortcomings are the slow rate of change in the voltage and the relatively long restoration time. The latter of these can be eliminated by making the circuit somewhat more complicated (by using cathode followers and clampers).

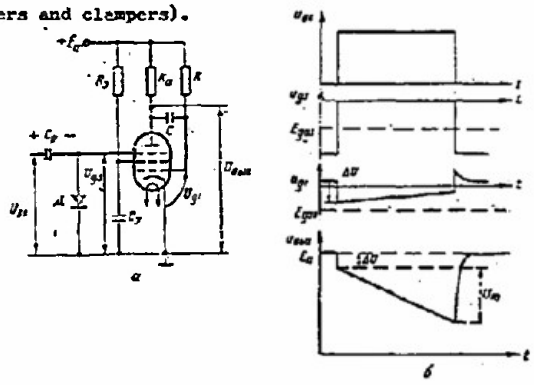


Figure 6.44. Sawtooth voltage generator with negative feedback.

a - schematic; b - voltage time diagrams.

The circuit considered can be combined with a trigger, in which case one must design an additional positive feedback loop. These devices are triggered by short pulses and the duration of the operating rate is fixed by circuit parameters.

The two-tube circuit (the sanatron), and particularly the one-tube circuit (the phantastron), are in use. One phantastron design is shown in Figure 6.45a. The cathode follower (tube T_2) reduces restoration time. Tube T_1 is initially blocked by the suppressor grid and unblocked by the first grid. The tube fires, and condenser C begins to discharge upon the arrival of a negative pulse. The discharge process is linearized by the negative feedback supplied the discharge circuit from the plate of tube T_1 through the cathode follower, T_2. When the voltage across the plate drops to the magnitude $U_{al\ min}$, at which the plate current ceases being controlled by the grid voltage, the action of the negative feedback ceases, the voltage across the control grid increases rapidly, and the screen grid current increases correspondingly. The voltage across the screen grid, and the suppressor grids

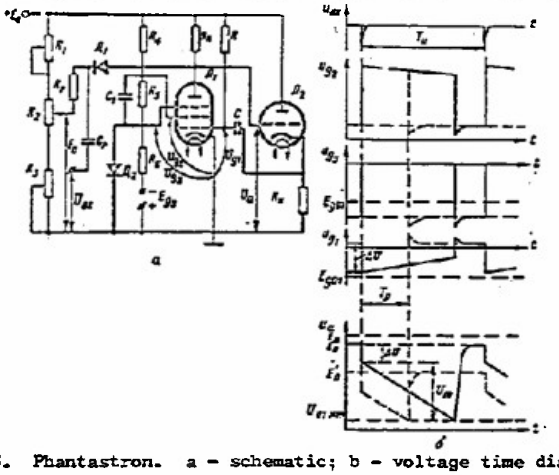


Figure 6.45. Phantastron. a - schematic; b - voltage time diagrams.

associated with it, decreases and the tube returns to the original condition in a jump, after which condenser C will charge quite quickly through the cathode follower. The phantastron is used primarily as a time delay device. The duration of the operating rate, which determines the magnitude of the delay, can be regulated by resistance R, capacitance C, and potentiometer R_2, by changing the original voltage across the anode (fig. 6.45b).

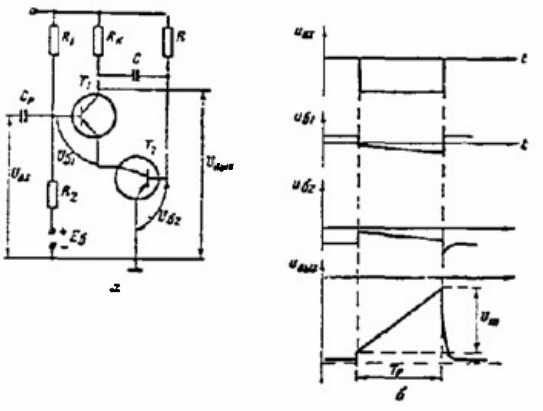


Figure 6.46. Sawtooth voltage generator with transistorized negative feedback.

a - schematic; b - voltage time diagrams.

Unitized transistors (two transistors connected in series), replacing the pentode, are usually used in sawtooth voltage generators with negative feedback (fig. 6.46). The operating principle of this circuit is similar to that of the tube circuit, for the most part. The sawtooth voltage generator can be built with one transistor if the trigger circuit is made more complicated.

Transistorized generators of this type provide a lower order of linearity of the sawtooth voltage than do the corresponding tube circuits (the minimum coefficient of nonlinearity has a magnitude on the order or tenths of a percent).

Sawtooth Current Generators

The chief purpose of sawtooth current generators is to create a sawtooth current in the deflection coil so as to obtain a time-base sweep in cathode ray tubes of radar scopes, in television, and the like. A trapezoidal voltage (fig. 6.1f) must be supplied in order to obtain a linearly changing current in the deflection coil. Accordingly, the sawtooth current generator consists of a trapezoidal voltage generator and an output stage (a current amplifier). A trapezoidal voltage generator will result from a sawtooth voltage generator if the capacitance is replaced by an RC circuit (fig. 6.47). The magnitude of the initial jump can be regulated by resistance R. Another way to get a trapezoidal voltage is to sum the sawtooth voltage and the square pulse in a mixer.

The output stages are current amplifiers, so they can be made using power tubes (beam tetrodes, ordinarily), or power transistors. The most widely used circuits are those with the deflection coil connected in the

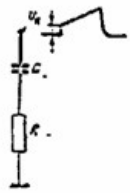


Figure 6.47. An RC circuit for obtaining a trapezoidal voltage.

plate, or the cathode, circuit of the tube (fig. 6.48). The chief advantage of inserting the coil in the plate circuit is the low amplitude of the input signal, while that of putting the coil in the cathode circuit is the smaller amount of current distortion. If slow sweeps are used the coil is usually inserted in the plate circuit, but if fast sweeps are used the coil will be found in the cathode circuit.

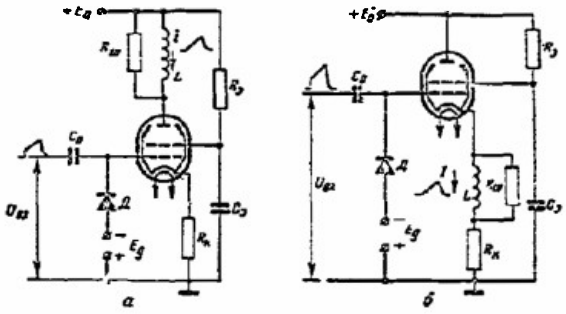


Figure 6.48. Output stages.

a - with the deflection coil in the plate circuit;
b - with the deflection coil in the cathode circuit.

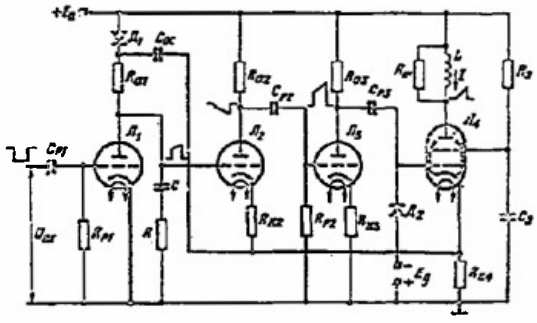


Figure 6.49. Sawtooth current generator with negative feedback.

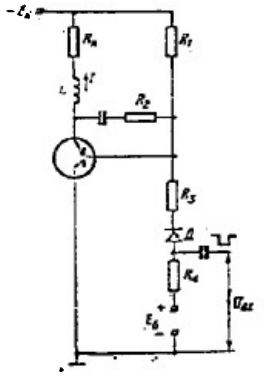


Figure 6.50. Transistorized sawtooth current generator.

Nonlinearity in the curves for the output stages of the tubes introduces marked distortions in the shape of the current. Great improvement in linearity can result from inserting an output stage in the trapezoidal voltage generator feedback circuit (fig. 6.49). In this circuit the positive feedback for linearization of the trapezoidal voltage is taken off the cathode output in the output stage. Amplifiers are inserted (tubes T_2 and T_3) to increase the gain in the feedback circuit, and as a result, the linearity. This circuit will yield a coefficient of nonlinearity for the current on the order of tenths of a percent.

The transistorized sawtooth current generator circuit can be put together in a manner similar to that used in the corresponding tube designs. However, there are difficulties, the result of the low input resistance of the transistor, so in practice output stages are combined with the trapezoidal voltage generator (fig. 6.50). The transistor is blocked in the initial mode. Delivery of an input pulse unblocks the transistor, and a current, increasing linearly, flows through the coil. Circuit linearity is relatively low (the coefficient of nonlinearity is 1 to 5%).

6.6 Pulse Control

Pulse Modulation and Demodulation

Modulation is defined as the change in one of the pulse parameters in accordance with a predetermined law. Modulation can take place with respect to amplitude, width, and phase. Demodulation is the process of separating out the modulating signal.

Pulse-amplitude modulation. A simple way in which to obtain pulse-amplitude modulation is to change the clipping level in the diode clipper. The advantages of this method are circuit simplicity and linearity in the modulation characteristic, while the chief shortcoming is low sensitivity. Pulse-amplitude modulation can also be obtained by changing the gain of the amplifier in accordance with the modulation law. Pulse-amplitude modulation in a sawtooth voltage generator can be accomplished by changing the source voltage, E_a.

Peak detectors can be used for demodulation. Bridge circuits yield good results (fig. 6.51a). The control voltage, U_c, is an unmodulated pulse with an amplitude in excess of the maximum amplitude of the modulated pulses, U_m (fig. 6.51b). A voltage equal to the control pulses in amplitude is set up in the autobias circuit, R_1C_1. Upon delivery of the control pulse the diagonal cd in the bridge is short-circuited and condenser C is charged from the source of modulated pulses (or is discharged through it if $U_{out} > U_m$). If there is no control voltage all the diodes are blocked by the voltage across the R_1C_1 circuit, and the voltage across the output remains practically constant. The result is separation of the envelope of modulated pulses at the output (fig. 6.51b).

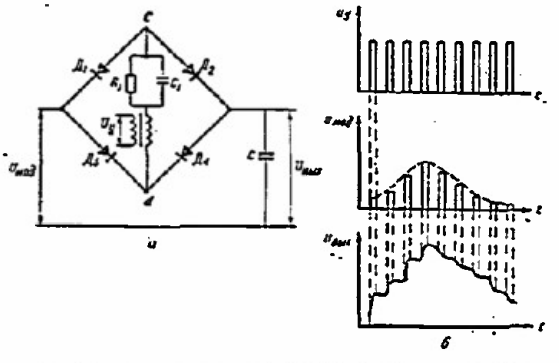


Figure 6.51. Bridge circuit for the demodulation of pulse-amplitude modulation.

a - schematic; b - voltage time diagrams.

Width modulation, or width-pulse modulation. Width-pulse modulation can be obtained by changing the width of the pulse from the pulse generator (trigger circuits, phantastron, and the like). Simple width-pulse modulation can be obtained by using nonlinear feedback in a transistorized blocking oscillator (fig. 6.52). Feedback speeds up the process of taking the transistor out of the saturation mode and reduces the width of the pulse generated. The modulating voltage changes the moment in time the feedback is cut in, and consequently controls pulse width.

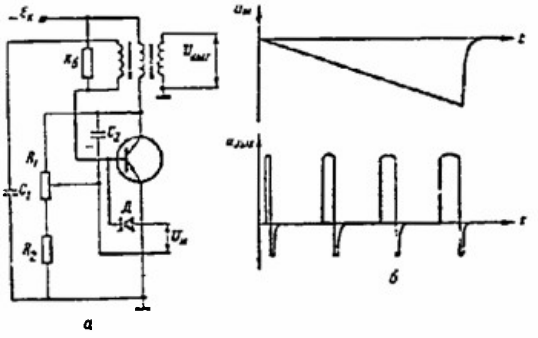


Figure 6.52. Blocking oscillator with width-pulse modulation.
a - schematic; b - voltage time diagrams.

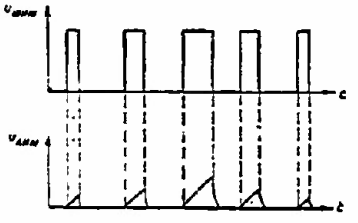


Figure 6.53. Schematic of the conversion of width-pulse modulation into pulse-amplitude modulation.

Width-pulse modulation demodulation is usually done by making a preliminary conversion into pulse-amplitude modulation, which can be demodulated by the methods described above. Conversion of width-pulse modulation into pulse-amplitude modulation can be done by using a generator that will produce a linearly changing voltage (fig. 6.53).

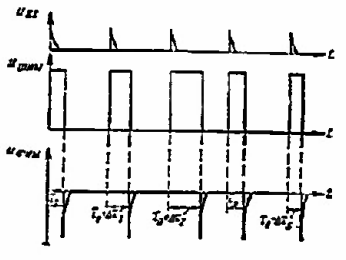


Figure 6.54. Schematic of the conversion of width-pulse modulation into pulse-position modulation.

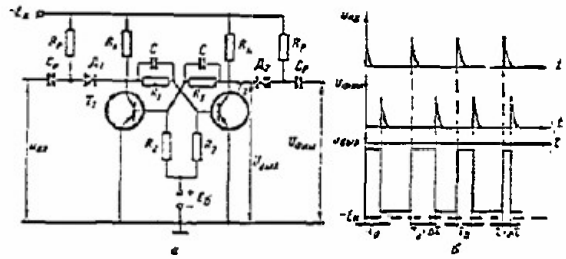


Figure 6.55. Conversion of pulse-position modulation into width-pulse modulation using a trigger.

a - schematic; b - voltage time diagrams.

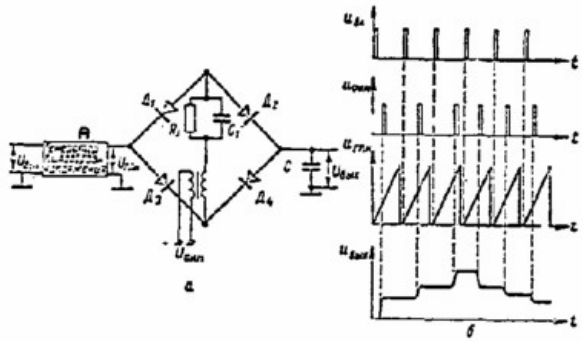


Figure 6.56. Conversion of pulse-position modulation into pulse-amplitude modulation.

a - schematic; b - voltage time diagrams; A - sawtooth voltage generator.

Phase modulation - pulse-position modulation. Pulse-position modulation involves changing the time position of each pulse in a modulated pulse train with respect to a periodically unmodulated pulse train. Pulse-position modulation can be obtained from width-pulse modulation by differentiation and subsequent clipping (fig. 6.54). Preliminary conversion of pulse-position modulation into width-pulse modulation, or into pulse-amplitude modulation, usually occurs during demodulation. A trigger can be used to convert pulse-position modulation into width-pulse modulation (fig. 6.55).

A device consisting of a sawtooth voltage generator, switch, and accumulator (capacitance) can be used to convert pulse-position modulation into pulse-amplitude modulation. The sawtooth voltage generator can be triggered by a periodic pulse train, with switching (connecting the accumulator to the sawtooth voltage generator) done by a modulated pulse train. This results in the envelope of modulated voltage (fig. 6.56) being separated at the accumulator.

Pulse Selection

Pulse selection is understood to mean the separation out of the pulse train of just those pulses that differ in accordance with a predetermined designation (predetermined amplitude, width, and the like). Selection can be used for the automatic logging of coordinates in radar, in pulse communications, in television, and the like.

Amplitude selection. Any of the following clippers can be used to select pulses, the amplitudes of which exceed a definite magnitude. The circuit shown in Figure 6.57a can be used to select pulses, the amplitudes of which are less than a specified magnitude. If the pulse amplitude is greater than a specified level, the pulse expander (the start-stop multivibrator) is triggered and the pulse blocking the amplifier is produced. If the amplitude of the input pulse is less than the specified magnitude the expander will not function, and a pulse will appear at the amplifier output (fig. 6.57b).

Width selection. A simple method for selecting pulses with widths in excess of a specified magnitude is the use of a sawtooth voltage generator with subsequent clipping (fig. 6.58). Used for this purpose is a circuit with an artificial chain line (fig. 6.59a), the advantage of which is high stability. If the input pulse exceeds twice the width of the artificial chain line delay, there is addition of the input pulse to the pulse reflected from the end of the line at the load, and a pulse will appear at the clipper (tube T_2) output (fig. 6.59b). If the width of the pulse is less than twice the width of the artificial chain line delay, there will be no addition, and there will be no pulse at the output.

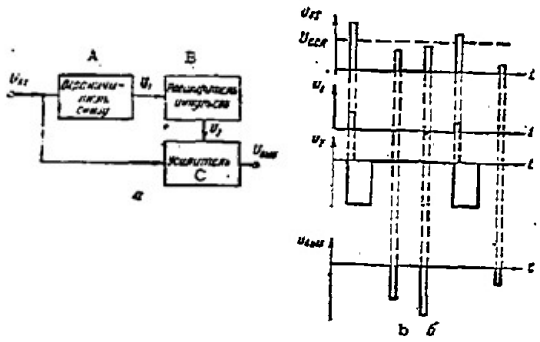


Figure 6.57. Amplitude selector.

a - block schematic; b - voltage time diagrams.
A - bottom clipper; B - pulse expander; C - amplifier.

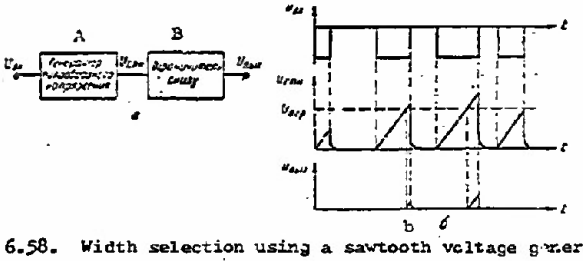


Figure 6.58. Width selection using a sawtooth voltage generator.
a - block schematic; b - voltage time diagrams;
A - sawtooth voltage generator; B - bottom clipper.

If it is necessary to select a pulse with a width less than a specified magnitude, the selection of a pulse with a width in excess of this magnitude as reviewed above, and a blocked amplifier, can be used. The functional construction of the circuit and its operation are similar to the circuit shown in Figure 6.57.

Time selection. In time selection those of the input pulses coinciding in time with the gate pulse are separated. Two methods are used in time selection. The first involves adding the input and gate pulses, and then clipping (fig. 6.60). The second method is based on the use of coincidence gate circuits (the "P" circuits) described in what follows.

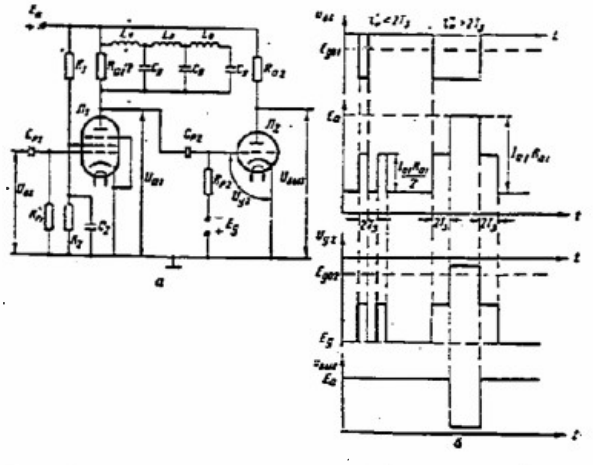


Figure 6.59. Width selection using an artificial chain line.
a - schematic; b - voltage time diagrams.

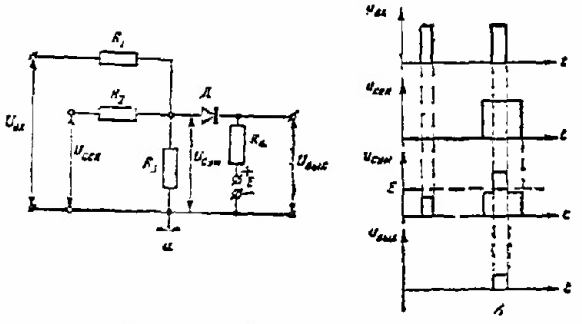


Figure 6.60. Time selection.
a - schematic; b - voltage time diagrams.

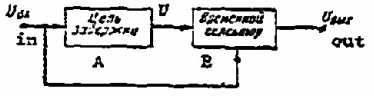


Figure 6.61. Block schematic of pulse repetition frequency selection.
A - delay circuits; B - time selector.

Repetition frequency selection. A circuit with a delay, $T_d = T_p$, and a time selector (fig. 5.61) can be used to separate those pulses with a predetermined repetition frequency, $F_p = 1/T_p$, from a pulse train.

Comparators (Amplitude Comparators)

These devices are used to bring about a drop in voltage (current) at the moment in time the input voltage and the reference voltage (constant, or changing slowly) are equal. The chief use of comparators is to obtain a time delay. Clipping amplifier circuits, or an unsymmetrical trigger (fig. 6.28), can be used as comparators.

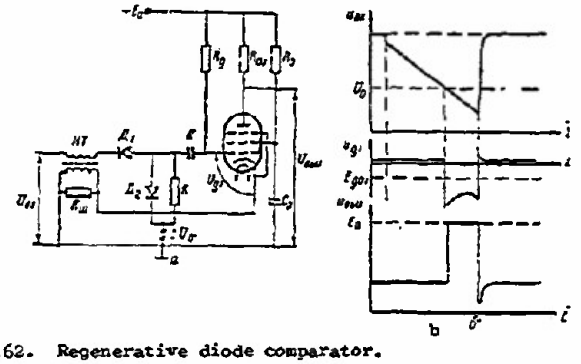


Figure 6.62. Regenerative diode comparator.
a - schematic; b - voltage time diagrams.

Great steepness in the drop at the moment the comparison is made is provided by a regenerative diode comparator (fig. 6.62). Prior to the moment the comparison is made, diode D_1 is blocked and the feedback circuit is open. At the moment the comparison is made between the input voltage and the reference voltage, U_o, the diode is unblocked, the feedback circuit is closed and a jump, like the one in a conventional blocking oscillator, takes place in the circuit.

Time Delay Devices

Time delay devices are used for time selection, pulse measurements, matching the operation of pulse devices, separating channels in the case of pulse communications, and the like. Delay lines, electronic delay circuits, and phase shifters, can be used for time delay.

Delay lines can be divided up into electrical and ultrasonic. Artificial chain lines (fig. 6.9) are used, for the most part, as electrical delay lines. Delay duration is

$$T_d = n\sqrt{L_c C_c}. \qquad (6.23)$$

Artificial chain lines are used primarily to cause no more than a 10 to 20 microsecond delay. Lines are cumbersome in the case of long time delays. Lines with distributed parameters (simple cables and cables with spiral windings) can be used to obtain very small delays (tenths of a microsecond).

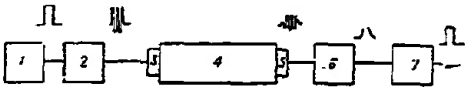


Figure 6.63. Piezoelectric delay line.

1 - modulator; 2 - generator; 3 - transmitting quartz plate; 4 - acoustic line; 5 - receiving quartz plate; 6 - detector amplifier; 7 - shaper.

Ultrasonic lines are used to obtain longer delays (hundreds and thousands of microseconds). Their action is based on converting an electrical signal into a sound oscillation for an acoustic line. In ultrasonic lines with piezoelectric converters the conversion is made by a quartz plate (fig. 6.63). Mercury (t_s = 6.7 ms/cm; attenuation δ = 0.083 db/cm), fused quartz (t_s = 1.8 ms/cm; δ = 0.007 db/cm), and magnesium alloys (t_s = 1.7 ms/cm; δ = 0.01 to 0.2 db/cm) are all used as acoustic lines.

Acoustic lines with multiple reflections (fig. 6.64) are used to increase the delay.

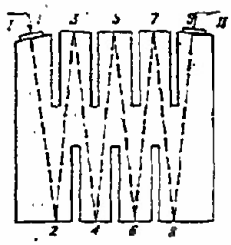


Figure 6.64. Piezoelectric delay line with multiple reflections.

The magnetostriction phenomenon (the change in the dimensions of ferromagnetic bodies when acted upon by a magnetic field) is used in ultrasonic magnetostriction delay lines (fig. 6.65). A thin nickel wire (a few tenths of a millimeter in diameter, t_s = 2 ms/cm, δ = 0.2 to 0.3 db/cm) is usually used as the acoustic line. Permanent magnetic fields are created at the transmitting and receiving ends in order to increase conversion effectiveness. Different time delays can be obtained by installing a few receiving coils along the line.

Electronic delay circuits make it possible to obtain a delay from a few microseconds to several seconds. The advantages of these circuits include simplicity and the possibility of adjusting the delay over broad limits. Low stability, compared with delay lines, is a shortcoming. The amplitude comparator with an input voltage changing in accordance with a linear law can be used as the electronic delay circuit. The time delay is adjusted by changing the comparison level. The time instability of these circuits is $\sigma = \Delta t_d/t_d$ and can be reduced to 0.1 to 0.05%.

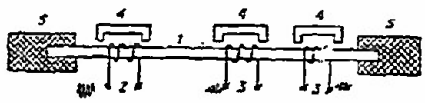


Figure 6.65. Magnetostriction delay line.
1 - acoustic line; 2 - transmitting coil; 3 - receiving coils; 4 - permanent magnets; 5 - sound absorbers.

Time delay can also be obtained by using trigger circuits (fig. 6.22) and a phantastron (fig. 6.45). The output pulse from these circuits is differentiated for this purpose. The pulse obtained as a result of differentiation of the cut will be delayed relative to the input by the magnitude $t_d = \tau_p$ (fig. 6.66). Regulation of the pulse width results in changing the delay time. Instability of the delay from a trigger circuit is σ = 1 to 5%;

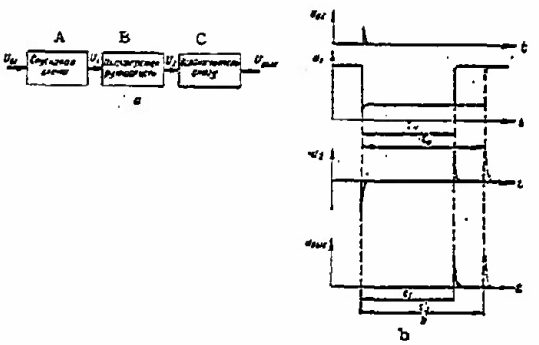


Figure 6.66. Obtaining time delay with a trigger circuit.

a - block schematic; b - voltage time diagrams; A - trigger circuit; B - differentiating circuit; C - bottom clipper.

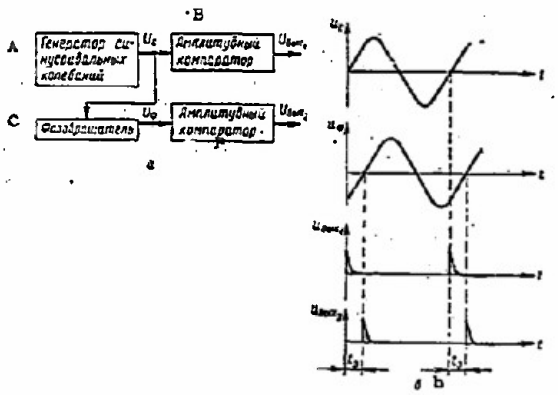


Figure 6.67. Time delay with amplitude comparators.

a - block schematic; b - voltage time diagrams; A - sine oscillation generator; B - amplitude comparator; C - phase shifter.

from the phantastron σ = 0.1 to 1%. A high degree of stability and good linearity of delay adjustment are basic principles of widespread use of the phantastron as a time delay circuit.

Time delay using phase shifters provide variable delay with a high degree of stability (σ = 0.01 to 0.001%). The delay principle is based on a phase shift in the sine voltage relative to the original voltage, with subsequent conversion of the sine voltage into a pulse, by amplitude comparators (fig. 6.67). The simplest type of phase shifter that can be used is a bridge circuit (fig. 6.68). Resistance R can be used to change the phase between 20° and 160°. Turning the rotor of an induction phase shifter (selsyn transformer) results in changing the phase 360°. The capacitance phase shifter

(fig. 6.69) has the greatest stability (σ = 0.005%). Sine voltages shifted 90° are fed to the fixed sectors. The phase of the output voltage can be changed over 360° by turning the rotor from the dielectric.

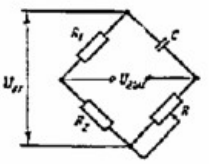


Figure 6.68. Bridge-type phase shifter.

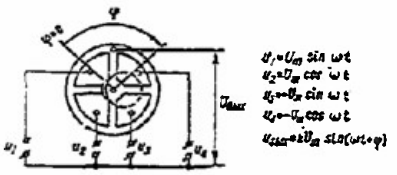


Figure 6.69. Capacitance phase shifter.

Pulse-Rate Division

Pulse-rate division can be used to match the operation of pulsers, for obtaining scale markers, for measuring time intervals, and the like. Most often used as dividers are synchronized generators, circuits with stable equilibrium states (triggers), and circuits with stepped energy accumulation.

Synchronized generators are dividers with synchronization multipleness of $n > 1$ (fig. 6.37).

The higher the synchronization multipleness (that is, the larger the divider's division factor) the lower the stability of the division factor. Practically speaking, a stable division factor in the case of synchronization does not exceed 5 to 7. When stabilizing devices are used (delay lines, impact excitation circuits) this magnitude can be increased to 8 to 10.

Dividers based on generators in the start-stop mode belong to this class, and in this case trigger pulses with a repetition period that is shorter than the length of the generated pulse are fed into the generator (such as the trigger circuit in fig. 6.22, for example), and circuit operation does not occur with each trigger pulse. After differentiation there are pulses at the output with a lower repetition rate. These dividers provide a stable division factor, equal to 10 to 15. The advantage of these over synchronized generators is that when there are no trigger pulses there are no output

pulses, indicating equipment fault. The generators continue to generate, although synchronization is upset.

A general shortcoming of this class of dividers is the change that takes place in the division factor with change in pulse synchronization frequency.

Counter triggering is used for trigger type pulse-rate division. The division factor equals two. The advantages of this divider are stability of circuit operation and independence of the division factor from the repetition rate of the triggering pulses. A shortcoming is the high value of the division factor.

Dividers with energy accumulators use stepped charging of the capacitance (fig. 6.70a). Originally the blocking oscillator tube is blocked by voltage U_c. Upon arrival of a triggering pulse the accumulator condenser, C, will charge and when the pulse ends will not, practically speaking, discharge. Condenser voltage, and consequently the voltage across the grid of the tube, will increase with the arrival of each pulse (fig. 6.70b). A jump will occur in the blocking oscillator when the grid-cathode voltage reaches the blanking voltage. At the same time a short pulse is formed at the output, the accumulator condenser will discharge the grid currents and the circuit will return to its original state. The stable division factor for the circuit equals 6 to 8. The advantage of the circuit is that the division factor is practically independent of input pulse rate; a shortcoming is the sensitivity to changes in the parameters of the trigger pulses and supply voltages.

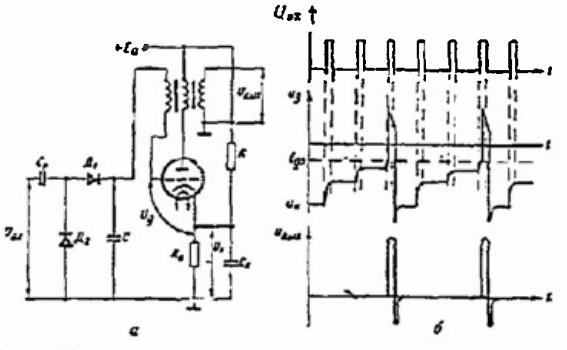


Figure 6.70. Divider with energy accumulator.
a - schematic; b - voltage time diagrams.

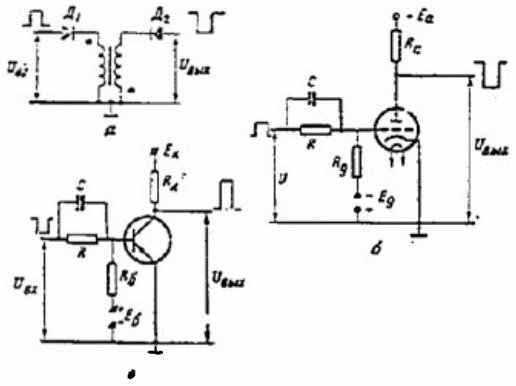


Figure 6.71. NO logic elements.
a - diode transformer; b - electron tube; c - transistor.

Decision Elements

Decision elements are widely used in electronic computers, in devices used for non-contact switching, and the like. They are gate circuits with several inputs and one output and make decisions. The signals at the input and output of gate circuits can have two values, so they are coded in a binary counting system with the numbers 0 and 1. Logic circuits can be divided into potential (signals at the input and output are potentials), pulse (signals at the input and output are pulses), and potential-pulse. The basic operations carried out by decision elements are NOT, OR, and AND. Any complex logic operation can be carried out on the basis of these operations.

The NOT decision element (inverter) realizes the operation of negation. In the case of the potential element there should be a low potential at the output when there is a high input potential, and in the case of the pulse element the output polarity will be opposite to that of the input polarity. The diode transformer circuit (fig. 6.71a) can use the coils hooked up to buck each other, while the tube and transistor circuits (figs. 6.71b and c) use the inverting properties of these devices. The diode transformer circuit is only a pulse logic element. The tube and transistor circuits can be used as potential or pulse elements.

The OR decision element (combining circuit) has two and more inputs and one output. There should be a signal at the output when a pulse is fed into any input. The diode rheostat (fig. 6.72a), diode transformer (fig. 6.72b),

tube and transistor circuits (fig. 6.7ac and d) work on a common load. A signal fed into any input results in a signal at the common output. In the ferrite diode element OR (fig. 6.72e) the amplitudes of the input pulses are selected such that each of them magnetizes the core, that is, any of them logs the information about the signal. The information is logged at the output by a clock pulse. The tunnel diode OR decision element is a trigger with several inputs. In the OR circuit the amplitudes of the input signals are selected with a magnitude such that each will individually cause the trigger to change over to a state with a high output potential.

The AND decision element (coincidence circuit) has two and more inputs and one output. There will only be a signal at the output when there is a simultaneous delivery of pulses to all inputs. In the diode rheostat circuit (fig. 6.73a) the entire diode will be blocked only when pulses are delivered to all inputs, and a high potential will appear at the output. If conditions are otherwise, just one of the diodes will be unblocked and there will be a low potential at the output. In the transformer circuit (fig. 6.73b) the amplitudes of the input pulses are smaller than the source voltage E, and their summed amplitude is larger than E. If one of the pulses is missing the diode is blocked, no current will flow in the primary winding, so there will be no pulse at the output. Simultaneous delivery of pulses unblocks the diode and there is a pulse at the output.

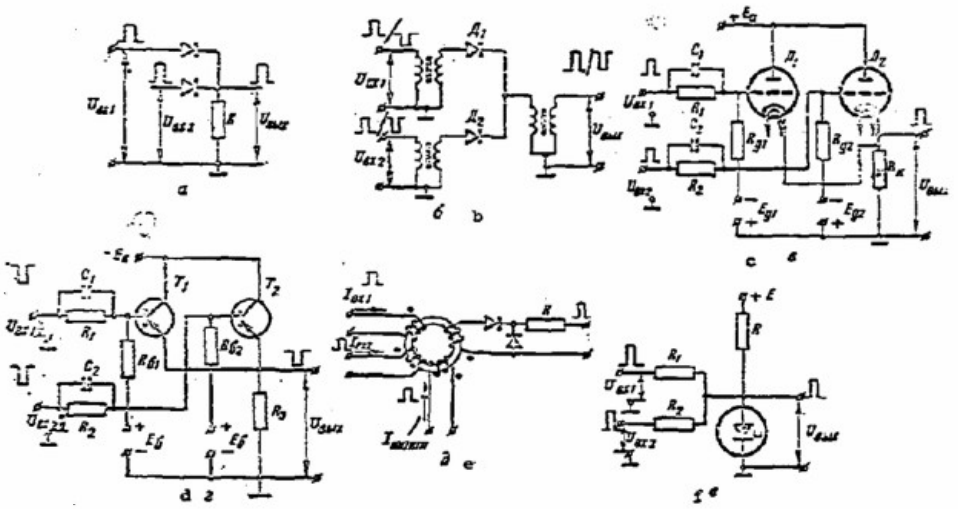


Figure 6.72. OR decision elements. a - diode rheostat; b - diode transformer; c - electron tubes; d - transistors; e - ferrite diode; f - tunnel diode.

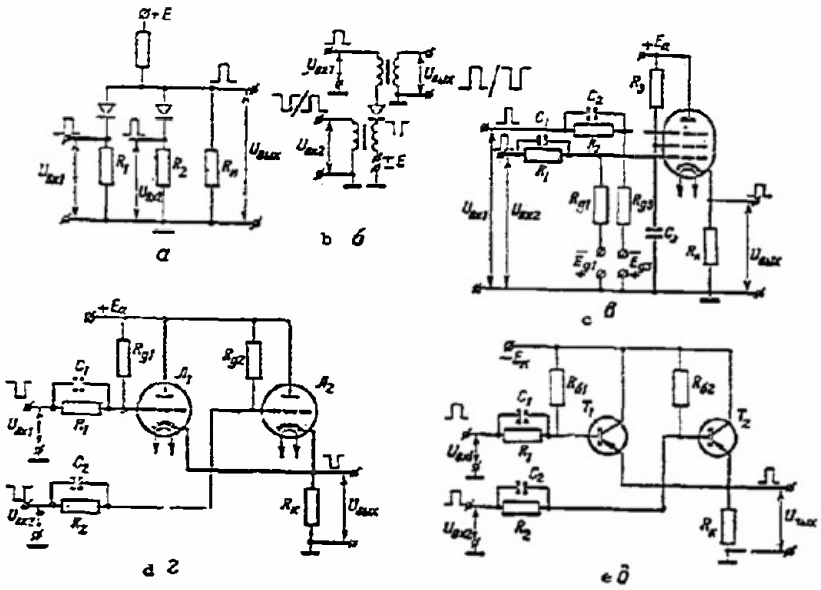


Figure 6.73. AND decision element. a - diode rheostat; b - diode transformer; c - pentode; d - triodes; e - transistors.

The tube in the pentode circuit (fig. 6.73c) is only unblocked upon simultaneous delivery of pulses to the first and third grids.

The triode and transistor circuits (figs. 6.73d and e) have the triodes (transistors) unblocked in the original mode. Only upon simultaneous blocking of all tubes (transistors) will there be a pulse at the output.

The AND decision elements using ferrites and tunnel diodes are no different than the OR elements (figs. 6.72e and f), but the amplitudes of the input signals are selected such that the polarity reversal of the ferrite core (or the flip-flop of the tunnel diode) occurs only upon the simultaneous delivery of pulses to all inputs.

New gate decision elements, parametrons, and elements based on the phenomenon of superconductivity, have recently begun to be introduced.

Parametrons are circuits in which oscillation with phases displaced 180° (π), occur with a periodic change in inductance (or capacitance) and the phase of the oscillation depends on the initial oscillations (signals) of small amplitude. Oscillations with zero phase are taken for the signal

"0" and those with phase π for the signal for "1" The advantage of these circuits are reliability, high speed operation, and noise stability, while the shortcoming is the need for powerful modulated sources of high frequency oscillations.

In the gate elements based on the phenomenon of superconductivity (the cryotron, the persistor) the state of normal conductivity is taken for the signal "0" and the state of superconductivity for the signal "1". The elements are switched from one state to the other by a magnetic field created by the current flowing through them. The advantages of these elements include reliability, efficiency, small size, while shortcomings include slow operating speed, and the need for deep cooling in the installations (on the order of units of degrees, absolute).

Chapter VII

Radio Transmitting Devices

7.1 Classification of Radar Sending Devices

Radar transmitters, depending on their operating range, are divided into meter, decimeter, centimeter, and millimeter transmitters.

Contemporary radar transmitters operate on pulses. According to modulation, distinctions are made among pulse, pulse-frequency, and pulse-position transmitters.

Depending on the construction of the generating stage, transmitters may be single stage or multistage.

Single stage generators operate in self-oscillating modes.

Multistage generators operate in amplifying modes with constant (external) excitation.

Transmitters are also classified by the generating element in the output stage. Accordingly, transmitters are divided into:

tube;

magnetron;

klystron;

platinotron, etc.

A block diagram of a single stage pulse transmitter is shown in Figure 7.1. Among the possibilities for SHF generators in this circuit are triode generators, magnetrons, backward wave tubes, stabilitrons, and other generating devices.

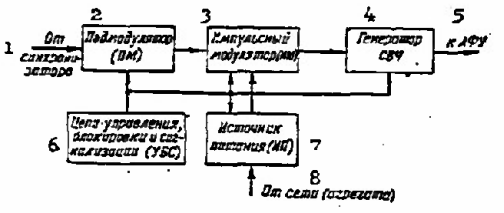


Figure 7.1. Block diagram of a single stage pulse transmitter.

1 - from synchronizer; 2 - driver; 3 - pulse modulator; 4 - SHF generator; 5 - to antenna feeder; 6 - control, interlock, and signaling circuits; 7 - supply source; 8 - from net (supply unit).

High frequency oscillations from the SHF generator are led through a feeder to the antenna.

The SHF generator is controlled by a pulse modulator.

A high-voltage rectifier is usually used for the supply source of the transmitter.

The pulse modulator is triggered by a driver.

Transmitter and indicator operation is synchronized by a common synchronizer.

The fundamental stages of all units of a radar device include circuits for control, interlocking, and signaling.

A multistage radar device is constructed as shown in the block diagram in Figure 7.2.

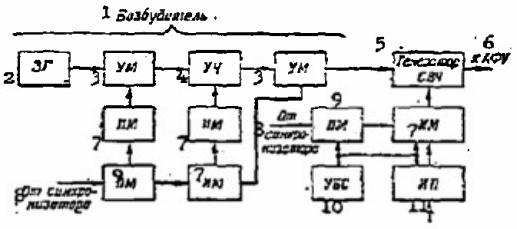


Figure 7.2. Block diagram of a multistage pulse transmitter.

1 - exciter; 2 - master oscillator; 3 - power amplification; 4 - frequency multiplication; 5 - SHF generator; 6 - to antenna feeder; 7 - pulse modulators; 8 - from synchronizer; 9 - driver; 10 - control, interlock and signaling circuits; 11 - supply source.

The output SHF oscillator, fundamental pulse modulator, supply source, and control, interlocking and signaling circuits in this transmitter are named analogously to the corresponding units of a single stage pulse transmitter and may be distinguished only by electrical and technical data.

The output oscillator of a multistage transmitter operates in an amplifying mode, and consequently, is excited by a constant source of SHF oscillations. In the given case, the exciter contains a master oscillator, stages for frequency multiplication and power amplification. These stages may operate continuously or pulsed. In the latter case, pulse modulation is effected in certain multiplication and amplification stages by pulse modulators, included in the exciter.

In the output oscillator, triodes, klystrons, amplitrons, backward wave tubes, and other amplifying devices are used. The same types of devices, but lower in power, are used in the amplifying and multiplying stages of the exciter.

Triodes, stabilitrons, and other generating devices are used in the master oscillators, which are SHF self-oscillators of low power with high frequency stabilization.

The pulse modulators may be triggered by various drivers or by one driver. Sequential pulse modulators may also be triggered by the preceding pulse modulators.

7.2 SHF Triode Oscillators

SHF triode oscillators are widely used in the meter and decimeter ranges of radio waves.

SHF oscillators are most often constructed with a common grid (fig. 7.3). Tube 1 with coaxial anode grid 2 and grid-cathode resonators forms a closed oscillations system. The oscillator is tuned by changing the length of the anode-grid resonator using plunger 4. Feedback factor is regulated, and consequently its operating mode is accomplished by changing the position of plunger 5 in the grid-cathode cavity. Naturally, shifting plunger 5 changes somewhat the operating frequency of the self-oscillator.

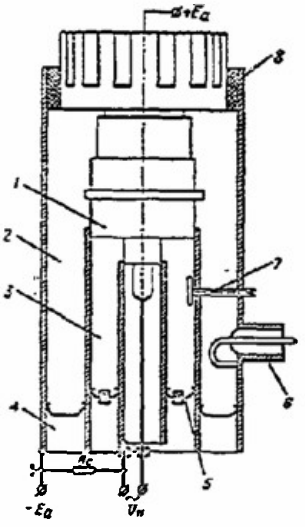


Figure 7.3. Schematic of a triode SHF generator.

1 - tube; 2 - anode-grid cavity; 3 - grid-cathode cavity; 4 and 5 - tuning plungers; 6 - high frequency output; 7 - capacitive stub; 8 - dividing capacitors.

Energy of the SHF oscillations is transmitted from the oscillator to the antenna through high frequency output 6.

Feedback between anode-grid and cathode grid resonators is effected by interelectrode capacitance of the tube. Inductive or capacitive elements are often used for its amplification.

In the oscillator under discussion, capacitive stub 7 with a small disk at the end is used for this purpose.

The outer cylinder of the oscillating system is usually grounded. In such a case, there should be a separating capacitor (which is shown in fig. 7.3 as a dielectric pacer) between this cylinder and the anode of the tube.

Anode voltage $+E_a$ is sually supplied to the tube radiator, structurally joined to its anode.

The tube is heated by ac voltage $\sim U_H$.

The grid is biased automatically with resistor R_g. Capacitive plunger 5 is used as a separting capacitance in the grid circuit.

In the pulsed mode in the meter range, power of SHF triode oscillators may reach several megawatts.

Efficiency of triode oscillators is comparatively low; up to 50% at wave lengths longer than 30 cm and lower at shorter wave lengths.

Frequency stability of tube SHF oscillators is high.

Data on SHF triode pulse oscillators is presented in Table 7.1.

Table 7.1

Basic data on pulse triode SHF generators

	1 Параметр	2 Тип лампы						
		3 ГИ-4А	4 ГИ-5Б	5 ГИ-6Б	6 ГИ-7Б	7 ГИ-14Б	8 ГИ-19Б	9 ГИ-24А
10	Напряжение накала, *в*	10	6,3	12,6	12,6	12,6	7,3	6,3
11	Ток накала, *а*	215±10	425±40	2,1	2	4,4	20	425±40
12	Напряжение на аноде, *кв*	35	26	9	9	20	14	27
13	Анодный ток, *а*	65±5	80±6	7,5	7,5	16±2	40±4	40±2
14	Напряжение приведения по аноду, *кв*	0,8	0,25	0,12	1,2	0,8	2,5	0,15
15	Крутизна лампы, *ма/в*	35±5	40±4	64	75	80±6	30	40
16	Крутизна линии критических режимов, *ма/в*	20±6	18±2	30	7,5	10	12±3	21±2
17	Проницаемость, %	2,5	4	1,5	1,5	,5±0,3	2	—
18	Колебательная мощность, *квт*	1200	1200	20	20	125	250	1000
19	Мощность рассеяния на аноде, *квт*	20	4,5	0,25	0,25	0,5	1	25
20	Мощность рассеяния на сетке, *вт*	800	400	2,5	2,5	5	30	1500
21	Максимальная частота, *Мгц*	150	202	1660	2700	1000	300	170
22	Максимальная длительность импульса, *мксек*	1000	11	10	10	8	10	800
23	Междуэлектродные емкости, *пф*:							
	C_{ak}	1,5	11±2	0,25	0,08	0,12	10	2
	C_{ag}	35	35	4,85	5	5,55	11	35
	C_{gk}	45	90	11,35	11	20	50	85

Key: 1 - parameter; 2 - tube type; 3 - GI-4A; 4 - GI-5B; 5 - GI-6B; 6 - GI-7B; 7 - GI-14B; 8 - GI-19B; 9 - GI-24A; 10 - heater voltage, V; 11 - heater current, A; 12 - anode voltage kv; 13 - anode current, A; 14 - anode reduction voltage, kv; 15 - tube transconductance, mA/V; 16 - slope of the critical mode lines, mA/V; 17 - permeability, %; 18 - oscillating power, kw; 19 - dissipated power on the anode; kw; 20 - dissipated power on the grid, w; 21 - maximum frequency, MHz; 23 - maximum pulse length, microseconds; 23 - interelectrode capacitance, pf.

7.3 Tetrode SHF Oscillators

In the lower part of the meter range and in the decimeter range of waves, tetrode oscillators are frequency used. They operate in a continuous mode and have large power output.

Tetrode oscillators use a common grid circuit. Ray tetrodes are often used. The second (screening) grid is connected by direct current to the anode, which produces great acceleration of the electrons in the region between the grids. This shortens the transit time of the electrons from the second grid to the anode, increases electronic efficiency, and consequently the overall efficiency of the oscillator. Tetrode oscillators have higher efficiency than triode oscillators.

Tetrode oscillators with all-metal ray tetrodes, the electrodes of which constitute part of a closed resonator, and oscillators as a whole, are a vacuum system, called resnatrons.

A diagram of a resnatron is shown in Figure 7.4. Its electronic system is a dual tetrode.

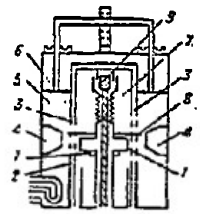


Figure 7.4. Diagram of a resnatron.

1 - cathodes; 2 - control grids; 3 - screen grid; 4 - anodes; 5 - anode-grid resonator; 6 - tuning plunger; 7 - cathode-grid resonator; 8 - capacitive stub; 9 - fine tuning element.

Capacitive feedback is accomplished with stub 8. Grid-cathode resonator has a capacitive fine tuning 9. Retuning is done with plunger 6, shortening the anode-grid resonator.

Power from resonators in the decimeter range reaches 60 kw with 40-60% efficiency in the continuous mode. The retuning range is about 20%.

7.4 Magnetrons

A multiresonator magnetron is an effective generating device in the decimeter and centimeter ranges.

A multiresonator magnetron (fig. 7.5) consists of a heated oxide cathode 1, anode block 2 with straps 3 and output 4 for high frequency energy output. Tunable magnetrons also have a tuning device. The space between cathode and anode block is called the interaction space.

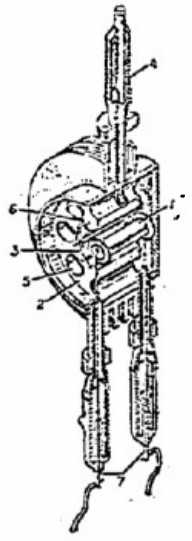


Figure 7.5. Design of a magnetron.

1 - cathode; 2 - anode block with resonators; 3 - straps; 4 - output device; 5 - resonator; 6 - slot; 7 - heater outputs.

Electron control in a magnetron is accomplished by the action of constant electric and magnetic fields on an electron stream. These fields act in mutually perpendicular planes (crossed fields). The electric field is directed radially from the anode unit to the cathode. The magnetic field, evenly distributed over the interaction space, is directed along the cathode.

The high frequency field of the resonators also takes part in controlling the electron flow. Electron flow in a generating magnetron is "spoke-shaped" fig. 7.6) and rotates in the interaction space. At the ends of the "spokes," electrons move in complex epicycloids.

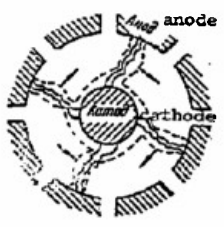


Figure 7.6. Rotating electron stream in a generating magnetron.

Moving electrons, acquiring kinetic energy from the pulse modulator, interact with the high frequency electric field of the resonators and supplement the energy of the field.

The degree of interaction of the electrons with the high frequency field, and consequently the energy which they give it, depend both on the amount of electrical voltage between anode and cathode and on the potential of the magnetic field. The high frequency field of the resonators in turn improves the conditions for energy transfer by the electrons to this field, grouping electrons around those resonators which most effectively interact with the resonator field.

The circuit of a magnetron generator is shown in Figure 7.7a. The anode unit is the magnetron casing, and is grounded, as a rule. High voltage rectangular pulses with negative polarity are applied from the pulse modulator to the magnetron cathode.

The magnetic field is created by an external permanent magnet.*

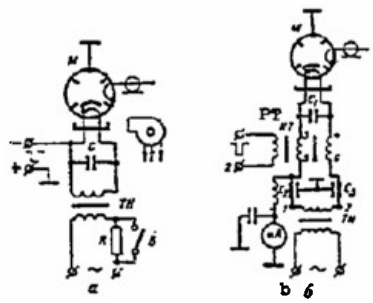


Figure 7.7. Schematic diagrams of a magnetron oscillator.
a - high voltage applied direct;y; b - high voltage applied from a modulator through a pulse transformer.

Magnetrons are cooled by ventilators or a liquid cooling system.

Other practical circuits for magnetron oscillators are more complex. Figure 7.7b shows a circuit where high voltage from the modulator is applied across a dual secondary winding (3-5; 4-6) of a pulse transformer. A milliammeter measures average current $I_{a\ av}$ of the magnetron, associated with the pulse equation

$$I_{a\ av} = \tau_p F_p I_a, \qquad (7.1)$$

* There are magnetrons in which permanent magnets are an integral part of the construction. Such magnetrons are called packed.

where

τ_p, F_p are pulse length and repetition frequency.

Operating characteristics of a magnetron. Electrical relationships in a magnetron determine its operating parameters: pulse anode voltage E_a and magnetic induction of the magnetic field B. These parameters influence anode current I_a, useful high frequency power P, power input P_0, efficiency η, and oscillating frequency f of the magnetron.

The operating characteristics are a family of graphical relationships of anode voltage E_a to anode current I_a with the following held constant:

magnetic induction;

oscillating power;

efficiency;

frequency.

They are determined with constant load on the magnetron.

Typical operating characteristics for a magnetron are shown in Figure 7.8.

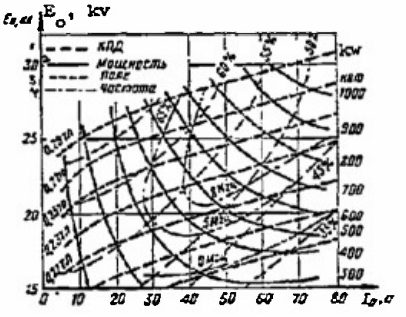


Figure 7.8. Operating characteristics of a magnetron.
1 - efficiency; 2 - power; 4 - field; 4 - frequency.

The relationship between magnetron current and voltage with constant magnetic induction is close to linear and is represented by almost horizontal lines, which are higher the higher the magnetic induction.

With constant B, I_a increases sharply with an increase in E_a. This is explained by the fact that with an increase in anode voltage, greater quantities of electrons leave the electron cloud surrounding the cathode, and take part in an energy exchange with the high frequency field of the resonators. Naturally, oscillating and output power increase also. Magnetron efficiency at first increases, and then, passing through its maximum, drops. Decrease in efficiency at low currents is explained by the fact that in these operating modes, the high frequency resonator fields are still comparatively weak and their bunching action on the spatial charge is small. Decrease in efficiency

at high currents is explained by the fact that, with a large current, mutual collisions of electrons in the spokes of the spatial charge increase; consequently overall grouping of the electrons is impaired.

Curves corresponding to P = const are reminiscent of hyperbolas, and indicate that, with an increase in voltage E_a and magnetic induction B, oscillating power increases.

Curves Δf = const show the relationship of magnetron frequency to its operating mode. They are complex in character and may vary widely from unit to unit of the same type of magnetron; for this reason they are not always given with the operating characteristics.

With the operating characteristics at hand, it is possible to foresee which changes in the operation of the transmitting device will involve change in this or that parameter.

Example. Let it be required to obtain oscillating power P = 900 kw from the magnetron whose operating characteristics are shown in Figure 7.8. Anode voltage E_a may be regulated from 0 to 30 kv. The magnetic system consists of a field with magnetic induction B = 0.27 tesla, which may be regulated within the limits ±0.02 tesla.

In such a case it is feasible to choose the following mode for the magnetron: B = 0.27 tesla; E_a = 27.5 kv; I_a = 62 amp.

This mode assures magnetron operation with sufficiently high efficiency (~53%) and a certain decrease in current.

Example. Let it be required to shift the magnetron of the preceding example from the mode specified to a mode with oscillating power P = 1000 kw.

It is best to increase anode voltage and magnetic induction simultaneously to E_a = 29 kv and B = 0.28 tesla. Now the pulse anode current will be $I_a \approx 64$ amp, and efficiency is held almost to its previous value.

If B remains equal to 0.27 tesla and only anode voltage is raised to E_a = 28 kv, the required oscillating power P = 1000 kw will only be obtained with a decrease in efficiency. Current through the magnetron in this case will rise to I_a = 74 amp, which exceeds the permissible limit ($I_{a\ lim}$ = 70 amp).

Using a magnetron with currents larger than the limiting values noticeably shortens its lifetime, and therefore is not recommended.

Electronic frequency shift. Electronic frequency shift (e.f.s) implies changing the generated frequency while changing the operating mode of the magnetron. It is evaluated by the e.f.s. coefficient k_f, which is determined as the ratio of frequency increments Δf to the corresponding pulse current increments, i.e., $k_f = \Delta f/\Delta I_a$.

In magnetrons it is estimated using the ratio

$$k_f = \frac{0.001 \div 0.003}{I_a} f\,[\text{MHz/amp}] \qquad (7.2)$$

where f and I_a are nominal values.

Depending on the range, in practice k_f = 0.05 to 0.3 MHz/amp. These values of k_f are considered large and reflect an intrinsic fault of magnetrons: the dependence of their frequency on the operating mode.

Load characteristics of a magnetron. Output power and oscillating frequency of a magnetron depend on the magnitude and character of the external load impedance.

A magnetron is always connected with the load by a coaxial or wave guide transmission line. In this case the load may be described by the modulus of the reflection coefficient (4.7) and its phase (4.6). Therefore, the relationship of generated power P and frequency f of the magnetron to load is graphically shown in polar coordinates as a Wolpert diagram (fig. 4.6).

A family of curves of P and f in polar coordinates with constant anode current and constant magnetic field potential is called the load characteristics (fig. 7.9). They are usually determined with lowered power, since large mismatches in the load produce large overloads in the transmission line. From the load characteristics it is evident that at the largest power output of the given magnetron, its frequency changes sharply with changes in phase and modulus of the reflection coefficient. In the range of smallest power for the given magnetron, its frequency stability is increased.

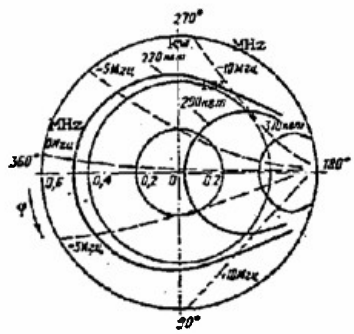


Figure 7.9. Load characteristics of a magnetron.

Considering equation (4.15), the load characteristics are often constructed in the plane, K_{swl}, phase of the reflection coefficient.

Frequency pulling of a magnetron. The phenomenon of a change in generated frequency produced by a change in load, is called pulling. It is evaluated by the frequency pulling factor F_{pull}.

The frequency pulling factor is the total change in frequency resulting from a 360° change in phase of the reflection coefficient where its modulus, equal to 0.2 (K_{swl} = 1.5) remains unchanged.

The factor F_{pull} is determined from the load characteristics. It is equal to the difference in frequencies of curves Δf = const, touching the circle which corresponds to a reflection factor p = 0.2. For example, for the magnetron whose characteristics are shown in Figure 7.9, F_{pull} = 12 MHz.

F_{pull} for magnetrons in the centimeter range is 7 to 15 MHz. This is an extremely large value and it illustrates the second fault of magnetrons: the strong dependence of their frequency on load.

When using magnetrons, the mode of the high frequency tract of the transmitting device must be carefully controlled so they do not operate with a reflection factor $|p| > 0.2$.

When a magnetron is replaced, the feeder system must be carefully adjusted to result in the lowest possible value of reflection factor.

Table 7.2 shows characteristic data for typical operating modes of pulse magnetrons.

Table 7.2

Characteristic data for typical operating modes of pulse magnetrons

1 Частота (ориентировочно), Мгц	2 Выходная мощность в импульсе, квт	3 E_a, кв	4 I_a, а	5 Минимальный к. п. д., %	6 Скважность
1200 (λ = 25 см)	10000	70	350	39	550
	2500	38	130	50	660
	600	30	60	38	500
3000 (λ = 10 см)	5000	70	125	56	1000
	1150	30	70	55	835
	250	22	30	33	500

Key: 1 - frequency (provisional), MHz, 2 - output power in a pulse, kv; 3 - E_a, kv; 4 - I_a, amp; 5 - minimum eff., %; 6 - off-duty factor.

Tunable magnetrons. The frequency of the oscillations generated by the magnetron is determined by the parameters of its oscillating system. The oscillating system of a magnetron consists of the resonators in the anode block, straps, the interaction space, and the end space. Its principal parts are the resonators and straps. Changing the parameters of the resonators and straps permits tuning of the magnetron in the frequency range.

High power magnetrons have a mechanical frequency retuning device, which consists of movable elements coupled with the resonant system and changing its reactive parameters. It produces changes in capacitance of the straps, in the inductance and capacitance of the resonators, the reactive parameters of additional tuning elements.

A tuning mechanism which operates by changing the capacitance of the straps is shown in Figure 7.10.

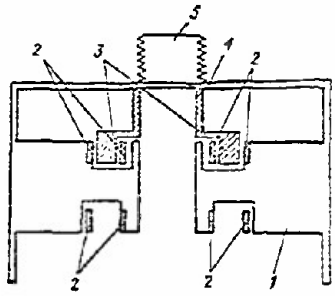


Figure 7.10. Tuning elements of a magnetron.

1 - anode block; 2 - dual straps; 3 - fine adjustment ring; 4 - movable vacuum seal; 5 - retuning screw.

In a magnetron with dual ring straps 2, the capacitance of the latter is changed by the ring 3, introduced between the straps. Inserting the ring between the straps is equivalent to decreasing the distance between them, and consequently it leads to an increase in capacitance of the straps and the whole oscillating system of the magnetron.

By changing the capacitance of the straps, the magnetron frequency may be adjusted within the limits of 5%. By simultaneously changing the inductance of the resonators, and the capacitance of the slot and straps, it is possible to tune magnetrons within 10-15% of the center frequency.

7.5 Klystron Oscillators

Multiresonator klystrons are widely used as SHF devices with very large power output.

A contemporary four-cavity klystron is diagrammed in Figure 7.11.

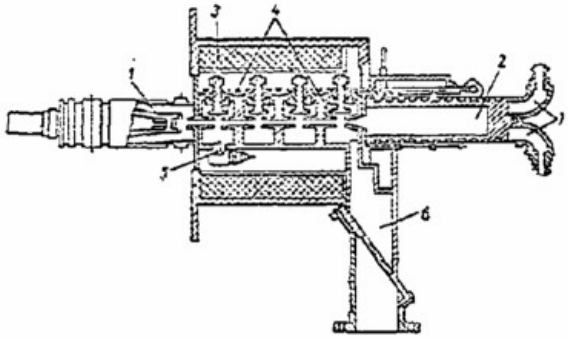


Figure 7.11. Multiresonator klystron.

1 - electron gun; 2 - catcher; 3 - focusing winding; 4 - resonators; 5 - input device; 6 - output device; 7 - input and output guides.

An electron stream, leaving electron gun 1, is influenced by the potential between catcher 2 and the cathode. The stream is focused by the magnetic field of the electromagnet windings 3. The walls of resonators 4 are transparent to the electron stream.

In the first resonator, the electron stream velocity is modulated by a high frequency field, created by a high frequency voltage applied to the klystron input. In the drift tube between the first and second resonators, the velocity modulation of the electron beam results in density modulation. The electron stream, whose density varies harmonically, sets up high frequency oscillations as it moves in the second resonator. The power of the oscillations, with given construction parameters and klystron mode, is considerably larger than the power of the oscillations in the first resonator.

In the second resonator, velocity modulation of the electron stream, resulting in density modulation in the drift tube between the second and third resonators, is intensified.

The electron flow, velocity modulated and with its energy increasing from resonator to resonator, produces high frequency oscillations in the last resonator which are fed through output device 6 to the feeder system.

Klystron amplification depends on the number of resonators and amounts to 15 to 90 db. Klystron efficiency is 40% and higher.

Klystrons operate in all ranges of SHF. Power in the 10-centimeter range reaches 40 Mw in the pulsed mode and several kilowatts in the continuous mode.

Multiresonator klystrons in the decimeter range are known to have an output power of 10 Mw pulsed, tunable over a 140 MHz band. Lifetime is on the order of 300 h.

Frequency stability of transmitters using klystron generators may be high, since it is determined by the frequency stability of the low power exciters.

Klystron oscillators also have inherent faults. The basic faults are: narrow band, difficulty in tuning over the band, very high operating voltages (to 400 kv), excessive size and weight.

The circuit of a klystron oscillator is shown in Figure 7.12. The klystron resonators and catcher are connected directly to the casing. AC heater voltage is applied through step-down transformer Tr_1.

A negative pulse of the modulating voltage is taken from the secondary winding of pulse transformer PT and led to the klystron cathode.

Focusing windings are supplied from a rectifier with voltage E_B.

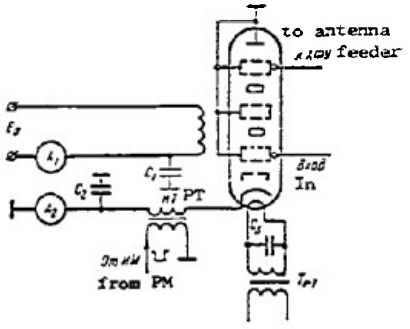


Figure 7.12. Schematic diagram of a klystron generator.

Ammeter A_2 serves to monitor average current of the klystron, and ammeter A_1 monitors current in the focusing windings. C_1, C_2, and C_3 are blocking capacitors.

Characteristics of multiresonator klystrons. The energy characteristics of amplifying klystrons are the operating and amplifying characteristics.

The operating characteristic of a klystron is defined as the relationship of output power to anode voltage E_a (fig. 7.13a) with constant excitation power. Output power is associated with the accelerating voltage by a cubic function. The same figure shows the relationship of efficiency η to voltage E_a.

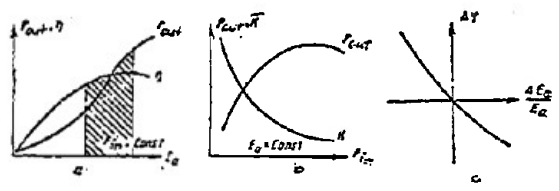


Figure 7.13. Characteristics of a multiresonator klystron.

a - operating characteristic; b - amplifying characteristic; c - phase characteristic.

Output power, efficiency, and accelerating voltage are related by the equation

$$P_{out} = \eta \mathcal{R} E_a^{3/2}, \tag{7.3}$$

where

$\mathcal{R}$ is conductivity of the electron beam in the klystron at saturation $\left[\mathcal{R} \approx (1 \div 1.1)10^{-6} \frac{a}{\sqrt{\frac{a}{c}}}\right]$

The region of practical operating modes is indicated on the operating characteristic. Outside this region, efficiency decreases, the catcher overheats, and x-ray radiation intensity increases.

Amplification characteristic is defined as the dependence of output power P_{out} on excitation power P_{in} (fig. 7.13b) with constant anode voltage E_a.

Since the power amplification factor, expressed in decibels, is determined by the ratio $K = 10 \log (P_{out}/P_{in})$, the amplification characteristic is sometimes graphed as $K = f(P_{in})$.

The phase characteristic of a klystron is defined as the dependence of the phase shift in oscillations at the input on anode voltage. It is usually represented as the relationship of a phase shift increment $\Delta\varphi°$ to a relative change in anode voltage (fig. 7.13c), i.e.,

$$\Delta\varphi = f\left(\frac{\Delta E_a}{E_a}\right),$$

where

ΔE_a is an increment of anode voltage.

The phase properties of a multiresonator klystron are determined by the phase pulling factor k_φ, indicating the ratio of an increment of phase shift $\Delta\varphi$ to an increment of anode voltage causing it, i.e., $k_\varphi = \Delta\varphi/\Delta E_a$.

For typical klystrons

$$k_\varphi = \frac{(4 \div 5)\,180}{E_a}\,[\text{deg/v}] \qquad (7.4)$$

It is evident from the phase characteristic that the stability of an advance in phase of oscillations at the output in relation to phase at the input in pulsed multiresonator klystrons depends on stability of the amplitude of anode voltage pulses.

Multiplying klystrons. Multipliers are klystrons in which the oscillation frequency at the output is a multiple of the exciting oscillations.

Klystron multipliers are similar in construction and operating principle to multi-cavity amplifying klystrons. The output resonator of a klystron is tuned to a frequency several times higher than the input frequency. Frequency multiplication in known klystrons amounts to 2 to 10. Their gain is usually close to one.

Klystron multipliers may be used in exciters for multistage radio transmitting devices.

Traveling Wave Devices

In recent years SHF generating devices in which an extended electron stream interacts with waves propagating along nonresonant decelerating systems have found increasingly wider application. These devices have been designated traveling wave tubes.

There are two classes of traveling wave tubes: type O (ordinary) and type M (magnetron) tubes.

In ordinary traveling wave tubes, the electron stream is focused by a longitudinal magnetic field, and in magnetron traveling wave tubes, it is focused by transverse (crossed) electric and magnetic fields. Each type includes forward and backward traveling wave tubes, depending on whether the direction of the electron beam deflection coincides with the direction of the traveling wave or is opposite to it.

7.6 Type O Forward Traveling Wave Tubes

In the literature, these devices are often simply called traveling wave tubes. This type of tube operates in an amplifying mode; its layout is shown in Figure 7.14.

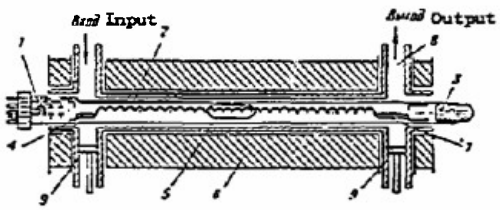


Figure 7.14. Construction of an ordinary traveling wave tube.
1 - electron gun; 2 - decelerating helix; 3 - collector; 4 and 7 - couplers; 5 - absorber; 6 - focusing winding (solenoid); 8 - output device; 9 - tuning plunger.

The cathode of the electron gun 1 emits electrons, which under the influence of a potential on the accelerating electrodes, helix 2, and collector 3, form an electron stream, moving inside the helix with velocity $v_e \ll c$.

Input signals and coupling element 4 set up in the decelerating helix an electromagnetic wave which, propagating with velocity $v_\phi \approx v_e$ toward the output end 7, interact with the electron stream and are amplified. High frequency oscillations are taken out of the tube through output waveguide 8. The extended electron stream is focused by solenoid 6. To prohibit self-oscillation of the tube, a local absorber 5 is mounted on several screws in the center of the helix.

At the present time many industrial types of these tubes have been developed which are used in amplifying stages of receiving devices. The wide practical application of traveling wave tubes is attributed to their wide-band operation, low noise levels, high gain, and also simple design.

Faults of traveling wave tubes are their low efficiency (less than 20%) and unwieldy construction.

Table 7.3 illustrates the basic parameters of some traveling wave tubes.

Table 7.3

Basic parameters of certain traveling wave tubes.

	Parameter Параметры	Tube type Тип лампы		
		УВ-5UV-5	УВ-6UV-6	УВ-7UV-7
1	Рабочий диапазон частот, *Мгц*	3400—4000	3400—4000	3400—4000
2	Коэффициент усиления, *дб*	18	30	26
3	Коэффициент шума, *дб*	8		
4	Мощность входного сигнала, *вт*	10^{-6}		
5	Наибольшая выходная мощность, *вт*	—	0,03	3
6	Напряжение накала, *в*	2—3	3—4	6,3
7	Ток накала, *а*	0,5—0,9	0,7—0,95	0,7—0,85
8	Напряжение управляющего электрода (=), *в*	12	—30	—50
9	Напряжение первого анода (=), *в*	5—150	150—500	1100—1400
10	Напряжение второго анода (спирали), *в*	400—500	800—1100	
11	Напряжение коллектора (=), *в*	600	1300	1500
12	Ток второго анода, *мка*	30	350	3000
13	Ток коллектора, *ма*	1	2,5—4,5	20—35
14	Магнитная индукция соленоида, *тл*	0,07	0,07	0,08
15	Долговечность, *ч*	1000	1000	1000

Key: 1 - operating frequency range, MHz; 2 - gain, db; 3 - noise factor, db; 4 - input signal power, w; 5 - greatest output power, w; 6 - heater voltage, v; 7 - heater current, amp; 8 - control electrode voltage (=), v; 9 - first anode voltage (=), v; 10 - second anode voltage (helix), v; 11 - collector voltage (=), v; 12 - second anode current, microamperes; 13 - collector current, ma; 14 - magnetic induction of the solenoid, tesla; 15 - lifetime, h.

Ordinary forward traveling wave tubes may be low, average, or high power. Average and high power tubes are used in multistage radio transmitting devices.

7.7. Type M Forward Traveling Wave Tubes

In the literature, these devices are often called magnetron amplifiers, as they operate in amplifying modes. Figure 7.15 shows a schematic layout of a tube with cylindrical construction.

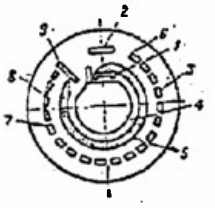


Figure 7.15. Magnetron traveling wave tube.

1 and 2 - cathode and accelerating electrode of the electron gun; 3 - deceleration system; 4 - negative electrode ("bottom"); 5 - electron beam; 6 - input; 7 - output; 8 - absorber; 9 - collector.

Electrons emitted by cathode 1 of the electron gun with the accelerating electrode 2, under the influence of the magnetic field set up by an external magnet and the voltage between the decelerating systems 3 and negative electrode 4, form the electron beam 5.

Electrons in the beam move in the same way as in a magnetron at the ends of the spokes.

Amplified oscillations are applied to input 6 of the decelerating system. Waves along the axis of the decelerating system propagate with phase velocity v_ϕ, approximately equal to the velocity of electron motion in the ray v_e. (This velocity is much smaller than the velocity of light). As a result of coupling the field of the traveling wave and the electron stream, the electromagnetic wave is amplified in proportion to its propagation from the input to the output of the decelerating system. Extremely powerful oscillations are obtained at output 7. Pulsed amplifiers of this type produce output power of several megawatts with efficiency greater than 60%. Their gain is 15 db within 25% limits.

The tubes function on comparatively low acceleration voltages, so multistage radio transmitters with excellent frequency stability can be built.

7.8. Type O Backward Traveling Wave Tubes

These tubes may be used as amplifiers and as generators.

In backward wave tubes* (fig. 7.16) periodic deceleration systems 3 are used, similar to the systems in a traveling wave tube. The field of the electromagnetic wave along the deceleration system is nonharmonic. It may be considered as the sum of spatial harmonic components with varying amplitudes and different phases.

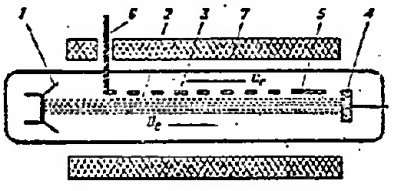


Figure 7.16. Layout of a type O backward wave tube.

1 - electron gun; 2 - electron stream; 3 - deceleration system; 4 - collector; 5 - absorber; 6 - output; 7 - focusing winding.

---

* Backward wave tubes are called carcinotrons from the Latin word "carcino" - back.

In a backward wave tube, electron beam 2 interacts with the first or second backward spatial harmonic. If the tube mode is correctly chosen disregarding the fact that the electromagnetic wave in the deceleration system is directed opposite to the electron beam, electrons are deflected synchronously with changes in phase of one of the backward harmonics.

Modulation of the electron beam in a backward wave tube is accomplished with both forward and backward harmonics. Percentage modulation of the beam increases in the direction from collector 4 to the electron gun 1. More dense bunches of electrons give greater energy to the electromagnetic wave in the deceleration system, and therefore at output 6 an amplified high frequency signal is received.

In generating and amplifying backward wave tubes, the physical processes of coupling the electromagnetic wave with the electron beam are similar, since after the tube is excited the spatial wave in the deceleration system will be similar to the wave in an amplifying tube.

Excitation of a backward wave tube, and consequently generation of high frequency oscillations are possible since the motion of an amplified electromagnetic wave opposite to the electron beam assures a feedback connection of the beam with the field of the deceleration system. Feedback in a backward wave tube is distributed in character and is effected along the whole tube.

Ordinary backward traveling wave tubes have wide-band properties. They permit electron tuning with a closed range of 3:1. These devices are practical even in the millimeter range. Most frequently generators of this type are used in a continuous-wave mode. Their efficiency is low, on the order of 1%.

7.9 Type M Backward Wave Tubes

Magnetron backward wave tubes (type M carcinotrons) are principally used both as generators and as amplifiers. Generating carcinotrons, like that shown in Figure 7.17, have undergone extensive development.

The structural elements of carcinotrons of this type are the same as in the ordinary backward traveling wave tube described above, although the deceleration systems are cylindrical. The magnetic field for controlling the electron beam is set up along the axis of the tube similarly to the field in a magnetron. The electron beam moves between the deceleration system and a special negative electrode ("bottom") opposite to the electromagnetic wave propagating along the deceleration system toward the tube output. The physical processes in a magnetron backward traveling wave tube are basically no different from the processes going on in a type O carcinotron.

Magnetron backward traveling wave tubes are quite effective generating devices in the decimeter, centimeter, and millimeter wave ranges. Their efficiency may be as high as 50%. They have high range characteristics, and

are easily retuned in a wide range of frequencies. In the 10-centimeter range their power output amounts to several hundred kilowatts pulsed and several kilowatts continuous wave.

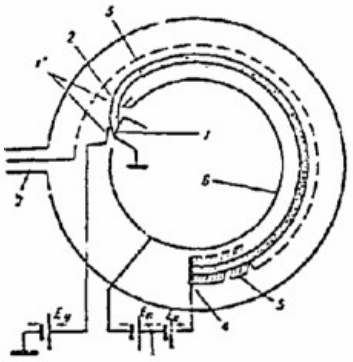


Figure 7.17. Type M backward wave tube layout.

1, 1' - cathode and acceleration electrode of the electron gun; 2 - electron beam; 3 - deceleration system; 4 - collector; 5 - absorber; 6 - negative electrode; 7 - output.

These devices are distinguished by exceptionally high noise levels (up to 25 db), therefore they are used as noise generators.

7.10 Carmatrons

On the basis of type M backward wave tubes and magnetrons, combined SHF generating devices have been developed and given the name carmatrons.*

In a carmatron, instead of an electron gun, cylindrical cathodes are used, positioned as in a multiresonator magnetron coaxial with the deceleration system. The advantage carmatrons have over type M backward wave tubes is a simplified construction since no electron gun is required. A more powerful cathode permits an increase in power in the device.

Bandwidth characteristics of carmatrons are somewhat worse than type M backward wave tubes.

Carmatrons inherited a fault of carcinotrons: strict dependence of frequency on supply voltage. For this reason, a highly stable voltage source is required for continuous-wave carmatrons, and, for pulsed carmatrons, pulse modulators are also required to form voltage pulses with an ideally stable base frequency.

---

* From the initial syllables of carcinotron and magnetron.

7.11 Platinotrons

Platinotrons were developed on the basis of magnetrons and are similarly constructed (fig. 7.18). Deceleration system 1 of a platinotron is similar to the anode block of a magnetron. Cylindrical activated cathode 2 is positioned coaxially with the deceleration system.

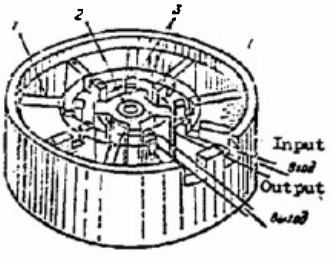


Figure 7.18. Platinotron construction.
1 - deceleration system; 2 - cathode; 3 - straps.

The electron beam is controlled by permanent magnets and magnetic fields are set up in exactly the same way as in a magnetron. The electron beam in platinotrons is usually coupled with a backward wave, but it may be used with a forward wave.

Platinotrons used as amplifiers are called amplitrons. Amplitrons are distinguished from magnetrons by the fact that their deceleration system is not closed. This is accomplished by openings in the straps 3. These openings are coupled to the input and output, which are carefully matched to the external mechanisms (transmission lines).

SHF oscillations applied to the amplitron input are amplified. Passing a signal through the device in the amplification process is accompanied by a predictable phase shift in the oscillations.

The dependence of output power on input power for various amounts of supplied power P_0 for one type of amplitron is shown in Figure 7.19. With the lowest supplied power (P_0 = 500 kw) the area a-b is applicable, where output power is only weakly dependent on input high frequency power.

If input power is too small, then frequency of the output oscillations does not depend on frequency of the input oscillations, i.e., control of the oscillations is lost.

The greater the power supplied from the modulator, the greater will be the required power of the input oscillations to control amplification.

In the region where the input signal controls the output, amplification is greater than 10 db with low values of input power. With an increase in power of the input oscillations, gain decreases, although dependence of output power on input power is more pronounced.

The operating characteristics of an ampliftron (fig. 7.20) are similar to those of a magnetron.

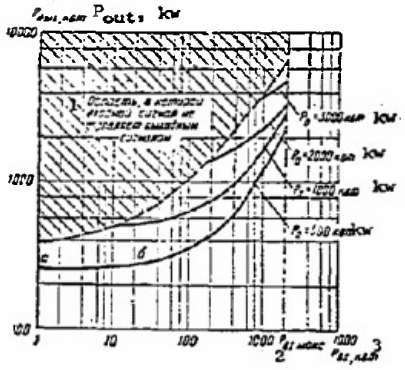


Figure 7.19. Amplifying characteristics of an amplitron.

1 - region in which input signal does not control output signal; 2 - $P_{in\ max}$; 3 - P_{in}, kw.

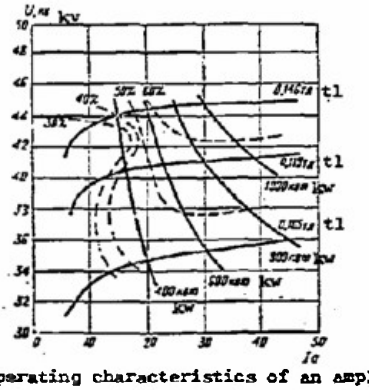


Figure 7.20. Operating characteristics of an amplitron.

Amplitron efficiency is determined as the ratio of the difference of output, P_{out}, and input, P_{in}, high frequency power to power supplied from the modulator, i.e.,

$$\eta = P_{out} - P_{in}/P_0, \tag{7.5}$$

where

$P_0 = E_a I_a$ is applied power.

It is important to notice that input power in an amplitron is not lost, but passes to the output and is completely used.

Amplitron efficiency increases with an increase in anode current and magnetic field. Practically values reach 65% and up, while remaining constant in a frequency band greater than 10%.

Table 7.4 lists the parameters of certain amplitrons, operating in the pulsed mode.

Table 7.4

Fundamental parameters of amplitrons

Parameters Параметры	Type of Тип амплитрона amplitron			
	QK 562	QK 653	QK 783	QK 622
1 Диапазон, *Мгц*	1225—1350 (механическая перестройка) 2	1280—1350	2700—2800	2000—3100
3 Выходная мощность, *квт*	800	4000	3000	3000
4 Анодное напряжение, *кв*	40	94	50	55
5 Анодный ток, *а*	35	78	65	65
6 Коэффициент усиления, *дб*	15	10	10	10
7 Скважность	300	1400	200	200

Key: 1 - tuning range, MHz; 2 - (mechanical tuning); 3 - output power, kw; 4 - anode voltage, kv; 5 - anode current, amp; 6 - gain, db; 9 - off-duty factor.

An amplitron reproduces an input signal very well. In the limits of possible operating modes, the signal is unchanged. Phase shift of high frequency oscillations on passing through the amplitron is only weakly dependent on anode current.

Investigations show that, even with a change in K'_{sw1} from 1 to 2.5 and variation in phase of the reflection coefficient in wide limits, the form of the spectrum of the amplified oscillations remains practically unchanged at the amplitron output; this is its greatest advantage over the magnetron.

Feedback elements may be introduced between output and input of a platinotron. Feedback in a platinotron assures its self-excitation. A high-Q resonator is introduced into the feedback loop to stabilize the frequency of the self-oscillations.

Platinotrons, together with external feedback elements and frequency stabilization are called stabilitrons.

Stabilitrons are new devices, and, as reported in the literature, their structural development is still in progress. In certain stabilitrons, feedback is accomplished with a special reflector at the output and frequency stabilization is accomplished with an external resonator at the input.

In such a case, feedback and frequency stabilization are accomplished as follows. Part of the high frequency energy is reflected from a mismatched output of the stabilitron and is returned through the device to a resonator. This energy sustains oscillations in the input resonator.

Because the input resonator has a high Q, it maintains oscillations of a strictly determined frequency. The external resonator may have a high Q, and its stabilizing properties will be extremely high.

Operating and load characteristics of a stabilitron are similar to magnetron characteristics. Electron displacement in it is 50-100 times less and frequency pulling is one order of magnitude less than in a magnetron.

Stabilitron parameters are shown in Table 7.5.

Table 7.5

Basic stabilitron parameters

Parameters	Type of stabilitron	
	QK 434	QK 526
1 Диапазон, Мгц (перестройка механическая)	1225—1350	1260—1350
2 Выходная мощность, квт	600	750
3 Магнитная индукция, тл	0,115	0,115
4 Анодное напряжение, кв	36	36
5 Анодный ток, а	37	40
6 К.п.д., %	—	52
7 Электронное смещение частоты, кгц/а	—	1—4
8 Коэффициент затягивания частоты, Мгц	—	0,04—0,5

Key: 1 - bandwidth (MHz) (mechanical retuning); 2 - output power, kw; 3 - magnetic induction, tesla; 4 - anode voltage, kv; 5 - anode current, amp; 6 - efficiency, %; 7 - electron frequency shift, kHz/Amp; 8 - frequency pulling factor, MHz.

7.12 Generators Using Semiconductor Devices

Junction semiconductor triodes and tetrodes find practical application in generators of high frequency oscillations.

Continuous-wave generators use common emitter, common base, and common collector circuits. The first two circuits are most often used.

Figure 7.21 illustrates a common emitter generator circuit using a p-n-p triode.

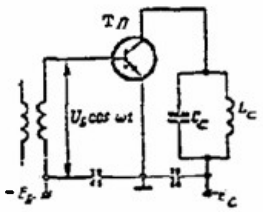


Figure 7.21. Schematic of a continuous-wave generator using a semiconductor triode.

Under the influence of excitation voltage U_{ex} on the first junction (p-n), barrier potential in the junction charges. The fundamental process in the first junction (emitter-base) is a repulsion of holes by the emitter and free electrons by the base. Forward current across this junction will be essentially hole current, since holes are injected into the base by the emitter.

The number of holes injected by the emitter is determined by the potential of the base relative to the emitter; i.e., emitter current depends on the voltage between base and emitter. If this voltage is harmonic, current in the emitter-base junction will also be harmonic.

Holes, diffusing through the base, reach the second junction (n-p), decrease the impedance of the base-collector junction, and collector current increases. In an energy exchange with the oscillating resonator L_c, C_c contributes a large number of electrons which supplement the energy of the cavity. Power of the oscillations in the collector circuit (in the cavity L_c, C_c) is tens and hundreds of times greater than the power expended by the excitation source. In this way high frequency oscillations are amplified.

Naturally, self-excited oscillators may be designed using semiconductor devices. The circuit of a self-excited oscillator with an autotransformer coupling with the common emitter, parallel base supply, and series collector supply is shown in Figure 7.22. A secondary negative voltage is applied to the base to choose the reference operating point on the part of the characteristic with a steep slope. It is taken off the supply source through potentiometer R_2, R_3. For available semiconductor triodes, a small initial bias is required, on the order of fractions of a volt; therefore resistance R_2 is small. The fundamental resistance in self-biasing the base is resistor R_1. Capacitor C_1 is an element of base bias. The remaining elements of the self-excited oscillator are designated the same way as elements from the theory of self-excited tube oscillators.

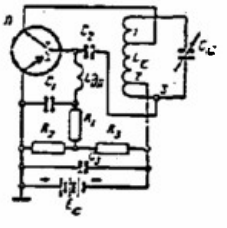


Figure 7.22. Schematic of a self-excited oscillator using a semiconductor triode.

Self-excited oscillators with capacitive and inductive coupling are also practical.

Semiconductor generators are coming into wider use. Small size and weight of semiconductors combined with miniature radio components permit compact designs. Their simplicity and reliability are high. Special designs can withstand accelerations produced during launch of a rocket or space vehicle.

Quantum Molecular Devices

Depending on the range of wave used, quantum devices are classed as SHF and optic quantum devices.

The following are classed among SHF quantum devices:

molecular (atomic) generators;

molecular amplifiers;

paramagnetic amplifiers.

In foreign literature, all these devices have the common name "maser."* In Soviet literature they are usually called quantum molecular generators and quantum molecular amplifiers.

Included in optic quantum devices are generators and amplifiers operating in the infrared (0.4 mm to 0.76 microns), visible (0.76 microns to 0.4 microns) and ultraviolet (0.4 microns to 20 Å) ranges. In foreign literature and often in Soviet literature, these devices are called "lasers"** or optic masers.

These devices are more often called optic quantum generators in Soviet publications.

For research in the field of quantum radiophysics and for building the first quantum molecular generators, Member Correspondents of the Academy of Sciences of the USSR N. G. Basov and A. M. Prokhorov won the Lenin prize in 1960. For the same scientific research these Soviet scientists (and the American physicist Charles H. Townes) were awarded the Nobel prize in 1964.

Optic quantum generators are developing rapidly at the present time. According to the operating principle and physical properties of the active substance, they are divided into many groups and types, the most important of which are solid body, gas, liquid, and semiconductor optic quantum generators.

7.13 Quantum Molecular Generators and Amplifiers

The operating principle of quantum devices is based on interaction of an electromagnetic field with atoms or molecules of certain substances which possess properties such that under certain conditions the internal energy of their atoms (molecules) is turned into electromagnetic energy. Such an energy transformation occurs where atoms (molecules) jump from one energy level to another, giving off quanta (portions) of electromagnetic energy of a given frequency.

In certain substances, which are called active, under certain conditions atoms (molecules), absorbing energy quanta, jump to a higher energy level, i.e., to an excited state. Excited atoms may spontaneously radiate (spontaneous radiation) their acquired energy and return to the original or some intermediate energy state.

---

* The word "maser" is formed from the initial letters of the English words "microwave amplification by stimulated emission of radiation."

** "Laser" is made up of the initial letters of the phrase "light amplification by stimulated emission of radiation."

Excited atoms (molecules) may be "forced" to radiate electromagnetic energy by the effect of an external electromagnetic field. Radiation caused by an external electromagnetic field is called inducted radiation.

Atoms (molecules) are shifted to an excited state by introducing electromagnetic energy into a system containing an active substance. This process is called pumping. Accordingly, the frequency of the excitation electromagnetic field is called the pumping frequency.

The quantity of atoms located at various levels or the population of the energy levels is determined by the temperature of the active substance. In quantum generators conditions are set up under which the population of higher energy levels is higher. In such a case even a weak external electromagnetic field at a given frequency will produce intensive radiation caused by transfer of atoms from a higher into a lower energy level.

The energy level of an active substance is quantitatively determined by the population of the level, i.e., by the number of atoms (molecules) and the energy of each atom (molecule). The usual distribution of atoms (molecules) in the energy levels of a substance at thermodynamic equilibrium determined at a specified temperature is shown in Figure 7.23a. It satisfies the Boltzmann equation

$$\frac{N_m}{N_n} = e^{-\frac{W_m - W_n}{kT}}, \qquad (7.6)$$

where

N_m is the number of molecules (atoms) in the upper level with energy W_m;

N_n is the number of molecules (atoms) in the lower level with energy W_n;

$k = 1.38 \cdot 10^{-23}$ j/1°K is the Boltzmann constant;

T is temperature of the active substance in degrees Kelvin.

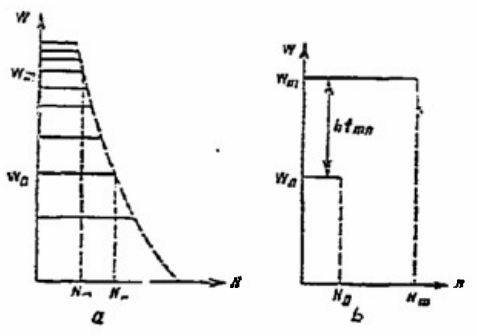


Figure 7.23. Energy levels of an active substance.

a - at thermodynamic equilibrium; b - when the substance is excited.

To create generation or amplification, thermal equilibrium of the substance is disrupted so that the population of the upper energy level is greater than that of the lower (fig. 7.23b).

Atoms in active substances have not two, but many energy levels. Accordingly, from one energy level an atom may jump to various energy levels, and each jump corresponds to a unique wave length of radiated oscillations. Consequently, spectra of energy radiated by atoms and molecules are studied.

The wave length for a given jump is equal to

$$\lambda_{m,n} = \frac{ch}{W_m - W_n}, \tag{7.7}$$

where

m, n are any possible (permitted) jumps;

W_m, W_n are energies of the corresponding levels;

c is propagation velocity of the electromagnetic energy;

$h = 6.6 \cdot 10^{-34}$ j·sec is Planck"s constant.

In the first quantum generator designed by N. G. Basov and A. M. Prokhorov molecules of ammonia were used as an active substance. Recently quantum generators have been developed which use atoms of hydrogen. Such a generator is sketched in Figure 7.24. Its basic elements are a source of atomic hydrogen 1, an atom sorting device 2, and a resonator 3.

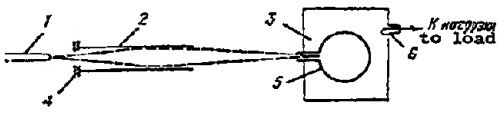


Figure 7.24. Design of a quantum generator using atoms of hydrogen. 1 - source of hydrogen atoms; 2 - device for sorting atoms according to their state; 3 - resonator; 4 - diaphragm; 5 - accumulating tube; 6 - output.

Atoms of hydrogen, obtained by high frequency discharge, pass through diaphragm 4 and fall in a narrow beam on the sorting device, which discriminates on the basis of energy level.

Two reference energy levels, conventionally designated 1 and 0, are used (fig. 7.25). In the field of a permanent magnet, the upper level is split into three. Atoms of the two upper levels (1-1; 1-2) are focused so that they enter the opening of accumulating tube 5 (fig. 7.24), which is located in the center of a high frequency resonator tuned to the frequency of the jump from level 1-2 to the zero level (1420 and 405 MHz).

Electromagnetic oscillations, set up in the resonator by induced radiation of hydrogen atoms, are led through the output device 6 to the load.

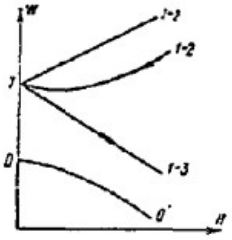


Figure 7.25. Energy levels of atomic hydrogen and their change due to the Zeeman effect.

The sorting device is a six-pole permanent magnet, which sets up a nonuniform field.

Accumulating tube 5 has a diameter of about 10 cm, and is made of quartz. Its inner surface is covered with teflon (a fluorine hydrocarbon compound), chemically inert to hydrogen atoms, which decreases the flow of atoms through the walls.

The high frequency resonator has the usual structure. Its inner surface is silvered.

The coupling loop with the load is regulated. It may be set up in such a manner that generation is pulsed or continuous-wave. Coupling can be increased to the extent that generation becomes impossible; then the quantum device will operate as an amplifier. Transition from oscillation to amplification may also be effected by changing the intensity of the hydrogen atom beam. Frequency stability of a hydrogen quantum generator is on the order of 10^{-13}. This stability is assured if the generator temperature is maintained with an accuracy of 0.01°C.

Hydrogen generators have the outstanding advantage that they can be used at room temperature and do not require complicated cooling mechanisms.

7.14 Optic Quantum Generators (Lasers)

Optic quantum generators have three basic elements: active substance, (possessing properties of induced radiation), pumping source (exciting the active substance), and a resonant system.

The active substance may be a solid body, gas mixtures, or liquid. Ruby with chromium additives and fluorite with uranium or samarium additives are the solids used.

The circuit of a ruby laser is shown in Figure 7.26. The active substance is a monocrystal of red ruby, which is aluminum oxide Al_2O_3 with about 0.05% chromium added. The ruby crystals are made in the form of a rod 1. The rod is usually about 4-5 cm long and 0.5-1.0 cm in diameter. The rod faces are plished and covered with silver. One face is opaque, the other is about 10% transparent.

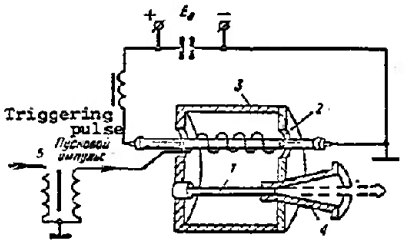


Figure 7.26. Construction of a ruby laser.

1 - ruby rod; 2 - xenon pulse tube; 3 - reflector; 4 - ruby holder; 5 - pulse transformer.

Pumping energy is produced by a pulsed xenon tube 2, which produces intensive light. The ruby is mounted parallel to the tube inside the same reflector 3 or inside the helix of the pulse tube.

The ruby rod itself serves as the resonant system of the generator.

The operating principle of an optic quantum generator is as follows. With each pulse from the pulse tube, chromium ions, absorbing light quanta with wave length 5600 Å, are excited and jump from reference level 1 to energy levels lying in region 3 (fig. 7.27).

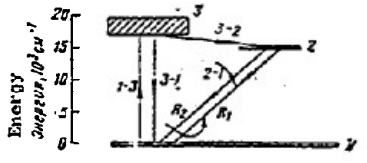


Figure 7.27. Energy levels in ruby.

1-3 - chromium ion jumps under excitation; 3-1 - spontaneous jumps of excited ions; 3-2 - jumps in an intermediate reference level; 2-1 - permitted jumps with radiation in two spectra R_1 and R_2.

Immediately after this, some of the chromium ions return to level 1, but the majority of them jumps to level 2. These ions remain in an excited state for a short time, and then jump to the reference state at level 1. Jumping from level 2 to level 1, the ions radiate light at about 6943 Å wave length.

Naturally, each ion emits a photon in going from level 2 to level 1, and each radiated photon causes other chromium ions to jump from level 2 to reference level 1. In other words, the photons induce (cause) radiation of other photons, and in a short time they form a large beam which propagates along the ruby rod. The opaque mirror on one face of the rod reflects the

light wave back into the ruby rod. When the light intensity level inside the rod is low, even the transparent face partially reflects the light. When the light reaches a determined intensity, its powerful beam pours out through the semitransparent face.

Photons leave the ruby through the side walls without reflection, and only those oscillations which propagate parallel to the axis of the ruby rod are amplified. The light beam from an optic quantum generator is highly directional and, because of the coherence (each succeeding photon is in phase with the preceding) of the processes in the ruby, its intensity is extremely high.

Ruby lasers are pulsed. Pulse length (fig. 7.28) is determined by the power of the pumping source and processes in the ruby itself. Pulse repetition frequency is limited to several pulses per minute, since it determines the temperature of the ruby crystal. In principle it is possible to make a continuous-mode crystal laser.

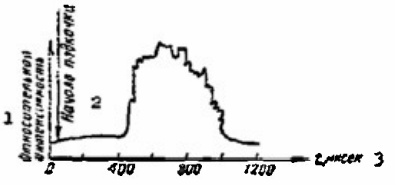


Figure 7.28. Pulse at the output of an optic quantum generator.
1 - relative intensity; 3 - beginning of pumping.

Laser oscillations are highly monochromatic, i.e., wave lengths of the oscillations which make up the light beam at the generator output are not dispersed. Thus, for example, for a ruby laser, depending on pumping power, the spectrum of the output beam varies as shown in Figure 7.29.

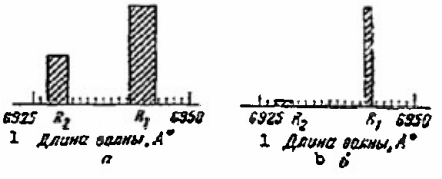


Figure 7.29. Radiation spectrum from a ruby laser.
a - low-power pumping; b - high-power pumping.
1 - wave length, Å.

A disadvantage inherent in solid state lasers, as well as gas and fluid lasers, is their low efficiency (1-2%).

7.15 Semiconductor Optic Quantum Generators

Crystals of gallium arsenide, gallium arsenide-phosphide, and other substances are used as the active substance in semiconductor optic quantum generators. Radiation in semiconductor crystals is caused by energy jumps of free electrons.

Lasers of this type employ electric pumping. Excitation of carriers (electrons) is accomplished by current through the diode in the forward direction. With low excitation current, radiation from the semiconductor diode is noncoherent and its intensity is low. With certain large values of current through the semiconductor, coherent induced radiation is observed; the radiation band converges, and radiation intensity increases.

The construction of a diode with noncoherent infrared radiation is sketched in Figure 7.30.

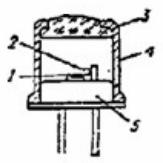


Figure 7.30. Diode with noncoherent infrared radiation.

1 - gallium arsenide wafer; 2 - wire contact; 3 - lens; 4 - kovar casing; 5 - base.

A diffusion diode of gallium arsenide, GaAs, with a p-n junction is placed in a standard casing with a glass lens in its upper part.

Diode excitation current of 100 ma at 25°C produces 1.2 V forward bias.

Reference radiation occurs at 0.9 microns wave length at 25°C. Diode radiation may be modulated by varying the excitation current at any frequency from audio to 900 MHz. Diode efficiency at room temperature is about 5%, and at liquid nitrogen temperature (77°K) it is 10 times higher, i.e., greater than 50%.

The construction of a diode with coherent radiation at 8400 Å is shown in Figure 7.31. Two wafers of p- and n-type gallium arsenide with zinc and tellurium additives respectively are divided by a 2.5 micron thick junction. The diode is shaped like a truncated pyramid. The junction and reverse sides of the pyramid are strictly parallel and highly polished. The semiconductor between these sides forms a resonator tuned to the radiation frequency.

A forward pulse current is used for pumping. Pulse length is 5 to 20 microseconds.

Semiconductor lasers operate pulsed and continuous-mode. They may be used independently or as pumping generators for solid stage lasers.

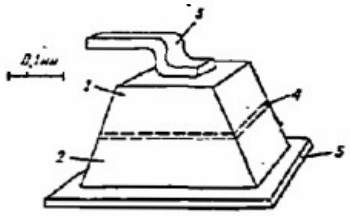


Figure 7.31. A semiconductor laser.

1 - p-type gallium arsenide; 2 - n-type gallium arsenide; 3 and 5 - electrodes; 4 - junction plane.

Pulse Modulators

Pulse modulators are radio devices designed to control the oscillations of SHF generators.

They are classified according to their fundamental elements: a commutating device and a storage element; and according to the operating mode of the latter.

The basic classes of pulse modulators are electron tube, linear and magnetic.

7.16 Electron Tube Pulse Modulators

Pulse modulators using electron tubes as switching elements are called electron tube modulators. Their storage element may be capacitive or inductive. The most widely used modulators have capacitive storage elements, operating on partial discharge. Such modulators are called pulse modulators with a partial discharge capacitor. A modulator circuit is shown in Figure 7.32.

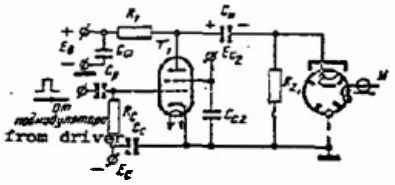


Figure 7.32. Circuit of a pulse modulator with partial discharge capacitor.

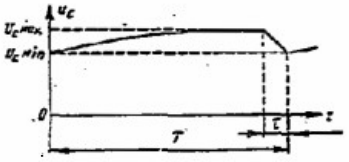


Figure 7.33. Change in voltage on the storage capacitor in the charging and discharging stages.

In the pauses between triggering pulses, the switching tube 1 is blocked by negative voltage $-E_c$. At this time storage capacitor C_H is charged from a high voltage rectifier with voltage E_B. Charging current flows through the following circuit: $+E_B$, resistor R_1, storage capacitor C_H, charging resistor R_2, to the chassis. Charge on C_H builds up exponentially from $U_{c\ min}$ to $U_{c\ max}$ (fig. 7.33).

The time constant of the charging circuit

$$T_c = (R_1 + R_2)C_H \qquad (7.8)$$

is chosen so that the capacitor will charge to the required value before the triggering pulse arrives.

The switching tube is unblocked by a positive triggering pulse coming from the driver. With the tube open, the voltage on capacitor C_H is applied through the tube to the magnetron. This voltage (and a constant magnetic field) causes the magnetron to begin generation. Of course, the energy of the storage capacitor is expended to supply the magnetron, and it partially discharges through the following circuit: $+C_H$, tube 1, magnetron M, $-C_H$. A small part of the current branches off through resistor R_2.

Capacitor C_H continues to discharge while the tube is conducting, and ceases discharge when the control pulse ends.

The discharge circuit time constant is

$$T_d = R_{eq}C_H \qquad (7.9)$$

where

$R_{eq} \approx R_g R_2/R_g+R_2$;

R_g is the dc resistance of the SHF generator (pulsed).

The time constant is made sufficiently large by choosing a large capacitance for C_H so that it will discharge very little while the pulse is formed. C_H discharges (fig. 7.33) to $U_{c\ min}$. In the following pause between pulses it again charges to $U_{c\ max}$.

Voltage on the capacitor changes by the amount

$$\Delta U_c = U_{c\ max} - U_{c\ min}, \qquad (7.10)$$

which, as a rule, does not exceed 5% of the minimum voltage on the capacitor. Hence the term modulator with partial discharge capacitor.

Leading and trailing edges of the pulse formed by the modulator are affected by parasitic capacitance of the circuit and its elements. Total parasitic capacitance is considerable (50 to 150 pF). It discharges principally through resistor R_2, which may not be chosen small. Therefore, pulse cutoff may be extended, and consequently, the high voltage pulse from the SHF generator is also extended. This is undesirable since prolonged generation may not be used, and the thermal load on the generator is increased, thus lowering its lifetime.

A high voltage choke L is often used instead of resistor R_2 (fig. 7.34) to improve the pulse trailing edge. In this case the basic pulse is followed by high intensity oscillations in the discharge circuit of the modulator, which may trigger the magnetron again. A diode is connected in parallel with the choke to eliminate these oscillations. It does not conduct current while the base pulse is being formed and does not affect the forming process. After the pulse is formed, as soon as the oscillations reverse polarity (fig. 7.35), the diode begins to conduct, shunts the modulator discharge circuit, and extinguishes the parasitic oscillations in it.

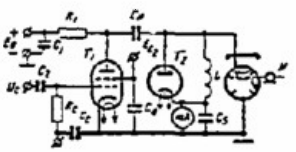


Figure 7.34. Circuit of an electron tube modulator with an inductance in the charging circuit.

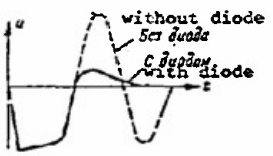


Figure 7.35. Output pulse of a modulator with post-pulse correction.

Resistor R_1 in both circuits limits forward current from the high voltage rectifier through the conducting switching tube.

A modulator with a discharge inductance and a diode forms a pulse with a sharp dropoff at the tip, since current through the inductor increases linearly during the pulse, which leads to an additional discharge by the storage capacitor. To minimize the drop at the pulse tip, the choke inductance is carefully selected, and in some cases the capacitance of the storage capacitor is increased.

An SHF generator and a modulator with partial discharge capacitor may be coupled by a pulse transformer, which fulfills its usual function and replaces the discharge inductance. The quenching diode may be connected in parallel with either the primary or the secondary winding of the transformer.

Switching electron tubes have comparatively small power (Table 7.6). For this reason they are often connected in parallel in modulators with a partial discharge capacitor. One of the variations of such a parallel tube

circuit is shown in Figure 7.36. Here the limiting resistors are R_1 and R_2. Two storage capacitors C_1 and C_2 are used in the circuit. They charge from one rectifier through their limiting resistors and the common primary winding of the pulse transformer. The grid circuits of the switching tubes are in parallel, but each tube discharges its capacitor through the following circuit: positive plate of the storage capacitor, tube, chassis, primary winding of the pulse transformer, negative plate of the capacitor.

Table 7.6.

Pulse modulator tube parameters

	1 Параметр	2 ГИ-30, двойной тетрод	3 ГМИ-2Б, тетрод	4 ГМИ-3, тетрод	5 ГМИ-5, тетрод	6 ГМИ-6, двойной тетрод	7 ГМИ-7, тетрод	8 ГМИ-10, тетрод	9 ГМИ-89, тетрод	10 ГМИ-90, тетрод
11	Напряжение накала, *в*	6,3/12,6	25	26	26	6,3/12,6	26	6,3	25	25
12	Ток накала, *а*	2,5/1,25	17,5	5	1,75	2/1	6,4	5,5	3,9	7,8
13	Напряжение анода, *кв*	5	32	28	20	4	22	9	26	33
14	Ток анода импульсный, *а*	10	90	30	16	8	52	13	20	40
15	Напряжение на управляющей сетке в импульсе, *в*	350	200	330	250	125	330	—	350	350
16	Напряжение смещения, *в*	—200	—600	—300	—800	—200	—900	—300	—600	—600
17	Ток управляющей сетки импульсный, *а*	2,5	7,5	5	1,7	1	10	2	5	5
18	Напряжение на экранирующей сетке, *в*	550	2000	1900	1250	800	2000	1000	1750	1750
19	Допустимая мощность рассеяния на аноде, *вт*	15	900	80	50	15	125	41	100	140
20	Длительность импульса, *мксек*	1	2	2	3	5	5	5	2	2,5
21	Емкость, *пф*:									
22	входная	15	350	110	57,5	13,5	80	40	60	110
23	выходная	7	125	15	8,5	5,2	11,5	6	12	16
24	проходная	0,1	10	1	0,5	0,2	0,9	0,7	6	10

Key: 1 - parameter; 2 - GI-30, dual tetrode; 3 - GMI-2B, tetrode; 4 - GMI-3, tetrode; 5 - GMI-5, tetrode; 6 - GMI-6, dual tetrode; 7 - GMI-7, tetrode; 8 - GMI-10, tetrode; 9 - GMI-89, tetrode; 10 - GMI-90, tetrode; 11 - heater voltage, v; 12 - heater current, amp; 13 - anode voltage, kv; 14 - pulse anode current, amp; 15 - control grid pulse voltage, v; 16 - bias voltage, v; 17-control grid pulse current, amp; 18 - screen grid voltage, v; 19 - permissible power dissipation on anode, W; 20 - pulse length, microsecond; 21 - capacitance; pF; 22 - input; 23 - output; 24 - transfer.

Resistors R_3, R_4, R_5, and R_6 in the anode and grid circuits are included to suppress parasitic high frequency oscillations. They are usually about 10 to ohms.

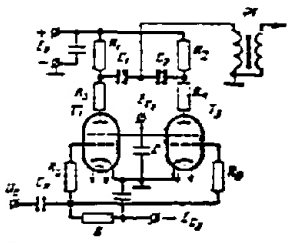


Figure 7.36. Circuit of an electron tube modulator with tubes in parallel.

7.17 Linear Pulse Modulators

Pulse modulators with storage elements in the form of an artificial line are called linear. Hydrogen thyratrons are used as commutating elements.

The principal circuit of a linear modulator is illustrated in Figure 7.37. It consists of an artificial line (AL), hydrogen thyratron Th, charging choke CC, and pulse transformer PT. The generator is loaded by an SHF generator, and supplied from a high voltage rectifier.

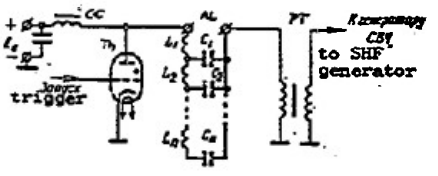


Figure 7.37. Linear pulse modulator circuit.

The modulator charging circuit contains the rectifier, charging choke, line, and pulse transformer. Figure 7.38 shows the equivalent circuit of the charging loop. The storage element begins to charge from the rectifier with voltage E_B when it is switched from discharge to charge, which corresponds to closing the switch in the equivalent circuit. In the discharge stage the artificial line behaves like a capacitor with capacitance equal to the total capacitance of the line C_{line}.

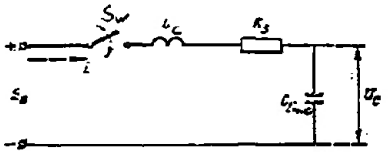


Figure 7.38. Equivalent circuit of the charging loop.

In addition to the choke (with inductance L_c), the line contains a series circuit. Resistor R_3 includes losses in the choke and other elements of the charging loop.

Inductance in the stages of the artificial line have practically no effect on the charging process since they are small in comparison with inductance in the charging choke. For the same reason, the pulse transformer does not affect the charging process.

The period of modulation usually does not exceed the period of intrinsic oscillations in the charging loop, and the charging process is transfer in character.

If there is no energy initially stored in the choke or in the capacitors of the line, then voltage and current in the charging loop are described by the equations

$$u_c = E_b(1 - e^{-\alpha t}\cos\omega_0 t); \tag{7.11}$$

and

$$i = I_m e^{-\alpha t}\sin\omega_0 t, \tag{7.12}$$

where $\alpha = \frac{\omega_0}{2Q}$;

$\omega_0 = 1/\sqrt{L_c C_{line}}$ is the angular resonant frequency of the charging loop;

Q is the figure of merit of the charging loop;

$I_m = E_B/\rho_c$;

ρ_c is characteristic impedance of the charging loop.

The charging process is often limited to the first half of the period T_0 of the intrinsic oscillations, i.e., by the ratio

$$T = T_0/2, \tag{7.13}$$

where

T is the period of modulation, equal to the pulse repetition period;

$$T_0 = 2\pi\sqrt{L_c C_{line}}. \tag{7.14}$$

In this case the charging voltage on the line U_{line} and charging current i vary as shown in Figure 7.39a.

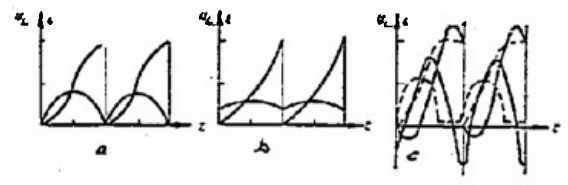


Figure 7.39. Graphs of charging voltage and current for various modes of the charging loop.

a - resonant charge; b - linear charge; c - oscillating charge.

The linear charging mode occurs when $T < T_0/2$. Figure 7.39b shows charging voltage and current under these conditions.

Oscillating charge mode takes place when $T > T_0/2$ (fig. 7.39c).

In all modes the line is charged to practically the same voltage

$$U_{line\ max} = 2E_B\left(1 - \frac{\pi}{4Q}\right). \tag{7.15}$$

Usually $Q \approx 10$ and $U_{line\ max} \approx 1.8\ E_B$.

Average charging current for the resonant mode

$$I_{av} = \frac{2I_m}{\pi}\left(1 - \frac{\pi}{4Q}\right). \tag{7.16}$$

The discharge loop of the modulator consists of the artificial line, the hydrogen thratron and the pulse transformer, loaded by the SHF generator. An equivalent circuit of the discharge loop is given in Figure 7.40. The artificial line is replaced by an ideal line with wave impedance

$$\rho_{line} = \sqrt{\frac{L_1}{C_1}}, \tag{7.17}$$

where

L_1, C_1 are inductance and capacitance of the line stages.

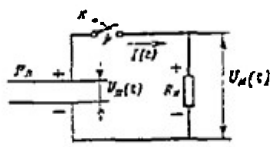


Figure 7.40. Equivalent circuit of the discharge loop.

The line is charged to $U_{line\ max}$ and is an energy source for the discharging loop.

The load resistance

$$R_L = R_g/n_p^2, \tag{7.18}$$

where

R_g is dc resistance of the SHF generator;

n_p is pulse transformer, PT, turns ratio.

The load impedance is considered matched to the artificial line if $R_L = \rho_{line}$. With a matched load, the discharge loop forms a single voltage pulse (fig. 7.31a) with amplitude $U_L = U_{line\ max}/2$ and length determined by equation (6.10).

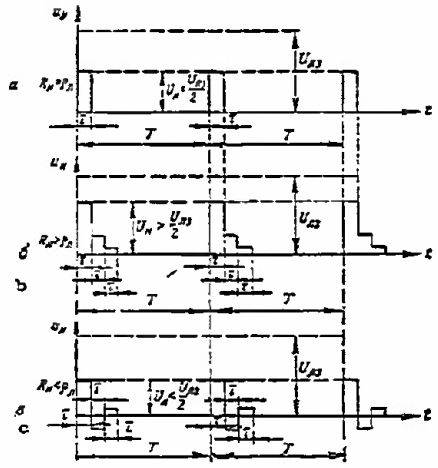


Figure 7.41. Formed voltage pulses for different load matchings of a linear pulse modulator.

a - matched load; b - load larger than matched; c - load less then matched ($U_x = U_{line\ max}$).

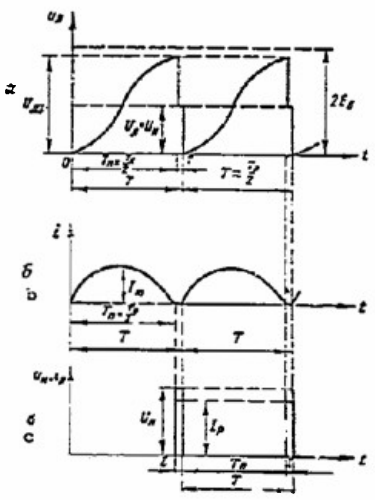


Figure 7.42. Voltages and currents in the circuits of a linear pulse modulator.

a - voltage in the charging and discharging stages of the AL; b - charging current; c - voltage on the primary winding of the PT and current through it.

A mismatched load may be larger or smaller than ρ_{line}. Voltage pulses for these cases are shown in Figure 7.41. The relationship of processes in the charging and discharging stages in a linear modulator may be explained with the sketch of voltages and currents shown in Figure 7.42. The artificial line is charged over the extended period $T_F \approx T$ by a small (average) current to voltage $U_{line\ max}$. At a certain moment a triggering pulse is applied to the grid of the hydrogen thyratron. The thyratron fires, and the artificial line discharges for a short time τ through the primary winding of the pulse transformer. As soon as the thyratron fires, the voltage on the line decreases to half its value, but it is held practically constant for the duration of the pulse. This voltage acts on the primary winding of the pulse transformer and is amplified into the anode circuit of the SHF generator.

Linear pulse modulators in many variations of circuits and constructions are being used in a broad range of applications.

Artificial lines. Artificial lines (AL) in pulse modulators are energy storage elements and forming elements. The customary ladder network is often used (fig. 6.9).

In actual conditions the AL as a forming two-terminal network is discharged into the load (an active impedance (7.18)), shunting the inductance of the pulse transformer. The forming voltage pulse thus has a steeply sloping tip, since current through the inductance rises linearly during the pulse.

To correct the pulse shape, an AL with unequal network parameters is used. Such lines are called unsymmetrical. Exponential type parameters of an unsymmetrical line (fig. 7.43) are calculated by the formulas:

$$L_1 = \frac{R_L \tau}{2n} \frac{1}{1 - \frac{R_L \tau}{2L_L}}; \tag{7.19}$$

$$C_1 = \frac{L_1}{R_L^2}; \tag{7.20}$$

$$C_k = C_1 = \text{const}; \tag{7.21}$$

$$L_k = L_1 e^{-2a(k-1)}, \tag{7.22}$$

where

$a = 2L_1/L_L$;

L_L is inductance shunting the load;

k is the number of sections.

The number of sections n usually satisfies the relationship (6.11) and

$$n = (2 \text{ to } 8)10^6 \tau. \tag{7.23}$$

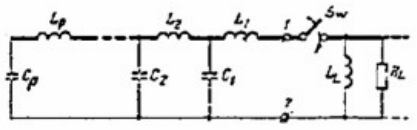


Figure 7.43. Artificial line with load.

Hydrogen thyratrons. Hydrogen thyratrons are almost exclusively used in linear pulse modulators. The construction of a thyratron is shown in Figure 7.44. It has three electrodes: cathode, anode, and control grid. The distance between anode and grid determines the value of breakdown voltage and consequently, of cutoff voltage.

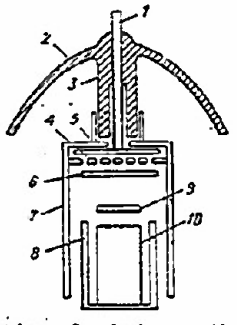


Figure 7.44. Construction of a hydrogen thyratron.

1 - anode input guide; 2 - bulb; 3 - glass baffle; 4 - grid; 5 - anode; 6 - grid shielding plate; 7 - grid frame; 8 - cathode screen; 9 - cathode shielding plate; 10 - cathode.

According to its construction, the grid may be a perforated plate or a woven grid located directly under the anode.

To obtain positive control characteristics, the cathode is completely screened from the anode field. Therefore, a thyratron includes shielding plates for the grid and cathode and a cathode screen. The grid shielding plate is electrically connected to the grid. When this plate is present, the electric field of the anode does not go beyond the edges of the plate 6, and consequently, the anode does not affect cathode emission. Electrons, emitted by the cathode are primarily located in the region bounded by the screen and the shielding plate of the cathode.

The thyratron fires when a positive voltage is applied to the control grid and the anode. Voltage on the grid starts current flowing between grid and cathode, thus producing electrons and ions in the region outside the cathode screen, which propagate up to the grid shielding plate. As soon as electron density at the edges of the shielding plate 6 becomes sufficiently high, the electric field of the anode causes ionization in the

region between the grid and anode plates, which leads to ionization in the whole space between the anode and cathode and to discharge in the thyratron. In the initial stages of discharge, the anode current is a glow discharge between anode and grid. Glow discharge leads to an increase in positive potential on the grid to a value at which the gas in the space between grid and anode is ionized rapidly. After the gas in this region is ionized, the anode electric field begins to affect electrons outside it, and gas ionization in the grid-cathode regions takes place, which leads to discharge in the thyratron. The whole process of initiating and carrying out discharge in a thyratron takes place in approximately 0.05 microsecond. De-ionization time in a hydrogen thyratron is on the order of tens of microseconds.

Hydrogen is one of the most active chemical elements. It eagerly absorbs impurities contained in the material of the electrodes, which may cause a decrease in hydrogen pressure and consequently, changes in the electrical characteristics of the thyratron. To eliminate this effect, thyratron electrodes are made of chemically pure nickel, and certain other measures are taken to compensate for the loss of hydrogen in a thyratron when it is in use.

Hydrogen thyratrons have a positive grid characteristic and are controlled by positive pulses whose amplitude is from 200 (for the majority of thyratrons) to 700 v, and whose length is several microseconds.

The rise time of grid voltage is given for each thyratron. For the majority of thyratrons it is equal to 300 to 600 volts/microsecond (see Table 7.7).

An anode current pulse is delayed in relation to the firing pulse of grid voltage. The delay amounts to several tens of microseconds and depends on the leading edge of the firing pulse, the value of anode voltage on the thyratron, and on the state of the cathode. The cathode state depends on the method of heater supply. If the heater is supplied by alternating current, the alternating field will affect discharge close to the cathode and also the time required to set up ionization stages of discharge in the thyratron. Delay of the anode current pulse in this case may be decreased considerably.

In using thyratrons, heater voltage must be held strictly constant, since cathode emission decreases sharply when heater voltage drops, and an increase in heating increases hydrogen reaction with the oxide cathode.

Service time of a hydrogen thyratron depends on its mode of operation and is determined primarily by the operating voltage, anode current, pulse length, and repetition rate. Thyratron service time is usually no less than 500 h. It may be extended if the thyratron is not continuously operated at its limits.

Table 7.7

Parameters of certain pulse hydrogen thyratrons

	1 Параметр	2 ТГИ1-260/12	3 ТГИ2-325/16	4 ТГИ1-400/16	5 ТГИ1-700/25
6	Напряжение накала, *в*	6,3	6,3	6,3	6,3
7	Ток накала, *а*	12	8,5	10,5	17
8	Напряжение анода, *кв*	12	16	16	20—25
9	Падение напряжения в импульсе, *в*	125	150	170	200
10	Ток анода в импульсе, *а*	260	325	400	700
11	Ток анода средний, *а*	0,4	0,2	0,5	1
12	Крутизна фронта импульса анодного тока, *а/мксек*	—	900	1000	1500
13	Выходная мощность в импульсе, *квт*	1560	2300	3200	8700
14	Длительность импульса, *мксек*	2—8	0,3—6	0,5—5	0,3—11
15	Число импульсов в секунду, *имп/сек*	4500	1700	450	500
16	Параметры поджигающего импульса:				
17	напряжение, *кв*	0,2	0,2	0,2	0,7—2
18	ток, *а*	0,5	0,5—1	0,5	3—8
19	длительность импульса, *мксек*	2—8	2	2—2,5	3—6
20	скорость нарастания, *кв/мксек*	0,3—0,6	0,3—0,6	0,3—0,6	1—2

Key: 1 - parameter; 2 - TGI1-260/12; 3 - TGI2-325/16; 4 - TGI1-400/16; 5 - TGI1-700/25; 6 - heater voltage, v; 7 - heater current, amp; 8 - anode voltage, kv; 9 - voltage drop in pulse, v; 10 anode current in pulse, amp; 11 - average anode current, amp; 12 - slope of the anode current pulse front, amp/microsec; 13 - pulse output power, kw; 14 - pulse length, microsec; 15 - pulses per second, p/sec; 16 - firing pulse parameters; 17 - voltage, kv; 18 - current, amp; 19 - pulse length, microsec; 20 - rise time, kv/microsec.

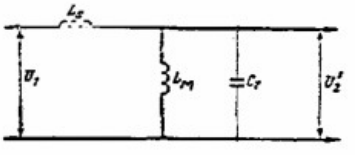


Figure 7.45. Simplified equivalent circuit of a pulse transformer.

Pulse transformers. Pulse transformers are widely used in pulse modulators and other devices. Their fundamental purpose is to increase the amplitude of the formed voltage pulse. Simultaneously they match the low impedance modulator output with the high impedance SHF generator input, change pulse polarity, provide dc decoupling with the circuit, etc.

The equivalent circuit of a PT is shown in Figure 7.45, where all elements are referred to the primary winding.

Pulse transformation should not be accompanied by a great deal of distortion in the pulse shape, distortion produced by distributed inductance, L_S, magnetization inductance L_M, and capacitance C_T of the transformer. To decrease L_S and C_T, the number of turns is limited, core volume is increased, and special coupling circuits are used for the windings. The turns ratio of pulse transformers is limited to 3-10. No more than several tens of turns are allowed in the primary and secondary windings.

Operating voltages on the windings, as a rule, are very high; consequently careful attention is given to their electrical stability.

Pulse transformer cores are made of high-grade steel with high magnetic permeability. They are wound of ribbons, 0.02-0.10 mm thick. Cores are often made of E310 steel (previously called XVP).

To improve electrical stability, pulse transformers are assembled in closed tanks filled with transformer oil. Transformers over 100 kw and 6 kv are provided with expansion tanks.

7.18 Linear Pulse Modulators with Dual Forming Lines

The circuit of a modulator with dual forming lines is illustrated in Figure 7.46.

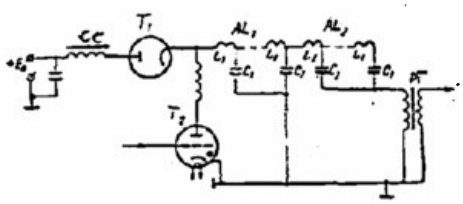


Figure 7.46. Basic circuit of a linear pulse modulator with dual forming lines.

The charging loop of the modulator includes charging choke CC, charging diode $Tube_1$, two-terminal forming networks AL_1 and AL_2, and the primary winding of the pulse transformer PT.

Charging current flows through the following circuit: $+E_B$, charging choke, charging diode, and then through the branches: first artificial line to the chassis; second artificial line to the primary winding of the pulse transformer to the chassis.

Processes in the charging loop are determined by inductance of the charging choke, total capacitance of the forming lines, total active losses in the circuit, and the charging diode.

An equivalent circuit of the charging loop is shown in Figure 7.47.

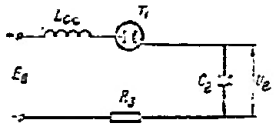


Figure 7.47. Equivalent circuit of the charging loop.

Parameters of the charging loop elements are selected so that the period of intrinsic oscillations in the loop, determined by equation 7.14, will satisfy the condition

$$T_0 \leqslant 2T_{min},$$

where

T_{min} is the minimum pulse sequence period.

The charging loop is adjusted so that at maximum pulse repetition rate (corresponding to T_{min}), it will operate close to resonance. Then at all other pulse repetition frequencies, the charging loop will rectify the charging process (fig. 7.48). In the latter case, the charging period of the storage element may be divided into two stages; an oscillating charge stage (from t_0 to t_1) and a rectified* charging process stage (from t_1 to t_2). Of course, charging current only flows in the first stage (from t_0 to t_1).

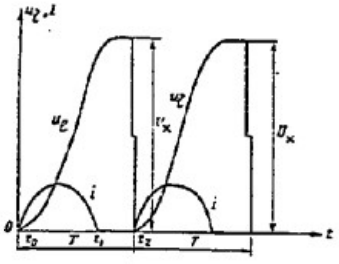


Figure 7.48. Voltage on the line and charging current in a rectified charging process ($U_x = U_{line\ max}$).

The first stage of the charging process is the same for all pulse repetition rates. The duration of the second stage varies inversely with the pulse repetition rate.

---

* The process in a charging loop which contains a diode with $T > T_0/2$ is called a rectified charging process.

Since voltage on the storage element does not change in the second charging stage from t_1 to t_2, then in the case of a rectified charge, the voltage to which the forming line is charged is the same as for resonant charging (7.15). Average charging current in the rectified mode is determined by the equation

$$I_{av} = \frac{I_m}{\pi}\left(1 - \frac{x}{4Q}\right), \tag{7.24}$$

where

$I_m = E_B/\rho_c$;

ρ_c is the characteristic impedance of the charging loop;

$x = T/T_0$ is the mode indicator.

The effective value of charging current is

$$I_{eff} = \frac{I_m}{\sqrt{2\pi}}\sqrt{1 - \frac{x}{2Q}}. \tag{7.25}$$

The discharging loop of a pulse modulator consists of the hydrogen thyratron $Tube_2$, first forming line AL_1, primary winding of the pulse transformer, and the second line AL_2.

A supplementary high frequency choke is usually included in the discharge loop to decrease the rise time of anode current through the thyratron in an attempt to decrease heat losses in the thyratron.

The outstanding feature of the discharge in this modulator is the two forming lines, and consequently the processes taking place in it are quite different from the processes in a discharge loop with one line.

The load of a pulse modulator with dual lines is considered matched if its impedance is twice as large as the wave impedance of the forming lines, i.e.,

$$R_L = 2\rho_{line}. \tag{7.26}$$

A simplified circuit of the discharge loop is shown in Figure 7.49a, and the equivalent circuit in Figure 7.49b.

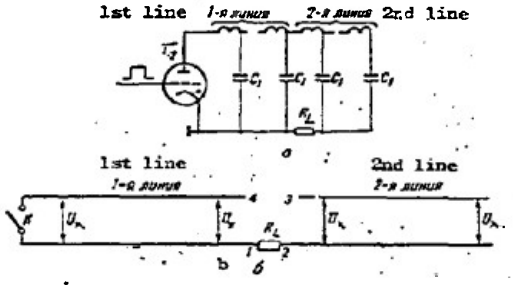


Figure 7.49. Simplified (a) and equivalent (b) circuits of the discharge loop of a linear pulse modulator with dual forming lines.

As soon as the line has charged to the required voltage $U_{line\ max}$, the thyratron fires, which is equivalent to closing switch K in the equivalent circuit in Figure 7.49h. The left end of the first line is now short-circuited.

This moment is taken as the initial (t = 0) moment for the process in the discharge loop.

With a short circuit, voltage and current waves are set up in the line. Waves propagating in both lines from left to right will be considered incident, and those propagating from right to left will be considered reflected.

The voltage reflection coefficient in a line in the general case is equal to

$$m = R_{LL}-\rho_L/R_{LL}+\rho_L, \tag{7.27}$$

where

R_{LL} is the total load impedance of the line.

Obviously, the coefficient of reflection from an open end of the line, when $R_{LL} = \infty$, is equal to +1, and from a short-circuited end of the line ($R_{LL} = 0$), is equal to -1.

From the moment switch K is closed, a current (fig. 7.50a)

$$I_r = U_{line\ max}/\rho_{line}, \tag{7.28}$$

begins to flow through the thyratron, and an incident wave appears at the left end of the first line. When the left end of the line is short-circuited, m = -1, the incident wave is negative and equal to $U_{line\ max}$. The incident wave, propagating from left to right, seems to strip voltage from the previously charged line. Energy previously stored in the line capacitances is transformed into magnetic energy.

Behind the front of the incident wave propagating from left to right voltage on the line becomes equal to zero, although the first line in front of the incident wave front and the second line (and consequently, the load) remain at the same state as before the thyratron was fired. There is practically no voltage on the load or current through it.

The line parameters are chosen so that the wave propagates one complete length in a time equal to half the length of the pulse formed by the modulator, i.e.,

$$\frac{\tau}{2} = \sqrt{L_2 C_l}. \tag{7.29}$$

All this time voltage on the right end of the first line remains equal to $U_{line\ max}$ (fig. 7.50b).

The first line is loaded on the right by impedance R_{LL}, equal to the sum of the modulator load impedance R_L, determined by equation 7.26, and the impedance of the second line.

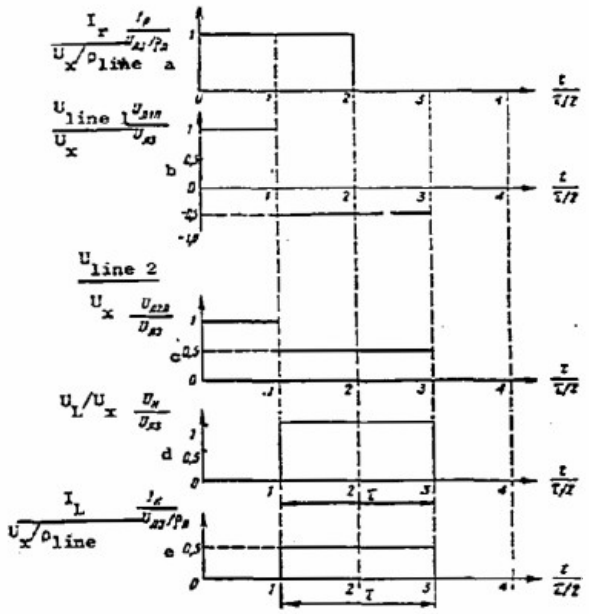


Figure 7.50. Graphs of currents and voltages in the discharge loop of a linear pulse modulator with dual forming lines.

Both lines are identical, accordingly $\rho_{line\ 2} = \rho_{line\ 1} = \rho_{line}$. Consequently, the total load impedance of the first line

$$R_{LL} = 3\rho_{line}. \tag{7.30}$$

After a time equal to $\tau/2$, the incident wave reaches the right (loaded) end of the first line. The line load from the right is not matched, so a reflected wave appears. For $R_{LL} = 3\rho_{line}$, the reflection coefficient is equal to 1/2. At the right end of the first line the voltage becomes equal to half of the former value and changes polarity, since a negative wave is reflected. This voltage on the right end of the first line will be maintained till the wave propagates from right to left and back (fig. 7.50b).

At $t = \tau/2$ the state of the first line at points 1-4 (fig. 7.49b) changes radically, which naturally causes a change in the boundary conditions on the left end of the second line. The second line seems to be connected to a load through the first line, consequently an indicent wave appears in

the second line. The load on the second line is the same as on the first, i.e., $R_{LL} = 3\rho_{line}$, accordingly the reflection coefficient from the left end of the line is also equal to 1/2. As a result, voltage on the left end of the first line at this moment will become equal to $U_{line\ max}/2$ (fig. 7.50c). The same voltage is established along the line as far as the incident wave propagates. At the moment $t = \tau$, the incident wave reaches the right open end of the line, is reflected, and begins to propagate back. Propagation of the wave from right to left indicates that the second line is also stripped of energy, transferring it to the load through the line sections located in front of the wave. Voltage on the line and current in it become equal to zero in the sections behind the front of the reflected wave.

Voltage $U_{line\ 2L}$ at the loaded end of the second line up to moment $t = 3\tau/2$ remains unchanged, equal to $+ U_{line\ max}/2$ (fig. 7.50b).

Thus, in the period from $t = \tau/2$ to $t = 3\tau/2$ voltage at the right end of the first line is constant and equal to $- U_{line\ max}/2$, and voltage at the left end of the second line remains constant and equal to $+ U_{line\ max}/2$. Voltage on the load at this time is equal to $U_{line\ max}$ (fig. 7.50d). Current in the load (fig. 7.50b) is constant:

$$I_L = U_{line\ max}/R_L = U_{line\ max}/2\rho_{line}. \tag{7.31}$$

For convenience, time on the abscissa of Figure 7.50 is given in units of half the duration of the formed pulse $\tau/2$, i.e., the time required for the wave to traverse the line in one direction. Values of voltage and current are plotted on the ordinate.

It must be emphasized that the duration of a pulse (see 6.10) formed by a modulator with dual lines is determined by a double transit of the wave along one of the lines, i.e.,

$$\tau = 2n\sqrt{L_1C_1},$$

where

n is the number of sections on one line;

L_1, C_1 are inductance and capacitance of one section.

Using dual artificial lines in the discharge loop of a pulse modulator has a number of advantages associated with the fact that, for given output data on the modulator, the voltage to which the line is charged is cut in half. The operating voltage on the hydrogen thyratron is thus reduced by half.

A linear modulator with dual lines is more complicated in its circuitry and construction than a modulator with one forming network.

7.19 Magnetic Pulse Modulators

Pulse modulators using nonlinear inductances as switching elements are called magnetic pulse modulators.

Nonlinear inductances, made in the form of chokes, autotransformers, or transformers with special alloy cores, are called magnetic switches.

The inductance of a choke L with a ferromagnetic core is equal to

$$L = N_t^2 S_c \mu / l, \tag{7.32}$$

where

N_t is the number of turns in the winding;

S_c is the core section;

l is average length of magnetic lines in the core;

μ is magnetic permeability of the core material.

The inductance of a magnetic switch depends on the degree of magnetization of its core. If a ferromagnetic material has a hysteresis loop as shown in Figure 7.51, its magnetic permeability (equal to the ratio of an increment of magnetic induction to an increment of field potential; i.e., $\mu = \Delta B/\Delta H$) in the saturated state μ_s (regions 1-2 and 1'-2' in fig. 7.51), is considerably less than permeability in the nonsaturated state μ (region 1-1'). Consequently inductance of a magnetic switch in the unsaturated state (L) will be considerably more than at saturation (L_s). The impedance of a choke in the unsaturated state will also be greater than at saturation. This fundamental property of nonlinear inductances is used in magnetic switches.

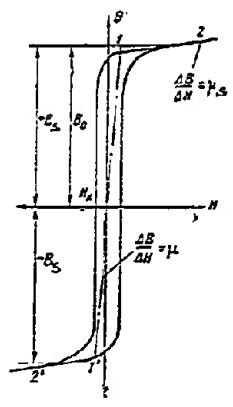


Figure 7.51. Hysteresis loop of special alloys.

The core material should have certain properties. Primarily, the hysteresis loop of the alloy should be almost rectangular, and its area as small as possible. With such properties, permeability of the core material varies drastically when the material saturates ($\mu/\mu_s \gg 2000$); the material switches from one state to another with low field potentials, and hysteresis losses are small.

A ferromagnetic material should also have the greatest possible magnetic induction at saturation B_s, thus permitting a reduction in the required volume of the core, and accordingly, in the dimensions and weight of the commutator.

The core material should have high resistivity to reduce eddy current losses. To further reduce these losses, cores are wound of ribbons from 10 to 80 microns thick.

Various ferromagnetic alloys with the required properties have been produced. The best of these are nickel permalloy type 60 NP, 65 NP, and others. Table 7.8 shows the characteristics of some alloys.

Table 7.8

Characteristics of some ferromagnetic alloys

1 Сплав	2 Толщина листа, мм	3 Индукция насыщения B_s, вб/м²	4 B_0/B_s, %	5 Коэрцитивная сила H_c, а/м	6 Удельное сопротивление $\rho_{ст}\cdot10^{-3}$, ом·м	7 Удельный вес, кг/м³
50 NP	0,05	1,5	88—92	24	45	8,2
65 NP	0,01 0,05—0,10	1,3	82—87 88—93	48 12	30	8,35
34 NKMP	0,02—0,04 0,35	1,3 1,55	83—90 86—95	16 5,6	52	8,3
80 NKhS	0,05—0,1 0,2—0,34	0,7		16—8 1,6	65	8,5
79 NMA	0,01 0,2	0,75	—	8,0 1,6	56	8,85
E310	0,05—0,09 0,03	1,6	40—42 71	4,0 32	32	8,5
	0,03	1,7	66	35		

Key: 1 - alloy; 2 - plate thickness, mm; 3 - saturation induction B_s, weber/m²; 4 - B_0/B_s, %; 5 - coercive force, H_c, a/m; 6 - specific impedance, $\rho_{ct}\cdot10^{-3}$, ohm-m; 7 - specific weight, kg/m³.

A schematic of a magnetic modulator is illustrated in Figure 7.52. Magnetic modulators usually have an ac supply, so ac voltage is supplied to input terminals 1-2. Transformer Tr steps up voltage to the required value. Charging choke CC assures the required charging mode for the first

storage capacitor C_1. Saturation chokes MK_1, MK_2, and MK_3 are magnetic switches. Artificial line AL and pulse transformer PT have the customary purpose.

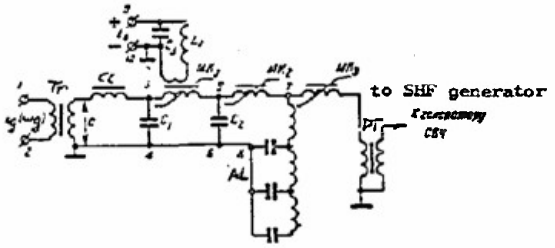


Figure 7.52. Circuit of a magnetic pulse modulator.

Capacitors C_1 and C_2 are equivalent and equal to the total capacitance C_{line} of the artificial line.

Inductances of the magnetic switches are chosen so that each is considerably smaller than the preceding one, i.e., in a nonsaturated state $L_1 \gg L_2 \gg L_3$. At saturation also $L_{1\,s} \gg L_{2\,s} \gg L_{3\,s}$. Inductance of the first switch in a nonsaturated state should be several times larger than inductance L_C of the charging choke. Usually $L_1 \gg (3\text{-}5)L_C$.

The first switch has an additional winding for initial magnetization from the dc source (through terminals 9-10).

Inductance of the final switch MK_3 in a nonsaturated state should be considerably larger than the inductance of the pulse transformer primary winding, and at saturation it should not be larger than the inductance of a segment of the AL.

In a magnetic modulator there are three principal parts: input circuit, transforming stages, and final stage.

The input circuit comprises capacitor C_1, charging choke CC, and the step-up transformer, which serves to charge C_1. The charging choke is chosen so that the circuit is tuned to the frequency of the generator ω_g supplying the modulator, i.e.,

$$\omega_0 = 1/\sqrt{L_C C_1} = \omega_g.$$

If there is no initial energy stored in the circuit, voltage (fig. 7.53a) on the secondary winding of the step-up transformer

$$e = E_m \sin \omega_g t. \tag{7.33}$$

Voltage on the capacitor will vary as shown in Figure 7.53b. Mathematically this voltage is described by the equation

$$u_{C_1} = QE_m(e^{-\alpha t} - 1)\cos \omega_0 t, \tag{7.34}$$

where

Q is the figure of merit of the input circuit;

α is the attenuation factor of the circuit;

ω_0 is the intrinsic angular frequency of the circuit.

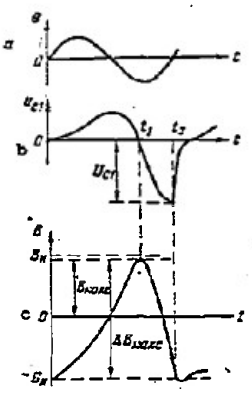


Figure 7.53. Graphs of voltage and magnetic induction in the input stage of a magnetic modulator.

In the initial state MK_1 is at negative saturation, i.e., initial magnetic induction in its core $B(0) = -B_s$.

An increment of magnetic induction ΔB in the core is associated with voltage on the choke by the equation

$$\Delta B = \frac{10^4}{N_2 S_c} \int_0^t u dt, \tag{7.35}$$

where

ΔB is in v-sec/m^2;

S_c in cm^2;

u is voltage in volts.

Voltage u_{C1} acts at points 3-4 of the circuit. It is applied simultaneously to the four sequentially switched-in elements MK_1, MK_2, MK_3, and PT. Since the inductance of the first switch is considerably larger than the inductance of all the following elements, voltage u_{C1} is applied almost entirely to the first switch.

With a variation in u_{C1} as shown in Figure 7.53b, at the very beginning of the C_1 charging stage, the first magnetic switch is taken out of negative saturation, and then its magnetic induction begins to rise.

The increment of magnetic induction satisfies (7.35), and total induction in the core varies as shown in Figure 7.53b.

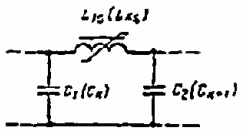


Figure 7.54. Equivalent circuit of the discharge loop of a transformer stage.

Magnetic induction is maximum at t_1, when u_{C1} changes polarity. At this time, the core should not be in a state of positive saturation. After t_1, the MK_1 core is demagnetized by a voltage of reverse polarity on capacitor C_1. Demagnetization takes place more rapidly than magnetization. It continues until the core returns to the original state of negative saturation (moment t_2). In the saturated state the impedance of MK_1 is radically reduced, and capacitor C_1 begins to discharge into capacitor C_2.

The equivalent circuit of the discharge is shown in Figure 7.54. At the initial moment of discharge, voltage on capacitor C_1 is equal to the amount to which it was charged. The process of discharging capacitor C_1 through the saturated first switch into capacitor C_2 is accompanied by a drop in voltage on both capacitors as shown in Figure 7.55. Note that MK_2 in this period is not saturated and does not permit capacitor C_2 to discharge into AL.

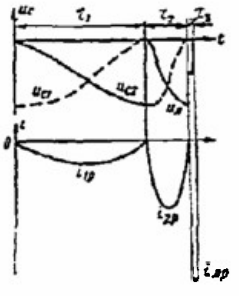


Figure 7.55. Voltage and current in the transformer and final stages of a magnetic modulator.

Analytically, voltage on the capacitors during recharging is described by the following equations

$$u_{C,k} = -\frac{U_{Ck}}{2}(1 + \cos\omega_k t) \tag{7.36}$$

and

$$u_{C,k+1} = -\frac{U_{Ck}}{2}(1 - \cos\omega_k t), \tag{7.37}$$

where

k is the number of discharging capacitors;

$k + 1$ is the capacitor following the discharging capacitor;

U_{Ck} is voltage to which the capacitor was charged;

$\omega_k = 1/\sqrt{L_{ks} C_{eq}}$ is the intrinsic angular frequency;

$C_{eq} = C_k C_{k+1}/C_k + C_{k+1} = C_k/2$ (assuming $C_k = C_{k+1}$).

In this case, $u_{C,k} = u_{C1}$, and $u_{C,k+1} = u_{C2}$. Voltage u_{C2} is applied across points 5-6, i.e., it is applied to three elements: MK_2, MK_3, and the primary winding of the pulse transformer. Since the inductance of MK_2 is much larger than the remaining inductances, the voltage acts principally on MK_2.

Voltage u_{C2}, in accordance with (7.35), causes a change in magnetic induction in the core of the second switch.

When voltage on the second capacitor reaches its maximum, the core of MK_2 saturates, and capacitor C_2 begins to discharge into the AL.

Voltage on capacitor C_2 varies in accordance with (7.36), and voltage on the line (with total capacitance $C_{line} = C_2$) varies in accordance with (7.37).

Voltage u_{line}, i.e., at points 7-8, principally acts on MK_3. When line voltage reaches its maximum, switch MK_3 saturates and begins to discharge the AL into the load through the pulse transformer.

In the circuits C_1, MK_1 and C_2, MK_2, voltage and current are transformed. Duration of the processes is reduced, but discharge current from stage to stage is increased. Accordingly, these circuits are called transformer stages.

Duration of the process in the k⁰ stage is equal to half the period of intrinsic oscillations, i.e.,

$$\tau_k = \pi \sqrt{L_{ks} \frac{C_k}{2}}, \tag{7.38}$$

where

L_{ks} is the inductance of the k⁰ switching choke at saturation.

The final stage consists of an artificial line, magnetic switch MK_3, and the pulse transformer. The line is discharged into the primary winding of the pulse transformer loaded by a SHF generator in the same way as for the customary linear modulator. This forms a voltage pulse of the given length, determined by the artificial line parameters.

This circuit is called a choke circuit, since its magnetic switches use a variety of choke. Composite circuits are sometimes used in practice, in which autotransformers and transformers are used together with chokes. In this case, an input transformer is not required, since the required voltage increase may be obtained in transformer stages with switching autotransformers or transformers.

There are usually three or four transformer stages in a modulator. If the modulator power is increased, the number of stages is raised to six.

Magnetic modulators are highly reliable, mechanically rugged, and require no warm-up time. They may be constructed with any desired output parameters; in particular, there are no factors which would limit the possibility of obtaining the necessary powers.

7.20 Radio Transmitter Circuits

In recent years multistage radio transmitters using floating drift klystrons and amplitrons in the output stages have come into wider use. A typical block diagram of a pulsed transmitter with a floating-drift klystron is shown in Figure 7.56. The klystron generator exciter is a multistage device with repeated frequency multiplication and amplification of the r.f. oscillations. Fairly powerful oscillations, stable in frequency and phase, are obtained at its output.

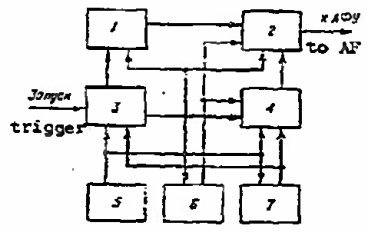


Figure 7.56. Block schematic of a transmitter with a floating drift klystron.

1 - exciter; 2 - klystron generator; 3 - pulse modulator of the exciter; 4 - pulse modulator of the klystron generator; 5 - control, blocking and signaling system; 6 - cooling system; 7 - supply source.

The oscillations are modulated in the exciter and the klystron generator by separate pulse modulators.

One possible circuit of a klystron generator is shown in Figure 7.12.

Quartz-crystal oscillators, stabilitrons, and reflex klystrons are used in the exciter master oscillator in the latter case (fig. 7.57). Reference voltage at the stable frequency is taken off after the oscillations of the quartz-crystal oscillator used in the AFC system have been frequency-multiplied.

Frequency modulation (fig. 7.57) is accomplished by applying a modulating voltage to the reflex klystron.

The complexity of the exciters in more powerful klystron transmitters is indicated in Figure 7.58. Here a quartz-crystal oscillator is used as a master oscillator. Electron tube amplifiers and frequency multipliers, using

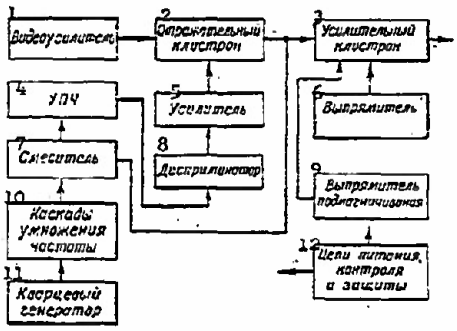


Figure 7.57. Block schematic of a klystron transmitter.

1 - video amplifier; 2 - reflex klystron; 3 - klystron amplifier; 4 - intermediate frequency amplifier; 5 - amplifier; 6 - rectifier; 7 - mixer; 8 - discriminator; 9 - magnetization rectifier; 10 - frequency multiplier stages; 11 - quartz-crystal oscillator; 12 - supply, control, and protection circuit.

SHF triodes, amplify the oscillations and raise their frequency. Amplitude modulation is accomplished by applying a modulating voltage to the output stage klystron.

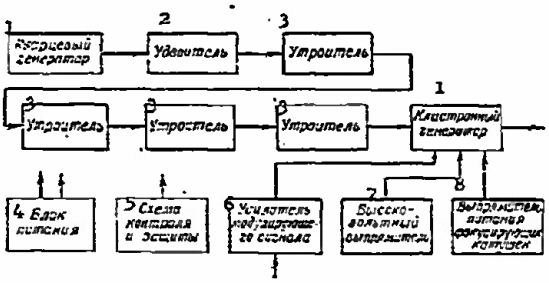


Figure 7.58. Block schematic of a klystron transmitter with electron tube exciter.

1 - klystron generator; 2 - doubler; 3 - tripler; 4 - supply unit; 5 - control and protection circuit; 6 - modulating signal amplifier; 7 - high voltage rectifier; 8 - focusing winding supply rectifier.

In pulsed radio transmitters pulse modulation in the electron tube frequency multipliers and the amplifiers of the exciter is carried on simultaneously with pulse modulation in the klystron generator.

Floating drift multiplier klystrons, amplitrons, and traveling wave tubes may be used in the exciter stages.

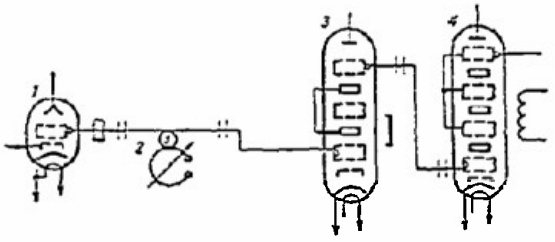


Figure 7.59. Schematic of the radio frequency channel of a klystron transmitter.

1 - reflex klystron; 2 - resonator; 3 - three-cavity floating drift klystron with magnetic focusing; 4 - four-cavity floating drift klystron with electromagnetic focusing.

Figure 7.59 shows a simplified schematic of the radio frequency channel of a three-stage klystron transmitter. A reflex klystron, 1, with a high-Q resonator, 2, serves as the master oscillator. Amplification and frequency multiplication are accomplished by a floating drift klystron, 3, with magnetic focusing. Floating drift klystron 4 with electromagnetic focusing is used in the output stage.

Figure 7.60 shows a block diagram of a pulsed radio transmitter using platinotrons.

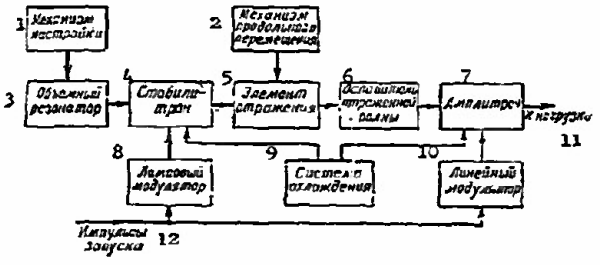


Figure 7.60. Block diagram of a pulsed transmitter using platinotrons.

1 - tuning mechanism; 2 - longitudinal deflection mechanism; 3 - cavity resonator; 4 - stabilitron; 5 - reflection element; 6 - reflected wave attenuator; 7 - amplitron; 8 - tube modulator; 9 - cooling system; 10 - linear modulator; 11 - to load; 12 - triggering pulses.

A stabilitron serves as exciter. A high-Q cavity resonator, tuned over a determined frequency range with the tuning mechanism, is connected to its input, and a reflecting element is connected to its output. The feedback necessary for self-excitation is accomplished by reflecting part of the energy. Feedback phase is controlled by moving the reflector along the coupled waveguide.

A decoupling attenuator is connected after the reflecting element to decrease the effect of the amplitron on the master oscillator. A ferrite rectifier which slightly attenuates the forward wave and blocks propagation of wave in the backward direction is often used as a decoupling element.

An amplitron is used in the output stage. Separate pulse modulators serve to control oscillations of the exciter and the amplitron output generator. An electron tube modulator is used to control the less powerful master oscillator, which is under rigid requirements relative to frequency stability and phase. The output amplitron generator is controlled by the less powerful linear modulator. The pulse modulators are synchronized from one source.

Figure 7.61 shows a simplified circuit of a radio frequency circuit of an amplitron transmitter. Traveling wave tube 1 is used in the first stages of the exciter. Amplitrons operate in the following stage 3 of the exciter and in the output stage 5 of the transmitter. Ferrite devices 2 and 4 are used for stage decoupling.

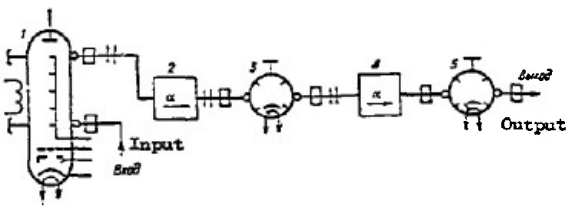


Figure 7.61. Circuit of the radio frequency channel of an amplitron transmitter.

1 - traveling wave tube; 2, 4 - ferrite devices; 3 - amplitron of the exciter following stage; 5 - amplitron of the transmitter output stage.

Carcinotrons, klystrons, and other generating devices may be used as master oscillators in amplitron transmitters. The frequency of the amplitron transmitter may be adjusted over a given range with electronic or mechanical tuning arrangements. Amplitrons cannot be tuned as wideband devices in this case.

Chapter VIII

Radio Receiving Devices

8.1 General Information on Radio Receiving Devices

Radio receiving devices are designed to separate signals of a certain frequency and shape, to subsequently amplify them, ano to coovert them to the form required to trigger the final stage (indicator).

The most common form of a radio receiving device at the present time is the superheterodyne receiver, such as that shown in the block diagram in Figure 8.1.

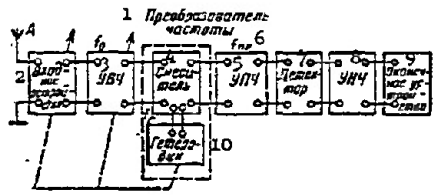


Figure 8.1. Block diagram of a radio receiving device.

1 - frequency converter; 2 - input device; 3 - high frequency amplifier; 4 - mixer; 5 - intermediate frequency amplifier; 6 - f_i; 7 - detector; 8 - low frequency amplifier; 9 - final stage; 10 - heterodyne.

Frequency selectivity is accomplished in high frequency and intermediate frequency amplifiers. Secondary conversion (detection) of the signals into a shape required for operation of the final stage is accomplished in a detector.

The frequency converter is required to simplify signal amplification; this is done principally at an intermediate frequency f_i which is lower than the received signal and which does not vary when the receiver operates in a range of frequencies. Amplification of the signals to a value required to operate the final stage is primarily effected in the intermediate frequency amplifier stages and partially in the high frequency amplifier and video amplifier (low frequency amplifier). The stages of the receiver (fig. 8.1) up to the detector comprise the linear part of the receiver.

Radio receivers are classified according to the form of the signal received (pulse, frequency modulated, etc.), the operating frequency range (meter, decimeter, centimeter, and millimeter wave length receivers), and other factors.

Depending on their purpose, receivers may be distinguished not only by the form of the received signal and the operating frequency range, but also by circuit, construction and performance figures.

8.2 Basic Performance Figures of Receivers

The sensitivity of a receiver characterizes its capability to receive weak radio signals. It is quantitatively described by the value of the limiting and actual sensitivity.

The limiting sensitivity of a receiver $P'_{r\ min}$ or $E'_{r\ min}$ is defined as that minimum power (or emf) of the signal in the antenna which, if the antenna and receiver are matched, produces at the output of its linear part a signal-to-noise ratio equal to one.

Actual sensitivity of the receiver $P_{r\ min}$ (or $E_{r\ min}$) is defined as that power (or emf) in the antenna which produces at the output of the linear part of the receiver a signal-to-noise ratio equal to the discrimination factor. For decimeter waves and shorter, sensitivity is measured in watts or decibels, for meter ano longer waves, it is measured in volts (microvolts).

Accordingly

$$P_{r\ min}[db] = 10 \log P_{ref}[w]/P_{r\ min}[w],$$

where

P_{ref} is the power of a reference level (usually $P_{ref} = 10^{-5}$ watts).

Receiver sensitivity varies inversely with the value of $P_{r\ min}$ or $E_{r\ min}$. In present-day decimeter and centimeter wave receivers, $P_{r\ min} = 10^{-13}$ to 10^{-14} watts, in meter wave receivers and above $E_{r\ min} = 2$ to 50 microvolts.

At SHF, sensitivity is primarily limited by internal noise of the receiver. Receiver noise is characterized by the noise figure N, which indicates by how many times the total noise at the input (or output) of the device is larger than total noise in the signal source at its input (output).

Since noise from the signal source is supplemented by internal noise of the receiver, the signal/noise ratio at the receiver output will be smaller than at its input. This deterioration is also characterized quantitatively by the noise figure.

The noise figure N of a device is defined as the ratio of signal/noise power at the input to the signal/noise power at the output:

$$N = \frac{(P_s/P_n)_{in}}{(P_s/P_n)_{out}} \qquad (8.1)$$

The noise figure is measured in relative units and decibels, where $N_{db} = 10 \log N_{dif}$. For an ideal (no-noise) stage N = 1. Actually, in modern radar receivers N = 1 to 10.

The noise properties of a receiver are also characterized by the effective noise temperature T_{eff}, associated with N by the equation

$$T_{eff} = T_0(N-1),$$

where

$T_0 = 300°K$.

The noise figure of a receiver is primarily determined by noise in the first stages. Thus, the first stages (high frequency amplifier stages) should be low noise in order to decrease the noise figure and increase sensitivity of the receiver.

Gain characterizes the amplifying properties of the receiver. A distinction is made between power gain (K_p) and voltage gain (K).

Power gain K_p is defined as the ratio of power at the output of the device P_{out} to power at the input, P_{in},

$$K_p = P_{out}/P_{in}.$$

Voltage gain is defined analogously:

$$K = U_{out}/U_{in}.$$

If the input and output stages are matched and the stage requires the maximum possible (i.e., nominal) power $P^n_{i\ in}$ from the signal source (the preceding stage or antenna) and delivers nominal power $P_{n\ out}$ to the load, the ratio of these powers is defined as the nominal power gain:

$$K_{np} = \frac{P_{n\ out}}{P_{n\ in}}$$

The greater the K_{np} of a given stage, the less the effect of noise in the following stages on the receiver sensitivity. Gain is given in relative units or decibels, where $K_{dB} = 20 \log K$ or $K_{pdB} = 10 \log K_p$. In common receivers total gain $K = (0.1\text{-}10) \cdot 10^6$; $k_p = (0.1\text{-}10) \cdot 10^{13}$; $K_{np} = (0.1\text{-}10) \cdot 10^{13}$.

Gain is a complex quantity, since between input and output voltages there is a phase shift due to reactive elements. The relationship of the gain modulus to the frequency is called the frequency characteristic, and the relationship of the phase shift angle between U_{out} and U_{in} (argument of K) to frequency is called the phase characteristic (Fig. 8.2).

The pass band of a receiver characterizes its selective properties and determines the region of frequencies passed at the same time by the receiver. Width of the pass band is determined (Fig. 8.2) as the difference of frequencies f_2 and f_1 for which K is decreased by $\sqrt{2}$ and K_p is half of its maximum value.

For pulse receivers there is an optimum or quasi-optimum pass band, for which the signal/noise ratio at the output of the linear part of the

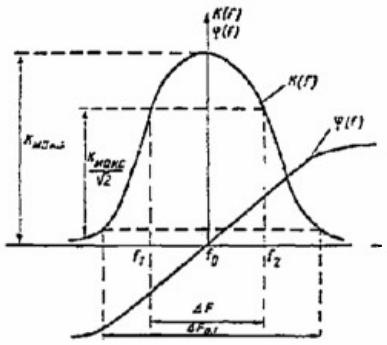


Figure 8.2. Frequency and phase characteristics of a receiver.

receiver is maximum. The optimum band depends on the form and length of the pulse and is determined by the equation

$$\Delta F_{opt} = \frac{0.8 \div 1.37}{\tau_p}, \qquad (8.2)$$

where τ_p - length of the received pulse;

0.8 for a bell-shaped pulse;

1.37 for a rectangular pulse.

In actual radar station receivers ΔF_{opt} = 0.1-1 MHz.

When the pass band of a radio receiver is shifted from the optimum (i.e., when $\Delta F_p > \Delta F_{opt}$ or $\Delta F_p < \Delta F_{opt}$), the signal/noise ratio at the output of its linear section becomes worse. This is caused by different spectral distributions in the frequency of the pulse signal and the noise.

Limiting sensitivity of the pass band and the noise figure are associated by the following relationships:

$$P'_{r\,min} = kT_0 N \Delta F\,[\text{MHz}]; \qquad (8.3)$$

$$E'_{r\,min} = 0{,}125\sqrt{NR_A \Delta F\,[\text{MHz}]}. \qquad (8.4)$$

where $E'_{r\,min}$ - in μV;

$P'_{r\,min}$ in W;

k Boltzmann constant

$T_0 = 300°K$, $kT_0 = 4 \cdot 10^{-21}$ W/s;

R_A radiation resistance of the antenna in Ω.

The range of a receiver is characterized by the ability of the receiver to operate in a range of frequencies and determines primarily the noise shielding of the receiver from antenna interference. It is quantitatively

determined by the dynamic range of the input D_{in} and output D_{out} signal.

Dynamic range of the input signal D_{in} is defined as the ratio of maximum input voltage $U_{in\ max}$ at which the receiver is not overloaded, to minimum input voltage $U_{in\ min}$, corresponding to the limiting sensitivity of the receiver:

$$D_{in} = \frac{U_{in\ max}}{U_{in\ min}} \tag{8.5}$$

Dynamic range of the output signal is found analogously:

$$D_{out} = \frac{U_{out\ max}}{U_{out\ min}} \tag{8.6}$$

These values are sometimes determined by power:

$$D_{p\ in} = \frac{P_{r\ max}}{P_{r\ min}}; \tag{8.7}$$

$$D_{p\ out} = \frac{P_{out\ max}}{P_{out\ min}} \tag{8.8}$$

They are measured in relative units or decibels, $D_{dB} = 20 \log D$; $D_{pdB} = 10 \log D_p$. In common receivers D_{in} should be greater than 70 to 80 dB. The relationship of U_{out} to U_{in} is called the amplitude characteristic of the receiver.

Radio receivers are also characterized by the operating stability, efficiency, and technical-economic indicators (cost, weight, dimensions, etc.).

8.3 Optimum Reception

Reception of useful signals in practice is accompanied by internal and external fluctuation interference (noise). A receiver which provides the best reception, evaluated by an arbitrary criterion, in the presence of noise is called an optimum receiver. Receiving a signal in the presence of noise requires a solution to one of the following two problems:

detecting the signal; i.e., establishing the presence or absence of a signal;

discriminating the signal; i.e., finding the form of the signal oscillations or one (sometimes several) of the parameters of this oscillation (amplitude, arrival time, etc.).

The problems of detecting and discriminating have much in common and in many cases may be solved at the same time. But these problems also have many differences. Thus an optimum signal detector may not be an optimum discriminator, and vice versa. The predominant problem in radar is the problem of detection.

Criteria for an optimum receiver. In the process of detecting a signal two types of errors are possible:

detecting a signal when in fact there is only noise (false alarm);

not detecting a signal when there actually is one (signal omission).

Criteria have been set up in relation to the probability of errors.

According to the criterion of an ideal observer, a receiver is considered optimum if it assures a minimum probability of error of any type.

According to the Newman-Pearson criterion, a receiver is considered optimum if it has a maximum probability of correct detection, a minimum probability of omitting a signal with a given probability of false alarm. The Newman-Pearson criterion is the most feasible for radar. Other optimum criteria are known (Bice criterion, mini-max criterion, information loss criterion).

In the general case, the criterion applied to a receiver depends on its construction and properties.

Two types of optimum receiver may be used for detecting a radar signal: a receiver with an optimum filter and a correlation receiver.

A receiver with an optimum filter may solve two problems:

reproduction of the information carried by the signal $U_s(t)$ with the minimum mean square error;

reproduction of the signal $U_s(t)$ with minimum mean square error.

For a radar station detector, detection is the main problem; consequently an optimum receiver amounts to a linear filter and a following detector (Fig. 8.3). In this case, that linear filter which provides maximum signal/noise ratio at the detector output is the optimum. Construction of the filter is strictly dependent on the signal spectrum. The filter is optimum whose complex frequency characteristic is proportional to the complex-conjugate spectrum of the received signal (see Chap. I). For example, for a sequence n of periodic radio pulses (for a series of signals reflected from the target) an optimum filter (Fig. 8.4) consists of two series blocks $K_1(j\omega)$ and $K_2(j\omega)$ and has a saw-tooth frequency characteristic (Fig. 8.5 b). Block $K_1(j\omega)$ is an optimum filter for a single radio pulse and is realized as a system (Fig. 8.6). Structure and frequency characteristic of block $K_2(j\omega)$ strictly depends on the number of pulses n. For $n>1$, this block has a saw-tooth frequency characteristic (Fig. 8.5 b) and for this reason is called a saw-tooth filter. For finite n, a saw-tooth filter is realized as a delay line DL with (n-1) taps and isolation stages (Fig. 8.7). For $n\to\infty$, a saw-tooth filter may be realized as a wide-band stage with delayed negative feedback (Fig. 8.8) with $\beta K_0 \to 1$. For a finite number of pulses, the latter filter (Fig. 8.8) is not optimum.

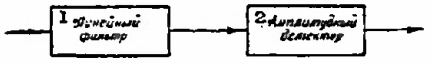


Figure 8.3. Block diagram of an optimum receiver with a filter.
1 - linear filter; 2 - amplitude detector.

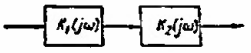


Figure 8.4. Block diagram of an optimum filter for a sequence of pulses.

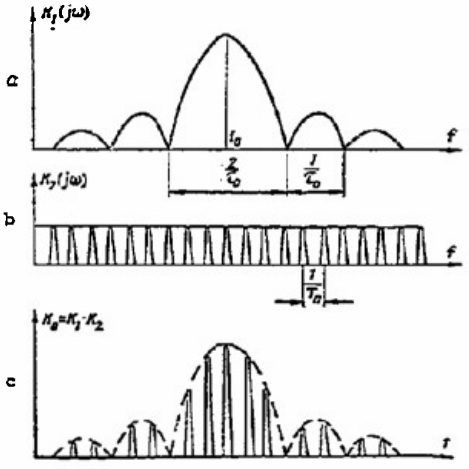


Figure 8.5. Frequency characteristics of an optimum filter:
a) for block K_1; b) for block K_2; c) for the filter as a whole.

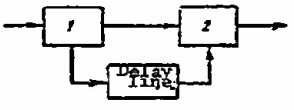


Figure 8.6. Structure of block K_1 in an optimum filter:
1) integrating circuit; 2) summing stage.

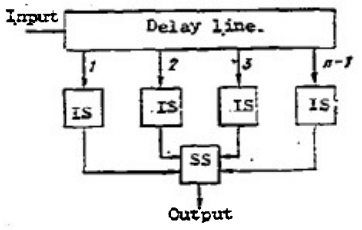


Figure 8.7. Structure of block K_2 in an optimum filter for finite n (IS - isolation stages; SS - summing stage)

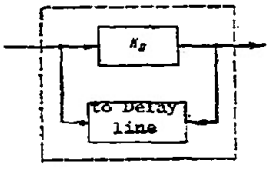


Figure 8.8. Structure of the K_2 block in an optimum filter for $n \to \infty$.

Since it is sometimes difficult to realize optimum filters, quasi-optimum filters have been studied by V. I. Siforov to operate on rectangular radio pulses, and by A. P. Belousov to operate on bell-shaped radio pulses. The form of the particular characteristic of the filter is assumed given and corresponding to the resonance characteristic of the system of circuits, and the maximum signal/noise ratio at the filter output is assured only by choosing the pass band. The latter is found from equation (8.2).

A correlation receiver permits optimum detection of signals with completely unknown parameters. For optimum correlation of a pulse radar receiver, besides the known parameters (signal frequency, pulse length, etc.), the position of the expected reflected signal in time must also be known. In this case there is minimum mean square error in reproducing the information. The block diagram of the simplest correlation receiver is shown in Fig. 8.9, where y(t) - received oscillation, x(t) - reference oscillation, associated in a particular way with the signal. For example, x(t) - useful signal delayed by time τ_d.

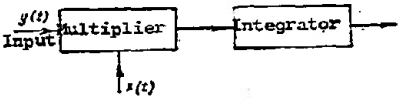


Figure 8.9. Block diagram of a correlation receiver.

8.4. Input Circuits

Input circuits are designed to match the antenna-feeder system to the high frequency amplifier, and, if there is no high frequency amplifier, to match directly to the converter and to carry out preliminary selection.

Separate oscillator systems or a set of several oscillator systems, coupled together are used as input circuits.

In the long, medium, and short wavelength ranges and in the first part of the meter range (up to 150-200 MHz) oscillator circuits with lumped parameters consisting of inductor windings and capacitors (Fig. 8.10) are used as input (and also interstage) circuits.

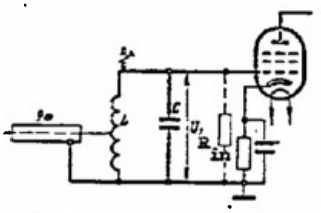


Figure 8.10. Input circuit with lumped parameters.

At frequencies above 200 MHz oscillator systems with distributed parameters are used, since with an increase in frequency the dimensions of inductor windings decrease, and at frequencies above 200 MHz they are impossible to construct. In addition, with an increase in frequency, selectivity of the oscillator system with lumped parameters drops rapidly as a consequence of an increase in active resistance in the conductors because of skin effect.

At frequencies from 200 to 1000 MHz, long line sections are used as input (and interstage) circuits, and at frequencies higher than 1000 MHz cavity resonators are used, since here sections of long lines are not highly selective as a consequence of losses due to skin effect. Cavity resonators have a greater surface area; consequently current density in the metal and losses in Joule heat are small. In centimeter range radar stations the input circuit is usually combined with one of the antenna T-R tubes.

Oscillator circuits in the frequency range are tuned by electrical or mechanical changes in one parameter or another. Electrical tuning is accomplished with ferrites, ferroelectric and semiconductor capacitors. Mechanical tuning is accomplished with variable capacitors, short circuit stubs, etc.

The input circuit may be coupled to the signal source capacitively, inductively, or with an autotransformer. For a system with distributed parameters, coupled lines, an opening, or a slot may be used.

The fundamental parameters of input devices are the voltage or power transmission coefficients, K_{in} and $K_{p\ in}$ respectively, the pass band ΔF_{in}, and the resonant frequency f_0.

Noise properties of the input circuits in an electron tube high frequency amplifier are taken into account in calculating the noise figure of the amplifier. In other cases, the noise properties of the input circuits are calculated separately and are characterized by the noise figure N, uniquely associated with the power transmission coefficient by the formula

$$N_{in} = \frac{1}{K_{p\ in}}.$$

The ratio of voltage (power) in the output circuit to the emf (nominal power) of the signal source is understood under the voltage (power) transmission coefficient of the input circuit.

The transmission coefficient is maximum when the antenna-feeder system is matched to the input circuit.

In practice, receiver input circuits have K = 0.7-3, K_p = 0.7-0.95; ΔF = 5-70 MHz.

8.5. High Frequency Amplifiers

The purpose of a high frequency amplifier is to provide minimum total noise figure, to suppress noise in the secondary channels of the receiver, and to amplify at the frequency of the received signal.

Two-stage electron tube high frequency amplifiers are used in the meter and decimeter ranges. Traveling wave tube high frequency amplifiers are used in the centimeter and shorter wave regions.

To decrease the noise figure at superhigh frequency, special low-noise amplifiers (parametric amplifiers, amplifiers using tunnel diodes, quantum amplifiers) are also used.

The basic parameters of a high frequency amplifier are: noise figure N, voltage gain K, nominal power gain K_{np}, and stable gain K_s.

Stable gain is defined as the maximum voltage gain at which the stage operates in a stable condition (does not oscillate). It is inversely proportional to frequency and to the capacitance, coupling the input and output circuits of the stage.

Furthermore, a high frequency amplifier is characterized by a resonant frequency, a pass band, and input and output admittances. In practice the high frequency amplifiers of contemporary receivers have, depending on frequency, N = 2-10; K = 10-70; K_{np} = 20-300; K_s = 10-100.

High frequency electron-tube amplifiers

The most common electron-tube circuits are: a two-stage circuit using a pentode with grounded cathode (pentode, grounded cathode - pentode, grounded cathode), a cascade circuit (Fig. 8.11) triode with grounded cathode - triode with grounded grid, and a two-stage triode with grounded grid - triode with grounded grid circuit (Fig. 8.12). All of these circuits using bantam tubes and oscillator systems with lumped parameters are used at frequencies up to 100-150 MHz; above 150-200 MHz, only the triode with grounded grid circuit using bantam tubes and cermet tubes and oscillator systems with distributed parameters are used.

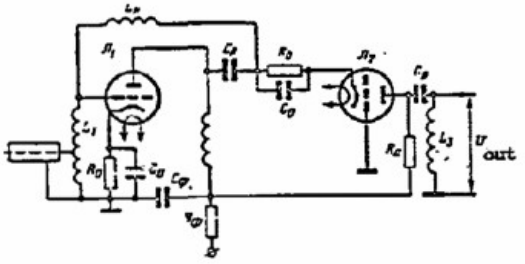


Figure 8.11. Cascade circuit (triode with grounded cathode - triode with grounded grid).

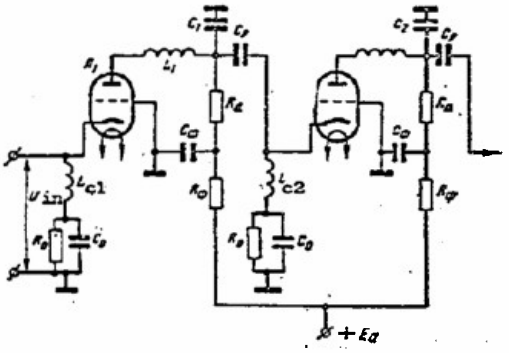


Figure 8.12. Triode with grounded grid - triode with grounded grid.

Figure 8.13 shows a schematic section of the construction and equivalent circuit of a high frequency amplifier stage using a lighthouse triode and sections of coaxial lines.

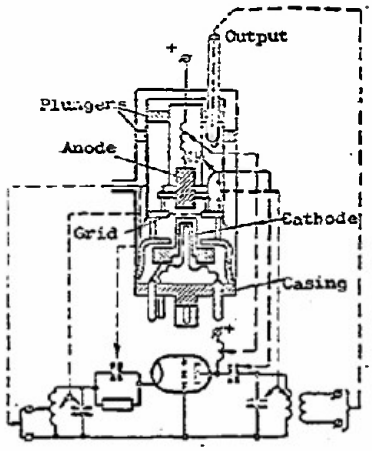


Figure 8.13. Schematic cross section and equivalent circuit of a lighthouse triode with grounded grid.

The triode with grounded cathode - triode with grounded grid circuit (cascade circuit) provides minimum noise figure and low input admittance. Because of the low output admittance of the first stage, a low noise figure is assured in the second stage. The second stage (triode with grounded grid) shunts the first (triode with grounded cathode), thus providing stable amplification of the first stage and sufficiently high nominal power gain, although the voltage gain is sharply reduced ($K \approx 1$).

A comparative appraisal of the basic parameters of electron-tube high frequency amplifiers is shown in Table 8.1.

Таблица 8.1

Свойства каскадов УВЧ

1 пэк – пэк	2 тэс–тэс	3 тэк – тэс (каскодная схема)
5 N — большой 6 $K_{пр}$ — наибольший 7 K — наибольший 8 $K_у$ — сравнительно большой	9 N — малый 10 $K_{пр}$ — сравнительно малый 11 K — большой 12 $K_у$ — большой	13 N — малый 14 $K_{пр}$ — достаточно высокий 15 K — малый 16 $K_у$ — сравнительно большой, только при малом усилении ($K \approx 1$) первого каскада

Table 8.1. Properties of high frequency amplifier stages.

1. pentode (grounded cathode) - pentode (grounded cathode)
2. triode (grounded grid) - triode (grounded grid)
3. triode (grounded cathode) - triode (grounded grid)

4. (cascade circuit); 5. N - large; 6. K_{np} - small; 7. K - largest; 8 - K_s - comparatively large; 9 - N - small; 10 - K_{np} - comparatively small; 11. K - large; 12. K_s - large; 13 - N - small; 14. K_{np} - sufficiently high; 15 - K - small; 16. K_s - comparatively high, but first stage gain low ($K \approx 1$).

For each stage of all electron-tube high frequency amplifier circuits the following relationships are valid:

$$N_1 = 1 + \frac{g_{oe}}{g_s} + \frac{S g_T}{g_s} + \frac{R_{neq} g_T^2}{g_s} \qquad (8.9)$$

$$K_{np} = \frac{S^2}{4 g_s g_\Sigma}; \qquad (8.10)$$

$$K = \frac{S}{g_\Sigma} m_a; \qquad (8.11)$$

$$\Delta F = \frac{g_\Sigma}{2\pi C_\Sigma}.$$

Stable gain is determined by the equations:

for triode with grounded cathode and pentode with grounded cathode circuits

$$K_s = 0{,}42 \sqrt{\frac{S}{\omega C_{ag}}}; \qquad (8.12)$$

for triode with grounded grid circuit

$$K_s = 0{,}5 \frac{S}{\omega C_{ak}}. \qquad (8.13)$$

Here

g_T, S, R_{neq}	input admittance, transconductance, and equivalent noise resistance of the tube respectively;
g_s	admittance of the signal source (antenna);
g_{oe}	input circuit admittance;
g_0	total admittance of the input circuit ($g_0 = g_{0e} + g_T + g_s$);
C_Σ, g_Σ	total capacitance and total admittance of the anode circuit of the stage respectively. ($g_\Sigma = g_{0e} + g_{out\ 1} + g_{in\ 2}$);
C_{ag}, C_{ak}	anode-gride and anode-cathode capacitance of the tube;
ω	frequency of the amplified oscillations;
m_a	coupling coefficient from the grid of the following stage. For complete coupling of the circuit $m_a = 1$.

Parametric amplifiers

Parametric amplifiers are devices in which the signal is amplified by the effect of a high frequency energy source periodically changing the reactive parameter (capacitance or inductance) of the circuit and consequently introducing energy into the circuit.

Depending on the method of changing reactance of the parametric amplifier circuit, the following types are recognized: semiconductor, electron ray, ferromagnetic (ferrite).

Operating principles of a semiconductor parametric amplifier. The operating principle of a semiconductor parametric amplifier in which the capacitance of the circuit is varied, may be explained on the model of an oscillating circuit containing a capacitor (Fig. 8.14) with movable plates.

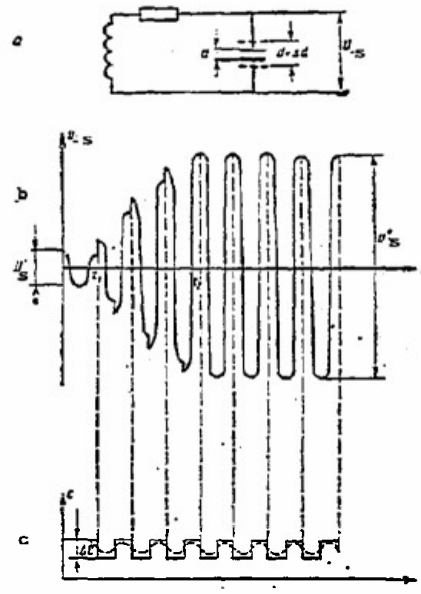


Figure 8.14. Operating principle of a parametric amplifier: a) model of a parametric amplifier; b) form of the signal at the parametric amplifier output; c) variation in capacitance of a parametric amplifier.

Until time t_1 (Fig. 8.14 b) the circuit parameters do not change, and the signal acting on the circuit has an amplitude U'_s. Beginning with t_1 the capacitor plates are mechanically separated by an amount Δd at the moments when the voltage and charge on them is maximum. Since an electric

field exists between the capacitor plates, energy must be expanded to shift the plates. This energy comes from a special source, called the pumping generator. Energy expanded in shifting the plates is transmitted to the electric field of the capacitor and increases its energy. This leads to an increase in voltage on the capacitor, since the capacitance of the component decreased and the charge remained unchanged.

If the capacitor plates are returned to the original position at those moments when their charge is equal to zero, then energy does not leave the circuit. Periodic reiteration of this process leads to an increase in the energy supplied to the system and an increase in the amplitude of the signal. The energy introduced into the circuit is characterized by a negative resistance r_-. The greater the amount of energy supplied, the greater the value of r_- and the greater the gain of the parametric amplifier.

However, the energy introduced into the circuit may not increase without limit. If energy introduced into the circuit compensates the total energy loss in the circuit, which corresponds to $r_- = r_0$, non-attenuated oscillations arise in the circuit when there is no signal, i.e., the system self-excites and begins to operate as a parametric generator. Signals cannot be amplified in such a system. Here r_0 - total resistance loss in the circuit.

Introducing energy into a circuit by changing its reactance is usually called parametric regeneration to distinguish it from the usual regeneration in tube amplifiers produced by positive feedback.

Oscillations of the pumping generator, which vary the capacitance of the circuit, should correspond in frequency and phase to the signal frequency. The plates must be shifted twice during the period of oscillation at the moments of maximum voltage and charge, and they must be shifted twice at the moments when voltage and charge are equal to zero (Fig. 8.14 b).

Thus, this system is a device in which external energy which changes the capacitance of the capacitor is converted to signal energy.

In actual semiconductor parametric amplifiers, a semiconductor diode whose capacitance is nonlinearly dependent on voltage and changes under the effect of the pumping voltage from C_{max} to C_{min} is used in place of a mechanical capacitor. To simplify the pumping generator, the voltage does not change in jumps, but varies sinusoidally.

A parametric diode is characterized by the following parameters: average capacitance C_D, modulation index m, resistance loss R_s, figure of merit Q, critical frequency f_{cr}, where

$$C_D = \frac{C_{max} + C_{min}}{2}$$

and modulation index

$$m = \frac{\Delta C}{2C_D} = \frac{C_{max} - C_{min}}{C_{max} + C_{min}}.$$

Loss resistance R_s is determined by the total volume resistance of the semiconductor and the resistance of the contacts; it varies from 0.3-6 Ω for germanium and from 1-10 Ω for silicon devices.

The quality of a parametric diode, from the point of view of its application, also depends on the figure of merit of the nonlinear capacitance, which is determined by the equation $Q = \frac{1}{\omega C_D R_s}$.

Consequently, to increase Q of the nonlinear capacitance, diodes with low resistance R_s must be chosen.

The critical frequency of a diode is equal to the frequency at which negative resistance of the diode r_- is equal to its loss resistance. The negative resistance introduced into the circuit is determined by

$$r_- = \frac{m}{2\pi f C_D}.$$

From which, considering $r_- = R_s$, we obtain the following equation for critical frequency:

$$f_r = \frac{m}{2\pi C_D R_s}.$$

For contemporary diodes, the critical frequency is around several thousand megahertz. The greater m and the less C_D and R_s, the greater amplification may be obtained from a parametric amplifier. Consequently, the better diode has the higher critical frequency.

Parametric amplifiers using semiconductor diodes may be separated into three basic groups:

single-circuit regenerating;

dual-circuit;

traveling wave amplifiers.

Regenerating amplifiers under certain conditions may operate in a superregeneration mode, forming a separate group of amplifiers.

Single-circuit parametric amplifiers. A parametric amplifier with only one circuit which is tuned to the signal frequency has been called a single-circuit regenerating parametric amplifier (Fig. 8.15).

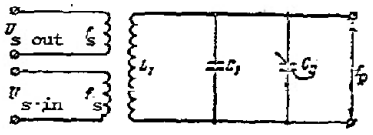


Figure 8.15. Single-circuit parametric amplifier.

To establish the regenerating amplifying mode in a single-circuit parametric amplifier, the circuit capacitance must vary with the pumping frequency $f_p = 2f_s$. A single-circuit regenerating parametric amplifier is simple in construction and tuning.

The major fault of a single-circuit parametric amplifier is the dependence of gain on the phase ratio between signal and pumping frequencies. If $f_s \neq f_p/2$, the ratio between the phase of these oscillations, and consequently the gain, varies periodically, which leads to parasitic amplitude modulation of the signal at a low frequency $F = |2f_s - f_p|$.

Dual-circuit parametric amplifier. A dual circuit parametric amplifier is a system of two oscillator circuits coupled by nonlinear capacitance C_D which varies with frequency f_p. One of the circuits is tuned to the signal frequency f_s, the other auxiliary (open-circuit or balanced) is tuned to the frequency difference $f_- = f_p - f_s$ or sum of $f_+ = f_p + f_s$. The most valuable property of dual-circuit parametric amplifiers is that pumping frequency drift and mutual phase shift between the signal and the pumping does not affect the amplifier operation. Furthermore f_s and f_p need not be multiples of each other.

Depending on the frequency to which the auxiliary circuit is tuned and the source of the amplified signal, dual-circuit parametric amplifiers may be separated into three types:

regenerating amplifier;
nonregenerating amplifier-converter;
regenerating amplifier-converter.

Dual-circuit regenerating parametric amplifier. A voltage from the signal mesh L_1C_1 (Fig. 8.16) and a voltage at the pumping frequency are applied to the nonlinear capacitor C_D. Under the effect of these two voltages, combined frequencies are produced on the nonlinear capacitor just as in the usual mixer.

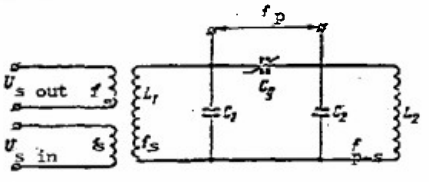


Figure 8.16. Circuit of a dual-circuit regenerating parametric amplifier.

Current at the frequency difference entering the second circuit L_2C_2 tuned to this frequency will produce a voltage increase in it. Oscillations at the remaining frequencies will, practically speaking, not be discriminated in the amplifier. Thus three voltages with frequencies f_s, f_p, and $f_2 = f_-$ will act on the nonlinear capacitor.

The capacitance of the nonlinear capacitor in this amplifier decreases at those moments when its voltage is maximum, and capacitance increases at those moments when voltage is minimum. This is equivalent to introducing a negative resistance into the signal circuit and into the frequency difference circuit, since at frequencies f_s and f_2 the capacitor is connected in parallel with these circuits.

In the given parametric amplifier, the signal is supplied to and taken from the first (signal) circuit.

Dual-circuit nonregenerating amplifier-converter. This amplifier (Fig. 8.17) is distinguished by the fact that its second circuit is tuned to the frequency sum ($f_2 = f_+$). In this parametric amplifier the signal is supplied to the first circuit and taken off the second circuit. There is no parametric regeneration in such an amplifier. Amplification takes place only because of parametric conversion of the signal frequency f_s to the higher frequency f_+, which has a greater amplitude. The absence of regeneration makes the amplifier exceptionally stable in operation (self-excitation is impossible). Its pass band may be 10% of the amplified frequency, which appears to be the fundamental advantage of this circuit.

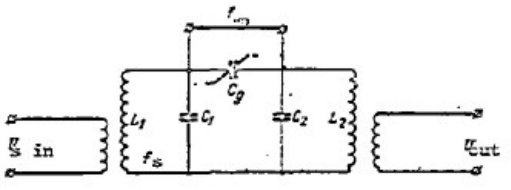


Figure 8.17. Schematic of a dual-circuit amplifier-converter.

A dual-circuit regenerating parametric amplifier-converter differs from the nonregenerating amplifier-converter (Fig. 8.17) only in that its second circuit is tuned to the frequency difference $f_2 = f_-$. Two principles are operative in such an amplifier: first, the principle of regenerating amplification, and secondly, the principle of nonregenerating amplification because of frequency conversion in the nonlinear reactive element from f_s to $f_2 = f_-$. The advantages of this type of amplifier are that greater gain may be obtained with high stability, a wide pass band is possible, and the amplifier input is decoupled from the output.

Traveling wave semiconductor parametric amplifier. This device is made in the form of a section of a feeder (long line or waveguide), in which parametric diodes are connected at given distances.

For such an amplifier to operate, the signal and pumping oscillations must propagate at a specific velocity in the line. Then the pumping frequency should be twice as high as the signal frequency (assuming proper phase relation). Thus when the signal passes down the line, conditions for transmitting energy from the pumping generator to the signal oscillations are set up in each diode. As a result, the power of the signal is increased as it passes from diode to diode. Power gain is proportional to the length of the line.

This type of amplifier has considerable advantages in comparison with the types discussed above; it has a wide pass band (up to 25% of the carrier) and does not require rectifying tubes and circulators. However, traveling wave parametric amplifiers have more complex tuning.

Operating principle of an electron ray parametric amplifier. Operation of this type amplifier is based on periodic changes in the reactive impedance of a cavity resonator when a bunched electron beam passes through it.

The schematic of an amplifier operating on this principle is shown in Fig. 8.18. The electron ray, leaving the electron gun, is accelerated toward the collector. The ray passes through the cavity resonator, supplied by a pumping voltage ($f_p=2f_s$) and is velocity modulated. Moving further, the electron ray passes through the second resonator, which has two identical gaps located a distance l from each other. Signal voltage with frequency f_s is applied to this resonator. With specific values of voltages, ray current, distance between both cavity resonators, and distance l, energy is introduced into the signal circuit compensating the energy loss in it, thus producing signal amplification. A shortcoming of this amplifier is the shot effect in the electron flow, which amplifies its noise.

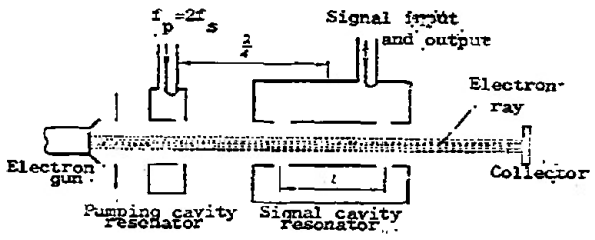


Figure 8.18. Schematic of an electron ray paramagnetic amplifier.

Operating principle of a ferromagnetic parametric amplifier. An amplifier of this type differs from those discussed previously with varying capacitors only in that energy is introduced because of varying inductance.

In these amplifiers, an equivalent self-inductance L_{eq} of a ferromagnet whose magnetic induction is a nonlinear function of current (power of the electromagnetic field) is used as the variable reactance. By varying current i with frequency f_2, the equivalent self-inductance is made to vary with time, and, under certain conditions, energy is introduced into the circuit.

Figure 8.19 shows the equivalent circuit of an amplifier. It consists of two circuits tuned to frequencies f_1 and f_2 and coupled by variable inductance L(t), varying with the pumping frequency $f_p = |f_1 \pm f_2|$. A signal with frequency equal to f_1 is amplified and supplied to the load. A shortcoming of this amplifier is the high noise figure and greater pumping power required to equal the gain of a semiconductor parametric amplifier.

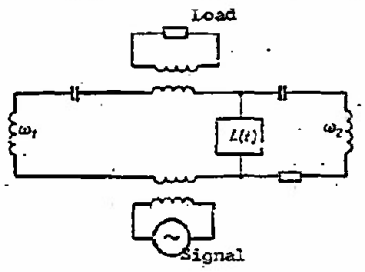


Figure 8.19. Schematic of a ferromagnetic parametric amplifier.

Circuits for connecting a parametric amplifier into the receiver channel. Resonant regenerating parametric amplifiers are reciprocal devices (devices which conduct energy in only one direction are usually called non-reciprocal) and may amplify signals even if the input and output circuits are interchanged. Therefore noise coming from the amplifier output will be amplified if special measures are not taken. Total noise of the parametric amplifier will be amplified, and receiver sensitivity will decrease.

To eliminate multiple regenerative noise amplification coming from the parametric amplifier output and to increase operating stability when the parametric amplifier is connected into the channel, non-reciprocal devices (tubes, circulators) and various connecting circuits are used. Depending on the circuit, parametric amplifiers are classed as pass and reflecting types.

The circuit of a pass type parametric amplifier is shown in Fig. 8.20 a. The input signal is applied to the amplifier through the tube, and the

amplified signal is applied to the following stage of the receiver through another tube. There may be no tubes in this stage. Then noise from the following stages of the receiver is amplified by the parametric amplifier, leading to a decrease in receiver sensitivity.

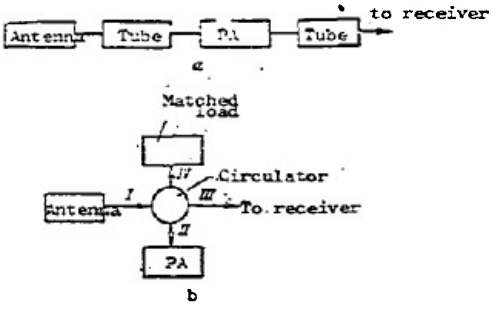


Figure 8.20. Circuit for connecting a parametric amplifier into the receiver channel:

a) pass type parametric amplifier; b) reflecting type parametric amplifier.

The circuit for connecting a reflecting type parametric amplifier is shown in Fig. 8.20 b. In this circuit the amplifier has a single coupling element. The signal from the antenna is directed by the circulator to the parametric amplifier through feeder II, and the amplified signal, leaving the amplifier through the same feeder, is applied to the following stages of the receiver (branch III). The matched load in branch IV of the circulator serves to absorb power reflected in the feeder because of imperfections in its matching.

The second circuit is considered preferable since the circulator introduces less noise than the tubes.

Performance figures of a parametric amplifier. Parametric amplifiers are characterized by the following basic performance figures: power gain K_p, pass band ΔF, effectiveness ϵ, and noise figure N.

The way in which these indices are associated with the amplifier parameters depends on the type of amplifier and the method used to connect it into the channel.

For pass regenerating amplifiers, the performance figures are determined by the following equations:

$$\left.\begin{aligned} K_p &= \frac{4g_1 g_s}{g_r^2}, \quad \Delta F = \frac{g_r}{2\pi C_0}; \\ \Theta &= \frac{\Delta F}{f_0}\sqrt{K_p} \quad N = 1 + \frac{g_1 + g_r}{g_s}, \end{aligned}\right\} \qquad (8.14)$$

where $g_r = g_s + g_{oe} + g_1 + g_L - g_-$ - regenerating admittance of the oscillating system.

For reflecting regenerating amplifiers

$$K_p = \frac{4g_L^2}{g_r^2}, \quad N = 1 + \frac{g_1}{g_L}. \tag{8.15}$$

ε and ΔF for these amplifiers are determined by Eq. 8.14.

Here g_L, g_s load and signal source admittance respectively;

g_1 admittance of diode losses.

For a regenerating parametric amplifier-converter, power gain

$$K_{p_{max}} = 1 + \frac{f_1}{f_2}.$$

The remaining parameters are determined from Eq. 8.14 with $g_- = 0$.

Amplifiers using tunnel diodes

An amplifier using a tunnel diode effects amplification through energy from a dc source introduced into the circuit by a semiconductor diode using the tunnel effect. Because of the tunnel effect, the voltampere characteristic of a tunnel diode has a decreasing segment where increasing forward bias leads to a decrease in current. The tunnel diode resistance in this segment is negative; thus energy is introduced into the circuit from a dc source (the bias source).

A tunnel diode is characterized by the following parameters: critical frequency f_{cr}, loss resistance R_D, capacitance C_D; and figure of merit Q, which are determined just as the corresponding parameters of a parametric diode.

Using tunnel diodes at SHF produces amplifiers with low intrinsic noise. Figure 8.21 shows a centimeter range amplifier with tunnel diode operating in reflection, and Fig. 8.22 shows its equivalent circuit.

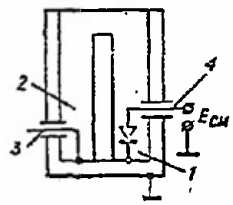


Figure 8.21. Schematic section of a SHF amplifier with a tunnel diode: 1) tunnel diode; 2) coaxial resonator; 3) input and output; 4) bias circuit.

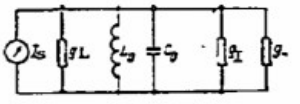


Figure 8.22. Equivalent circuit of an amplifier with a tunnel diode: g_L - load admittance; g_1 - admittance of the losses in the p-n junction of the diode; g_- - negative admittance; L_0, C_0 - equivalent inductance and capacitance of the circuit; I_s - equivalent signal current generator.

The basic performance figures of the amplifier are K_p, ΔF, and N, which are determined by Eq. 8.15 for the given circuit.

A comparative characteristic of the basic performance figures for the different types of high frequency amplifier is shown in Table 8.2.

Таблица 8.2

Качественные показатели различных типов УВЧ

1 Тип усилителей высокой частоты	2 T_e (°K)	3 N	4 $K_{пр}$	5 $K_{пр}$ дб	6 ΔF в % от f_c	7 λ_c
8 УВЧ на металлокерамических лампах . . .	150—600	1,5—3	50—100	20	2—7	1—3 м
9 УВЧ на ЛБВ	600—1200	2—5	100—10⁴	20—40	10	10 см
10 УВЧ на туннельном диоде	100—450	1,3—2,5	30—100	15—20	10—15	10 см
11 Электронолучевые ПУ	150—400	1,5—2,3	30—100	15—20	10	10 см
12 Полупроводниковые ПУ	100—300	1,2—2	—	—	0,1—5	10 см
13 Квантовые усилители	10—50	1,03—1,18	30—300	15—25	0,05—0,1	10 см

Table 8.2. Performance figures for various types of high frequency amplifiers.

1 - type of high frequency amplifier; 2 - T_{eq} (°K); 3 - N; 4 - $K_{пр}$; 5 - $K_{пр}$ dB; 6 - ΔF in % of f_s; 7 - λ_s; 8 - high frequency amplifier with cermet tubes; 9 - high frequency amplifier with TWT; 10 - high frequency amplifier with a tunnel diode; 11 - Electron ray parametric amplifier; 12 - semiconductor parametric amplifier; 13 - quantum amplifier.

8.6. Frequency Converters

Frequency converters are devices in which the carrier frequency of the accepted signal is converted to a lower, so-called intermediate frequency without changing the modulation or spectrum of the signal.

A frequency converter consists of a mixer and a heterodyne. Basic elements of a mixer are: nonlinear element, input device, and output filter, tuned to f_i.

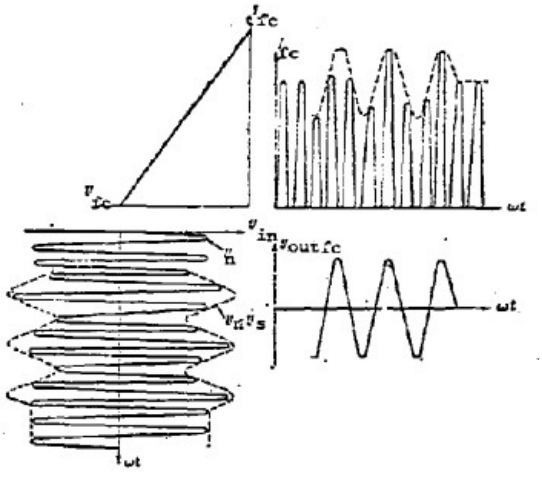


Figure 8.23. Frequency conversion principles

In the input device of the mixer (Fig. 8.23), heterodyne oscillations with frequency f_h and signal oscillations with frequency f_s form a beat frequency, that is, an oscillation modulated in amplitude and phase which carries information about the signal. This oscillation is applied to a nonlinear element. The current in the nonlinear element is a sequence of pulses, whose amplitude changes according to the beat frequency envelope. The frequency spectrum of this sequence contains oscillations at the signal and heterodyne frequencies, their harmonics, and oscillations at the combined frequencies ($nf_s \pm mf_h$), where m, n - arbitrary whole numbers. Oscillations at the combined frequencies carry information about the signal. To obtain oscillation with the same shape as the applied signal, the oscillation at only one combined frequency must be separated out. Usually oscillation at the difference frequency $f_h - f_s$ or $f_s - f_h$ is separated out with an output resonant filter. The difference frequency is called the intermediate frequency. The choice of f_i is considered below.

A converter using a harmonic of the heterodyne is sometimes used to provide decoupling between the circuits of the heterodyne and the input circuit of the mixer. In this case $f_i = mf_h - f_s$ or $f_i = f_s - mf_h$, where m is the number of the harmonic.

A frequency converter is a linear element of a receiver for signal oscillations, where the envelope of the heat frequency, acting in the

linear range of the voltampere characteristic of the nonlinear element, carries information about the signal. At the same time the converter is a nonlinear element for heterodyne oscillations used for beating and which operate in the nonlinear region of the mixer characteristic.

A frequency converter is characterized by the following basic parameters: conversion coefficient K_{fc}; noise figure N_{fc}, pass band ΔF_{fc}, input and output admittance $g_{in\ fc}$ and $g_{out\ fc}$.

The concept of conversion coefficient K_{fc} encompasses the ratio of voltage at the output at frequency f_i to voltage at the input at frequency f_s. The remaining parameters are the same as the analogous parameters of the amplifier.

The value of K_{fc} and N_{fc} depend on the amplitude of the herodyne voltage U_h and bias voltage U_b. U_h and U_b are considered optimum when the noise figure is sufficiently small, and the conversion coefficient is sufficiently large. Note that in converting harmonics of the heterodyne, K_{fc} is decreased while the noise figure N is increased.

In radar receivers single-grid and crystal type frequency converters are used. Furthermore, mixers with tunnel diodes or traveling wave tubes and semiconductor parametric amplifier-converters may also be used.

Single-grid frequency converters. In single-grid converters, which are used in the meter and decimeter ranges, pentodes or triodes are used as the nonlinear element of the mixer. Heterodyne and signal voltages are applied to the control grid.

The principal circuit of a single-grid frequency converter is shown in Fig. 8.24. In contrast to the amplifying mode, the operating point of the tube in the converting mode is chosen on the low bend of the plate-grid characteristic. Conversion at the first harmonic of the heterodyne is optimum if $U_h = U_b$, $U_h = 0.7\ E_{g0}$, where E_{g0} - cutoff voltage of the tube.

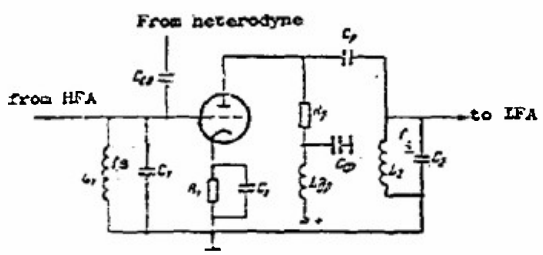


Figure 8.24. Schematic of a single grid frequency converter.

The required value of U_b is established, as a rule, by the constant component of cathode current, which provides the least shift in the conversion mode from optimum for a change in voltage. The tube is sometimes biased by the constant component of grid current set up by the heterodyne voltage from a fixed source. The required heterodyne voltage is provided by the amount of coupling between heterodyne and mixer. Capacitance, autotransformer, transformer, and cathode coupling are used. The most common, as well as the simplest, is capacitive coupling.

To insure stable and reliable heterodyne operation, and also to eliminate reciprocal effects during tuning, coupling between heterodyne and mixer should be low.

Conversion parameters are determined by the following:

$$K_{fc} = \frac{S_c}{g_\Sigma}; \qquad (8.16)$$

$$\Delta F_{fc} = \frac{g_\Sigma}{2\pi C_\Sigma}; \qquad (8.17)$$

$$N_{fc} = 1 + \frac{g_{or} + g_{in\,fc} + R_{neq} g_o^2}{g_{out\,HFA}} \qquad (8.18)$$

where S_c transconductance of conversion in the optimum mode, equal to $(0.25\text{-}0.28)S_{max}$;

g_Σ total admittance of the anode circuit of the mixer at the intermediate frequency, equal to the sum of plate circuit admittance, plate load admittance, converter output admittance, and input admittance of the IFA stage;

g_o total admittance of the mixer grid circuit at the signal frequency, equal to the sum of the converter input circuit admittance, mixer input admittance ($g_{in\,fc} = 0.5g_T + S_o \frac{ag}{C_\Sigma}$), output admittance of the last stage of the HFA, calculated at the mixer input;

$R_{n\,eq}$ equivalent noise resistance of the tube in the conversion mode; this resistance is considerably larger than the resistance of the same tube in the amplifying mode;

C_Σ total capacitance of the mixer plate circuit.

Crystal frequency converter. In crystal frequency converters, a semiconductor diode is used as the nonlinear element of the mixer. These converters are used in the centimeter and decimeter wave ranges. At decimeter wave lengths, vacuum tube diodes are sometimes used as the nonlinear elements,

and the converter is then called a diode converter.

Coaxial (Fig. 8.25 a) and waveguide (Fig. 8.25 b) constructions are most commonly used to transmit signal energy and heterodyne energy to the nonlinear element (diode). In some cases strip lines are used, which have the advantages of simple preparation, compactness, and wide band width.

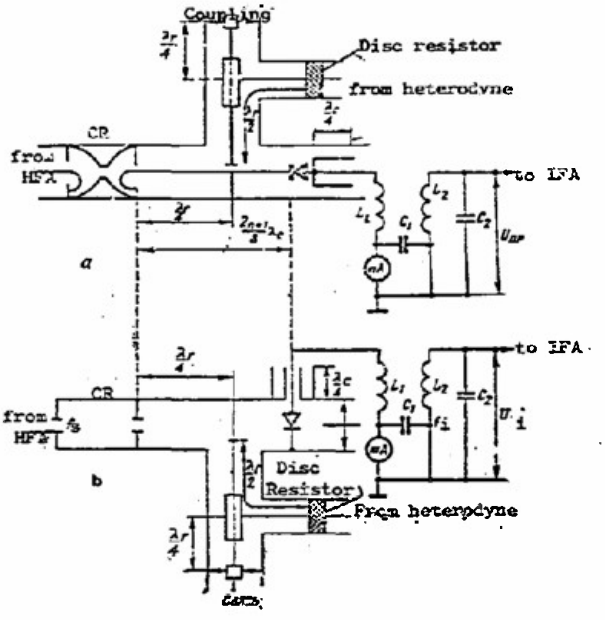


Figure 8.25. Schematic section of the construction of a crystal frequency converter: a - coaxial construction; b - waveguide construction

The construction of a converter should provide:

transmission of maximum signal power from the cavity resonator (CR) to the diode with minimum losses in the heterodyne circuit, HFA circuit, and minimum variation in diode parameters in the cavity resonator;

transmission to the diode of heterodyne power required for optimum conversion mode, while simultaneously providing stable operation in the heterodyne and no power transfer from the heterodyne to the receiver input.

These problems can be solved by placing the diode in the maximum current in the coaxial construction and in the maximum electric field in the waveguide construction and by matching the mixing chamber with the cavity resonator.

Minimum loss of signal power is obtained by providing low coupling between mixer and heterodyne and by including a quarter-wave filter in the intermediate frequency circuit. The effect of diode parameter breakdown is decreased by choosing the section from the diode up to the resonator equal to an odd number of $\lambda_s/8$ and assuring that its active impedance varies little with variations in the load over wide ranges.

The second problem is solved by varying the coupling between heterodyne and mixer. To assure stable operation of the heterodyne, a disc resistor, equal to the wave impedance of the feeder is included, and the sections from the probe to this resistor are chosen somewhat less than $\lambda_h/2$ to compensate for the reactance of the probe and the disc resistor. As a result, a traveling wave is set up in the feeder, coupling the heterodyne with the mixer, and the mixer does not affect heterodyne operation.

To prohibit energy transmission from the heterodyne to the receiver input, the distance from the probe to the cavity resonator is chosen equal to a quarter wave length of the heterodyne. This segment in the cavity resonator section is short-circuited; therefore the probe section has a large impedance which prohibits energy from the heterodyne from entering the cavity resonator. The cavity resonator presents a low impedance to the heterodyne frequency, since its pass band is small (high Q), and the resonator is sharply de-tuned relative to f_h.

The optimum conversion mode is controlled by the value of the constant component of diode current which is about 150-250 μA for germanium diodes and 500-800 μA for silicon diodes.

For mixing diodes, losses L_{dB} and relative noise temperature t_{fc} (see Chap. V) are introduced, which are used to express uniquely the conversion coefficient and noise figure of the mixer:

$$K_{fc} = 10^{-0.1\,L_{dB}}$$

$$N = \frac{t_{fc}}{K_{fc}} = t_{fc} \cdot 10^{0.1\,L_{dB}}$$

Heterodynes

Heterodynes are self-exciting low-power generators. They should provide suitable power to the mixer for obtaining optimum conversion and have sufficiently high frequency stability, sufficiently wide tuning range, and low intrinsic noise.

In the meter and decimeter range low-power electron tube generators with capacitive feedback are used as heterodynes. In the meter range circuits with lumped parameters are used, and in the decimeter range, circuits of long line segments are used. Reflex klystrons are usually used in the centimeter range.

The operating principle of a reflex klystron is based on interaction of the electron flow with the electromagnetic field of the resonator (Fig. 8.26). The electron gun sets up an electron flow which is accelerated by a constant voltage between cathode and resonator 2, and which penetrates the grid of the cavity resonator. The electromagnetic field in the gap between grids affects velocity modulation, that is, the velocity of an electron leaving the gap will be greater or less, depending on the phase of the field in the gap. In the space between resonator 2 and reflector 3, the electrons are turned back because the reflector potential is negative relative to the cathode potential. As the electrons continue to move along, velocity modulation becomes density modulation: the slower electrons which were accelerated in the resonator begin to overtake the electrons moving with average velocity, while the faster electrons, which have been slowed down, overtake the remaining electrons. Electron bunches are formed as a result.

The electron bunches again enter the space between resonator grids and under certain conditions (certain dimensions and voltages between the klystron electrodes) decellerate, giving off energy to the electromagnetic field of the resonator, and maintaining sustained oscillations in the resonator. Oscillations generated by the klystron are led to the mixer over coupling circuits and a coaxial cable.

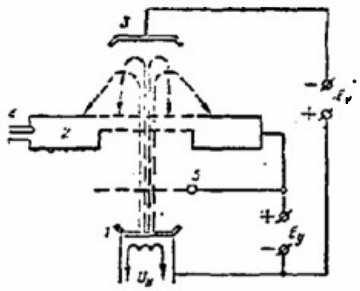


Figure 8.26. Construction of a reflex klystron: 1- electron gun; 2 -resonator; 3 -reflector; 4 -output; 5- accelerating electrode.

Frequency of the oscillations generated by a klystron may vary smoothly with variations of the reflector potential within certain limits. This property of a klystron is used for automatic frequency control (AFC) of a heterodyne. In this way the AFC system varies the reflector potential so that the difference between heterodyne and signal frequencies is equal to the nominal intermediate frequency.

In case a wide range of electron tuning is required, and also in the millimeter wave range where klystrons are not effective, backward wave tubes (BWT) are used as heterodynes.

Auxiliary reception channels

A superheterodyne receiver has, along with substantial advantages, certain shortcomings in comparison with direct amplifier receivers. The major shortcoming is the presence of auxiliary channels of reception, also called parasitic, since all kinds of noise are produced in them.

Parasitic reception channels are caused by the presence of signals with frequencies different from the signal frequency which, acting in the frequency converter, create oscillations at an intermediate frequency at its output. Combined oscillations are formed in the converter because of interaction between the input oscillations nf_{in} and heterodyne harmonics mf_h. This produces oscillations at the IFA input whose frequency is determined by

$$|nf_{in} \pm mf_h| = f_i$$

$$f_{in} = \frac{m}{n} f_h \pm \frac{1}{n} f_i$$

where m = 0, 1, 2, 3,...; n = 1, 2, 3...

BLANK PAGE

One of the values of f_{in} corresponds to the useful signal f_s (n=1; m=1 or 2, sign + when f_h is chosen less than f_s, or - when $f_h > f_s$), and all the rest are parasitic signals. Since the amplitude of the combined oscillations drops sharply with increasing m and n, practically speaking the most dangerous parasitic channels are the mirror symmetrical (n = 1, m the same as for f_s, and opposite in sign) and the direct pass channel (n=1, m=0) oscillations with frequency f_i.

Since the frequencies of the parasitic channels are transformed at the converter output to the same intermediate frequency as the useful signal, the useful signal cannot be separated from (filtered out of) noise at the converter output and in the IFA.

Up to the converter, the frequencies of the parasitic signals differ from the frequency of the useful signal; consequently frequency selectivity in the circuits preceding the converter (HFA, input circuit) is a likely method for eliminating parasitic channels.

The degeee to which parasitic channels are reduced is quantitatively characterized by the attenuation factor σ.

The attenuation factor of a stage σ is equal to the ratio of useful signal amplitude to the parasitic channel amplitude at the stage output, when the ratio of these amplitudes at its input is equal to one.

The value of σ for a single oscillating circuit is found by the ratio

$$\sigma = \sqrt{1 + \left(\frac{2\Delta f}{f_s} Q\right)^2},$$

where $\Delta f = |f_s - f_k|$;

f_k - frequency of the k-th parasitic channel. For a mirror channel, $\Delta f = 2f_i$.

The total attenuation of several stages is equal to the products of their attenuation factors.

8.7. Intermediate Frequency Amplifiers

Intermediate frequency amplifiers carry out the primary amplification of the signal. Using a constant intermediate frequency considerably lower than the frequency of the received signal makes it possible to obtain a high, undistorted amplification of the oscillations at the required frequency band for stable operation of the amplifier. The number of stages in the amplifier, commonly called a ruler IFA, may be from six to 10.

Single-circuit and multi-circuit intermediate frequency amplifiers are most widely used. In single-circuit amplifiers, the circuits of all stages may be tuned to one frequency or mutually staggered; in both cases the circuits are the same. Such a circuit is shown in Fig. 8.27.

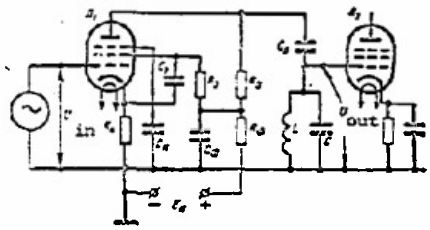


Figure 8.27. Schematic of a single-circuit intermediate frequency amplifier stage.

Dual-circuit schemes (Fig. 8.28) are the most widely used multi-circuit intermediate frequency amplifier devices. IFAs with dual-circuit staggered stages and IFAs where one of the stages has several (three or four) coupled circuits (the so-called concentrated selection filter) are sometimes used.

Intermediate frequency amplifiers are characterized by the following parameters: nominal intermediate frequency f_i, gain K, pass band ΔF, effectiveness ε, and rectangularity coefficient $K_{r0.1}$.

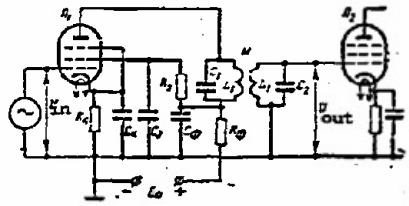


Figure 8.28. Circuit of an intermediate frequency amplifier stage with two coupled circuits.

In considering the nominal intermediate frequency, it must be noted that there are reasons for choosing either a higher or a lower frequency f_i. The fundamental advantages of a higher frequency are the greater selectivity in the mirror channel, smaller inductance windings, better separation of the intermediate and video frequencies. The advantages of a lower f_i are the greater amplification stability and the fact that the breakdown parameters of the circuit are not so critical. The intermediate frequency for radio pulses should also satisfy the inequality

$$f_i \geqslant \frac{20 \div 40}{\tau_p}.$$

In practice f_i = 10-100 MHz; $K_{r0.1}$ = 1.5-5; $K_t = 10^4$-10^6.

The product of the resonant gain times its pass band is called the effectiveness of one stage.

The product of the gain of one stage times the pass band of the whole amplifier is called the effectiveness of amplification ϵ.

The greater the effectiveness of a stage, the greater pass band it will have at constant gain, or the greater the amplification which can be obtained with constant pass band. Amplifiers are classed according to effectiveness. Amplifiers with tuned (resonant) circuits have the greatest effectiveness, multi-circuit IFAs have high effectiveness. Amplifiers with coupled circuits have higher effectiveness than amplifiers with staggered pairs.

The selective properties of an intermediate frequency amplifier are characterized by the coefficient of rectangularity $K_{r0.1}$, which indicates the ratio of the frequency band at a level of 0.1 of the maximum gain (or output voltage) to the pass band. For an ideal amplifier with a rectangular frequency characteristic, $K_{r0.1} = 1$. In practice $K_{r0.1} > 1$. An amplifier is more selective the more closely its frequency characteristic approximates the rectangular, and consequently, the more closely $K_{r0.1}$ approaches one.

Amplifiers with single circuits have lower selectivity than amplifiers with two circuits, etc. The remaining amplifiers are approximately equivalent.

In practice IFA with single tuned circuits are more widely used for amplifying radio pulses longer than 2μs, since these amplifiers, except for the faults discussed above, are quite superior in comparison with other types of IFA: they are simpler to build and tune, have the shortest delay time, do not have parasitic blips in the radio signal, and are least critical toward breakdown of the circuit elements, especially the tube capacitance.

For amplifying radio pulses less than 2μs in length when the amplifier should have high effectiveness, for which the IFA using tuned circuits should have 10 to 15 stages, six to eight stage IFAs with staggered tuning (in duplicates) or IFAs with two coupled circuits are used. Table 8.3 shows the calculated relationships for n stages of these types of IFA and a comparative estimate of their effectiveness.

8.8. Detectors

Detection is the process of transforming the modulated signal to reveal the information. This process is the reverse of modulation.

Depending on the form of the modulated signal, detectors are classed as amplitude, frequency, and phase detectors.

Amplitude detectors. Amplitude detection comprises two fundamental processes:

Таблица 8.3

Сводная таблица формул параметров УПЧ

1 Параметр \ 2 Тип УПЧ	3 УПЧ с одиночными настраиваемыми контурами	4 УПЧ на парах расстроенных контуров	5 УПЧ со связанными контурами
6 Коэффициент усиления $K_{0\text{вых}}$	$K_{01}^{n} = (SR_0)^n$	$K_{01}^{n} = \left(\frac{S\sqrt{R_{01}R_{02}}}{\sqrt{2}}\right)^n$	$K_{01}^{n} = \left(\frac{S\sqrt{R_{01}R_{02}}}{2}\right)^{2n}$
7 Полоса пропускания $\Delta F_{0\text{вых}}$	$\frac{f_0}{Q}\sqrt{\frac{1}{2^{\frac{1}{n}}}-1}$	$\frac{\sqrt{2}f_0}{Q}\sqrt[4]{\frac{2}{2^{\frac{1}{n}}}-1}$	$\frac{\sqrt{2}f_0}{Q}\sqrt[4]{\frac{1}{2^{\frac{1}{n}}}-1}$
8 Эффективность усилителя $Э$	$\frac{S}{2\pi C_0}\sqrt{\frac{1}{2^{\frac{1}{n}}}-1}$	$\frac{S}{2\pi C_1}\sqrt[4]{\frac{2}{2^{\frac{1}{n}}}-1}$	$\frac{S\sqrt{2}}{4\pi\sqrt{C_{01}C_{02}}}\sqrt[4]{\frac{1}{2^{\frac{1}{n}}}-1}$
9 Коэффициент прямоугольности $K_{0.1}$	$\sqrt[n]{\frac{\sqrt[n]{10^2}-1}{\sqrt[n]{2}-1}}$	$\sqrt[4]{\frac{\sqrt[n]{10^2}-1}{\sqrt[n]{4}-1}}$	$\sqrt[4]{\frac{\sqrt[n]{10^2}-1}{\sqrt[n]{2}-1}}$
10 Отношение эффективностей УПЧ $\frac{Э}{Э_1}$	1	$1{,}3\sqrt[4]{n}$	$1{,}55\sqrt[4]{n}$

Table 8.3. Summary table of formulas for IFA parameters.
1 - parameter; 2 - type of IFA; 3 - IFA with single tuned circuits; 4 - IFA with staggered pairs; 5 - IFA with coupled circuits; 6 - gain K_t; 7 - pass band ΔF_t; 8 - amplifier effectiveness ε; 9 - rectangularity coefficient $K_{0.1}$; 10 - ratio of IFA effectivenesses.

creation of a low frequency signal (envelope of the amplitude-modulated oscillation or video pulse) on the basis of a high frequency AM oscillation or radio pulse;

separating (filtering out) useful low frequency signal from the high frequency oscillations.

An amplitude detector consists of three basic elements: a nonlinear element, a load resistance R, and a filter (capacitive load) C. The most common are amplitude diode detectors. Their circuit is shown in Fig. 8.29.

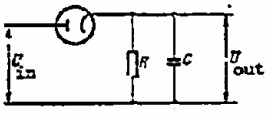


Figure 8.29. Schematic of a diode detector

The first part of the process in amplitude detection is carried out by the diode and the resistance R, the second is carried out by capacitor C. When the high frequency signal modulated by a low frequency oscillation acts on the nonlinear element, current with a complex form flows in the diode circuit, creating a voltage on the load resistance R with the frequency of the envelope of the AM oscillation, and the remaining current components which have higher frequency flow through capacitor C, which presents a low impedance to these frequencies. If a radio pulse acts on the nonlinear element (Fig. 8.30), the diode conducts current during the positive half-period of the high frequency oscillations, setting up a voltage on resistor R, and the capacitor C charges rapidly. In the negative half-period of the high frequency oscillations, the diode is blocked, capacitor C slowly discharges through R, creating a voltage of the same polarity on it as that created by the current flowing through the diode in the positive half-period. Consequently a positive video pulse appears on R with pulsations at the tip which are removed by a special LC filter connected between the detector and the video amplifier. The polarity of the output signal depends only on the way the diode is connected.

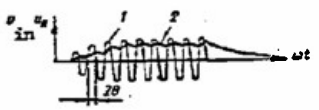


Figure 8.30. Graph of voltages in an amplitude detector: 1 - voltage at the detector input; 2 - voltage on the load

An amplitude detector appears as a nonlinear element of the receiver to the input signal. It deforms the signal spectrum and impairs the signal/noise ratio because of the interaction of the frequency components of the signal spectrum and the noise spectrum.

The latter phenomenon decreases for $U_{in} \geq (0.5-2)$ V, where the detection process becomes linear, i.e., U_{out} is proportional to U_{in}. At low signals U_{out} is proportional to U_{in}^2, and the detection process is quadratic.

A radio pulse detector is characterized by the transmission coefficient K_D, rise time τ_s and decay time τ_d, input and internal impedance, R_{in} and R_i respectively.

These parameters are uniquely determined by the slope of the voltampere characteristic S of the nonlinear element and by the load parameters R and C:

$$\left.\begin{aligned} K_D &= \cos\theta, \quad \tau_0 = 2{,}2R_iC; \\ \tau_d &= 2{,}2RC, \quad R_i = \frac{\operatorname{tg}\theta - \theta}{\pi} R; \\ R_{in} &\approx \frac{R}{2K_D}, \quad R_{in} = \frac{\pi R_i}{S(\theta - \sin\theta\cos\theta)}, \end{aligned}\right\} \qquad (8.19)$$

where θ - cutoff angle, by definition equal to half the period of oscillation in the course of which the diode conducts current (Fig. 8.30), and uniquely determined by:

$$\operatorname{tg}\theta - \theta = \frac{\pi}{SR}. \qquad (8.20)$$

The load elements are usually found from the conditions:

$$RC < \frac{\tau}{5 \div 10}; \quad C \geqslant 10C_{ak} \qquad (8.21)$$

where C_{ak} anode-cathode capacitance of the diode.

Electron-tube and semiconductor triodes which, depending on the method of connecting the detector load, may operate as grid, plate, and cathode detectors, are used for the detection of AM signals. For details see [L2] and [L3].

Synchronous detector. Detection of AM oscillations may be done by linear elements (L, C, R) with varying parameters. These detectors have a valuable characteristic: they permit synchronous (commutator or amplitude-phase) detection. Detection is considered synchronous when the amplitude and polarity of the output signal depend on the difference in phase of the oscillations received by the detector and the oscillations changing the detector parameters. This permits these detectors to be used in radar stations to distinguish between moving objects and stationary objects or passive noise. The diode detector is the simplest of this type, with its internal resistance changed by a special generator which has the same frequency as the signal received by the detector.

A coherent heterodyne is generally used for this purpose in radar stations. To obtain maximum output signal from the detector, the oscillations of the coherent heterodyne and the input signal should have a certain phase shift, depending on the ratio of their amplitudes.

Asynchronous detector. Since in many cases it is difficult to synchronize the phase of these oscillations, detectors which do not require synchronization are often used. These detectors are called asynchronous. In the simplest case, an asynchronous detector is a diode detector, with two signals applied to its input:

the signal from the intermediate frequency amplifier output;

oscillations from an arbitrary low power generator with frequency $f_h = f_i$.

These detectors have many advantages over the usual diode detector (larger transmission coefficient, better signal/noise ratio, and others).

Frequency detector. A frequency-modulated signal carries information in the variations of its frequency. The process of detecting this information is called frequency detection (FD).

Frequency detection in the majority of detectors consists of two processes:

converting frequency-modulated oscillations into an amplitude-modulated signal;

amplitude detection of the latter and exposure of its information.

Any linear system whose transmission coefficient depends on frequency may be used for the first part of the process. The simplest system is an oscillating circuit detuned relative to the center frequency f_c of the FM oscillations. In this case, depending on the ratio of signal frequency f_s and intrinsic frequency f_o of the circuit, the amplitude of the oscillations varies at the output (Fig. 8.31). This preserves the frequency modulation in the output signal, i.e., the signal at the output has both amplitude and frequency modulation, and is called amplitude-frequency modulation (AFM). Amplitude detection of AFM oscillations is usually done by diode detectors (Fig. 8.32).

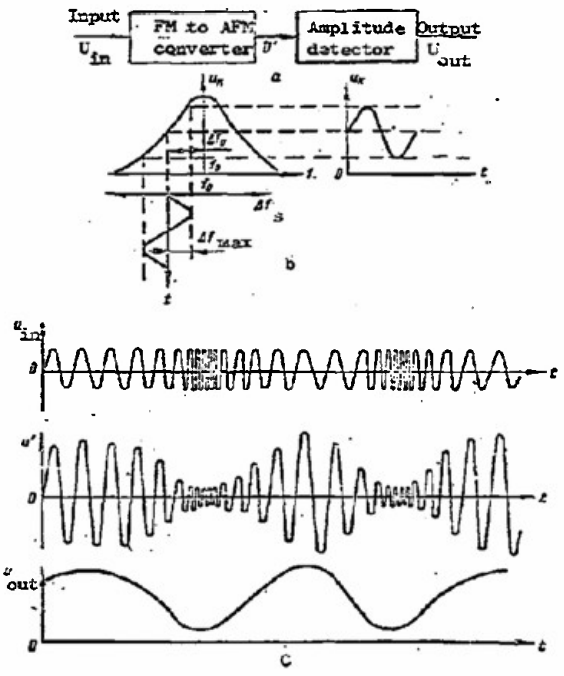


Figure 8.31. Process of converting frequency-modulated oscillations into amplitude-frequency-modulated oscillations.

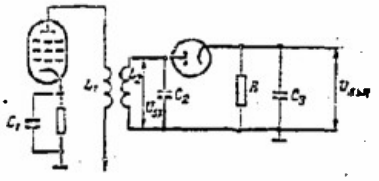


Figure 8.32. Circuit of the simplest frequency detector.

A frequency detector is characterized by the slope of the amplitude-frequency characteristic S_{fd}, indicating voltage change at the output for a 1 MHz (or kHz) change in frequency from f_c:

$$S_{fd} = \frac{2U_{02}QK_D}{f_\phi(1+\alpha_0^2)^{1/2}}, \qquad (8.22)$$

where U_o - maximum voltage amplitude on the circuit at resonance;

K_D transmission coefficient of the amplitude detector;

initial detuning of the circuit;

Q figure of merit of the circuit;

$\Delta f = |f_c - f_s|$.

Two-cycle frequency detector circuits are employed to increase transconductance S_{fd} and the allowable detuning (deviation) Δf of the FM oscillations (Fig. 8.33) when the circuits are detuned symmetrically relative to the center frequency of the FM oscillations and the output voltages of the detectors are mutually subtracted. If the signal frequency is not f_c, voltages on the detectors will be different and a signal with positive or negative polarity, depending on the type of detuning ($f>f_c$ or $f<f_c$) will appear at the output.

Detectors operating analogously to phase detectors (fractional frequency detectors, heterodyne frequency detectors, see [L3] for details) may be used for frequency detection. A frequency detector is one of the basic elements of automatic frequency control system.

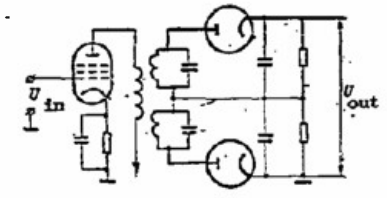


Figure 8.33. Schematic of a two-cycle frequency detector

Phase detector. In phase modulation, information is carried in the phase variations of the high frequency oscillation. The phase, varying with time, represents the speed with which frequency changes. Of course speed must be measured with reference to another object. Therefore, to evaluate phase changes, the phase of the modulated oscillations must be compared with the phase of another oscillation taken as reference.

Phase detection (PD) involves two processes:

addition of phase-modulated U_s and the reference oscillation U_{ref}, forming an amplitude-phase modulated oscillation;

amplitude detection of the oscillation obtained.

The simplest circuit of a single-cycle phase detector and the vector diagram of the addition at its input are shown in Fig. 8.34.

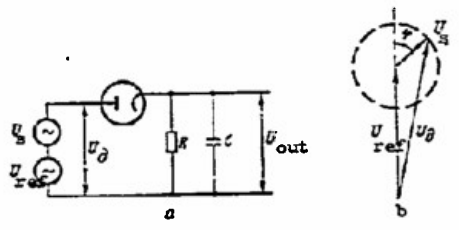


Figure 8.34. Phase detector:

a - circuit; b - vector diagram of voltages at the input.

A phase detector is characterized by the slope of the amplitude-phase characteristic S_{fd}, indicating the voltage change at the output for a one degree change of phase.

A balanced circuit PD (Fig. 8.35) is used to increase S_{fd}. The outstanding feature of this circuit is that the signal is applied to the diodes out of phase, and the reference voltage is in phase. Thus, if the phase of the signal and the phase of the reference voltage do not coincide, a voltage difference will appear on the diode, and the polarity of the signal may change.

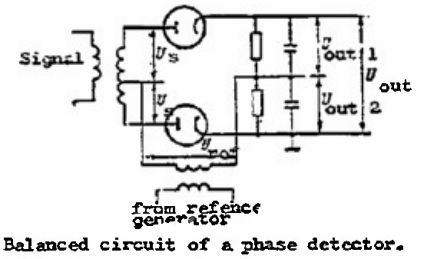


Figure 8.35. Balanced circuit of a phase detector.

The value of S_{fd} depends on the relative amplitudes of the reference voltage U_{ref} and the phase-modulated signal U_s as well as the phase shift φ between them.

For $m = U_s/U_{ref} = 1$

$$S_{fd} = K_D U_s \left(\left| \cos \frac{\varphi}{2} \right| + \left| \sin \frac{\varphi}{2} \right| \right); \qquad (8.23)$$

for $m \leqslant 1$

$$S_{fd} = 2K_D U_s \sin \varphi.$$

Phase sensitivity (slope) is most even when $U_s = U_{ref}$, and under these conditions is maximum for $\varphi = (2n+1)\frac{\pi}{2}$.

Amplitude limiters are used to limit the signals so as to eliminate the effects of variations in signal amplitude on the output voltage.

8.9 Video Amplifiers

The purpose of the video amplifier in a receiver is to boost the detected signal (video pulse) to the level required for normal operation of the indicator devices.

One of the possible variations of a two-stage video amplifier is shown in Fig. 8.36.

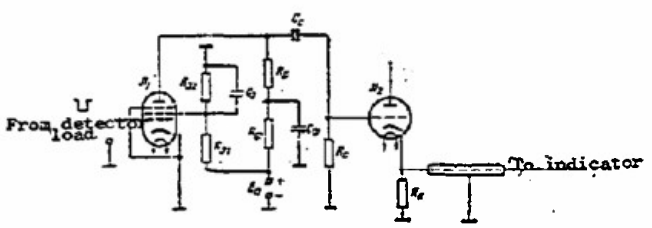


Figure 8.36. Video amplifier circuit

As a rule the first stage operates as a limiter to prevent overloading the indicator devices.

The second stage (a cathode follower) matches the output impedance of the receiver to the wave impedance of the cable and transmits the amplified signal along the cable to the indicator device without distortion.

First stage gain

$$K_1 = SR_a. \qquad (8.24)$$

where S - slope of the tube characteristic;

R_a plate load impedance.

First stage gain is usually equal to 10-20.

For non-distorted amplification, the stage should have a sufficiently

wide pass band. The pass band of a video amplifier is the difference of its upper and lower limiting frequencies.

The upper boundary frequency is found from the resistance and capacitance of the plate circuit:

$$f_{max} = \frac{1}{2\pi C_0 R_a} \qquad (8.25)$$

Capacitance C_o, composed of the tube capacitance and the assembly capacitance, is usually equal to 15-30 pF.

The value of f_{max} uniquely determines the pulse rise time

$$t_r = \frac{0.35}{f_{max}}.$$

To increase f_{max} (decrease rise time of the pulse) high-frequency correction is sometimes employed, connecting an inductor in series with R_a.

The lower boundary frequency is determined by the transfer capacitance C_g and resistance R_g:

$$f_{min} = \frac{1}{2\pi R_g C_g}. \qquad (8.26)$$

The value of f_{min} uniquely determines drop G at the tip of a pulse with duration τ_p:

$$G = 2\pi \tau_p f_{min}$$

Accordingly, to decrease f_{min} (drop G at the pulse tip) R_g and C_g should be quite large. Sometimes low frequency correction, choosing C_g of a certain value depending on R_a and R_g, is used to decrease G.

The gain of a cathode follower is less than one, since all output voltage from the resistor is again applied to its control gird (100% negative feedback).

However, a cathode follower with low output impedance may feed into a cable with high capacitance, and this is one of the major reasons for its use.

The performance figures on a cathode follower are determined by:

$$\left.\begin{aligned} K_{kf} &= \frac{SR_k}{1+SR_k}; \\ f_{max\,kf} &= \frac{1+SR_k}{2\pi C_0 R_k}; \\ f_{min\,kf} &= \frac{1}{2\pi R_g C_g}, \end{aligned}\right\} \qquad (8.27)$$

where R_k - impedance of the cathode load.

If necessary, the signal may be further amplified before it is applied to the indicator by video amplifier stages in the indicator block.

8.10 Amplifier Control in the Receiver

When a station is operating, radio pulses of differing intensity, depending on the distance to the target, its reflecting surface and other factors, are applied to the receiver input. Weak reflected signals, as well as strong ones, must be observed. This requirement can be met if the receiver amplification can be changed rapidly, which requires the use of automatic gain control.

Over a period of time, the tubes in the receiver age and are replaced; consequently its gain changes. But for a given signal at the input, a given signal intensity should be obtained at the output. This is accomplished by using manual gain controls (MGC). Manual gain controls allow the operator to choose the gain depending on the intensity of the reflected signal.

The most common method of automatic and manual gain control in radar receivers is based on changing the transconductance of the tube in one or more stages of the intermediate frequency amplifier. Control is sometimes applied to the video amplifier.

Transconductance of a tube may be changed by changing the voltage on one or more of its electrodes. Bias voltage and voltage on the screen grid are most often used for control. Changing plate voltage and voltage on the shielding grid are not very effective.

The disadvantage inherent in controlling gain by changing voltage on the screen grid (Fig. 8.37) is that, with bias constant and gain decreasing (decreasing screen voltage), the grid characteristic shifts to the right, increasing the danger of overloading the stage.

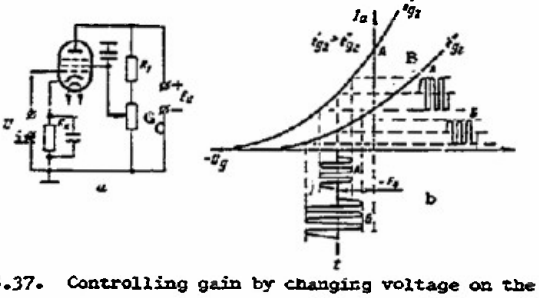


Figure 8.37. Controlling gain by changing voltage on the screen grid: a -schematic of the stage; b -graph of input and output voltages of the stage for two operating modes

Bias may be changed by changing the potential of the cathode or the control grid. A control circuit which varies cathode potential has large inertia and requires considerable energy from a constant source, which is dissipated as heat in the control elements. The control circuit with

variable control grid voltage shown in Fig. 8.38 eliminates these faults.

Protecting a receiver from active and passive noise is discussed in Chapter XI.

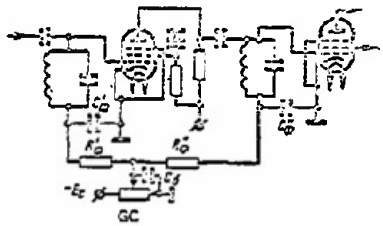


Figure 8.38. Schematic of an IFA stage with gain controlled by varying control grid potential

Chapter IX

Displays

Radar utilizes various types of displays to observe the air situation, detect targets, and establish their coordinates.

The type of display used depends on the purpose for which the radar is intended. PPI displays can be used for detection and for making rough approximations of coordinates. Range-azimuth and range-height type sector displays can be used if more accurate measurements of coordinates are required, and if target characteristics must be found more accurately. Special displays, on which conventional signals and symbols will glow along with the radar signals, can be used when the radars are linked with automated control systems.

9.1. Effect of the Display on Radar Characteristics. Conditions for Signal Observation and Visibility

The display is the final link in the radar system. It reproduces all the signals received by the antenna and produced by the receiver. It is the task of the operator watching the screen to make a careful analysis of all the information on the display in order to separate the useful signals from the interference.

Once the operator has detected the targets, he has a second task to perform, that of establishing the range, R, the azimuth, β, and the height, H, after which, and in so far as this is possible, he must establish target characteristics.

Consequently, the most important stage of radar operation is associated with observation of targets on displays, and, naturally, the efficiency with which the displays function is a very important factor in the successful completion of an assigned mission.

Modern radars make widespread use of displays with intensity markers and afterglow. Displays with amplitude markers are also used.

Receiver noise causes a characteristic "noise background" on the screens of intensity marker displays. The background consists of a great many luminous spots of different widths, brilliance, magnitudes, and shapes, with many of them resembling targets, at least at first glance, thus making the work of the observer difficult. The structure of the noise background is the result of many factors, the most important of which are the pass bands of the IF and video amplifiers, sweep speed, antenna rotation rate, and some others. The pass band, ΔF, will be close to optimum in this case if it is associated with the pulse width (τ_p) by the relationship at (8.2).

The so-called visibility factor, v_v (See Chapter I) is used to make a qualitative assessment of signal visibility against a background of noise.

The smaller the visibility factor, v_v, the lower the power of the detected signal.

The different factors bearing on the fundamental engineering characteristics of the radar have a very definite effect on the magnitude of the visibility factor.

Those with the greatest effect on the magnitude of the visibility factor are:

pulse repetition frequency, F_p;

receiver pass band, ΔF;

sweep speed, v_s;

pulse width, τ_p;

the quality with which the electron beam in the tube used in the display is focused;

external conditions prevailing when the observation is made;

antenna rotation rate.

The dependence of the visibility factor on the pulse repetition frequency is shown in Figure 9.1a. Signal observability improves with an increase in the pulse repetition frequency and, consequently, in the number of pulses handled in the receiver.

The effect of receiver pass band at the intermediate frequency on observability can be followed from the graph in Figure 9.1b.

The curve in this figure is for the case when the sweep speed, v_s, and pulse width, τ_p, remain constant and the only facts that changes is the pass band, ΔF.

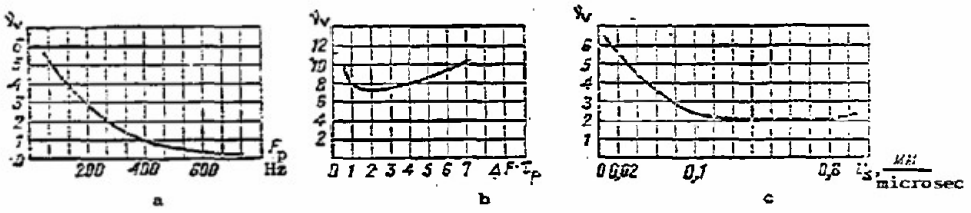


Figure 9.1. Dependence of v_v on repetition frequency (a), pass band (b), and sweep speed (c).

The trace of the curve shows that there is a clearly defined minimum for the visibility factor when $\Delta F_{\tau p} \approx 1$.

The dependence of the change in observability on sweep speed is shown in Figure 9.1c.

Degradation of observability with increase in sweep speed results because at high speeds of passage of the beam over the surface of the screen the conditions under which the screen is excited are degraded, and besides there is an increase in the "graininess" of the noise background, that is, the markers from the noise blips increase in magnitude.

So far as observability is concerned, the optimum magnitude of a marker for screens with intensity markers is considered to be

$$l_{opt} = (1 \text{ to } 2)\, d_s \qquad (9.1)$$

where

d_s is the spot diameter.

Widely used displays have d_s = 1 mm, l_{opt} = 1 to 2 mm.

The effect quality of focusing of the electron beam has on observability is to degrade contrast, because when focusing is poor the target and noise markers "dissolve."

Observability is effected significantly by the average brilliance of the noise background brightening the screen, by the external conditions under which the observations are being made. This is fixed on the one hand by the properties of the cathode ray tube, and on the other by the special features inherent in the human eye (see #9.10).

Antenna rotation rate too has a definite effect on observability because when all other conditions are equal, the rotation rate is what determines how many pulses will be present at a definite point on the screen.

The number of pulses in the train will decrease with increase in antenna rotation rate, and this will result in a reduction in marker brilliance.

Effect of Display on Radar Detection Range

The radar equation at (1.48) defines the dependence of radar detection range on inherent parameters and target characteristics.

Detection range for a surveillance radar can be found through the following dependence on the visibility factor

$$R = \frac{k_o}{\sqrt[4]{v_v}}, \qquad (9.2)$$

where

k_o is a coefficient that takes radar and target parameters into consideration.

The smaller v_v, that is, the smaller the value of the threshold signal, the longer the radar detection range.

Widely used three-scale PPIs have approximately the following visibility factor magnitudes:

on small scale - $v_v \approx 3.5$;

on medium scale - $v_v \approx 3$;

on large scale - $v_v \approx 2$.

The change in the magnitude of the visibility factor with change in the scale on the screen is the result of the change in sweep speed.

The relationship at (9.1) links the magnitude of the spot, d_s, and the length, l, of the marker on the screen in the optimum case. At small scale settings for the sweep this condition cannot be satisfied, and shows up in the magnitude of the visibility factor.

The linear dimension of the marker on the screen can be computed through the formula

$$l \approx R_{scale}/L_s \cdot d_s, \tag{9.3}$$

where

R_{scale} is the scale used on the screen;

L_s is the linear dimension of the sweep.

Effect of Display on Radar Resolution in Range and Azimuth

The relationship at (2.9) is used to establish the range resolution for the radar. It can be written in the following form

$$\delta R \doteq \frac{p(\tau_p + \Delta\tau)}{2} + m\frac{R_{scale}}{L_s} d_s \tag{9.4}$$

where

m is a coefficient that takes the magnitude of the "gap" between two adjacent markers into consideration.

The second term in the right hand side of the expression at (9.4) defines the display's resolution, δR_d, or, putting it another way, it demonstrates the degradation in radar resolution attributable to the display.

An analysis of the second term will lead to the conclusion that improvement in display resolution can be arrived at in the following ways:

reduce the range sector displayed on the screen;

increase the linear dimension of the sweep; that is, increase screen dimensions;

reduce the dimension of the spot on the screen.

The only way resolution can be improved for a selected type of cathode ray tube is to use several range scales. Two or three different scales are usually used.

The first scale, the smallest, covers the entire scanning zone at maximum range. The second scale is a one and one-half to two-fold amplification of the first, while the third is a four- to five-fold amplification of the second.

The display's resolution changes accordingly as the shift is made from one scale to another.

At long ranges the PPIs include a ring scan, that is, individual sectors (rings) covering the whole of the zone of observation, can be looked at at large scale. In this case range sweep begins after some delay corresponding to the distance to the beginning of the sector being looked at, rather than at the moment the main pulse is radiated.

Resolution in azimuth can be found through the formula

$$\delta\beta = \varphi_{0.5} + m \ \beta_{scale}/L_s \cdot d_s, \qquad (9.5)$$

where

β_{scale} is the magnitude of the azimuth sector in degrees;

L_s is the linear dimension of the azimuth sweep;

m is a coefficient that takes the magnitude of the gap between markers into consideration.

The second summand in the right hand side of the expression at (9.5) characterizes the resolution in azimuth for the display, $\delta\beta_d$. In the case of the PPI the magnitude of the sector $\delta\beta_{scale}$ is always 360°, while the length of the sweep will be variable, depending on the distance the marker is from the center of the screen.

The linear dimension of the sweep can be calculated as the circumference of a circle, that is,

$$L_s = 2\pi r,$$

where

r is the distance the marker is from the center of the screen.

Analysis of the second term in the expression at (9.5) leads to the conclusion that in order to increase the resolution in azimuth it is necessary to reduce the magnitude of the sector brought to the screen for the type of cathode ray tube selected. The use of the sector scan mode is sought out for this purpose.

It is also possible to simultaneously reduce the magnitude of the sector β_{scale}, and increase the linear dimension of the sweep, L_s. This is done in special scopes of the range-azimuth type, in which the azimuth and range sweeps form a rectangular raster.

Effect of Display on Coordinate Measurement Accuracy

Errors resulting from the inaccuracy in distinguishing the centers of target pips and scale markers because of their finite magnitudes. Since target pips as well as range and azimuth scale markers have finite dimensions, the operator has to make a visual determination of their centers, corresponding to the true position of the target, or of the scale marker. This operation always contains an error. The mean square error, determined experimentally, is about 3% of marker magnitude.

Errors associated with poor focusing. The error in fixing the center increases with increase in the marker resulting from poor focusing. The mean square error in range is on the order of tens of meters, in azimuth on the order of minutes.

Errors in interpolating the position of the target pip between the scale divisions. Interpolation involves a visual, proportional division of the segment between two scale markers within the limits of which the target pip is located. The mean square error in interpolation can be computed through the formula

$$\sigma_{int} = k \cdot 0{,}005\Delta_{scale} \tag{9.6}$$

where

$k = 1$ to 2, and depends on the scale setting;

Δ_{scale} is the distance between azimuth or range markers.

Errors caused by nonlinearity in sweep. Nonlinearity in the sweep causes errors in measuring the coordinates, and these errors are associated with those of interpolation in various sectors of the sweep. Nonlinearity in range sweep results in range determination errors.

In this case the mean square error can be found through the formula

$$\sigma_{int} = \frac{\alpha}{8}\Delta_{scale} \tag{9.7}$$

where

α is the nonlinearity in the sweep.

As will be seen from the foregoing, the errors in interpolation have a significant dependence on the distance between scale markers, expressed in terms of the scale setting used.

9.2 Plan Position Indicators

The PPI can be used to make a rough determination of target range and azimuth, as well as to make an approximation of target characteristics. On the PPI the air situation is in the form of a projection on the horizontal plane. A PPI screen is shown in Figure 9.2. The PPI usually can be used to scan in three operating modes:

circular scan, during which the entire radar coverage zone is looked at;

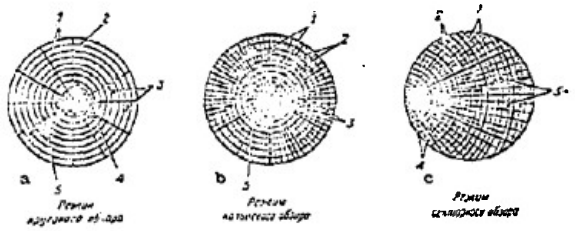


Figure 9.2. Images on the PPI screen in three operating modes. a - circular scan mode; b - ring scan mode; c - sector scan mode. 1 - range markers; 2 - azimuth markers; 3 - reflected signals; 4 - signals reflected from local objects; 5 - signals reflected from clouds.

ring scan, during which the entire radar coverage zone is looked at by sectors at enlarged scale;

sector scan, when only a specific sector is looked at.

A range sweep delay circuit, inserted in the range sweep channel, can be used to obtain the ring scan mode.

A circuit for displacing the center of the sweep can be used to obtain the sector mode.

The channel for forming the azimuth sweep provides range sweep synchronized with antenna rotation. All PPIs can be divided into two groups with respect to the method used to obtain a rotating sweep:

those with a rotating deflection system;

those with a fixed deflection system.

The distinguishing feature of those in the first group is that rotation of range sweep with respect to the azimuth can be synchronized with the antenna by rotating the deflection system mechanically, causing the range sweep to rotate.

The deflection system in the PPIs in the second group is fixed. Sweep rotation occurs as a result of rotating the magnetic field.

The block diagram of a PPI with a rotating deflection system is shown in Figure 9.3.

The trigger stage is the source forming the range sweep and can be used in any PPI operating mode, not only to eliminate pulse pickups, but to get stable trigger pulse amplitude. A wideband resistance-coupled amplifier with bottom clipping can be used as the trigger stage. Chapter VI contains the circuits of similar clipping amplifiers.

Specific requirements imposed on circuits for trigger pulse delay are:

minimum time to restore the original condition, particularly important in the case of long delays;

linearity of the delay over the entire range of the scale.

These requirements are best satisfied by circuits in which the voltage drops linearly and which are, moreover, the most efficient.

A circuit providing smooth trigger pulse delay, and widely used in displays, is shown in Figure 6.9. Jump delay circuits can also be used. A block diagram of a possible variant of the jump trigger pulse delay, and voltage curves, are shown in Figure 9.4.

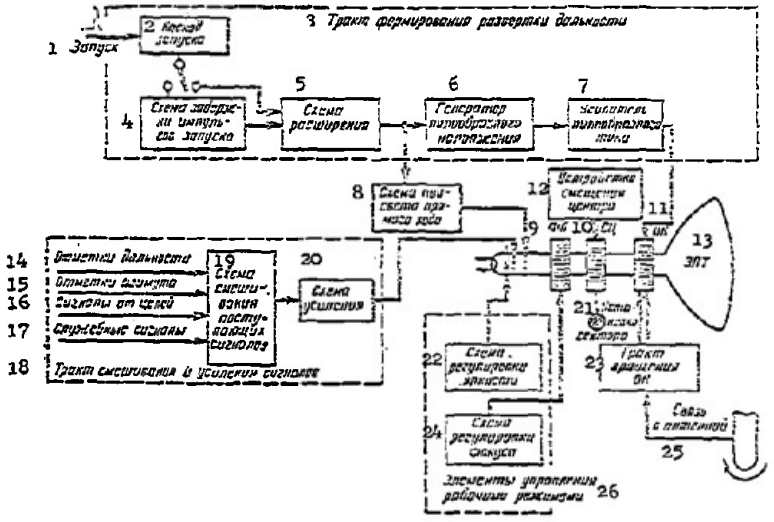


Figure 9.3. Block diagram of a PPI with rotating deflection system. 1 - trigger; 2 - trigger stage; 3 - channel for forming range sweep; 4 - trigger pulse delay circuit; 5 - expansion circuit; 6 - sawtooth voltage generator; 7 - sawtooth current amplifier; 8 - outgoing trace intensifier circuit; 9 - focus coil; 10 - center shift; 11 - deflection coil; 12 - center shift device; 13 - cathode ray tube; 14 - range markers; 15 - azimuth markers; 16 - target signals; 17 - procedure signals; 18 - signal mixing and amplifying channel; 19 - circuit for mixing incoming signals; 20 - amplifier circuit; 21 - sector setting; 22 - brilliance control; 23 - deflection coil rotation channel; 24 - focus control; 25 - antenna coupling; 26 - elements for controlling operating modes.

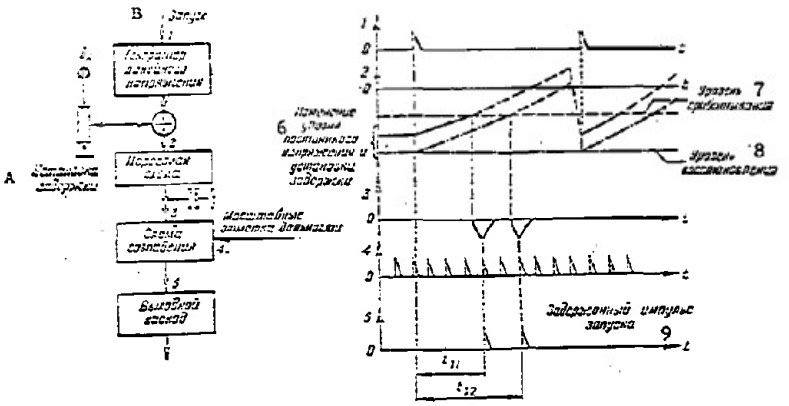


Figure 9.4. Block diagram of jump trigger pulse delay and voltage time diagrams for various points in the delay. A - delay setting; B - trigger; 1 - linc voltage generator; 2 - threshold circuit; 3 - coincidence circuit; 4 - range scale markers; 5 - output stage; 6 - change in level of constant voltage and delay setting; 7 - firing level; 8 - restoration level; 9 - delayed trigger pulse.

The trigger pulse excites the generator, the sawtooth voltage from which, added to the constant regulated voltage, acts across the grid of a tube with a clipped operating level (the threshold circuit). If a change is made in the magnitude of the constant component of the voltage acting across the tube grid, the tube will fire at different sawtooth voltage levels.

The magnitude of the constant voltage can be changed smoothly by rotating the arm of the delay setting potentiometer, and correspondingly changing the moment in time the tube fires. The pulse corresponding to the moment of firing is fed into the coincidence circuit, into which range scale markers are also fed for synchronization purposes.

The coincidence circuit produces a pulse which can be used as the delayed trigger pulse.

One of the fundamental requirements imposed on expansion circuits is a minimum time requirement for restoring to the original condition.

Various types of triggers, in which cathode followers are used to cut restoration time, can be used as expansion circuits.

Regardless of the circuit used, there is distortion in the shape of the expanded pulse, the result of the transition processes that take place during current and voltage drops. This has a negative effect on the operation

of the sawtooth voltage generator. Different types of clamping clipping circuits eliminating these distortions can be used to improve the shape of the expanded pulse. Chapter VI contains a description of clipping circuit operation.

A voltage that changes linearly, and produced by the sawtooth voltage generator, is used to deflect the cathode ray tube beam from the center to the edge of the screen.

Generators with positive feedback (see Chapter VI) have been widely used in practical circuits.

Special features of a PPI with a fixed deflection system. In order to obtain a radial-circular sweep with a fixed deflection system, it is necessary to create a magnetic field that changes linearly and rotates in synchronism with antenna rotation in the throat of the tube.

A fixed deflection system consists of two pairs of coils positioned mutually perpendicular with respect to the cathode ray tube axis. A sawtooth current, modulated in accordance with the antenna rotation law, and displaced with respect to the phase in these coils by 90°, flows through these coils. The currents in the deflection coils are

$$i_1 = f(\tau) \sin \Omega_A t,$$
$$i_2 = f(\tau) \cos \Omega_A t,$$

where $f(\tau)$ is the law in accordance with which the sawtooth current changes;
Ω_A is the angular velocity of antenna rotation;
$\Omega_A t$ is the current azimuth of the antenna.

Magnetic fluxes, changing in accordance with this same law, occur as a result of the effect of currents i_1 and i_2,

$$\Phi_1 = zf(\tau) \sin \Omega_A t,$$
$$\Phi_2 = zf(\tau) \cos \Omega_A t,$$

where z is a coefficient fixing the relation between the current in the coil and the magnetic flux.

The resultant field, which deflects the electron beam, acts in the tube throat

$$\Phi_{res} = \sqrt{\Phi_1^2 + \Phi_2^2} = zf(\tau). \tag{9.8}$$

The position of the resultant field vector can be calculated as follows

$$\text{arc tg} \left(\frac{\sin \Omega_A t}{\cos \Omega_A t}\right) = \Omega_A t = \beta \tag{9.9}$$

that is, it corresponds to the present azimuth. The resultant field therefore rotates in synchronism with antenna rotation. Consequently, formation of a rotating sweep requires the generation of a sawtooth voltage and modulation

of that voltage by the sine and cosine law for the antenna rotation angle. The sawtooth voltage is generated by the generator in the range sweep channel. This voltage is then split into two components, 90° apart. Each of these components is modulated by the antenna rotation law.

Most often used for splitting are sine-cosine rotating transformers and sine-cosine potentiometers. They are mechanically reliable and can be used to load the deflection coils directly.

The block schematic of a PPI with a fixed deflection system is shown in Figure 9.5. The only difference between it and that for the PPI with a rotating deflection system is in the azimuthal sweep rotation channel.

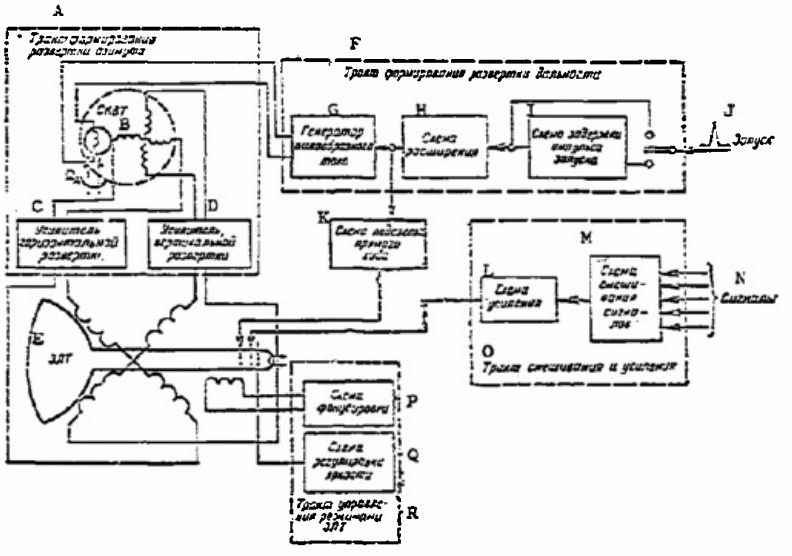


Figure 9.5. Block schematic of a PPI with a fixed deflection system. A - azimuth sweep generation channel; B - sine-cosine rotatable transformer; C - horizontal sweep amplifier; D - vertical sweep amplifier; E - cathode ray tube; F - range sweep generation channel; G - sawtooth current generator; H - expansion circuit; I - trigger pulse delay circuit; J - trigger; K - outgoing trace intensifier circuit; L - amplifier circuit; M - signal mixing circuit; N - signals; O - mixing and amplifier channel; P - focusing circuit; Q - brilliance control circuit; R - cathode ray tube control channel.

The sine-cosine rotatable transformer in this channel has a rotor winding and two stator windings located perpendicular to each other. The rotor is mechanically coupled to the antenna shaft and rotates in synchronism with it.

The rotor winding is the sawtooth current amplifier load in the range sweep generation channel.

The magnetic flux around the rotor winding, the result of the flow of sawtooth current, induces emfs in the stator windings, the magnitudes of which depend on the mutual positioning of the rotor and stator windings.

It is highly important that there be a fixed central point, corresponding to the origin of the sweep, in scopes with fixed deflection systems because positive and negative sawtooth voltages appear across the amplifier tube grids. Controlled level clampers, designed to restore the original potential, are used for this purpose.

PPIs for combining situational information. The data received from several radars can be combined on the screen of one PPI.

A single system of synchronous and cophasal antenna rotation, and a single trigger for the transmitting and indicating devices, can be used to combine the information.

Figure 9.6 is the block schematic of a PPI for combining situational information from two unsynchronized radars.

As will be seen from the schematic, a cathode ray tube and a fixed deflection system, are used in the indicator.

The indicator utilizes the time division of signals method wherein the sweep voltage relative to the X and Y axes, corresponding to the position of the antenna associated with radar No. 1, is fed to the deflection coil during one time interval, with the sweep voltage from radar No. 2 fed to, and the sweep voltage from radar No. 1 disconnected from, the deflection coil during the succeeding time interval. The result is the appearance of two sweeps on the indicator screen, the positions of which correspond to the azimuthal positions of the antennas for radars No. 1 and No. 2 (fig. 9.7).

The echo signals from the radars are switched simultaneously with the sweeps, and at the moment the sweep from radar No. 1 is generated, the echo signals from that radar only are fed to the cathode ray tube modulator. Echo signals from radar No. 2 are cut out at this time.

The functions of signal time division are carried out by a special circuit.

If the radars are quite a distance apart an additional circuit for taking the base into consideration is used, and this circuit superimposes the origin of the sweep at a point on the screen corresponding to the position of the radar on the terrain.

When the radars have identical repetition frequencies, or very nearly so, the sweep trigger frequency of one of the radars is reduced when information is combined on one indicator.

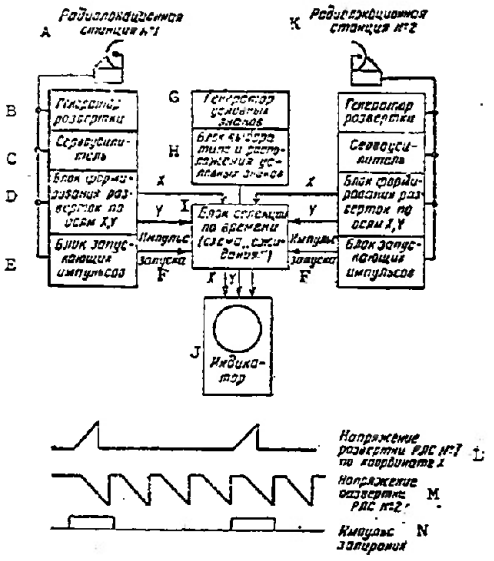


Figure 9.6. Block schematic of a PPI for combining situational information. A - Radar No. 1; B - sweep generator; C - servo-amplifier; D - unit for generating the X and Y axis sweeps; E - trigger pulse unit; F - trigger pulse; G - conventional symbol generator; H - unit for selecting and positioning conventional symbols; I - time selection unit ("wait" circuit); J - indicator; K - Radar No. 2; L - Radar No. 1 sweep voltage, X coordinate; M - Radar No. 2 sweep voltage; N - blanking pulse.

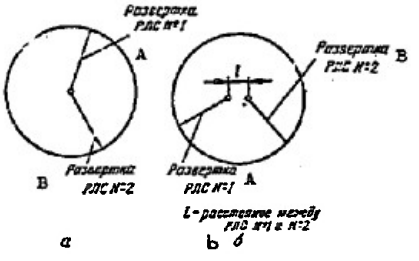


Figure 9.7. View of a screen of an indicator showing the combined situation (a) and the screen of an indicator when the information from two remote radars is combined (b). A - Radar No. 1 sweep; B - Radar No. 2 sweep; l is the distance between radars no. 1 and no. 2.

9.3. The Azimuth-Range Sector Display Indicator

These indicators are used:

to improve azimuth and range resolution;

to increase the accuracy with which coordinates are read;

to find the height of a target with a radar producing a V-beam.

Expanding the sweep scales in the azimuth-range indicator is unavoidably connected with decreasing the scan zone. A view of the screen of an indicator of this type is shown in Figure 9.8. The reflected signals are reproduced on the indicator screen as vertical, or horizontal, lines; the scale markers in the form of rows of vertical and horizontal lines.

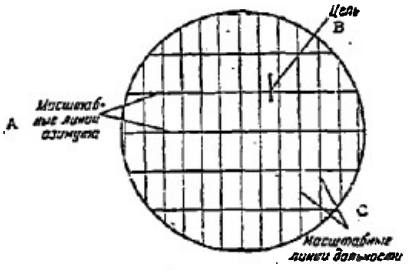


Figure 9.8. View of an azimuth-range indicator screen; A - azimuth scale lines; B - target; C - range scale lines.

Usually used to obtain an image on the screen of a cathode ray tube is fast range sweep (horizontal or line scan), and slow azimuth sweep (vertical or frame scan), controlled by the antenna. The deflection system in these indicators therefore consists of two mutually perpendicular, fixed deflection coils, one creating the range sweep, the other the azimuth sweep.

The vertical deflection coil is supplied with current from the range sweep channel, similar to the range sweep channel in the PPI with a rotating deflection coil.

The horizontal deflection coil is supplied with current from the azimuth sweep channel, changing in proportion to the angle, or to the sine of the angle, of rotation of the antenna.

A typical azimuth-range indicator consists of the following main elements:

a cathode ray tube with a fixed deflection system;

a range sweep channel;

an azimuth sweep channel;

an amplification channel;

circuits for supplying and controlling tube operation.

Figure 9.9 is a functional diagram of a typical indicator.

The azimuth sweep channel is designed to develop the current, the magnitude of which changes with the angle of rotation of the antenna in azimuth, that is

$$i = k\beta \quad \text{or} \quad i = k \sin \beta$$

where

k is a constant;

β is the angle of rotation of the antenna in azimuth.

The channel consists of two main elements; the modulator, and the deflection system supply circuit.

Azimuth sweep channel using an azimuth pulse transmitter. The block schematic of a sweep channel using an azimuth pulse transmitter and the voltage curves, are shown in Figure 9.10.

A special device (a transmitter) generates short pulses that are related to the angular position of the antenna as the antenna rotates in azimuth.

The pulses from the transmitter output are fed into the gating circuit, which also receives the intensification pulse voltage for the sector selected. Thus, a "bundle" of azimuth pulses is formed at the output of the gating circuit. The width of the "bundle" is fixed by the width of the azimuth intensification pulse and fixes the magnitude of the working sector in azimuth.

The pulses obtained at the gating circuit output are then fed into the counter and then into the circuit for converting the code into voltage. The result is to generate a stepped voltage which is then smoothed out by a filter.

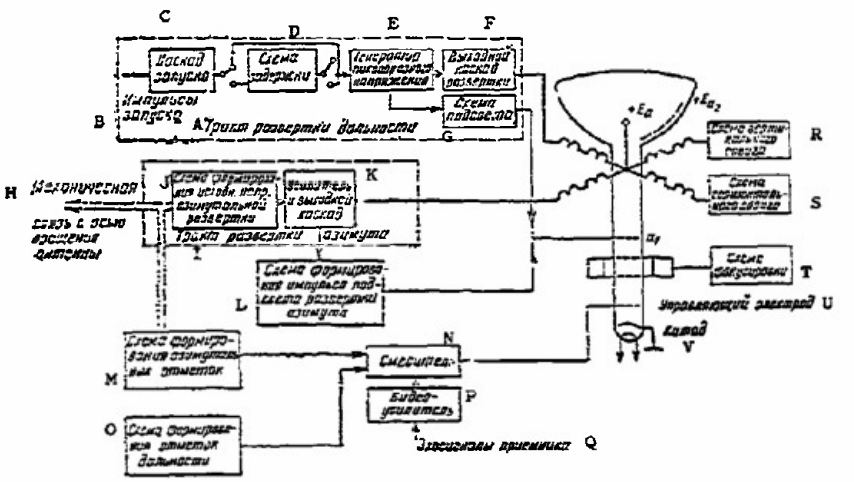


Figure 9.9. Block schematic of an azimuth-range indicator.

Figure 9.9. Block schematic of an azimuth-range indicator. A - range sweep channel; Б - trigger pulses; C - trigger stage; D - delay circuit; E - sawtooth voltage generator; F - sweep output stage; G - intensification circuit; H - mechanical coupling with shaft for rotating antenna; I - azimuth sweep channel; J - circuit for generating the initial azimuth sweep voltage; K - amplifier and output stage; L - circuit for generating azimuth sweep intensification pulse; M - circuit for generating azimuth markers; N - mixer; O - circuit for generating range markers; P - video amplifier; Q - echo signals from receiver; R - vertical displacement circuit; S - horizontal displacement circuit; T - focus circuit; U - control electrode; V - cathode.

A "north" pulse can also be generated at the instant the antenna passes through the north point, thus providing for correct selection of the operating sector by the transmitter.

The azimuth sweep channel with selsyn modulators. The most widely used circuit for generating azimuth sweep is that using a selsyn couple (see fig. 9.11).

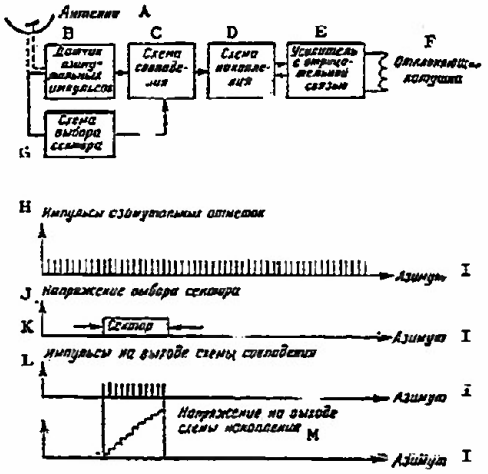


Figure 9.10. Block schematic and azimuth sweep voltage curves when an azimuth pulse transmitter is used. A - antenna; B - azimuth pulse transmitter; C - gating circuit; D - storage circuit; E - amplifier with negative coupling; F - deflection coil; G - sector selection circuit; H - azimuth marker pulses; I - azimuth; J - sector selection voltage; K - sector; L - pulses at the gating circuit putput; M - voltage across the storage circuit output.

The variable voltage used for selsyn supply is usually called the carrier frequency voltage, and the selsyn that modulates this voltage in amplitude is called the modulator.

The carrier frequency should be a great deal higher than that of selsyn rotor rotation in order to simplify the process of detection and voltage envelope discrimination.

A selsyn transmitter can be used to modulate the carrier frequency in accordance with the antenna rotation law. The rotor winding of the selsyn transmitter rotates in synchronism with the antenna. A voltage is induced in the stator winding, the load on which is the three-phase winding of the stator of the selsyn transformer being used. The amplitude of the voltage

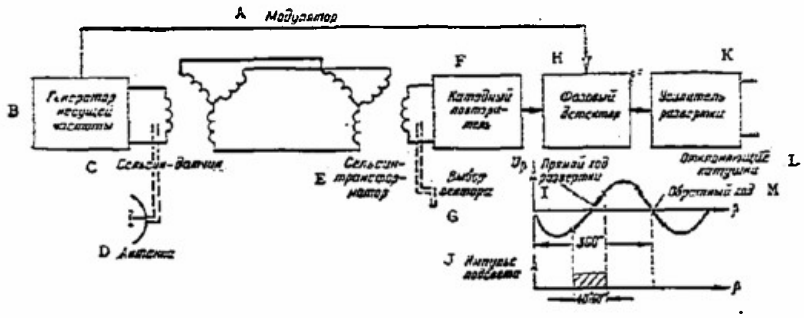


Figure 9.11. Block schematic of an azimuth sweep channel with a selsyn modulator. A - modulator; B - carrier frequency generator; C - selsyn transmitter; D - antenna; E - selsyn transformer; F - cathode follower; G - sector select; H - phase detector; I - outgoing sweep trace; J - intensification pulse; K - sweep amplifier; L - deflection coils; M - return trace.

across the rotor winding of the selsyn transformer depends on the mutual positioning of the rotors of the selsyn transmitter and the selsyn transformer. Since the selsyn transmitter rotor rotates at antenna rotation frequency, a voltage, amplitude modulated by the antenna rotation frequency, is induced in the selsyn transformers braked winding.

Zero voltage amplitude means that the positions of the selsyn rotors match. By rotating the selsyn transformer rotor we can change the moment zero is established for the output voltage envelope. At the same time the zero amplitude for this voltage will correspond to the different angular positions of the antenna in space.

The phase detector in the azimuth sweep channel is used to make the rectified voltage proportional to the sine of the antenna rotation angle. When a conventional detector is used this dependence could only be arrived at within 180° limits. The linear section of this voltage in the 20 to 60° range of antenna rotation in azimuth (fig. 9.11) is used to generate azimuth sweep.

Use of the cathode follower is the result of a need to generate a high input resistance in the azimuth sweep channel, because it is customary in radar practice to connect several selsyn transformers to one azimuth sweep voltage selsyn transmitter.

A powerful output stage is needed to convert the voltage of the sinusoidal envelope obtained across the phase detector output into the current for deflecting the tube beam.

Switching the video amplifier input by intensifying only the outgoing sweep trace can be resorted to in order to eliminate the superimposition of images of the outgoing and return azimuth sweep traces.

9.4 The elevation-position type indicator for measuring altitude

The elevation-position type indicator is included among the indicators with rectangular sweep, and in design is little different from the azimuth-range indicator.

The horizontal sweep current in this indicator is proportional to the range, and the vertical sweep current is proportional to the angular position of the beam in the vertical plane.

Those radars with a V-beam radiation pattern have the vertical sweep current in the selected sector proportional to the angle to which the antenna is turned in azimuth.

The elevation scanning voltage in the radar with a "rocking horse" beam in the vertical plane can be generated by devices that cause the beam to wobble.

Lines of different heights, H1, H2, H3, etc., are constructed using the range and elevation coordinates, in order to read target height. These equal height lines, as well as the range and elevation lines, are plotted on a transparent light filter and placed atop the tube screen.

When a scale such as this is used the determination of target height can only be made after a careful matching of the electrical scale lines for range and elevation with similar lines of the plotted scale. The accuracy in determining the height depends on the accuracy with which the match is made.

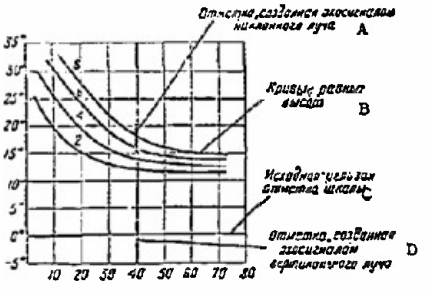


Figure 9.12. View of a height indicator with scales for the different heights and the markers for a radar with a V-beam. A - marker generated by the echo signal from the tilted beam; B - curves of equal heights; C - initial angular scale marker; D - marker generated by the echo signal from the vertical beam.

Lines of equal height are in the form of hyperbola in the elevation-position indicators (or angle of rotation-range for V-beam radars), and the constant current lines are in the form of horizontals (fig. 9.12).

There are two basic shortcomings in these indicators; at long ranges the equal height lines converge and a considerable error in reading the height results, and much of the useful area of the screen is wasted.

Two scales, range and elevation, are often used in order to eliminate these shortcomings.

As was indicated above, the vertical sweep voltage for the V-beam radar is proportional to the angle of rotation of the antenna in azimuth, not to the angle of elevation of the antenna.

Since echo signal pips can appear at different places on the screen, it is necessary to match the original zero line of the angle of the mechanical scale with respect to which the grid of antenna rotation angles was constructed with the echo signal pip from the vertical beam in order to determine target height. The height reading is taken from the position of the echo signal pip from the incident beam relative to the lines of equal height.

The zero line on the scale can be matched with the echo signal pip from the vertical beam in one of the following ways:

by correcting the position of the sector selection setting; that is, changing the position of the signal pip from the vertical beam with respect to the lines on the fixed mechanical scale, a procedure complicating the

work of the operator and resulting in substantial measurement errors;

by shifting the mechanical scale until its zero line matches the center of the trace of the signal from the vertical beam.

This method permits determination of target height on the second antenna revolution.

But it is only possible to shift the scale when the vertical sweep voltage can be changed linearly from the angle of rotation of the antenna.

The most effective way in which to establish a vertical linear sweep based on the angle of rotation of the antenna is to use the azimuth sweep circuit with pulse accumulation reviewed above.

The method that utilizes summing of the sine voltages can be used in the circuit with selsyns to obtain a linear azimuth sweep. For example, it is possible to obtain an output voltage from the two selsyns in Figure 9.13 that will be proportional to the angle of antenna rotation.

One of the selsyns (the primary) can be rotated in proportion to the angle of rotation of the antenna in azimuth Θ, while the other (the linearizing selsyn) -2Θ, with both selsyns fed from the same carrier frequency generator.

If 1/8th of the output voltage from one selsyn is subtracted from the output voltage of the other the summed voltage will be a linear function of the angle Θ, and will be accurate to within 1°.

In the case of mechanical scale shifting it is convenient to use a scale of equal height, projected optically. This makes it easy to shift from one scale to another in the event the indicator uses two scales and also reduces the parallax error.

Figure 9.14 is the schematic diagram of an elementary optical device with one mirror.

The operator observes the tube beyond the screen through a mirror with partial reflection. Mirror 1 is set up at an angle of 45° to the tube axis. Scale 5, which is engraved on a plastic plate, is end lighted.

The plate is set up perpendicular to the tube screen and at some distance from the mirror such that the operator sees the image of the scale on the surface of the screen.

When the switch is made to the other scale, the lighting for scale 6 is cut in, and that for scale 5 is cut out.

Lighting can be switched by the same knob that switches scales. When the plates are moved the image of the scale on the indicator screen is moved.

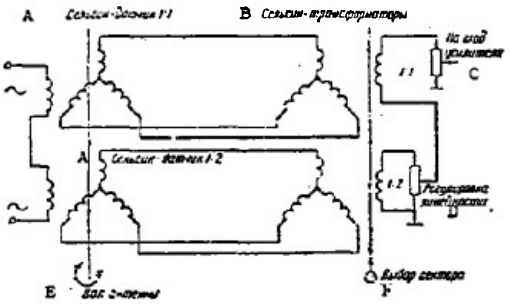


Figure 9.13. Schematic diagram of linearization of the azimuth voltage. A - selsyn transmitter; B - selsyn transformer; C - to amplifier input; D - linearity adjustment; E - antenna shaft; F - sector select.

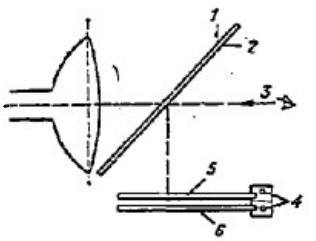


Figure 9.14. Schematic diagram of an elementary optical device with one mirror. 1, 2 - mirror; 3 - observer; 4 - scale lighting; 5, 6 - scales.

9.5 The Range-Height Type Indicator for Measuring Height

As distinguished from the foregoing, in the range-height indicator the scale of the elevation sweep expands with increase in the range and thus simplifies reading the height.

In order to create sweep in height-range coordinates in an indicator with fixed deflection coils, it is necessary to feed current into the horizontal deflection coil, and the magnitude of this current is proportional to the horizontal range to the target:

$$i_{hor} = k_1 R_{hor} = k_2 t \cos \epsilon$$

where

k_1 is a coefficient fixing the sensitivity of the deflection system;

$k_2 = k_1 \frac{c}{2}$;

c is the rate at which the electromagnetic energy is propagated;

ϵ is the target elevation.

At the same time, the vertical deflection coil must be supplied with current, the magnitude of which is proportional to target height

$$i_{vert} = k_3 H = k_4 t \cdot \sin \epsilon$$

where

$$k_4 = k_3 \frac{c}{2} .$$

The diagrams of the currents flowing in the deflection coils for this case are shown in Figure 9.15a, and the lines of equal height are depicted on the screen by a family of horizontal lines.

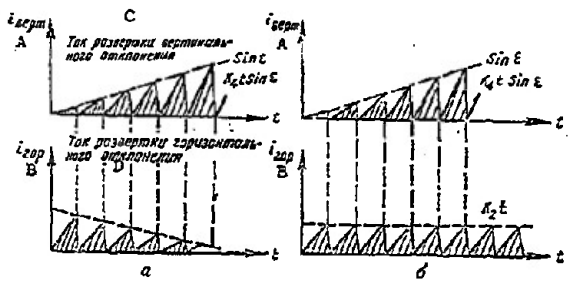


Figure 9.15. Diagrams of currents flowing in the deflection coils. A - vertical; B - horizontal; C - vertical deflection sweep current; D - horizontal deflection sweep current.

If a small sector of the elevation is depicted on the indicator screen, time t is often used as the horizontal coordinate (sweep current is proportional to slant range), and $\sin \epsilon \approx \epsilon$ is used as the vertical coordinate. The reading of range data can be simplified in this case, but the lines of equal height bend upward somewhat with an increase in angle ϵ.

Curvature in the lines of equal height can be eliminated if the horizontal deflection coil is fed a current proportional to the range, and the vertical deflection coil is fed a current proportional to the height (fig. 9.15b), that is

$$i_{hor} = k_1 R = k_2 t,$$

$$i_{vert} = k_3 H = k_4 t \sin \epsilon$$

A sawtooth voltage across the input to the output stage will result in a sweep current proportional to the range.

In order to obtain the sweep current to feed to the vertically deflected coil, it is necessary to form a sawtooth voltage and multiply it by the sine of the angle of rotation of the antenna in the vertical plane.

The earth's curvature, and the refraction of radio waves, must be taken into consideration in determining the height relative to the earth's surface at long ranges. The formula for so doing is in the form

$$H = R \sin \epsilon + \frac{R^2}{2r_e}$$

where

r_e is the effective radius of the earth.

In this case the sweep current flowing in the deflection coils must have been changed in accordance with the following law

$$i_{hor} = k_2 t, \quad i_{vert} = k_4 t \sin \epsilon + k_5 t^2,$$

that is, the current flowing in the vertical deflection coils must have an additional parabolic component.

Integration is one of the simplest methods to use to obtain a parabolic-shaped voltage. Integration changes the sawtooth into a parabolic voltage. Chapter VI discusses integration circuits.

In range-height indicators, the sweep lines go off the tube screen at high elevations and heavy sweep current is required to create them. During sweep flyback the high rate of decay can cause excessive overloading and damage to deflection coil insulation. Hence, rise in current at high elevations must be eliminated in indicators. This is done by using circuits limiting amplitude sweep voltage amplitude or by using circuits automatically changing the sweep voltage duration.

The level of limitation is fixed by the moment in time the beam reaches the upper edge of the scale on the indicator screen.

9.6 Indicators for Semiautomatic Pickoff

Semiautomatic information pickoff, automatic transmission, and reproduction of target pips is the procedure used to reduce the time required to transmit and reproduce measured data. Semiautomatic pickoff can be accomplished with conventional radar indicators of various types, but the indicator must be equipped with a special optical-mechanical pickoff, as well as with special indicators for semiautomatic pickoff, using the electronic-optical pickoff method. This method involves the reproduction on the indicator screen of a spot of light which the operator mechanically matches with the target pip, thus "pinning down" the coordinates. The most widely used practical method is the electronic-optical one, the essence of which involves reproducing an electronic marker on the screen, the position of which is fixed by the values of the constant components of the voltage picked off the arms of the potentiometers in the pickoff device. The

electronic-optical method is realized in special indicators, the principle of construction of which differs substantially from that discussed in the foregoing. In these indicators the glow of the primary radar situation is superimposed on the secondary signals, which later can be target pips, markers, symbols, figures, and the like. Constant voltages must be used to reproduce the secondary signals. One fixed deflection system will reproduce the primary radar situation and the secondary signals. The block schematic of the apparatus for reproducing the primary situation is no different in principle from the PPI with the fixed deflection system discussed in the foregoing, one in which the situation is reproduced in polar coordinates. Special equipment is usually needed in the indicator because the pickoff is usually in rectangular coordinates. Figure 9.16 is a functional schematic included to explain the principle of the electronic-optical method. An electronic marker, the position of which is fixed by voltages U_{xM} and U_{yM}, picked off the potentiometers in the pickoff divide, R_x and R_y, is brought out on the screen. The arms of these potentiometers are moved by the operator through a mechanical transmission, and the marker moves on the screen. At the moment the marker and the target pip match the pickoff button is pressed and relay P functions. The contacts of this relay connect the outputs from the coordinate potentiometers to the storage devices, C_x and C_y, in which voltages U_{xM} and U_{yM} are clamped. These voltages can then be transmitted to the reflecting device directly, or it can be coded first and then transmitted over communication channels in code. Marker pip reproduction requires cutting off, for some period of time, the scanning sawtooth voltages fed to the deflection coils, and, at this same time, feed in the constant voltages for the marker. The frequency at which it is necessary to cause the marker to glow should be such that on the one hand the marker can be observed as a non-flickering point, and on the other, that the information lost as a result of the curtailment of the sweep be a minimum. It is considered that a marker glow frequency of 15 to 16 hertz satisfies this requirement. Curtailment of the delivery of the sawtooth voltage sweeps is carried out by a special inhibitor which can be synchronized off the common synchronization system. In addition to inhibiting sweep intensification, a marker intensification pulse must be delivered to the tube's control electrode.

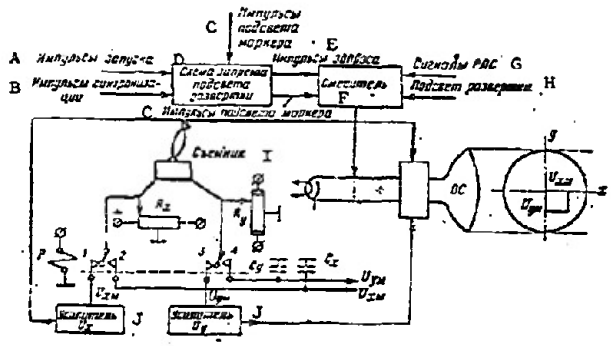


Figure 9.16. Functional schematic of the electronic-optical pickoff method.

A - trigger pulses; B - synchronization pulses; C - marker intensification pulses; D - circuit for inhibiting sweep intensification; E - interrogation pulses; F - mixer; G - radar signals; H - sweep intensification; I - pickoff; J - amplifier.

Figure 9.17 shows the voltage diagrams explaining the electronic-optical pickoff method. When the semiautomatic pickoff indicator is mated with a short-range radar, the intensification of the primary situation and the secondary signals can be carried out in one sweep period. A variety of conventional symbols and numbers can be intensified on the semiautomatic pickoff indicators, simplifying the work. Two methods are used to intensify numbers and symbols: the first is by the use of Lissajous figures, the second by the use of the small raster format method. The block schematic explaining the principle involved in the first method is shown in Figure 9.18. Basic components in the circuit are the sine voltage generator, producing a frequency oscillation, f_g, and a doubler, producing a frequency oscillation $2f_g$. The sine oscillations produced are fed into a special deflection system, and an image resembling the number 8 will appear on the screen. By intensifying the individual sections of this number we can obtain all numbers from 1 to 9, as well as several conventional symbols. The sensitizing pulses can be shaped from these sine voltages, shifted the required magnitude in phase, as well as by the sensitizing pulse shaping circuit. The small raster format method is one in which a television type raster is created at the point at which the symbol or number is reproduced, and the individual sections of this raster are intensified. This method has greater possibilities for reproducing numbers and symbols than does the method which forms from the Lissajous figures.

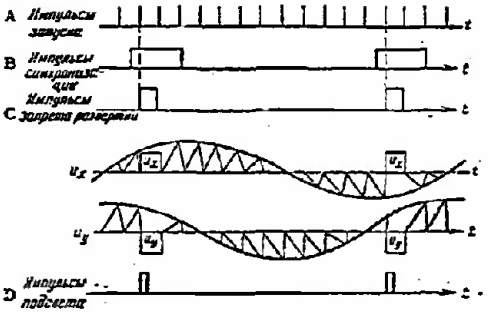


Figure 9.17. Voltage curves associated with Figure 9.16.

A - trigger pulses; B - synchronization pulses; C - sweep inhibiting pulses; D - sensitizing pulses.

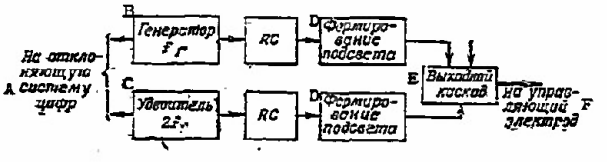


Figure 9.18. Block schematic of symbol reproduction.

A - to deflection system for numbers; B - generator, f_g; C - doubler, $2f_g$; D - intensification shaping; E - output stage; F - to control electrode.

Symbol Indicators. The semiautomatic pickoff indicator lets us combine the properties of the conventional indicator with the possibilities for reproducing special symbols and numbers. However, there are indicators which can reflect a significantly greater amount of information by using a variety of symbols. These are called symbol indicators. The basis of indicators is a symbol-printing tube, with a special matrix with symbols inside the neck. The electron beam, passing through the matrix at the required place, "prints" the symbol on the screen. Symbol indicators are most often used as terminals in electronic digital computers in control points.

9.7. Scale Marker Formation Methods

Azimuth Marker Formation Methods

Azimuth scale markers must be fixed, and rigidly synchronized with antenna rotation, if target azimuth is to be established accurately.

Each azimuth marker must be intensified by just one cycle of the outgoing trace of the radial sweep, and must always begin with the beginning of the sweep. If this requirement is not satisfied, the scale markers will appear chaotically, they will be seen to flicker, and will be defocused. Azimuth determination is difficult in this case. This can be eliminated by rigidly synchronizing the beginning of the azimuth scale markers with the beginning of the range sweep.

Marker width, t_β, must satisfy the condition

$$t_\beta \leq T_p$$

where

T_p is the pulse repetition frequency for the radar.

If this condition is not satisfied, the markers will be intensified by more than one range sweep, and this will cause some of the markers to appear more brilliantly than others.

A typical circuit for obtaining azimuth scale markers is shown in Figure 9.19. Here the transmitter is connected to the antenna shaft and produces the primary signals as the antenna passes through predetermined angles, the magnitudes of which can be fixed by the required resolution for the azimuth markers. The primary signals are triggered by the circuit for shaping the expanded pulse and producing a pulse with a width somewhat greater than the pulse repetition period for the radar. This pulse is fed into the synchronization stage, which is a coincidence circuit. Trigger pulses are fed into the other input of the synchronization stage. At the moment pulses coincide the circuit functions and feeds a pulse into the scale marker generator, where the synchronized azimuth marker is formed.

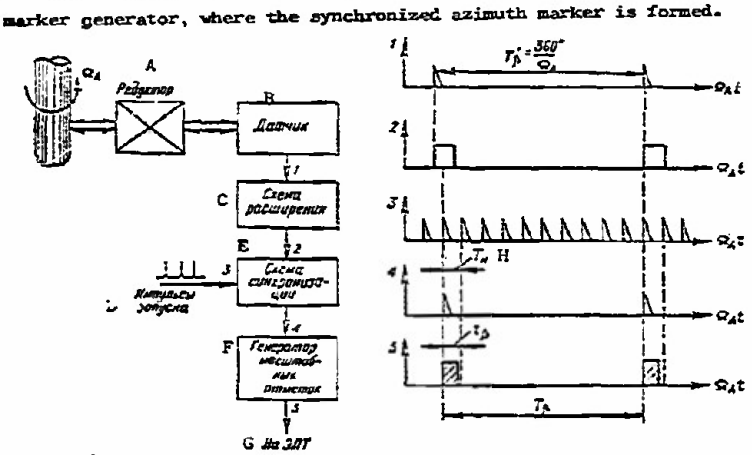


Figure 9.19. Block schematic of the formation of azimuth markers and voltage time diagrams at various points in the formation.

Figure 9.19. Block schematic of the formation of azimuth markers and voltage time diagrams at various points in the formation. A - reduction; B - transmitter; C - expansion circuit; D - trigger pulses; E - synchronization circuit; F - scale marker generator; G - to cathode ray tube; H - T_p.

The electric-mechanical method. In this method the transmitter is a cam (fig. 9.20a) connected to the antenna shaft through a reduction and closes contacts 1-2 as many times during one antenna revolution as is required to produce the azimuth markers for one antenna revolution. For example, if the 5° scale markers are to be formed, the contacts must close 72 times (360/72 = 5°). At the moment the contacts close there is a pulse across resistance R, and this pulse is also the primary signal for the subsequent formation. The advantage of this circuit is its simplicity. It does have significant shortcomings, however, including short contact life, and a broad scatter in times of occurrence of the primary signals, and this, in the final analysis, reflects on the accuracy with which the scale markers chart.

The photoelectric method. The principle involved in obtaining markers by the photoelectric method is explained in Figure 9.20b.

Basic components are the light source, the opaque disk with radial apertures, the photo electric cell, and the amplifier.

The disk is connected to the antenna shaft through a reduction. As the antenna rotates the light beam passes through the apertures, striking the photoelectric cell, and producing a pulse of current. The amplified pulse is then used as the primary signal. The pulse repetition period, T_β depends on the reduction ratio and the number of slots in the disk, and equals

$$T_\beta = \frac{360° i_r}{n_s}, \tag{9.10}$$

where

i_r is the reduction ratio;

n_s is the number of slots in the disk.

This is a more complicated circuit to manufacture, but can be used when antenna rotation rates are high.

Selsyns and the Zero Reading Method

This method is widely used in circuits for forming azimuth markers because it is simple and dependable. The basic components (fig. 9.20c) are the selsyn transmitter, the selsyn transformer, and the shaper.

The selsyn transmitter rotates with the antenna. The selsyn transmitter rotor will rotate at a rate that depends on the reduction gear transmission ratio.

The selsyn transformer, the rotor of which is fixed, is electrically connected to the selsyn transmitter. A voltage with a frequency that of the voltage supplied to the stator of the selsyn transmitter, but modulated by the amplitude of antenna rotation frequency (or a multiply of it) appears across the rotor windings. This is the voltage that is applied across the shaper, which detects and forms the envelope.

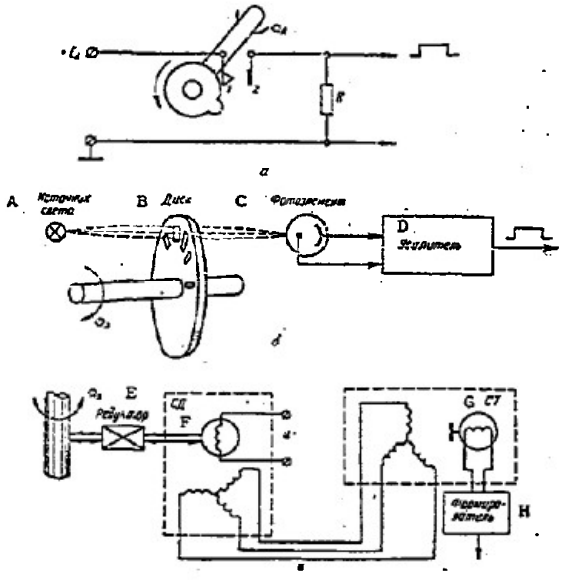


Figure 9.20. Azimuth pulse transmitters.

a - electric-mechanical; b - photoelectric; c - selsyn.

A - light source; B - disk; C - photo electric cell; D - amplifier; E - reduction; F - selsyn transmitter; G - selsyn transformer; H - shaper.

Every time the envelope thus formed passes through the zero value, a triangular pulse is formed at the shaper output, and the width of this pulse can be regulated by changing the shaper clipping level (fig. 9.21). These pulses are also used as the original pulses for further shaping of the azimuth markers.

The reduction transmission ratio, i_r, can be found from antenna rotation rate and marker graduations through the formula

$$i_r = \frac{2\Delta\beta^\circ}{360^\circ}, \tag{9.11}$$

If the scale is to have two, or more, graduations, there must be several channels, differing only in the i_r ratio.

Duplex synchronization is required for the formation of azimuth scale pulses. First, the markers for the lowest graduation must be synchronized with the trigger pulses, and, second, the marker for the highest graduation must be synchronized with the markers for the lowest graduation.

Range Marker Formation Method

For convenience in reading the coordinates, the range scale markers should be shaped in the form of short electrical pulses, the amplitudes of which differ according to the marker graduations. Ordinarily, the amplitude of the marker used with the highest graduation exceeds that of the marker used with the lowest graduation by 30 to 50%.

Range scale pulses should be formed while the outgoing trace of range sweep lasts, and the accuracy with which they are plotted should remain constant over the entire range scale.

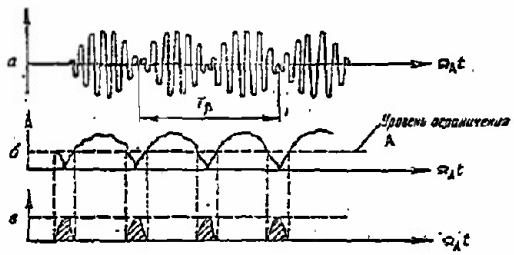


Figure 9.21. Shaper operating principles.
A - clipping level.

A finely controlled crystal oscillator, which functions continuously (fig. 9.21) is used as the marker oscillator. The crystal oscillator frequency is divided into stages which make the division into magnitudes corresponding to the repetition period for the markers, and this is done in order to obtain the synchronization pulses.

In this method the range scale markers are formed from the IF voltages generated during crystal oscillator frequency division. In this case, the crystal oscillator frequency must not be lower than the repetition frequency for the scale markers on the lowest graduation, that is,

$$f_{co} = n\frac{c}{2t_{res}}, \qquad (9.12)$$

where

$n = 1, 2, 3 \ldots;$

t_{res} is the specified resolution for the scale markers.

The principle involved in division of the repetition frequency for the scale markers for the lowest graduation to the required magnitude is used to obtain the scale markers for the highest graduation. For example, if the markers for the lowest graduation are 10-km markers the recurrence frequency of 15 kHz must be reduced by a factor of 5.

9.8. Three-Dimensional Indicators

There are two basic methods that can be used to obtain three-dimensional indications:

the use of the stereo effect which creates visibility of three-dimensional images;

building three-dimensional indicators using optical and mechanical equipment.

The first method utilizes the principle of a special optical system superimposing two flat images to obtain a three-dimensional image. Figure 9.22 shows a schematic diagram explaining the functioning of an indicator of this type.

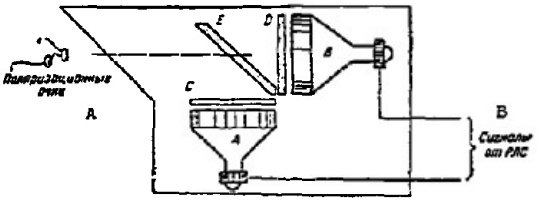
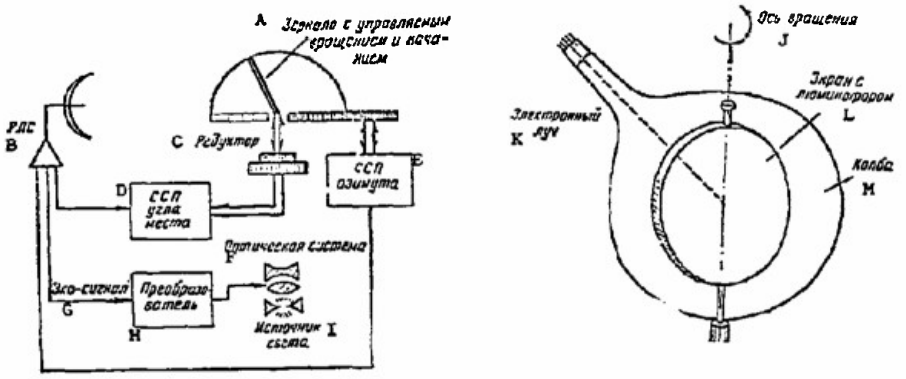


Figure 9.22. Three-dimensional indicator using stereo effect. A - polarized glasses; B - signals from radar.

Two cathode ray tubes, positioned at an angle of 90° to each other are used. The images on the two tubes are displaced somewhat relative to each other. Polarized filters, C and D, are placed in front of each of the screens. The image can be superimposed by a semi-transparent mirror, E, which is half silvered. Polarized glasses are used for observations, with the left eye seeing screen B, the right eye screen A.



a b

Figure 9.23. Three-dimensional indicator with tilting (a) and rotating (b) mirrors.

A - mirror with controlled rotatioo and tilt; B - radar; C - reduction; D - elevation synchronous tracking transmitter; E - azimuth synchronous tracking transmitter; F - optical system; G - echo signal; H - converter; I - light source; J - axis of rotation; K - electron beam; L - phosphor screen; M - tube envelope.

Optical parallax is introduced in each of the images to reproduce the third dimension (depth).

A voltage, proportional to the sine of the elevation is added to the deflection system in indicator A to achieve this. This same voltage is subtracted from the indicator B sweep voltage. Advantages of indicators such as these include simplicity of design and use of conventional indicators; a shortcoming is the narrow observation angle.

Tilting, or rotating screens or mirrors, are used in the second of the methods. The image is projected onto them by an optical system. The optics most often create a brilliant marker, which is the reflection of the object in space.

The screen is lighted up at that moment in time when the angle of rotation of the screen corresponds to the target elevation.

The mirror is connected to the radar antenna by a synchronous tracking system, and repeats radar beam movement. The two other coordinates lie in the plane of the screen.

Figure 9.23 illustrates the operating principle of these indicators.

What makes these indicators complex is getting the mirror to revolve inside the cathode ray tube, and at a high enough speed.

A good effect can be obtained by placing a screen covered with a fluorescing composition inside the cathode ray tube (see fig. 9.23). Here

the electron beam is deflected in accordance with the plane coordinates of the target, and the screen repeats antenna rotation with respect to elevation. A phosphorescent marker forms when the beam hits the screen.

Indicators such as these can only operate off radars using beam scanning in the vertical plane.

9.9. Color Indicators

Marker discrimination in the known types of receivers is brought about by using one of the properties of the human eye - that of sensitivity to change in brilliance. While the eye can distinguish no more than four to six gradations of brilliance, the radar signal can have a very great many more such gradations. So conditions are created wherein not all the details of the situation brought out on the indicator can be perceived by man, or, putting it another way, not all the potentials of the radar are used to the maximum. Press reports suggest that one way to increase observation efficiency is to use color indicators, so the human eye is reacting to change in color, as well as to change in brilliance. This improves signal visibility and increases the effectiveness of operation under interference conditions. The essence of color indication is that signals differing in level (amplitude) are reproduced on the screen in different colors. If two colors are used, the indicators are called two-color indicators, if more, they are called multicolor indicators.

A special two-color tube, or two separate, single-color tubes, with subsequent optical superimposition of the image, are required to reproduce signals of different levels in the form of two colors. As distinguished from conventional radars, the amplitude curve for the receiver-amplifier channel in the radar with color indication must have a section in the form of two intersecting straight lines (rising and falling). Therefore, individual channels with predetermined amplitude curves for each color are built to operate an indication system such as this and the functions of separating the signals of different levels and converting them into brilliant markers of different colors are combined. The amplitude curves for the channels are selected so weakly reflected targets will glow in one color, and strongly reflected ones will glow in another, or in an intermediate, color. Experiments have proven that the human eye is most sensitive to green and red, so these two colors are the ones most widely used in color indication systems. A color indication system can also be built by selecting complementary colors, that is, mixing them in a predetermined proportion; to yield a definite color. An indication system such as this is called a

complementary color system. Let us, briefly, consider the circuitry used to design a radar receiver-indicator channel for a two-color indication system.

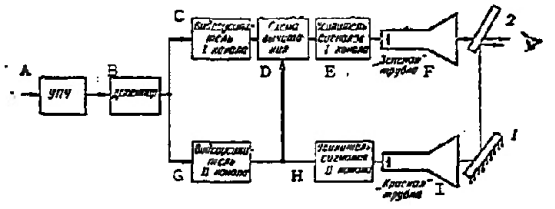


Figure 9.24. Block schematic of two tube color indication.

A - IF amplifier; B - detector; C - Channel I video amplifier; D - subtraction circuit; E - Channel I signal amplifier; F - "green tube"; G - Channel II video-amplifier; H - Channel II signal amplifier; I - "red tube".

Figure 9.24 shows one way in which a receiver-indicator channel with two different color tubes can be designed. In this circuit the output from the channel II video amplifier is connected into the subtraction circuit, thus providing amplitude curves of required shape for both channels. Signals from Channel II are subtracted from signals in channel I as the signals flow through the subtraction circuit. The signals thus formed control the brilliance of the tubes with red and green persistence. Mirrors 1 and 2 are used to make an optical match of the two-color image obtained. It should be noted that there are definite technical difficulties involved in designing this system, primarily those involved with the bulk of the apparatus. This is why a circuit with one two-color tube and two control beams is sometimes used. Color indicators with a three-color persistent kinescope (red, green, blue) can be designed on approximately this same principle.

9.10. Lighting Engineering Units and Elements of the Physiology of Vision.

The capability of an indicator to reproduce a situation, and the capacity of an observer to evaluate it, are used in making visual observations of radar signals. The images reproduced on the indicators, as well as the observation conditions, should be such that maximum operating efficiency is ensured.

The human eye is most sensitive to light oscillations with a wave length of $\lambda_0 = 0.556$ micron, a wave length corresponding to that of the color green. The sensitivity of the eye to other colors can be evaluated by a parameter called the relative luminous efficiency and showing the percentage of the

power of light with wave length λ_0 that can constitute the power of light with wave length λ. For example, the relative luminous efficiency of white (day) is 0.14. Special lighting engineering terms and units are used to make a quantitative evaluation of the result of the effect of light on the human eye.

Luminous flux. The unit used is the equivalent luminous flux with wave length λ_0 = 0.556 micron, and called the light watt. The fluxes of light with other wave lengths of the same power will differ in accordance with the relative visibility factor. In practice, units smaller than the light watt by a factor of 621, and called the lumen (lm), are used to measure luminous flux. An electric lamp rated at 55 watts yields a luminous flux of about 650 lumens.

Luminous intensity. Characterizes the angular density of the luminous flux, that is, it defines the magnitude of the luminous flux radiated in a given direction.

The unit of luminous intensity adopted is the candle, corresponding to a source uniformly radiating a luminous flux of 1 lumen at a solid angle equal to a steradian.

Illumination. Defined as the magnitude of the luminous flux incident per unit area. The unit of illumination is 1 lx (lux), and is the illumination of a section of a surface with an area of 1 m^2 covered by a luminous flux of 1 lumen.

Examples of illumination (in lux) are:

minimum illumination needed for reading	20-30
illumination in a lighted room in the daytime	100
illumination in an open area in cloudy weather	1,000
illumination in the summer time under the direct rays of the sun	100,000

Luminance characterizes the source of radiation and equals the luminous intensity arriving per unit area of the projection of the luminous surface on a plane perpendicular to the given direction. The unit of luminance is the nit (nt), that is, the luminance of a surface yielding a luminous intensity of 1 candela per square meter in a direction perpendicular to it. Luminance is the basic magnitude to which the human eye reacts when looking at an illuminated or luminous surface.

Examples of luminance values (in nits) are:

sun at noon	$10^9 - 2 \cdot 10^9$
projector arc	$4 \cdot 10^8 - 15 \cdot 10^8$
filament of a lighting lamp	$3 \cdot 10^6 - 20 \cdot 10^6$
luminescent lamps	$5 \cdot 10^3 - 10 \cdot 10^3$

normal motion picture screen	60
image on indicator screen	10-30

Contrast characterizes the relationship of the luminance of the lightest section to the luminance of the darkest section. Good photographic contrast is 100 and higher. The radar image has infinitely less contrast.

The eye is a recording device and possesses a number of properties which must be taken into consideration when observations are being organized. The capabilities of the eye must also be represented in evaluating the observability of subjects, in evaluating accuracy in determining coordinates, the basic operations carried out by the operator during his observations.

Visual evaluation of luminance. The sensitivity of the eye to luminance can be characterized by three magnitudes of sensitivity; absolute, difference, and upper thresholds of sensitivity. The eye is not satisfactory for making accurate quantitative evaluations. It can distinguish weak luminances when the distinction between small subjects is associated with intensities, and strong luminances associated with the painful sensations of being blinded.

The absolute threshold of sensitivity defines the least luminance causing light sensation. It is determined by individual properties of the eye and the degree of its preliminary preparation (adaptation) to observation conditions. Extremely significant is the preliminary adaptation to darkness to cope with the conditions under which a radar indicator screen is observed. For the dark adapted eye the absolute threshold of sensitivity is 10^{-4} to 10^{-5} nits for a luminous surface with an area on the order of a few square centimeters at a distance of 25 to 35 cm, and at least several seconds of observation time. For these same conditions the absolute threshold of sensitivity for the daylight adapted eye is 0.3 nit.

The difference threshold of sensitivity establishes the least difference between two different luminances being compared by the eye. The upper threshold of sensitivity establishes the greatest luminance of a luminous field the eye can withstand without being blinded.

Visual evaluation of observation results. The evaluation of dimensions is determined by visual acuity, that is, the capacity of the eye to distinguish small details. The quantitative sensitivity of the eye to the difference in details can be characterized by three thresholds; non-separated vision, split vision, and recognition of shape. The threshold of non-separated vision is defined as the least angle at which the image detail becomes visible as a non-separated spot.

The threshold of split vision is defined as the least angle between two adjacent objects under observation at which they are still seen separately.

For most people this angle is 35 to 40" (about 0.04 to 0.05 mm). The threshold of split vision depends on the magnitude, shape, and position of the objects, as well as on their luminance, color, and observation conditions.

The threshold of recognition of shape is defined as the least angle characterizing definite dimensions of the details of an object subject to evaluation in connection with observation missions.

In radar indicators the task of visual evaluation arises in determining the position of the marker with respect to the lines on the scale, during visual determination of the center of the marker, its origin, or end, and in the visual relating of these points to the lines on the scale, in the behavior of the amplitude markers with respect to the noise path, in target resolution, and the like. Errors associated with visual evaluation of results depend on the degree of focusing, luminance, adaptation, and many other external factors.

Vision fatigue. Insufficient luminance of the image, and unevenness in lighting, lead to strain on vision and as a result to fatigue. Fatigue causes a reduction in contrast sensitivity, in visual acuity, and in stability, that is, a reduction in the efficiency with which the observer functions. It is very important that observation conditions for the operator be created (darkening of the compartment, average luminance, and the like) such that a minimum of visual strain occurs.

Chapter X

Automatic Control Systems for Electronic Devices

Automatic control systems for electronic devices are systems of automatic regulation and control. An automatic control system consists of a set of devices which control an object without direct human participation.

The problem of control is considered to be the regulation of one or several physical values characterizing the stage of the object to be controlled.

An automatic control system is a system which operates automatically to maintain a certain physical value constant or to change it according to a given or previously unknown rule.

The very simplest as well as extremely complex automatic control systems of every type, construction, and purpose have certain design principles and functional rules in common.

10.1. Operating Principle and Functional Diagram of an Automatic Control System

An automatic control system may be one of two types: open-loop or closed-loop.

The block diagram of an open-loop system is shown in Fig. 10.1. In open systems, there is no feedback from the output to the input and consequently, the control process does not depend on the results of the system operations, on how the system carries out its function.

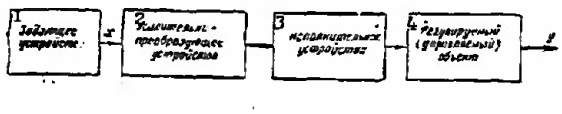


Figure 10.1. Block diagram of an open-loop system.

1 - Drive; 2 - amplifying-converting device; 3 - actuator; 4 - regulated (controlled) object.

The device used to produce the required input value x (given value) is called the drive. In the amplifying-converting device, the input signal is amplified and converted into a controlling action, which is applied to the actuator. The actuator acts on the controlled object, controlling the value y (the regulated value) which characterizes the state of the given object.

All open-loop systems have great faults, which are caused by the absence of feedback; there is no way of controlling the regulating process, and there is no way of controlling the regulating process, and there is no correction

if the controlled value does not correspond to the given value.

Closed-loop automatic control systems are more feasible. A typical automatic control system block diagram is shown in Fig. 10.2.

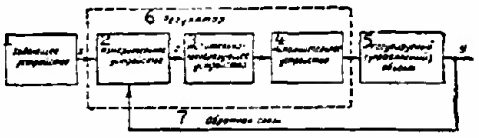


Figure 10.2. Block diagram of a closed-loop automatic control system.

1 - Drive; 2 - measuring device; 3 - amplifying-converting device; 4 - actuator; 5 - regulated (controlled) object; 6 - regulator; 7 - feedback.

In these systems, there is feedback between output and input systems. The controlled value y is applied through the feedback loop to the measuring device, where it is compared with the given value x. As a result of comparison of these two values, an error signal (mismatch) is produced, proportional to the difference z=x-y.

In the amplifying-converting device, the error signal is amplified and converted to a form suitable for applying to the actuator.

The actuator acts on the controlled object and changes the regulated value in such a way that its deviation from the given value is completely eliminated or reduced to an allowable amount.

Separate elements in some system may be expressed implicitly or may be absent, or may be combined into units. A special drive device may be lacking. In some systems, the actuator may be completely combined with the amplifying-converting device and may be called the regulator (controller). Correction signals are generated in this device to provide stability and the required indicators of the quality of the regulating process.

The control device is sometimes lumped together with the measuring device and called the regulator. Then any automatic control system consists of a regulator and regulated object. The regulator is used to hold constant or to vary according to a given rule one or several physical quantities characterizing the processes in the controlled objects, their position, or parameters of motion.

Under actual operating conditions, an automatic control system may be subject to disturbances as well as to a useful signal. Disturbances (interference) may affect the controlled object or the regulator.

A normally functioning automatic control system should meet the specifications of operational stability, control quality, and operational accuracy.

10.2. Stability of Automatic Regulation and Control Systems

The stability of an automatic control system is taken to mean the ability of the system to return to the original or a new equilibrium state after a change in the driving (controlling) or disturbing signal.

Stability depends on the system parameters and above all on gain.

The stability of an automatic control system may be judged by introducing the free component of a transient process. If the free component Y_{fr} is attenuated in the course of time, i.e., at $t \to \infty$, $Y_{fr} \to 0$, the system is stable. If Y_{fr} increases with time or shows undamped oscillations, then the system is unstable.

Figure 10.3 shows graphs of a transient process in an unstable system (a), a stable system (b), and a system on the borderline of stability (c).

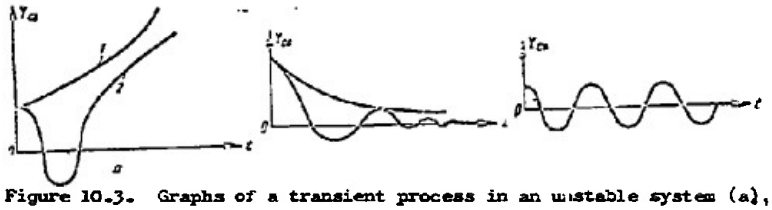


Figure 10.3. Graphs of a transient process in an unstable system (a), a stable system (b), and a system on the borderline of stability (c).

The required margin of stability is provided by choosing the system parameters and connecting correction (stabilizing) devices to deform its amplitude-phase characteristic in a suitable way. The stability margin may be estimated by the amplitude and phase stability margins.

The margin of amplitude stability (modulus) ΔN_s indicates by how many times (how many decibels) the modulus $N(\omega_\pi)$ of the amplitude-phase characteristic of an open loop system at frequency ω_π must be increased so that it passes through the point (-1, j0) (Fig. 10.4).

$$\Delta N_m = 20\lg \frac{1}{N(\omega_\pi)} \text{ [dB]}. \qquad (10.1)$$

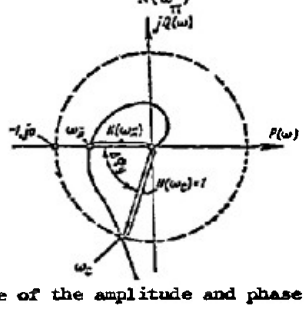


Figure 10.4. Estimate of the amplitude and phase stability margin.

The phase stability margin $\Delta\varphi_s$ is taken to mean the angle between the amplitude-phase characteristic vector with modulus $N(\omega_c) = 1$ and the negative real axis.

Acceptable values for stability margins of amplitude and phase lie within the following limits: $\Delta N_s = 10\text{-}15$ dB; $\Delta\varphi_s = 30\text{-}50°$.

10.3 The Quality of Automatic Control Systems

In a stable automatic control system, the free component of a transient process dies out with time.

A transient process is a process of change with time of the state of a dynamic system from the moment it is acted upon until a stabilized process is begun.

The quality of a system may be judged by its behavior in the transient mode in response to a single step action (Fig. 10.5) on the initially quiescent system.

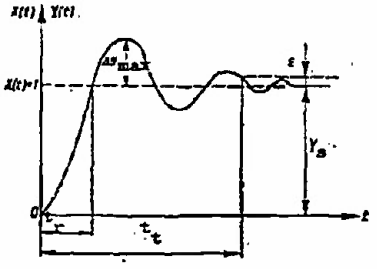


Figure 10.5. Characteristics of a transient process.

Systems are qualitatively determined by a set of characteristics of the transient process, called quality indicators.

The following direct quality indicators are usually used:

1. Regulation time (time of the transient process) t_t - the time in the course of which the output signal does not deviate more than 5-10% from the stabilized value, i.e., $\varepsilon = (0.05\text{-}0.1)Y_s$.

2. Rise time t_r - the time interval in the course of which the output attains its stable value for the first time.

3. Overcontrol $\sigma\%$ - the ratio of maximum deviation ΔY_{max} to the stabilized value in a direction opposite to the initial deviation, expressed in %:

$$\sigma\% = \frac{\Delta Y_{max}}{Y_s} 100\% \qquad (10.2)$$

Permissible values of $\sigma\%$ lie within the limits 20-40%.

4. Number of oscillations n and frequency of oscillations f of the output Y in the course of the transient process.

Usually n should be beld to 0.5-3.

The accuracy of an automatic control system operating in the presence of a useful, regular signal x(t) and a random disturbance (interference) p(t) is determined, on one hand, by the dynamic error z_d, and on the other hand, by random error z_r.

Dynamic error is taken to mean an error of the system due to inexact termination of the regular component of the input signal, and random error is an error produced by the effect of interference.

When a slowly changing signal x(t) is applied to the system, the ultimate-state dynamic error is determined by the series

$$z_d(t) = C_0 x_0(t) + C_1 \dot{x}(t) + \frac{C_2}{2!}\ddot{x}(t) + \ldots, \qquad (10.3)$$

where C_0, C_1, C_2 ... - coefficients of dynamic error;

$\dot{x}(t)$, $\ddot{x}(t)$ ... - corresponding derivatives of the input signal.

The terms of the series are usually called:

$C_0 x_0(t)$ - position error (signal error);

$C_1 \dot{x}(t)$ - velocity error;

$C_2 \ddot{x}(t)$ - acceleration error, etc.

To determine dynamic error, the error coefficients C_i and the derivatives of the input signal must be calculated.

The error coefficients C_i are determined by the transfer function of the system in the open state $K_p(p)$. The table below shows some first error coefficients for a system with various numbers of integrating elements.

1 Число интегрирующих звеньев	2 Коэффициент ошибки	3 Формула	4 Примечание
0	C_0	$\frac{1}{1+K_n}$	5 K_n, K_v, K_a — коэффициенты усиления системы в разомкнутом состоянии
1	C_0	0	
	C_1	$\frac{1}{K_v}$	
2	C_0	0	
	C_1	0	
	C_2	$\frac{2}{K_a}$	

Table 1. First error coefficients for a system with various numbers of integrating elements.

1 - Number of integrating elements; 2 - error coefficient; 3 - formula; 4 - note; 5 - amplification factors of the system in the open state.

Example: For a system with one integrating element and open state gain

K_V = 200/sec, it is required to determine dynamic error if the input signal changes with constant velocity V = 800 m/s, i.e., x(t) = Vt = 800 t.

Solution. The first derivative of x(t) = V = 800 m/sec, and the second and higher order derivatives of the input signal are equal to zero. From (10.3) and the table, dynamic error will be equal to

$$z_d(t) = C_0 x_0(t) + C_1 x(t) = \frac{V}{K_V} = \frac{800}{200} = 4\ м.$$

System error produced by the first derivative (velocity) of the input signal is usually called velocity error, and the coefficient K_V is called the velocity gain or velocity Q (dimensions of Q are 1/sec).

If the useful control signal x(t) and the disturbance p(t) represent a stationary random process, then total error Z(t) and its dynamic and random components, $Z_d(t)$ and $Z_c(t)$ respectively, will also be a stationary random process. In this case the mean square error or its root-mean-square value is taken as the measure of error.

The mean square error of an automatic control system may be calculated by the correlation function $R_z(t)$ or the spectral density $S_z(\omega)$ of error:

$$\overline{z^2(t)} = R_z(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_z(\omega)\,d\omega. \tag{10.4}$$

If the control and disturbance are applied to the system input and there is no mutual correlation between them, the mean square total error

$$\overline{z^2(t)} = \overline{z_d^2(t)} + \overline{z_c^2(t)} =$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} |1 - K_{cl}(j\omega)|^2 S_x(\omega)\,d\omega + \frac{1}{2\pi}\int_{-\infty}^{\infty} |K_{cl}(j\omega)|^2 S_p(\omega)\,d\omega, \tag{10.5}$$

where $K_{cl}(j\omega)$ - amplitude-frequency characteristic of the closed-loop system;

$S_x(\omega)$ - spectral density of the control signal;

$S_p(\omega)$ - spectral density of the disturbance signal.

The integrals of (10.5) may be calculated analytically or graphically as the area bounded by the functions under the integrals.

The mean-square error σ of an automatic control system is defined as the square root of the average of the error squared.

$$\sigma = \sqrt{\overline{z^2(t)}}. \tag{10.6}$$

Example. The transfer function of a servo system takes the form

$$K_p(p) = \frac{K_V}{p(1+Tp)}, \tag{10.7}$$

where K_V = 180 1/sec - velocity Q;

T - time constant.

At the input of the servo system there are both a useful regular signal of the form

$$\alpha(t) = \alpha_0 + \Omega t \tag{10.8}$$

(where Ω = 100 deg/sec - angular velocity) and interference, constituting white noise with spectral density

$$S_\beta(\omega) = G = 0{,}1\,\frac{\text{deg}^2}{\text{Hz}}. \tag{10.9}$$

It is required to determine the mean-square err[illegible]he system.

Solution. From the expression for the transfe[illegible]ion (10.7), it is evident that the system includes one integrating ele[illegible]t (the denominator contains p to the first degree) and therefore, according to the data in the table above,

$$C_0 = 0,\ C_1 = \frac{1}{K_V}.$$

The dynamic component of error

$$x_d = C_0\alpha_0 + C_1\Omega = \frac{\Omega}{K_V}. \tag{10.10}$$

The transfer function of the servo system in the closed state

$$K_{cl}(p) = \frac{K_p(p)}{1 + K_p(p)} = \frac{K_V}{Tp^2 + p + K_V}.$$

Substituting p = jω into the transfer function, we obtain, in accordance with (10.5), the mean square interference component of error*:

$$\overline{x_c^2} = \frac{1}{2\pi}\int_{-\infty}^{\infty} |K_{cl}(j\omega)|^2 S_\beta(\omega)\,d\omega =$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{K_V^2 G}{|T(j\omega)^2 + j\omega + K_V|^2}\,d\omega = \frac{K_V G}{2}. \tag{10.11}$$

The mean square total error

$$\overline{x^2} = \overline{x_d^2} + \overline{x_c^2} = \frac{\Omega^2}{K_V^2} + \frac{K_V G}{2}. \tag{10.12}$$

Substituting numerical values gives the root-mean-square error

$$\sigma = \sqrt{\overline{x_d^2} + \overline{x_c^2}} = \sqrt{\frac{100^2}{180^2} + \frac{180 \cdot 0{,}1}{2}} \approx 3^\circ.$$

10.4. Systems for Remote Synchronous Transmissions of Coordinates

Synchronous indicator transmissions are often used in radar stations for transmitting data on range, azimuth, and elevation of a target over great distances.

Synchronous transmission indicates a system which provides continuous match in the position of two or several shafts (which are not connected mechanically) as they rotate.

A synchronous electric transmission includes the following elements (Fig. 10.6):

---

* The method of calculating the integrals is shown in appendix II of [L.1].

detector D - a device which converts the angle of rotation of the drive shaft DS into electrical signals;

receiver R - a device which accepts the electrical signals of the sensor and transforms them into an angle of rotation of the receiving shaft RS;

connection line CL - for transmitting electrical signals from the sensor to the receiver.

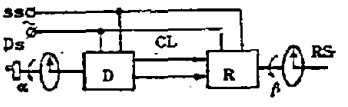


Figure 10.6. Structural diagram of a synchronous transmission.

In transmitting coordinates with synchronous transmission, the angular position α of the drive shaft determines the transmitted angle, and angular position β of the receiving shaft determines the transmitted angle reproduced at the receiving point. Thus for example, if the drive shaft is connected by a gear transmission to the lead of an antenna, then the pointer connected with the receiving shaft will indicate the azimuth or angular position of the antenna. The range may be also transmitted by connecting the gear transmission of the drive shaft to a range potentiometer.

Automatic radar systems most often use indicator synchronous transmissions. Selsyns are used as detectors and receivers in indicator synchronous transmission. Figure 10.7 shows the circuit of one variant of coupled selsyns.

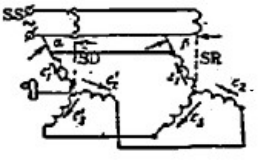


Figure 10.7. Circuit of an indicator synchronous transmission

The circuit consists of selsyn-detector SD, whose rotor-position gives the transmitted angle, and selsyn-receiver SR, whose rotor rotation reproduces the transmitted angle.

Single-phase rotor excitation windings are connected to supply source SS, and corresponding phases of the three-phase rotor windings are interconnected.

Single-phase alternating current in the stator windings sets up a pulsating magnetic excitation flux in each selsyn, which indicates emfs e_1', e_2', e_3' in the rotor phase windings of the detector, and emfs e_1, e_2, e_3 in the receiver phase windings. If the transmitter and receiver rotors are positioned identically with respect to the excitation flux and the selsyns are identical, equal emfs will be induced in the rotors. Since these emfs are opposing, there will be no current in the rotor phases.

If the transmitter rotor turns through a certain angle relative to the receiver rotor, the corresponding emfs in the transmitter and receiver phases will no longer be equal; compensating currents and the magnetic fluxes associated with them will arise in the rotor windings, and their interaction with the stator field will create in both the transmitter and the receiver a rotational moment which will try to match the positions of the transmitter and receiver rotors. The receiver rotor will assume the same angular position as the transmitter rotor, with a given degree of accuracy.

The discrepancy between the angle of rotation of the transmitter α and of the receiver β is called the angle of error $\theta = \alpha - \beta$. This angle in the ultimate state determines the error of the synchronous transmission.

The angular error θ depends on the external moments M_{ex} constituting the load on the receiver shaft;

$$\theta = \frac{M_{ex}}{M_s} \tag{10.13}$$

where M_s - specific synchronous moment, whose value is given in the certified data of the selsyns.

A deviation in the operating conditions of the selsyns from nominal (change in friction of the rings and bearings due to scale on the rings and impurities of the lubricant, change in voltage or frequency of the supply source, length of the connecting line) leads to reduced operational accuracy of the synchronous transmission. Two-speed (dual channel) synchronous transmissions are used to increase transmission accuracy. The circuit of such a transmission is shown in Fig. 10.8. It uses two parallel synchronous transmissions for coarse and fine readings. The drive shaft is directly connected to the coarse selsyn from the transmitter selsyn and through a reducing gear with accelerating transmission q to the fine detector. This provides highly accurate data readout through the fine channel (see servomechanisms for the deflection system of a plan position indicator.

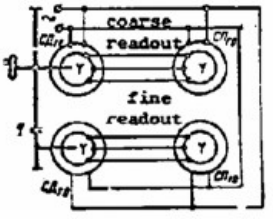


Figure 10.8. Circuit of a two-speed indicating synchronous transmission

One detector selsyn often transmits the elevations of two or more selsyn-receivers. Figure 10.9 shows the circuit of an indicator transmission, where two receivers SR_1 and SR_2 are connected in parallel to one detector selsyn SD.

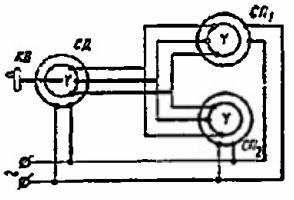


Figure 10.9. Parallel operation of receiver selsyns in a synchronous transmission

Special detectors (series DI) are produced industrially for operation with various numbers of receiving selsyns.

Differential selsyns are sometimes used for remote indicator transmission of an angle.

Differential selsyns are used to indicate the algebraic sum of angular displacement of two shafts which are not mechanically connected.

A differential selsyn (diff S) is connected in the rotor winding circuit of selsyns S_1 and S_2 of a remote angle indicator transmission, as shown in Fig. 10.10. In this circuit the differential selsyn may operate as detector or as receiver.

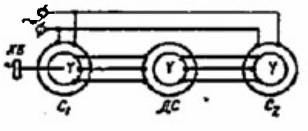


Figure 10.10. Circuit of an indicator synchronous transmission with a differential selsyn.

When selsyns S_1 and S_2 operate as detectors and the differential selsyn as a receiver, the rotor of the differential selsyn will rotate through an angle equal to the algebraic sum of the angular positions of selsyns S_1 and S_2.

Contact-less selsyns may be used in indicator synchronous transmissions. A synchronous transmission comprises either two contact-less selsyns, or one contact and one contact-less selsyn.

In the latter case, the receiver will be a contact-less selsyn (CSR) and the detector will be a contact selsyn. The three-phase windings of the DS rotor and the CSR stator are connected together, as in the usual induction transmission (Fig. 10.11).

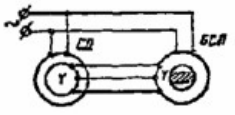


Figure 10.11. Circuit of a synchronous transmission with a contact-less selsyn

Selsyns may be used in the transformer mode in radio electronic devices.

In transformer (measuring) synchronous transmission (Fig. 10.12), the secondary winding of the detector is connected to the three-phase winding of the selsyn-transformer ST. The load is connected to the single phase winding of the transformer. Supply voltage is applied to the excitation winding of the detector.

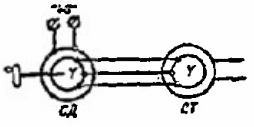


Figure 10.12. Circuit of a transformer synchronous transmission

When the detector rotor turns through a certain angle, an ac emf is induced in the single phase transformer winding, with its amplitude proportional to the angle of mismatch between the angular position of the detector rotor and the transformer rotor, and the phase of the carrier depends on the direction (sign) of the mismatch.

The fundamental indicator of quality of remote synchronous transmissions is the error in transmitting coordinates, which depends mainly on the accuracy class of the selsyns and the load on the selsyn receiver. Contact selsyns are produced in three classes of accuracy.

Maximum error:

detector selsyns	of the first class:	up to 0.25°;
	of the second class:	up to 0.5°;
	of the third class:	up to 1.0°;
receiver selsyns	of the first class:	up to ± 0.75°;
	of the second class:	from ± 0.75 to ± 1.5°;
	of the third class:	from ±1.5 to ±2.5°.

10.5 Servomechanism for the Deflecting System of a Plan Position Indicator

In a plan position indicator (PPI), the scan line should rotate synchronously and in phase with the antenna rotation. This is accomplished by rotating the deflecting coil of an electron ray tube with a servomechanism.

Figure 10.13 shows a simplified circuit of a servomechanism for rotating the deflecting system of an electron ray tube (ERT).

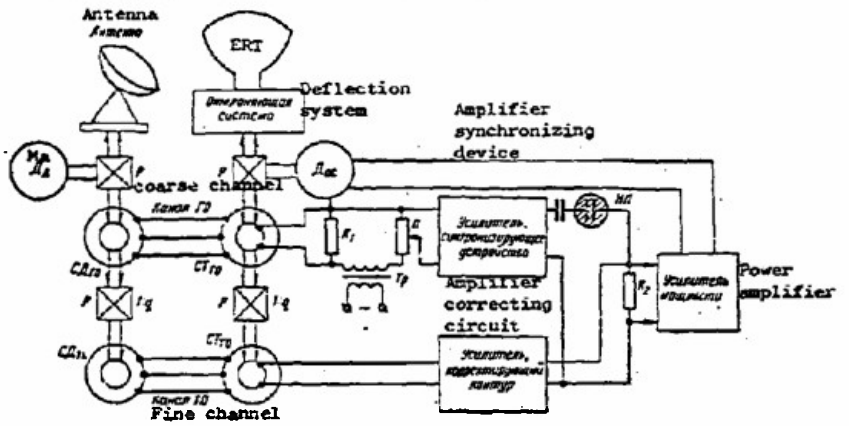


Figure 10.13. Simplified circuit of a servomechanism for rotating the deflection system of a plan position indicator.

To increase accuracy, a two-speed measuring device with coarse and fine readout channels and an additional synchronization device are used as the measuring element.

The rotors of the coarse-readout selsyn detector SD_{co} and the selsyn transformer ST_{co} are connected to the rotors of the selsyns in the fine-readout channel SD_f and ST_f through the reducing gear P with increasing mechanical speed 1:q.

When the antenna motor M_a rotates the antenna in the horizontal plane, the matched position of the selsyn pairs is disturbed.

For small mismatches between the angular position of the antenna and the deflecting winding, a voltage greatly exceeding the output voltage of the ST_{co} appears at the output of the ST_f.

After amplification in the power amplifier and conversion in the correcting device, the fine channel voltage is applied to the drive motor of the deflecting system M_{def}, which, through the reducer, swings the deflecting system and the rotors of the selsyn-transformers to decrease the mismatch. The scan on the screen of the electron ray tube moves to a new position.

However, the discussion here cannot be limited to the fine channel, because the transmission mismatch may occur in different ways. When the SD_{co} completes one rotation, the magnetic flux of ST_f makes q revolutions, that is, it will occupy the position in which its output voltage is zero a total of q times. Figure 10.14 a, shows a graph of the voltage variation on the signal windings ST_{co} and ST_f in relation to the angle of mismatch for an uneven transmission number q.

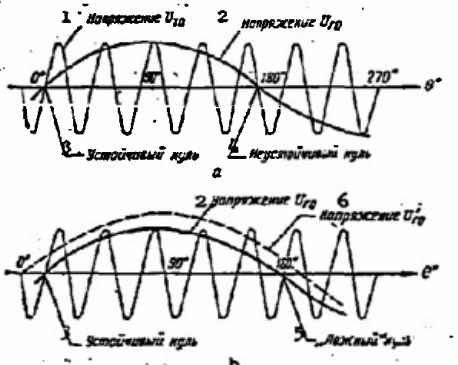


Figure 10.14. Voltages on coarse and fine selsyn-transformer channels for uneven (a) and even (b) transmission number

1 - Voltage U_f; 2 - voltage U_{co}; 3 - stable null; 4 - unstable null; 5 - "false" null; 6 - voltage U'_f

The presence of q positions with zero voltage at the ST_f output may lead to spurious matching in any of these positions. To prevent this, the system operates in the coarse channel for large mismatches, thus providing only one position for stable matching (stable null); and for small mismatches, it operates in the fine channel. The channels are switched automatically by the synchronization (switching) device.

There are a great number of synchronization devices using relays or diode circuits or other nonlinear elements. Many servo systems for radio electronic devices use a synchronization circuit with a neon lamp (NL).

When the mismatch angles exceed $\theta = 2\text{-}5°$ (for $q = 36$), output voltage

of the coarse channel after amplification in the synchronizing device is sufficiently high to ignite the neon lamp (Fig. 10.13). After the neon lamp is lit, voltages for both fine and coarse readout appear at the power amplifier input. Since the coarse channel voltage is larger in amplitude, it will be the controlling voltage until, in the process of mismatching, the neon lamp is extinguished, and the fine channel voltage will become the controlling value.

The value of the angle of mismatch at which the neon lamp is ignited may be changed by adjusting the drive with potentiometer P.

The supplementary synchronizing device cannot eliminate the stable false equilibrium (false null) which appears on the fine channel at the mismatch angle of 180° and with even transmission ratio 2 (Fig. 10.14, b).

To eliminate constant error at 180°, an additional voltage U_d is connected in series with U_{co}, the voltage on ST_{co}; U_d is taken from the secondary winding of the step-down transformer Tr, which is connected to the supply source of the SD excitation windings. With the supplementary voltage, the graph of total voltage

$$U'_{co} = U_{co} + U_d \qquad (10.14)$$

of the coarse channel is shifted upward (Fig. 10.14, b); and at the "false" null point, it will be ample to set the synchronizing device into operation.

As a result of introducing U_c, voltage U'_{co} differs from zero at $\Theta = 180°$ and $\Theta = 0°$. For the system to operate normally, signal voltages on the coarse and fine channels should remain equal to zero at $\Theta = 0°$. To fulfill this condition, the ST_{co} stator rotates through a certain angle φ, close to one fourth the period of variation of the fine channel voltage.

Electronic or magnetic amplifiers may be used in the servomechanism for rotating the deflection system of a plan position indicator.

In some radar stations, the servo system is supplemented by a differential device or by setting the scan manually in any position or at the basic voltage.

Other methods of rotating the deflection windings use a contact-less selsyn with two stator windings connected to the corresponding coils 90° away. Scan current passes through the rotor of the contact-less selsyn, which is mechanically connected to the antenna. In place of a selsyn-detector, a sin-cosin potentiometer is sometimes used.

The basic indicators of the quality* of a servomechanism for rotating

---

* Here and subsequently in calculating the technical indicators of the systems under discussion, it is assumed that they have acceptable stability margins.

the deflection system of a plan-position indicator are:

tracking accuracy;

time for the system to enter synchronization.

10.6. Servomechanisms for Controlling Antennas

Servomechanisms using selsyns with an electrical machine power amplifier are most often used for controlling antennas or the cab of a radar station.

Figure 10.15 shows the circuit of a servo with an electrical machine power amplifier.

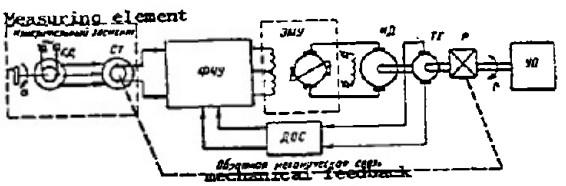


Figure 10.15. Servomechanism with an electrical machine power amplifier

The elements of the system are:

a measuring device, consisting of selsyn-detector SD and selsyn-transformer ST;

phase sensing amplifier PSA;

electrical machine power amplifier EMA;

actuator A;

reducer P;

controlled object O (antenna, cab);

supplementary feedback F, effected through dc tachometer-generator TG.

The selsyn-detector of the measuring device is located at the control point or is connected to the drive mechanism. Depending on the type of radar or its operating mode, the rotor of the selsyn-detector is mechanically connected to the command (drive) shaft and may rotate with constant speed or may be turned manually with a control wheel. The rotor of the selsyn-transformer is connected by mechanical feedback to the receiver shaft-axis of the controlled object. When there is a mismatch angle between the command and receiving shafts, an error signal in the form of an ac voltage, whose amplitude depends on the amount of mismatch and whose phase depends on the sign of the mismatch angle, appears at the single-phase winding of the selsyn-transformer.

In the phase-sensing amplifier, the error signal is transformed into a dc voltage, which is applied to the electrical machine amplifier after filtration. The amplified control voltage is supplied to the actuator. The

actuator will turn the controlled object and the selsyn-transformer rotor through the reducer to decrease the mismatch angle until the angular position β of the receiving shaft becomes equal to the angular position α of the command shaft.

The system is stabilized by feedback, supplied from tachometer-generator TG and including the supplementary circuit F, which may be a differentiating element. To increase accuracy of the servo, a two-speed measuring device is often used, consisting of two selsyn-detectors and two selsyn-transformers for coarse and fine readout. This type of measuring device has been discussed in the case of a servo for rotating the deflection system of a plan position indicator.

Two input channels are sometimes required, which may influence the actuator independently of each other. Then a differential selsyn is connected between the selsyn-detector and the selsyn-transformer, as shown in the circuit of Fig. 10.16.

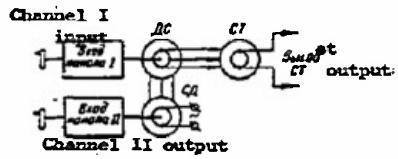


Figure 10.16. Circuit for including a differential selsyn in a servomechanism.

The stator windings of the selsyn-detector SD supply the stator windings of the differential selsyn, and the rotor windings of the detector are connected with the stator windings of the transformer. When the input shaft of channel I or channel II rotates, a mismatch voltage corresponding to the given channel appears on the terminals of the ST rotor windings. If both input shafts rotate simultaneously, the mismatch voltage and the antenna rotation voltage associated with it will be proportional to the algebraic sum of the rotations of both input shafts.

Radar display cabs are often rotated by electric motors, whose rotational velocity is varied with special relay-contact switches or with automatic devices. In some systems, the angular velocity of an asynchronous three-phase motor is varied by switching the supply windings from a delta to a double Y with an automatic machine.

The basic quality characteristic of a servo antenna control is the maximum dynamic error for a given controlling effect. It is most often required that at a given angular velocity Ω deg/s the velocity error θ of the servo with gain K_v should not exceed the given value of θ_{max}:

$$\Theta_V = \frac{\hat{\varepsilon}}{K_V} \leq \Theta_{max} \qquad (10.15)$$

The basic characteristic of a cab-rotating servo is the range or number of stages of regulated speed. Two-speed control is most often used.

10.7. Automatic Range Tracking System for Radar Stations

A system of automatic target ranging is designed for continuously determining any of the target coordinates.

Figure 10.17 shows the functional diagram of an analog automatic range tracking system (ARS). Oscillograms of voltage at principal parts of the circuit are shown in Fig. 10.18.

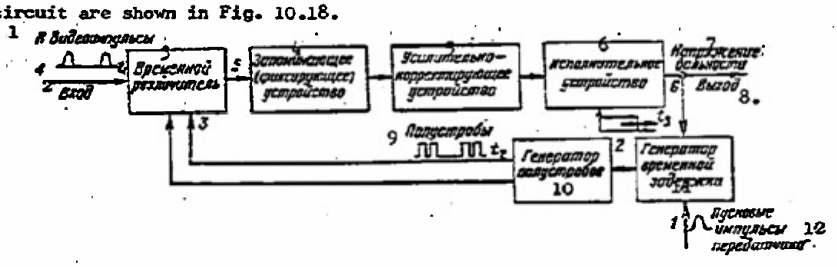


Figure 10.17. Functional diagram of a range tracking system in a pulsed radar station.

1 - videopulses; 2 - input; 3 - time discriminator; 4 - memory (fixing) device; 5 - amplifying-correcting device; 6 - actuator; 7 - range voltage; 8 - input; 9 - gate pulses; 10 - gate pulse generator; 11 - time delay generator; 12 - transmitter triggering pulses.

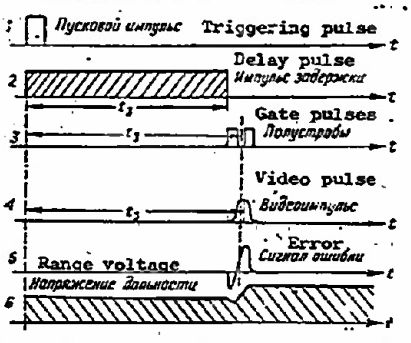


Figure 10.18. Oscillograms of voltages at certain points of the automatic range tracking system.

Range to the target is determined by measuring the time interval t_1 between the moment the pulse (1) is emitted by the transmitter and the moment

the video pulse (4) is received. To this end, the system develops a selector pulse (3) in the form of two gate pulses, whose time position is controlled by a special time delay circuit. A range scale is associated with the gate pulses.

The problem of automatic range tracking is to automatically eliminate mismatch between the time positions of the video pulse and the tracking gate pulse $\Delta t = t_1 - t_2$.

A time discriminator is used as the mismatch measuring element, to whose input the video pulses (4) from the receiver output and the tracking gate pulses (3) are applied. In the discriminator the areas of the video pulse overlapped by the first and second gate pulses are compared, and an error signal (5) is generated, proportinnal to the difference in areas, taken over the left and right gate pulses. The error voltage in the form of short two-polar pulses is applied to the memory (fixing) device, where it is converted into a continuous control signal.

A differential detector or some other fixing circuit may be used as a memory device, in which the error pulses are recorded on the first period of alternation T and are read out one after the other. The sign of the resulting error voltage is determined by the bias voltage of the gate pulses relative to the center of the video pulse.

The amplifying-correcting device, in addition to amplifying, stabilizes the system and decreases tracking error.

Range finder circuits usually include one or two integrating elements and a focusing element for stabilization.

The amplified and converted error signal is applied to the actuator, which changes the range voltage (6) and duration of the delay pulse(2).

Actuators in an automatic range tracking system can be low-power (3-5 W) two-phase asynchronous motors or electronic integrators. Accordingly, the time delay generator may be either a circuit with capacitive phase reversal or an electronic circuit using a phantastron.

The gate pulse generator includes a blocking oscillator and a delay line or two blocking oscillators.

The delay pulse circuit generates a rectangular pulse which begins simultaneously with the transmitter triggering pulse (1).

The trailing edge of the delay pulse in the gate-pulse generator is formed into the first gate pulse, and the trailing edge of the first gate pulse is formed into the second gate pulse.

The time position of the trailing edge of the delay pulse is varied by the actuator voltage until the gate pulse is located symmetrically relative to the video signal.

Mismatch signal decreases, and range, measured by the system and proportional to the interval t_2, approaches the actual range which is proportional to interval t_1. Basic indicators of the quality of an automatic range tracking system are the accuracy in determining range, and the range at which coordinates are determined accurately.

With pulse length on the order of 1μs, range error does not exceed 25 m. Depending on the type of radar, maximum range at which coordinates are determined accurately may amount to tens to hundreds of kilometers, with the minimum 400-500 m (for a target angle of 6-8°).

10.8. Automatic Azimuth Tracking System with Simultaneous Spatial Scanning

The functional diagram of one of the possible variations of an automatic azimuth tracking system with spatial scanning by a linear sweep of the directionality characteristic is shown in Fig. 10.19.

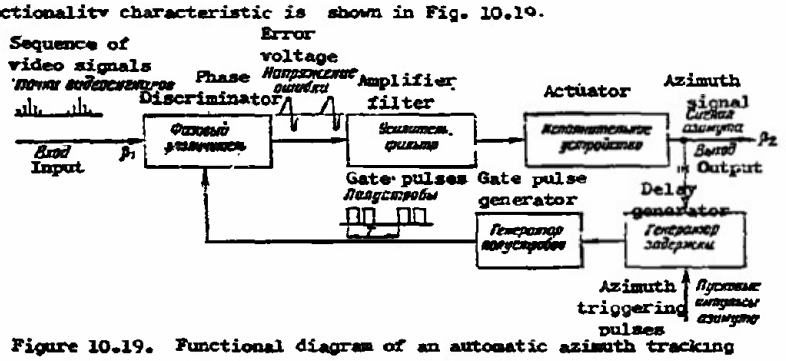


Figure 10.19. Functional diagram of an automatic azimuth tracking system with scan.

In principle the given system is analogous to the automatic range tracking system. The differences are only in the role played by each video signal in the given case of a sequence of video signals, and in the fact that the delay generator is triggered by azimuth triggering pulses, which are generated at the moment the radiation pattern of the antenna passes its maximum, taken after the beginning of the azimuth readout.

A sequence of video pulses arrives at the input to the phase discriminator with each illumination of the target. In the phase discriminator, the time positions of the azimuth gate pulses and the "center of gravity" of the signal sequence is compared, and an error signal is generated. The amplified and smoothed error voltage affects the delay generator so as to decrease the difference between the positions of the azimuth gate pulses and the sequence of signals. Azimuth β_2 measured by the system approximates the true azimuth β_1, determined by the time position of the center of the video pulse sequence.

This method of determining azimuth is used both in the case of circular scanning by the antenna ray, as well as for sector scanning. An analogous method is used for determining target elevation.

10.9. Automatic Direction Tracking System with Final Scan

The principle of final scan of the directional diagram is often used with automatic direction tracking (ADT) in precision radar stations. Figure 10.20 shows the functional diagram of an automatic direction tracking system with final ray scan.

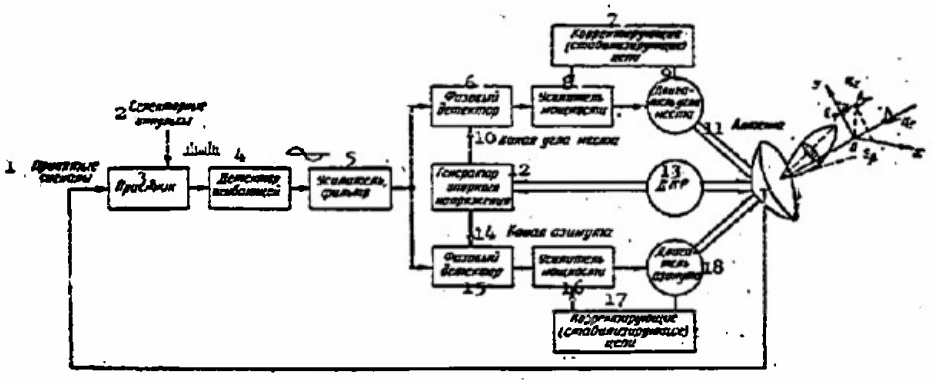


Figure 10.20. Functional diagram of an automatic direction tracking system with final scan.

1 - Received signals; 2 - selector pulses; 3 - receiver; 4 - envelope detector; 5 - amplifier, filter; 6 - phase detector; 7 - correcting (stabilizing) circuit; 8 - power amplifier; 9 - elevation motor; 10 - elevation channel; 11 - antenna; 12 - reference voltage generator; 13 - FSM; 14 - azimuth channel; 15 - phase detector; 16 - power amplifier; 17 - correcting (stabilizing) circuit; 18 - azimuth motor.

When the target is found at the axis of the cone (T_1), reflected signals at the receiver output are constant in amplitude, and when the axis is shifted relative to the axis of the paraboloid (T_2) by an angle ϵ, the amplitudes of the reflected pulses will be modulated approximately sinusoidally with the frequency of the directional pattern rotation.

For the target T_2 to appear on the equisignal line, the axis of rotation must be turned in the horizontal plane through an angle ϵ_β, and in the vertical plane through an angle $\epsilon\varphi$. The phase detectors, using reference voltages, reduce the error signal ϵ to azimuth and elevation components. Thus the automatic direction tracking system should include two servo systems

to track azimuth and elevation.

Output signals from the receiver pass through the envelope detector, amplifier, and band filter with central frequency Ω equal to the frequency of rotation of the directional pattern. The filter serves to separate the first harmonic of the signal arriving from the detector. Then the mismatch signal is applied to the phase detectors of the azimuth and elevation channels.

The reference voltage generator is connected to the final scan motor (FSM) often in such a way that two rectangular or sinusoidal voltages are generated at its output, shifted in phase by $\pi/2$.

Output voltages from the phase detectors are applied to the input of the servos in the azimuth and elevation channels. Each system contains a power amplifier, correcting and stabilizing circuits, and actuators. The latter turn the antenna through azimuth and elevation so that the axis of the paraboloid is directed with the required accuracy at the target T_2.

Systems of this type are corrected by sequential correcting circuits or by flexible feedback, including a motor and a power amplifier.

Selector pulses are applied to the special receiver channel (self-tracking channel), so that only the signal chosen for tracking the target will be separated out of all the signals reaching the receiver.

The basic indicators of the quality of an automatic direction tracking system are:

accuracy in determining angular coordinates;

range at which coordinates are determined accurately.

With a radiation pattern about 4° wide in both planes, accuracy in determining angular coordinates varies from 4 to 8'; with a pattern 5° wide, it varies from 6 to 14'.

The range at which coordinates are determined accurately is the same as for the ARS systems.

10.10. Automatic Range Tracking Systems Using an Electronic Computer

Radar tracking stations with electronic computers are used for automatic tracking of a large number of targets.

The operating principle of the automatic range tracking system using a computer is illustrated in the block diagram shown in Fig. 10.21. Two fundamental problems are involved in the process of tracking: automatic measurement ("plot") of range of all targets during the scanning period (primary information processing) and identification, or determining which of the range measurements belongs to the trajectories of each tracked target (secondary information processing).

Range of the target R is determined by the time delay t_1 of the reflected pulse relative to the triggering pulse of the transmitter:

$$R = \frac{c}{2}t_1 \qquad (10.16)$$

where c - is the propogation velocity of electromagnetic waves.

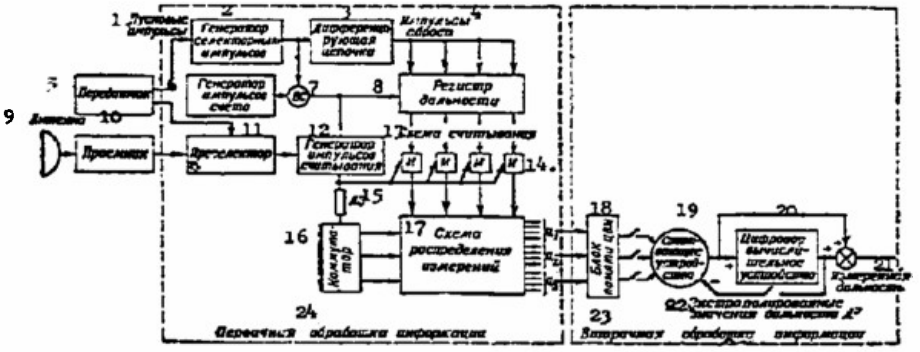


Figure 10.21. Block diagram of an automatic range tracking system with an electronic computer

1 - Triggering pulses; 2 - selector pulse generator; 3 - differentiating loop; 4 - discharge pulses; 5 - transmitter; 6 - counting pulse generator; 7 - CG; 8 - range register; 9 - antenna; 10 - receiver; 11 - preselector; 12 - read-out pulse generator; 13 - read-out circuit; 14 - AND; 15 - DL; 16 - commutator; 17 - measurement distribution circuit; 18 - digital computer memory; 19 - comparing unit; 20 - digital computer; 21 - range measurement; 22 - extrapolated range value R^*; 23 - secondary information processing; 24 - primary information processing.

By calculating the number of pulses n generated by the counting pulse generator with stable frequency F_p during time t_1, $n=F_p t_1$, target range may be found from

$$R = \frac{c}{2F_p} n, \qquad (10.17)$$

where $\frac{c}{2F_p}$ - discreteness of the range read-out.

Counting pulses are applied to the range register through the coincidence gate CG. Pulse count in the circuit begins with the arrival of a triggering pulse from the transmitter and continues until the receiver is blocked towards

the end of the sending period. The coincidence gate is opened by a selector pulse which is admitted during the whole time the receiver is open. With the end of the selector pulse, the range register discharges due to a pulse obtained by differentiating the trailing edge of the selector pulse.

Information on the time the transmitter is triggered is also given in the preselector, where the useful signal is separated out of the noise background. From the preselector output, the pulses are admitted to the range resister through a distribution circuit for readout pulses, thus opening the range readout circuit, and to the commutator which controls the operation of the circuit for distributing measurements in the computer memory blocks.

To determine the pertinence of the range measurements to the actual targets, the computer calculates coordinates extrapolated (advanced) one period. The extrapolation process generally leads to a solution of a difference equation of the form

$$R_k^e = \alpha_1 R_{k-1}^m + \alpha_2 R_{k-1}^m + \dots + \alpha_n R_{k-n}^m + \\ + \beta_1 R_{k-1}^e + \beta_2 R_{k-2}^e + \dots + \beta_m R_{k-m}^e \qquad (10.18)$$

where α_i and β_i - coefficients of the "weight" with which the measured and extrapolated values of the coordinates, R_i^m and R_i^e respectively, are computed in the preceding periods.

The separate channels of the secondary information processing system, which provide running coordinates of the target, may be schematically represented as closed digital automatic systems.

Figure 10.21 shows one channel of the secondary information processing system, which functions in the following manner.

After extrapolation has been done in the digital calculating device of the channel, the value R_k^e is compared with the measured range value R_k^m of all targets which are recorded in the computer memory in the kth period of the information sequence.

If the comparison unit finds the inequality

$$(R_k^m - R_k^e) \leqslant \Delta R_{max} \qquad (10.19)$$

where ΔR_{max} - possible extrapolation (gate) error, then the measurement is considered to be attached to the given target and the approximate value R_k^e is substituted for the precise R_k^m after extrapolating on the following period of the sequence.

If the inequality is not fulfilled for all targets, the target is considered to be new, and its data are brought to a new section of the memory.

In the full circuit of the secondary information processing unit, the

In the full circuit of the secondary information processing unit, the target is identified in all three coordinates on the basis of the region of possible (probable) target position. In the identification process, the absolute difference between extrapolated and measured values of coordinates is compared not only in the kth period, but also during several preceding periods.

10.11. Automatic Azimuth Measuring System Using a Computer

The operating principle of an automatic azimuth measuring system for radar station displays using a computer is illustrated in the block diagram shown in Fig. 10.22.

A series of circuit elements (receiver, preselector, selector pulse generator, counting pulse generator) is common to the range and azimuth measuring system, and the azimuth register, measurement distribution circuit, and commutator are analogous in their operating principle to the elements of the range measurement circuit.

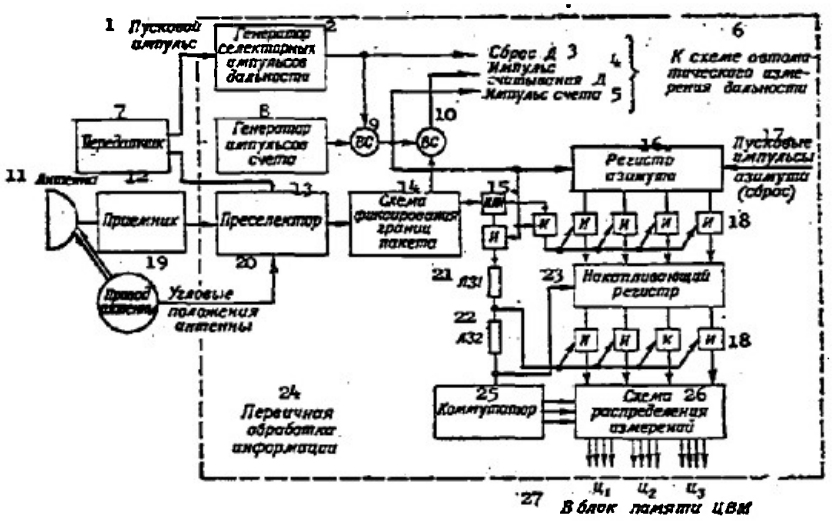


Figure 10.22. Block diagram of an automatic azimuth measuring system using a digital computer.

1 - Triggering pulse; 2 - range selector pulse generator; 3 - range discharge; 4 - range read-out pulse; 5 - counting pulse; 6 - to automatic range measuring circuit; 7 - transmitter; 8 - counting pulse generator; 9 - CG; 10 - CG; 11 - antenna; 12 - receiver; 13 - preselector; 14 - packet boundary fixing circuit; 15 - OR; 16 - azimuth register; 17 - azimuth triggering pulses (discharge); 18 - AND; 19 - antenna drive; 20 - antenna angular position; 21 - DL1; 22 - DL2; 23 - storing register; 24 - primary information processing; 25 - commutator; 26 - measurement distribution circuit; 27 - to digital computer memory unit.

With a round looking radar, a packet of reflected signals arrives from each target. Azimuth measurement requires that the location of the center of the packet of signals reflected from each target be found. This is accomplished by using a boundary forming circuit for the packet with subsequent calculation of the arithmetic mean of the time delay of the pulses in the packet boundaries in relation to the moment they cross the antenna axis in the primary direction.

The calculated time interval is then converted to digital code.

To convert to digital code, the counting pulses produced by the generator are fed to the azimuth register. The register is a digital counter of the time elapsed from the moment the signals passed the primary direction antenna. At that moment, azimuth triggering pulses are produced which discharge the azimuth register.

When a pulse at the beginning of the packet arrives at the data distribution circuit, the storage register begins to take in information in binary code through the upper stages of the AND gates.

When a pulse at the end of the packet crosses the antenna axis, read-out from the azimuth register is again passed into the storage register. The sum obtained is proportional to the azimuth of the target.

The pulse from the end of the packet is passed through a short delay line DL1 into the lower stages of the AND gates, passing azimuth information through the measurement distribution circuit into the corresponding computer memory unit. After further delay in DL2, this same pulse is used to discharge the storage register and to switch the commutator into the position required to pass azimuth information on the following target into the computer memory.

The problem of determining pertin nce of the measured azimuth to the specific target (identification problem) is also accomplished by the secondary processing system, analogous to the automatic range tracking system using a computer.

10.12. Automatic Frequency Control Systems

Automatic frequency control is required to hold the intermediate frequency of a receiver constant during variations in frequency of the transmitter and heterodyne under the effect of random factors (change in temperature, load, etc.) or during frequency retuning.

AFC circuits are automatic control systems (ACS). They vary considerably in their structure and design; however, a common block diagram of an often-used AFC circuit takes the form shown in Fig. 10.23.

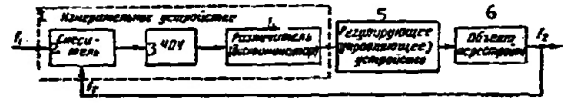


Figure 10.23. Block diagram of an automatic frequepcy control

1 - Measuring device; 2 - mixer; 3 - IFA; 4 - divider (discriminator); 5 - regulstor (controller); 6 - tuned object.

The intermediate frequency (such as $f_i=f_H-f_T$) may be held constant by adjusting the transmitter frequency f_T or the heterodyne frequency f_G. If f_G is adjusted, the tuned object in the circuit (Fig. 10.23) will be the heterodyne, and the respective input and output signals of the system will be $f_1=f_T$, $f_2=f_H$. If f_T is varied, the tuned object will be the transmitter, and the signals will be $f_1=f_H$, $f_2=f_T$.

To measure drift of the intermediate frequency from the given value, the AFC circuit has a measuring device, usually including a mixer, intermediate frequency amplifier IFA and divider (discriminator). See Chapt. VIII for operating principles of the circuit and characteristics of the measuring circuit elements (mixer, IFA, discriminator).

Frequency and phase AFC systems are possible. In frequency self-tuning systems, the measuring circuit reacts directly to the running value of the frequency difference, while in phase systems, it reacts to the phase shift, that is, to the difference in phase between reference and stabilizing oscillations.

In frequency systems, a frequency discriminator is used as discriminator. In phase systems, a phase discriminator and reference frequency generator are used in place of the frequency discriminator. Otherwise, a phase adjustment system is the same as a frequency adjustment system, and therefore, we shall consider the operation of AFC systems using frequency systems as an example.

Heterodyne and transmitter voltages enter the mixer, which produces a new voltage, whose frequency, called the running intermediate frequency, is

equal to the difference of frequencies f_H and f_T. This voltage enters the IFA and after amplification is led to the frequency discriminator. The frequency discriminator compares the running intermediate frequency with the nominal intermediate frequency, and produces a mismatch voltage, with value and polarity depending on the value and sign of the intermediate frequency shift (frequency mismatch). All elements located between the discriminator and tuned object are usually called regulating or controlling units.

The regulating element amplifies and converts the mismatch voltage into a controlling voltage whose value and form are appropriate for the tuned element. Under the effect of the controlling voltage, the heterodyne (or transmitter) frequency changes so that the difference f_H-f_s (or f_s-f_G) is approximately equal to the nominal intermediate frequency.

In many cases, the regulating device determines the type and behavior of the whole AFC system.

Depending on the type of regulating device, AFC circuits are classed as electronic, electromechanical, and thermal.

In electronic circuits, all elements of the regulating unit are electronic devices. Most frequently, dc amplifiers, thyratron and phantastron circuits are used.

Electromechanical circuits use low-power electric motors or relays as actuator element to change the heterodyne or transmitter frequency mechanically.

Thermal AFC circuits vary the frequency of special klystrons by varying temperature.

Regulating units of the majority of present-day AFC systems are fitted with an automatic search circuit which is triggered when the frequency difference increases to limits exceeding the attenuation band and ceases to search when the frequency difference is reduced to a value lower than the set band. Thus the AFC system operates in two modes: search and tuned (tracking). In the search mode, the circuit generates oscillations of a special form, which cause the klystron frequency to vary periodically within wide limits, carrying out "search" for the transmitter frequency.

In the tuned mode, the search generator is eliminated from the control circuit or operates as an inertial amplifier with a large time constant, and the system returns to the tracking mode.

To stabilize the system and improve its characteristics, additional elements may be introduced into the regulating unit as integrodifferential elements, fixing, and smoothing circuits (filters), supplementary feedback loops.

Electronic AFC systems usually have good regulation without using any kind of special correction methods. Supplementary feedback loops with tachometer generators are often used in electromechanical circuits.

The fundamental quality indicators of an AFC circuit are the following:

tuning accuracy; the value of residual error in the meter range is usually equal to 25-100 kHz, and in the centimeter range, 200-300 kHz;

search range (band of heterodyne electronic tuning) usually comprises 0.5-1.0% of the carrier frequency.

10.13. Automatic Gain Control Systems

Automatic gain control systems are designed to maintain the output voltage of a radio receiver constant during considerable changes in the amplitude of the input signal. These systems are classified as ordinary automatic gain control systems (AGC) and special systems. Instantaneous automatic gain control (IAGC) systems and automatic time gain control (ATGC) systems belong to the special systems.

Ordinary AGC systems are divided into systems without feedback (AGC "forward") and systems with feedback (AGC "backward").

Block diagrams of these two classes of AGC are shown in Fig. 10.24. Complex systems of "forward" and "backward" control, multi-branch systems, and others are all possible.

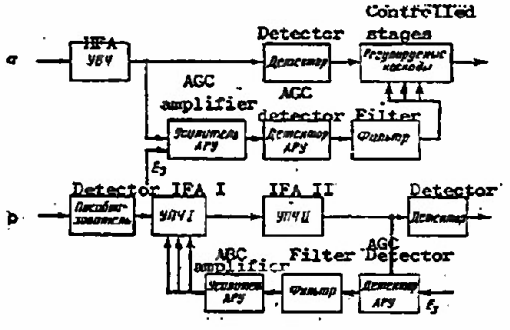


Figure 10.24. Block diagram of AGC "forward" (a) and AGC "backward" (b)

The basic elements in all these systems are the AGC detector and the controlled amplifier tube.

Amplifiers in the AGC circuits improve the stabilizing characteristics of the system. If amplification is ac, the AGC amplifier precedes the detector; if amplification is dc, it is located behind the detector.

The filter serves to filter out high and low frequency components which may exist in the voltage rectified by the detector.

A delay voltage E_d, which determines the triggering threshold of the AGC, is introduced into one of the AGC stages. In the delay circuit, receiver amplification remains unchanged until input voltage reaches a determined level.

The operating principle of an AGC circuit depends on establishing a feedback effect from the detector output to the controlled amplifier stages in the intermediate, high, and sometimes low frequency circuits.

In AGC "forward" circuits (Fig. 10.24, a) the signal from the detector output passes through the filter and enters the following (forward) amplifier stages as a negative bias.

For example, in changing (amplifying) a signal, negative bias increases and gain of the regulated stages drops. The gain can be made to vary in such a way that the output signal remains approximately constant during extensive variations in the input signal.

An AGC "forward" system operates in an open cycle. It has limited use, since, with no feedback, any changes in the conditions of gain in the stages behind the detector of the primary channel will not be compensated by the system.

An AGC "backward" system has feedback (Fig. 10.24, b), and therefore, it is an automatic control system, whose regulating action is determined by the deviation of the output voltage from a given level.

The output signal from the receiver IFA stages arrives at the AGC amplitude detector, and then after filtration and amplification is led to the preceding (backward) regulated IFA stages. Gain is most often changed by applying negative bias to the control grids of the regulating stages, although it is also possible to apply the controlling signal to other grids.

When the amplitude of a reflected signal at the receiver input increases, the rectified voltage at the AGC detector output increases. This voltage increases negative bias on the grids of the controlling tubes, decreasing gain to the point that voltage at the receiver output remains almost constant.

An instantaneous AGC circuit differs from an AGC circuit primarily in that the IAGC circuit has a smaller time constant. Consequently, its regulating voltage remains practically constant during rapid variations in the useful signal, and an additional negative bias is produced on the grid of the amplifier tube as a consequence of detecting persistent interference voltage. Thus, the instantaneous automatic gain control circuit holds the gain of useful signals constant and decreases the receiver gain for interference.

The block diagram of an IAGC is shown in Fig. 10.25. The circuit includes an IAGC detector and an IAGC amplifier. The amplifier is usually a cathode follower with positive feedback, used to stabilize operation. To eliminate extra bias at low interference voltages, the circuit operation is often delayed by supplying delay voltage E_d. The principal circuit elements (detector, amplifier) are described in Chapt. XI.

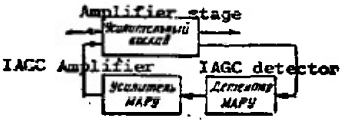


Figure 10.25. Block diagram of an instantaneous automatic gain control.

An automatic time gain control circuit (ATGC) eliminates receiver overload when it is subject to powerful pulses from the generator, transient interference, and also powerful reflected signals from objects located close to the radar station. The ATGC blocks the receiver during transmission of pulses from the generator and unblocks it when gain is decreased within several microseconds after each pulse is ended.

An ATGC circuit differs in complexity and the arrangements of its elements, and it operates in an open cycle (Fig. 10.26).

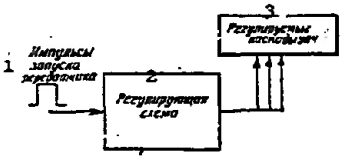


Figure 10.26. Simplified block diagram of an automatic time gain control
1 - Transmitter triggering pulses; 2 - regulating circuit; 3 - controlled IFA stages.

When a transmitter triggering pulse acts on the ATGC regulating circuit, a large negative bias voltage appears on the grid of the IFA regulating tubes, causing a sharp decrease in the gain of the IFA stages, whereupon normal reception of comparatively strong pulses becomes possible. Chapter XI contains an example of a possible ATGC circuit.

The quality of an automatic gain control system is appraised by the degree of decrease (contraction) of the dynamic range of voltage change U_{out} at the receiver output in relation to the range of voltage change U_{in} at the receiver input (Fig. 10.27).

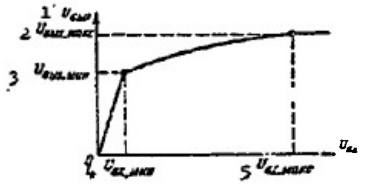


Figure 10. 27. Appraising the effectiveness of an automatic gain control system.

1 - U_{out}; 2 - $U_{out\ max}$; 3 - $U_{in\ min}$; 4 - $U_{in\ max}$; 5 - U_{in}.

To appraise the effectiveness of an AGC system in the ultimate state, we use the coefficient

$$D_o = \frac{U_{o\ max}}{U_{o\ min}} \tag{10.20}$$

indicating by how many times the amplitude of the receiver output voltage is increased if the amplitude of the input voltage increases a given amount

$$D_i = \frac{U_{i\ max}}{U_{i\ min}} \tag{10.21}$$

The ratio D_i/D_o, called the contraction coefficient, indicates how many times the AGC system decreases the dynamic range of voltage change at the receiver output in relation to its input.

The dynamic range of change in the input signals of contemporary radar receivers may be on the order of 80 to 100 dB. Note that, for the final stages and the output units to operate normally, this range must be decreased to 4 to 10 dB.

10.14. Optimum and Self-Tuning Systems

In general, optimum control is taken to mean control which provides the best solution to the problem set for the system under given actual conditions and limitations; for example to achieve maximum speed, highest accuracy in reproducing the useful signal, minimum fuel expenditure, etc.

The latest development in optimum control systems is the self-tuning system.

Self-tuning control systems provide the required amount of control (regulation) during changes in the characteristics of an external effect or the properties of the controlled objects and the regulator elements. A typical block diagram or a self-tuning system is shown in Fig. 10.28.

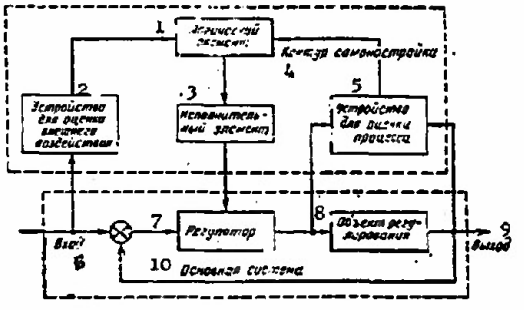


Figure 10.28. Typical block diagram of a self-tuning system

1 - Logic element; 2 - device for estimating external influence; 3 - actuator; 4 - self-tuning loop; 5 - device for estimating the process; 6 - input; 7 - regulator; 8 - controlled object; 9 - output; 10 - primary system.

A self-tuning system has two essentially differen [illegible] mary system and the self-tuning loop.

The self-tuning loop may include several elements wi[illegible]rent pu[illegible]poses. A device for estimating the process serves to identify the dynamic properties of the object. The device for estimating an external influence serves to determine the features of the external influence. The logic element (calculator) produces or stores the criteria for estimating the completeness of the system. The actuator is meant to transfer the signal (command) from the logic element to the regulator. Under the effect of this signal, the parameters or structure (characteristics or function program) of the primary system vary in accordance with varying features of the external influences or the dynamic properties of the object, and a criterion of completeness is accepted.

Those enumeration elements are most common which may be modified in actual systems and which are not strictly necessary.

Chapter XI

Countermeasures and Protecting Radar Stations Against Interference

Countermeasures include those measures, the purpose of which is to interfere with the operation of enemy radar stations, or to reduce the enemy's effectiveness.

These measures include the following:

searching for the enemy's radar equipment;

creating interference with radar stations;

radar-proof camouflage;

destruction of the enemy's radar equipment.

11.1. Radio Search for Radar Equipment

The most effective type of search by radar is search made using special radio search equipment.

Radio search for a radar system involves the detection and interception of radio signals, direction finding, and the recording, measuring, and analyzing of radar signal parameters.

As reported in the foreign press, the basic missions assigned when searching for radar stations are the following:

obtain information on the engineering characteristics of the radar in order to set up effective countermeasures;

obtain information on the enemy's radar equipment;

warning crews of aircraft, ships, and ground stations that they have been illuminated by enemy radar stations;

obtain information on the number, types, disposition, and tactics in using radar stations in order to expose the actions and the intentions of the enemy, how his equipment and forces are disposed, and other information needed in planning and conducting combat operations.

A typical search station includes a detection system, a direction finding system, a signal analysis system, and a variety of recording devices (magnetic tape recorders, photographic attachments, and others).

Search stations can be installed on the ground (transported, carried), aboard aircraft and ships, and on earth satellites.

11.2. Methods Used to Detect Radar Operation

Antennas with narrow beams are usually used in search stations in order to increase the search range, as well as for purposes of accurate direction finding. In this case directional search is made in order to detect radar operations.

The search for the frequency on which a radar is operating can be conducted by using a non-search method, one involving a non-retunable multichannel search receiver, or by tuning the frequency of a super-heterodyne receiver over a specified frequency range. This is the difference between search and non-search detection methods.

Search methods can utilize:

directional scanning only;

frequency scanning only.

It is not the practice to use frequency and directional scanning simultaneously because the possibility of missing some of the incoming signals increases.

Directional scanning is that process of making a successive inspection of space by the major lobe of the beam from the antenna of the search station, carried out in order to detect radar operation and establish the bearing.

Frequency scanning is that process of successive retuning of the receiver within the limits of a specified frequency band, carried out in order to detect the operation of a radar and to establish its carrier frequency.

Directional Scanning

It is usual in search stations to select the w[illegible] beam in the vertical plane such that scanning will not take place [illegible]d in this way scanning occurs only in the azimuth plane.

Let us conduct a search for a radar operating in the circular directional scan mode (fig. 11.1). φ_A is the effective beam width for the radar, φ_{AS} is the effective beam width for the search station. The dimensions of the φ_A and φ_{AS} sectors are established by conditions for normal signal reception. The φ_{AS} sector usually coincides with the width of the major lobe in the search station antenna pattern. The φ_A sector can coincide with, or be much broader than, the major lobe in the radar antenna pattern because of side lobe reception.

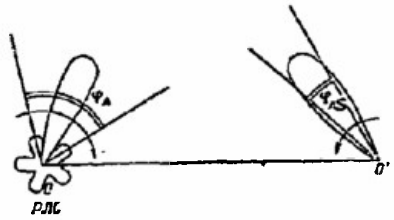


Figure 11.1. Incidence conditions for rotating patterns.

Radar signals can be detected only when φ_A and φ_{AS} are incident. Incident is understood to mean the case when both sectors intersect the line OO' for a period of time sufficient to provide normal signal indication.

Directional scanning can be assured, or probable. Assured scanning provides reliable incidence of the rotating sectors after some finite time interval. In the case of probable scanning one can only refer to the magnitude of the probability of incidence of the rotating sectors after a fixed time interval.

Assured scanning. There are two types of assured scanning: slow and fast.

In the case of slow scanning the antenna of the search station rotates more slowly than does the target radar antenna and completes a rotation of angle φ_{AS} after some time interval that is longer than, or equal to, the time required for one revolution of the radar antenna. At the same time, after one revolution of the search station's antenna there is reliable incidence of the rotating sectors, φ_A and φ_{AS}. This condition can be expressed mathematically in the form

$$T_{AS} \geqslant \frac{2\pi}{\varphi_{AS}} T_A, \tag{11.1}$$

where

T_{AS} is the period of rotation for the search station antenna;

T_A is the period of rotation for the antenna of the target radar.

There is a second condition to be satisfied if radar operation is to be detected; the incidence time, t_i, for the rotating sectors must be at least that of the time, t_{ind}, required for normal functioning of the indicator in the search station, that is

$$t_i \geqslant t_{ind} \tag{11.2}$$

When reception is of pulse signals the number of pulses supplied from the target radar during the incidence time should be at least the number required for a normal indication.

A shortcoming of slow scanning is the long scanning time required, fixed by the time it takes for the search station's antenna to make one revolution. This time becomes longer the slower the target radar antenna rotation rate. Slow scanning can be used when a search is being made for a radar with a long operating period, as well as when detection is of a radar with rapidly rotating antennas.

In the case of fast scanning the search station antenna rotates faster than does the target radar antenna and completes one full revolution in a period of time not in excess of the time required for the target radar antenna to rotate through the sector φ_A, that is

$$T_{AS} < \frac{\varphi_A}{2\pi} T_A. \tag{11.3}$$

Given this condition, there will be reliable incidence of the rotating sectors φ_A and φ_{AS} after one revolution of the target radar antenna.

The second condition, $t_i \geqslant t_{ind}$, must also be satisfied in order to detect radar operation during rapid scanning. The advantage of rapid scanning is the short scan time, fixed by the time it takes the target radar antenna to make one revolution.

The shortcomings of rapid scanning include:

a requirement for high antenna rotation rates for the search station and the engineering difficulties involved in meeting the requirement;

a requirement for a sufficiently broad beam from the search station, reducing the radio search range.

Rapid scanning can be used when searching for a radar with slowly rotating antennas.

Probable scanning. Medium-speed scan, or probable scanning, is sometimes used to reduce the detection time as compared with slow scanning, as well as to reduce the search station antenna rotation rate as compared with rapid scanning.

If scanning at average speed with the search station antenna results in a slower rotation rate for the antenna than is the case for the target radar antenna, the probability of incidence of the rotating sectors of the beams during one search station antenna revolution can be found through the formula

$$p_1 = \frac{\varphi_{AS} T_{AS}}{2\pi T_A}. \tag{11.4}$$

The probability of detecting target radar operation in n revolutions is

$$p_n = 1 - (1 - p_1)^n. \tag{11.5}$$

But if, during probable scanning, the search station antenna rotates faster than does the target radar antenna, the probability of incidence of the rotating sectors of the beams during one target radar antenna revolution can be found through the expression

$$\bar{p}_1 = \frac{\varphi_A T_A}{2\pi T_{AS}}. \tag{11.6}$$

Frequency Scanning

Retunable superheterodyne search receivers are usually used in precision search stations. Their principal advantage is a high degree of sensitivity and fine frequency resolution, determined by the receiver IF amplifier band width.

The sensitivity of superheterodyne search receivers is usually 10^{-9}–10^{-11} watt.

The error in establishing the frequency when a superheterodyne receiver is used is usually taken to be equal to half the IF amplifier pass band.

Superheterodyne search receivers for use in the frequency range up to 1000 MHz have a pass band of about 4 MHz, in the frequency range up to 3000 MHz about 10 MHz, and for the higher frequency 20 MHz and more.

The tuning band of the search receiver can cover thousands of megahertz and so is divided up into subbands. Retuning is done in the subband selected.

Frequency scanning, like directional scanning, can be assured, or probable. Slow, or fast, frequency search can be used for assured detection of radar operation.

Slow frequency scanning. Let us carry out the procedure involved in detecting signals from a radar operating in the pulse mode on frequency f_s. The antenna of the search station is aimed at the radar. Scanning is in frequency only. Retuning of the search receiver is in accordance with a symmetrical sawtooth law (fig. 11.2) in the band f_2-f_1, T_r is the retuning period, ΔF_s is the search receiver band width.

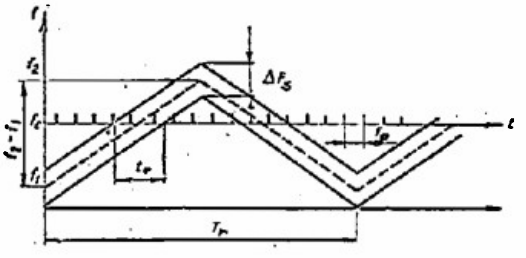


Figure 11.2. Slow frequency scanning

Slow frequency scanning is characterized by the fact that the search receiver is retuned over the entire search band twice in a period of time shorter, or equal to, the illumination time, t_{ill}. In this case the retuning rate γ_f, should satisfy the condition

$$\gamma_f \geqslant \frac{2(f_2-f_1)}{t_{ill}} = \frac{2\pi}{\gamma_A T_A} 2(f_2-f_1). \qquad (11.7)$$

At the same time, the retuning rate should be such that during the time required to retune over the band width the search receiver can take at least N_p pulses required for the normal operation of the indicators

$$\gamma_f \leqslant \frac{\Delta F_s}{N_p T_p}. \qquad (11.8)$$

Detection of radar operation is mandatory if these two conditions are satisfied during one retuning period, T_r.

A shortcoming of slow frequency scanning is that the search band must be taken quite narrow in order to satisfy both of the conditions stipulated at (11.7) and (11.8).

Fast frequency scanning. Detection of radar operation should occur with one pulse when retuning for the carrier frequency of the radar from pulse to pulse. This is when fast frequency scanning is used. But if the search receiver is retuned in accordance with a symmetrical sawtooth law, and fast scanning is used, the time for double retuning of the receiver over the entire search band must be shorter than the pulse duration τ_p, for the target radar; that is, the retuning rate must satisfy the condition

$$\gamma_f \geqslant \frac{2(f_2 - f_1)}{\tau_p}. \qquad (11.9)$$

Now reliable detection of radar operation will occur after each pulse. However, the search receiver retuning rate is limited by the time required for the oscillations in the receiver to rise to a sufficient magnitude. Consequently, the permissible retuning rate is fixed by the receiver pass band

$$\gamma_f < \Delta F_p^2. \qquad (11.10)$$

Shortcoming of fast scanning include a requirement for a wide band search receiver in order to scan a broad band, and this reduces receiver sensitivity, as well as the accuracy with which the frequency is established.

If the fast and slow frequency scanning conditions cannot be satisfied, detection of radar operation will occur after a finite time interval with some degree of probability.

11.3. Non-Scanning Methods of Searching for the Operating Frequency

Non-scanning methods of searching for the operating frequency are used when short search time is a requirement.

The bandpass filter method is used in coarse search stations. Each such filter has a fixed pass band, providing coarse determination of the incoming signal frequency. The error is half the filter pass band. A separate antenna, and a separate amplifier channel can be used for each bandpass filter to eliminate the mutual effect of the filters and provide separate indications for the signals passing through the different filters. Figure 11.3 is the block schematic of a search system using straight amplification and providing simultaneous reception in the 1 to 10 cm band.

The high frequency devices in the receivers (antenna, waveguides, filters, detectors) are designed to have the limit frequencies of adjacent subbands overlap somewhat (fig. 11.4a). Absence of these overlapping zones results in attenuation, or even total loss of reception where the subbands meet (fig. 11.4b). However, these overlapping zones should be kept to a minimum, because broadly overlapping zones can dictate the need for an

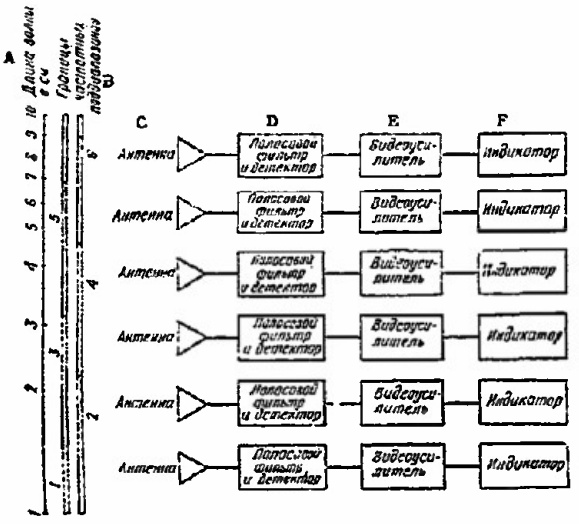


Figure 11.3. Multichannel direct amplification search receiver.

A - wave length in cm; B - limits of frequency subbands; C - antenna; D - bandpass filter and detector; E - video amplifier; F - indicator.

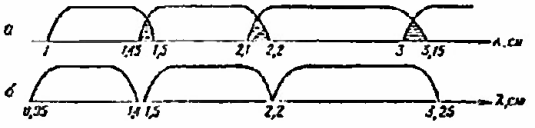


Figure 11.4. Use of overlapping zones to eliminate nulls where adjacent subbands meet.

increase in the number of subbands, increasing the bulk of the equipment accordingly. Moreover, reception ambiguity will occur in an overlapping zone because signal reception will be noted simultaneously on adjacent channel indicators.

11.4. Radio Search Range

The range of search equipment is understood to mean the maximum distance at which the source of radiation can be detected when in operation.

Radio search range in free space can be found through the formula

$$R_r = \frac{\lambda}{4\pi}\sqrt{\frac{P_{rad}\, g\, g_s \gamma \eta}{v P'_{p\,min}}} \tag{11.11}$$

where

P_{rad} is the power radiated by the target radar;

g is the antenna gain factor for the radar in the direction toward the search station;

g_s is the antenna gain factor for the search station;

γ is a coefficient taking into consideration the non-coincidence of polarization of search and target station antennas (taken as equal to 0.5 in calculations);

η is the signal power propagation factor between the antenna and the input to the receiver in the search station (taken as equal to 0.5 in calculations);

$P'_{p\,min}$ is the cutoff sensitivity of the receiver in the search station, fixed by the power of own noise, equated to the input;

v is the coefficient of excess of signal over noise level in the receiver needed for dependable operation of the analyzing and display equipment used with the search station.

This formula does not consider attenuation in the propagation of radiowaves in the atmosphere, something that weighs particularly heavily in the millimeter and shortwave portions of the centimeter bands.

When attenuation is considered the range can be found through the formula

$$R_r = \lambda/4\pi\sqrt{P_{rad} g\, g_s \gamma \eta / v P'_{p\,min}} \cdot e^{-\alpha R_r/20} \tag{11.12}$$

where

α is the coefficient of absorption of electromagnetic energy per kilometer of path in the atmosphere, expressed in db/km (its value is taken from special tables or graphics).

In the millimeter, centimeter, and decimeter bands, in which the diffraction phenomenon is very weak, the maximum radio search range is limited to horizon range.

Horizon range can be found through the following formula when atmospheric refraction is normal

$$R_v[\text{km}] \doteq 4.1(\sqrt{h_t[\text{m}]} + \sqrt{h_s[\text{m}]}), \tag{11.13}$$

where

h_t is the height of the target station antenna above ground level;

h_s is the height of the search station antenna above ground level.

As has been noted in the press, the utilization of the phenomenon of long-range tropospheric propagation is a recent innovation in the field of super-long-range radio search in the meter and centimeter bands.

11.5. Target Radar Parameters

Modern fine search stations are usually equipped with analyzing equipment for use in establishing many radar parameters, such as:

carrier frequency;

type of modulation;

pulse width, shape, and repetition frequency;

radiation pattern width and shape;

search character and sector;

antenna rotation rate;

width and structure of code trains;

signal spectrum structure;

direction to source of radiation.

These data provide a sufficiently firm basis for deriving the type and tactical purposes of target radars. This information can be used for operational search targets, as well as for purposes of organizing effective electronic countermeasures.

These target parameters can be measured directly, at the time of search, by analyzers, as well as recorded on special recorders using tape, photography, punched tape, or punched cards.

11.6. Radar Jamming

Any radio signals arriving at a receiver input, and hampering separation of useful signals, can jam a radar. The different types of jamming are listed, in brief, in Figure 11.5. All types of jamming can be divided into two large groupings:

organized (electronic);

unorganized (non-electronic).

Unorganized jamming includes mutual radar interference, atmospheric and man-made interference, reflections from local objects and clouds, interference of cosmic origin, and internal noise in radio receivers.

Organized jamming is produced specifically for purposes of reducing the effectiveness of radar operation, or to jam the radar completely.

Organized jamming can be divided into active jamming, generated by special transmitters, and passive jamming, the result of the reception of electromagnetic waves reflected from a variety of artificial reflectors, depending on the method used to jam.

The manner in which the jamming effects radar can be conventionally divided into two types: masking and simulative.

Masking jamming can be generated by narrow-band and wide-band transmitters

of continuous and strobed noise, by transmitters of pulse noise, false targets (traps), and by parasitic dipole reflectors.

Simulative (misinformation) jamming can be generated by transmitters of response type jamming and by special simulators. In addition, range tracking and angle tracking can be broken up, and false targets can be imitated. Misinformation jamming can also be generated by corner reflectors and false targets (traps).

Jamming can be barrage, or selective, depending on the width of the spectrum of radiated frequency.

Selective jamming has a spectrum width of the same order as that of the pass band of the receiving device in jammed radar. This type of jamming is characterized by a comparatively high spectral power density. Spectral power density is the ratio of full transmitter power to the width of the frequency spectrum of the oscillations generated by the transmitter and is usually expressed in watts per megahertz. The shortcoming in selective jamming is the requirement that the jamming transmitter be finely tuned to the frequency of the radar being jammed.

Moreover, when selective jamming is used one transmitter can only jam one radar frequency channel at any given moment in time.

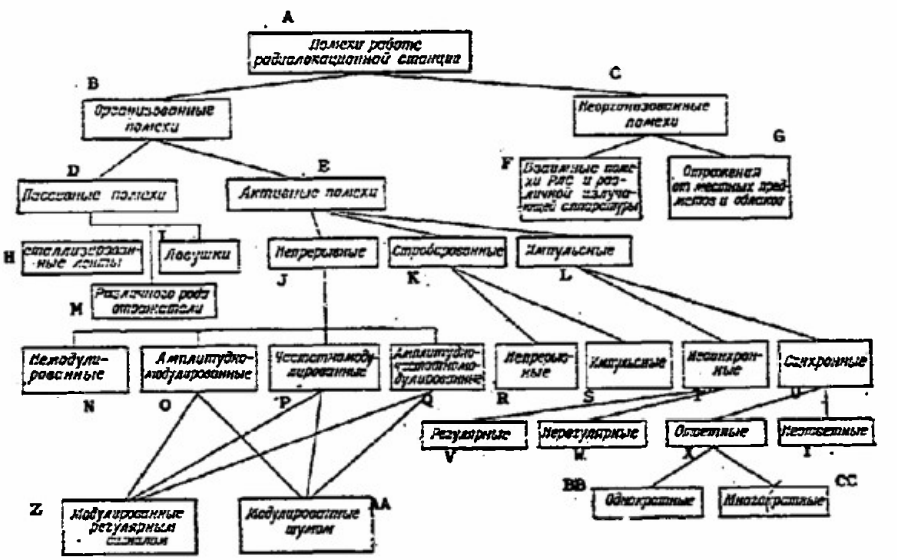


Figure 11.5. Jamming classifications.

A - radar jamming; B - organized jamming; C - unorganized jamming; D - passive jamming; E - active jamming; F - mutual

jamming of radars and various radiating equipment; G - reflections from local objects and clouds; H - window; I - traps; J - continuous; K - strobed; L - pulsed; M - various types of reflectors; N - unmodulated; O - amplitude-modulated; P - frequency-modulated; Q - amplitude-frequency modulated; R - continuous; S- pulsed; T - non-synchronous; U - synchronous; V - regular; W - irregular; X - response; Y - non-response; Z - simulated by regular signals; AA - simulated by noise; BB - simple; CC - multiplex.

The difficulties involved in generating selective jamming increase greatly when there are a great many radars involved, as well as when rapid radar retuning is utilized.

Barrage jamming is generated over a broad band of frequencies to simultaneously jam several radars operating on different frequencies. Barrage jamming requires no accurate matching of jamming transmitter frequencies and target radar frequencies. The shortcoming of barrage jamming is that high-powered jammers are needed to generate the high spectral power density.

Selective jamming with frequency switching (sliding jamming) is considered to be a compromise between these two types of jamming. In this method the transmitter frequency is changed very rapidly over a broad band of frequencies. It concentrates a high power density from the jamming transmitter in all possible radar operating channels in a very short period of time. Given the corresponding selection of frequency switching speed, the jamming transmitter can attempt to prevent the radar receiver from successfully restoring full sensitivity in the time required to retune the jammer to the operating band of frequencies for the radar after the effect of power jamming has been used.

11.7. Continuous Unmodulated Jamming

The mathematical expression for oscillations with fixed amplitude and frequency is in the form

$$u_j = U_{mj} \cos \omega_j t \ , \tag{11.14}$$

where

U_{mj} and ω_j are the amplitude and carrier frequency of the jamming, respectively.

The useful signal during the time the RF pulse reflected from the target acts is

$$u_s = U_{ms} \cos (\omega_s t + \varphi_k). \tag{11.15}$$

where

U_{ms} and ω_s are amplitude and carrier frequency of the reflected RF pulse;

φ_k is the phase of the fill frequency for the k-th RF pulse.

When jamming and signal act on the receiver at the same time, the resultant voltage at the IF amplifier output has an amplitude established through the expression

$$U_{mr} = K_{IFa}\sqrt{U_{mj}^2 + U_{ms}^2 + 2U_{mj}U_{ms}\cos(\omega_b t + \varphi_k)}\ , \qquad (11.16)$$

where

K_{IFa} is the gain in the IF amplifier stages;

ω_b is the beat frequency, equal to the difference between signal and jamming frequencies ($\omega_b = \omega_s - \omega_j$).

The nature of the effect of unmodulated jamming on the receiver depends on the jamming amplitude, U_{mj}, and the degree of detuning, which is established by the beat frequency ω_b.

The Effect of Weak Unmodulated Jamming

Unmodulated jamming, in which the amplitude of U_{mj}, although greater than the signal amplitude, U_{ms}, still does not overload the receiver stages, is called weak unmodulated jamming.

In the case of slight detuning of the jamming with respect to the signal frequency, when the beat period is many times greater than the pulse width,

$$T_b = \frac{2\pi}{\omega_b} \gg \tau_p,$$

the amplitude of the resultant voltage can be written in the form

$$U_{mr} = K_{IFa}\sqrt{U_{mj}^2 + U_{ms}^2 + 2U_{mj}U_{ms}\cos\varphi_k}\ , \qquad (11.17)$$

where

φ_k = constant when $0 < t \leqslant \tau_p$.

Since the phase φ_k will change from pulse to pulse and can take any value in the range 0 to 2π, the amplitude of the resultant voltage at the output of the IF amplifier will change in the range from $K_{IFa}(U_{mj}-U_{ms})$ to $K_{IFa}(U_{mj}+U_{ms})$. There is virtually no distortion in pulse shape (fig. 11.6).

A constant component and negative (when $\varphi_k = 0$), or positive (when $\varphi_k = \pi$), polarity (fig. 11.7) are obtained across the detector load, R_{load}.

In the case of galvanic coupling of the video amplifier to the detector, the presence of a large constant component at the detector output causes a displacement of the operating point of the video amplifier stage with the result that the video amplifier gain factor changes, and clipping of the signal from the target can occur as a result of anode current cut-off, or as a result of the grid currents flowing in the video amplifier tubes. The useful signal will be weakened, or completely disappear.

The insertion of a transition circuit in the video amplifier first stage input eliminates the effect of the constant component of the output voltage from the detector. But the useful signal with polarity opposite to that for which the video pulses amplifier was designed, can be weakened. Consequently, pulses of reverse polarity will be dimly seen on the screen of an indicator with amplitude markers, and will not be seen at all on a tube with brilliance markers.

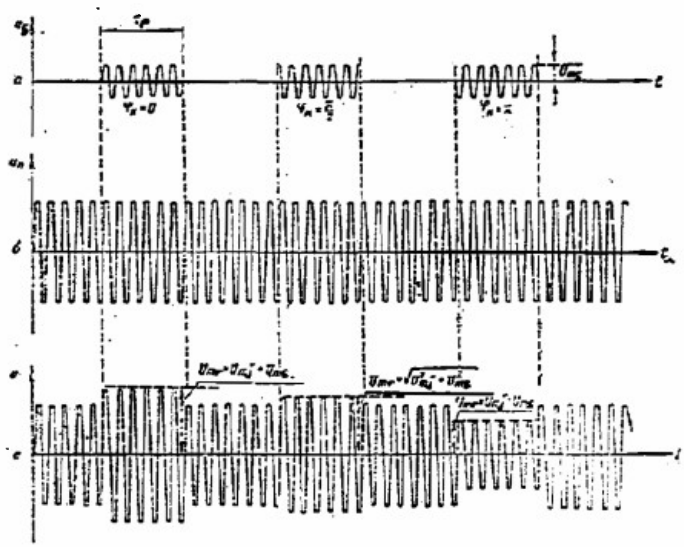


Figure 11.6. The appearance of the pulsating pulses at the IF amplifier output as a result of the random relationship between the phases of weak unmodulated jamming and signal.

Since probability of the appearance of positive and negative pulses is identical, the presence of unmodulated jamming reduces the number of pulses creating the target pip by a factor of approximately two.

In the case of severe detuning of the jamming relative to the signal frequency (beat period $T_b = \frac{2\pi}{\omega_b} \leqslant \tau_p$) the value of $\cos(\omega_b t + \varphi_k)$ while the reflected signal is acting will change from -1 to +1, regardless of the magnitude of φ_k. And the envelope of the amplitude of the resultant voltage will change in accordance with a harmonic law, with the initial phase of the envelope established by the magnitude of φ_k.

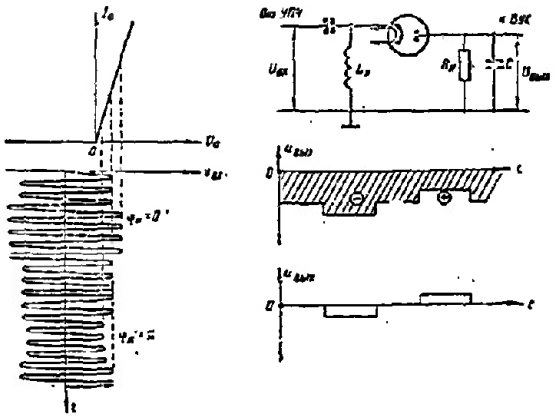


Figure 11.7. Effect of weak unmodulated jamming on the detector.

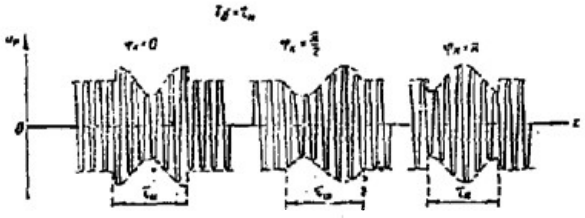


Figure 11.8. Distortion in the shape of pulses in the IF amplifier as a result of the effect of weak unmodulated jamming when $T_b = \tau_p$.

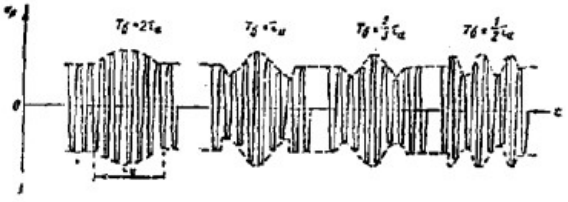


Figure 11.9. Distortion in the shape of the pulses in the IF amplifier as a result of the effect of weak unmodulated jamming for various T_b

Figure 11.8 shows the graphics of the resultant voltage at the IF amplifier output for $T_b = \tau_p$ and three different values of φ_k. Figure 11.9 shows the graphics of the resultant voltage across the IF amplifier output for $T_b = 2\tau_p$, $T_b = \tau_p$, $T_b = 2/3\tau_p$, $T_b = 1/2\tau_p$. In this case distortion of the shape of the video pulse (its breakdown) will occur at the detector output, in addition to the appearance of the constant component and upset in the functioning of the video amplifier. Useful signal attenuation will also take place.

The Effect of Strong Unmodulated Jamming

Strong unmodulated jamming overloads the IF amplifier. The overload is the result of saturation of the plate circuits of the amplifier tubes, or by the grid-circuit clipping, or by the simultaneous action of both. Complete jamming of the useful signal (fig. 11.10) can result. When strong unmodulated jamming is acting on a scope with amplitude markers one sees

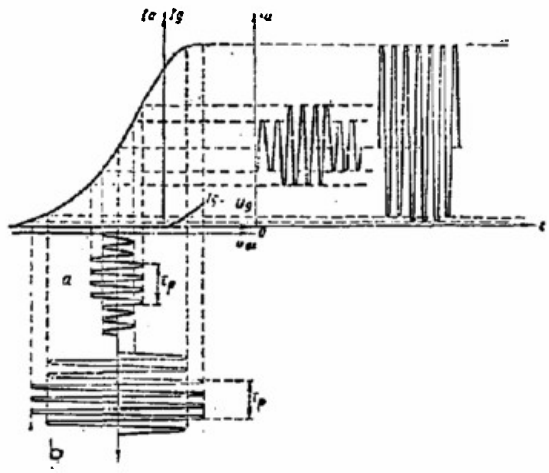


Figure 11.10. Jamming of the useful signal by strong unmodulated jamming in the IF amplifier stages. a - weak unmodulated jamming does not jam the signal; b - strong unmodulated jamming overloads the IF amplifier stages.

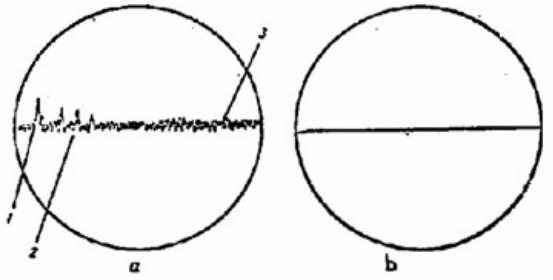


Figure 11.11. View of range scope.

a - no jamming (seen on the scope is the main pulse from the transmitter - 1; pulses reflected from local objects - 2; pulses reflected from the target - 3; inherent receiver noise); b - strong unmodulated jamming.

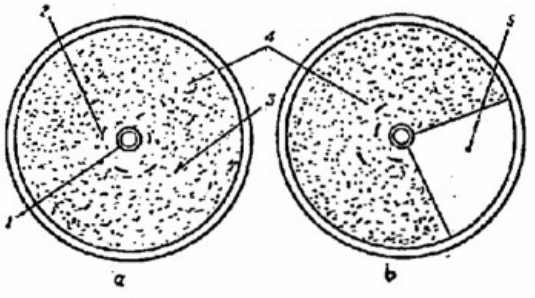


Figure 11.12. View of PPI.

a - no jamming (1 - main pulse from transmitter; 2 - pulses reflected from local objects; 3 - pulses reflected from target; 4 - surface of PPI intensified by noise); b - strong unmodulated jamming (darkened sector, 5, corresponding to the main lobe of the antenna pattern, is seen on the screen).

the target pip vanish and the inherent noise in the receiver, as well as reflections from local objects will diminish, or virtually disappear (fig. 11.11). A dark sector will appear on the PPI: one corresponding to the main lobe in the antenna pattern (fig. 11.12).

It has been reported that continuous unmodulated jamming is not widely used against pulse radars because it can be overcome by relatively simple means.

11.8. Continuous Jamming Amplitude Modulated by Sinusoidal Oscillations

The mathematical expression for oscillations amplitude modulated by harmonic oscillations of frequency Ω, is in the form

$$u = U_{mj}(1 + m \cos \Omega t) \sin \omega_j t \tag{11.18}$$

or

$$u = U_{mj} \sin \omega_j t + m\frac{U_{mj}}{2} \sin (\omega_j - \Omega)t + m\frac{U_{mj}}{2} \sin (\omega_j + \Omega)t,$$

where m is the modulation factor.

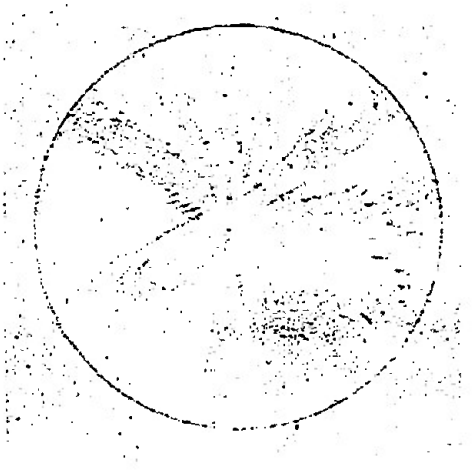


Figure 11.13. PPI screen showing continuous jamming modulated by low frequency oscillations.

A constant component and a modulating oscillation with frequency Ω will be separated at the detector output if the IF amplifier is not overloaded.

If the video amplifier is protected against the constant component, a modulating oscillation with frequency Ω will be reproduced at the video amplifier output, and this will be fed into the indicator and reproduced on the screens of cathode ray tubes.

When the frequencies of the modulating voltage of the jamming and the scanning voltage of the indicator are multiples, the oscillogram will be fixed. Practically speaking, because of instability in the frequencies,

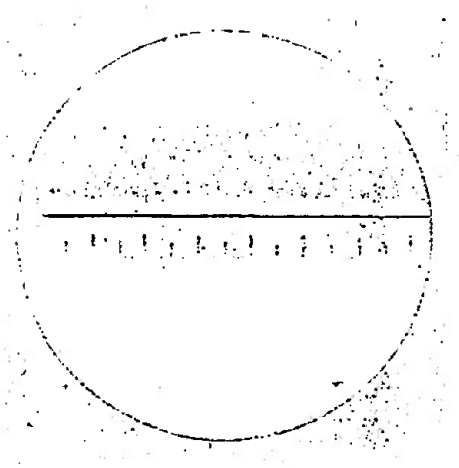


Figure 11.14. Range indicator scope showing continuous jamming modulated by high frequency oscillations.

the multiple nature of their relationship is upset, and the oscillogram moves along the sweep, forming a flickering image. Observation of the useful signal is difficult. Figures 11.13 and 11.14 are views of the screens of indicators so affected. The higher the frequency of the non-multiple nature Ω/Ω_{sweep}, the greater the modulation factor.

Moreover, amplitude modulated jamming can cause the same phenomena as can unmodulated:

overloading of the IF amplifier and jamming of the signal before it reaches the detector as a result thereof;

appearance of dual-polarity pulses at the detector output as a result of the random ratio between the phases of the carrier frequencies of signal and noise;

distortion of the shape of the pulses as a result of the difference in signal and noise frequencies.

The jamming discussed is narrow band. The equipment for coping with it is simple. Hence it has not been widely used either. But it can be used when conical scanning by the antenna beam is used to angle track. In this case the modulation is by the spatial scanning frequency. An angular error is introduced in the antenna control system when jamming of this type is in use, or automatic tracking is completely upset.

11.9. Pulse Noise

Pulse noise is a sequence of unmodulated, or modulated, high frequency pulses generated by the transmitters tuned to the frequency of the target radars. The synchronous pulse noises at which the pulse repetition frequency is equal to, or is a multiple of, the pulse repetition frequency of the target radar, and the non-synchronous pulse noises, differ.

The noise pulses on the screen will remain fixed when synchronous pulse noise with pip amplitude is working on the screen, for in this case the pulse noise is very similar in appearance to target pips (fig. 11.15). If the noise pulses are close to target pulses in amplitude and width, it is very difficult to distinguish the useful signals from the interference.

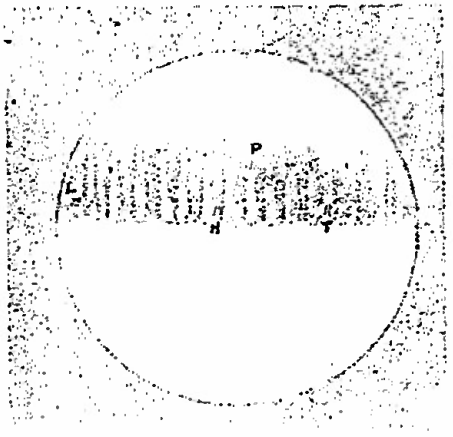


Figure 11.15. View of a range scope screen with synchronous pulse noise showing (L - pulses from local objects; N - receiver noise; P - pulse noise; T - target pulse).

Non-synchronous pulse noise moves the noise pulses about on the screen, and it is very difficult to distinguish the target pip (fig. 11.16).

Large amplitude pulse noise overloads the stages, causing a loss of receiver response for some time after it has acted.

When the radar is operating in the target range tracking mode, and in order to drain off the tracking strobe, response jamming with a false delay time can be used, leading to an interruption in range tracking of the target.

Simplex and multiplex response pulse noise can also be used to imitate false targets.

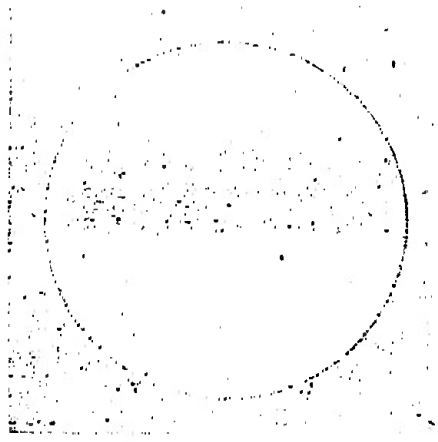


Figure 11.16. View of a range scope showing unsynchronized pulse noise.

Noise with a regular structure is of little effect. First of all, it is usually narrow band. Second, since the modulating signal has a definite periodic behavior pattern, selection of the corresponding filters can weaken considerably, or even do away with, the effect of the noise.

The greatest effect is by noise with an irregular structure, so called noise jamming. Noise jamming is similar to the internal fluctuating noise inherently present in the receiver, so if no special measures are undertaken, it is difficult to distinguish from set noise.

11.10. Fluctuation (Straight Noise) Jamming

Straight noise jamming is a limited band of frequencies of fluctuation noise amplified in the corresponding devices and radiated in order to jam electronic equipment. Moreover, fluctuation jamming can be formed in the equipment itself because of fluctuations in the currents flowing in the tubes (the shot effect) and as a result of the chaotic movement of the electrons in resistors and in other components.

Fluctuation jamming is a random process, and probability laws (so called probability distribution laws) are resorted to to describe it. However, there are many problems dealt with when knowledge of certain statistical characteristics of the process is sufficient. A process is said to be

stationary when the distribution laws and all the other probability characteristics associated with it do not depend on the origin of the time reading.

Stationary processes possess ergodidic properties, in accordance with which any realization of the process has the same statistical properties. Therefore, in the case of ergodidic random processes any statistical characteristic obtained by averaging many possible realizations with a probability as close to unity as possible can be obtained by averaging a single realization of the random process over a sufficiently long time interval.

In the case of stationary fluctuation jamming, the basic characteristics, which can be found by time averaging, are:

the average value, or the constant component of the noise voltage

$$\bar{u} = \lim_{T\to\infty} \frac{1}{T} \int_0^T u\,dt; \tag{11.19}$$

the effective, or mean square value of the voltage

$$U_{eff} = \lim_{T\to\infty} \sqrt{\frac{1}{T} \int_0^T u^2\,dt}; \tag{11.20}$$

the effective squared value of the voltage, or the mean square noise voltage

$$\overline{u^2} = U_{eff}^2 = \lim_{T\to\infty} \frac{1}{T} \int_0^T u^2\,dt \tag{11.21}$$

(the latter magnitude is numerically equal to the average power for the process, separable by a resistance of 1 ohm);

the mean square variable component (dispersion) of the process

$$\sigma^2 = \overline{u_{\sim}^2} = \lim_{T\to\infty} \frac{1}{T} \int_0^T (u - \bar{u})^2\,dt. \tag{11.22}$$

The dispersion is numerically equal to the average power of the variable component of the process, separable by a resistance of 1 ohm.

The Correlation Function

Most completely characteristic of fluctuation jamming is the correlation function, which is the time average of the product of two running values of fluctuation voltages displaced by an interval of time τ:

$$R(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_0^T u(t)\,u(t+\tau)\,dt. \tag{11.23}$$

The correlation function characterizes the association between the preceding and succeeding values of the fluctuation voltage time interval τ apart.

Figure 11.17 shows a typical curve of the correlation function for a random stationary process.

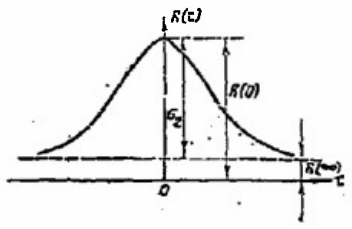


Figure 11.17. Graphic of the correlation function of a stationary random process.

Basic properties of the correlation function.

1. The correlation function is even

$$R(\tau) = R(-\tau).$$

2. When $\tau \to \infty$, the dependence between the magnitudes u(t), u(t + τ) weakens. At the same time, if there is no constant component i the process

$$\lim_{\tau \to \infty} R(\tau) = R(\infty) = 0.$$

When there is a constant component, R(∞) is numerically equal to the power of the constant component of the process, separable by a resistance of 1 ohm.

$$R(\infty) = (\overline{u})^2$$

3. When $\tau = 0$, the correlation function takes its maximum value and is numerically equal to the total average power for the process, separably by a resistance of 1 ohm.

Dispersion of the stationary process is

$$\sigma^2 = R(0) - R(\infty).$$

If the correlation function is reduced to an extremely small value (say, to 1% or 5%) over a period of time $\tau \sim \tau_c$, the magnitude τ_c is called the time, or the interval, of correlation. The magnitudes u(t), u(t+τ_c) can be taken as practically independent.

The Energy Spectrum

The concept of an energy spectrum, or of spectral power density, is introduced for the characteristic curve of the spectral composition of the fluctuation jamming. By definition, the following magnitude is called the spectral power density

$$G(f) = \lim_{\Delta f \to 0} \frac{p(f, f + \Delta f)}{\Delta f} = \frac{p(f, f + df)}{df} \qquad (11.24)$$

where

df is an infinitely narrow band of frequencies;

p(f, f+ df) is the power contained within this infinitely narrow band.

The spectral power density is a function of the frequency and indicates the distribution of the power of the fluctuations with respect to the frequency band.

The spectral power density has the dimensionality watts/hertz, or watts/megahertz.

It is sometimes suggested that there is some ideal source, from the output of which one can take noise voltage with an infinitely uniform energy spectrum, the purpose of which is to study the properties of fluctuation noise. This noise is called white noise, by analogy to white light, the spectrum of which in the visible portion is continuous and uniform. In fact,

there is no such thing as white noise because any electronic device has a finite frequency pass band and a limited dynamic range.

The nature and width of the spectrum of fluctuation depends on the noise source, as well as on the circuit through which the jamming flows.

Some sources generate noise with a uniform spectrum. Noise such as this is created by electronic tubes (because of the shot effect), pure resistances, and other sources.

The spectral density of the noises generated by pure resistance R, can be found through the expression

$$G(f) = 4kTR, \tag{11.25}$$

where

k is the Boltzmann's constant, equal to $1{,}38 \cdot 10^{-23}$ joules/degree;

T is the absolute temperature

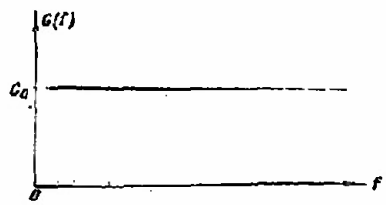


Figure 11.18. Graphic of the white noise energy spectrum.

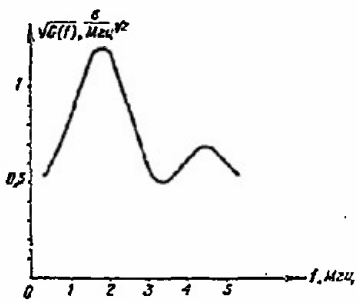


Figure 11.19. Spectrum of noises generated by a thyratron
Vertical: volts/MHz$^{1/2}$; horizontal: MHz.

The energy spectrum of the noises generated by pure resistance does not depend on the frequency in the range to 10^{13}-10^{14} hertz, but does serve as a kind of approximation of the energy spectrum of white noise.

The graphic of the energy spectrum of white noise is shown in Figure 11.18.

Other sources, such as a thyratron, for example, produce noise with a nonuniform spectrum (fig. 11.19).

Atmospheric noise, the intensity of which decays with increase in frequency, has a nonuniform spectrum.

The structure of the fluctuation noise depends on the width of the energy spectrum. Jamming is called narrow band if the average frequency in the spectrum is much greater than its width (fig. 11.20a). Jamming is called broad band if the width of its energy spectrum is comparable to the average frequency of this spectrum (fig. 11.20b). The average duration of the voltage spikes is longer for narrow band jamming than it is for jamming with a broad energy spectrum. The duration of the noise spikes can be approximated by the magnitude

$$\tau_{\text{noise spike}} \approx \frac{1}{\Delta f_{\text{width}}} = \tau_c \tag{11.26}$$

where

Δf_{width} is the width of the energy spectrum, found through the formula

$$\Delta f_{\text{width}} = \frac{\int_0^\infty G(f)\,df}{[G(f)]_{\max}}; \tag{11.27}$$

where τ_c is the correlation time.

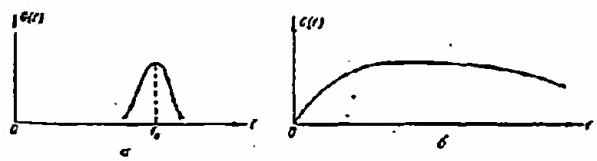


Figure 11.20. Energy spectrums for narrow band (a) and broad band (b) fluctuation jamming.

Jamming without a constant component, the spectrum of which is in the radio frequency band is called high-frequency jamming (high-frequency noise). If the spectrum is adjoining the coordinate origin, the jamming is called low-frequency jamming (video noise).

If fluctuation jamming with an energy spectrum $G_{in}(f)$ is acting at the input to a linear system the energy spectrum at the output of this system will be found through the formula

$$G_{out}(f) = G_{in}(f) \cdot K^2(f) \tag{11.28}$$

where

$K(f)$ is the amplitude-frequency curve for the system.

When it is white noise that is acting at the input to a linear system, the energy spectrum of the noise at the output will coincide in shape with the square of the amplitude-frequency curve (fig. 11.21).

If fluctuation voltage is fed into the input of a nonlinear component, such as a detector, the energy spectrum will vary greatly in shape.

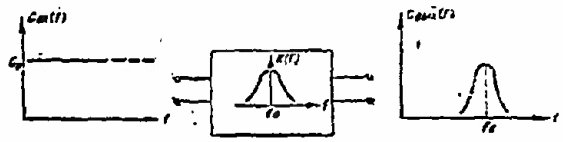


Figure 11.21. Conversion of a white noise spectrum into a linear selective system.

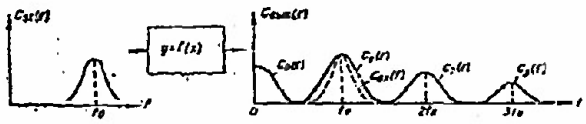


Figure 11.22. Energy spectrum of a narrow-band random process after conversion into a nonlinear system.

For example, if a narrow-band process, the spectrum of which is concentrated in a relatively narrow band of frequencies near high frequency f_o (an example of this process is noise at the output of a linear system, the band of which is much narrower than its resonant frequency) is acting at the input of a nonlinear system with $y = f(x)$, the output spectrum will consist of several parts (fig. 11.22):

a low-frequency part of the energy spectrum $G_o(f)$;

part of the energy spectrum located near the carrier frequency $G_1(f)$.

The other parts of the energy spectrum are located near the odd and even harmonics of frequency f_o.

From the standpoint of detection, the main interest is the low-frequency portion of the energy spectrum because the high-frequency component of the spectrum is usually filtered out in the detector load. The spectral band, $G_1(f)$, is the main interest in the case of modulation in radio transmitters.

The energy spectrum and the correlation function are not independent characteristics of a random process since both describe the process from the standpoint of the rate at which it occurs.

The energy spectrum and the correlation function of a random stationary process are related to each other through

$$R(\tau) = \int_0^\infty G(f) \cos 2\pi f \tau \, df, \tag{11.29}$$

$$G(f) = 4\int_0^\infty R(\tau)\cos 2\pi f\tau\, d\tau. \tag{11.30}$$

The correlation function of white noise equals

$$R(\tau) = \int_0^\infty G_0 \cos 2\pi f\tau\, df = \frac{G_0}{2}\delta(\tau), \tag{11.31}$$

that is, it is a delta function at the coordinate origin. By definition, the delta function $\delta(\tau)$ equals zero when $\tau \neq 0$ and is unlimited when $\tau = 0$.

$$\delta(\tau) = \begin{cases} 0, & \tau \neq 0 \\ \infty, & \tau = 0. \end{cases}$$

In this case the correlation function is instantaneously damped, indicating the complete independence of the instantaneous values when $\tau \neq 0$. The correlation function will be damped to varying degrees for the intermediate values of the width of the spectrum of random processes ($0<\Delta f<\infty$).

Figure 11.23 shows the correlation function of the noise voltage across the output of an amplifier with a bell-shaped frequency curve. The solid line is the correlation function for the case when the resonant frequency is $f_0 \neq 0$, and the dotted line for the case when $f_0 = 0$ and the system band pass equals $\Delta f/2$ (low frequency filter). Since the correlation function is even, the image of this function is only given for positive τ.

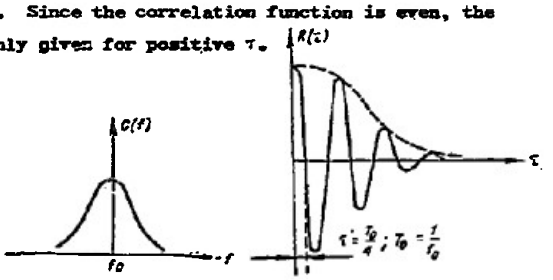


Figure 11.23. Energy spectrum and correlation function for fluctuation noise across the output of an amplifier with a bell-shaped frequency curve.

Thus, the correlation function contains traces of all the manifestations of a process fixing its frequency spectrum. Oscillations in the correlation function with predetermined frequency denote the average spectrum frequency. Attenuation in the oscillations characterizes the width of the fluctuation spectrum, and the shape of the envelope of the correlation function is related to the shape of the frequency spectrum.

Direct noise jamming is created by special jammers providing a comparatively high spectral power density over a broad band of frequencies and is used to create barrier jamming.

11.11. Noise Jamming with Amplitude, Frequency, and Amplitude-Frequency Modulations

Spot frequency noise jamming is generated by modulating high frequency generators with noise voltage. The modulation can be done in amplitude, frequency, or by both parameters simultaneously.

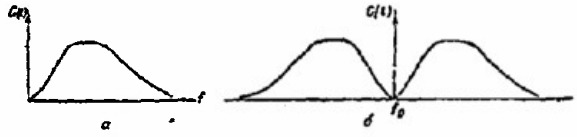


Figure 11.24. Energy spectrums of modulating noise (a) and of amplitude-modulated noise jamming in the case of linear modulation (b).

The energy spectrum of amplitude modulated noise jamming consists of a discrete component at the carrier frequency and two side bands, the shape of the envelope of which in the case of linear modulation coincides with the envelope of the low-frequency spectrums of the modulating noise (fig. 11.24). The side components of the spectrum of amplitude-modulated noise jamming has symmetrical amplitude and phase. Amplitude-modulated noise jamming is close to straight noise jamming in effectiveness. This type of jamming is widely used in the meter and decimeter wave bands.

Frequency-modulated noise jamming has been the type most widely used. In the case of the frequency modulation of the noise-jamming transmitter, the width and the nature of the jamming spectrum depend on the effective index of the frequency modulation and on the law for the probability of the distribution of the modulating noise.

The effective index of modulation can be found through the expression

$$m_{eff} = \frac{\Delta D_{eff}}{\Delta f_{width}},$$

where

ΔD_{eff} is the effective deviation (the mean square deviation) in the frequency;

Δf_{width} is the width of the spectrum of the modulating noise.

In the case of a linear frequency-modulation characteristic curve and sufficiently high modulation indexes ($m_{eff} \gg 1$), the envelope of the spectrum of modulated oscillations repeats the law for the probability of the distribution of modulating noises. The width of the spectrum is fixed by the dual

deviation of the frequency. If the modulation is done by noise with a normal distribution law, the envelope of the energy spectrum has a shape corresponding to the normal law (fig. 11.25). Amplitude limited noise is used to obtain a more uniform spectrum.

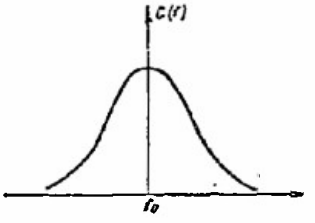


Figure 11.25. The energy spectrum of frequency-modulated noise jamming in the case of linear modulation by noise with a normal distribution law.

Combined frequency and amplitude modulation is the case in many generators using self-excitation. For example, in the case of frequency modulation of magnetrons and backward wave tubes additional amplitude modulation occurs because the generated power depends on the frequency.

The shape and width of the modulated oscillation spectrum desired can be obtained by selecting the coefficient of amplitude modulation, the index of frequency modulation, and the spectrum and structure of the modulating noises.

11.12. Methods of Generating Noise Jamming

Photoelectron multipliers, thyratrons in a magnetic field, and noise diodes are used in modern transmitters as sources for the modulation of noise voltage.

Photoelectron multipliers provide an adequately uniform spectrum of noise voltage up to frequencies of several tens of megahertz, with a noise intensity of approximately 10^3-10^4 microvolts/megahertz.

Thyratrons in a magnetic field have quite a wide noise spectrum (up to 6 to 7 megahertz). Their principal advantage is a high level of noise voltage.

The types of oscillator tubes used in jamming transmitters are selected according to the purpose for which the transmitter is intended, wave band, and power required.

Conventional oscillator tubes, triodes and tetrodes, are used in jamming transmitters for the shortwave and meter wave bands.

100 to 1500 watt, continuously oscillating magnetrons, are widely used in airborne and shipboard, as well as ground, jammers in the centimeter and decimeter bands.

A recent development is a more powerful tube, the barratron, which is a magnetron operating on unstable types of oscillations. The effective power of barratrons is approximately ten times greater than that of modern magnetrons. Barratrons have been developed in two versions; one with a fixed frequency for barrier jammers, and one with smooth retuning (electronic or mechanical) for spot jammers. Barratrons are used in the decimeter and centimeter wave bands.

Backward wave tubes have also been widely used. The tubes in use at the present time operate in the band from 400 to 200,000 MHz and can be quickly retuned electronically over a range of several tens of percent. Their power ratings range from a few watts to tens of kilowatts.

11.13. Characteristics of Noise Jamming

The effect of continuous straight noise jamming is similar to that created by receiver noise. This type of jamming forms a characteristic noise track on the screen of an indicator with an amplitude marker, and its image depends on the sweep rate and repetition frequency for a given receiver IF and video frequency pass band. A low amplitude useful signal can be seen in some cases as a darkening of the sweep line at the base of the signal against the noise background.

In PPIs the noise modulates the electron beam intensity. At the same time, the noise leaves traces in the form of bursts of random brilliance and size out of synchronism with the sweep. Noise reduces the brilliance interval, distorts the signal image, and masks it.

The manner in which noise jamming with amplitude, frequency, or amplitude-frequency noise modulation acts is, basically, the same as in the case of straight noise jamming. Amplitude-modulated noise jamming differs from straight noise because it contains a carrier frequency and symmetry with respect to amplitude and phase of the side components relative to the carrier frequency. As distinguished from the case of straight noise jamming, there will be no darkening of the image near the sweep when there is a signal caused by signal pulsation because of the beating of the carrier, which degrades signal observability.

In the case of frequency modulated noise jamming, frequency modulation is converted into amplitude modulation in the receiver because the IF amplifier's transmission coefficient is a function of the frequency, and an

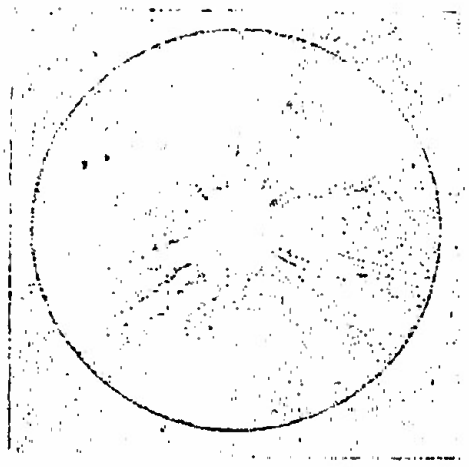


Figure 11.26. PPI scope showing continuous noise jamming.

Figure 11.27. Range scope showing continuous noise jamming.
(J - jamming; T - target pulse)

envelope of converted amplitude modulated oscillations will be separated out at the output of the amplitude discriminator.

If, in this case, the level of the noise jamming is high enough, the IF amplifier will be overloaded, and the result will be partial, or total, jamming of the useful signal, depending on jamming power. Figures 11.26 and 11.27 show the effects of continuous noise jamming on the receiver-indicator channel.

11.14. The Jamming Factor. The Countermeasures Equation

The concept of a jamming factor, K_j, will be introduced for masking type jamming. The jamming factor will be understood to mean the magnitude equal to the minimum ratio of jamming power to useful signal power at the receiver input at which the probability of a radar's carrying out its combat mission is reduced to a predetermined level.

The jamming factor depends on the type of radar being jammed (mainly on the availability of protection against jamming in the equipment), and on the form of jamming selected.

Straight noise jamming is the most effective, and in this case the jamming factor is at its least value.

Let us suppose the jamming is created by an on-board jammer.

Disregarding attenuation in the atmosphere, the reflected signal power at the receiver input is

$$P_s = \frac{P_{rad} g^2 \sigma_t \lambda^2}{(4\pi)^3 R^4}, \tag{11.32}$$

where

P_{rad} is the radar transmitter power;
g is the antenna gain for the radar;
σ_t is the target scatter area;
R is the target range.

Jammer power at the receiver input is

$$P_j = \frac{P_{jt} g_{jt} g \lambda^2}{(4\pi R)^2}, \tag{11.33}$$

where

P_{jt} is the jammer power;
g_{jt} is the jammer antenna gain.

Jamming will be effective if

$$P_j \geqslant K_j P_s. \tag{11.34}$$

This condition is called the countermeasures condition. By substituting the relationships at (11.33) and (11.32) in the equation at (11.34), we obtain the expression

$$P_{jt}g_{jt} \geq K_j P_{rad} g \sigma_t / 4\pi R^2 \qquad (11.35)$$

which is known as the countermeasures formula. The product of $P_{jt}g_{jt}$ is called the equivalent jammer power.

An analysis of these relationships indicates that as a target carrying a jammer approaches, the signal power at the receiver input increases very much more rapidly than does the jammer power. This can be explained by the fact that signal power at the receiver input is inversely proportional to the fourth power of the distance, but the jammer power is inversely proportional to the square of the distance. Therefore, at some range, called the minimum jammer operating range, R_{min}, the countermeasure condition at (11.34) ceases to prevail and the target begins to appear in the noise background (fig. 11.28).

This leads to another conclusion, namely, that if the target is using jamming for self-protection, a lower powered on-board jammer is needed for long-range radar jamming.

The absolute value of the jammer power at the receiver input increases as the jammer approaches the radar and, beginning at some range R_j, begins to overload the receiver-indicator channel. When this is the case, light-struck sectors, the sizes and number of which increase continuously as the receiver is overloaded by the jamming acting on the side lobes in the antenna pattern, form on the PPI screen. Beginning at some range, the indicator can be light-struck in a circular direction. If the receiver gain is reduced, the jamming will only effect the main lobe in the pattern. Only a narrow flare section will be observed in this case.

The jamming range is selected greater than the target radar range. If this is done the target radar will be jammed before the jammer comes within range of the radar.

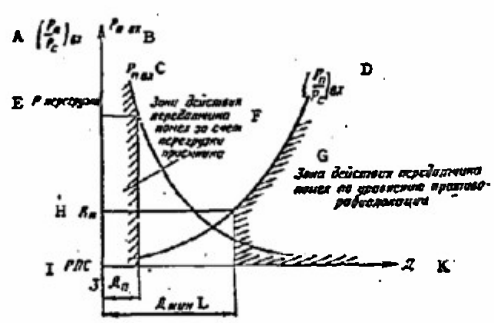


Figure 11.28. Superposed operating ranges for a jammer.

A - (P_j/P_s)in; B - $P_{j\ in}$; C - $P_{j\ in}$; D - $(P_j/P_s)_{in}$;
E - overload P; F - jammer range resulting from receiver overload; G - jammer range controlled by countermeasures; H - K_j; I - radar; J - R_j; K - R; L - R_{min}.

11.15. Parasitic Jamming

The application of parasitic jamming is based on the use of the principle involved in the secondary emission of radio waves. In the case of ground radar, this parasitic interference is currently generated by so-called "traps" (false targets), inflated and corner reflectors, and parasitic reflectors. In the case of airborne radars the jamming can be generated by using corner reflectors producing a mask for individual objects, imitating false objects (factories, bridges, railroad stations, and the like), or even changing the radar map of the terrain. Unorganized parasitic jamming results from reflections from local objects, thunderheads, rain, and snow. In the millimeter and centimeter bands, reflections of this kind can be very intense so that thunderheads, rain, and fog can greatly reduce radar range, or even hide whole areas within radar range from observers.

Parasitic reflectors to generate parasitic jamming are made in the form of dipoles of length $l = \lambda/2$ or $l \geqslant \lambda$. If a half-wave dipole is within the zone illuminated by a radar, the operating frequency of which coincides with the dipole's resonant (natural) frequency, the dipole will generate more intensive oscillations. The natural wave length of the dipole is approximately equal to twice its length. The length of the dipole is taken as something less than half the length of the wave illuminating it in order to obtain resonance, that is, a contraction factor is introduced

$$k = \frac{l}{\frac{\lambda}{2}} \approx 0{,}95,$$

where

l is the length of the dipole;

λ is the wave length.

Small deviations from the resonant frequency (±5 to 10%) do little to reduce the jamming effectiveness, so chaff of a single size is used to create jamming of stations the frequencies of which are 10% apart.

Parasitic reflectors can be made of aluminum foil, foil glued to paper, metallized paper, glass fiber, and glass plastic.

Chaff is dropped from aircraft in bundles, and the amount of this chaff per bundle is selected so that each bundle will offer a reflection equivalent to one that would be obtained from a target. The bundles are cut diagonally in order to increase the range.

Figure 11.29. PPI screen showing parasitic jamming by the corridor method.

If the target is to be hidden from radar observation, enough chaff must be released from the aircraft to provide an effective reflecting surface in the pulse volume equal to, or greater than the effective reflective surface offered by the target.

Dipole jamming today is generated by the discrete, or the corridor, method. In the first of these the bundles are dropped from the aircraft at predetermined time intervals. The reflections from these bundles look like real target pips on radar screens. Package life is short, however, because the packages quickly lose forward motion and scatter as they fall.

In the corridor method the aircraft providing the jamming drops chaff continuously to form a long cloud (a corridor) of jamming. Following aircraft can use this cloud to cover them so they can fly through undetected (fig. 11.29). The aircraft providing the jamming can also fire the packages of chaff ahead of it by using rockets.

Automatic devices that make drops at a previously set tempo are used to drop the bundles of chaff.

11.16. Countermeasure Masking

The purpose of countermeasure masking is to hide various ground targets, aerial and space guidance equipment, and military equipment from radar observation, as well as to create false target pips and objects on radar screens.

This masking can be done by using:

corner reflectors;

narrow-band interference radio-frequency absorbing coatings;

broad band radio-frequency absorbing coatings;

an object shape offering little reflection towards the radar.

A corner reflector consists of mutually perpendicular flat faces cut in triangles, sectors, and rectangles (fig. 11.30) and rigidly fastened together. It reflects electromagnetic energy falling on it in the opposite direction.

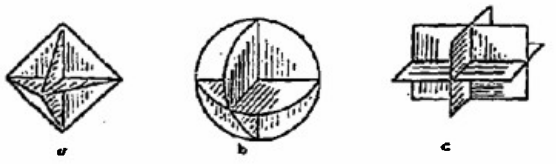


Figure 11.30. Shapes of corner reflectors.

a - with triangular faces; b - with sector faces;
c - with rectangular faces.

The maximum reflecting surface of a triangular reflector with triangular faces can be computed through the formula

$$\sigma_{\max} = \frac{4\pi a^4}{3\lambda^2},$$

where

a is the length of a rib of a corner face rib.

For the corner reflector with square faces

$$\sigma_{\max} = \frac{12\pi a^4}{\lambda^2}.$$

Corner reflectors are used to mask ground targets against radar observation, as well as to create false targets on the ground, in the water, and in space.

Anti-radar coatings. The use of anti-radar coatings reduces the effectiveness of target reflecting surfaces, that is, it attenuates the power of reflected signals. Radar range is thus reduced, and, by way of an example, a 12 db reduction in reflected signal strength results in halving the radar range.

The coatings are of two types; interference and absorbing.

A wave reflected from the external surface of the interference coating are compensated for by a wave reflected from the surface of the object. Coating thickness is selected equal to one-fourth the wave length, or to a

multiple of an odd number of quarter waves. Interference coatings can be made of rubber, mixed with carbonyl iron. The principal drawback is lack of band coverage.

There is no reflection of electromagnetic energy from the external surface of broad band absorbing coatings because of the predetermined structure of their material. The electromagnetic waves enter the coating and are completely damped in it. The materials used are diolectrics, or fibrous mats impregnated with a mixture of neoprene and conducting carbon black, iron fillings or chips in combination with ordinary fur, with layers of sheets of corrugated paper board, rubber mixed with carbon dust, and others.

The external surfaces of absorbing coatings are not made smooth, but are given a sort of relief of pyramidal protrusions (horns) covering the entire surface. This provides additional reduction in residual reflection of incident electromagnetic energy.

Absorbing coatings are applied in several layers.

The development of anti-radar coatings is following a path leading to a reduction in residual reflection, in reducing their weight and overall dimensions.

Anti-radar coatings can be used on ground, surface, submarine, air, and on space targets offering a radio frequency contrast.

Specifically, the effective reflecting surface of the warhead of a ballistic missile can be reduced by metal screens shaped to provide little reflection in the direction to the radar, or by materials absorbing radio waves.

These measures can reduce the detection range of a missile warhead by a factor of 2 or 3, so that target detection is only possible at the moment the warhead enters the dense layers of the atmosphere, when the coatings mentioned burn up, or are dropped with the nose cone.

False targets ("traps") are small guided missiles launched by conventional bombers beyond the range of radar detection. They form pips similar to those presented by real targets on radar screens so can draw active AA defense equipment away from the real targets. The missiles can be fitted with corner reflectors to increase their effective reflecting surfaces. These missiles can also carry equipment for creating active jamming of electronic equipment

11.17. Destruction of Radar Equipment

The means for destroying radar can be used very widely and this is in addition to jamming the radar and using anti-radar camouflage. Radio search, air and space search, military scouts and agents can all be used to establish radar coordinates.

Radar within short ranges of the front line can be destroyed by artillery and mortar fire. Radar installations deep in the rear can be destroyed by air strikes and missiles.

Aircraft use special homing devices to get at the radar installations.

The "air-ground" missiles used to destroy radar installations have homing guidance heads.

The United States has developed a special tactical missile of the "air-ground" class known as the "Shrike," designed to suppress AA radars. The missile's launching range is from 50 to 75 km.

The missile has a solid fuel engine and a passive guidance system that uses radar radiations.

United States Navy and United States Air Force aircraft are armed with "Shrike" missiles.

11.18. Methods Used to Protect Radar Against Jamming

The capacity of a radar to provide dependable detection and determination of target coordinates (reception of signals) despite jamming is called its antijamming ability (its immunity to jamming). The antijamming ability can be characterized by the ratio of signal power to jamming power for which the probability of detection and the accuracy in determination of target coordinates is no lower than may be required. It is obvious that the smaller this ratio, the greater the antijamming ability.

Antijamming devices are not all-purpose devices because each is only effective against a specific type of jamming. Hence it is obvious that the same radar has different antijamming ability with respect to different types of jamming depending on the protective devices used in the radar.

Organized and engineered measures comprise the protection given the radar against jamming.

Organized measures include operation on separated frequencies, covert operation of radar equipment, the training given operators in working through jamming, and the like.

All engineered measures can be divided into two groups.

The first group includes methods to protect against jamming, those that prevent the jamming from arriving at the receiver input:

spatial selection (narrowing the antenna pattern, reducing side lobe levels);

polarization selection (based on the difference in the polarized structure of signals and jamming);

frequency selection.

The second group includes methods and circuitry to protect against the jamming entering the receiver:

those aimed at preventing overloading the receiving channel;

those protection circuits based on the differences in the structure of useful signals and jamming (difference in carrier frequency, in spectral composition, in phase, amplitude, width, and repetition frequency of the pulses).

Among the engineered measures used to provide protection against jamming is an increase in the radar's energy potential.

11.19. Preventing Overloading of the Receiving Channel

Protecting the receiver against overloading by reflections from local subjects, as well as by active jamming, means increasing the receiver's dynamic range. The characteristics of some of the methods used to prevent overloading the receiving channel are given below.

Expansion of the dynamic range of the IF amplifier stages. Receiver overloading occurs first of all in the last stages of the IF amplifier. The dynamic range of the IF amplifier can be expanded by using more powerful tubes, and by supplying higher plate and screen voltages.

The gain in the IF amplifier channel can be reduced and that in its video channel can be increased in order to expand the dynamic range. This has limited possibilities because the voltage across the detector input has to be on the order of one volt in order to ensure a linear detection mode.

Instantaneous automatic gain control. Instantaneous automatic gain control makes it possible to retain a fixed gain for the receiver for short-duration useful signals, and considerably reduced the gain in the case of lengthy high amplitude jamming. The operating principle of instantaneous automatic gain control is as follows. When high amplitude jamming appears across the grid of the regulated stage in the IF amplifier there is an additional shift, ΔE_g, with a magnitude approximately equal to the jamming amplitude, $U_{mj\ in}$. This prevents suppression of the signal, which "sweeps" over the linear section of the tube's characteristic curve (fig. 11.31).

Figure 11.32 shows an IF amplifier stage that is part of an instantaneous automatic gain control circuit. The voltage from the IF amplifier stage (the tube T_1) is fed into the detector. The rectified voltage of negative polarity is fed to the grid of the IF amplifier stage tube through the cathode follower (tube T_2). The time constant for the instantaneous automatic gain control filter, primarily established by the RC circuit, is selected several times larger than the signal pulse duration. Delay voltage E_d is

fed into the instantaneous automatic gain control detector, otherwise the detector will begin to function on weak voltages. The magnitude of resistance R_k establishes the selection of the operating point on the tube's characteristic curve in the original condition.

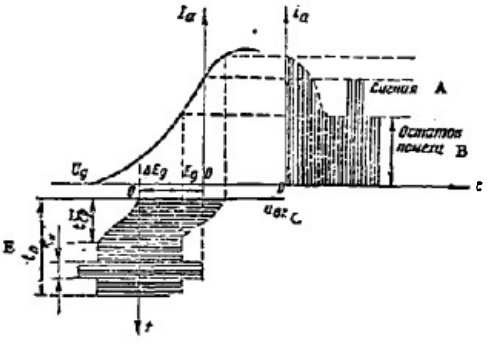


Figure 11.31. Strong pulse jamming with duration t_j at the input and output of a stage in the instantaneous automatic gain control (t_{op} is the instantaneous automatic gain control time of operation).
A - signal; B - remainder of jamming; C - u_{in}; D - t_{av}; E - t_j;

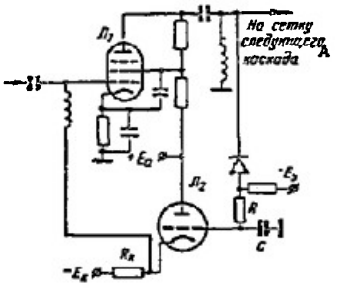


Figure 11.32. Amplifier stage, tube T_1, part of the instantaneous automatic gain control circuit. A - to grid of next stage.

A differentiating circuit with a small time constant is sometimes inserted between the detector and the video amplifier in order to suppress the long jamming pulses remaining after the instantaneous automatic gain control.

Use of an IF stage with a linear logarithmic amplitude response curve. The amplitude response curve for a conventional receiver looks like the one shown in Figure 11.33. A receiver such as this is readily overloaded, that is, after the input voltage has reached some value, $U_{in\ O}$, a further increase in that value will not result in any increase in the output voltage.

If the receiver is to function normally in the face of high powered jamming, it is desirable to have an amplitude response curve that is linear for weak, and logarithmic for stronger, signals (fig. 11.34).

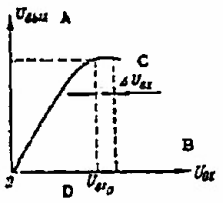


Figure 11.33. Amplitude response curve for a conventional receiver. A - U_{out}; B - U_{in}; C - ΔU_{in}; D - $U_{in\ O}$.

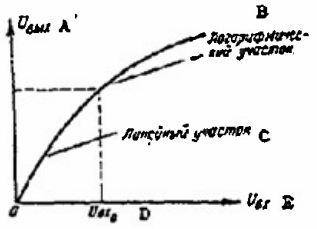


Figure 11.34. Linear-logarithmic amplitude response curve. A - U_{out}; B - logarithmic section; C - linear section; D - $U_{in\ O}$; E - U_{in}.

A curve such as this can be obtained by the use of feedback, by the delivery of negative voltage to the tube with variable steepness, by shunting the plate loads with a non-linear component (by a diode), and by using other methods.

Automatic time gain control. In order to avoid overloading the receiver when powerful pulses from the oscillator and powerful reflected signals from the targets close to the radar arrive directly across the receiver input, it is necessary to block the receiver during the time the oscillator is transmitting pulses, and to cause it to operate at reduced gain for a few microseconds after each pulse has terminated. This is done by using an automatic

time gain control. Figure 11.35 is one possible automatic time gain control circuit. Tube T_1 is shut down by the voltage from source E_1 in its original condition. When the main pulse is radiated the tube opens up and condenser C rapidly charges to voltage $U_1 = I_1R_2$, the magnitude of which can be adjusted by resistance R_2. When the pulse is over tube T_1 is blocked and condenser C discharges through resistances R_4 and R_5. A negative voltage, exponential in shape, is supplied to the grid of one or two IF amplifier stages from condenser C through the cathode follower (tube T_2). Resistance R_4 can be used to change the rate of discharge of condenser C, and R_5 can be used to change the initial voltage across the capacitance. Figure 11.36 is the curve for the change in the regulating voltage in the automatic time gain control circuit. An effort can be made to attain a situation such that the receiver will generally be blocked for some period of time after the main pulse has been radiated.

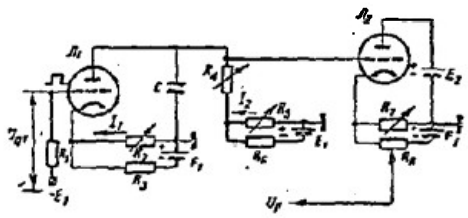


Figure 11.35. Automatic time gain control circuit.

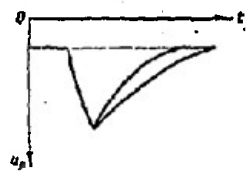


Figure 11.36. Curve for change in regulation voltage across the output of an automatic time gain control circuit for different time constants for the discharge of condenser C.

11.20. Antijamming Methods Based on Signal Selection

Selection is understood to mean the separation of the signal from the jamming using its known parameters (width, pulse interval, pulse amplitude, and carrier frequency and phase). Let us take up some of the methods that can be used to select signals.

Frequency selection is based on the use of the selective properties of the IF amplifier, and of various types of filters.

Rejection filters, used to separate out of the receiver pass band the section of frequencies containing the jamming, will suppress the narrow band jamming (unmodulated jamming, or jamming with very simple types of modulation). An example of a circuit with a rejection filter is shown as Figure 11.37.

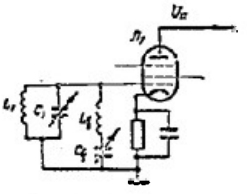


Figure 11.37. Rejection filter circuit (L_1, C_1 - circuit tuned to the intermediate frequency; L_f, C_f - rejection filter).

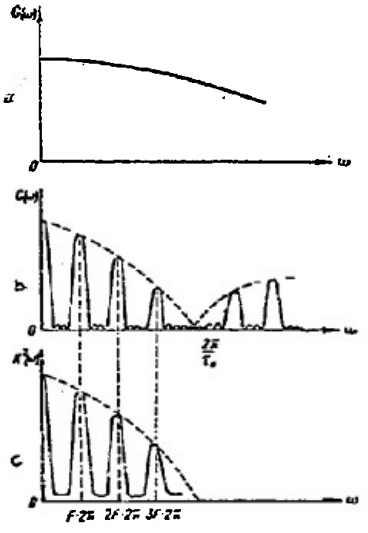


Figure 11.38. (a) Noise energy spectrum density; (b) spectrum for a clipped video pulse train; (c) square of the amplitude-frequency response curve for a comb filter.

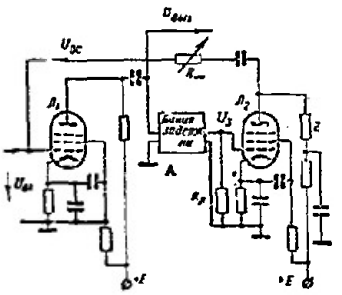


Figure 11.39. Simplified schematic of an amplifier with an amplitude-frequency response curve close to the comb filter curve. A - delay line

So-called comb filters can be used to separate a pulse signal from noise modulated jamming. Figure 11.38 presents the energy spectrum for noise, the spectrum for a clipped pulse train with repetition frequency F_p, and the square, or the amplitude-frequency, response curve for a comb filter. As will be seen from that figure, the comb filter separates the signal spectrum components, and suppresses the jamming spectrum components.

The filter should pass the basic portion of the signal spectrum if the result is to be an undistorted signal. In the case of an overall filter pass band $\Delta f_f = 1/\tau_p$, and of an elementary band for passing one spectral component, Δf_1, the increase in the signal/noise/voltage ratio will be approximately equal to $\Delta f_f/n \cdot \Delta f_1$, where n is the total number of signal spectrum components that must be passed.

A feedback amplifier circuit will provide an amplitude-frequency response curve close to the comb filter curve. Figure 11.39 is a simplified circuit of an amplifier of this type. The load on tube T_1 is the delay line. Delay time, T_d, is equal to the pulse sequence period. The line load is resistance R_{load}, equal to the line's characteristic impedance. The tube T_2 stage amplifies oscillations U_d, tapped from the line. Load Z in its plate circuit can be resonant, or aperiodic (depending on the nature of the amplified oscillations). The feedback voltage, U_{fb}, is fed into the first stage input from tube T_2 output. Resistance R regulates the magnitude of the feedback. The output voltage is picked off at the plate of tube T_1.

The comb filter is also an antijamming device for use against unsynchronized pulse jamming.

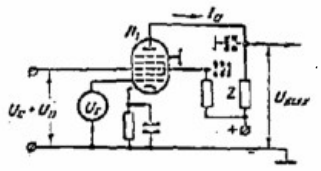


Figure 11.40. Heptode synchronous detector circuit.

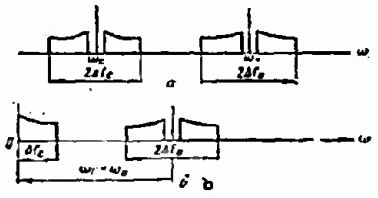


Figure 11.41. Signal and noise spectrums at input (a) and output (b) of a synchronous detector.

Phase selection is based on separating the signal from the noise by using its known phase, as distinguished from the phase of the noise. Phase selection circuits are based on the use of the synchronous detection method. Figure 11.40 is a heptode synchronous detector circuit.

The summed voltage of signal and noise at frequencies ω_s and ω_j respectively is fed to the signal grid of tube T_1. Voltage from the oscillator, U_o, the frequency of which is equal to the signal frequency and is rigidly synchronized with it, is fed to the oscillator grid. The result is to cause the synchronous detector for signal voltage to act like a detector, and as a converter for the jamming voltage. The result of the detection is to shift the signal and jamming spectrums by the magnitude ω_o (fig. 11.41). A low frequency filter is put in the synchronous detector output to separate the signal and suppress the jamming.

Since at low frequency one can find a filter with quite a narrow pass band and very steep slopes in its resonance curve, the use of a synchronous detector makes possible quite a suppression of the jamming, if its spectrum does not intersect the signal spectrum.

A direct signal from the transmitter can be used to synchronize the oscillator.

Synchronous detection, combined with compensation circuits, provides protection against organized passive jamming and reflections from local objects.

The accumulation method is based on the use of the property of periodicity of the signal in terms of time to distinguish the signal from the jamming. It is equivalent to using the comb filter method, but considered from the time point of view. The engineering is quite simple, and has found widespread use in radar. Signals reflected from a target are a clipped periodic train of pulses. There is no regularity in the case of noise voltage, which is a random process.

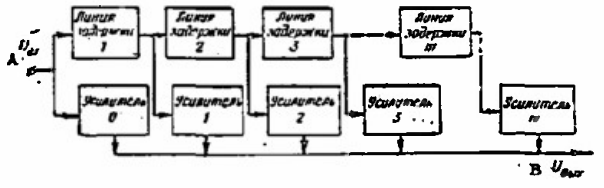


Figure 11.42. A possible block schematic for accumulation of signals by delay lines (adder circuit).
A - U_{in}; B - U_{out} (top all delay line, bottom all amplifier).

Signal order selection is used in accumulation circuits. One such block schematic of an accumulator is shown in Figure 11.42. It consists of amplifiers, the inputs of which are fed with voltages from the delay lines. Each line has a delay equal to the period of pulse repetition, T_p. All amplifiers have a common output, across which the summed voltage is picked off. This circuit has a number of accumulation cycles equal to the number of delay lines (m = 1, 2, 3...). The signal across the output of the nth amplifier appears with a time delay of nT_p relative to the signal fed to the receiver input (n = 0, 1, 2, 3... - the number of amplifiers).

Let us suppose a signal, together with the noise, is fed from the receiver output to the accumulation circuit. At the moment the mth pulse in the train arrives the accumulation circuit sums the signal and noise for m periods in the sequence. At the same time the signal voltage increases m times, and the noise voltage, because of its random nature, only increases by the $\sqrt{m}$. The signal/noise ratio increases by a factor of m in power across the accumulator output.

Given a sufficiently large number of accumulation cycles, accumulation circuits will separate very weak signals out of a background of strong jamming. The drawback in these circuits is the long integration time.

Used as accumulators in radar are cathode ray tubes with persistence, with charge accumulation, with dark recording, integrators of various types, delay lines, as well as the lag (memory) in the operator's vision.

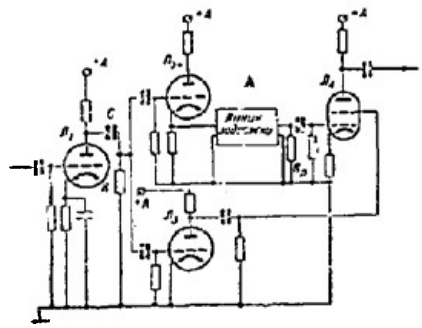


Figure 11.43. Pulse width selection circuit.
A - delay line.

Spectral analysis of the operation of accumulators demonstrates that their amplitude-frequency curve has the shape of that for a comb filter, and the greater the accumulation cycles for the accumulator used, the closer its frequency curve will approach the ideal comb filter.

Amplitude and pulse width selection. Pulse jamming differs from useful signals in its structure and frequency spectrum. Amplitude and pulse width selection is used to cope with it. The difference in target and jamming pulse widths is used in the pulse width selection circuit (fig. 11.43). The incoming signal with τ_1 is fed to the RC differentiating circuit from the output of the video amplifier in the receiver (tube T_1) The differentiated signal is fed into two channels. The first of these channels consists of the cathode follower (tube T_2) which only passes positive pulses, delay line for a time delay of t_d equal to the width of the radiated pulse, and the coincidence circuit (tube T_4). The second channel consists of an amplifier, tube T_3, operating in the grid-circuit clipping mode, and a coincidence stage, T_4. Positive pulses are fed to both grids in the tube in the coincidence stage at the moment in time corresponding to the end of the pulse. Tube T_4 is triggered, and the signal is picked off its plate load. The voltage curves for this case are shown in Figure 11.44a. If the width of the input signal is longer than the delay time in the line ($\tau_2 > t_d$), the positive pulses are not fed into the coincidence stage simultaneously. Tube T_4 will not be triggered and no signal will appear across its output. The voltage curves for this case are shown in Figure 11.44b.

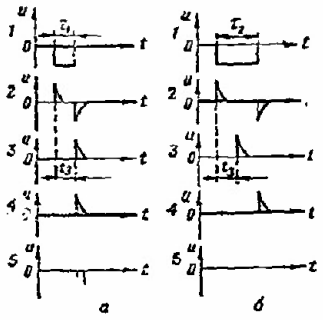


Figure 11.44. Voltage curves for a pulse width selection circuit.

a - widths of received and radiated pulses equal; b - width of received pulse longer than that of radiated pulse.
1 - input signal; 2 - signal from differentiating circuit; 3 - signal across control grid of tube in coincidence stage; 4 - signal across screen grid of tube in coincidence stage; 5 - signal across the plate of the tube in the coincidence stage.

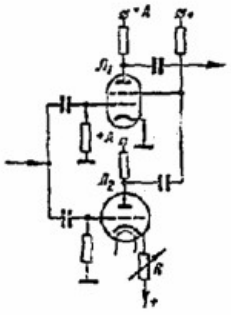


Figure 11.45. Amplitude selector circuit.

A possible circuit that can be used as an amplitude selector is shown in Figure 11.45. The incoming positive signal is fed simultaneously to the control grid of tube T_1 in the coincidence stage, and to the amplitude clipper, T_2. If the jamming signal exceeds the level of clipping of the amplitude selector, fixed by resistance R, there is a negative pulse fed to the screen grid in the tube in the coincidence stage from the plate circuit

of tube T_2. Tube T_1 is not triggered by the incoming pulse and there is no jamming at the receiver output. The useful signal, which does not exceed in amplitude the clipping level, triggers tube T_1 and passes freely to the receiver output.

11.21. Correlation Reception

The correlation method of separating the signal from the noise is based on the use of the properties of the correlation function. Let a sinusoidal signal and fluctuating (noise) jamming be acting at the receiver input

$$u_{in}(t) = u_S(t) + u_n(t) = U_m \sin(\omega_0 t + \varphi) + u_n(t).$$

The autocorrelation function of the input process is

$$R(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_0^T [u_S(t) + u_n(t)]\cdot[u_S(t+\tau) + u_n(t+\tau)]\,dt = R_s(\tau) + R_n(\tau) + R_{sn}(\tau) + R_{ns}(\tau). \quad (11.36)$$

where

$R_s(\tau)$ and $R_n(\tau)$ are noise and signal autocorrelation functions;

$R_{sn}(\tau)$ and $R_{ns}(\tau)$ are the mutual correlation functions of noise and signal.

Since signal and noise are independent functions,

$$R_{ns}(\tau) = R_{ns}(\tau) = 0.$$

Then

$$R(\tau) = R_s(\tau) + R_n(\tau).$$

The autocorrelation function of a periodic signal is periodic. For a sinusoidal signal

$$R_s(\tau) = \frac{U_m^2}{2}\cos\omega\tau.$$

The autocorrelation function of the noise is non-periodic, and diminishes with increase in the argument τ.

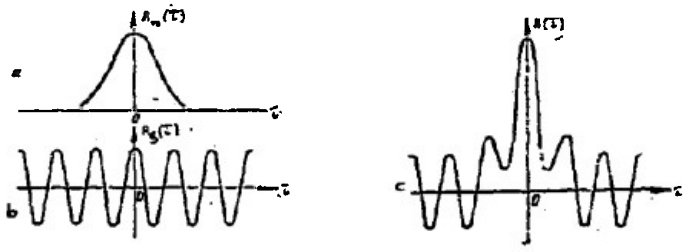


Figure 11.46. Correlation functions:

a - noise; b - periodic signal; c - sum of periodic signal and noise.

Figure 11.46 depicts the correlation functions of noise, a periodic signal, and the sum of signal and noise.

If the receiving device includes a correlator (delay circuit, multiplier, integrator, and recorder) the periodic component will be separated when there is a signal across the output of the correlator. There will be no signal if, by virtue of an increase in τ, the values of the correlation function tend to a constant (if there is no constant noise component, to zero).

Figure 11.47 is the block schematic of a correlation receiver. Two trains of short pulses, shifted with respect to each other by a circuit with a variable delay in time τ, are fed from the pulse generator to the selecting device. The pulse repetition period, T_p, is selected as the integration interval. The selecting device separates n pairs of discrete values for the input process a_i, b_i, for each value of τ. The pairs of discretes are divided by time interval τ (fig. 11.48). The number n fixes the observation time, t_o, since

$$t_o = n \cdot T_p.$$

The approximate value of the correlation function obtained across the correlator output is

$$R(\tau) \doteq \frac{1}{n}\sum_{i=1}^{n} a_i b_i,$$

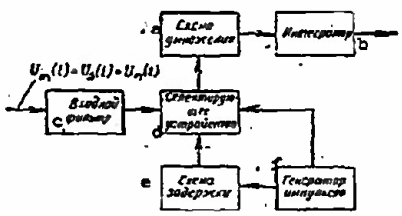


Figure 11.47. Block schematic of a correlation receiver.
a - multiplier circuit; b - integrator; c - input filter; d - selector; e - delay circuit; f - pulse generator.

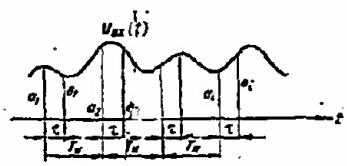


Figure 11.48. Separation of discrete values from the envelope of the input process.

Theoretically, correlation reception permits detection of signals regardless of the signal/noise ratio, but a long observation time is required to do so ($n \to \infty$).

When signal/noise ratios at the receiver input are small, a long correlation interval τ is needed in order to separate $R_s(\tau)$. But the correlation interval cannot be taken as longer than the pulse width, τ_p, in the case of pulse radar, because

$$R_s(\tau) = 0 \text{ when } \tau > \tau_p.$$

Consequently, we use the correlation reception method in radar with continuous radiation and in pulse radars with long pulse widths. The literature on the subject notes that the use of a correlation receiver in centimeter band radar, in which pulse widths are 1 to 2 microseconds, presents known difficulties.

11.22. Other Antijamming Methods

Shift in Radar Working Frequencies

Shifting the operation of a radar to another wavelength is an effective method of getting away from active jamming, the spectrum of which falls in the receiver's band, or intersects it. However, this method runs into extensive technical difficulties, resulting from the fact that in the centimeter wave bands there are no magnetrons and klystrons that will function dependably over a broad band of frequencies. Replaceable oscillators, and high frequency heads for receivers, are used to increase band width.

Retuning is adequate over a sufficiently broad band of frequencies in the decimeter and meter bands.

Continuous change in radar working frequencies

An effective method for protecting against active jamming is continuously changing the working frequency of the radar in accordance with some law. A "sliding" working frequency makes it very difficult to detect the operation of a radar and to analyze the signal radiated by the radar, and thus generate the active jamming to hamper the radar's operation.

Receivers with changing tuning usually have quick-acting automatic frequency control systems.

Coding a radar signal

Radar signal coding is sometimes used to make it difficult to generate active jamming. The number of pulses, and the time intervals between them, can be changed automatically during operation, thus making it even more difficult to jam.

The drawback in coding a radar signal is that the range resolution of the radar is reduced.

The pulse compression method

Radar range depends on the pulse energy

$$W_p = P_p \tau_p,$$

where

P_p is the pulse power.

Consequently, the long pulse radiation mode can greatly improve the radar's immunity to jamming. But at the same time the pulse frequency must be modulated in accordance with a predetermined law in order to obtain good range resolution. The incoming signal is compressed by an optimum filter. At the same time the pulse energy can be retained and thus lead to an increase in signal power at the filter output and to an increase in resolution.

Chapter I reviewed the pulse-frequency method of radar operation in detail.

11.23. Coping with Mutual Jamming

Acting on the radar in addition to organized jamming is the interference created by neighboring radars, or other equipment operating in the same band.

If such interference is to be avoided, working frequencies must, in so far as this is possible, be separated, positions must be properly selected, and the working sectors for the individual radars established. Particular attention is given to the locations of radars operating on the same wavelength. These radars must be set up at definite distances from each other and working sectors assigned such that the transmitting antenna of one is not aimed into the receiving antenna of a nearby radar.

Pulse width and amplitude selection circuits, as well as radar retuning to another working frequency, can be used to cope with mutual interference.

It is desirable to change the repetition frequency to cope with synchronous interference caused by nearby radars with the same pulse repetition frequencies, or a repetition frequencies, or a repetition frequency close to that used by the radar near by.

Chapter XII

Automation of the Gathering and Processing of Radar Information

12.1. Fundamental Concepts

Gathering and processing of radar information is the process of obtaining possible evidence of targets located in the range of vision of a radar station.

The physical basis for obtaining this evidence is the radar station signals, including the following:

signals from the output of the radio receiver;

triggering pulse from the radar station transmitter which determines the beginning of the time delay count of the reflected signal;

signals characterizing the spatial position of the radar station antenna (signals of the azimuth and elevation sensors of the antenna system).

An indication is obtained from processing the radar information. This indication is a set of all possible information about the target: coordinates, height, velocity, bearing, location time. In addition, a series of other characteristics may be included in the indication, such as nationality, number of targets, importance, type, authenticity, etc. On plan position indicators, the indication is an illuminated point which carries information on the presence of a target and its coordinates. In automatic gathering and processing, the indication may be represented by the amount of voltage, current, or rotation angle of a selsyn, and, in units based on computer technology, it is represented in binary numbers. In this case, the indication may include manifold information.

Signals which carry the information required for the observer (operator) are called useful signals. All those sources which distort the useful signals and disturb the target information are interference. They include noise in the receiving channel and various artificial active and passive interferences. Furthermore, automatic information processing can produce other interference factors, such as errors in measuring coordinates, reductions in the transmission channels, disturbances in the function of specific elements of the processing system, etc. As a rule, all useful signals are accompanied by interference, so that the problem of separating useful signal from interference must be dealt with in the processing stage. The deleterious effect of interference cannot be completely eliminated, so that the results of processing always contain errors.

Information processing may be semiautomatic, where the human operator takes part and carries out certain functions, or completely automatic, where

all functions are carried out without human participation.

Radar information processing is divided into three stages: primary processing, secondary, and tertiary.

12.2. Receiver Output Signal

In the general case, the signal $u(t)$ at the output of the receiver is a mixture of useful signal $u_s(t)$ and interference $u_i(t)$. The mixture may be represented by the sum

$$u(t) = u_s(t) + u_i(t).$$

Interference makes the output signal become essentially random, and therefore it can only be described statistically. When the useful signal does not fluctuate and interference conforms to the normal law, distribution of the amplitude of the sum "signal + interference" at the output of the receiver channel may be described by the generalized Rayleigh law, whose distribution density takes the form

$$w_{si}(x) = xe^{-\frac{x^2+v^2}{2}} I_0(vx), \tag{12.1}$$

where $x = \frac{u(t)}{\sigma_i}$ is the normalized output signal;

$u(t)$ is the signal at the receiver output;

σ_i is mean square noise;

v is the signal/noise ratio;

$I_0(vx)$ is a modified Bessel function of the first kind, zero order on the argument vx.

The expression for the probability density of pure noise $w_i(x)$ is found from formula (12.1) if $v = 0$ (i.e., there is no useful signal present):

$$w_i(x) = xe^{-\frac{x^2}{2}} \tag{12.2}$$

Graphs of the function $w_{si}(x)$ for various signal/noise ratios (including $v = 0$) and the Bessel function $I_0(vx)$ are shown in Fig. 12.1 and 12.2. It is evident that the distributions of the amplitudes of pure noise and the total differ more, the greater the signal/noise ratio.

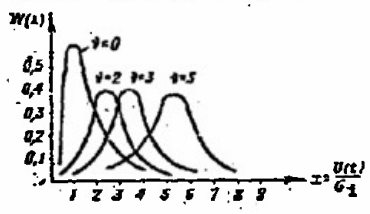


Figure 12.1. Graphs of probability density of generalized Rayleigh law

for amplitudes of the sum "signal + noise". Probability density for pure noise corresponds to the curve v = 0.

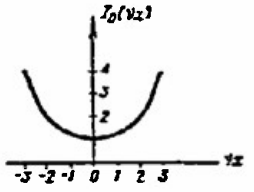


Figure 12.2. Graph of the modified Bessel function of the first kind, zero order.

12.3 Primary Processing of Radar Information

Primary processing is the first step in receiving target information. At this stage, the following functions are fulfilled:

detecting the useful signal and deciding on the presence or absence of a target in the observed region;

measuring the target coordinates;

converting the received information into a form suitable for further processing and transmitting it along the communication channels; this process is called coding.

The input signals involved in the primary processing operation are: the signal from the receiver channel output (in particular from the detector), triggering pulse for the radar, and signals from the angular coordinate sensors. Primary processing is accomplished in one or several successive range scans; in principle, this is sufficient to detect and measure the target coordinates. With regard to plan position indicators, primary processing may be taken to mean processing during one scan.

12.4. Principles of Automatic Target Detection

Devices for automatic target detection belong to a class of decision devices which, as a result of processing the signal from the receiver channel, decide "target" or "no target". When the sum "signal + noise" is observed at the output of the receiver channel, the decision should be "target"; in the case of pure noise, the solution should be "no target". Thus, the device should be capable of distinguishing pure noise from noise + signal. However, because of the random nature of these two signals, both noise + signal and noise may take on various values (see Fig. 12.1). But on the average,

amplitudes of the sum noise + signal will be higher than the average value of noise, so that in tha statistical sense there will be a definite difference between them.

One of the variations of the procedure for producing a decision consists of the following. In the region of the signal x(t) observed at the receiver output, a certain threshold (limit) x_0 (Fig. 12.3) is established. All signals exceeding the threshold x_0 are considered noise + signal, the device signals "target." Signals whose amplitudes are lower than the x_0 level are considered noise, and the device signals "no target"; i.e., if $x(t) \geq x_0$, then "target"; if $x(t) < x_0$, then "no target".

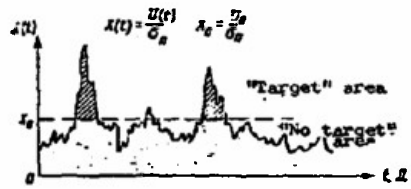


Figure 12.3. Detection of a useful signal by establishing a threshold x_0.

The functional diagram of a suitable resolving device and oscillograms of the input signal x(t) and the resolution signal at the output circuit are shown in Fig. 12.4.

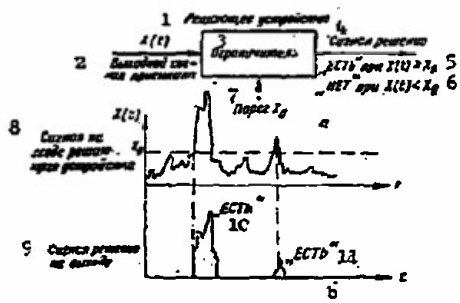


Figure 12.4. Functional diagram of a resolving device for detecting a target (a) and oscillograms of the input and output signals (b).

1 - Resolution device; 2 - Receiver input signal; 3 - Limiter; 4 - Resolution signal; 5 - "YES" at $x(t) > x_0$; 6 - "NO" at $x(t) < x_0$; 7 - Threshold x_0; 8 - Resolver input signal; 9 - Resolution signal at the output; 10 - "YES"; 11 - "NO".

A diode limiter (Fig. 12.5) may be used as a resolver. The required limiting level (threshold) x_0 is established by potentiometer R_0, and the resolution signal is taken from resistor R_r.

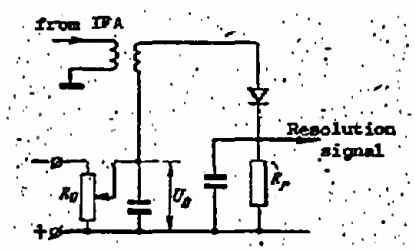


Figure 12.5. Diode limiter circuit

12.5 Detection Errors

Noise is a random process; thus some spikes of noise signal may be so large as to exceed the limiting level, and the device will signal "target" when there actually is no target. Such an erroneous signal is called a "false alarm", and the probability of a false alarm is qualitatively designated P_{fa}.

A similar error may be observed for the sum "signal + noise" when fluctuations and noise combine to make the sum signal so small that it falls below the x_0 level, and despite the actual presence of a target, the resolver will show "no target". This type of error is called target omission, and the probability of target omission is designated P_{to}.

The probabilities of erroneous resolutions may be calculated using the equations for probability density (12.1) or its graph. Figure 12.6 shows the graph of the probability density for noise and for the sum with $v = 3$ with range scan oscillograms at the same $v = 3$. The dotted line shows the chosen threshold x_0. The probability of false alarm, i.e., the probability that noise will exceed the limiting level x_0 is calculated from

$$P_{fa} = \int_{x_0}^{\infty} w_i(x)dx \tag{12.3}$$

Graphically P_{fa} appears as the area under the curve of noise probability density ($v = 0$), shifted to the right from x_0.

The probability of omission

$$P_{to} = \int_{0}^{x_0} w_{si}(x)dx \tag{12.4}$$

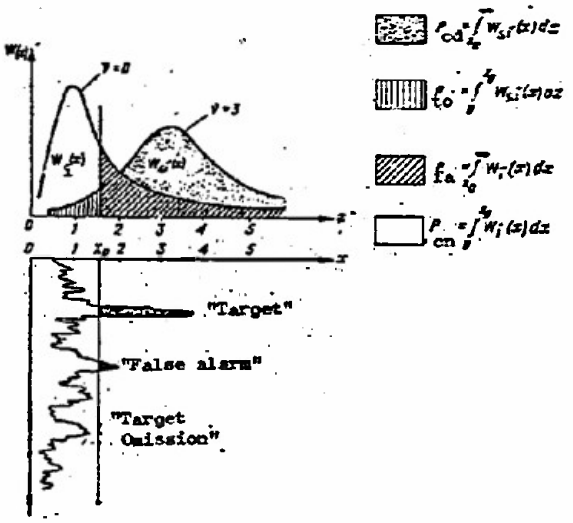


Figure 12.6. Graph of probability density for noise and for the sum at $v = 3$ with oscillograms of range sweep. The regions for calculating probabilities of correct and erroneous resolutions are shown on the graph.

This is the probability that the sum signal will not exceed the x_0 level. On the graph, P_{to} represents the area under the curve $w_{si}(x)$ in the interval between 0 and x_0.

Correct resolution is estimated as the probability of correct detection P_{cd} and the probability of correct nondetection P_{cn}.

These probabilities are given by

$$P_{cd} = 1 - P_{to} = \int_{x_0}^{\infty} w_{si}(x)dx, \tag{12.5}$$

$$P_{cn} = 1 - P_{fa} = \int_{0}^{x_0} w_i(x)dx. \tag{12.6}$$

12.6 The Concept of Optimum Limiting Level and Criteria for Determining Optimum

Turning to Fig. 12.6, it is evident that the probability of a false alarm decreases with an increase in the limiting level x_0 (shifting the dotted line to the right), but along with this, the probability of detection decreases, which is of course undesirable. If the limiting level x_0 is decreased, the

probability of detection increases, but the probability of false alarms also increases, which is also undesirable. Thus there are contradictions which may be resolved by choosing an optimum limiting level x_{opt}.

To find the optimum, the criteria of an optimum must be understood, that is, the measure used to determine the best or worst limiting level. Criteria from mathematical statistics are used for this purpose and for a large class of optimization problems. In the theory of signal detection, the most commonly used are the following.

Weight criterion. Here it is assumed that the device will be the better the greater the difference between the probability of detection and the probability of a false alarm, taking into account its importance (weight) in the resolution. Thus, of all possible devices with different thresholds the one is chosen for which the difference has its maximum:

$$P_{cd} - tP_{fa} = \max \tag{12.7}$$

where t is a weight coefficient. It is chosen on the basis of the undesirability or danger of false alarms.

Neyman-Pearson criterion. This criterion requires that the allowable level of the probability of false alarm $P_{fa\ all}$ be designated. Then the device will be considered best which will give the maximum probability of correct detection under the condition that the probability of false alarms will not exceed the allowable amount:

$$P_{cd} = \max \text{ for } P_{fa} \leq P_{fa\ all} \tag{12.8}$$

The allowable level of false alarms is set, as in the preceeding case, on the basis of the importance and undesirability of false alarms.

Ideal observer criterion. This criterion is feasible if information on the presence or absence of a target in the scan area covered by the radar is known or can be obtained in some way up to the beginning of detection. Such information is expressed by the apriori probability of the presence of a target p and the apriori probability of the absence of a target $q = 1-p$. Then the probability of accepting an erroneous resolution on the presence of a target is expressed by the product qP_{fa}, and the probability of another erroneous resolution (target omission) will be equal to pP_{to}. According to the given criterion, the best detector will be the one which gives a minimum sum of the probabilities of erroneous resolutions:

$$qP_{fa} + pP_{to} = \min \tag{12.9}$$

Minimum average risk criterion (Bayes criterion). Here not only the a priori probabilities, but also the "payment for risk" (cost of an erroneous resolution) must be known. If the "payment" for false alarm is designated C_{fa} and the "cost" of omission is designated C_{to}, then in generating a large

number of detection operations, the "payment" for false alarm on the average will be equal to $C_{fa}qP_{fa}$, and payment for omission will be $C_{to}pP_{to}$. Of course, the best detection system will be that which requires the least average payment for these and other errors, i.e.

$$C_{fa}qP_{fa} + C_{to}pP_{to} = \min \tag{12.10}$$

The meaning of payment for risk is a difficult factor and should be considered a statement of how dangerous or undesirable a particular error is.

The criterion of minimum average risk is the most common, and all preceding criteria may be found as particular cases of it.

Probability ratio criterion. The probability ratio is defined as the value $w_{si}(x)/w_i(x)$, which is the ratio of the probability densities of one and the same signal x under the condition that it is created by a noise + signal sum or only by noise. The probability ratio criterion is a corollary of the preceding criteria and has the following expressions:

$$\frac{w_{si}(x)}{w_i(x)} \geqslant 1, \tag{12.11}$$

$$\frac{w_{si}(x)}{w_i(x)} \geqslant \frac{q}{p}, \tag{12.12}$$

$$\frac{w_{si}(x)}{w_i(x)} \geqslant \frac{C_{fa}q}{C_{to}P}. \tag{12.13}$$

These ratios are found from Eqs. 12.7, 12.9, 12.10, if probabilities from Eqs. 12.3, 12.4, and 12.5 are substituted into them and the conditions of maximum or minimum are found.

All these criteria are used for analyzing information processing at various stages (primary, secondary, and tertiary). The most convenient criteria for practical calculations are the Newmann-Pearson and weight criteria.

12.7. Finding the Optimum Limiting Level (Threshold)

To find the optimum limiting level x_{opt} (or U_{opt}), it is necessary to chose a criterion and use it to find a mathematical expression suitable for calculating the value of x_{opt}. As an example we shall use the weight criterion (12.7) and its corollary (12.11). If, in place of $w_{si}(x)$ and $w_i(x)$ we substitute their values from (12.1, 12.2) and perform the simplest conversion, we obtain

$$I_0(vx) \geqslant 1e^{v^2/2}. \tag{12.14}$$

If the equals sign is put into this expression, we obtain an equation for the value of x which is the threshold required for resolution. Considering

that the inequality (12.14) was found from the conditions for optimum operation, we obtain a mathematical expression for finding the optimum limiting level (threshold) x_{opt}:

$$I_0(vs_{opt}) = 1e^{\frac{v^2}{2}} \tag{12.15}$$

To calculate x_{opt} it is necessary to have the signal/noise ratio v and the importance of false alarms $1, \frac{\partial}{\partial}$ at which the detector will operate. After obtaining these, the value of $1e^{\frac{\partial}{\partial}}$ is calculated and the product vx_{opt} is determined graphically (Fig. 12.2) or from tables, from which the threshold x_{opt} is found.

To establish the determined optimum threshold in the resolver circuit shown in Fig. 12.5, the noise level must be measured. Afterwards, we find that $U_0 = \sigma_i x_{opt}$. This voltage is the optimum limiting level. It is set up on potentiometer R_0.

12.8. Binary-Quantized Signals (BSQ)

The essence of binary quantization consists of the following: A signal u(t) from the receiver output (from the detector) is applied to the threshold device. At the moment that the signal exceeds the established threshold u_0, a standard pulse ("one") is generated, whose length, amplitude, and shape are identical each time (Fig. 12.7).

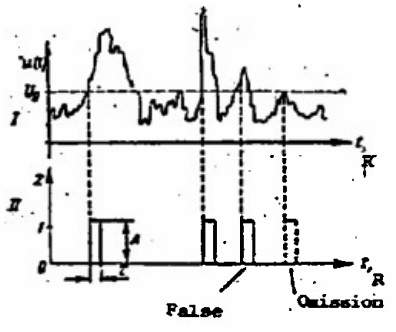


Figure 12.7. The principle of binary quantization of signals.

Thus with binary quantization, resolution is a function of Z where Z can be 0 or 1:

$$Z = \begin{cases} 0: & \text{"no target"} \\ 1: & \text{"target".} \end{cases}$$

As a result of binary quantization, the resolution signal is transformed into digital form, which makes it possible to carry out subsequent information processing on a digital computer.

The effect of interference in binary quantization may result in false ones ("false alarms") or false nulls ("target omission"). Therefore, the threshold U_0 is established on the principles of optimum operation, as explained in the preceeding section.

A blocking oscillator with delay (Fig. 12.8) may be used as the quantizing element, where the delay is established in regard to the required threshold with the variable resistor R_0. The blocking oscillator must have a short rise time.

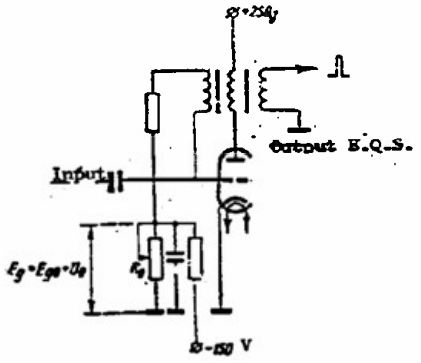


Figure 12.8. A blocking oscillator quantizer schematic. Quantizing level U_0 is controlled by R_0.

12.9. Optimum Algorithm for Detecting Sequences of Binary Quantized Signals

As a rule, radar stations operate in such a way that, in the process of scanning a space, each target is illuminated several times; thus the signal from the target is not one reflected pulse, but a sequence of pulses. Detection of the target with a sequence of pulses gives better results, since a much larger quantity of information is used for resolution than in the case of a single signal. Repeated appearance of a pulse in the same scan region essentially increases the probability that there is a target. Furthermore, the probability that a false pulse will occur in the same place each time is small, the smaller the greater the length of this sequence.

Sequences of quantized signals are an azimuthal sequence of standard pulses, and in digital form they are an azimuthal sequence of nulls and ones (Fig. 12.9). The value N_s is called the width of the sequence: it is equal to the number of positions occupied by one signal, assuming no noise. The width of the sequence changes considerably, depending on the power of the received signal and the threshold U_0.

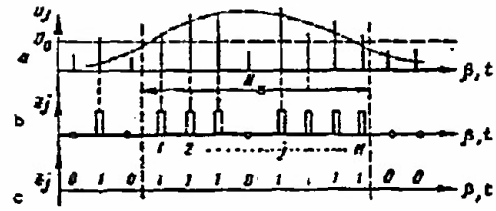


Figure 12.9. Radar signal: a) sequence of nonquantized signals; b) sequence of binary quantized signals; c) sequence in digital form

An algorithm for optimum detection of a sequence of binary quantized signals is given by

$$\sum_{j=1}^{N_s} x_j \alpha_j \geq C, \tag{12.16}$$

where C and α_j are coefficients;

$j = 1, 2, 3 \ldots N_s$ is the number of positions in the sequence.

Detection with the given algorithm is done by counting the total of one signal in the sequence, where each jth unit is multiplied by its coefficient α_j. Further, the total is compared with the coefficient C, and, if the sum is greater than or equal to C, the decision is given as "target".

The coefficients are calculated from the equations:

$$\alpha_j = \ln \frac{P_{sij}(1-P_i)}{P_i(1-P_{si})};$$

$$C = \ln\gamma - \sum_{j=1}^{N_s} \ln\frac{1-P_{sij}}{1-P_i},$$

where γ is a coefficient determined by the chosen criterion of optimum detection; for example, for the weight criterion, $\gamma = 1$;

$P_{si\,j}$ is the probability of a true one signal appearing at the jth position; it is calculated from Eq. 12.5, taking into account the fact that at each position in the sequence there is a unique signal/noise ratio;

P_i is the probability that a false one signal appears; it is calculated from Eq. 12.3.

12.10. Detection by the "k of m" Method

In many cases the whole sequence is not used for resolution, but only a part of m positions, such that $m < N_s$ (Fig. 12.10). In this case, $\alpha_j \approx \alpha =$

const may be considered valid on these m positions, and Eq. 12.16 takes the form

$$\sum_{j=1}^{m} z_j \geqslant k, \tag{12.17}$$

where k is a whole number coefficient equal to C/α, rounded off to a whole number.

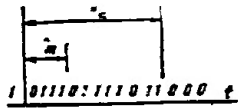


Figure 12.10. m positions are taken from the sequence of length N_s

Equation 12.17 is an algorithm of optimum detection by the "k of m" method. Detection by the "k of m" method amounts to calculating the number of one signals at m adjacent positions, and if the number of ones is larger than a certain set number k (or even equal to k), the device signals "target".

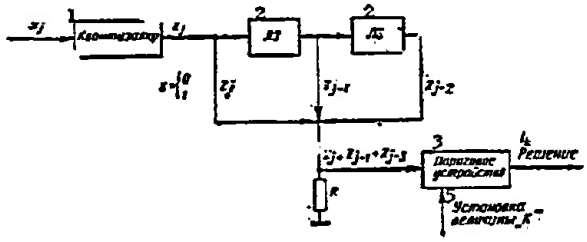


Figure 12.11. Functional diagram of a detector on the k of m logarithm. Here x_j is the receiver output signal, observed on a fixed part of the range.
1 - Quantizer; 2 - DL; 3 - threshold device; 4 - resolution; 5 - setting for "k".

The algorithm is quite simply executed. A possible functional schematic of a resolution device for m=3 is shown in Fig. 12.11. From the quantizer, "zero" or "one" signals enter the delay lines DL, each of which delays the signals by an amount equal to the pulse repetition rate of the radar station T_r. Subsequently the signals are summed on resistor R, and the resulting voltage is applied to the threshold device. The threshold is established by a regulator at such a value that, when k or more pulses ("ones") are applied to the resistor simultaneously the device operates and signals "target".

12.11. Probability of Detection and False Alarm in the "k of m" Method

The probability of detecting a target by the "k of m" method is given by the formula

$$P_{cd}(k, m) = \sum_{i=k}^{m} C_m^i P_{si}^i (1-P_{si})^{m-i}$$

where $C_m^i = \frac{m!}{i!(m-i)!}$ is the number of combinations of m elements i at a time;

P_{si} is the probability of a one appearing initially.

The probability of a false alarm in the "k of m" method is equal to

$$P_{fa}(k, m) = \sum_{i=k}^{m} C_m^i P_i^i (1-P_i)^{m-i},$$

where P_i is the probability of a false one appearing.

The probability of a false alarm is reduced by the "k of m" method.

12.12. Automatic Reading and Coding of Range

Range is proportional to the delay time of the signal reflected from the target relative to the triggering pulse; therefore, measuring range amounts to measuring delay time t_d.

Figure 12.12 illustrates a variation of a functional schematic for measuring range. It includes a time pulse generator (TPG), time pulse counter with n trays and a switch with the same number of elements. The circuit is triggered by a pulse from the radar station, which fires the pulse generator and sets the counter at 0. Time pulses from the generator are applied to the first element of the counter. Consequently, the counter forms a binary number, proportional to the time calculated relative to the triggering pulse. Each element of the counter is connected through the switch to a memory device of the information processing computer. Range is read at the moment a "target" pulse arrives from the detector. This pulse opens the switch, and the binary number, proportional to range, located at that moment on the counter, is transferred into the computer memory for further processing.

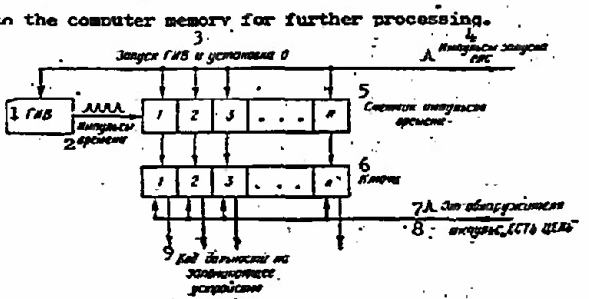


Figure 12.12. Variation of a functional circuit for automatic range measurement.

1 - TPG; 2 - time pulses; 3 - TPG trigger and 0 setting; 4 - triggering pulse; 5 - pulse counter; 6 - switch; 7 - from detector; 8 - "TARGET" pulse; 9 - range code to memory.

The following factors affect the accuracy of range measurement:

frequency of the pulse generator F_G; in the circuit of Fig. 12.12, maximum range error is

$$\Delta R[m] \approx 150 \frac{1}{F_G[Hz]};$$

to increase accuracy frequency F_G must be increased;

frequency stability of the pulse generator; any instability leads to a change in the proportionality established for the automatic equipment between the counter code number and range; to decrease errors of this type, the generator should be highly stable;

fluctuations in the "target" pulse along the time axis; the moment the quantizer which produces the "target" pulse fires depends on the slope of the leading edge of the reflected signal and the noise level distorting the leading edge; mean square error of range measurements associated with this factor may be calculated by

$$\sigma_R[m] = 150 \frac{1}{v\Delta f[MHz]}$$

where v is the signal/noise ratio;

Δf is the spectral width of the reflected signal.

12.13. Automatic Reading and Coding of Azimuth

Measuring azimuth of a target by a plan position indicator radar facility amounts to measuring the angle of rotation of the antenna at the moment when the center of the sequence is fixed. In this case, the azimuth of the center of the sequence should coincide with the geometric axis of the antenna. Accurately determining the center of the sequence is a complex problem, since interference distorts the sequence, and its boundaries become indefinite, thus producing errors in azimuth measurements.

Several procedures for measuring azimuth have been published.

Measuring azimuth from the beginning of the sequence. The azimuth of the beginning of the sequence β_b is fixed, and the angle equal to half the width of the directional pattern of the antenna is added:

$$\beta = \beta_b + \frac{\varphi_{0.5}}{2},$$

where $\varphi_{0.5}$ is the width of the directional pattern, assumed to be constant.

This method is the simplest to put into practice, since it requires a minimum memory volume to store the constant $\varphi_{0.5}$. However, measurement errors

are quite large and are comprised of errors in calculating β_b and errors arising from assuming the sequence width to be constant, although in reality it varies.

Measuring azimuth from the beginning of the sequence taking its width into consideration. This method eliminates error due to assuming $\varphi_{0.5}$ constant. The azimuth of the beginning of the sequence β_b is measured and the width of the sequence N_a is calculated. Azimuth is computed by

$$\beta = \beta_b + \frac{3n_A}{F_r} N_s,$$

where n_A is the antenna rpm;

F_r is the repetition rate of the triggering pulses.

The additional calculation of the width of the sequence results in certain complexities in the computer.

Measuring azimuth by the beginning and end of the sequence. In this case azimuth is calculated as one half the sum of the azimuth of the beginning β_b and end β_e of the sequence

$$\beta = \frac{\beta_b + \beta_e}{2}.$$

This method entails fewer errors in comparison with the first method, but it is complex to carry out, since it requires additional memory space for storing azimuth β_b in the calculating process.

Figure 12.13 shows a possible variation of a circuit for automatic reading and coding of azimuth for the first method. It consists of three parts: an azimuth pulse sensor, azimuth counter, and switches. Two magnetic discs, rigidly fastened to the axis of the antenna, may be used as a sensor. Notches (calibration lines) are placed around the edge of disc A, evenly spaced. Magnetic tips are placed above the surface in line with the notches. When the antenna and disc rotate, pulses are generated in the magnetic tips by the notches in the discs; these pulses are subsequently amplified and sent to the azimuth counter as azimuth markers (pulses). Thus a binary number, proportional to azimuth of the antenna, is formed on the counter.

Sensor disc B has only one calibration line. It is oriented on the antenna axis so that at the moment when the antenna passes through an azimuth equal to zero, the disc is located under the magnetic tip and generates an "azimuth 0" pulse. This pulse is necessary to set the azimuth counter to "0" and to calculate azimuth from the null position of the antenna.

The binary number is applied to the switches that are opened by the "beginning of sequence" pulse coming from the detector. The "beginning of sequence" pulse coming from the detector. The "beginning of sequence" signal

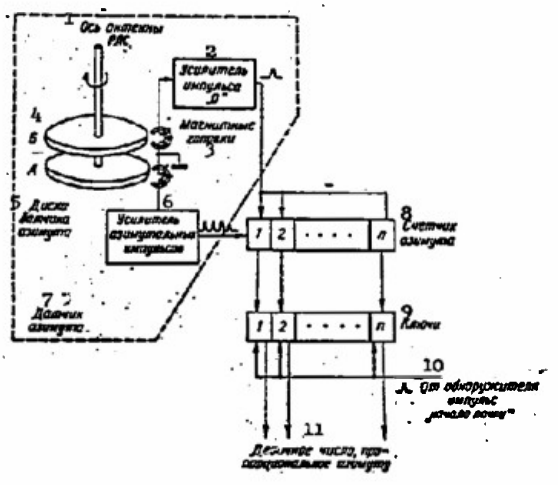


Figure 12.13. Operating principle of an automatic azimuth,reading and coding device (variant).

BLANK PAGE

Figure 12.13. Operating principle of an automatic azimuth reading and coding device (variant)

1 - Antenna axis; 2 - "0" pulse amplifier; 3 - magnetic tip; 4 - B; 5 - azimuth sensor discs; 6 - azimuth sensor discs; 6 - azimuth pulse amplifier; 7 - azimuth sensor; 8 - azimuth counter; 9 - switches; 10 - "Beginning of sequence: pulse from detector; 11 - binary number, proportional to azimuth.

may be formed by a logic circuit "k of m". At the moment when the switches open, the azimuth code number is taken from the counter and transferred to the computer memory for processing.

The accuracy of the azimuth reading depends primarily on two factors. It depends on the number of calibration lines on disc A. For increased accuracy, the number of calibration lines per unit length of the circumference may be increased. Accuracy also depends on the precision with which the beginning of the sequence is determined by the detector circuit. Random appearance and disappearance of pulses on the edges of the sequence due to interference has the result that the "k of m" logic circuit cannot detect precisely the beginning of the sequence. Therefore the switches will open prematurely or too late, causing azimuth to be read with random errors.

12.14. Secondary Information Processing

Information obtained during primary processing is not freed from the effects of interference which may cause the primary processing facility to give false data on the target or to lose data. Furthermore, interference causes errors in measuring coordinates.

Secondary information processing is the following stage of obtaining information on the target, and it is designed to eliminate false data, restore filtered data from the target and decrease errors in measuring coordinates. In addition, other functions may be fulfilled, associated with calculating the velocity and course of the target.

These functions may be accomplished with the results of several scans by the radar station: thus secondary information processing may be considered to be the processing of information during several scans.

12.15. Principles of Secondary Processing

The regularities of positioning false indications and indications from a target in each succeeding scan by a radar station are completely different. False indications appear with no relation to each other from scan to scan; therefore they are shown as random spikes of noise. Indications from a

target are positioned by the mechanics of the motion of the target in space; consequently there is a correlation between the indications of a preceding and succeeding scans. Actually, it can be seen from a suitable picture of the indicator screen, that indications from a target are spread out along the trace of the target motion, while false indications occur chaotically at various places on the screen.

This difference in the regularities of the positions is at the basis of secondary processing.

Secondary processing operates as follows. Assume that the observer locates three indications from one target during three scans. Figure 12.14 illustrates these indications, numbered according to the scan numbers N, N-1, N-2. Some indications show up in the following N+1 scan, and the problem is to determine which of them belongs to the target (trace) and which are false. This is done in two steps.

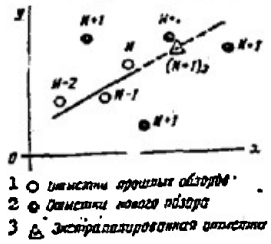


Figure 12.14. Secondary information processing.

1 - Preceding scan indications; 2- New scan indications; 3- Extrapolated indications.

First, the point at which the future N+1 indication is expected to appear is calculated. To do this, a trace is laid out in the best possible way along the three preceding indications, velocity is calculated, and the probable position of the future indication is determined. It is designated $(N+1)_e$ on the figure. This step is called extrapolating (forecasting), and the calculated indication is the extrapolated point.

Secondly, the positions of all indications in the new scan are analyzed relative to the extrapolated indication, and the signal from the target is chosen from them. This step is called comparison. The choice is based on comparing the probability that the indication belongs to the trace. That indication which has the greatest probability is considered to be the indication from the target. For the case shown in Fig. 12.14, it is possible to consider that indication to be from the target which is located close to the extrapolated indication, because the probability of its belonging to the trace is greatest in comparison with the others.

12.16. Model of the Motion of the Target

For the extrapolating step, the regularities of the motion of the target must be known.

In general, the trajectory of a target is a complex and often random function, exceptionally difficult to describe. Therefore in many cases a simplified model of target motion is used in practice, most frequently a rectilinear, uniform model. It is always used when the whole trajectory is not observed, but only a small portion of it, which is sufficiently accurately described by the equation of a straight line.

In other cases it is assumed that the target may maneuver. Then in the limited observation interval, the trajectory of the maneuvering target is represented by a second order polynomial.

Algorithms of the extrapolation step determine the choice of the model for target motion. The calculations are simplest for rectilinear uniform motion.

12.17. Extrapolation of Indications

The essence of extrapolation is predicting from the previously obtained data the coordinates of the next indication. This problem is complex because the data used for extrapolation contain errors; consequently the results of extrapolation also contain errors, and in many cases have more errors than the original data.

From many methods of calculating extrapolated data, the two most widely used may be pointed out.

Extrapolation on parameters. This method requires that the parameters of the trajectory, such as the component velocity vectors V_x and V_y, be found. If extrapolation is carried to one scan forward, the x component of velocity is approximately equal to

$$V_{xN} \approx \frac{X_N - X_{N-1}}{T_s},$$

where N is the number of the scan;

T_s is the scan period.

The coordinates of the extrapolated X indication X_e are calculated from

$$X_e = X_N + \overline{V}_{xN} T_s$$

The Y coordinate is calculated analogously.

Extrapolating on two indications gives more errors because of errors in calculating velocity. Therefore velocity is refined (averaged) in each succeeding scan by the formula

$$\overline{V}_N = \overline{V}_{N-1} + \frac{V_N - \overline{V}_{N-1}}{N-1}$$

where V_{N-1} is the average velocity during the preceeding scans without the Nth indication;

V_N is average velocity including the Nth indication.

This method gives good results if the target is moving rectilinearly and regularly.

Extrapolation on several indications. The algorithm of this method can be used for maneuvering targets also. Generally

$$X_e = \sum_{i=1}^{n} \eta_i x_i, \tag{12.18}$$

where n is the number of indications (scans) used for extrapolation; the indication of the preceding scan is given the index n=1;

η_i is the weight coefficient of the ith indication.

For extrapolating on three indications for a nonmaneuvering target one scan forward

$$X_e = \frac{4}{3} X_1 + \frac{1}{3} X_2 - \frac{2}{3} X_3.$$

For a maneuvering target

$$X_e = 3X_1 - 3X_2 - X_3.$$

The Y coordinate is calculated analogously.

This type of algorithm is quite simply realized in a digital computer. but it requires a large memory if the number of indications n used for processing is increased.

12.18. Extrapolation Errors

The following factors affect extrapolation errors:

errors in measurement; the greater the error in measuring coordinates, the worse the prediction; these errors are directly proportional;

the number of indications n used for processing; increasing the number of indications reduces the prediction error, and with a very large number of indications, it may become lower than measurement error (Fig. 12.15); however, it requires an excessively large computer memory;

the prediction interval τ; the farther into the future the indications are extrapolated, the sharper the rise in error (Fig. 12.16); this effect is most severely pronounced in the case of a maneuvering target.

12.19. Smoothing

Smoothing is a special case of extrapolation where the extrapolation interval is equal to zero. It is used when the coordinates of the indication

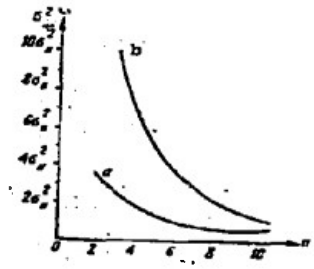


Figure 12.15. Relationship of extrapolation errors to number of indications n used for extrapolation: a- for a nonmaneuvering target; b- for a maneuvering target

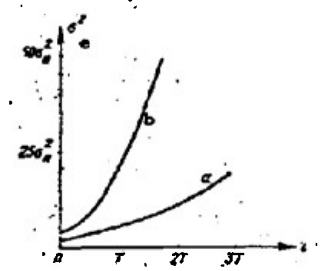


Figure 12.16. Relationship of extrapolation errors to extrapolation interval τ: a- for a nonmaneuvering target; b- for a maneuvering target

must be known more accurately. Equation 12.18 is used for smoothing, and formulas for calculating have the form:

for a nonmaneuvering target using three indications

$$X_s = \frac{5}{6} X_1 + \frac{1}{3} X_2 - \frac{1}{6} X_3;$$

for a maneuvering target using four indications

$$X_s = -\frac{1}{5} X_1 + \frac{1}{10} X_2 + \frac{3}{5} X_3 + \frac{7}{10} X_4.$$

12.20. Comparison

The results of comparing extrapolated indications with newly received data tells which indications belong to the trace. This problem is solved with a strobe, which is defined as an area constructed around the extrapolated point in which an indication from the target may be expected to appear (Fig. 12.17). The method is based on the following logic:

if the indication falls inside the strobe, then it is considered to be an indication from the target; all remaining signals are assumed false or not belonging to the trace;

if no indications fall within the strobe, the extrapolated point is accepted as the signal from the target; if no indications fall within the strobe several times in a row, the target is considered to be lost or the trace was false;

if several indications fall within the strobe, the point lying closest to the extrapolated point is considered to originate from the target.

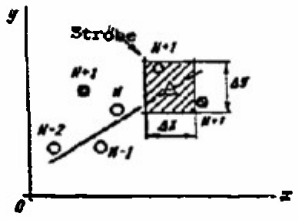


Figure 12.17. Comparison with a strobe.

Choosing the size of the strobe is a contradictory problem. In the interest of raising the probability that an indication from the target will fall within the strobe, its dimensions should be increased, but this increases the probability that false signals will occur within the strobe, thus making a smaller strobe more desirable. The contradictions are solved by choosing the optimum strobe dimensions by using the optimum criteria (see 12.6. "The concept of optimum limiting level and criteria for determining optimum"). For the case of a normal distribution of measurement errors and equally accurate measurements (where $\sigma_x = \sigma_y$) the dimensions of a rectangular strobe are determined by:

$$\Delta X = \Delta Y \approx 2.5\sigma\sqrt{-\ln \gamma\xi\sigma^2 + 1.8},$$

where γ is a coefficient determined by the choice of the optimum criterion;

ξ is the density of false indications; it is measured by the number of indications per unit area during one scan;

σ is the scattering of the indications relative to the extrapolated indication; calculated by

$$\sigma = \sqrt{\sigma_m^2 + \sigma_e^2}$$

where σ_e is the extrapolation error;

σ_m is the error in measuring coordinates.

12.21. Capture of Signals for Processing

Capture is the first moment of secondary processing. It may be semi-

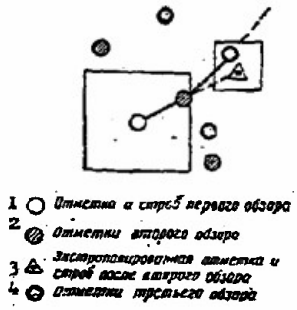


Figure 12.18. Diagram of the process involved in target capture for processing.

1 - first scan signal and strobe; 2 - second scan signal;
3 - extrapolated signal and strobe after second scan;
4 - third scan signal.

BLANK PAGE

automatic or automatic. In the first case, the operator introduces several signals from the same target which he has chosen into the secondary processing machine, and he sends the command that initiates secondary processing of this group of signals.

In the second case, the process of automatic capture is initiated by the appearance of the first signal (Fig. 12.18). A large strobe is constructed around it. When a signal of the second scan falls inside the strobe, an extrapolated point and a small strobe are calculated from it and from the first signal. If a signal from the third scan falls inside this strobe, capture may be considered to have taken place, and secondary processing is carried out in the usual order. If no signals occur inside the strobes, the beginning of the trace is considered to be false and processing is terminated. The possibility exists that the computer repeatedly captures false signals and that the trace they represent is false. Usually, a false trace breaks off quite rapidly.

12.22. Tertiary Information Processing

Tertiary processing is analysis of information from several radar stations. The purpose of tertiary analysis to collect information from several stations and to combine it, forming a general picture of the aerial situation.

Since the scanning zones of the radar stations often overlap, the same target may show up on the displays of several stations. In the ideal case, these indications should coincide. In practice however, they do not coincide because of systematic and random errors in measuring the coordinates of the targets and because of different location times. For these reasons, complexities arise in combining the information, where it must be decided which target is the real one. Of course, the indications may not coincide not only because of the errors mentioned above, but also because there are several targets creating these indications.

This problem is the major one in tertiary analysis. The following functions must be accomplished to solve it:

collect the data from the radar stations;

reduce the indications to one system of coordinates and to a single reference time;

identify the indications to establish their relationship to the target;

average the coordinates of several identified indications to obtain one indication with more accurate coordinates.

These functions are accomplished using all components of the indications: coordinates, height, velocity, location time, number of targets, relationship to the trace, etc.

In practice, tertiary analysis devices require special computers.

Tertiary analysis is the concluding step in obtaining information on the aerial situation.

12.23. Collecting Messages about the Target

Signals sent from a radar station through the communication channel for further analysis and processing are usually called target messages. Collecting messages includes accepting the greatest possible amount of data with minimum losses. Losses occur because the analyzer cannot collect all the data supplied to it, and part of the data is unused.

The theory of mass accommodation is used to appraise possible data collection devices. The basic characteristic of a collector is its relative throughout ρ, calculated as the ratio of the average number of processed messages $\overline{Q}_p$ to the average number of messages applied to the collector input $\overline{Q}_i$:

$$\rho = \frac{\overline{Q}_p}{\overline{Q}_i};$$

Messages are usually transmitted in code, such as a binary number. In this case, the message consists of several groups of digits (Fig. 12.19) which assume a certain meaning. Such a group of digits is called a "word". The message includes supplementary symbols, determining the beginning and end of each message. The number of digits required to form the message is called the length of the message L. It is easy to find the minimum transmission time T_m for a message, if the transmission time of one digit t_d is known:

$$T_m = Lt_d.$$

	1		2			3				4		5		6	
a	Служебный признак		№ цели			Координаты				Характеристика цели		Время локации		Служебный признак	
b	9	9	5	1	2	6	8	1	1	4	0	1	6	0	0
c	1001	1001	0101	0001	0010	0110	1000	0001	0001	0100	0000	0001	0110	0000	0000

Figure 12.19. Code message (variation): a - meaning of a word; b - digital representation of a work; c - digit code.

1 - Supplementary symbol; 2 - target No.; 3 - coordinates; 4 - target characteristic; 5 - location time; 6 - supplementary symbol.

There may be pauses of length T_p between messages. In some data transmitting systems, the pauses are of constant length, in other systems, the pauses may be of random length. In these systems, the average message transmission speed is equal to

$$\overline{V}_m = \frac{60}{T_m[\text{sec}] + \overline{T}_p[\text{sec}]} \text{ [messages/min]},$$

where $\overline{T}_p$ is the average length of random pauses.

Messages entering the information processing computer through several channels form a stream of messages at its input with a density of

$$\delta = n_k \overline{V}_m \text{ [messages/min]},$$

where n_k is the number of channels.

The flow is called stationary if its density remains constant with time.

Each incoming message should be subject to preliminary processing: decoding, correction, and entering in the computer memory. These operations require a certain time T_o in which the machine is occupied. When the density of the message flow is sufficiently great, the machine cannot carry out preliminary processing, and some messages are lost.

Collecting systems may be of the following types.

Collecting systems with rejections in which, when the machine is occupied at the moment another message arrives, the message receives a "rejection" and leaves the processing.

Collecting system with limited hold. In such a system, the entering message is held in its turn a certain time T_h if the machine is occupied. If in this time the message is not accepted for processing, it receives a "rejection" and is lost. Systems with hold capability have fewer losses than systems with rejections. Figure 12.20 shows the block diagram of a collecting system with hold. The message enters the hold register through the receiver, where it is stored a time T_h. After finishing the previous information processing, the computer replies ready to the inquiry of the holding register. The interrogation device interrogates the holding registers in turn. If a register holds a message, it is collected into the computer and processed, after which the cycle repeats itself again.

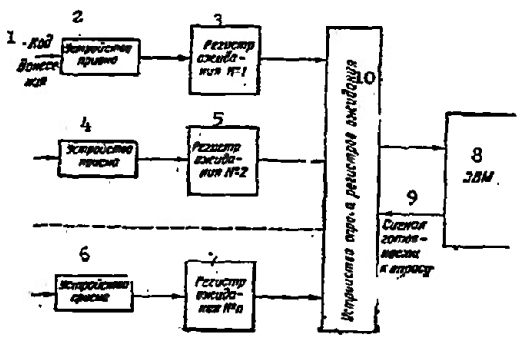


Figure 12.20. Variation of a block diagram for a message collecting system with hold.

1 - Message code; 2 - receiver; 3 - No. 1 hold register; 4 - receiver; 5 - No. 2. hold register; 6 - receiver; 7 - No. n hold register; 9 - ready signal to interrogation; 8 - computer; 10 - interrogator for hold registers.

When hold time T_h is much less than busy time T_o, the collecting system may be considered to be a system with rejection.

For delay systems, the relative throughput may be calculated from

$$\rho = \frac{1}{T_o \delta + 1}.$$

The equation is valid when the message flow is considered stationary and messages arrive independently in a random fashion. This is called the simplest flow.

12.24. Reducing Indications to a Single System of Coordinates and to a Single Reference Time

When indications enter the computer with coordinates measured relative to points of the radar station standing, they must be referred to a single system of coordinates.

A geodesic system (latitude, longitude, height above sea level) may be used as the single set of coordinates. Curvature of the earth's surface must be completely taken into account; however, calculations with such a system are complicated, and therefore they are used when the radar stations are separated by large distances and errors due to curvature of the earth cannot be tolerated.

When distances between radar stations are small, a rectangular system of coordinates with correction in height is usually used. Calculations in this system are relatively simple, and errors from substituting a plane surface for a sphere may be completely acceptable for solving many tactical and operative problems.

The algorithm for converting in this system of coordinates is illustrated in Fig. 12.21 and represented by the formulas:

$$X = R\cos\varepsilon\sin(\alpha+\beta) + X^*;$$
$$Y = R\cos\varepsilon\cos(\alpha+\beta) + Y^*;$$
$$H = R\sin\varepsilon + \frac{R^2}{2R_e},$$

where X, Y, and H are target coordinates in a single system of coordinates;

X^* and Y^* are coordinates of the standing point of the radar station relative to the point of tertiary information processing;

R, β, and ε are target coordinates in the polar system of coordinates of the radar station (slant range, azimuth, elevation);

R_e is the equivalent radius of the earth;

α is the angle of approach of the meridians to the width of the processing point; it is considered positive if the point of the radar

station standing is to the left of the axis, and negative if the point of the radar station standing is to the right of this axis.

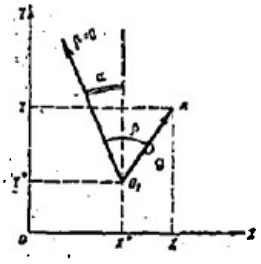


Figure 12.21. Conversion of coordinates.

The readings are converted to a common time when the indications have different location times. This is necessary to determine the position of the processed indications at a certain moment in time. Similar conversion simplifies identification of the indications.

As an example we shall examine two indications A and B with different location times t_a and t_b, which arrive for processing (Fig. 12.22). It is quite possible that these indications are from the same target taken at different times. To confirm this statement, we must calculate their position at the same moment. If after this, they are quite close to each other, it is quite possible that they refer to the same target.

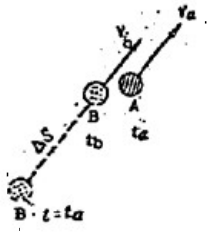


Figure 12.22. Reduction of two indications to a single time when $t_b > t_a$.

Reduction to a common time is done as follows:

Assume a time t, to which the indications are to be referred. Further find the difference in time for each in indication

$$\Delta t_i = t - t_i.$$

Then extrapolate the indications to time Δt_i, computing sections of the path ΔS_i to which each indication must be shifted along its trace:

$$\Delta S_i = V_i \Delta t_i.$$

If Δt_i is negative, the indication is shifted in a direction opposite to the direction of motion of the target. For this example (Fig. 12.22) the time $t=t_s$ is chosen as the common time. Assume that indication B has a later location time, i.e., $t_b > t_a$ The difference Δt_b is then negative, and indication B is shifted in a direction opposite to the target motion by an amount $\Delta S = V_b \Delta t_b$. As a result, we obtain the mutual positions of the indications at time $t=t_a$.

After reducing to a single system of coordinates and a single time, the indications can be analyzed for identification.

12.25. Identification of Indications

In the identification process a decision is produced which establishes:

how many targets there actually are, if the indications come from several radar stations;

how the indications are located as to targets.

Identification is usually accomplished in two steps:

Step I: comparison, as a result of which a first approximation identification (rough identification) is obtained;

Step II: distribution, permitting a more precise solution among roughly identified indications.

Comparison. The basis of this step is that indications from the same target should be identical in all components (coordinates, height, velocity, etc.). Therefore, deciding whether one indication is related to the other may be based on comparing the components of the indications. If comparison shows a decided coincidence of all corresponding components, it can be assumed that the indications are identical, i.e., that they refer to the same target. However in practice, due to errors in measuring and converting to a single system of coordinates and a single time, there is no complete coincidence in the indications, in spite of the fact that they come from the same target. Consequently, an uncertainty exists, and two competing hypotheses arise:

hypothesis H_1 assumes that noncoincidence arose from errors in the indications although they come from the same target;

hypothesis H_0 assumes that noncoincidence arose because the indications are from different targets.

Choice of the correct hypothesis is based on evaluating the amount of

noncoincidence. The differences in all components must be established:

$$X_{ij} - X_{rz} = \Delta X;$$
$$Y_{ij} - Y_{rz} = \Delta Y;$$
$$H_{ij} - H_{rz} = \Delta H;$$
$$\ldots \quad \text{etc.}$$

where i and r are the numbers of the radar stations relaying indications;

j and z are the number of the indications.

Subsequently the allowable deviation in all components is established

$$\Delta_{all} = \{\Delta X_{all}, \Delta Y_{all}, \Delta H_{all}\} \ldots .$$

The decision rests on the following logic: if even one of the components ΔX, ΔY, ΔH exceeds its allowable value, the indications are considered to be from different targets. Otherwise they are identical.

If acceptance of hypothesis H_1 is designated 1, and acceptance of hypothesis H_0 is designated 0, then the identification decision may be written as the product

$$H = z_x z_y z_H \ldots,$$

where

$$z_x = \begin{cases} 1, & \text{if } \Delta X \leq \Delta X_{all}, \\ 0, & \text{if } \Delta X > \Delta X_{all}, \end{cases}$$

$$z_y = \begin{cases} 1, & \text{if } \Delta Y \leq \Delta Y_{all}, \\ 0, & \text{if } \Delta Y > \Delta Y_{all}, \end{cases}$$

Then for H=1, hypothesis H_1 is used, and for H=0, hypothesis H_0 is used.

Designating allowable deviations Δ_{all} is a contradictory problem: increasing Δ_{all} raises the probability that indications from the same target will be identified correctly, but at the same time, the probability of erroneous identification of indications from different targets increases also. In this regard, only rough, approximate identification is done in this step. The value Δ_{all} is calculated by

$$\Delta_{all} = k\sqrt{\sigma_m^2 + \sigma_c^2},$$

where k is a coefficient chosen in accordance with the required probability of correct identification; for k=3, this probability is close to 1;

σ_m and σ_c are mean square errors of measuring and converting the indication components.

Distribution. In many cases, the required region of allowable deviations is so large that it exceeds the resolution capability of the radar station by

several times. Consequently indications from different targets located close to each other are usually considered to be identified. In the second stage it must be decided how many targets are located in the region of allowable deviations and how the indications should be distributed by targets.

Distribution is done by the rules of logical analysis of the indication positions. These rules may be as follows:

If indications in the region of allowable deviations are obtained from the same radar station, then the number of targets is equal to the number of indications, for it is impossible for one station to transmit different indications from one target at a given moment of time.

If an equal number of indications is received from each radar station, it is most probable that the number of targets is equal to the number of indications from one radar station, since it is not probable that, in the boundaries of a small area, each radar station "sees" only its own number of targets and "does not see" a target observed by another radar station.

Indications are distributed by targets by choosing the most probable variation. For example, if two radar stations observe two targets, four indications will be received at the processing point. They may be distributed in three ways, shown in Fig. 12.23. Variation III is immediately excluded by the first rule of logical distribution given above. The method of a minimum sum of the squares of distances is used to decide between variations I and II. This requires the following operations:

1. Calculate the distance between indications (see Fig. 12.23) S_{1121}, S_{1222}, S_{1122}, S_{2112}.

2. Determine the sum of the squares of the distances for each variation:

Variation I: $S^2_{1121} + S^2_{1222}$;

Variation II: $S^2_{1122} + S^2_{2112}$.

3. Choose that variation for which the sum is the least. This variation is considered more probable.

The rules of comparison and grouping given above are not the only ones, and, depending on the required processing accuracies, they may be complicated or simplified in calculations.

12.26. Averaging Indications

As a result of identification, information about the target is represented by a group of indications obtained from different radar stations, usually having different characteristics of accuracy. An averaging operation is performed to make one (combined) indication out of them.

The first method of averaging amounts to computing the average arithmetic

coordinate by

$$X = \frac{\sum_{i=1}^{m} X_i}{m},$$

where m is the number of indications in the given group.

The formulas are analogous for the Y, and H coordinates, velocity V, and bearing θ.

This method is quite simple in operation with a computer, but it does not consider accuracy characteristics of the radar stations. The best results are obtained by averaging by weighted means according to the equation

$$X = \frac{\sum_{i=1}^{m} b_i X_i}{\sum_{i=1}^{m} b_i},$$

where a value inversely proportional to the square of the error of measurement and conversion of the im indication coordinate is taken as the weight coefficient:

$$b_i = \frac{1}{\sigma_i^2}.$$

Error σ is calculated from

$$\sigma = \sqrt{\sigma_{\text{изм}}^2 + \sigma_{\text{пр}}^2}.$$

In some simplified cases, it is permissible to use only the error in measuring coordinates, which is given in the technical data of the radar station, projected to the im indication.

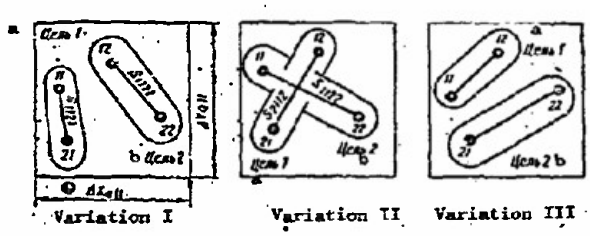


Figure 12.23. Variations of the distribution of the indications with with respect to the targets. Indications are described by a two-digit number: the first digit is the number of the radar, the second is the number of the indication received from the given radar.

a - Target 1; b - target 2.

Chapter XIII

Radar Reliability and Operational Fundamentals

13.1 Basic Concepts

Systems and components are the subjects of reliability engineering research. A system is understood to mean the set of objects designed to work together to carry out a predetermined task. Objects are understood to mean equipments, and sometimes even the environment and the operating personnel. Components are parts of a system designed to function in a predetermined manner.

Reliability is the capacity of the system (component) to carry out assigned functions under predetermined operating conditions while retaining the values of basic parameters within established limits.

A failure is an event causing the system (component) to cease carrying out assigned functions, or the basic parameters to exceed permissible limits.

A broader concept than failure is the concept of fault. A fault is an event during which any of the initial properties of a system (component) are disturbed. To be distinguished are the "primary faults," those resulting in failure, and the "secondary faults," those that do not cause failures.

Reliability contains two systems (component) properties; failure-free operation, and restorability.

Failure-free operation is understood to mean the property of a system (component) to retain its ability to function for a specified period of time under predetermined operating conditions.

Restorability is understood to mean the capacity of a system (component) to have its ability to function restored after a failure has occurred.

Restorability includes repairability and serviceability. Repairability characterizes the internal (engineering, design) properties of the system (component).

Serviceability characterizes factors external to the system (component) (qualifications of personnel, organization of service, supply, repair, and the like).

Also to be distinguished are "restorable" (used more than once) and "nonrestorable" (used only once) systems and components. A radar, for example, is included among the systems used more than once, whereas, an on-board missile electronic system is included among the systems used but once.

Name of index	Designations used	Measurement unit
Indexes characterizing failure-free operation		
Probability of failure-free operation	$p(t)$, $P(t)$	
Probability density of failure-free operation	$f(t)$	
Failure rate	$a(t)$, $A(t)$	$hours^{-1}$
Failure density	$\lambda(t)$, $\Lambda(t)$	$hours^{-1}$
Mean operating time between failures - mean time to failure	T_f	hour
Average service life for component	T_{av}	hour
Indexes characterizing restorability		
Restoration probability	$V(\tau)$	
Restoration rate	$v(\tau)$	$hours^{-1}$
Restoration density	$\mu(\tau)$	$hours^{-1}$
Average restoration time	T_r	hours
Generalized indexes		
Availability	C_a	
Probability of normal functioning	$P_{nf}(t)$	
Downtime ratio	C_d	

The probability of failure-free system operation is understood to mean the probability that the system will function correctly for a predetermined time interval, t, under specified operating conditions. The probability density of failure-free operation, $f(t)$, indicates the change in the probability of failure-free operation with respect to time.

The failure probability, $Q(t)$, is the probability of the opposite event taking place

$$Q(t) = 1 - P(t). \tag{13.1}$$

The following formula can be used to arrive at a practical, approximate, determination of the probability of failure-free operation

$$P(t) \approx N_f - n(t)/N_f, \tag{13.2}$$

where

N_f is the number of systems (components) tested;

$n(t)$ is the number of systems (components) failing in time t.

The accuracy with which $P(t)$ is established through the use of the formula at (13.2) will depend on N_f and increases with increase in N_f.

Failure rate is the differential law for the distribution of time of failure-free operation. Failure rate is the rate of "drop" in reliability.

The statistical value of failure rate can be established through the formula

$$a(t) = dq(t)/dt \approx \Delta n/N_f \Delta t, \qquad (13.3)$$

where

Δn is the number of elements failing in time Δt in the interval $(t, t + \Delta t)$.

Failure rate is a characteristic of restored equipment. Components failing during tests should be replaced with new ones when a statistical determination of failure rate through the formula at (13.3) is made.

Failure density can be expressed as the ratio of the increase in the number of failures, Δn, in the time interval from t to $t + \Delta t$, to the number of components remaining in good working order during the interval of time considered

$$\lambda(t) \approx \Delta n/[N_f - n(t)]\, \Delta t, \qquad (13.4)$$

where

$N_f - n(t)$ is the number of components remaining in good working order during the interval of time considered;

$n(t)$ is the number of components failing from the beginning of the test to time t.

The statistical value of failure density is established through the formula at (13.4). Failure density is a characteristic of nonrestorable equipment.

Failure density is sometimes expressed as a percentage of 100, or 1000, hours of operation.

If a failure density per 1000 hours of operation (λ') is specified, a conversion formula

$$\lambda = \lambda'/1000,$$

must be used to obtain the failure density (λ) per hour.

Experience in the operation of electronic systems and components indicates that a change in failure density with time occurs in accordance with the law shown in Figure 13.1.

The dependence $\Lambda = f(t)$ has three clearly defined curve sections.

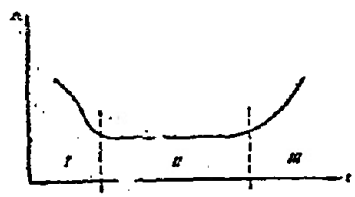


Figure 13.1. Dependence of failure density, Λ, on time in operation.

Section I is the breaking in period. Here an increased failure density because of failure of the weakest components with hidden defects is characteristic. The duration of the period is sometimes hundreds of hours, and depends on the production process used.

Section II is the normal operating period, or the established operating condition for the system. A reduced level, and constancy in the failure density, Λ = const, is characteristic here. This section covers several thousand hours.

So far as Section II is concerned, the dependence of the probability of failure-free operation on time is fixed by an exponential law

$$P(t) = e^{-\Lambda t}. \tag{13.5}$$

Section III is the period during which wear and aging occur. Here an increase in the failure density as a result of failure of mass application components is characteristic. The normal law applies when establishing the probability of failure-free operation over this section.

Mean time to failure, T_f, can be established as the average length of time of operation between failures, using the formula

$$T_f \approx \frac{1}{n} \sum_{i=1}^{n} t_i, \tag{13.6}$$

where

n is the number of failures during the test period, t;

$t = \sum_{i=1}^{n} t_i$.

t_i is the time of failure-free operation from the $(i - 1)^{th}$ failure to the ith failure.

When the exponential law applies, the mean time to failure can be established as a magnitude the inverse of the failure density

$$T_f = 1/\Lambda.$$

An approximate determination of average service life for components can be made through the formula

$$T_{av} = \frac{\sum_{i=1}^{N_f} t_i}{N_f} \tag{13.7}$$

When there is a known dependence of the probability of failure-free operation of a system on time, P(t), the mean time to failure can be established through the formula

$$T_f = \int_0^{\infty} P(t)\,dt. \tag{13.8}$$

The restoration process simply means putting a failed system, or component, back in service. How long this process takes is established by the elapsed time between failure and the completion of restoration.

Restoration probability, $V(\tau)$, provides a quantitative index of the probability that the system will be restored in a period of time not in excess of a specified value, τ.

Average restoration time is the average time spent on restoring the system, and is found through the expression

$$T_r = \frac{1}{n}\sum_{i=1}^{n} t_{ri} \tag{13.9}$$

where

t_{ri} is the time spent on restoring the ith failure;

n is the number of failures (completely restored).

If the average restoration time for individual sub-systems is known, the average restoration time for the system can be found through the formula

$$T_p = \frac{\sum_{i=1}^{N} \Lambda_i T_{ri}}{\sum_{i=1}^{N} \Lambda_i} \tag{13.10}$$

where

T_{ri} is the average time to restore the ith sub-system;

Λ_i is the failure density in the ith sub-system.

When the restoration probability is known, the average restoration time can be found through the expression

$$T_r = \int_0^{\infty} [1 - V(\tau)]\, d\tau \tag{13.11}$$

Availability is the probability that an arbitrarily taken system will be in proper working order at any moment in time. The steady-state (fixed) value of the availability can be found through the formula

$$C_a = T_f/T_f + T_r \tag{13.12}$$

The probability of normal functioning is a generalized index of reliability for multipurpose use systems. This index takes into consideration the initial state of the systems as well as its failure-free operation and restorability, and is the probability of a complex event, including the fact that the systems will be in good working order at the beginning of use and will function without failing for specified time t

$$P_{nf}(t) = C_a P(t). \tag{13.13}$$

The downtime ratio characterizes the forced downtime of equipment because of failures and is established as the probability that at any arbitrary moment in time the equipment will be inoperative.

Quantitative indexes are reliability measures. The correct selection of a quantitative index is the basic, correct, evaluation of equipment

reliability when designing and evaluating equipment reliability. For example, the probability of failure-free operation during time of flight will completely characterize the reliability of aircraft equipment. The probability of normal functioning will be the index for missile equipment for a missile that must be ready to launch at any time. The mean time to failure can serve as a characteristic of reliability for unattended equipment such as artificial earth satellite equipment, and the like.

13.3. Reliability Characteristics and Typical Faults in Radio Components

The determinant factor for typical radio components is the failure density. Generalized data on component failure density, as published in the press from time to time, are listed in Table 13.1.

Table 13.1. Failure Density for Typical Radio Components

Names of components	Failure density/hour of operation
Vacuum tubes	$(0.001 \text{ to } 0.345)10^{-3}$
Resistors	$(0.00001 \text{ to } 0.015)\ 10^{-3}$
Condensers	$(0.00001 \text{ to } 0.164)\ 10^{-3}$
Transformers	$(0.00002 \text{ to } 0.064)\ 10^{-3}$
Chokes, induction coils	$(0.0002 \text{ to } 0.044)\ 10^{-3}$
Relays	$(0.005 \text{ to } 1.01)\ 10^{-3}$
Selsyns, electric motors	$(0.001 \text{ to } 0.33)\ 10^{-3}$
Semiconductor instruments:	
diodes	$(0.00012 \text{ to } 0.5)\ 10^{-3}$
germanium and silicon triodes	$(0.0001 \text{ to } 0.9)\ 10^{-3}$
Switch gear	$(0.000003 \text{ to } 0.28)\ 10^{-3}$
Plugs	$(0.00001 \text{ to } 0.091)\ 10^{-3}$
Connections, soldered and wired	$(0.00001 \text{ to } 0.01)\ 10^{-3}$
Gyroscopes	$(1 \text{ to } 6)\ 10^{-3}$

Typical faults in receiver-amplifier tubes (in % of total number of failures):

change in parameters and loss of emission - 43 to 51%;

filament breaks and burning - 20 to 28%;

cracks in envelopes, socket breakage, and others, 9 to 12%.

Typical resistor faults:

breakage of, or broken contacts - over 50%;

burn out of conducting layer - 35 to 40%;

sharp changes in resistance compared with rated - 5 to 8%.

Typical condenser faults:

breakdown in the dielectric - almost 80%;

reduction in capacity - 15%;

reduction in insulation resistance - 5%.

Typical relay faults:

closed and burned contacts;

contact springs out of adjustment;

breakdown of insulation between windings and housing;

open windings.

Typical faults in transformers and chokes:

burned out windings;

windings shorted between turns;

breakdown of insulation between windings and housing;

breakdown between turns.

13.4. Factors Affecting Equipment Reliability

Factors affecting equipment reliability can be divided up into two groups: subjective and objective. The subjective factor depends on human activity. Included here is everything related to circuit selection and design decisions during the design stage, the selection of components and materials to provide normal operating conditions, the organization of engineering service, and the operation of the equipment.

The objective factors are the result of the effect created by the environment and related to climatic, meteorological, biological, mechanical, and other activities. The degree to which these factors act depends on the efforts made by man to weaken them. The factors can be broken down into design-fabrication and operational, according to the nature of their activity. Design-fabrication factors include those related to the development, planning and building of the equipment. They have the greatest effect on reliability. Operational factors include those affecting equipment use reliability. They include the objective factors that result from the influence exerted by the environment, and the subjective factors related to the organization of the engineering service system, repair, provision of spare parts, qualifications of the operating personnel, and on other questions.

The probability of normal operation is one of the most complete quantitative characteristics of reliability, taking into consideration as it does failure-free operation and restorability of the equipment. Its magnitude is the result of the effect of two diametrically opposed resultant factors, shown diagrammatically in Figure 13.2.

Destabilizing factors of a subjective and objective nature reduce reliability. At the same time, the efforts of developers and operating personnel are directed at increasing reliability.

Measures leading to an increase in the reliability level can be divided up into two groups in so far as they affect reliability parameters:

measures directed at increasing failure-free operation;

measures directed at improving restorability.

All these measures should be operable in the development and production stage, as well as during operation. Of particular importance in operation are preventive, or adjustment measures, the scope of which and when done, is established by the corresponding instructions. The purpose of preventive maintenance is to avert some of the failures, and at the same time, to increase failure-free operation and reduce equipment downtime during restoration.

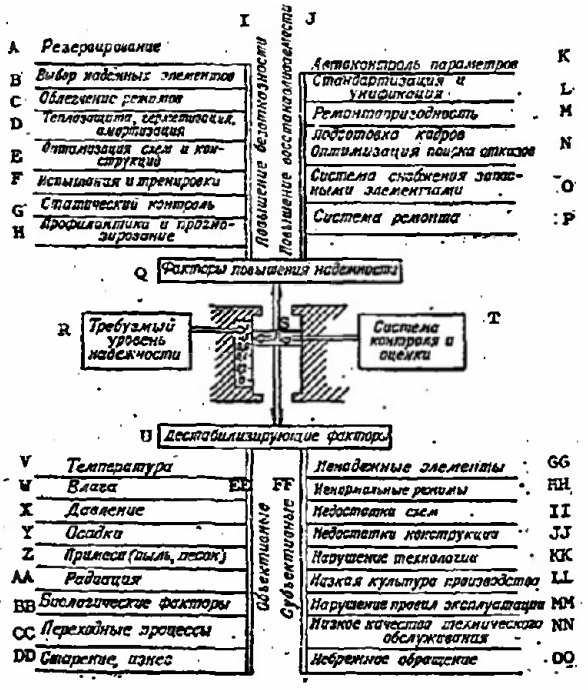


Figure 13.2. Factors affecting reliability

A - Redundancy

B - Selection of reliable components

C - Simplification of modes

D - Heat protection, sealing, shock mounting

E - Optimization of circuits and construction

F - Tests and preaging

G - Static monitoring

H - Preventive maintenance and forecasting

I - Increase in failure-free operation

J - Increase in restorability

K - Automatic monitoring of parameters

L - Standardization and unification

M - Repairability

N - Training cadres. Optimization of trouble shooting

O - System of supplying spare components
P - Repair system
Q - Factors for increasing reliability
R - Required level of reliability
S - P_{nf}
T - Monitoring and evaulating system
U - Destabilizing factors
V - Temperature
W - Moisture
X - Pressure
Y - Precipitation
Z - Impurities (dust, sand)
AA - Radiation
BB - Biological factors
CC - Transitional processes
DD - Aging, wear
EE - Objective
FF - Subjective
GG - Unreliable components
HH - Abnormal modes
II - Circuit shortcomings
JJ - Design shortcomings
KK - Engineering violations
LL - Poor production methods
MM - Violations of operating rules
NN - Poor quality of mechanical servicing
OO - Careless handling

13.5 Computing Reliability Characteristics

The essence of the reliability computation is a determination of the numerical values of the reliability indexes in accordance with the initial statistical data for the components making up the system. The initial data for computing equipment reliability are:

the failure rate, λ_i;

number of components N_i, for each type used in the equipment;

average restoration time for components, T_{ri}.

Established as a result of the reliability computation are:

for one-time use systems - system failure rate

$$\Lambda = \sum_{i=1}^{n} \lambda_i \qquad (13.14)$$

and the probability of failure-free operation of a system P(t) on a normal operation section;

for continuous duty, or repeated use systems - average restoration time, T_r, availability, C_a, and probability of normal functioning, $P_{nf}(t)$, are computed.

The basic assumptions used in computing reliability are:

failure of any component will cause system failure;

failure of any component has no effect on the operation of the rest of the components; that is, component failures are considered mutually independent.

An approximate reliability computation is made in the initial stage of equipment design. Characteristic of it is that the precise number N_i and λ_i in the groups are as yet unknown because the component operating modes have not yet been selected and the final electrical computations for

the circuits have not yet been made. The purpose of this computation is to inject operational interference into the equipment design in the initial stage and take steps to increase the reliability of the weakest units.

The sequence in which the approximate reliability computation is made

1. All system components are broken down into several groups with approximately the same failure rate.

2. The number of elements, N_i, in each group is computed.

3. Tables are used to find the values, λ_i, for the components in each group (by averaging, or taking the extremes).

4. The system failure rate is computed through the formula at (13.14).

5. The average failure-free operating time is found

$$T_f = 1/\Lambda.$$

6. The probability of failure-free operation for operating time t is computed through

$$P(t) = e^{\Lambda t}.$$

In addition, the following are computed for continuous duty systems:

7. the average restoration time

$$T_r = \sum_{i=1}^{m} \frac{\lambda_i}{\Lambda} T_{ri}$$

8. the availability

$$C_a = T_f/T_f + T_r; \text{ and,}$$

9. the normal functioning probability

$$P_{nf}(t) = C_a P(t).$$

Data on the number of identical components are not always available in the initial design stage. Sometimes only the total number of stages is known. Then, in order to make the computation, what must be known is the average stage failure rate, λ_s, and the number of stages in the equipment, N_s.

The final reliability computation is made with design and operational factors based on the above explained methodology for making the approximation computation taken into consideration. Temperature and duty have the greatest effect on component failure rate. Consequently, component failure rate is finally computed with these factors taken into consideration.

The dependence of failure rate on temperature for various load factor, C_1, values is of the greatest practical significance. The load factor is understood to mean:

for condensers

$$C_1 = U_o/U_r,$$

where

U_r is the rated voltage for the particular type of condenser;

U_o is the condenser operating voltage in the circuit;

for resistors

$$C_1 = P_d/P_r,$$

where

P_d is dissipated power;

P_r is rated power;

for vacuum tubes

$$C_1 = P_f + P_p + P_g/P_{fr} + P_{pr} + P_{gr},$$

where

P_f, P_p, and P_g are the powers dissipated by the tube electrodes under the operating conditions (filament, plate, and grid power, respectively);

P_{pr}, P_{fr}, and P_{gr} are the rated values of the plate, filament, and grid powers, respectivaly;

for a relay

$$C_1 = I_o/I_r,$$

where

I_o is the contact operating current;

I_r is the rated contact current value.

The values for λ_i are found from graphics according to the operating temperature and the load factor for each component when the final reliability computation is made with duty taken into consideration.

As before, the formula at (13.14) can be used to find the summed failure rate for the system.

The sequence in which the final reliability computation is made

1. Component load factors, C_1, are computed.

2. The $\lambda_i = f(C_1, t°)$ graphics are used to find the actual values of λ_i for the specified load factors and operating temperatures.

3. The precise number of components in each group is computed.

4. The failure rate, and other characteristics, are computed as in the approximation method.

The reliability computation with component aging taken into consideration

There is, in addition to the unexpected failures characteristic over a normal operating period (section II, fig. 13.1), an inherently gradual change in the original properties and parameters for each component with time, leading to a so-called gradual failure. When these will occur is a random event, described by the normal law. As a practical matter, it is extremely difficult to distinguish between unexpected and gradual failures,

but it generally is accepted that in a normal period of operation, when λ = constant, almost all failures are unexpected, and the reliability computation is made on the basis of an exponential law. When the period of general wear and aging sets in on the main body of components, the failure rate will begin to increase sharply because of gradual failures. Moreover, the dominant failure in many components is the gradual, rather than the unexpected. Included among these components as a rule are the electromechanical units and parts, the lives of which are fixed by wear and aging phenomena.

The probability of failure of a component such as this in time t can be found through the formula

$$q(x) = 0{,}5 - \Phi(x), \quad x = \frac{t - T_{av}}{\sigma}, \tag{13.15}$$

where

T_{av} is the average life for the component;

σ is the mean square deviation from the average life.

The function $\Phi(X)$ is called the probability integral, and its value is found in tables [L.1]. The computation for the probability of failure-free operation reduces to finding the function $\Phi(X)$ in the tables.

The probability of failure-free operation of a system equals

$$P(t) = \prod_{i=1}^{N} P_i(t) = \prod_{i=1}^{N} [0{,}5 + \Phi_i(x)]. \tag{13.16}$$

For the case of identical components

$$P(t) = [0{,}5 + \Phi_i(x)]^N. \tag{13.17}$$

For a large number of components in a system

$$P(t) \simeq e^{-N[0{,}5 + \Phi_i(x)]^N} \tag{13.18}$$

The sequence in which the computation is made is:

1. find the magnitude $X_i = t - T_{avi}/\sigma_i$ from the original values of T_{avi} and σ_i;
2. find the probability integral from the tables [L.1];
3. find the probability of failure-free operation, $P(t)$.

13.6 Evaluating Reliability in Operation and During Tests

Tests establish the data for equipment operating time between successive failures, as well as the data on restoration time. Naturally enough, the greater the number of readings on operating time between failures, the more accurate will be the computed average value.

The accuracy which the evalustion is made can be established through the accuracy factor δ, which depends on the number of experiments and the required authenticity γ. The authenticity of the evaluation indicates the magnitude of the assurance that the true value of the parameter evaluated is within specified limits.

For example, the mean time to failure is found through the expression at (13.6)

$$T_f \approx \frac{t_\Sigma}{n},$$

where

t_Σ is the total time of serviceable operation;

n is the number of failures occurring in time t_Σ.

The accuracy in the evaluation can be established by the factors δ_1 and δ_2

$$T_{f\ min} = \delta_1 T_f \le T_f \le \delta_2 T_f = T_{f\ max}.$$

Authenticity of evaluation is usually given in the range from 0.9 to 0.99. Proceeding from the specified authenticity from the curves in Figure 13.3, we find the accuracy factors δ_1 and δ_2, then the maximum and minimum values for the sought for magnitude,

$$T_{f\ min} = \delta_1 T_f,\ T_{f\ max} = \delta_2 T_f. \qquad (13.19)$$

The magnitudes $T_{f\ min}$ and $T_{f\ max}$ are the numerically established boundaries of the evaluation accuracy, and are called the trustworthy boundaries of the sought for parameter, T_f.

The same methodology can be used to evaluate the magnitude of the average restoration time, T_r. The accuracy factors for the evaluation of T_f can also be established from the curves in Figure 13.3. The scale is for the case when the probability of restoration is subordinate to

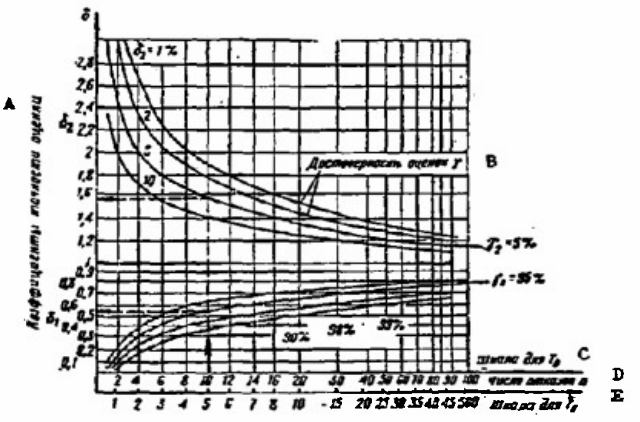


Figure 13.3. Dependence of the evaluation accuracy factors δ_1 and δ_2 on the number of tests for different evaluation authenticities.

A - evaluation accuracy factors; B- evaluation authenticity;

C - T_f scale; D - number of failures, n; E - T_r scale.

Erlang's law

$$v(\tau) = 1 - \left(1 + \frac{2\tau}{T_r}\right) e^{-\frac{2\tau}{T_r}}.$$

The downtime, or the availability, ratio can be evaluated similarly. Table 13.2 lists the evaluation accuracy factors for the downtime ratio for an authenticity $\gamma = 0.9$.

Table 13.2

Table of accuracy factors for evaluating the C_d parameter with an authenticity of $\gamma = 0.9$

n	1	2	3	4	6	8	10	12
z_1	0,15	0,22	0,28	0,32	0,42	0,46	0,5	0,54
z_2	19	7	4	32	2,43	2,18	1,96	1,86
n	14	16	18	20	25	30	35	40
z_1	0,56	0,58	0,6	0,62	0,65	0,68	0,7	0,72
z_2	1,77	1,7	1,63	1,6	1,53	1,47	1,42	1,39
n	45	50	60	70	80	90	100	200
z_1	0,73	0,74	0,76	0,78	0,79	0,81	0,82	0,86
z_2	1,36	1,33	1,28	1,27	1,25	1,24	1,23	1,16
n	300	400	500	600	700	800	900	1000
z_1	0,89	0,9	0,91	0,92	0,925	0,93	0,935	0,94
z_2	1,13	1,11	1,1	1,09	1,08	1,075	1,07	1,069

Example 1. After some period of operation of two identical radar systems, the following statistical data were compiled:

for the 1st system: $n_1 = 20$; $t_{\Sigma 1} = 1600$; $\tau_{\Sigma 1} = 150$;

for the 2d system: $n_2 = 20$; $t_{\Sigma 2} = 1000$; $\tau_{\Sigma 2} = 50$.

Required with an authenticity of $\gamma = 0.9$ are the operational reliability of the system and the average number of failing systems if $N = 100$ sets of equipment are in operation.

Solution.

1. Let us find

$$T_{f1} = t_{\Sigma 1}/n_1 = 1600/20 = 80 \text{ hours}; \quad T_{r1} = \tau_{\Sigma 1}/n_1 = 150/20 = 7.5 \text{ hours};$$

$$T_{f2} = 1000/10 = 100 \text{ hours}; \quad T_{r2} = 50/10 = 5 \text{ hours};$$

$$T_f = t_{\Sigma 1} + t_{\Sigma 2}/ n_1 + n_2 = 1600 + 1000/20 + 10 \approx 87 \text{ hours};$$

$$T_r = \tau_{\Sigma 1} + \tau_{\Sigma 2}/n_1 + n_2 = 150 + 50/20 + 10 \approx 6.7 \text{ hours};$$

$$C_d = T_r/T_f + T_r = 6.7/87 + 6.7 \approx 0.07.$$

2. Let us find the accuracy factors, δ_1 and δ_2, for T_f from Figure 13.3 for $n = 30$ and $\gamma = 0.9$

$$\delta_1 = 0.7; \ \delta_2 = 1.3.$$

The trustworthy boundaries for the parameter T_f will equal

$$T_{f\ min} = 0.7 \cdot 87 = 61 \text{ hours};$$

$$T_{f\ max} = 1.3 \cdot 87 = 113 \text{ hours}.$$

3. Let us compute the possible minimum and maximum values for the probability of failure free operation, P(t) for any time t

$$P_{max}(t) = e^{-t/T_{f\ max}}; \ P_{min}(t) = e^{-t/T_{f\ min}}.$$

4. Let us find δ_1 and δ_2 from Figure 13.3 to evaluate the parameter T_r

$$\delta_1 = 0.788, \ \delta_2 = 1.22.$$

The trustworthy boundaries are

$$T_{r\ min} = 0.788 \cdot 6.7 \approx 5.3 \text{ hours},$$

$$T_{r\ max} = 1.22 \cdot 6.7 \approx 8.2 \text{ hours}.$$

5. Let us compute the possible minimum and maximum values for the probability of restoration for any time τ.

6. Let us find the evaluation accuracy factors δ_1 and δ_2 for the downtime ratio, C_d, from Table 13.2 for $n = 30$

$$\delta_1 = 0.68, \ \delta_2 = 1.47.$$

The trustworthy boundaries for the availability equal

$$C_{a\ min} = 1 - C_{d\ max} = 1 - 1.47 \cdot 0.07 \approx 0.9,$$

$$C_{a\ max} = 1 - C_{c\ min} = 1 - 0.68 \cdot 0.07 \approx 0.95.$$

7. The most probable average number of constantly operating systems, N_o, will be within the following boundaries

$$N_{o\ max} = NC_{a\ max} = 100 \cdot 0.95 = 95 \text{ system},$$

$$N_{o\ min} = NC_{a\ min} = 100 \cdot 0.9 = 90 \text{ systems}.$$

13.7 Electronic Equipment Redundancy

The methods used to provide redundancy can be divided into those using some volume of stand-by equipment (a redundancy "level"]and those using the condition of stand-by components while main components are functioning properly.

The first of these distinguishes three basic ways in which the main and the stand-by components can be included:

general redundancy for an entire system by using identical, duplicating, components;

individual (component by component) redundancy for a system by providing redundancy for individual components in that system;

mixed redundancy, in which individual sub-systems, as well as certain of the primary components, are provided with stand-bys.

The second of these distinguishes two different conditions for the stand-by components when the main components are in operation:

constant redundancy, wherein the stand-by components (systems) are connected to the main components throughout the operating period and wherein both function under the same conditions;

redundancy by replacement, wherein the stand-by components (systems) do not replace the main ones until the latter fail.

Characteristic of replaceable redundancy is a condition in which one finds the stand-bys under light, or no load conditions, that is when the stand-by components are not in operating condition.

The ratio of the number of stand-by components (systems) to the number of corresponding main components (systems, installations), is called the redundancy ratio.

Failure of a main unit, or block will not result in system failure when redundancy is built in. Consequently, the main and stand-by components must be considered as being connected in parallel for redundancy. Failure will occur only when the main, and all stand-by components, have failure.

The probability of failure-free operation of a system with redundancy can, in the case of general redundancy, be found through the expression

$$P_{gen} = 1 - \prod_{i=0}^{m} (1 - P_i), \qquad (13.20)$$

where

m is the redundancy ratio;

P_i is the probability of failure-free operation of each of the parallel circuits (main and stand-by).

In the case of individual redundancy, the probability of failure-free operation can be found through the expression

$$P_{ind} = 1 - \prod_{i=0}^{m} \left(1 - \prod_{j=1}^{n} P_{ij}\right), \qquad (13.21)$$

where

P_{ij} is the probability of failure-free operation of the ith component in the jth circuit;

n is the number of components individually provided with stand-bys.

What follows from the expressions at (13.20) and (13.21) is that individual redundancy provides a great gain in reliability, but realization demands quite a large number of switches. Unreliability in this devices can lead to a considerable reduction in this gain. The most effective approach is to use an unloaded stand-by, but this cannot always be done because this type of redundancy requires a good deal of time to switch to stand-by.

13.8 Restorability of Electronic Equipment

Restorability is a broad concept, encompassing the property of the equipment itself, as well as the operating procedure, and depends on the following principles:

the training of operating personnel;

the organizational and engineering measures taken to service the equipment and provide spare parts.

Restorability is a combination of repairability and the maintenance measures used. Repairability is understood to mean the design and engineering properties of the system characterizing its adaptability to the detection and elimination of failures.

Maintenance measures can be characterized by the qualifications of the operating personnel, how the repair supply system is organized and others.

There are many systems in which the danger is not the fact that functioning has ceased, but how long the system will be shutdown, and this latter is determined by restoration time. The basic components of the restoration time are the time required to find the fault, the time required to eliminate it, and the time required to tune and cut in the equipment after restoration. Of these, the time required to find the fault is the longest, so proper organization of the process used to find the fault will result in a significant reduction in restoration time.

The best result can be obtained by using systems that automatically monitor operability and that will provide information on the failure site immediately after a failure occurs. But failures in the monitoring system used must also be taken into consideration when automatic monitoring systems are used. The components in the monitoring system are designed so their failure will have no direct bearing on the operation of the main equipment, but there is always a logical connection between the functioning of the monitoring system and the main equipment. Hence there are optimum relationships between the complexity and the reliability characteristics of the monitoring equipment and the complexity and the reliability of the main equipment.

When there is no system monitoring, detection of a faulty component reduces to making a series of successive tests. Each test provides definite information on equipment condition. The process of finding the fault can be represented as a series of tests, as a result of which the faulty component is found.

13.9 The Methodology Used to Shoot Trouble

The methodology used to shoot trouble is a set of methods and a sequence of actions to arrive at the most desirable and fastest way to shoot troubles.

There are two basic trends in solving this problem at the present time:

develop and use a subjective methodology for shooting trouble (a manual method);

automate the troubleshooting procedure by designing special systems.

The most logical sequence of tests for shooting trouble in any system is one in which the boundaries of the area in which the trouble may be are gradually narrowed down until the trouble is localized to a concrete, damaged component. In other words, each successive step is taken on the basis of information obtained from the preceding test showing which part of the system can be excluded from consideration because it is functioning properly. The first thing to be found in an electronic system is the faulty channel (receiver, indicator, etc.), then the faulty block in the channel, the stage in the block, the component in the stage. Thus, trouble shooting involves a successive division of the system from the initial large section and the elimination of properly functioning components, into smaller sections, or areas of that system.

This logical, planned, succession of tests (verifications) has come to be called the successive approximations method.

The successive approximations method

Let some system consist of $N = 8$ components connected in series.

Let us assume that the failure probability is equal for all components, that is, there is an equal probability of failure on the part of any of the components.

The probability that the failed component will be found by using a sequence of random tests equals 1/7. This means the time required to find the trouble is proportional to the number of components less one (the last component is not tested)

$$\tau_{st} = k(N - 1),$$

where

k is the average time required to make one test;

τ_{st} is the time required to find the trouble by using the sequence of tests method.

The successive approximations method can be reduced to the so-called mean point method in the special case when we have no advance information (the indications are of normal operation prior to testing). This means we make the first test in the oiddle of the system, between components 4 and 5. This cuts the boundary of the faulty section in the system down to components 5, 6, 7, and 8. The properly functioning half of the system (components 1, 2, 3, and 4) is immediately ruled out of the consideration.

The second test is made by halving the rest of the faulty sections in the system. Thus, to trouble shoot in a system of eight components, no more than three tests will be required.

The maximum number of tests, m, in this method can be computed through the formula

$$m = \log N / \log 2.$$

The time required to find the faulty component is

$$\tau_s = km = 1.43\ k \log N. \qquad (13.22)$$

The successive approximations method is a necessary element of the methodology, but is not enough. If the methodology is to be fruitful, definite test procedures (methods) are needed.

The principal test methods are:

the switching monitor and control method;
the intermediate measurements method;
the replacement method;
the external inspection method;
the comparison method;
the characteristic troubles method.

The switching monitor and control method

What this method amounts to, in essence, is the use of an evaluation of the external indications of the appearance of trouble to successively eliminate from consideration the properly functioning sections of the system by an analysis of the channel circuits and the use of the operating switching and controlling devices, as well as the operating monitoring elements (signal lights, built-in instruments, automatic protective devices, and the like).

The channel is understood to mean the aggregation of a set of components carrying out one of the system's predetermined functions.

The channel schematic must show the concrete cabinets, connecting cables, plugs, stages, contacts, and switches in the system carrying the particular type of signal. It differs from the block schematic and the functional schematic in that it has numbering and installation designations for the connecting plugs, cables and contacts. It differs from the installation diagrams in that it does not contain the cable runs and the structural

members, and finally, differs from the principal schematic in that it does not contain stage components. The channel schematic provides adequately complete information on the paths followed by the signal. We should note, however, that the channel schematic will not be completely satisfactory if it does not contain the control and monitoring components (switches, adjustments and monitoring components, and the like).

Let us take some examples of how the switching monitor and control method is used.

When a signal arrives at the output of a broadcast receiver, the receiver must be switched to the record player mode on all bands. The appearance of sound in this mode lets us rule out of consideration the sections of the circuit from the LF amplifier input to the speaker.

If there are no signals from one of the channels in the system at the receiver output, the "Automatic Frequency Control - Manual Frequency Control" switch must be used, setting it to the MFC position until satisfied with the appearance of signals. The tests made with this switching monitor rule out of consideration the signal channel (the IF amplifier, the detector, the video amplifier), and the decision is made to make the next test of the faulty automatic frequency control channel.

The switching monitors method is an effective way to use the successive approximations method. The method is fast and simple for checking assumptions as to the status of sections of a system. The disadvantage of the method is the limited use that can be made of it (only sections can be separated, and not the concrete site of the damage).

The intermediate measurements method

The essence of this method is to compress the faulty area, or to find the failing components, by measuring circuit resistances, supply conditions, high-frequency circuit parameters, oscillographing voltages and currents at the pin jacks, and at other points in the circuit. Measurement results are compared with charts of resistances, voltages, with mode tables, and with oscillograms. This method is particularly desirable to use in the final stage of the trouble shooting, when the boundary of the faulty section of the system has been compressed into a block, stage, or circuit, and it is necessary to find the failed component, or the reason for the faulty operation. Built-in, as well as accessory monitoring and measuring, equipment is used to make the tests by this method.

The replacement method

Individual components in the system (blocks, sub-blocks, cable connections, vacuum tubes, and the like) are replaced by components known to be functioning properly in order to shoot trouble by this method. If normal operating indications are observed after the replacement, the conclusion has to be that the component replaced was not functioning properly.

Just where this method will be used when making operating repairs is determined by the availability of spare blocks, units, or vacuum tubes. The method is most often used when trying to find faulty vacuum tubes. When the boundaries of the trouble are reduced to a block the tubes in that block are the center of attention (since they are the least reliable of the components). As a matter of fact, a whole new set of tubes can be installed if tube matching is not required. The effectiveness of this method is enhanced in the case of sub-block design, and if a second, identical, system is available.

The advantages of the replacement method are the speed at which the check can be made, and the simplicity of concluding where the faulty section (component) in the system is located. The disadvantages are the possibility of damaging replacement components, as well as a certain amount of ambiguity if normal operating criteria are not restored after the component is replaced.

The external inspection method

This method assumes that a visual inspection of the installation, of the cables, and of the other components will be made, as well as a check to see if these components, as well as the housings of electric motors, generators, reduction gears, and tubes are hot, listening for sparking in dischargers and in components in the high-frequency channel. This method is particularly effective when the trouble is accompanied by indications that a failure has occurred. The advantages of this method include simplicity and the use of visual means for inspections, but they can be overevaluated now and then and become disadvantages when operating personnel spend entirely too much time trying to trouble shoot using the external inspection method.

The tentative nature of the findings is a disadvantage. The trouble can only be found when the external indications are clearly defined, something that does not happen very often.

The comparison method

In essence, this method resorts to using a properly operating, identical component and comparing its operation with the faulty section in the system. This method supplements and simplifies the intermediate measurements method, and can be used when the components, or systems, are identical.

The characteristic troubles method

The essence of this method is to look for the trouble by using known indications that uniquely set the particular fault apart from all others. A list of these faults, and their indications, is compiled in the form of a table that can be used for making repairs, or for advance studying. The table lists the frequently recurring faults characteristic of the particular system cropping up in operation. These tabulated faults are usually included in the operating instructions.

The disadvantage of this method is the difficulty involved in using it in practice since the entire list of fault indicators must be reviewed until the necessary combination is found. The effectiveness of the method depends on the experience of the technicians using it.

13.10 The Troubleshooting Test Sequence

Three test sequences can be distinguished as applicable to the task of troubleshooting in radar equipment. The need for so doing is related to the difference in the circumstances preceding the troubleshooting, and to the nonidentity in the probability of one, or of several faults occurring simultaneously. Knowledge of the characteristic circumstances very much simplifies the initial stage of the troubleshooting.

The following premises (cases) can be distinguished:

premise I - when signs of trouble appear as soon as the system is cut in;

premise II - when signs of trouble appear during normal system operation;

premise III - when the trouble, without regard for time of appearance, is accompanied by indications of an emergency.

The troubleshooting test sequence when trouble occurs after the system is cut in

Trouble associated with premise I above, occur most often because transportation, opening, and adjustment operations are not carried out according to the rules. Too, there is quite a good probability that several faults will show up simultaneously. Hence, the troubleshooting sequence should be as follows.

If one is to retain one's perspective when analyzing the criteria associated with the trouble, the correctness of the installation in accordance with the original provisions for switching and monitoring devices must be checked very carefully as a first step. Improper installation can be the cause of false information with respect to abnormal system operation.

For example, if one were to simply set the "Automatic Frequency Control - Manual Frequency Control" switch on the radar receiver panel in the "Manual Frequency Control" position, one would be able to reject out of hand the appearance of an indication of false trouble, that of loss of observability over this channel.

If some of the systems are not cut in in the proper sequence, false indications of trouble can appear, and in individual cases, so too can failures.

The first thing that is necessary, therefore, is to eliminate the effect of possible detunings; and to check the supply source voltages, mindful of safety regulations while so doing.

If the results are negative and normal equipment operation is not restored, the search for the damaged component in made as follows.

Using the switches, adjustment devices, monitoring equipment, and the channel schematic, as well as the switching monitors method, reduce the boundaries of the faulty section in the system to the maximum extent possible (right down to the stage).

Having rid the preceding tests of the indications of false faults, the effect of detunings, and of abnormal supply conditions, and having separated the faulty section of the system (channel, block, stage), it is now necessary to turn to the principal schematic of the section (component) and make an analysis of the indications of damage, and the possible causes thereof. At the same time, the ensuing test sequence is planned and the most rational test methods selected. The damaged component is found, and the cause of the failure established, using the intermedinte measurements method, or other methods. After the damage has been eliminated, the monitor measurements in the circuits that have been repaired must be made in order to check conditions and parameters and to tune and adjust the restored component to its full correspondence with the specified operating instructions. This set of rules is sufficiently general for use in troubleshooting in different types of systems.

The troubleshooting test sequence for troubles arising during normal system operation

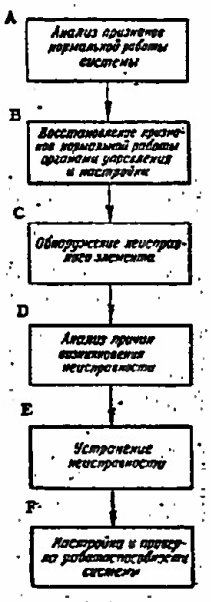
In the case of premise II, the probability of several faults occurring at the same time is slight. Obviously, then the 1st and 2d stages of the sequence considered can be omitted. Troubleshooting must begin by interrogating operating personnel as to the circumstances that brought on the trouble. This information can be useful becuase it will develop incorrect actions taken by operators or technicians. Then an analysis of the symptoms of abnormal operation, and further tests, are made in accordance with the sequence set forth above.

The troubleshooting test sequence for troubles accompanied by indications of an emergency

The first step to be taken when there are unmistakable signs of emergencies is to cut out the system. Once cut out, the section in which the emergency occurred must be given an external inspection. There are individual cases when the signs of the emergency are ill-defined, and in such cases, it is necessary to pinpoint the damaged section, and then cut out the system (or part of it). The damaged component cannot be replaced until the reason for the faulty operation has been established.

The general methodology does not free one from taking a subjective approach to the troubleshooting process. The basis for the process is logical thought and a knowledge of the system in need of repair.

Figure 13.4 is the troubleshooting sequence (algorithm) for the general case (premise I above). The algorithm makes it possible to program an all-purpose troubleshooting electronic computer for concrete examples.



A - Analysis of indications of normal system operation

B - Restoration of indications of normal operation by control and tuning organs

C - Detecting the faulty component

D - Analysis of the reason for the fault

E - Elimination of the fault

F - Tuning and checking system operability

Figure 13.4 The troubleshooting sequence in the general case.

13.11 Safety Precautions When Servicing Radar Equipment

Table 13.3 lists the nature of the action of an electrical current on man, according to the type and magnitude of the current flowing through him.

A voltage of 40 volts is taken to be a safe voltage.

Systematic, long-term and continuous irradiation of the human organism by high-frequency electromagnetic fields in the form of high- and superhigh-frequency pulses, as well as in the form of simple high-frequency oscillations, can have an unfavorable reaction on the human organism. The hygienic standards for superhigh-frequency irradiation:

must not be in excess of 10 microwatts/cm^2 in the case of irradiation for an entire working day, which figure is the magnitude of the power density;

no more than 100 microwatts/cm^2 in the case of irradiation for 2 hours;

no more than 1000 microwatts/cm^2 for irradiation for 12 to 15 minutes under conditions in which the use of protective goggles is mandatory;

when protective clothing and goggles are used the level of the electromagnetic field can be attenuated 30 db.

Table 13.3

Nature of the effect of an electrical current on man

Current, ma	Nature of the effect of	
	50-60 cycle AC	DC
0.6 - 1.5	Beginning of sensation, slight shaking of the hand	No sensation
2 - 3	Fingers shake badly	No sensation
5 - 7	Spasms in hands	Itching, sensation of heat
8 - 10	Difficult, but still possible to tear the hands away from the electrodes. Severe aching of fingers, wrists and hands.	Intensification of heat
20 - 25	Hands become paralyzed slowly. It is impossible to tear the hands away from the electrodes. Very severe aching. Difficulty in breathing.	Even greater intensification of heat, slight contraction of hand muscles.
50 - 80	Breathing paralyzed. Onset of palpitations of the heart ventricles.	Great sensation of heat. Contraction of hand muscles. Spasms, difficulty in breathing.
90 - 100	Breathing paralyzed. When for 3 seconds, or longer, paralysis of the heart, palpitation of its ventricles.	Breathing paralyzed.
3000 and more	Breathing and heart paralyzed when the action is for longer than 0.1 of a second. Destruction of body tissues by the heat from the current.	

The irradiation doses are determined by the power density of the superhigh-frequency oscillations.

The power density Π, on the electrical axis of an antenna in the far zone can be computed through the formula

$$\Pi = P_{av}G/4\pi R^2 \text{ [microwatts/cm}^2\text{]}, \qquad (13.23)$$

where

P_{av} is the average radiated power in microwatts;

R is the distance from the antenna to the site where the operating personnel are working, in cm;

G is the antenna gain.

The formula at (13.23) is only correct for the far zone, that is, when

$$R \geq L_1L_2/\lambda,$$

where

L_1 is the horizontal dimension of the antenna mirror, in cm;

L_2 is the vertical dimension of the antenna mirror, in cm;

λ is the wavelength, in cm.

With the antenna radiation pattern taken into consideration, for any direction in the far zone

$$\Pi = P_{av}G/4\pi R^2 F(\varphi^\circ, \theta^\circ), \qquad (13.24)$$

where

φ° is the azimuth to the personnel irradiated with respect to the maximum irradiation;

θ° is the elevation angle of the personnel irradiated.

Accidents can be avoided if:

maintenance is not permitted on equipment carrying dangerous voltage (over 40 volts);

zones in which it is assumed, or found, that radiation power is in excess of 1 milliwatt/cm^2 are considered dangerous;

danger zones are surrounded by warning signs and the presence in them of operating personnel is forbidden except in cases of extreme emergency (and then only for the minimum time needed);

radiators (antennas) are not directed at living spaces;

protective devices (equipment for collective and individual protection) against irradiation are used;

protective goggles, gloves, and aprons are used when replacing cathode ray tubes.

Chapter XIV

Electrical and Radio Measurements

14.1. General Provisions for Measurements and Measurement Units. The International System of Units (SI). Bels and Decibels.

Electrical and radio measurements are a division of the general science of measurements, metrology, and study methods for use in making experimental determinations of various electrical and radio engineering characteristics, and the corresponding equipment.

Measurement is a perceptional process involving the making of a comparison between a physical experiment concerned with some specific magnitude and certain of its values adopted as the unit of comparison. In principle, it is impossible to make a measurement without the establishment of units.

If one selects at least three physical magnitudes, X, Y, and Z, arbitrarily and independently, to represent certain basic properties of the material world (say length, mass and time), all other magnitudes could then be expressed in terms of these arbitrarily selected magnitudes and represented in the form of a product of the type

$$G = kX^pY^qZ^r, \qquad (14.1)$$

where,

p, q, r are whole numbers, or fractions, positive, or negative;

k is the proportionality factor.

Given an arbitrary selection of units for measuring basic physical magnitudes, one can find the derived units for measuring all physical magnitudes through equations such as these.

A set of measurement units is called a system of units.

On 1 January 1963, State Standard 9867-61 introduced the International System of Units (SI) in the Soviet Union. This is the International System adopted at the XIth General Conference on Measure and Weights (1960).

In accordance with this Standard, the International System of Units should be given preference over other systems of units (the CGS, the MKGFS, and others).

In the SI system the basic units are the meter, kilogram, second, degree Kelvin, ampere, and candle.

The electrical and radio measurement system of electrical units known as the MKSA (meter, kilogram, second, ampere), and widely used in practice, is part of the SI system.

Introduced in the so-called rationalized MKSA system has been the conversion of the formulas for the electromagnetic field equations, involving

the introduction of the factor 4π in the denominator of Coulomb's law. This factor has disappeared from the many formulas encountered in practice, however.

Basic Units in the MKSA System

The meter (m) is a unit of length (a), a length equal to 1650763.73 wavelengths in a vacuum of radiation corresponding to the transition between levels $2p_{10}$ and $5d_5$ of an atom of krypton 86.

The physical basis for expressing the meter in lengths of light waves is the radiation by an excited atom of a quantum of light with constant frequency

$$\nu = \frac{E_2 - E_1}{h},$$

where

h is the Planck constant;

E_1 and E_2 are the possible values of the energy of the atoms, called energy levels.

The selected unit of length (the meter, in particular) can be expressed in terms of the wavelength corresponding to a particular frequency, ν.

The transition from the international platinum-iridium standard to the newly established meter (1960) resulted in increasing the accuracy with which units of length can be reproduced by a factor of almost 100.

The kilogram (kg) is a unit of mass (m) represented by the mass of the international prototype of the kilogram (a platinum weight with a mass close to the mass of one cubic decimeter of water at 4°C).

The second (sec) is a unit of time (t), the 1/31556925.9747 part of a tropical year for 1900, January, 0 to 12 hours ephemeral time.

The transition from the determination of the second as 1/86400 part of mean solar days to the present determination (1960) is related to the establishment of the presence of irregular changes in the period of rotation of the earth on its axis. The tropical year is the interval between two vernal equinoxes. The astronomical determination of the date corresponds to noon and 31 December 1899, while ephemeral time is taken as the time used to compute the ephemerides, the coordinates of heavenly bodies.

The ampere (a) is a unit of current strength (I), the strength of an unchanging current which, when flowing along two straight parallel conductors of infinite length and of negligibly small cross section, located at a distance of 1 meter apart in a vacuum, would create a force between these conductors equal to 2×10^{-7} units of force in the International System for each meter of length.

Table 14.1

Generally Used Derived Units in the MKSA System (rationalized)

Name of physical magnitude	Formulas for dependency	Dimension of Units	Name of unit
	ELECTRICAL MAGNITUDES		
Quantity of Electricity	Q=It	sec x a	Coulomb (c)
Electrical Voltage	U=A*/Q	$m^2 \cdot kg \cdot sec^{-3} \cdot a^{-1}$	volt (v)
Electrical Power	P=UI	$m^2 \cdot kg \cdot sec^{-3}$	watt (w)
Electrical Capacity	C=Q/U	$m^{-2} \cdot kg^{-1} \cdot sec^4 \cdot a^2$	farad (f)
Electrical Resistance	R=U/I	$m^2 \cdot kg \cdot sec^{-3} \cdot a^{-2}$	ohm (ω)
Electrical Conductivity	g=1/R	$m^{-2} \cdot kg^{-1} \cdot sec^3 \cdot a^2$	Siemens (S)
Electrical Field Intensity	E=U/a	$m \cdot kg \cdot sec^{-3} \cdot a^{-1}$	v/m
Dielectric Constant	ϵ_a=aC/S*	$m^3 \cdot kg^{-1} \cdot sec^4 \cdot a^2$	f/m
	MAGNETIC MAGNITUDES		
Magnetic Induction Flux	Φ	$m^2 \cdot kg \cdot sec^{-2} \cdot a^{-1}$	Weber (W)
Magnetic Induction	B=Φ/S*	$kg \cdot sec^{-2} \cdot a^{-1}$	Tesla (T)
Magnetic Field Intensity	H=I/2πa	$m^{-1} \cdot a$	a/m
Inductance	L=-E_i*/di/dt	$m^3 \cdot kg \cdot sec^{-2} \cdot a^{-2}$	henry (h)

Complete tables of derived units are contained in GOST 8033-56, "Electrical and Magnetic Units."

Multiples and fractions of units (Table 14.2) can be formed by multiplying, or dividing, the basic or the derived, units by a number which is a power of base 10.

---

* A-work; S-area; E_i-emf of self-induction.

Table 14.2

Prefixes for Forming Multiples and Fractions of Units (GOST 7663-55)

Prefixes for multiple units	Relationship to basic unit	Designation	Prefixes for multiple units	Relationship to basic unit	Designation
tera	10^{12}	T	deci	10^{-1}	d
giga	10^{9}	G	centi	10^{-2}	c
mega	10^{6}	M	milli	10^{-3}	m
kilo	10^{3}	k	micro	10^{-6}	M
hecto	10^{2}	h	nano	10^{-9}	n
deca	10	da	pico	10^{-12}	p
			femto*	10^{-15}	f
			atto*	10^{-18}	a

Bels and Decibels

Bels and decibels are logarithmic units of measurement in which it has been accepted to express power or potential gradients (magnitudes of amplification, attenuation, decay, and the like).

The number of bels, S, can be found as the common logarithm of the ratio of two compared powers

$$S = \log P_2/P_1. \tag{14.2}$$

When $P_2 > P_1$ (amplification), S is +, when $P_2 < P_1$ (attenuation), S is -.

A decibel is 1/10th of a bel. The power gradient in decibels can be computed through the formula

$$S = 10 \log P_2/P_1. \tag{14.3}$$

A magnitude in decibels is often read from a conditional initial level, such as 1 watt, 1 megawatt, or 10 microwatts.

The power gradient in terms of potential (given the condition of equality of the resistances on which these powers are based) can be found through the formula

$$S = 20 \log U_2/U_1. \tag{14.4}$$

---

* The prefixes "femto" and "atto" will be brought in as additional prefixes in accordance with the recommendations of the International Committee on Measures and Weights (1962).

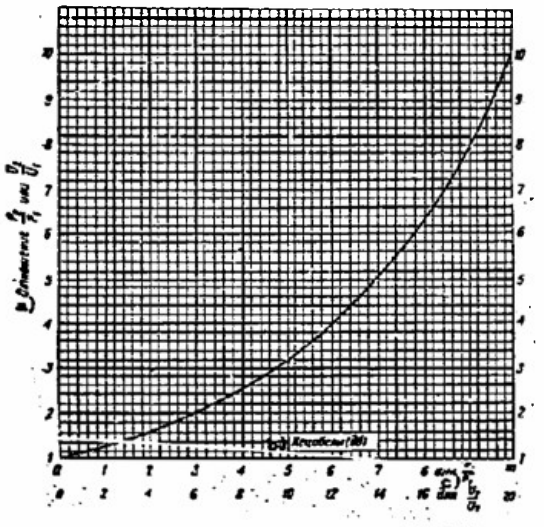


Figure 14.1. Graphic for finding power gradient in decibels.

a) ratio P_2/P_1 or U_2/U_1; b) decibels (db); c) for.

A log slide rule, or tables, can be used to find the logs of the numbers when making the computations.

Special tables and graphics can also be used to find the power gradient directly in decibels (see fig. 14.1 and Table 14.3).

In this case the power ratio must be represented as the product of a factor (less than 10) by a power of 10. Then the power gradient can be found as the sum of decibels corresponding to the factor (from the graphic) and the decibels corresponding to the powers of ten (from the table).

For example

$$P_1 = 300 \text{ watts}, \quad P_2 = 50 \text{ microwatts};$$

$$P_2/P_1 = 1.67 \cdot 10^{-7},$$

$$S = 2.2 - 70 = -67.8 \text{ db}.$$

Table 14.3

Finding the Power Gradient in Decibels

a) или $\frac{P_2}{P_1}$ ($\frac{U_2}{U_1}$)	b) Перевод мощности $S=10\lg\frac{P_2}{P_1}$ [дб] c)	b) Перевод мощности $S=20\lg\frac{U_2}{U_1}$ [дб] c)	a) или $\frac{P_2}{P_1}$ ($\frac{U_2}{U_1}$)	b) Перевод мощности $S=10\lg\frac{P_2}{P_1}$ [дб] c)	b) Перевод мощности $S=20\lg\frac{U_2}{U_1}$ [дб] c)
10^{10}	100*	200	10^{-1}	— 10	— 20
10^{9}	90	180	10^{-2}	— 20	— 40
10^{8}	80	160	10^{-3}	— 30	— 60
10^{7}	70	140	10^{-4}	— 40	— 80
10^{6}	60	120	10^{-5}	— 50	—100
10^{5}	50	100	10^{-6}	—60	—120
10^{4}	40	80	10^{-7}	— 70	—140
10^{3}	30	60	10^{-8}	— 80	—160
10^{2}	20	40	10^{-9}	— 90	—180
10	10	20	10^{-10}	—100	—200
1	0	0			

a) or; b) power gradient; c) [db].

14.2. Measurement Errors

Any measurement, no matter how accurate, and how carefully made, will have errors. Conversely, identical results from repeated measurement of a single magnitude indicate low accuracy on the part of the meters used, for it is assumed that small changes are hard to detect.

The absolute error in a measurement, x_i, is the difference between the instrument reading, a_i and the true value of the magnitude being measured, X

$$x_i = a_i - X. \tag{14.5}$$

The absolute error in metrology is designated by the symbol Δ.

A correction factor is the absolute error taken with the opposite sign

$$-\Delta = X - a_i. \tag{14.6}$$

The relative error is the ratio of the absolute error to the true value of the magnitude being measured

$$\delta = \frac{\Delta}{X}\cdot 100\% \approx \frac{\Delta}{a_i}\cdot 100\%. \tag{14.7}$$

When the relative error is small it is more convenient to express it directly in relative magnitudes, using the form of inscription $\pm 3\cdot 10^{-5}$.

A reduced error is the ratio of the absolute error to the upper (face) value of the readings for the particular meter

$$\delta_r = \Delta/a_f \cdot 100\%. \tag{14.8}$$

Errors can be broken down into three groups on the basis of the characteristic in the behavior pattern of their occurrence.

1. Systematic errors remaining during the measurement process, constant or changing in accordance with some definite law. They can be broken down into instrumental, the result of the limited accuracy of the meters, personnel, the result of the imperfections in the sense organs of the observer, external, the result of changes in the external environment (temperature, moisture, pressure, etc.), and errors in method, the result of inadequate development of the method, or the result of incomplete knowledge of all the factors associated with the particular measurement.

These must be avoided by eliminating the reasons why they occur, or by the introduction of correction factors. The presence of systematic errors fixes the correctness of the measurements.

2. Random errors, unavoidable when making measurements, and relatively small in magnitude. In the case of precise measurements the mathematical processing of a series of individual results (a_i) of the measurement of the same magnitude should be taken into consideration. The most probable result of the measurement is the average of an arithmetical series of measurements

$$a_{av} = \frac{\sum_{i=1}^{n} a_i}{n}, \tag{14.9}$$

where

n is the number of repeated measurements.

The mean square error in the average arithmetical value equals

$$\sigma_{series} = \pm \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n(n-1)}}, \tag{14.10}$$

where

$\varepsilon_i = a_i - a_{av}$ is the residual error which, in the case of a sufficiently large number of measurements, fixes the random error as a result of the measurement.

The presence of random errors fixes the accuracy of the measurements

3. Blunders (gross errors) occur as a result of miscountings considered probable during measurements. Blunders are measurements with

$$\varepsilon_i > 3\sigma_{series} \tag{14.11}$$

Blunders must be discarded. Their absence determines the validity of the measurements.

The error listed on the meter's certificate is used to evaluate the error contained in one-time engineering measurements.

In the case of indirect measurements, when the result is found by computations based on direct measurements of auxiliary magnitudes, the error can be evaluated through the formula

$$\delta = \pm \sqrt{\sum_{k=1}^{m} \delta_k^2}, \qquad (14.12)$$

where

δ_k is the error in measuring one of the m auxiliary magnitudes.

For example, when voltages are measured with a vacuum tube voltmeter type V4-2 (VLI-3), with a 1:10 divider there are two independent errors; the basic error in the measurement (±3% of the upper limit of the scale) and the error in the voltage divider (±2% of the magnitude measured). Now, if the measurement is of 400 volts on the 50 x 10 = 500 volts scale, we have a basic measurement error of

$$\delta_U = \pm \frac{3 \cdot 500}{100} = \pm 15 \text{ volts}$$

and a divider error of

$$\delta_d = \pm \frac{2 \cdot 400}{100} = \pm 8 \text{ volts.}$$

The summed error is $\delta = \pm \sqrt{15^2 + 8^2} = \pm 17$ volts and the result of the measurement should be written U = 400 ± 17 volts.

In many cases the relative errors are expressed in decibels, rather than in percentages. In order to convert the values of the errors from one of the magnitudes to the other we can use the formulas

$$\delta[\text{db}] = 10 \lg \left(1 + \frac{\delta[\%]}{100}\right); \qquad (14.13)$$

$$\delta[\%] = 100 \left[\text{antilg}\left(\frac{\delta[\text{db}]}{10} - 1\right)\right] \qquad (14.14)$$

As is often the case in practice, the values are tabulated as in Table 14.4.

Table 14.4

Conversion of errors from decibels into percentage

db	0,01	0,1	0,15	0,2	0,5	0,8	1,0	1,2	1,5
%	0,2	2,3	3,5	4,7	12,2	20,2	25,9	31,8	41,3

14.3. Direct-reading electrical instruments

Direct-reading electrical instruments are the group of devices and mechanisms for making direct measurements of some electrical magnitude. These instrumenst have a needle, or an optical device, and a scale for reading the indication.

The instruments in the electrical-mechanical groups are the ones most widely used in practice. They can be broken down into systems such as moving-coil, moving-iron, electrodynamic, induction, and electrostatic.

Galvanometers (the most sensitive of all the instruments) and logometers (meters for measuring the relationship between two magnitudes) too are built on these systems.

Basic engineering features of the various systems are listed in Table 14.5.

The accuracy class to which a direct-reading instrument belongs is characterized by the maximum basic reduced error, expressed in precent. Instruments are broken down into eight accuracy classes: 0.05; 0.1; 0.2; 0.5; 1.0; 1.5; 2.5; 4.0.

The maximum absolute error in measurements made by a direct-reading instrument is constant for the entire scale, but the relative error increases with reduction in the magnitude measured.

The system to which the instrument belongs, the accuracy classification, type of current, and many of the other of the instrument's characteristics are designated by conventional symbols on the face (usually on the scale).

The main conventional symbols, conforming to GOST 1845-59, are listed in Figure 14.2

The conventional symbols from preceding GOSTs, differing from those indicated, are shown in Figure 14.3.

Table 14.5

Basic Engineering Characteristics of the Various Systems

System to which meter belongs	Type of current measured	Scale	Area of Application	Advantages	Disadvantages
moving-coil	direct	uniform	most direct current measurements. In combination with converters, measurements in high frequency circuits	highly accurate and sensitive. Low power requirement	high cost
moving-iron	direct and low frequency alternating	nonuniform	engineering measurements	simplicity and low cost, capacity to withstand overloading	low accuracy
electrodynamic	direct and low frequency alternating	nonuniform	precision alternating current measurements	highly accurate for direct and alternating currents	low stability in the face of overloads, energy requirement high
induction	low frequency alternating	—	electrical energy counters	large rotational moment, not sensitive to overloads	low accuracy
electrostatic	direct and low and high frequency alternation	nonuniform	measurement of high voltage direct current and high frequency voltages	low energy requirement	low accuracy, cumbersome
galvanometers in the moving-coil	direct	uniform	measurement of extremely low currents	high degree of sensitivity	difficult to use under field conditions
logometers	the relationship between two direct or alternating currents	depends on system	measurement of magnitudes not carriers of electrical energy (resistance, capacitance, induction, frequency, phase and the like)	—	—

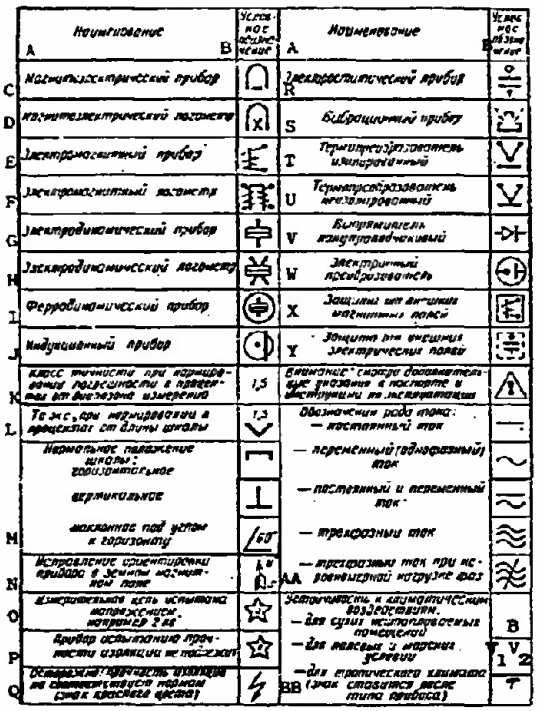


Figure 14.2. Main conventional symbols on instrument scales (GOST 1845-59).

A. Item

B. Con sym

C. Moving-coil instrument

D. Moving-coil logometer

E. Moving-iron instrument

F. Moving-iron logometer

G. Electrodynamic instrument

H. Electrodynamic logometer

I. Ferrodynamic instrument

J. Induction instrument

K. Accuracy class for standardization of errors in percentage of measurement range

L. Same, for standardization in percentage of length of scale

M. Normal scale position:
horizontal
vertical
tilt at an angle to the horizon of

N. Direction in which instrument is oriented in the earth's magnetic field

O. Measurement circuit for testing with a voltage of approximately 2 kilovolts

P. Instrument will not be required to test insulation strength

Q. Danger! Insulation strength does not correspond to standards (symbol in red)

R. Electrostatic instrument

S. Vibration meter

T. Thermal converter, insulated

U. Thermal converter, non-insulated

V. Rectifier, semiconductor

W. Electronic converter

X. Protection against external electrical fields

Y. Protection against external electrical fields

Z. Attention! See supplemental readings on the instrument card and in the operating instructions

AA. Symbol for type of current:
direct current
alternating (single phase) current
direct and alternating currents
three-phase current
three-phase current with unequal phase loads

BB. Climatic stability:
for dry, unheated spaces
for field and marine conditions
for a tropical climate (symbol placed after instrument type)

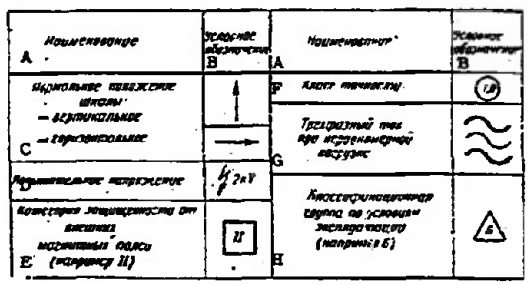


Figure 14.3. Conventional symbols on the scales of instruments from preceding GOSTS.

A. Item

B. con sym

C. Normal scale position:
vertical
horizontal

D. Test voltage

E. Category of protection against external magnetic fields (example II)

F. Accuracy class

G. Three-phase current in the case of uneven load

H. Classification group according to operational conditions (example B)

14.4. Measurement of electrical parameters

Table 14.6 lists data on the principal methods used to measure the various electrical parameters.

14.5. Oscillography

Loop (up to 10,000 hertz) and electron oscillographs (up to tens and hundreds of megahertz) are used to observe and investigate the shapes of complex oscillations.

Electron oscillographs are used to observe the shapes of, and to measure, the amplitudes and durations of periodic and pulse voltages and currents, the amplitude modulation factor, the oscillation frequency and the phase shift, and to investigate various characteristics.

The modern electron oscillograph is a complicated instrument and a set of basic rules must be observed when it is used:

the type and frequency (duration) of the sweep must be selected in accordance with the process being investigated;

amplifier must not be overloaded; it is always best to work with the maximum division factor at the input;

the magnitude of the oscillogram should be set by adjusting the gain to 2/3 to 3/4 the size of the screen;

minimum synchronized amplitude should be used;

should always work with minimum brilliance and good focus, since this will not only preserve the tube, but will also improve measurement accuracy.

The only way to investigate pulse voltages with high pulse ratios is by using a driven sweep.

Radio frequency pulses with a pulse rate in excess of the upper limit of the oscillograph pass band can only be observed after preliminary detection. The envelope of the radio frequency pulse is observed.

Voltage is measured by making a comparison with a known, calibrated, sinusoidal voltage forming the same size vertical oscillogram as the voltage under investigation.

Current is measured indirectly by the voltage drop across a known resistance.

Width is measured by counting the number of calibrated markers with known width formed by the sweep during the supply of a sinusiodal voltage from the width calibrator to the control grid (or cathode) of the tube.

The measurement error in the case of the oscillograph is an average ±5 to 10%.

Table 14.6

Principal Methods Used to Measure Various Electrical Parameters

Parameter	Measurement Method	Instruments Used	Advantages	Disadvantages
direct current and low frequency current	direct measurement with direct-reading instruments	ammeters, moving-coil (direct current only), moving-iron and electrodynamic systems	simplicity of measurement	dependence of error on frequency
high frequency current	conversion into direct current through the use of thermocouples	ammeters, thermoelectric system	little dependence of readings on frequency	squareness of scale, unacceptability of overloads
constant voltage and low frequency voltage	direct measurement with direct-reading instruments	voltmeters, moving-coil (constant voltage only), moving iron, electrodynamic, and electrostatic systems	simplicity of measurement	dependence of error on frequency
	compensation method	potentiometers (compensators)	small error, high input impedance	instrument complexity, time required to take measurements
high frequency voltage	direct measurement	voltmeters, electrostatic system	low energy requirement	accuracy poor
	conversion into a constant voltage through the use of solid rectifiers and vacuum tubes	voltmeters with solid rectifiers	simplicity of measurement	accuracy poor, dependence of readings on frequency and temperature
		vacuum-tube voltmeters	highly stable, high input impedance	accuracy poor, need for source of supply

Table 14.6 (continued)

Principal Methods Used to Measure Various Electrical Parameters

Parameter	Measurement Method	Instruments Used	Advantages	Disadvantages
power of direct current and low frequency current	direct measurement	wattmeters, electrodynamic	highly accurate	dependence of readings on frequency
	indirect measurement	direct reading voltmeters and ammeters	no need for special instruments	accuracy poor
power in the high and super-high frequency	measurement of absorbed power			
	indirect	vacuum-tube voltmeters and thermal ammeters	measurement limits 10 milliwatts-500 watts at frequencies up to 100 megahertz	error ± (10 to 15)%
	photometric	photometric wattmeters	measurement limits up to 150 watts at frequencies up to 3000 megahertz	error ± (10 to 15)%
	calorimetric	calorimetric wattmeters	measurement of high powers over a broad band of frequencies, error ±(3 to 10)%	cumbersome, inertness
	bolometric	bolometer and thermistor bridges, vacuum-tube voltmeters	measurement of low powers over a broad band of frequencies	error ± (10 to 20)%
	measurement of power passing through			
	directional branchings	voltmeters with solid rectifiers	simplicity of measurement	dependence of reading on match to line

Table 14.6 (continued)

Principal Methods Used to Measure Various Electrical Parameters

Parameter	Measurement Method	Instruments Used	Advantages	Disadvantges
	directional branchinge	bolometer bridges	highly accurate	narrow operating band of frequencies
	absorbing wall	direct current bridges	simplicity of measurement	high degree of inertness
	through the use of pondermotive forces	specially built wattmeters	highly accurate	only high powers can be measured under field conditions
	through the use of Hull effect	spscially buily wattmeters	independence of readings from match to line	reduction in the electrical strength of the line
low frequencies	direct measurement	frequency meters, vibrating, ferrodynamic, and vacuum-tube systems	simplicity of measurement	accuracy poor
	bridge method	alternating current bridges	highly accurate	complexity and highly time consuming
	stroboscopic method	stroboscopes and sources of reference frequency	highly accurate	complexity and highly time consuming
	Lissajou figures method	electron oscillograph and sources of reference frequency	highly accurate	complexity and highly time consuming
high and super-high frequencies	resonance method	resonance-frequency meters (frequency meters)	simplicity of measurement	accuracy poor

Table 14.6 (continued)

Principal Methods Used to Measure Various Electrical Parameters

Paramoter	Measurement Method	Instruments Used	Advantages	Disadvantages
	beat method	heterodyne frequency meters, crystal calibrators	highly accurate	complexity and highly timo consuming
direct current resistance	indirect measurement	ammeters and voltmeters	no special instruments required	accuracy poor
	direct measurement	ohmmeters and ammeter-voltmeters	simplicity of measurement	dependence of readings on supply source voltage
		megohmmeters based on logometers	independence of readings on supply source voltage	complexity of the instrument
	bridge method	direct current bridges	highly accurate	complexity and highly time consuming
capacitance in the low frequency range	direct measurement	low frequency capacity meters	wide range of measurements	accuracy poor
		faradmeters based on logometers	highly accurate	need for alternating current net
capacitance and inductance in low frequency range	bridge method	alternating current bridges	highly accurate	complexity and highly time consuming

Table 14.6 (continued)

Principal Methods Used to Measure Various Electrical Parameters

Parameter	Measurement Method	Instruments Used	Advantages	Disadvantages
resistance, capacitance, inductance and Q in the high frequency range	resonance method	Q-meters	measurement at operating frequency	accuracy poor

14.6. Measuring generators

Special sources of supply, called measuring generators, are widely used as auxiliary devices for making electrical and radio frequency measurements.

A brief listing of data for the main groups of measuring generators is contained in Table 14.7.

14.7. Basic measurements in electronic equipment

Measuring instruments are used during the operation and troop repair of electronic equipment for:

monitoring the match between the main parameters of the equipment as a whole, and of the individual units, with their rated values;

tuning the equipment;

detecting faulty operation.

The results of the measurements are compared with conditions and parameters set forth in the service manuals, or in the operating instructions, and particularly with the so-called voltage and resistance charts which show the rated values of these parameters for different points in the circuit. It should be borne in mind that the voltage charts are made up for taking measurements with an instrument with a definite internal resistance (usually one of the types of ampere-voltmeters), so when the measurements are taken with other instruments the readings can differ considerably from those indicated on the chart.

One of the basic requirements imposed on electrical and radio frequency measuring devices is that there must be no interference with the circuit under investigation when the devices are inserted in the circuit. This requirement can be satisfied when the input, or output, impedance of the instrument, Z_{in}, is equal to the resistance of the circuit when disconnected at the time of measurement, Z_{dis}:

$$Z_{in} = Z_{dis}. \tag{14.15}$$

When voltmeters and oscillographs are inserted in a circuit from which nothing has been disconnected ($Z_{dis} = \infty$) this requirement reduces to a need to increase the instrument input impedance. This impedance should be minimal for ammeters ($Z_{dis} = 0$).

When other measuring devices are used the requirement that $Z_{in} = Z_{dis}$ is satisfied if special transient matching devices are used.

Transient matching devices in the low and high frequency ranges are made in the form of a combination of lumped pure resistance and capacitance, computed accordingly. Examples of transient matching devices for this range are shown in Figure 14.4

Table 14.7

Main Groups of Measuring Generators

Name of Group	Oscillation Generated	Purpose	Basic Engineering Characteristics
audio and video frequency generators	sinusoidal unmodulated oscillations	check, tune audio and video frequency channels, investigate load and characteristics	the frequency range from one hertz to 10 megahertz. Set accuracy and stability ± 1 to 2%
signal generators	sinusoidal unmodulated and amplitude modulated oscillations	investigate radio frequency receivers and the supply to the circuits when the measurments are made in the high and superhigh frequency band	the frequency range from tens of kilohertz to tens of gigahertz. Set accuracy and stability ± 1 to 3%; Modulated by low frequency and square pulses
frequency-sweep generators	frequency modulated sinusoidal oscillations	visual tuning and investigation to load ends in the frequency band	the frequency range from tens of kilohertz to thousands of megahertz. Frequency band up to tens of megahertz
noise generators	uniform level noise ("white noise")	measurement of radio frequency receiver sensitivity	the frequency range from tens of megahertz to tens of gigahertz. Spectral power density of noise up to 100 kT_o
pulse generators	square video pulses	investigate and tune pulsers	the pulse width range from units nanoseconds to hundreds of milliseconds, in repetition rate from tens of hertz to tens of kilohertz, in amplitude to hundreds of volts.

When the transient matching devices contain R_d or C_d, the voltage across the transient device output will differ from the voltage across its input, and this has to be taken into consideration in many measurements.

When the measuring device (such as a measuring generator, for example) must be used under a definite load this too must be taken into consideration when computing the transient matching device.

Two types of transient matching devices are used in the superhigh frequency:

through matching devices (linear sonde, directional coupler, measuring antenna);

direct matching devices (waveguides, coaxial, and combination transitions).

The following instrument measurements, general in nature, are made during operation and troop repair of electronic equipment:

a check to see that fuses are in good condition;

measurement of supply voltages;

a check to see that circuits and contracts are in good condition;

measurement of the resistances between individual points in the circuit;

a check of the manner in which vacuum tube and semiconductor meters are functioning;

investigation of the oscillograms at the monitoring jacks and at other points in the circuit;

verification of the serviceability of vacuum tube and semiconductor meters;

measurement of the parameters of resistors, condensers, and other components.

There are individual devices in electronic equipments that require the making of a number of specific measurements.

Basic information on these measurements is listed in Table 14.8.

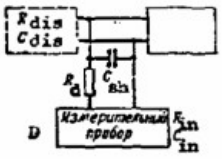
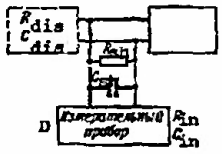
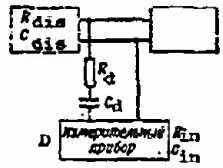
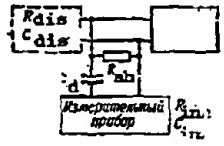
A Соотношение сопротивлений и емкостей	B Расчетные формулы	C Схема ПСУ
$R_{in} < R_{dis}$ $C_{in} < C_{dis}$	$R_d = R_{dis} - R_{in}$ $C_{sh} = C_{dis} - C_{in}$	
$R_{in} > R_{dis}$ $C_{in} < C_{dis}$	$R_{sh} = \frac{R_{dis} R_{in}}{R_{in} - R_{dis}}$ $C_{sh} = C_{dis} - C_{in}$	
$R_{in} < R_{dis}$ $C_{in} > C_{dis}$	$R_d = R_{dis} - R_{in}$ $C_d = \frac{C_{dis} C_{in}}{C_{in} - C_{dis}}$	
$R_{in} > R_{dis}$ $C_{in} > C_{dis}$	$R_{sh} = \frac{R_{dis} R_{in}}{R_{in} - R_{dis}}$ $C_d = \frac{C_{dis} C_{in}}{C_{in} - C_{dis}}$	

Figure 14.4. Circuits and formulas for transient matching devices.

A - resistance and capacitance relationships

B - formulas

C - transient matching device circuits

D - read instrument

Table 14.8

Measurements in Electronic Equipment

Name of Device	Parameter Measured	Measured Method	Equipment Used
trans-mitting	electrical conditions of pulsed magnetrons	direct measurement of voltage and average magnetron current	built-in, or assigned instruments for making direct measurements
	magnetic induction in magnet gaps	direct measurement of magnetic induction	magnetic induction instruments
	frequency of generated oscillations	resonance method	absorption frequency meters (wavemeter)
	frequency stability	frequency measurement	absorption frequency meters
		investigation of the shape of the radio frequency pulse	oscillographs
		analysis of energy spectrum	spectrum analyzers
	generated power	indirect	vacuum tube voltmeters, Nadenko voltmeters
		directional couplers	voltmeters with solid rectifiers, bolometer bridges
		absorbing walls	direct current bridges
	pulse shape	oscillographic investigation	cathode ray oscillographs
	energy spectrum	sequential analysis with a narrow band device	reference resonators, oscillographic spectrum analyzers
antenna-feeder	channel match	measurement of the standing-wave ratio	directional couplers with thermistor bridges measuring lines
	electrical conditions of resonant dischargers:		

Table 14.8 (continued)

Measurements in Electronic Equipment

Name of Device	Parameter Measured	Measured Method	Equipment Used
	- current in the ignition circuit	direct measurement	microammeters in the moving coil system
	- leakage power	direct measurement	thermistor bridges
	- losses under reception conditions	substitutioo method	indicators with solid rectifiers and calibrated attenuators
	antenna pattern	measurement of incoming signal level while rotating the antenna under investigation	signal generators and indicators with solid rectifiers
receiving	traveling-wave tube electrical conditions	direct measurement of filament voltage, voltages across tube elements including their stability and pulsations, collector and second anode currents	built-in or assigned direct reading instruments
	basic parameters of crystal mixers	direct measurement of forward and back resistances, measurement of sensitivity at the operating frequency	crystal detector testers, ampere-voltmeters, signal generators, and microammeters in the moving coil system
	reflex klystron electrical conditions	direct measurement of filament voltage, voltages across resonator and reflector including their stability and pulsations, cathode current	direct reading instruments
		direct measurement of power and frequency in accordance with voltage across the reflector	thermistor bridges and resonance frequency meters (absorption frequency meters)

Table 14.8 (continued)

Measurement in Electronic Equipment

Name of Device	Parameter Measured	Measured Method	Equipment Used
	input circuit frequency characteristics (meter and decimeter bands)	direct measurement of the voltage across the output in the case of constant voltage across the input and a changing frequency	signal generators and vacuum tube voltmeters
	amplitude characteristic, pass band, and gain of the intermediate-frequency amplifier	direct measurement of the voltage across the output in the case of a change in the voltage across the input or of the frequency	signal generators and vacuum tube voltmeters
	gain, amplitude and frequency characteristics of video amplifiers. Quality of reproduction of short and long square pulses	direct measurement of the voltage across the output in the case of a change in the voltage across the input or of the frequency, oscillographic investigation	video frequency generators and vacuum tube voltmeters or oscillographs, pulse generators and oscillographs
	receiver sensitivity	comparison method	signal generators and microammeters in the moving coil system or oscillographs
	receiver noise figure	comparison method	noise generators and microammeters in the moving coil system
indicating	pulse width and amplitude of the sawtooth voltage in the sweep units	oscillographic investigation	cathode ray oscillographs
	scanning linearity	oscillographic investigation	range calibrators or own marking units
	mark frequency	comparison method	specimen quartz-crystal oscillators and oscillographs

Table 14.8 (continued)

Measurement in Electronic Equipment

Name of Device	Parameter Measured	Measured Method	Equipment Used
	amplitude of marks in mark units	oscillographic investigation	cathode ray oscillographs
	length and limits of change in delay in delay units	oscillographic investigation	cathode ray oscillographs
supply source	electrical parameters of primary supply sources	direct measurement of output voltage, summed load current, and frequency of generated voltage	built-in direct reading instruments
	parameters of secondary supply sources		
	-circuit insulation resistance	direct measurement	megohmmeters based on logometers
	- rectified voltage and load current	direct measurement	ampere-voltmeters, electrostatic voltmeters, vacuum tube direct current voltmeters
	- voltage and ripple factor	direct measurement of the variable component across the filter output	vacuum tube millivoltmeters
	stabilization factor	direct measurement of the voltage across the output in the case of a change in the voltage across the input	ampere-voltmeters
	power demand	indirect or direct measurement in the case of normal load	voltmeters and ammeters of the moving coil type, wattmeters of the electrodynamic type
	idle current	direct measurement in the case of a disconnected load	ammeters of the moving iron type

14.8. Operational monitoring of basic electronic equipment parameters

Since the basic engineering parameters of electronic equipment undergo changes in operation, they must be monitored systematically. The monitoring system fulfills the following two requirements:

the monitoring of the basic engineering parameters with the required accuracy;

in the event the parameters being monitored deviate from rated values, it finds the reason for the deviation.

The methods most widely used to monitor radar are;

individual parameter measurements;

direct monitoring of the power gradient.

The individual parameter measurement method ("instrumental moni oring") is carried out by using a specific set of measuring devices capable of making the necessary measurements with specified accuracy.

The set includes general purpose instruments (assigned to the radar, or available in the monitoring and repair shops), special combination radar measuring instruments, as well as built-in measuring instruments.

The radar measuring instruments include:

a signal generator for the specified frequency range;

power measuring device;

frequency meter;

spectrum analyzer;

standing wave ratio measuring device.

The measuring devices built directly into the monitoring equipment can be broken down into two groups:

built-in radio frequency measuring devices (frequency meters, oscillographs, power measuring devices, and the like);

built-in direct reading measuring devices for monitoring currents and voltages.

The built-in devices can be used to monitor a single parameter, or, what is more convenient, to monitor several parameters by appropriate switching.

It is desirable to monitor the following when the radar is in operation:

transmitter pulse power;
receiver sensitivity or noise figure;
} while the radar is in operation

the match in the antenna-feeder channel;
transmitter and oscillator operating frequencies;
magnetron conditions;
} immediately prior to operation

the pulsa energy spectrum;

the shape of the modulating and high-frequency pulse;

detector currents (mixer and automatic frequency control system);

heterodyne conditions;

output, and in part the intermediate, voltages and currents in the units

} when doing adjustment work

The instruments most widely used, as well as those used for the greatest variety of purposes, during the detection process are the direct-reading instruments. Measuring instruments are usually used to check assumptions as to the sites of faults, as well as to make final detection.

There is a need to automate monitoring during the operational process because of the complexity of modern electronic equipment, a complexity that has converted the equipment into a set of interconnected units with tens of thousands of instruments and components, as well as for purposes of improving requirements imposed for operational reliability and battle readiness.

Only an automatic monitoring system, and the widespread use of built-in measuring equipment, can successfully solve the problem of providing timely and objective checks of modern sets of electronic equipment.

Automatic monitoring systems carry out the following functions:

continuous or discrete (periodic) monitoring of output parameters of the equipment as a unit, as well as of the parameters of the most important of the functional units (accessible, and the possibility of making periodic measurements of the absolute values of individual parameters);

establish the faulty functional removable units (in the case of modular design, the removable modules);

check the conditions in functional units and stages (establishment of the site of and reasons for the faulty operation);

forecase faulty operations that by their nature can be classed as gradual failures.

The method that directly monitors the power gradient involves the use of so-called reference resonators (cavity resonators with high Q) connected to the radar's antenna-feeder channel through a transient matching device. As the main pulse is radiated impulsing of the resonator occurs and attenuated oscillations are seen on the indicator screen prior to merging into the receiver noise. The ring time (the duration of the damping signal) is directly proportional to the radar's jump in power, expressed in decibels:

$$t_{ring} = k\ P_p/P_{in\ min} = kS. \qquad (14.16)$$

These instruments, in addition to monitoring the jump in power, monitor (with what is comparatively low accuracy):

transmitter power and frequency;

pulling of the magnetron;

pulse energy spectrum;

receiver tuning and sensitivity;

local oscillator power and frequency;

tuning and proper functioning of the antenna transfer switch.

14.9. Classification of electrical and radio frequency measuring instruments. Radio repair shop measuring instruments.

Soviet industry mass produces hundreds of different types of electrical and radio frequency measuring instruments.

In accordance with standard specifications contained in No.009.000, issued by the State Committee for Radioelectronics of the Council of Ministers of the USSR, radio frequency measuring instruments for general use are classified by purpose into groups that are assigned specific letters. The instruments within each group are broken down by specific features, and a number is assigned to each subgroup. A concrete model of an instrument receives it's own number, indicated in the instrument designation after the subgroup (after a hyphen); V7-2, Ye-12-1, S1-6, etc., for example.

Table 14.9 lists the designations by groups.

Table 14.9

Groups of Measuring Instruments

Group Designation	Group Name	Subgroup No.	Instrument Name
A	instruments for measuring current	1	installation for checking ammeters
		2	ammeters, direct current
		3	ammeters, alternating current
		4	all purpose ammeters
V	instruments for measuring voltage	1	installation for checking voltmeters
		2	voltmeters, direct current
		3	voltmeters, alternating current

Table 14.9. (continued)

Groups of Measuring Instruments

Group Designation	Group Name	Sub-Group No.	Instrument Name
		4	voltmeters, pulse
		5	voltmeters, phase-sensitive
		6	voltmeters, selective
		7	voltmeters, all-purpose
M	instruments for measuring power	1	installations for checking power meters
		2	passing power meters
		3	absorbed power meters
		4	bridges for measuring thermistor and bolometer power
		5	heads, thermistor and bolometer
Ye	instruments for measuring parameters in devices with lumped constants	1	installations for monitoring and checking parameter meters
		2	pure resistance measures
		3	inductance measures
		4	capacitance measures
		5	conductivity measures
		6	pure resistance meters
		7	inductance meters
		8	capacitance meters
		9	Q-meters
		10	impedance and conductance meters
		11	meters for measuring the electrical and magnetic properties of materials
		12	parameter meters, all-purpose
R	instruments for measuring parameters in devices with disturbed constants	1	lines, measuring
		2	standing wave ratio and reflection factor meters
		3	impedance and conductivity meters

Table 14.9. (continued)

Groups of Measuring Instruments

Group Designation	Group Name	Sub-Group No.	Instrument Name
		4	attenuation meters
		5	cable line meters and testers
Ch	instruments for measuring frequency	1	installations for checking frequency meters and for reproducing specimen frequencies
		2	frequency meters, absorption
		3	frequency meters, vacuum tube-counter
		4	frequency meters, local oscillator
		5	calibrators, quartz
F	instruments for measuring phase and delay time	1	installations for monitoring and checking instruments used for phase measurements
		2	phase meters
		3	phase shifters, measuring
		4	correlation meters
		5	group time delay meters
S	instruments for observing and investigating signal and spectrum shapes	1	oscillographs
		2	amplitude modulation factor meters
		3	frequency deviation meters
		4	spectrum analyzers
		5	harmonic analyzers
		6	nonlinear distortion factor meters
Kh	instruments for observing and investigating the characteristics of radio devices	1	instruments for investigating frequency characteristics
		2	instruments for investigating transient characteristics
		3	instruments for investigating phase characteristics
		4	instruments for investigating amplitude characteristics

Table 14.9. (continued)

Groups of Measuring Instruments

Group Designation	Group Name	Sub-group No.	Instrument Name
		5	noise figure meters
I	instruments, special, for pulse measurements	1	installations for checking pulse instruments
		2	time interval (shifts, fronts, pulse width, etc.) meters
		3	pulse counters
		4	pulse analyzers
		5	delay lines
U	amplifiers, measuring	1	constant voltage amplifiers
		2	alternating voltage amplifiers, selective
		3	alternating voltage amplifiers, broadband
		4	amplifiers, all-purpose
P	instruments for measuring field intensities and interference and antenna measurements	1	installations for checking meters used to measure field intensity and noise and for checking measuring receivers
	measuring receivers	2	field indicators
		3	field intensity meters
		4	radio frequency interference meters
		5	receivers, measuring
		6	antennas, measuring
		7	instruments for antenna measurements
D	attenuators and voltage dividers	1	installations for checking attenuators
		2	resistance attenuators
		3	capacity attenuators
		4	attenuators, limit
		5	attenuators, absorbing

Table 14.9. (continued)

Groups of Measuring Instruments

Group Designation	Group Name	Subgroup No.	Instrument Name
		6	voltage dividers
Е	components in the coaxial and waveguide channels	1	transformers, matching
		2	junctions
		3	transfer switches
		4	phase shifters
		5	couplers, directional
		6	Y-s and ring bridges
		7	heads, detector and mixer
		8	components, connecting
		9	resistances, load
		10	equivalents, antenna
G	generators, measuring	1	installations for checking measuring generators
		2	generators, noise signal
		3	generators, signal
		4	generators, standard signals
		5	generators, pulse
		6	generators, signal, special shape
L	instruments for measuring the parameters of tubes and semiconductor instruments	1	meters for measuring tube parameters and cathode-ray curve tracers
		2	meters for measuring the parameters of semiconductor instruments and cathode-ray curve tracers
		3	meters for measuring superhigh frequency vacuum tubes (klystrons, traveling-wave tubes, magnetrons, and others)

Combination instruments are included in the group to which they are basic, and have the letter "K" added after the group designation, VK7-4,

GK4-49, for example. Models that have been modernized have a letter added, in alphabetical order, after the number, such as G4-10A, for example.

Old instrument names are being replaced with new ones, with the exception of those taken out of production prior to the introduction of standard 019.000. The replacement names for some of the instruments found in use are listed in Table 14.10.

Table 14.10

New and Old Names of Instruments

Name of Instrument	Conventional Designations for Instrument Types	
	New	Old
V. Instruments for measuring voltages		
millivoltmeter, pulse	V4-1A	MVI-1M
voltmeter, pulse	V4-2	VLI-3
all-purpose electrical measuring instrument	VK7-1	TT-3
voltmeter, all-purpose tube	V7-2	VLU-2
voltmeter, tube	VK7-3	A4-M2
voltmeter, all-purpose, tube	VK7-4	VOLU-1
voltmeter, number	VK7-5	VOTs-1
M. Instruments for measuring power		
meter, low power	M3-1	IMM-6
meter, low power	M3-2	VIM-1
meter, power	M3-4	IM-4
meter, high power	M3-6	IM-5
meter, high power	M3-7	IBM-4
Ye. Instruments for measuring the parameters of devices with lumped constants		

Table 14.10 (continued)

New and Old Names of Instruments

Name of Instrument	Conventional Designations for Instrument Types	
	New	Old
megohmmeter, tube	Ye6-2	MOM-3
megohmmeter, tube	Ye6-3	MOM-4
meter, quality	Ye9-1	KV-1
meter, quality	Ye9-2	UK-1
meter, magnetic induction	Ye11-1	IMI-1
meter, magnetic induction	Ye11-3	IMI-3
meter, inductance and capacitance, high frequency	Ye12-1	IIYeV-1
bridge, all purpose	Ye12-2	UM-3
R. Instruments for measuring the parameters of devices with distributed constants		
line, measuring, spiral	R1-1	ISL-1
line, measuring, coaxial	R1-2	IKL-112
line, measuring, coaxial	R1-3	IKL-111
line, measuring, waveguide	R1-4	IVLU-140
Ch. Instruments for measuring frequency		
frequency meter, absorption	Ch2-2	UVR-1
frequency meter, average accuracy	Ch2-3	VST-1
frequency meter, average accuracy	Ch2-4	VST-2
frequency meter, low accuracy	Ch2-5	VMT-1
frequency meter, average accuracy	Ch2-6	VST-DM
frequency meter, average accuracy	Ch2-9	VST-10
frequency meter, average accuracy	Ch2-9A	VST-10P

Table 14.10 (continued)

New and Old Names of Instruments

Name of Instrument	Conventional Designations for Instrument Types	
	New	Old
frequency meter	Ch3-1	ICh-6
frequency meter	Ch3-1A	ICh-7
frequency meter, heterodyne	Ch4-1	VG526-U
frequency meter, precision	Ch4-3	VVT-D
frequency meter, precision	Ch4-4	GVShD
calibrator, frequency, crystal	Ch5-1	KK-6
S. Instruments for observing and investigating the shapes of signals and spectra		
oscillograph, electron	S1-1	EO-7
oscillograph, electron	S1-2	25-I
oscillograph, electron	S1-3	IO-4
oscillograph, electron low frequency	S1-4	ENO-1
oscillograph electron	S1-5A	SI-1
oscillograph, electron miniature	S1-6	EMO-2
spectrum analyzer	S4-1	IV-46
spectrum analyzer	S4-2	IV-66
distortion factor meter	S6-1	INI-12
Kh. Instruments for observing and investigating the frequency characteristics of radio installations		
oscillating frequency generator (sweep generator)	Kh1-1	102I
noise-figure meter	Kh5-1	IKSh-1
noise-figure meter	Kh5-2	IKSh-2

Table 14.10 (continued)

New and Old Names of Instruments

Name of Instrument	Conventional Designations for Instrument Types	
	New	Old
I. Instruments, special, for pulse measurements		
range calibrator	I2-1A	27IM
U. Amplifiers, measuring		
amplifier, measuring	U2-1A	28IM
P. Instruments for measuring field intensities and radio noise		
meter, field intensity	P3-1	INP-5
meter, radio noise	P4-4A	IP12-2M
meter, radio noise	P4-5A	IP-26M
D. Attenuators and voltage dividers		
voltage divider	D6-1	DNYe-3
voltage divider	D6-2	DNYe-9
voltage divider	D6-3	DNYe-6
voltage divider	D6-4	DNYe-7
voltage divider	D6-5	DNYe-8
G. Generators, measuring		
generator, audio frequency	GZ-2	ZG-10
generator, audio and supersonic frequencies	GZ-3	ZG-11
generator, audio and supersonic frequencies	GZ-4A	ZG-2M
generator, signal, video frequency	GZ-7	100-I
generator, signal	GZ-8	GMV
generator, signal	GZ-9	GS-6
generator, signal	GZ-12	GS-23
generator, audio frequency	GZ-18	ZG-14

Table 14.10 (continued)

New and Old Names of Instruments

Name of Instrument	Convertional Designations for Instrument Types	
	New	Old
generator, audio and supersonic	GZ-33	ZG-16
generator, standard signal	G4-1A	GSS-6A
combination tester (radar tester)	GK4-3	RT-10
combination tester (radar tester)	GK4-3A	RT-10A
generator, signal, combination	GK4-4	GSK-2
generator, standard signal	G4-5	GSS-12
generator, standard signal	G4-6	GSS-17
generator, standard signal	G4-7	GSS-7
generator, standard signal	G4-8	GSS-15B
generator, standard signal	G4-9	GSS-27
generator, pulse	G5-1	26-I
generator, pulse, miniature	G5-8	MGI-1
generator, pulse, miniature	G5-15	MGI-2
L. Instruments for measuring tube and semiconductor meter parameters		
tester, radio tube, small	L1-1	IL-13
tester, radio tube	L1-2	IL-14
tester, radio tube, all-purpose, small	L1-3	MILU-1
tester, flat crystal triode parameters	L2-1	IPT-1

Radio shops, equipped with a wide variety of measuring instruments, are widely used in connection with the operation and repair of electronic equipment. Basic data on some of the instruments are listed in Table 14.11 (the old factory designations for the instruments are given in parenthesis).

Table 14.11

Basic Data On Selected Measuring Instruments

Type of Instrument	Name of Instrument	Basic Engineering Characteristics
PMT-70	switchboard ammeter, thermoelectric system	measurement limits 1 to 5 amps, accuracy class 2.5, operating frequency 50 Hz to 7.5 MHz
AVO-5M	ammeter-voltmeter ("avometr")	measurement limits up to 6000 volts, AC and DC, 60 microamps to 12 amps DC, 3 milliamps to 12 amps AC, resistance 3 ohms to 3 megohms, measurement error on DC ± 3%, on AC ± 4%, when measuring resistance ±2.5% the length of the operating part of the scale, input resistance on DC 20000 ohms per volt, on AC 2000 ohms per volt
Ts-57	ammeter-voltmeter ("avometr")	measurement limits up to 600 volts, AC and DC, up to 1.5 amps AC and DC, resistances up to 3 megohms, capacitance up to 0.3 microfarads, level of transmission, attenuation, and gain from -10 to +12 db, measurement error on DC ± 1.5%, on AC ± 2.5%, when measuring resistance ± 1.5% the length of the scale, capacitance, level of transmission, attenuation, and gain ± 2.5% length of scale; input resistance on DC 20000 ohms per volt, on AC 2000 ohms per volt
V3-10A (IVP-3A)	meter, receiver output (voltmeter with solid rectifiers)	limits of measurement from 50 millivolts to 300 volts, basic error in measurement in the frequency range from 50 Hz to 10 kHz not in excess of ± 4%, in the frequency range from 10 to 20 kHz not in excess of ± 8%, operating frequency 50 Hz to 20 kHz
V4-2 (VLI-3)	voltmeter, pulse, vacuum tube, with 1:10 divider and additional capacitive dividers D6-1 (DNYe-3) and D6-2 (DNYe-9)	limits of measurement up to 150 volts, with a 1:10 divider up to 500 volts, with a D6-1 up to 5000 volts, with a D6-2 up 50000 volts, basic error ±(3 to 4)%, divider error up to ± 10%, pure input resistance 0.5 megohm up to 4 MHz, and 0.2 megohm up to 10 MHz, input capacitance 12 picofarad
VK7-9	all-purpose vacuum tube voltmeter	limits of measurement of DC 30 millivolts to 500 volts (up to 20 kilovolts with a DN-1 divider), of AC 100 millivolts to 1000 volts, resistance 2 ohms to 1000 megohms, measurement error of DC ± 2.5%, AC ± (4 to 6)%, operating frequency 20 Hz to 700 MHz, input resistance on DC 15 megohms, on AC 3 megohms
M1103	meter, resistance, ground	limits of measurement 0.1 to 50 ohms, basic error within the 0.1 to 10 ohms range ± 5% +0.1 ohm in the 0.5 to 50 ohm range ±10% + 0.5 ohm
M1101	megohmmeter, logometer base with hand generator	limits of measurement 100, 500, 1000 megohms (for different modifications), operating voltages 100, 500, 1000 volts, respectively, basic

Table 14.11 (continued)

Basic Data On Selected Measuring Instruments

Type of Instrument	Name of Instrument	Basic Engineering Characteristics
		error ±1% of length of scale
Ch2-6 (VST-DM)	frequency meter, average accuracy	range of measured frequencies 350 to 675 MHz (85.7 to 44.4 cm), basic error ± 0.05%, sensitivity no worse than 0.2 milliwatt
S1-20	oscillograph with all-purpose sweep	sweep 0.025 microsecond/cm to 10 microseconds/cm, pass band 10 Hz to 20 Mhz and 15 Hz to 2 MHz, sensitivity 100 and 10 millivolt per cm, input resistance 0.5 megohm, input capacitance 40 picofarads, error in amplitude and width measurements ± 5%
Kh1-1 (102I)	frequency sweep generator	frequency range 10 to 100 MHz, frequency deviation 0 to 15 MHz, frequency modulation 2220 Hz, output signal level 10 microvolts to 0.1 volt
GZ-33 (ZG-16)	audio and supersonic frequency generator	frequency range 20 Hz to 0.2 MHz, output 0.5 to 5 watts, error in frequency for installation (0.02f + 1) Hz, output ±2.5% (level), ± (0.5 to 1) db (attenuator)
G4-42	generator, standard signal	frequency range 0.012 to 10 MHz, amplitude modulation possible, output 0.1 microvolt to 0.1 volt, 75 ohms, error in frequency for installation ±1%, output ±4% (level), ±11% (attenuator)
G4-44	generator, standard signal	frequency range 10 to 400 MHz, amplitude and pulse modulation possible, output 0.1 microvolt to 0.1 volt, 75 ohms, error in frequency for installation ± 1%, output ±10% (level), ±(0.7 to 1.5) db (attenuator)
G4-31 (GSS-15A)	generator, standard signal	frequency range 0.15 to 1 gigahertz, amplitude and pulse modulation possible, output 10^{-8} to 100 microwatts, 1 watt, 75 ohms, error in frequency for installation ±1%, output ±1 db (level), ±1.3 db (attenuator)
G5-15 (MGI-2)	small pulse generator	duration of generated pulses of both polarities 0.1 to 10 microseconds, repetition frequency 40 Hz to 10 kHz, output 10 to 100 volts, installation error 10% ± 0.03 microsecond

Table 14.11 (continued)

Basic Data On Selected Measuring Instruments

Type of Instrument	Name of Instrument	Basic Engineering Characteristics
L1-3 (MILU-1)	small tube tester, all purpose	measurement of following tube parameters: diodes - emission current, or anode current; triodes and combination tubes - anode current, second grid current, reverse current of first grid, steepness of the characteristic curve for the anode current, internal resistance; stabilitrons - ignition potential, stabilization voltage, change in stabilization voltage with change in current; kenotrons - rectified current
L2-1 (IPT-1)	tester, flat crystal triode parameters	measurement of the following parameters: current gain in event of a short circuit in the collector circuit a = 0.9 to 1, measurement error ±5%; reverse conductivity idle in emitter circuit h_{22} = (0.4 to 4)10^{-6} ohm, measurement error ±10%; initial collector current when no emitter current is flowing I_{co} = 0 to 50 microamps, measurement error ±2.5%
UPU-1M	all-purpose test installation	electrical strength can be tested at the operating voltage (AC and DC) up to 10 kilovolts, power transformer rating 1 kva.

The chief possibilities for making measurements and tests of equipment with these instruments, in the shop, as well as in the field, are listed in Table 14.12.

Table 14.12

Principal Parameters and Measurement Limits for Miscellaneous Instruments

Parameters and Measurement Limits	Measuring Instruments
Direct current:	
75 microamps to 1.5 amps	Ts-57
60 microamps to 12 amps	AVO-5M

Table 14.12 (continued)

Principal Parameters and Measurement Limits for Miscellaneous Instruments

Parameters and Measurement Limits	Measuring Instruments
Alternating current:	
0.5 milliamp to 1.5 amps	Ts-57
3 milliamps to 12 amps	AVO-5M
DC and AC voltages:	
75 millivolts to 600 volts	Ts-57
0.1 volt to 6 kilovolts	AVO-5M
30 millivolts to 20 kilovolts	VK7-9
Voltage, high frequency, up to 700 MHz-100 millivolts to 1000 volts	VK7-9
Pulse voltages, 0 to 50 kilovolts	V4-2
DC resistance:	
3 ohms to 3 megohms	AVO-5M, Ts-57
up to 1,000 megohms	M1101, VK7-9
Electrical strength of insulation up to 10 kilovolts	UPU-1M
Correct functioning of electrical circuits	M1101, AVO-5M, Ts-57
Frequency measurement 350 to 675 MHz	Ch2-6
Pulse measurements	S1-20
Correct functioning of pulse circuits	G5-15, S1-20
Correct functioning of audio frequency channels	G3-33, S1-20
Response and correct functioning of high frequency amplifiers	G4-42, G4-44, G4-31
Visual tuning of broad band IF amplifiers	Kh1-1
Measurement of modulation depth	G4-42, S1-20
Check of receiver amplifier and low powered oscillator	L1-3
Check of crystal triodes	L2-1

14.10. Maintenance of Measuring Instruments and Checking Measurement Accuracy.

State supervision of the condition of, and checking on, measures and measuring instruments, is exercised by the State Committee on Standards, Measures, and Measuring Instruments. The corresponding supervision in the armed forces of the USSR is exercised by the Inspection for the Supervision

of Measures and Measuring Instruments of the Ministry of Defense of the USSR.

Military units can use only those measuring instruments that are in good working order, read correctly, and are authorized for use in the armed forces of the USSR.

All measures and measuring instruments are broken down into standard and working types.

Standard instruments are instruments that are specially selected, tested, and supplied with correction charts. These are of the highest precision, and used only to check working instruments.

Working instruments are those used to make direct measurements.

All measuring instruments are subject to a mandatory (working instrument) or state (standard instrument) check at predetermined times set for each type of instrument (usually once every year or two years). The check is made by organizations designated by the Ministry of Defense of the USSR, or by the State Committee on Standards, Measures, and Measuring Instruments.

Military unit personnel and equipment can be used to deliver the instruments for check and to ship them back. A seal is affixed to checked instruments, and units must not tamper with it.

Faulty and inaccurate instruments, those not checked at the times specified, as well as instruments about which there are doubts as to correctness of readings, are to be withdrawn from use and repaired, adjusted, and checked for correct readings.

Measuring instruments require particularly careful handling. Instruments must be protected against jolts and shocks, and dust, moisture, and dirt must be kept out.

Instruments should be used in the temperature range from +10° to +30°C, and in humidity of from 60 to 75% (exceptions are special purpose instruments).

It is permissible to expose an instrument to temperatures from -40° to +40°C, and to humidity as high as 95%, for brief periods of time, but in such cases the instrument should have been subject to normal conditions for at least 24 hours prior to use.

Military units may decide to make minor repairs, those that do not require breaking the seal.

Measuring instruments should be stored in dry, heated spaces in cabinets, or storage boxes. Acids, lubricants, and storage batteries will not be stored in these spaces. Such storage is in fact prohibited.

Measuring instruments are transported in their storage boxes which, in turn, are stored in packing boxes.

Chapter XV

Cybernetics and Digital Computers

15.1. The Subject of Cybernetics

Cybernetics is the science of control processes. It studies the general laws involved in the translation of information into complex controlling systems (machines, living organisms, society).

Questions concerned with control are the subject of investigation by a number of sciences, including, for example, the theory of electronic computers, communication theory, the theory of automatic regulation, and others. As distinguished from the sciences indicated, cybernetics studies control processes in general, rather than in concrete systems, and as it applies to any control process, regardless of the physical nature of that process. Whereas the above mentioned sciences study what is primarily the internal structure of a particular controlling system, cybernetics is primarily interested in the general laws governing the functioning of complex control systems.

The approach to the study of the processes involved in the translation of information into controlling systems that is characteristic of cybernetics is sometimes called the "black box" method in the literature, and the essence of this method is as follows.

The control system is represented in the form of some device (the "black box") with inputs through which the external medium can act on the system, and outputs for the reverse action of the system on the external medium (fig. 15.1).

From the point of view of cybernetics, the system studied can be fully described by laws establishing the dependence of the output information on the input information.

If two control systems have identical laws for the translation of information they are considered as equivalent, so far as cybernetics is concerned, even though they can, generally speaking, be significantly different in their physical nature.

Thus, cybernetics studies control processes from the informational point of view.

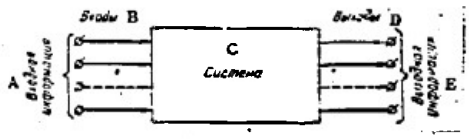


Figure 15.1. Conventional depiction of a control system. A - input information; B - inputs; C - system; D - output information.

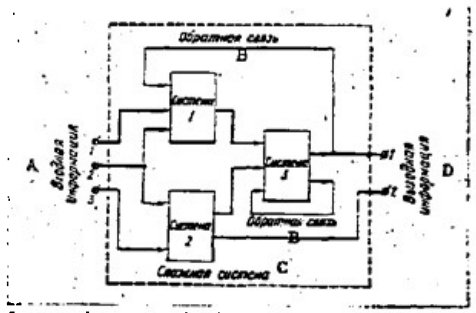


Figure 15.2. Conventional depiction of a complex control system. A - input information; B - feedback; C - complex system; D - output information; systems numbered 1, 2 , 3.

Systems used to process information can be interconnected to form a new, complex, control system.

Figure 15.2 shows a complex control system consisting of three systems, 1, 2, and 3. The individual systems within the framework of a complex control system are called its elements.

Information translation systems can be connected up in three different ways: in series, in parallel, and with feedback. In Figure 15.2 the systems (elements) numbered 2 and 3 are connected in series, 1 and 2 are connected in parallel, and 1 and 3 are connected with feedback.

Feedback is of particular significance in cybernetics, because most control processes are processes with feedback in them. In the broad sense of the word, feedback can be considered as taking place when two systems interact. Let us take the example of the operating principles of the automatic pilot (control system) installed in an aircraft (controlled system). When the aircraft deviates from its assigned course a signal, characterizing the magnitude of such deviation (the error signal), is fed into the automatic pilot input. The automatic pilot produces output signals that are fed into the aircraft rudder control system, causing the aircraft to change course in a direction such that the error signal is reduced.

Feedback plays a big part in living organisms as well because many physio-

logical processes can be considered to be processes taking place in a feedback system. The processes involved in the formation of conditioned reflexes and training, for example, are based on obtained feedback signals confirming the correctness of the action taken.

Because cybernetics studies control processes in systems in the abstract, as opposed to their physical essence, one must introduce some procedure for describing the laws that are involved in the information translation process that does not depend on the structures of the systems under investigation. The theory of algorithms deals with the study of formal methods of information transformation.

The concept of the algorithm is the central concept in this theory. Without pretending to strictness, algorithms can define the totality of formal rules in accordance with which information translation is carried out.

From the point of view of cybernetics, the information translation algorithm realized by some control system (or its elements) completely characterizes the system (its elements).

Let us made a summation. What follows from the definition of cybernetics as a science is that the object of its investigation is complex control systems. At the same time, cybernetics digresses from the physical essence of the system under study and presents it in one of the following two forms:

in the form of an abstract device for translating information ("black box"), the functioning of which can be described by some algorithm;

in the form of some combination of simple systems for translating information (of elements), the functioning of which can be described by definite algorithms.

This formalization of the object of investigation on the part of cybernetics can prove to be extremely fruitful. It does not enable us to study each concrete control system individually, but to investigate the common properties of a great many systems. This is precisely why cybernetics is defined as the science of general laws for translating information in control systems.

Data on the control system can be incomplete in the general case. For example, the structure of the system can be known, that is, the algorithms for the functioning of its elements, and ways in which they combine with each other, but the algorithm for the functioning of the system as a whole is not known. In other cases the algorithm for the functioning of a complex system can be known, but the algorithms for the functioning of its elements, or the

ways in which they combine with each other, are not known.

And as a result, one of the central tasks of cybernetics is to investigate control systems in order to determine the unknown characteristics of those systems.

15.2. Basic Tasks of Cybernetics.

The cybernetic approach to the study of control systems can be broken down into three main tasks:

system analysis;
system synthesis;
system modeling.

Analysis of Control Systems

The analysis task is one that arises in the study of already existing systems for information translation. The purpose of the analysis is to uncover unknown characteristics of control systems. If a system with a known structure is given (a set of elements, and the ways in which they combine with each other, is given), the analysis task involves a determination of the information translation algorithm for that system.

Analysis tasks can also include the separation of the component parts of a system within the limits of that system, the exposure of certain unknown links between them, and others. Different analysis tasks are widely known in many fields of science and engineering.

The special feature of analysis in cybernetics is that the analysis process deviates from the physical nature of the elements in the system, and investigates the general laws covering information translation in the system as a whole, and in its component parts.

Synthesis of Control Systems

In some respects the task of synthesis of control systems is the reverse of the analysis task in its relationships.

The synthesis task can be formulated as follows. A certain set of elements (of simple systems for information translation) is known, and must be used to construct a control system that will provide the specified information translation algorithm.

The synthesis task usually has several solutions, giving rise to the problem of selecting the optimal solution.

Criteria ordinarily used to establish the quality of control systems are system simplicity, reaction time, reliability, and accuracy. The important task of equivalent conversions of control systems arises in connection with the problem of optimization.

Control systems are said to be equivalent when they have the same laws for information translation. Questions concerned with synthesis are of great practical value for the designing of various types of control systems.

Modeling of Control Systems

When the control system is extremely complex, or little access can be had to it so analysis and synthesis methods can be used for investigative purposes, the replacement of the system by some model can often be desirable. A model is a comparatively simple system that functions in a manner similar to the functioning of the original, and the process of designing the model is called modeling.

Cybernetic modeling is most often done by programming a universal digital computer. Various random factors (noise, interference, equipment failures, and others) can effect the operation of control systems. In such cases, the statistical tests method, also known as the Monte Carlo method, can be used for modeling. Random number transmitters are used to produce the random magnitudes in models, or they are formed by using a special program for the electronic computer on which the modeling is done.

The modeling of various biological processes, including certain functions of the nervous system and the brain, are of great interest to cybernetics.

Modeling, as a method for investigating control systems, is widely used in various branches of economics, science, engineering, and in military affairs.

15.3. The Mathematical Apparatus of Cybernetics.

The mathematical apparatus of cybernetics, which includes a number of branches of mathematics, has been created to investigate the general laws in accordance with which information is translated.

Information Theory

Any control system that is the object of investigation by cybernetics has inputs and outputs by means of which it is related to the environment.

This relationship is brought about by signals. A signal is any action that can be fed into the system inputs or observed at the system outputs. For

the input signals to a television set are the radio waves, and the outputs are the picture and the sound. The input and output signals associated with different systems can be mechanical forces, electrical voltages, or currents, and the like. However, in the cybernetic approach to the investigation of control systems the physical nature of the input and output signals is rejected, and the one characteristic of the signals, called the amount of information, is used.

The amount of information is a measure of signal variety.

The amount of information is primarily dependent on the number of different signals the sources of information can form.

Cybernetics has adopted the coding of each signal with a predetermined symbol, such as a letter, number, or some sequence of letters, or numbers. If each symbol can appear with the same probability, the number of different symbols, N, uniquely determines the amount of information from the signal source. And in this case, the amount of information, I, is equal to the logarithm of the number of possible symbols

$$I = \log_a N.$$

The base of the logarithm is what establishes the unit by which the amount of information is measured. Information is most often measured in binary units, so in the case cited, the base of the logarithm, a, is equal to two. By way of an example, the high and low levels of voltages are coded with the symbols 1 and 0, respectively, in electronic digital computers. An information source producing one of two possible symbols with equal probability of doing so can be characterized by an amount of information equal to one binary unit.

If a source forms eight equally probable symbols, the amount of information equals three binary units

$$\log_2 8 = 3.$$

If the probability of the appearance of different symbols is not the same, and as an example let us take the probability (frequency) of the appearance of different letters in words, the amount of information can be established with the distribution of the magnitude of the probability with respect to the symbol taken into consideration.

Thus, using the apparatus comprising the theory of information, cybernetics abstracts from the physical sense of signals and considers only the information those signals contain.

The information is presented by a set of symbols called an alphabet. Most widely used in cybernetics is a two-letter alphabet containing the symbols 0 and 1.

Any sequence of symbols is called a word in the particular alphabet.

In cybernetics, questions concerned with the transmission and translation of information reduce to questions concerned with the transmission and translation of words in some alphabet.

Theory of Algorithms

We have defined the concept of the algorithm as a system of rules in accordance with which information is translated. Known in mathematics, for example, are algorithms for doing arithmetical tasks with numbers, square rooting algorithms, differentiation of functions, and others.

The technical sciences, a good part of which is devoted to the study and development of algorithms for computing various structures, and for designing equipment, provide us with a multiplicity of examples of algorithms.

Algorithms equated to the effective control of troops are of great importance in military affairs in arriving at successful solutions to combat missions. The anti-aircraft defense system (PVO) uses algorithms for guiding missiles to enemy air targets, algorithms for processing information dealing with the air situation, and target distribution algorithms.

Universal algorithms are of particular interest to cybernetics. Universal algorithms are those that can be converted into any other algorithms. Units realizing a universal algorithm can, in principle, perform any information processing established by known algorithms.

Modern universal electronic digital computers have the remarkable property of algorithmic universality. It is because of this property that universal computers are used in all fields of science and engineering. Electronic digital computers can solve computational, as well as logical, problems, exercising control and planning in economics, helping commanders to arrive at a decision, translating texts from one language to another, and the like.

The human brain too is a universal information conversion device. In this connection, cybernetics has raised an extremely important, and interesting, problem, that of the relationship between the brain and a machine.

Mathematical Logic

The basic task of mathematical logic is one of analyzing mathematical considerations and mathematical concepts.

The founder of mathematical logic, George Boole, developed one of the basic divisions of the subject, that of the algebra of logic, which, today, is also called Boolean algebra.

The algebra of logic analyzes different formulations (propositions). All propositions in the algebra of logic are broken down into two classes; true, and false. True propositions are assigned the symbol 1, false ones the symbol 0.

The logical operations performed on propositions, conjunction, disjunction, inversion, and others, are included in the algebra of logic.

The conjunction of two propositions is the conventional speech equivalent of connecting these propositions by the conjunction "AND," disjunction the corresponding connection by the link "OR," while inversion signifies the negation of the particular proposition.

Various propositions can be represented in the form of formal mathematical expressions that can be subjected to a variety of conversions, given the assistance of the algebra of logic.

Mathematical logic is used in cybernetics in modeling certain thought processes.

The division of mathematical logic known as Boolean algebra is widely used to solve problems concerned with the synthesis and analysis of control systems.

Optimal Solutions Through Mathematical Methods

Tasks of the best possible construction of control systems are often carried out in cybernetics. In so doing, widsspread use is made of optimal solution methods developed in such divisions of mathematics as game theory, linear and dynamic programming, and the queueing theory.

The game theory studies situations in which opponents with opposing goals collide. It is the task of the game theory to establish those courses of action providing the best advantage for each of the protagonists. The game theory must take into consideration in its solution counteractions by the protagonists, elements of randomness, and the degree of completeness of information on the behavior of the opposing sides.

So-called two-person games with a zero sum are of the greatest interest in military affairs. In these games we have two sides, and the gain made by one is precisely equal to the loss inflicted on the other.

Game theory is what makes it possible to determine the optimal strategy, that is, that method of conducting the game that will yield a maximum average gain to one of the sides after a great many repetitions of the game.

Linear programming theory leads to a solution to the problem of optimal control of various processes that can be specified by a system of linear inequalities and equations.

Typical of linear programming are transportation problems in which the most advantageous variants in the delivery of cargo from several points to different consumers can be established.

Linear programming methods can also solve problems concerned with the optimal distribution of forces against enemy targets.

As often happens when arriving at the solution to optimal control, there is no immediate decision, but rather a gradual, step by step, one. The study of step-by-step solutions, optimal in some particular sense, is the subject of dynamic programming.

Dynamic programming methods are used to solve problems concerned with establishing the optimal sequence in which combat forces are to be used during an operation, the optimal use of reserves during the various stages of the battle, and others.

The queuing theory enables us to find the most rational method of building a system designed to satisfy requirements arising on a random basis.

The queuing theory helps establish the most effective method of reflecting the flight of an air enemy, the optimal forces, communication channels, the optimal reserve method, and others.

Control System Theory

Just as in the case of cybernetics, the objectives of control system theory research are the control systems. But as distinguished from cybernetics, which studies abstract systems represented by mathematical models, classic control system theory investigates concrete systems used to provide automatic regulation.

The theory of discrete control systems is highly important in cybernetics. In this connection, one must set apart a special mathematical discipline, that of the theory of finite automatons, studying a special class of discrete systems for translating information.

A finite automaton is a system that can shift from one condition to another under the influence of the input information, and, using these shifts, remember and reprocess the information feed into its input.

Finite automatons are systems the synthesis and analysis methods for which have been the most completely worked out, and thus have great value for cybernetics.

15.4. Cybernetics in Military Affairs.

Today the use of new methods for controlling troops and military equipment through the achievements in the fields of electronics and cybernetics is becoming more and more widespread.

Military cybernetics, because it is one of the applied trends in cybernetics, is concerned with the study of automating troop control processes.

The goal of military cybernetics is to achieve the maximum effectiveness in actions taken by troops and in the use of weapons, with scientific methods of conducting operations and the use of the most modern technical media for controlling, assembling, and processing information, as the basis.

Methods used to arrive at the optimum decisions concerned with control of troops are the content of the division of military cybernetics called the operations research theory. *

* Methods encompassed in operations research theory are not only used in military affairs, but in various other areas of science and engineering as well. Generally speaking, operations research is referred to as a science concerned with rational methods of organizing purposeful human endeavors.

Operations, from the standpoint of military cybernetics, are any measures united by a single thought, and aimed at achieving a definite goal.

The first step in using methods encompassed by operations research theory is to construct a simplified arrangement, or model, of the operation. A model is a mathematical description of the operation that takes into consideration the basic factors that will affect the outcome of the operation, and discards those factors that can be said to be secondary in importance.

The effect of random factors of different kinds are usually taken into consideration in the course of building the model of the operation. Consequently, the basic mathematical apparatus used in operations research is the probability theory, a science that studies random phenomena and the behavior patterns inherent in them.

Different criteria are used to evaluate the effectiveness of an operation. The basic criterion of effectiveness is the probability of carrying out the combat mission.

Investigation of a model of an operation by using the criterion of effectiveness involves a determination of the most rational methods for using manpower and equipment.

One of the important trends in military cybernetics is the automation of the control of troops by using electronic digital computers. The computers are used to plan and prepare for the operation, to assemble, process, and reflect information for use in arriving at decisions for the different operational and tactical missions in question.

Very high-speed electronic computers make it possible to read out a great many variants for conducting combat operations in a very short period of time, to compare them with each other, and to select the best decision.

15.5. Digital Computers.

One of the most significant technical achievements of our time has been the development of universal digital computers.

The appearance of high-speed electronic digital computers has made it possible to automate the mental activities of man in the most diverse fields of science and engineering.

The wide-ranging capabilities of modern computers are the result of three outstanding characteristics:

universality;
very high operating speed;
high degree of accuracy.

The universality of computers is what makes it possible, at least in principle, to automate any field of human endeavor based on information translation.

Universal computers can solve problems that are purely computational in nature, problems concerned with the planning and control of production, they can translate one language into another, and they can solve a variety of military problems.

The universality of computers is based on the use of the programmed control principle, and on a set of operations they can carry out.

The speed at which digital computers function can be evaluated by the number of operations completed in unit time.

Modern electronic digital computers complete tens and hundreds of thousands of operations per second.

All magnitudes with which operations are performed in electronic digital computers are represented by numbers. The computational accuracy depends on the ordered state of the numbers. A highly accurate computation can be obtained by selecting the number of orders large enough.

15.6. Block Schematic of an Electronic Digital Computer.

Modern universal digital computers have five basic devices:

the memory;

the arithmetic device;

the control device;

the input device;

the output device.

Figure 15.3 is a simplified block schematic of an electronic digital computer.

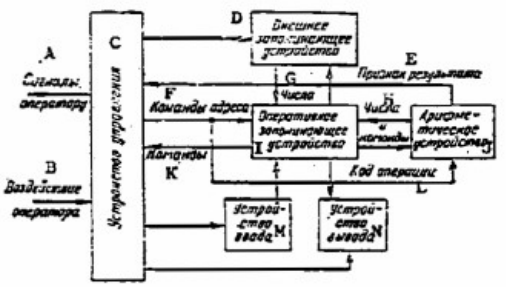


Figure 15.3. Simplified block schematic of an electronic digital computer. A - signals to operator; B - action by operator; C - control device; D - external memory; E - sign of result; F - address commands; G - numbers; H - numbers and commands; I - operative memory; J - arithmetic memory; K - commands; L - operation code; M - input device; N - output device.

The memory is designed to store the original information, the intermediate and final results of solutions to problems, and the program for making the computations. The memory consists of separate cells which assume numbers called addresses.

The basic characteristics of a memory are capacity, and access time.

The capacity of a memory is the number of numbers that can be stored in

in the memory at any one time.

The access time is the time required to select numbers from the memory, or to record them in the memory.

The memories of modern universal electronic digital computers have high capacities, and short access times. This is what results in the electronic digital computers being so high speed, thus making it possible to use them to solve complicated problems requiring the storage of great quantities of data.

However, there are major difficulties involved in engineering memories with high capacities and short access times, and it is for this reason that electronic digital computers usually have two types of memories:

an external memory, sometimes called the store; and,

an operative memory.

The external memory has a high capacity (up to several million numbers), but the access time is long, as long as several minutes for certain types of external memories. External memories are usually in the form of magnetic tapes and magnetic drums.

The operative memory has a comparatively low capacity (several thousand numbers), and a short access time (in units of microseconds). The operative memory in modern electronic digital computers is made up of ferrite cubes.

The arithmetic device is designed to complete the arithmetical and logical operations with the numbers. The information to be processed is fed from the operative memory into the arithmetic device. The results of the arithmetical and logical operations are back-recorded in the operative memory. The operational sequence in the arithmetic device is established by signals supplied by the control device. In turn, certain signs for the results obtained (the sign for a negative number, the sign that the order network is overflowing, and others) are fed from the arithmetic device into the control device.

The control device ensures operation of the electronic digital computer in accordance with the program specified.

The basic functions of the control device are to:

control the processes involved in access to the operative memory;

control the exchange of information between the operative memory and the external memory;

generate the control signals for carrying out the operations in

the arithmetic device;

control the processes involved in information input and output.

The input device is designed to transmit the original data and the program for the computations to the computer memory. The information going into the computer is put in by the operator on perforated tape, or on punched cards, in the form of a system of holes. The input devices translate this information into electrical signals which are fed into the computer memory.

The output device is designed to generate the results of the solution to the problem in the form of numerical, alphabetical, and graphical material printed on paper.

The programmed control principle. A problem committed to an electronic digital computer is made up in the form of a program.

A program consists of a sequence of commands (orders) that establish the course of the computational process in the computer.

A command is a special code that establishes the type of operation, and the numbers, to be used to do the given operation. The command has two parts, the operational part, and the address part.

The operational part of the command contains the operation code, which, when acted upon, delivers the command given. The operation code is the ordinal number of the command in the electronic digital computer's command system.

The address part of the command is used to allocate the addresses of the cells in the operative memory in which the information subject to translation is stored.

Commands are broken down into single-address, and multiaddress, depending on how many addresses they contain. Two- and three-address commands are the most widely used of the multiaddress commands.

A schematic representation of a three-address command is shown as Figure 15.4.

Operation	A_1	A_2	A_3
Operation code	1st address	2nd address	3rd address

Figure 15.4. Schematic representation of a three-address command.

The operation code (multiply, add, shift, and so forth, for example) is written down in the operational part of the command.

The numbers (addresses) of the cells in the operative memory in which the numbers to be used in the operation are stored are written down in the first and second parts of the command (A_1 and A_2). Written down in the third part (A_3) is the address of the cell in the operative memory in which the results of the operation are located.

In the single-address computer the command contains the operation code and one address. So, in order to carry out an operation with two numbers, what is needed in the general case are three commands, two to select the numbers for use in the operation, and one to register the result.

The electronic digital computer utilized two methods of carrying out commands, natural and forced.

In the natural method, all the commands in the program are sent to the cells in the operative memory one after the other in an ascending sequence of numbers. When the command placed in the cell with address o has been carried out, the computer automatically shifts to the cell with the address n + 1, etc., until such time as a special command changing the natural sequence of carrying out the command (commands for conditional and unconditional transfer, for example) is received.

In the forced method of carrying out a command, each command, in addition to its basic content, indicates the number of the cell in the operative memory from which the next command must be selected. The forced method of carrying out commands is most often used in two- and four-address computers.

15.7. Principal features of the electronic digital computer.

The principal features of modern electronic digital computers include:

- a numerical system;
- a method for representing the numbers;
- a form for representing the numbers;
- a system of operations;
- high speed;
- orderedness;
- memory capacity.

Numerical System

A numerical system is understood to mean a method for coding numbers by a cipher, or digit.

The number of different digits used in a numerical system is called the basis of the numerical system. For example, ten different digits, 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9, are used in the generally accepted decimal number system. A ternary number system with digits 0, 1, 2, or -1, 0, +1, is used in the "Setun'" electronic digital computer.

Each number in a numerical system can be coded by a predetermined sequence digit.

Most existing electronic digital computers use a binary system, in which two digits, 0 and 1, are utilized. This is done primarily because of the ease with which the arithmetical and logical operations can be carried out within the framework of the binary system, and because of the widespread use of two-position components (triggers, ferrites, and others)

In the binary system a number is represented by a predetermined sequential digit, 0 and 1. Each position assumes a serial number, called a bit.

A binary number has n whole, and m fractional, bits, written in the following form

$$a_{n-1} \dots a_i \dots a_1 a_0, a_{-1} \dots a_{-m},$$

where a_i equals 0 and 1.

All that need be done to translate a binary number into a decimal notation is to write it in the form

$$a_{n-1} \dots a_1 \dots a_1 a_0, a_{-1} \dots a_{-m} = a_{n-1} 2^{n-1} + \\ + \dots + a_i 2^i + \dots + a_1 2^1 + a_0 2^0 + a_{-1} 2^{-1} + \dots + a_{-m} 2^{-m}$$

<u>Example.</u> Represent the binary number 11011, 101, in a decimal code.

Solution.

$$11011{,}101_{(2)} = 1 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + \\ + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} = 27{,}625_{(10)}.$$

A sequence of integral decimals from 0 to 10 in the binary system can be written in the form

0; 1; 10; 11; 100; 101; 111; 1000; 1001; 1010 .

The arithmetical action on the numbers in a binary system is carried out in accordance with rules similar to those developed for the decimal system. Addition and multiplication tables are used in the binary system as follows:

Addition Table	Multiplication Table
$0+0=0$	$0 \times 0 = 0$
$0+1=1$	$0 \times 1 = 0$
$1+0=1$	$1 \times 0 = 0$
$1+1=0$	$1 \times 1 = 1$

Example. Add the binary numbers 11101, 01, and 1001, 101.

```
      11101.01
+      1001.101
  ----------
    100110.111
```

Example. Multiply the binary numbers 101.1 and 10.01.

```
        101.1
x        10.01
        -----
          1011
         0000
        0000
       1011
      -------
      1100.011
```

The binary-decimal notation is the one most often used with electronic digital computers. This system is, in essence, the conventional decimal system in which all the digits are represented by four-digit binary codes, called tetrads,

0=0000	5=0101
1=0001	6=0110
2=0010	7=0111
3=0011	8=1000
4=0100	9=1001

Example. Write the number 85.3 in the binary-decimal notation.

$$85{,}3_{(10)} = \underset{8}{\underline{1000}}\underset{5}{\underline{0101}},\underset{3}{\underline{0011}}_{(2-10)}.$$

Methods Used to Represent Numbers in an Electronic Digital Computer

Three methods are used with electronic digital computers to represent positive and negative numbers:

direct code;

one's code;

two's code.

A sign digit, placed in front of the high-order numeric digit, is introduced when numbers are represented in the direct code. A plus sign usually represents zero, and a minus sign usually represents the units in the sign digit.

For example, numbers $A = +13.5_{(10)} = +1101{,}1_{(2)}$, and $B = -13.5_{(10)} = -1101{,}1_{(2)}$ will be written in the form $[A]_d = 0.1101{,}1_{(2)}$, $[B]_d = 1.1101{,}1_{(2)}$.

A summer and subtracter must be available in the electronic digital computer's arithmetic device when addition and subtraction of numbers represented in a direct code is done. Representation of numbers in the one's and two's codes results in replacement of the subtraction by addition.

The one's code for positive numbers coincides with their direct code. When the translation of negative binary numbers is from the direct code into the one's code, the one must be left in the sign position, and the zeros in the digit positions in the direct code must be replaced by ones , and the ones by zeros.

For example

$$[A]_d = 0.1011,01_{(2)}, \quad [A]_{one's} = 0.1011,01_{(2)},$$
$$[B]_d = 1.1011,101_{(2)}, \quad [B]_{one's} = 1.0100,010_{(2)}$$

The addition of numbers represented in the one's code is carried out in an electronic digital computer in a summer with end-around carry. In these summers, when ones of the carry from the sign position crop up they are translated into a low-order digit and summed with the results already obtained.

Example. Add two numbers, $1.1010,\ 1100_{(2)}$ and $0.1011,\ 0101_{(2)}$, represented in the one's code

$$\begin{array}{rl} [A]_{one's} & = 1.1010,1100_{(2)} \\ [B]_{one's} & = \underline{0.1011,0101_{(2)}} \\ & \quad 10.0110,0001 \\ \text{end-around carry} & \underline{\qquad\qquad 1} \\ [A]_{one's}+[B]_{one's} & = 0.0110,0010_{(2)} \end{array}$$

The two's code for positive numbers coincides with their direct code. In order to represent a negative binary number in the two's code it is necessary to make up the one's code, add a low-order unit to it and write the unit in the sign position.

Example. Represent the number $[A]_d = 1.1010,\ 1100_{(2)}$ in the two's code

$$\begin{array}{rl} [A]_d & = 1.1010,1100_{(2)} \\ [A]_{one's} & = 1.0101,0011_{(2)} \\ & + \underline{\qquad\qquad 1} \\ [A]_{two's} & = 1.0101,0100_{(2)} \end{array}$$

The addition of numbers represented in two's codes is carried out in summers without end-around carry. When ones of the carry from the sign position crop up they are dropped.

Example. Add the numbers $0.10101,01_{(2)}$ and $1.10101,10_{(2)}$ represented in the two's code

$$[A]_{two's} = 0.10101,01_{(2)}$$
$$[B]_{two's} = 1.10101,10_{(2)}$$
$$[A]_{two's}+[B]_{two's} = 10.01010,11_{(2)}$$
$$1 \text{ dropped}$$
$$[A]_{two's} + [B]_{two's} = 0.01010,11_{(2)}$$

The number of positions available for representing numbers in an electronic digital computer is limited, so it is possible to obtain a result that exceeds the maximum number that can be written in the computer's position network when two numbers with identical signs are added. The computer's position network overflows.

Modified codes are introduced to establish the fact that the position network has overflowed. Code modification involves the introduction of two sign positions. Any of the three codes, direct, one's, two's, can be modified. Positive numbers in the modified codes have two zeros in the sign positions, while negative numbers have two ones. Different numbers in the sign positions indicate overflow of the position network.

Example. Add the number $11.0101,01_{(2)}$ and $11.1011,11_{(2)}$, represented in a modified two's code

$$[A]^{mod}_{two's} = 11.0101,01_{(2)}$$
$$[B]^{mod}_{two's} = 11.1011,11_{(2)}$$
$$[A]^{mod}_{two's} + [B]^{mod}_{two's} = 11.0001,00_{(2)}$$
$$\text{not overflowing}$$

Example. Add the numbers $00.0010_{(2)}$ and $00.0110_{(2)}$, represented in a modified two's code

$$[A]^{mod}_{two's} = 00.1101_{(2)}$$
$$[B]^{mod}_{two's} = 00.0110_{(2)}$$
$$[A]^{mod}_{two's} + [B]^{mod}_{two's} = 01.0011_{(2)}$$
$$\text{overflowing present}$$

Forms for representing the numbers

Numbers are represented in two ways in electronic digital computers. One is the natural form of representing numbers, used in computers with a fixed point, the other is the normal form, used in a computer with a floating point.

In the case of the natural form of representing numbers, the position of the point in the position network of the computer is fixed. In order to simplify the selection of the scale factors during programming, as well as in order to eliminate certain possible cases of overflowing (when multiplying, for example), the point is fixed between the sign digit and the more significant digit. Only proper fractions can be written in the position network in this case.

Figure 15.5 shows a schematic representation of how the number $-0.101101_{(2)}$ is written in the position network of a computer with a fixed point. Scale factors such that any numbers cropping up during the computation process will be less than one are introduced during the programming of an electronic digital computer with a fixed point.

1	1	0	1	1	0	1

Figure 15.5. Schematic diagram of how the number $-0.101101_{(2)}$ is written in the position network of a computer with a fixed point.

The so-called normal form is often used to write quite large, or very small, numbers

$$-390000_{(10)} = -39 \cdot 10^4 ,$$

$$0.00000125_{(10)} = 0.125 \cdot 10^{-5} .$$

When this manner of writing is used the value of the number depends on the mantissa and on the number order.

In the above-cited examples, the mantissas are -39 and 0.125, and the orders are 4 and -5.

Numbers in normal form are not uniquely expressed. For example, the number $31_{(10)}$ can be written

$$31_{(10)} = 31 \cdot 10^0 = 3.1 \cdot 10^1 = 0.31 \cdot 10^2 = 0.031 \cdot 10^3 .$$

The position of the decimal point in the mantissa changes according to the magnitude of the order. Hence a computer using the normal form for representing numbers is called a computer with a floating point.

When binary numbers are represented in the position network of a computer with a floating point, the magnitude of the order is usually selected such that the mantissa of the number is a proper fraction, in which one can be written as the high-order digit. A number in this form is called normalized.

1	1	1	0	1	0	1	1	0	1	0	0	0	1	0	0

Figure 15.6. Schematic diagram of how the numbers -1101, $01101_{(2)}$ in normalized form are written in the position network of a computer with a floating point.

Figure 15.6 is a schematic representation of how the binary number -1101, $01101_{(2)}$ in the normalized form $-0.110101101 \cdot 2^4$ is written in the position network of a computer with a floating point.

Computers with a floating point have a wide range of represented numbers, so scale factors are not introduced in the usual course of programming tasks for these computers.

When arithmetical operations are performed with the numbers in electronic computers with a floating point it is necessary to translate the mantissas, as well as the orders. This complicates the structure of the arithmetical device and reduces machine speed. These disadvantages are not present in computers with a fixed point.

System of operations

Digital computers can perform a limited number of very simple operations. The totality of the very simple operations performed by the computer is called the system of operations. The system of operations performed by a universal computer should satisfy two basic requirements:

realization of any algorithms for translating information (the property of universality);

simplicity of programming the electronic digital computer.

Universality on the part of an electronic digital computer is provided by a comparatively small set of operations.

For example, it is possible to build a universal electronic digital computer with a set of operations as follows:

the operation of forwarding the contents of any cell in the memory to any other cell in the memory;

the operation of discrimination;

the operation of readdressing;

the operation of shutting down the computer.

Any universal computer can also have the capacity to perform the information input and output operations, in addition to those listed above.

But in general the program for a small number of operations is cumbersome and the programming process is significantly complicated.

This is why modern universal electronic digital computers have a branching system of operations.

The "Ural-2," an electronic digital computer, has 41 operations, for example.

All operations can be broken down into four main groups: arithmetical; logical; control operations; and forwarding operations.

Let us list the most characteristic operations in these groups.

Arithmetical operations:

- addition;
- subtraction;
- multiplication;
- division.

Logical operations:

- number shift;
- separation of parts of a number;
- comparison of two numbers.

Control operations:

- discrimination;
- unconditional transfer of control;
- readdressing;
- stops.

Forwarding operations:

- transfer from the arithmetic device to the operative memory;
- transfer from the operative memory to the arithmetic device;
- transfer from the operative memory to the store;
- transfer from the store to the operative memory.

The principal characteristics of Soviet electronic digital computers are listed in Table 15.1.

15.8. Electronic Digital Computer Elements.

Modern digital computers contain a great many radio components. For example, the number of transistors (tubes) is in the thousands, and the number of diodes, resistors, capacitors is in the hundreds of thousands. Nevertheless, the number of different types of elements used in the electronic digital computers is small.

Table 15.1. Principal Characteristics of Soviet Universal Electronic Digital Computers

Computer Name	Average speed, operations/ second	No. of digits	Form of number representation	No. of addresses	Memory capacities operative	Memory capacities store
BESM	10000	39	Normal, m = 32, p = 5	3	Ferrite cube 2048	Magnetic tape 12000
"Strela"	3000	43	Normal, m = 35, p = 5	3	Cathode ray tubes 2048	Magnetic tape 20000
"Ural-1"	100	36	Natural	1	Magnetic drum 1024	Magnetic tape 40000
"Ural-2"	5000	40	Natural and normal, m = 32, p = 6	1	Ferrite cube 2048	Eight magnetic drum 8192 each. Magneti tape 100000.
"Dnepr"	10000	26	Natural	2	Four ferrite cubes 512 each	-
"Minsk"	3000	31	Natural	2	Ferrite cube 1024	-
"Minsk-2"	5000	37	Natural	2	Two ferrite cubes 4096 each	-
M-2	3000	34	Natural and normal, m = 27, p = 5	3	Cathode ray tubes 512	Magnetic drum 512. Magnetic tape 50000.

NOTE: m is the number of digits in the mantissa (without the digit for the sign of the number); p is the number of digits in the order (without the digit for the sign of the order).

The elements contained in an electronic digital computer can be broken down into two groups; logical elements, and memory elements.

So far as the logical elements are concerned, the input signals at some moment in time are uniquely fixed by the values of the input signals acting at that very same moment in time.

So far as the memory elements are concerned, the output signals depend on the input signals, as well as on the condition of the elements.

Every element in an electronic digital computer forms two types of signals, one of which corresponds to the number 0, the other to 1.

The physical analogues of 0 and 1 can be low and high potentials, or the absence or presence of pulses (in tube and transistor circuits), oscillations with phase 0, and phase π (in parametron circuits), and others.

The set of elements from which the digital electronic computer circuits are assembled should have the property of functional completeness.

Functionally complete sets of elements are those from which it is possible to build any circuit for the conversion of digital information.

One of the most widely used sets of elements contains three logical elements; a coincidence circuit (AND element), a separation circuit (OR element), an inverter (NOT element), and one memory element, such as a trigger, for example.

<u>AND Element.</u> The operation of the AND element with two inputs (x and y) is described in Table 15.2. The output signal from the AND element is only equal to 1 when both input signals have a value equal to 1.

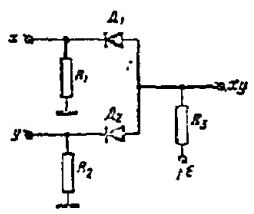


Figure 15.7. Diagrammatic layout of the AND element.

Table 15.2
Operation of the AND Element

x	0	0	1	1
y	0	1	0	1
xy	0	0	0	1

Figure 15.7 is the diagrammatic layout of the AND element.

If a low potential (code 0) is fed into one of the circuit inputs, the corresponding diode opens and a low potential (code 0) is picked off the circuit output. It is only when high potentials are present at both inputs that the diodes will be closed and that a high potential (code 1) will appear at the output.

OR Element. The operation of the OR element with two inputs is described in Table 15.3.

High potential fed to any element input (fig. 15.8) is transmitted to its output.

Table 15.3
Operation of the OR element

x	0	0	1	1
y	0	1	0	1
$x \vee y$	0	1	1	1

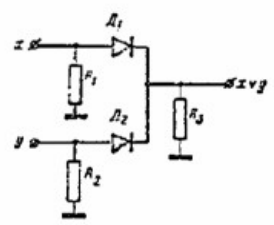


Figure 15.8. Diagrammatic layout of the OR element.

NOT Element is an inverter made in the form of an amplifier with an odd number of stages. When a high (low) potential is delivered to the inverter input there will be a low (high) potential at its output.

Triggers are elements with two stable states.

When the trigger has three inputs, one input is used to set the trigger in the 0 state, another input is used to set the trigger in the 1 state, and the third input is the inverting input. When pulses are fed into the inverting input the trigger changes state with each pulse.

A trigger with two inputs only has adjusting inputs, and a trigger with one input only has an inverting input.

15.9. Electronic Digital Computer Blocks

The following blocks are those most widely used in electronic digital

computers:

pulse counters;

registers;

decoders;

summers.

Pulse counters are used to count the number of pulses fed into the counter input. The results of the pulse count are usually recorded in the form of a binary code for the number of pulses fed into the input.

Figure 15.9 is the structural layout of a three-digit summing counter.

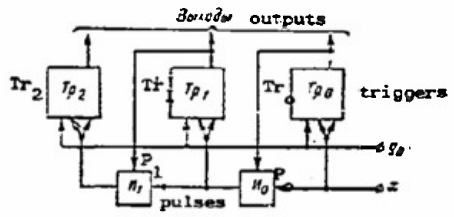


Figure 15.9. Structural layout of a three-digit summing counter.

Prior to operation all the counter's triggers are zeroed by a y_o pulse. Pulses, x, to be counted are fed into the inverting input of the low-order digit trigger, Tr_o. Trigger Tr_o changes state with the arrival of each pulse. Trigger Tr_1 will only change state when trigger Tr_o is in state 1, and a pulse is fed into the counter input because it is only in this case that a pulse to be fed into the inverting input of trigger Tr_1 can be formed at the output of the coincidence circuit, P_o. If triggers Tr_o and Tr_1 are in the 1 state at the time of arrival of x pulses, circuit P_1 causes trigger Tr_2 to function.

As the pulse counting process proceeds the trigger outputs provide an ascending sequence of numbers in a binary code.

This type of counter is called a summing counter.

Electronic digital computers also use subtract counters that count pulses in a descending sequence of numbers, and reversible counters that sum, as well as subtract.

Registers. A storage device designed for storing one number is called a register.

One trigger cell is used to store one digit of a number. The number of trigger cells in a register equals the number of digits in the stored binary number.

Registers can be parallel or series, depending on the method used to feed the number into the register.

A number is recorded in the parallel register (fig. 15.10) by using the P_1 circuit, into one input of which is fed the digits of the number, and into the other the number record control signal (y_{nr}). If some one of the digits of the number is 1, the output of the P_1 circuit, into the input of which the particular digit had been fed, will develop a signal that will set the corresponding trigger cell in the 1 state.

Delivery of the code stored in the number register is by the use of circuit P_2, controlled by the number delivery signal (y_{nd}).

The trigger cells are zeroed by the y_o signal prior to the reception of a new number.

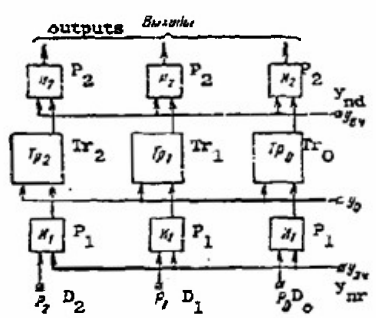


Figure 15.10. Structural layout of a parallel register.

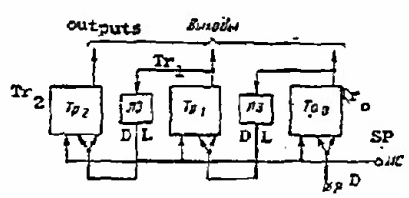


Figure 15.11. Structural layout of a series register.

The series register (fig. 15.11) has one input that receives a series code

in the number record mode. The register is fed a shift pulse, SP, after delivery of each digit in the number, and this sets all trigger cells in the 0 state. If one of the triggers is in the 1 state the delivery of the shift pulse changes that particular trigger to the 0 state and a signal forms at the trigger output that shifts the next trigger in line to the 1 state. This is what provides for the movement of the digits in the number through the register's cells. The delay lines between trigger cells in the register preclude the pulses coinciding in time at the trigger inputs.

A series, or a parallel, code can be used to deliver the number from a series register.

n pulse shifts are fed into the register in order to obtain delivery of a series coded n-digit number.

The register has additional circuits, P_2 (see fig. 15.10), controlled by the number deliver signal for purposes of delivering a parallel coded number.

Decoders. Decoders, or selective circuits, function as follows.

The delivery of some combination of binary signals to the decoder input will result in a signal appearing at but one output. The number of decoder outputs equals the number of different combinations of binary signals it is possible to deliver to the decoder's input.

If it is possible to feed any combination of binary signals (the number of such combinations equals 2^n) into n decoder inputs, the number of decoder outputs will equal 2^n. Decoders such as these are classed as complete decoders.

Diode decoders, and ferrite core decoders, are most widely used in electronic digital computers.

The decoder has an input register and a matrix that is a set of coincidence circuits.

Figure 15.12 shows the structural and diagrammatic layouts of a three input decoder.

There is only one coincidence circuit, at all inputs to which there are high potentials, for each combination of input signals. There will, therefore, be a high potential at the output of this circuit.

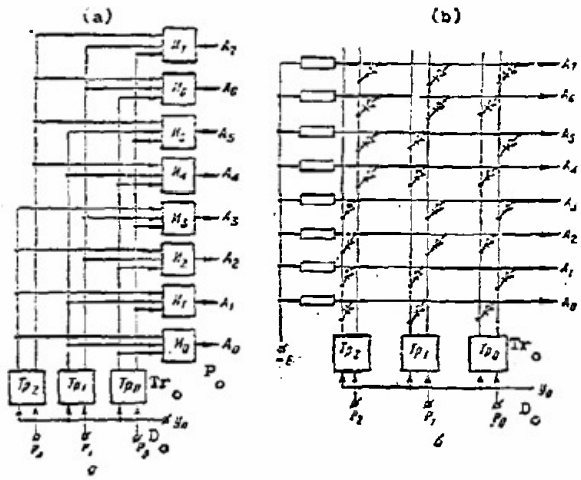


Figure 15.12. Three input decoder. a - structural layout. b - diagrammatical layout.

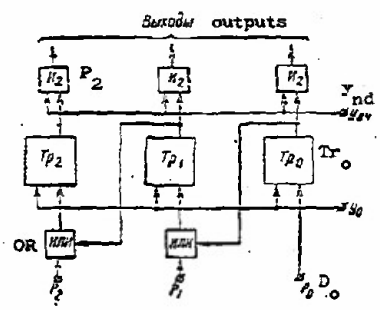


Figure 15.13. Structural layout of a summer with parallel input of summand digits.

Decoders are used in electronic digital computers to select the memory cells for recording and information readout by specific address, for processing control signals by codes used in the particular operations, and in the output device for printing control.

Summers. The summer is the basic component in the electronic digital computer's arithmetic device. Summers with a parallel input of summand digits are the type most widely used (fig. 15.13).

The summer circuit consists of triggers into the inputs of which are fed the summand digits, and carry circuits.

Each trigger of the cumulative summer has two basic functions; summing of numbers in a given order, and storing the results obtained.

Two numbers are added in the summer in four time steps.

During the first such step all summer triggers are zeroed by a y_o pulse.

During the second step the digits of the first summand are fed into the trigger inputs. All triggers receiving a signal corresponding to 1 at their inputs are shifted to the 1 state.

During the third step the digits of the second summand are fed into the trigger inputs. Triggers signalled corresponding to code 1 change state. A carry signal, fed into the trigger inputs from the high-order digits, appears at the outputs of the triggers that have shifted from state 1 to state 0.

During the fourth step the results are removed by signal y_{nd}.

The major shortcoming in the summer shown in Figure 15.13 is its comparatively slow speed, explained by the fact that the carry signal moves from trigger to trigger in turn.

This is why summers with simultaneous carry are quite frequently used in electronic digital computers. Special logical circuits are inserted in the carry circuit to provide for simultaneous carry.

If a sequence of several numbers is fed into the summer input the summer will accumulate the results of the addition of these numbers. This is why a summer of this particular type is called a cumulative summer.

There are combination type summers as well, based on logical circuits (without triggers). Both summands are fed into the input of the combination summer simultaneously, and the result of the addition will appear at the output at the moment the summands are delivered.

Summers with a series input of summand digits are sometimes used in slow speed electronic digital computers in order to save on equipment.

15.10. Principles on Which the Main Units in the Electronic Digital Computer are Based

Memories

There are a great many different types of memories, differing from each other with respect to operating principles, purpose, and parameters.

Ferrite, magnetic drum, and magnetic tape memories are the types most widely used in modern electronic digital computers.

The ability to use ferromagnetic materials for remembering information is based on the fact that these materials have two stable magnetic states, one of which can be used to store the 0 code, the other to store the 1 code.

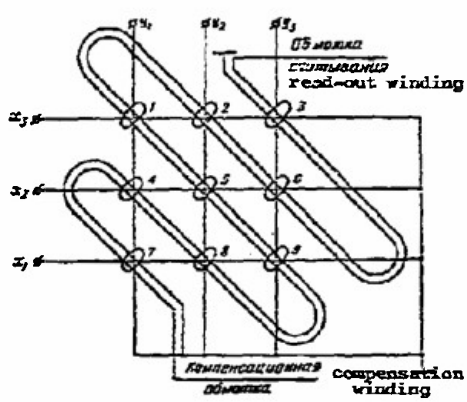


Figure 15.14. Schematic diagram of an individual matrix for a matrix memory.

Ferrite cube. The ferrite cube is the main type of operative memory used in modern electronic digital computers. This type of memory has a short access time, and this is constant during access to any cell in the memory.

The ferrite cube memory is a ferrite torus with control windings.

Toroidal cores are connected up into a cell, and the number of cores in a cell is equal to the number of digits in the numbers in the electronic digital computer.

The number of cells in a cube is what determines its capacity.

Matrix memories and type z memories differ in the manner in which they record and are read out.

Design-wise, the matrix memory is made in the form of individual matrices each of which combines the cores of all the cells designed for storing the same number order. The number of matrices is equal to the number of digits in the numbers in the electronic digital computer.

Four buses, two coordinate (x and y), a read-out, and a compensating, pass through each core of the matrix (fig. 15.14).

The matrix memory functions on the coincidence of currents principle.

Let us review the principle involved in recording and reading out information with the matrix memory.

Let us say, by way of an example, that we must record code 1 in the sixth core (fig. 15.14). The record signals are fed over the coordinate buses x_2 and y_3 to the particular core found at their intersection. The record signal fed over each of the coordinate buses has two pulses, one negative, the other positive, with amplitude $I_m/2$. The amplitude of the current pulses is selected such that only that core found at the intersection of the selected coordinate buses is remagnetized.

The sixth core is zeroed by the negative record pulse, while the positive pulses change it over to the 1 state.

A negative pulse with amplitude $I_m/2$ is fed to the compensating bus while code 0 is being recorded by the positive pulses along the coordinate buses. At the same time, one of the positive pulses is cancelled out and the core will remain in the 0 state.

Negative and positive current pulses are fed to the corresponding coordinate buses for read-out of information from the core selected.

The negative pulses cause the read-out to occur, while the positive pulses restore information previously recorded in the core.

If the core selected were found to be in the 0 state, the negative pulses would have no effect on its state, and no emf would be induced in the read-out winding.

If the core is storing code 1, the negative pulses cause the core to change to the 0 state, and a pulse will appear in the read-out winding.

Since ferrite hysteresis loops are not strictly rectangular, cores induce an interference emf in the winding under the effects of the half-currents. Compensating cores, strobing, and integration of the output signal are used to reduce the level of interference.

The type z memory is made in such a way that the read-out current will only act on the core of one numerical cell at a time. A simplified structural layout of this device is shown in Figure 15.15.

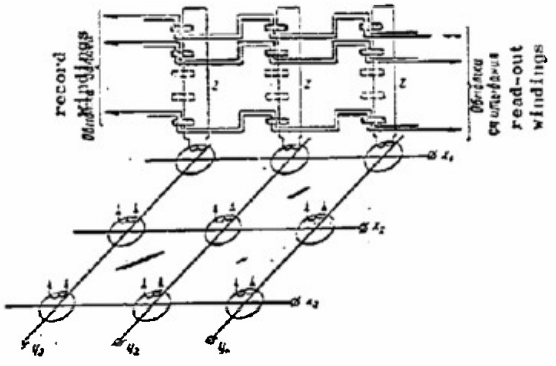


Figure 15.15. Simplified structural layout of the type z memory.

All the cores in each of the numerical cells thread a single winding connected to a ferrite core transformer tap and serving to select the cell, and called a coordinate transformer. All coordinate transformers are combined into a matrix so a cell can be selected on the current coincidence principle to record, or read-out, a number.

Each coordinate transformer is threaded by three buses; two coordinate (x and y), and an output (z).

Two pulses are formed in the z winding of the selected coordinate transformer; a negative one with amplitude I_m, and a positive one with amplitude $I_m/2$.

The negative pulse zeroes all the cores. At the same time, pulses occur in those read-out windings that pass through the cores and are in the 1 state.

The positive pulses record or restore information.

In order to record code 1 in the core (or to restore this code after read-out), a positive pulse with amplitude $I_m/2$ is fed to the record winding at the same time that the positive pulse with amplitude $I_m/2$ is acting in the z winding. Current coincidence will cause the selected core to convert to the 1 state.

There is no pulse in the record winding when 0 is being recorded (or restored).

The special feature of the operation of the cell in the type z memory is that the load resistance of the coordinate transformer will change with the code recorded in the cell. Compensating cores (fig. 15.16) are inserted in each cell, in addition to the working cores, in order to stabilize the load.

The capacity of an electronic digital computer ferrite cube is several thousand numbers.

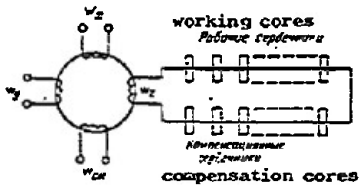


Figure 15.16. Simplified layout of an individual cell in a type z memory.

Magnetic drum. The magnetic drum is used as an external memory in modern electronic digital computers, as well as an operative memory, but a slow-speed one.

The magnetic drum is a cylinder, the surface of which is covered with a magnetic medium. Information is stored on the surface of a magnetic drum in the form of magnetized segments.

Magnetic heads record and read-out the information.

When information is being recorded the current flowing through the magnetic

heads creates a magnetic field, and this in turn forms a magnetized segment on the surface of the drum.

During read-out, the magnetized segment, passing under the magnetic heads, induces a voltage in them corresponding to the 0 or 1 signal.

The information is in the form of tracks around the periphery of the magnetic drum. The density with which the binary signs are recorded along the tracks determines the speed of rotation of the magnetic drum and the length of the code pulses.

The rate of rotation of the magnetic drum in modern electronic digital computers is between 100 and 2000 rpm, and the density of the recording along the tracks is from 2 to 30 pulses/mm.

The transverse density of the recording (the density of the recording with respect to the cylinder's generator) is 2 to 10 tracks/cm.

These characteristics make the capacity of the magnetic drum type memory several hundred thousand binary signs.

Spot marks are made on the surface of the magnetic drum and are used to form synch pulses which are fed to the cell address counter in order to gain access to, or record, a number at a particular address. The reading on the address counter corresponds to the number of the cell that is positioned under the heads of the magnetic drum at the particular moment in time.

The magnetic drum is included among the cyclic access memories. The period during which access is had to the cell is fixed by the drum diameter and the revolution rate.

Magnetic tape. Magnetic tape memories are the most widely used type of external accumulator associated with the electronic digital computer.

The chief advantages of the magnetic tape memory are:

- high capacity memory, reaching hundreds of thousands of numbers;
- highly reliable storage and delivery of information;
- simplicity.

The magnetic tape is an elastic base with a ferromagnetic coating. Mag-

netic heads are used to record, read-out, and erase information on the magnetic tape.

The tape is divided up into zones for convenience of access. The address code of the zone is recorded at the beginning of each zone.

The information is located on tracks on the tape under the magnetic heads.

The chief disadvantage of the magnetic tape memory is the long access time.

Arithmetic Device

The arithmetic device in the electronic digital computer is designed to carry out arithmetical and logical operations. The arithmetic device can be of the parallel, or series, type, depending on the method used to introduce the numbers.

The parallel arithmetic device operates on all the orders of numbers simultaneously. The series arithmetic device operates on the numbers in turn, usually beginning with the low order ones.

High-speed, parallel devices are the ones most widely used.

The main unit is the summer. The device also contains registers for storing the original numbers, as well as the intermediate and final results.

The structure of the arithmetic device depends on the manner in which the numbers are represented. The arithmetic device has two summers in a machine with a floating point, a mantissa summer, and a sequence summer.

The arithmetical and logical operations performed by the device can be broken down into a series of elementary operations called microoperations. For example, microoperations include zeroing the arithmetic device's registers and summers, receiving the number code, shifting the number code to the low, or high, orders, and delivery of the result code.

The microprogram is the predetermined sequence of microoperations performed by the arithmetic device. For example, the following microprogram can be used for addition in the arithmetic device:

setting the receiving register and the summer on 0;
transmitting the augend code from the operative memory to the register;
transmitting the augend code from the register to the summer;

setting the receiving register on 0;

transmitting the addend code from the operative memory to the register;

transmitting the addend code from the register to the summer;

transmitting the sum obtained from the summer to the operative memory.

Completion of the microprogram is through the use of signals generated in the local control unit of the arithmetic device.

Control Device

The control device is designed to automatically carry out the program and coordinate the operations of all the devices in the computer.

The control device carries out the following basic functions:

computer starting and stopping;

inputs of original data and programs in the operative memory;

automatic completion of the program;

delivery of results;

monitoring the processes involved in carrying out the program.

Commands are used to enable the control device to carry out programs automatically. As the process is carried out the control device selects the command from the operative memory, generates the control signals needed to carry out the command, and prepares to select the next command.

The following basic units are contained in the control device:

the command register;

the operation code register, and the operation decoder;

clock pulse unit;

command counter.

The control device also includes a control and signal panel, a start-stop unit, cyclic operation block, and other blocks.

Let us see how the control device carries out a three-address command.

The address of a command for access is stored in the command counter. The command code for this address is fed from the operative memory to the command register. The command register has two sections, one operational, the other address. The operation code is fed from the command register to the operation register, and the address code is fed to the address register in the operative memory. The numbers concerned with the operation are fed from the operative memory to the arithmetic device by address codes. The

operation code is fed from the operation register to the decoder. The number of decoder outputs equals the number of operations carried out by the electronic digital computer.

The signal for the particular operation, subsequently fed to the local control block in the arithmetic unit, is formed at one of the decoder's output buses according to the operation code. The local control block in the arithmetic device forms the microprogram for carrying out the particular operation in accordance with the operation signal. When the arithmetic device has completed the operation the result is recorded in the operative memory according to the address indicated in the particular command.

At the end of the cycle the reading on the command counter, which is storing the command in progress, increases by one. It can be anticipated that it is possible to change the command counter reading by a number different from one. This is necessary when carrying out a command for conditional and unconditional transition.

Control signals, formed by the clock pulse unit, ensure that the sequence in which the control device functions to carry out commands is followed.

Input and Output Devices

The input device is designed to transcribe the original information from punched tape (punched cards) to the electronic digital computer memory.

The punched tape is a carrier of information. Numbers and commands, represented by combinations of 0 and 1, are present on the punched tape in the form of a system of holes (perforations) positioned in several rows. The presence of a hole in a corresponding position usually signifies 1 in a predetermined order of numbers, with absence of a hole depicting zero. A special electromechanical device, a perforator, punches the holes to transcribe the information onto the tape.

Photoelectric equipment is used to read-out information from modern electronic digital computers. This method moves the translucent tape in front of an illuminated slot for the read-out. Every time a hole passes in front of the slot the light flux strikes the photoelectric cell and is converted into an electrical pulse. Read-out can take place at a rate of 600 signs/second.

Punched cards are used in some electronic digital computers for inform-

ation input. The advantages of the punched card over the punched tape is that damaged cards can be replaced, and it is more convenient to rearrange them when changes in the information they contain are made.

The output device usually consists of three main blocks:

the register block;
the local control block;
the printer.

The register block is designed to receive the numbers, convert them into an octonary, or a binary-decimal, code, and produce the control signals for the electromagnets in the printer.

The local control block selects the type of printing to be used, in the octonary code, or in the decimal code, and others.

Two types of printers are used, bar and wheel. Bar printers can print 1 or 2 numbers per second, while the wheel printer can print 30 to 40 numbers per second.

Photographic adapters, cathode ray tubes, electronic recording potentiometers, and other instruments, can be used as the output devices in special purpose electronic digital computers.

Chapter XVI

Fundamentals of General Information Theory

16.1 Basic Definitions

Communication system. A communication system is understood to mean all of the devices and facilites required to transmit messages from sender to receiver. A generalized communication system can be shown in block schematic form (fig. 16.1).

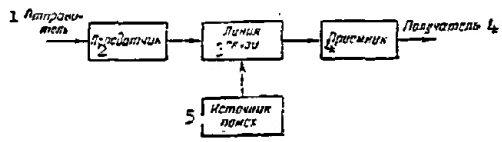


Figure 16.1. Block schematic of a communication system. 1 - sender; 2 - transmitter; 3 - communication line; 4 - receiver; 5 - noise source.

Transmitter. The transmitter is a device that processes messages and produces communication signals in a manner established by the particular design.

Receiver. The receiver is a device that converts the communication signal received and restores the signal message.

Noise source. Noise is always present in a communication system. The effect of noise on the useful signal is to distort the message transmitted when the message is received at the receiver input. The degree to which the incoming message corresponds to that sent establishes communication authenticity, and the capacity of the system to provide authentic communications in the presence of noise is called system noise immunity.

Communication channel. A communication channel is understood to mean the totality of equipment used to ensure the independent transmission of the particular message over a common communication line in the form of corresponding communication signals.

Communication signal. In electric communication the signal is understood to mean the electrical distrubance uniquely reflecting the message. Communication signals vary greatly in shape, and are voltages, or currents, that changes in a definite manner with the passage of time.

Message. In information theory the message is understood to mean the totality of the information that can be transmitted to a receiver, or in other words, the message is everything that lends itself to transmission.

In remote control systems the message is a limited number of previously conventionalized commands, and in telemetry the result of a measurement in some magnitude.

In telegraph communication the message is any text written in a predetermined alphabet, and in telephone communication the message is any sound (speech, music) in a predetermined frequency band.

In facsimile communication the message can be any text, drawing, photograph or picture, or images on ordinary or special paper, in a definite format. Television communication is any fixed or moving image projected onto the transmitting tube.

When messages are transmitted, the random nature of their appearance is intrinsic. Actually, if there were laws enabling the receiver to visualize how the message was going to appear at any given moment, the transmission of messages would provide no new information.

Information. Information, insofar as information theory and the transmission of messages is concerned, is understood to mean data concerning the results of some event which should take place, or has already taken place, but the outcome of which had not been known ahead of time. Information delivered from the message source is included in human speech. in the reading of a measuring device, in a telephone call, telegram, or radio message. in facsimile broadcasts, television and radar images, and the like.

Any message contains some useful data (information)

Information Content. General information theory measures the information content in binary digits. It is abbreviated as b.d., or bit. The word bit stems from the abbreviation of the words binary digits in the English language.

When the case is one of equally probable messages, the numerical measure of the information content, I, can be taken as the binary logarithm of the number of possible messages N:

$$I = \log_2 N \text{ [bits]}. \tag{16.1}$$

If N = 2,

$$I = \log_2 2 = 1 \text{ bit}.$$

Therefore, it is accepted that the unit of information content is that information content that does away with the ambiguity in the selection of one of two equally probable outcomes. This unit for the information content is called the binary digit, or bit.

Each of N messages can be transmitted in the form of a coded combination of length n, which is a combination of m possible symbols, that is $N = m^n$. Now the information content in a message made up of n symbols will be equal to

$$I = n \log_2 m \text{ [bits]}. \qquad (16.2)$$

One binary pulse contains one bit of information, that is, if $m = 2$, and $n = 1$, then $I = 1$ (because $\log_2 2 = 1$).

In binary systems the information content is equal to the number of pulses, $I = n$.

In the general case the probability of the appearance of different elements (different messages) can be unequal. For example, the letters most often encountered in words in the Russain language are O, A, E, H, and those least often encountered are X, Ф, Щ

The following formula can be used to find the information content

$$I = n \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} \text{ [bits]} \qquad (16.3)$$

where

n is the number of elements in the message;

m is the total number of possible element states;

p_i is the probability of the appearance of the ith element state.

The formula at (16.3) is correct for sufficiently long messages.

It is correct as an average for short messages.

If the probability of some element is equal to 1 ($p_i = 1$), the information content will be equal to zero ($I = 0$). This is only natural, because in this case the situation is completely clear ahead of time, and the message will provide nothing new. On the other hand, when all the elements are equal in probability, the situation has its greatest ambiguity and the information content becomes maximum. The formula at (16.3) is converted into the expression at (16.2).

Entropy. Information theory uses a magnitude that defines the specific information content arriving per element in the message

$$I' = \frac{I}{n} = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i} \text{ [bits/element]}. \qquad (16.4)$$

In the special case of equally probable elements

$$I' = \frac{I}{n} = \log_2 m \text{ [bits/element]} \qquad (16.5)$$

The specific information content is only dependent on the properties of the message source. Conventionally, this magnitude is called entropy of the message source. The dependence of the entropy, I', on the probability p_i, of one of two messages ($m = 2$) is shown in Figure 16.2. System entropy (the entropy of the message source) is maximum when the events are of equal

probability, and equals zero when one of the probabilities is equal to one.

Thus, nonhomogeneity in message elements reduces the information content of a message, and this makes it possible to speed up message transmission.

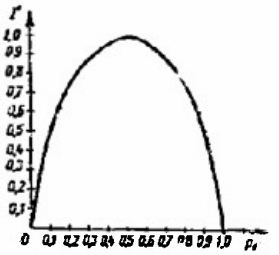


Figure 16.2. Dependence of the information content arriving per message element on the probability on one of two messages.

16.2 Principal Characteristics of Signals and Communication Channels

Signal volume and channel capacity. When solving practical problems stemming from general information theory, the signal is characterized by the volume, V_s, equal to the product of three of its characteristics; signal length τ_s, signal spectrum width, ΔF_s, and excess H_s:

$$V_s = \tau_s \Delta F_s H_s. \tag{16.6}$$

The product of these three magnitudes plotted parallel to the axes in a Cartersian system of coordinates is, geometrically, a parallelepipod with sides τ_s, ΔF_s, H_s, and hence the produce is called the volume of the signal.

Signal length defines the interval of time over which the signal exists.

Signal spectrum width is the interval of frequencies within which the frequency spectrum-limited communication signal is located. The real communication signal is contained within the spectrum of frequencies within the communication channel pass band.

The excess of average signal power, P_s, over average noise power, P_n, is found through the expression

$$H_s = \log P_s/P_n. \tag{16.7}$$

The minimum average signal power should be greater than the average noise power.

A communication channel, by its physical nature, is in a condition to pass effectively only signals, the spectra of which are in the limited band ΔF_c, for a permissible range of change in power, H_c, and which is determined by the difference in the maximum permissible signal level in

the channel and the noise level, normalized for the particular type of communication channel. Moreover, the communication channel is made available to the message sender for a completely established period of time, τ_c. Consequently, the concept of channel capacity, V_c, is introduced in the general theory of communications by analogy with the signal, and is defined as the product of the three channel characteristics mentioned

$$V_c = \tau_c \Delta F_c H_c. \tag{16.8}$$

A necessary condition for the transmission of a signal with volume V_s over a communication channel, the capacity of which is equal to V_c, is $V_c \geq V_s$, or

$$\tau_c \Delta F_c H_c \geq \tau_s \Delta F_s H_s. \tag{16.9}$$

The physical characteristics of a signal can be changed, but if they are, a reduction in one is accompanied by an increase in another. For example, if a signal is recorded on a magnetic tape at slow speed and is then reproduced at high speed, the signal length will be reduced, but its spectrum width will be increased by an equal factor.

Traffic handling capacity and transmission speed. Traffic handling capacity of a communication channel is understood to mean the number of messages that can be transmitted in unit time with the required degree of transmission accuracy. In other words, the traffic handling capacity is the limiting speed at which information can be transmitted.

The limiting traffic handling capacity depends on the channel bandwidth, as well as on the signal noise power ratio, and is found through the formula

$$C_{max} = \Delta F_c \log_2 (1 + P_s/P_n) \text{ [bits/second]}, \tag{16.10}$$

where

P_s and P_n are the average signal and noise powers, respectively;

ΔF_c is the channel bandwidth.

The expression at (16.10) is called the Shannon formula and is correct for any communication system when fluctuation noise is present.

The following expression is convenient to use for practical computations

$$C_{max} = 1.44\, \Delta F_c \log (1 + e^{2\Delta P}), \tag{16.11}$$

where

ΔP is the difference in the signal and noise levels, in nepers.

Modern telephone channels, which have ΔP = (4.5 to 5.5) nepers, can theoretically reach a traffic handling capacity of 40 to 50 thousand bits per second. However, modern systems used to transmit information over

telephone channels have a capacity of no more than 3400 bits/second. Thus, the Shannon formula suggests an extremely significant future increase in the traffic handling capacity of information transmission systems. Moreover, what also follows from the formula is that authentic transmission of messages can take place even when $P_s < P_n$.

In telegraphy the traffic handling capacity is characterized by telegraphic speed, and this is defined as the number of elementary code pulses that can be transmitted in one second. The unit of speed in telegraphy is the baud, that is, the telegraphic speed at which one elementary pulse of current is transmitted in one second.

Accordingly, telegraphic speed is

$$B = 1/\tau_o \text{ [baud]}, \tag{16.12}$$

where

τ_o is the length of the elementary current pulse, in seconds;

So far as the telegrephic speed limit is concerned, the length of an elementary pulse, τ_o, can be equal to its rise time, t_r.

The length of the rise is fixed by the channel band width, ΔF_c, and is approximately equal to:

when modulated signals with two side bands are transmitted

$$t_r = 1/\Delta F_c;$$

when transmission is with one side band, or is of unmodulated direct current pulses

$$t_r = 1/2\Delta F_c.$$

Accordingly, in telegraphy, the maximum transmission speed is equal to

$$B_{max} = 2\Delta F_c. \tag{16.13}$$

This formula, called the Nyquist criterion, is of great value in communication engineering.

Since one elementary pulse carries one bit of information, the traffic handling capacity of a binary channel will equal

$$C_{max} = B_{max} = 2\Delta F_c \text{ [bits/sec]}. \tag{16.14}$$

Practical transmissions are made at slower speed, so a satisfactory shape of incoming signals can be expected without the use of special circuits. It is believed that it is sufficient to provide for the transmission of three harmonics of the fundamental signal frequency, that is

$$\Delta F_{c\ min} = 3F_1, \tag{16.15}$$

where

$\Delta F_{c\ min}$ is the minimum channel bandwidth;

F_1 is the fundamental keying frequency, equal to half the telegraphic speed(F_1 [hertz] = B [bauds]/2).

Consequently

$$\Delta F_{c\ min} = 3\ B/2 = 1.5\ B$$

Transmission speed in bauds, and traffic handling capacity in bits per second, connot coincide in binary systems because of the transmission of service information (of synchronizing pulses) and of excess information during coding. Therefore, the traffic handling capacity in terms of useful information is usually lower than the telegraphic speed

$$C_{use} = n_{use}/T_t, \qquad (16.16)$$

where

n_{use} is the number of pulses of useful information transmitted;

T_t is the duration of the transmission.

Transmission levels. In wire communication the power at the Bending end, equal to one milliwatt, and called the zero level (P_o = 1 milliwatt = 10^{-3} watt), is taken as the original magnitude for power comparison purposes. If the line input resistance equals R_{in} = 600 ohms, the zero voltage level is U_o = 0.775 volt, and the zero current level is I_o = 1.29 milliamps.

It is convention to use a logarithmic scale to compare powers, and the level, expressed in nepers (units of level), can be found through the formula

$$p[\text{nepers}] = 1/2 \log P_t/P_o = 1/2 \log P_t/1 \text{ milliwatt}, \qquad (16.17)$$

where

P_t is the power at the termination, expressed in milliwatts.

If the level is expressed in decibels, then

$$p\ [\text{db}] = 10 \log P_t/1 \text{ milliwatt}. \qquad (16.18)$$

The conversion from one set of units to the other can be made through the following equalities

$$1 \text{ db} = 0.115 \text{ neper},$$
$$1 \text{ neper} = 8.68 \text{ db}.$$

The difference in levels at the input and output of a communication channel is called the overall line attenuation. According to prevailing standards the overall line attentuation for telephone channels whould range between 0.8 and 1.0 neper.

If the power at the sending end is P_s, and the power at the termination is P_t, the attenuation in nepers will be

$$b\ [\text{nepers}] = 1/2 \log P_s/P_t. \qquad (16.19)$$

Similarly for the voltages

$$b\ [\text{nepers}] = \log U_s/U_t,$$

and for currents

$$b\ [\text{nepers}] = \log I_s/I_t.$$

Amplifiers are inserted in the line to compensate for signal attenuation on long lines. The overall line attenuation, b_o, for the entire communication channel can be found as the algebraic sum of the attenuation, b_i and the gain, K_i, for the individual sections.

$$\Delta b = b_o = \sum b_i - \sum K_i \qquad (16.20)$$

Channel frequency response curve. The plot of the overall line attenuation in terms of the frequency is called the communication channel frequency response curve.

In a telephone channel the frequency band bounded by the frequencies on which the overall channel attenuation exceeds the channel's overall attenuation at a frequency of 800 hertz by 1 neper is called the band of effectively transmitted frequencies.

Figure 16.3 shows the plot of the overall line attenuation in terms of the frequency (the limits of permissible deviations for a standard telephone channel with a frequency band from 300 to 3400 hertz).

Channel amplitude response curve. The plot of the transmission level at the output of a channel (or the increase in its overall line attenuation) in terms of the magnitude of the level at the input is understood to be the channel amplitude response curve. Inflection of the amplitude response curve is permissible for values of the level at the input of at least +0.8 nepher if nonlinear distortions are to be avoided in the standard telephone channel.

Channel phase response curve and group porpagation time. The channel phase response curve is understood to be the plot of the phase shift, φ, between the oscillations at the channel input and output in terms of the carrier frequency ω.

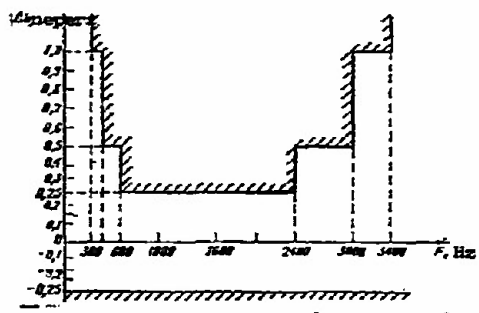


Figure 16.3. Dependence of the overall line attenuation on the frequency for a normal telephone channel. The channel frequency response curve should fall between the hatched boundaries.

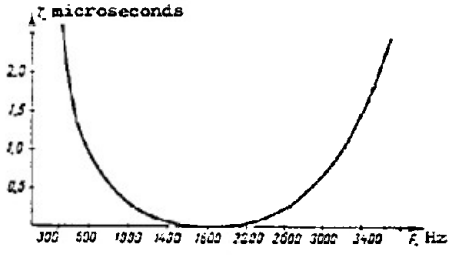


Figure 16.4. Phase response curve for a telephone channel.

If the phase response is linear over the entire spectrum of transmitted signal frequencies the signal will be transmitted without distortion and will appear at the channel output after time interval τ_g, which is called the group propagation time

$$\tau_g = d\varphi/d\omega.$$

Group propagation time is the time interval between signal delivery to the input to appearance at the channel output of the maximum energy of some group of oscillations within a sufficiently narrow band of the spectrum.

If the phase response curve is nonlinear, the group propagation time will depend on the frequency and on the different groups of signal frequency components delivered to the channel output at different times. The result will be signal distortion.

In real communication channels the minimal nonuniformity in the group propagation time takes place in the middle of the curve (fig. 16.4). This is why the center of the channel band is used to transmit telecode signals.

16.3 Converting the Message Into a Signal

Messages to be transmitted are usually nonelectrical in nature. A text, for example, is a set of characters (letters and punctuation marks), and sound is a change in pressure in terms of time, etc. The communication channel can only transmit the electrical reflection of the message, the signals.

The porcess of converting a message into a signal involves three operations; conversion, coding, and modulation.

Conversion

The conversion of a message into a communication signal involves translating nonelectrical magnitudes, those including in the original message, into electrical ones. For example, a photo electric cell is used to convert the image reflected from an object by a beam of light into the corresponding electrical oscillations.

Coding

Coding is the building of the message, or signal, on a definite principle. A code is a combination composed of different elementary signals. The particular code selected must provide:

diversity of code elements so the receiver can react to the individual elementary signals contained in the code;

simplicity in building the code, that is, use of a minimum number of code elements (of a base) and the maximum number (index n) of possible combinations of electrical transmissions, in accordance with the needed number, N, of meaningful messages

$$N = m^n. \tag{16.21}$$

Telegraph type printers use a uniform, five-unit code for each character. The number of elements in the code is m = 2 (binary code). The number of elementary electrical transmissions is n = 5 (five-element code).

Then

$$N = m^n = 2^5 = 32.$$

This five-element binary code can transmit 32 different characters (messages).

Uniform codes. Uniform codes are those in which each code combination consists of the same number of elements. A five-element telegraphic code (fig. 16.5) is an example. The digits can be transmitted by a positive pulse, zero by negative pulses of the same length. The transmission by a positive pulse, zero by negative pulses of the same length. The transmission time for any code combination in a five-element code equals 5_{T0}

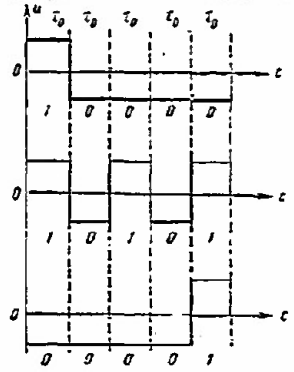


Figure 16.5. Selected combinations of a uniform five-element code.

A six-element unifrom code can be built in the same way, so $2^6 = 64$ different combinations can be transmitted. The process for a seven-element, etc., code is similar.

Uniform codes make it comparatively simple to build transmitter and receivers for automated and automatic discrete communication systems.

Nonuniform codes. Nonuniform codes are those in which the code combinations differ from each other, not only with respect to the interposition of zeros and digits, but also in their numbers.

A typical example of the nonuniform code is the Morse code, in which the elements in the code combinations of digits and zeros are used in but two combinations; singly (1 and 0), or in triplets (111 and 000). The signal corresponding to one digit is called a dot, that corresponding to three digits a dash. The zero element is used to separate a dot from a dash, a dot from a dot, and a dash from a dash. A set of three zeros ends each code combination, thus simply separating one code combination from another. Figure 16.6 is an example of a Morse code combination.

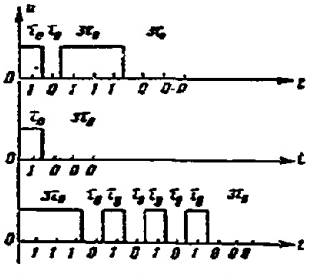


Figure 16.6. Selected combinations of a nonuniform Morse code.

The time required to transmit each of the characters in the Morse code is different. The shortest combination in the code is the letter E, with a length equal to 4_{TO}, and the longest is 22_{TO} (the number 0).

Given the structure of the Russian language, the average for transmitting by Morse code is some 9.5 elementary pulses per character, making it a less economical code as compared with the five-element uniform code. The advantage of the Morse code is that it can be received by ear, hence is widely used in radio communications for sending by key and receiving by ear.

Modulation

Modulation is the effect on one of the parameters of the signal carrier resulting from law governing the change in the function reflecting the message transmitted over the communication channel. The transfer agent used in electric communication is direct current, high or low frequency alternating current, or a periodic sequence of short pulses.

Pulses of code combinations can be transmitted directly over lines passing extremely low frequencies, such as physical lines, or cables, for

example. If the transmission must be made over telephone channels, which will not pass very low frequencies, recourse is had to modulation of one of the parameters of a harmonic oscillation, called the carrier

$$u(t) = U_m \sin(\omega t + \varphi_c), \tag{16.22}$$

where

U_m is the amplitude;

ω is the angular frequency;

φ_c is the initial phase of the oscillations.

Modulation can be amplitude (AM), frequency (FM), or phase (PM), depending on the parameter modulated.

Amplitude modulation. A radio-frequency oscillation, the amplitude of which changes according to the law governing the modulating voltage, is said to be amplitude modulated. For example, an oscillation with carrier frequency f can be changed by the amplitude of the oscillations occurring in accordance with the law governing the modulating frequency f (fig. 16.7).

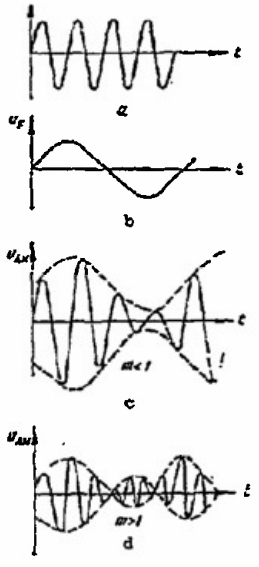


Figure 16.7. Oscillograms of amplitude modulated oscillations.
a - carrier frequency oscillations;
b - modulating frequency oscillations;
c - amplitude modulated signal with modulation factor less than one;
d - amplitude modulated signal with modulation factor greater than one.

The modulation factor, $m = \Delta U/U_m \leq 1$, must be less than one, or at the very worst equal to one, in order to transmit the signal without distortion. When $m > 1$, overmodulation occurs, and the envelope of the amplitude modulated oscillation does not repeat the shape of the curve of the modulating voltage, and signal distortion will result. The practical modulation factor is taken as equal to 0.8 to 0.9.

The amplitude modulated oscillation consists of three components in the simplest of cases, when the modulating voltage is a low frequency sinusoidal oscillation; a carrier frequency oscillation, and two side frequencies (fig.16.8). If the modulation is by the frequency spectrum, as is the case when speech is transmitted, the amplitude modulated oscillation will consist of the carrier frequency oscillation and two side bands (fig 16.9).

If the modulating signal is designated x (t), the expression for an AM signal will be in the form

$$u(t) = U_m [1 + mx(t)] \sin(\omega t + \varphi_0). \tag{16.23}$$

When binary signals are transmitted m = 1, so

$$u(t) = U_m [1 + x(t)] \sin(\omega t + \varphi_0). \tag{16.24}$$

The amplitude of the modulated signal will change from $2U_m$ to zero.

The width of the frequency spectrum of an amplitude modulated signal is twice that of the maximum modulating frequency. For example, if the maximum transmitted speech frequency is equal to F = 3.4 kHz, the width of the spectrum of the amplitude modulated signal will be $\Delta F = 2F = 6.8$ kHz.

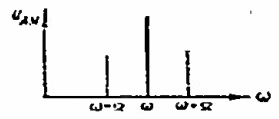


Figure 16.8. Frequency spectrum of an amplitude modulated signal when the carrier is modulated by a single audio frequency.

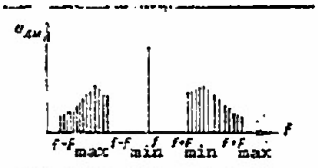


Figure 16.9. Frequency spectrum of an amplitude modulated signal when the carrier is modulated by the audio frequency spectrum.

Single-sideband modulation. Single-sideband modulation is a type of signal transmission in which the transmitter radiates only a single sideband with the suppressed carrier (the pilot signal), or without it (fig. 16.10).

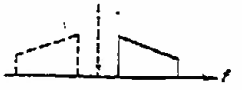


Figure 16.10. Frequency spectrum in the case of single-sideband modulation.

In the receiver the carrier frequency from the local oscillator is added to the incoming single sideband signal and the audio frequency spectrum is restored in the demodulator.

The advantage of single-sideband modulation is that the band of frequencies radiated by the transmitter is halved so transmitter power can be used more effectively because it is expended on radiating the single sideband and this, in turn, increases reception noise resistance because of the narrowing of the passband.

A disadvantage that must be pointed out is the extremely regid demands imposed on the frequency stability for the transmitter and the receiver's local oscillator, because this serves to complicate the equipment and increase its cost.

Frequency modulation. When frequency modulation is used, the amplitude of the oscillations of the carrier frequency remains constant, but the frequency changes according to the law reflecting the nature of the modulating voltages (fig. 16.11).

A frequency modulated oscillation can be represented in the form of s vector of fixed length swinging around its initial position with an angular frequency Ω (fig. 16.12). The maximum angle of deflection is $\varphi_m = m_f$ (the modulation index) and it can be defined as the ratio of the amplitude of the change in the frequency, ω_d, to the angular frequency of the modulation Ω.

$$m_f = \omega_d/\Omega = f_d/F. \qquad (16.25)$$

The modulation index characterized the modulation depth. If the amplitude of the change in the frequency, f_d, or, and this is the same thing, the frequency deviation is less than the modulation frequency, that is, if $f_d < F$, the modulation index will be $m_f < 1$. In this case the vector, swinging near its middle position, will be deflected from it by an angle less than 57°. This is called narrow band frequency modulation and occupies the same frequency band as is found in amplitude modulation.

But if $f_d \gg F$, the maximum angle of deflection, $\varphi_m = m_f$, can be larger

than 360°, with the result that the vector, rotating around its middle position, will complete more than one full turn to some particular side. This is called broad band frequency modulation and covers the frequency band equal to approximately double the deflection frequency, $2f_d$.

The width of the spectrum of a frequency modulated signal can be defined accurately enough for practical use through the approximate formula

$$\Delta F \approx 2F(1 + m_f + \sqrt{m_f}), \qquad (16.26)$$

where ΔF is the width of the spectrum of the frequency modulated signal;

F is the modulation frequency;

m_f is the modulation index.

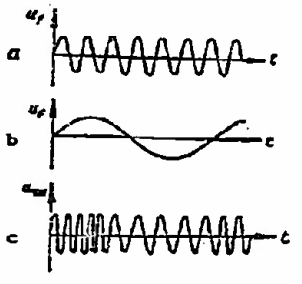


Figure 16.11. Oscillograms of frequency modulated oscillations.
a – carrier frequency oscillations;
b – modulating frequency oscillation;
c – frequency modulated signal.

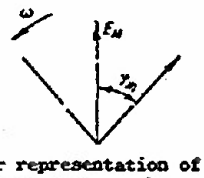


Figure 16.12. Vector representation of a frequency modulated oscillation.

Phase modulation. Phase modulation can be considered to be a variety of frequency modulation, with the phase of the high-frequency oscillation changing. Where phase modulation differs from frequency modulation is in phase modulation being defined by the angular deflected of the vector resultant from the position of the carrier frequency, whereas frequency modulation is defined by the rate of this deflection, that is, by the derivative of the phase in terms of time.

When a pure tone is modulated it is difficult to conclude how it was modulated, in frequency, or phase, just from the nature of the oscillation and its properties. In both cases the vector (fig. 16.12) swings relative to its initial position in a manner such that angle φ will change with respect to time according to the law

$\varphi = m_f \sin \Omega t$ in the case of phase modulation;

$\varphi = \omega_d/\Omega \sin \omega t \sin \Omega t$ in the case of frequency modulation.

The difference between frequency and phase modulation will only show up in the event there is a change in the frequency of the modulation, or when the frequency band is modulated. Then too, frequency and phase modulation differ in the manner in which the modulation is carried out. In the case of the former the direct effect of oscillator oscillations on the frequency is used, whereas in the case of phase modulation the oscillator is modulated in a phase modulator.

In general, if the modulating signal is x (t), the expression for finding the signal in the case of phase modulation is in the form

$$u(t) = U_m \sin [\omega t + \Delta\varphi x(t)], \tag{16.27}$$

where

$\Delta\varphi$ is the phase deviation.

The condition $\Delta\varphi \leq 90°$ must be satisfied in order to avoid overmodulation.

Phase modulation is accompanied by a change in signal frequency, because, by definition, frequency is the rate of change in phase in terms of time

$$\omega(t) = \frac{d\varphi(t)}{dt}. \tag{16.28}$$

In the case of phase modulation of the frequency

$$\omega(t) = \frac{d}{dt}[\omega t + \Delta\varphi x(t)] = \omega + \Delta\varphi \frac{dx(t)}{dt}. \tag{16.29}$$

The changing frequency is called the instantaneous frequency, and the changing phase is called the present phase

$$\varphi(t) = \int \omega(t)\, dt. \tag{16.30}$$

Phase modulation is not used for transmitting speech, and it is only recently that it has been used in discrete communication systems to transmit binary signals (see keying, or on-off modulation).

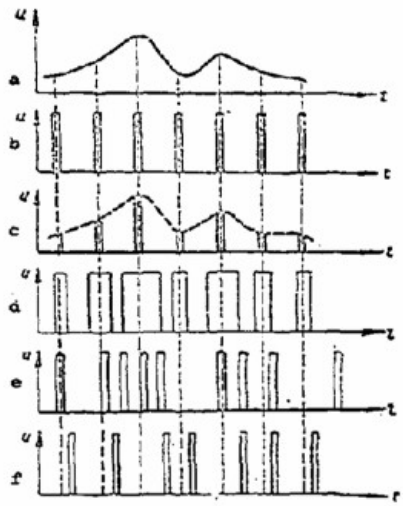


Figure 16.13. Oscillograms of different types of pulse modulation. a - modulating voltage; b - clock pulses; c - amplitude modulated pulses; d - pulse length modulation; e - frequency modulated pulses; f - phase modulated pulses.

Principal types of pulse modulation. A distinguishing feature of pulse communication engineering is the use of a periodic pulse train as the message transfer agent, characterized by the following parameters:

pulse height (amplitude), h;

pulse length, τ;

pulse repetition rate, $F = 1/T$;

position of the pulses in terms of time relative to the position of the pulses of an unmodulated train, that is, the pulse phase.

If we change one of the above parameters by the modulating function we can obtain four principal types of pulse modulation (fig. 16.13):

amplitude pluse modulation APM;

frequency pulse modulation, FPM;

phase pulse modulation, PPM;

pulse length modulation, PLM.

Signal quantization. Signal quantization is the presentation of a continuous signal in the form of a discrete one. In practice, this is done by replacing the continuous physical magnitude by a finite set of its values (quantized levels).

There are two quantization cases; in terms of level (fig. 16.14) and in terms of time (fig. 16.15). In the figures X(τ) is the continuous signal and S(τ) represents the discrete values of the continuous signal

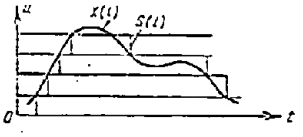


Figure 16.14. An example of level quantization.

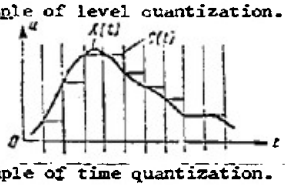


Figure 16.15. An example of time quantization.

In the case of level quantization, the discrete signal remains unchanged until the value of the continuous signal reaches some present magnitude, and this can be at any moment in time.

In the case of time quantization, the magnitudes of a continuous signal that exist at a present moment in time are recorded. The discrete magnitude can be rounded to the nearest quantization level in this case.

In the case of combination quantization, the signal can be quantized in terms of time and, in addition, its value can be level quantized at time points.

V.A. Kotel'nikov's theorem. If the function X(t) contains no frequency higher than F_m hertz, it can be completely defined by the sequence of its values at a moment in time read, one to another, by $1/2F_m$ second.

Kotel'nikov's theorem points out that given these conditions, the transmission of a continuous signal can be reduced to the transmission of individual pulses, or of code combinations.

A continuous communication signal is a function with a limited spectrum, so, in accordance with Kotel'nikov's theorem, there is no need to transmit all the values of the function in order to completely define the curve of the function X(t) at the receiver. It is sufficient to transmit the individual, instantaneous, values (fig. 16.16) read at time intervals

$$\Delta t = 1/2\Delta F,$$

where

ΔF is the spectrum width for the function.

The readings are most usually taken as the width of the spectrum for the function. The number of sample values can be computed through the formula

$$N = \frac{T}{\Delta t} = 2\Delta FT, \tag{16.31}$$

where

T is the limited interval during which the function X(t) is considered.

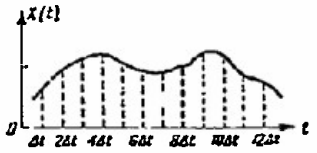


Figure 16.16. An example of the transmission of a continuous function by its discrete values.

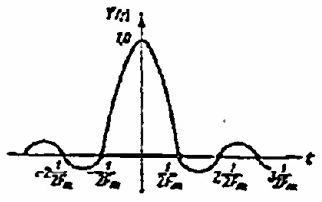


Figure 16.17. Curve of the function of the readings.

The auxiliary function (the function of the readings)

$$Y(t) = \frac{\sin 2\pi F_m t}{2\pi F_m t} = \frac{\sin \Omega_m t}{\Omega_m t}. \tag{16.32}$$

is used for an analytical definition of the function X(t), using its values at the moment of the reading.

The curve of this function is shown in Figure 16.17.

The function of the readings has a uniform spectrum in the frequency band from 0 to F_m. At point t_0 Y(0) = 1, and at point $t = n\Delta t = n1/2F_m$ the function $Y(n\Delta t) = 0$. Here n is any integer (positive or negative).

On the basis of what has been said in the foregoing, the present signal X(t) can be represented by the following sum

$$X(t) = \sum_{n=-\infty}^{+\infty} X(n\Delta t)\, Y(t - n\Delta t) = \sum_{n=-\infty}^{+\infty} X(n\Delta t) \frac{\sin \Omega_m (t - n\Delta t)}{\Omega_m (t - n\Delta t)}, \tag{16.33}$$

where

nΔt are the reading points along the t axis.

The expression at (16.33) accurately defines the function X (t) at any moment, t.

The value of Kotel'nikov's theorem for communication theory and engineering is that based on it, the transmission of a continuous message reduces to the same situation that pertains in the case of the transmission of a discrete message, or, putting it another way, in both cases the transmission reduces to one of transmitting a discrete sequence of numbers. This theorem is the basis for all pulse communication.

The necessary pulse repetition rate (the clock rate) can be found through the formula

$$F_1 = \frac{1}{\Delta t} = 2\Delta F. \qquad (16.34)$$

Pulse-code modulation. Pulse-code modulation is understood to mean the transmission of continuous functions by a binary code.

In order to obtain pulse-code modulation, the signal being transmitted (fig. 16.18a) is level (or time) quantized. The quantized level values (fig. 16.18b) in this case are represented by a sequence of numbers, 5, 7, 9, 6, 4, 2, recorded in a binary counting system. This series can be written in the form of four-digit numbers as follows: 0101, 0111, 1001, 0110, 0100, 0010.

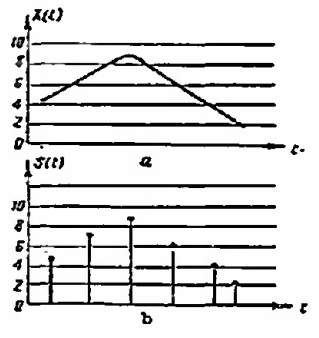


Figure 16.18. An example of pulse-code modulation
a - continuous function of a message;
b - quantized values of the function.

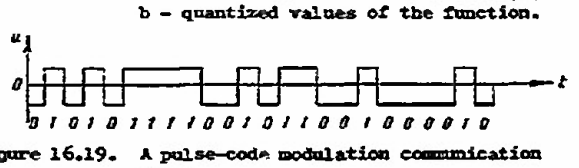


Figure 16.19. A pulse-code modulation communication signal.

Assuming that 1 designates the positive train, and that 0 designates the negative, we obtain a combination of direct current transmissions (fig. 16.19) transmitted over the communication channel.

Accordingly, quantization provides us with a way to reduce the number of different numbers subject to transmission to some finite magnitude N, which expresses the number of permissible levels there are on the quantization scale.

If a quantization step is taken as 1, (N-1) will be the notation for the largest quantized number. The necessary number of characters, n, in a binary code combination can be found from the relationship $N = 2^n$, from whence

$$n = \log_2 N. \tag{16.35}$$

If n is not a whole number it is rounded to the nearest larger whole number.

The selection of the number of levels, N, must give consideration to the fact that the fewer the quantization steps, that is, the greater the number of quantization stages, the more accurately the particular function (the continuous signal) will be transmitted. However, increasing the number of stages requires that the code combinations be increased, and this is undesirable. A compromise solution is arrived at experimentally. For example, it has been found that satisfactory speech quality in voice transmissions can be had when $n \geq 100$. Hence, it follows that a telephone conversation can be held when the pulse-code modulation uses a seven-unit binary code because $N = 2^7 = 128$.

Pulse-difference modulation. Pulse-difference modulation is understood to be the transmission of the increments of the quantized levels of a continuous signal through the medium of pulses with a calibrated shape and based on the "pulse - no pulse" principle. Here it is not the signal values proper that are quantized and transmitted, but their increments relative to the preceding values, that is, the difference between these values. The quantization is done over calibrated time intervals. This type of modulation is sometimes referred to as "delta modulation," because in mathematics the increment is designated by the Greek Δ ("delta").

The principle on which the modulation is based is as follows. The voltage of the primary signal is compared with the auxiliary voltage generated in the equipment in the form of a step function that approximately reproduces the signal shape (fig. 16.20).

The length and height of the step are constants. The comparison is made automatically at a time interval, the frequency of which is a multiple

of the clock frequency. If the signal voltage (curve 1) exceeds the step level at the time the step is completed, that is, there is a positive increment, the auxiliary voltage will increase in a bound to another step (curve 2). Accompanying the positive bound of auxiliary voltage is the appearance of a calibrated pulse positive polarity at the output of the pulse modulator. If, on the other hand, when the step is completed the signal voltage is lower than the step level, that is, the increment is negative, the auxiliary voltage will drop one step, and the pulse at the modulator output will have a negative polarity (fig. 16.20b).

Only the positive pulses will be transmitted over the communication channel (fig. 16.20c).

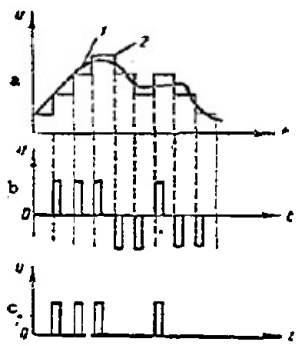


Figure 16.20. Signal conversion in the case of pulse-difference modulation.
a - signal subject to transmission (curve 1) and auxiliary voltage (curve 2); b - voltage at the pulse modulator output; c - communication signal.

Keying

If information is coded in binary codes, the modulated parameter can only take two different values. One corresponds to the transmission of 0, the other to the transmission of 1. This special case of modulation is called keying.

Figure 16.21 shows the characteristics of the various types of keying.

If the information transfer agent is direct current, the signal will have the shape shown in Figure 16.21a, provided the keying is done by the magnitude of the voltage, and will have the shape shown in Figure 16.21b if the keying is done by changing the direction of the current.

The transmission of this combination when alternating current is used as the transfer agent is brought about in a different way, and depends on the type of keying:

in the case of amplitude-shift keying, by the transmission of the carrier frequency for T_0 if 1 is being transmitted, and by the absence of the carrier frequency if 0 is being transmitted (fig. 16.21c);

in the case of frequency-shift keying, by the transmission of the carrier frequency, f_1, corresponding to 1, and by the transmission of frequency f_2, corresponding to 0 (fig. 16.21d), with the result that here the beginning of the transmission of an element in the coded combination different from the preceding one can be characterized by a transition from frequency f_1 to f_2, or from f_2 to f_1;

in the case of phase-shift keying, by changing the phase of the carrier with each transition from state 1 to state 0, or vice versa (fig. 16.21e).

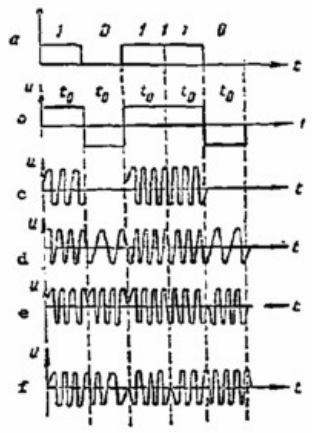


Figure 16.21. Types of keying.

a - amplitude-shift keying of direct current in one direction; b - amplitude-shift keying of direct current in both directions; c - amplitude-shift keying of alternating current; d - frequency-shift keying of alternating current; e - phase-shift keying of alternating current; f - phase-difference keying of a alternating current.

The last of the foregoing types of keying listed is the most resistant to noise.

There are practical difficulties involved in the realization of phase-shift keying because when there is a change in the phase of a linear signal by more than 90° the phase of the incoming pulses will change 180° (negative work).

This drawback can be eliminated by using relative phase-shift keying (fig. 16.21 f), in which the phase of the carrier changes as each 1 is transmitted, rather than when the character transmitted changes.

This method of signal formation provides for stable operation of the receiver when there are phase jumps in the signal on the channel.

When discrete signals are transmitted by conventional AM, FM, and PM methods the reception process reduces to comparing the signal parameters with the receiver parameters. In the case of AM, for example, the signal amplitude is compared with the receiver threshold; in the case of FM the signal frequency is compared with the center frequency of the detector; and in the case of PM the signal phase is compared with the local oscillator phase.

Relative phase-shift keying results in a comparison being made in the receiver between two adjacent transmissions, and not in a determination being made as to the absolute value of each transmission, so in this type of keying there can be no "reverse operation" (no negative work).

Spectrum Width and Telegraphy Speed

This section will discuss the basic relationships existing between signal spectrum width and telegraphy speed for the various types of keying.

The Amplitude-shift Keyed Signal Spectrum Width

When transmission is by direct current transmissions over wire lines, and the transmission of the third harmonic, inclusive, is taken into consideration

$$\Delta F = 3F_1 = 1.5B \text{ [hertz]}, \tag{16.36}$$

where

F_1 is the fundamental keying frequency, in hertz;

B is the telegraphy speed, in bauds.

In the case of amplitude-shift keying in radio communications

$$\Delta F_{AM} = 6F_1 = 3B \text{ [hertz]}. \tag{16.37}$$

The greatest signal spectrum width corresponds to the transmission of a series of alternating transmissions of 0 and 1 (fig. 16.22).

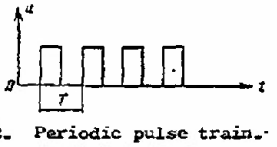


Figure 16.22. Periodic pulse train.- Vertical: Hertz.

The signal band can be reduced if some method that will lessen the effect of stretching the transmission front is used. This includes the "contracted contact method" and the integral reception method.

The first of these envisages strobing the center, that is, the least distorted, section of the transmission and deciding on which character in the incoming transmission to take as a result of the strobing.

In the second method the decision with respect to the character is made on the basis of the results obtained from integrating the transmission as a whole.

Quality reception can be had for a signal spectrum width

$$\Delta F_{AM} = (1.1 \text{ to } 1.2)\ B \text{ [hertz]}. \qquad (16.38)$$

In single-sideband systems for the transmission of binary signals, the value of the required channel band when one sideband is completely suppressed is

$$\Delta F_{AM\ ssb} = (0.5 \text{ to } 0.6)\ B \text{ [hertz]}. \qquad (16.39)$$

The frequency modulated signal spectrum width. In the case of frequency-shift keying the transmission of carriers f_1 and f_2 correspond to elements 1 and 0. The difference $\Delta f_{shift} = f_2 - f_1$ is called the spacing, or the shift in frequencies, and the magnitude $f_d = \Delta f_{shift}/2 = f_2 - f_1/2$ is called the frequency deviation. The ratio of the deviation to the fundamental keying frequency is the frequency-shift keying index, $m_f = f_d/F_1$.

The FM signal spectrum width, with the transmission of the third harmonic taken into consideration is

$$\Delta F_{FM} = 2\ (f_d + 3F_1) = 2f_d + 3B = \Delta f_{shift} + 3B \text{ [hertz]} \qquad (16.40)$$

The greater the frequency deviation the easier it is to distinguish the transmissions from each other upon reception. But this increases the band occupied by the signal, increasing the noise level.

It is recommended that the deviation be selected on the basis of

$$f_d > F_1/2k \text{ or } m_f > 1/2k,$$

where

$k = \tau_{dis}/\tau_o$ is the permissible time dominance, that is, the maximum ratio of length of the distorted section of the transmission, τ_{dis}, to the entire length of the transmission, τ_o.

m_f is selected small in systems for transmitting data over narrow band channels in order to increase the traffic handling capacity, and the maximum capacity in the case of narrow band FM is achieved when

$$\Delta F_{FM} = \Delta F_{AM} = (1.1 \text{ to } 1.2)\ B \text{ [hertz]}. \qquad (16.41)$$

Consideration must be given to the fact that resistance to noise will be low, however. The effective band in the case of narrow band FM can be found through the formula

$$\Delta F_{FM} = 2\Delta F_{AM} = (2.2 \text{ to } 2.4)\ B \text{ [hertz]}. \qquad (16.42)$$

FM radio lines have a deviation selected several times greater than is the case for narrow band FM, and the frequency band can be found through the formula

$$\Delta F = 2F_1\sqrt{\frac{200}{\pi} m_f + m_f^2} \text{ [hertz]}. \qquad (16.43)$$

The phase-shift keyed signal spectrum width. Modern data transmission systems using the phase-shift keyed method limit the channel pass band to that ensuring transmission of the first harmonic of the keying frequency, F_1. Given these conditions, the band covered by the signal is equal to

$$\Delta F_{PM} = 2F_1 = B \text{ [hertz]}. \qquad (16.44)$$

An even greater reduction in the band can be arrived at by using single-sideband transmission of phase-shift keyed signals.

16.4. The Multichannel Communication Principle

Multichannel communication is communication between several pairs of correspondents (senders and receivers) over a common line.

Figure 16.23 shows, in general form, a block schematic of multichannel communication. Signals from all n channels are summed and transmitted over the communication line. The receiving end must have a device for separating the signals corresponding to the different channels.

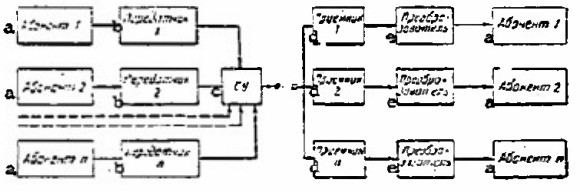


Figure 16.23. Block schematic of multichannel communication.
a - subscribers; b - transmitters; c - summer;
d - receivers; e - converter;

Separation of the signals in the different channels must clear the signals in the particular channel from the interference created by the signals in the other channels, in so far as this is possible, hence the shape of the signal used for multichannel communication is selected so the interference between channels will be minimal.

Frequency Separated Channels

The essence of the frequency method of separating channels is as follows. Since the real communication signal contains the overwhelming portion of its power within the limits of the width-limited frequency spectrum, the organization of multichannel communication for transmitting signals

assigns a definite section of the common band of frequencies used to each channel.

Thus, the transmitter for each sender should send signals over the communication line, the frequency spectrum of which will fit completely within the frequency band assigned to the particular channel.

There will be a set of voltages, or currents, at the receiving end of each channel on all frequencies forming the linear signal of multichannel communication. The receiver contains a frequency filter, the purpose of which is to separate out the voltage of the frequencies that reflect the message for the specific sender, and to suppress the voltage of the other frequencies. The frequency filter in each channel will only pass the frequency spectrum for its channel, and will not pass the frequencies on the other channels. The separation of signals by using frequency filters is called frequency separation.

Figure 16.24 shows the spectra of signals in a two-channel communication system with frequency separation.

The first channel carries the frequency band from f_1 to f'_1, and the second the frequency band from f_2 to f'_2; that is, the signals in the first channel have spectrum ΔF_1, while the signals in the second channel have spectrum ΔF_2.

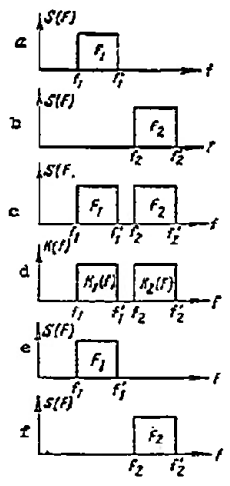


Figure 16.24. Spectra of signals in a two-channel communication system with frequency separation.
a - first channel signal frequency spectrum;
b - second channel signal frequency spectrum;
c - linear signal; d- filter frequency curves;
e - signal at the output of the first filter;
f - signal at the output of the second filter.

The condition of separability of signals in this case reduces to preventing the frequencies in the ΔF_1 spectrum from entering the ΔF_2, spectrum, and vice versa. In other words, spectra ΔF_1 and ΔF_2 cannot overlap.

Transference of the signal upward through the spectrum during transmission, and the reverse conversion upon reception, is carried out by subcarrier frequency oscillators and modulators. Oscillator frequency drift should not exceed ±2 hertz.

The attenuation of filters in the pass band cannot be greater than 0.3 neper, but outside the pass band should be 7.55 nepers.

Time Separated Channels

Based on V. A. Kotel'nikov's theorem, any continuous function with a limited spectrum can be represented by a train of discrete pulses. Pulses received at the point of reception can be completely restored to the transmitted function. The use of pulse methods for purposes of transmission (fig. 16.25) make the organization of multichannel communication with time separated channels feasible.

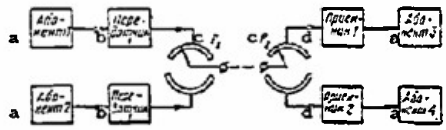


Figure 16.25. Block schematic of multichannel communication with time separated channels.
a - subscriber; b - transmitter; c - distributor; d - receiver.

In time separated channel systems the communication line utilizes a distributor, D_1, to permit signals from different senders (subscribers) to be transmitted in turn. Distributor D_2 is at the receiving end, and has the capacity to make a time selection, or, in other words, to separate the signals from the different channels. Each channel is thus assigned a definite section of the total time the line is in use.

Complete separation of the signals requires that distributors D_1 and D_2 rotate at identical speeds (in synchronism). Moreover, the distributors should connect to the line either the first pair of subscribers, or the second, or, in other words, they should operate in phase.

When time separation is used the signals belonging to a particular channel are transmitted during a time interval that is free of signals from other channels. The condition of separability of signals in the case of time separation reduces to preventing signals in the different channels from overlapping in time

In the case of the transmission of a telephone signal, the spectrum

of which is between 300 and 3400 hertz, the frequency at which the communication channel is cut in should equal 6.8 kHz, based on Kotel'nikov's theorem. In practice, the switching frequency is taken equal to 8 kHz. Electronic distributors are used at these switching frequencies.

Time separated channels are widely used in communication and remote control systems. The communication systems with time separation in use at the present time do not have more than 24 channels. This limitation is associated with the need to use extremely short pulses. In remote control systems the number of channels are many times greater because the signals can be limited to low frequency on the order of 10 hertz.

Combination Separation

Combination separation of signals is based on the difference in the combination of signals on the different channels. Let us suppose we have a two-channel system with a binary code with elements 0 and 1 in use on both channels. It will then be possible to have four different combinations of signals in both channels:

Communication Channel	Combination of Signals			
	№ 1	№ 2	№ 3	№ 4
Channel 1	0	1	0	1
Channel 2	0	0	1	1
Sum of signals . .	0	1	1	2

As will be seen from the table, if we take a signal equal to 1 we do not know the channel to which it belongs, but by referring to the tables we see that all four combinations differ from each other. Therefore, instead of a summed signal we must transmit the number of the combination because this number uniquely determines the signal in each channel. Thus, the task reduces to one of transmitting four numbers, and they can be transmitted in a variety of ways (by any code and modulation). This type of transmission makes the linear signal a reflection of a fixed combination of signals for the different channels.

Two-channel frequency telegraphy is a known example of combination separation. Here four different frequencies, f_1, f_2, f_3 and f_4 are used for the transmission. In the general case of an n-channel system with code base m, it will be required that a linear signal consisting of $N=m^n$ different combinations be transmitted. Each combination will correspond to a signal for a predetermined channel.

Two-channel frequency telegraphy increases the traffic handling capacity of the channel by a factor of two as compared with single-channel

operation. But this also requires an expansion in the channel pass band by a factor of approximately two, and this leads to some reduction in the noise resistance.

16.5. Transmission of Radar Information Over Communication Channels.

Transmission Over Broad Band Channels

Radar information can be transmitted over broad band communication channels by direct translation, or by television.

In the case of direct transmission of radar signals the information contained in each individual pulse reflected from the target can be transmitted over the communication channel. The frequency band needed for the undistorted transmission of a pulse of lenght τ_p can be found through the relationship

$$\Delta F_c \approx 1/\tau_p. \tag{16.45}$$

Transmission of pulses with length τ_p = 2 to 0.2 microsecond requires a communication channel frequency band of ΔF_c = 0.5 to 5 MHz, respectively.

Since other signals besides target pulses should be transmitted over the communication channel, the total band of the combination signal in broad band systems using direct transmission can reach 10 MHz and higher.

In the television method of transmitting a radar image the frequency spectrum band can be established by the highest transmitted frequency, and this can be computed through the formula

$$f_{max} = 1/2\ nf_f \tag{16.46}$$

where

n is the total number of television image resolution elements;

f_f is the frame frequency.

The number of television resolution elements in standards accepted for modern television definition is approximately equal to $n = (3 \text{ to } 4) \times 10^5$, and the frame frequency, is $f_f = \Delta F_c$ = 4 to 6 MHz.

The use of broad band methods of transmitting radar information requires special, broad band communication channels.

Transmission Over Narrow Band Channels

When a narrow band communication channel is used to transmit radar information there are various methods used to convert this information into a narrow band signal, most of which are based on the integration of the train of pulses received by the radar receiver during the time the target is illuminated. This train of pulses lights up a definite section of the indicator scope, the limits of which are fixed by the radar's resolution in range, δR and azimuth $\delta\theta$. Each light-struck section of the scope can be considered to be an image element subject to transmission over a communication channel, one per radar scan period, for example.

If the image of an entire PPI scope must be transmitted, the number of image elements can be computed through the formula

$$n = 360°/\varphi_{0.5P} \cdot 2R_{max}/c\tau_p, \qquad (16.47)$$

where

R_{max} is the maximum radar detection range;

$\varphi_{0.5P}$ is the width of the pattern in the horizontal plane at half power;

τ_p is the pulse length;

c is the speed of light.

If a radar is considered as a device establishing the presence, or absence, of a target in each element of the image, the total amount of information equals the number of image elements, $I = n$, and the required pass band will then be equal to

$$C = n/T_s = I/T_s, \qquad (16.48)$$

where

T_s is radar scan time, that is, the time required to transmit the information.

Shannon's formula yields the pass band

$$C = \Delta F_c \log_2 (1 + P_c/P_n). \qquad (16.49)$$

Equating the right hand sides of these last two equations to each other, we can find the required frequency band for the communication channel as

$$\Delta F_c = I/T_s \log_2 (1 + P_s/P_n)[\text{hertz}]. \qquad (16.50)$$

Thus, with the integration of the train of pulses reflected from the target during the time it is illuminated taken into consideration, the required width of the communication channel pass band can be reduced to a magnitude not exceeding several tens of kilohertz.

Devices that narrow the frequency band make it possible to use telephone communication channels with pass bands of 0.3 to 3.4 kHz to transmit information.

Another method of transmitting radar information over narrow band communication channels is one based on the transmission of target coordinates in the form of coded digital values.

Digital systems, or discrete systems, for transmitting binary information are readily mated to computers that establish target movement parameters. The communication channel in digital systems in only loaded when there are targets, that is, it will not transmit excessive information. It is this latter that permits the designing of channels for transmitting digital systems at slow transmission speeds, and, as a result, with a narrow pass band.

The communication channels in the United States semiautomatic antiaircraft defense systems known as SAGE are an example of the use of digital methods to transmit radar information. Figure 16.26 shows the interconnection between the individual elements of the guidance center of this system. The number of communication lines for one guidance center is on the order of 600. Data are transmitted at a rate of 1300 to 1600 bits per second over telephone channels with an error probability of 10^{-5}

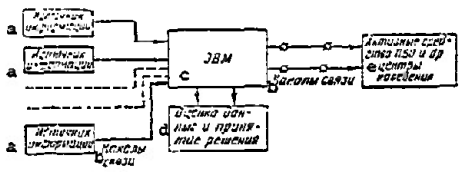


Figure 16.26. Simplified schematic diagram of the interconnection between the individual elements in a guidance center. a - information source; b - communication channels; c - electronic computer; d - data evaluation and decision; e - active AA defense equipment and other equipment in the guidance center.

16.6. Immunity of Communications to Interference

The immunity of communications to interference is the capacity of the communication system to resist the distrubing effect of interference. Interference is any effect on a receiver (other than the effect of the useful signal) that causes the output device to function.

The quality of a discrete communication channel is characterized by the authenticity factor, expressed in terms of the ratio of the number of correctly received characters, A, to the total number of characters transmitted, N:

$$K_a = A/N. \tag{16.51}$$

Immunity to interference depends on the type of modulation used, the coding method, the receiver circuitry, and many other factors.

Interference with radio reception takes many forms. It can take the form of stray electromagnetic fields acting on the antenna at the reception site, distortion in transmission caused by unstable conditions in the propagation of radio waves, thermal noises in the receiver, and other causes.

So-called fluctuating interference of the "white noise" type, consisting of individual, very short pulses (lengths on the order of 10^{-12} second) with amplitudes that changes at random with time, has a very special place among all of the possible types of interference. An oscillogram of white noise is shown as Figure 16.27.

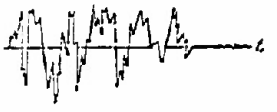


Figure 16.27. Oscillogram of interference of the "white noise" type.

White noise has a uniform power spectrum over a very broad (practically infinite) frequency band.

White noise results from the thermal movement of elementary particles of matter. Since this movement is always in evidence, practically speaking, white noise is, in principle, a source of interference that is impossible to eliminate. The special part played by white noise can be expalined too by the fact that it is the principal source of interference that bears on receiver sensitivity. It is for this reason that in general information theory the effect on the transmission of messages is considered to be due primarily to fluctuating interference of the white noise type.

At the basis of all methods used to imporve immunity to interference is the principle of increasing the excessiveness in the message transmitted, or, and this is the same thing, the principle of increasing the signal volume

$$V_s = \tau_s \Delta F_s \log_2 P_s/P_n.$$

As will be seen from this formula, it is possible to increase signal volume by increasing the length of the signal (the time of transmission), the signal frequency band, and the ratio of the average signal to noise powers.

The improvement in immunity to interference through the frequency band can be achieved by using broad band types of modulation (frequency, and all types of pulse), and by increasing the signal/noise ratio by reducing the level of noise and increasing transmitter power.

Moreover, an improvement in immunity to noise can be arrived at by using noise immune codes and special methods of receiving weak signals, such as the filtration of periodic signals, accumulators, narrow band reception, broad band - clipper - narrow band system, and others.

The method that uses the filtration of a period signal is based on the difference in the power density spectra (fig. 16.28) of the useful periodic signal and the white noise type of interference.

As will be seen from the figure, signal separation requires that the mixed signal and noise be fed into a narrow band filter with a frequency curve (fig. 16.28 d). The narrower the filter pass band, the more effective

will be the separation of the signal from the noise.

If the filter pass band is ΔF_f, and the noise power in the unit element of the band is G_p, the power of the noise at the filter output will equal

$$P_n = G_p \Delta F_f. \tag{16.52}$$

The signal/noise ratio at the filter output will equal

$$(P_s/P_n)_{out} = P_s/G_p \Delta F_f. \tag{16.53}$$

The filter pass band, ΔF_f is selected in practice on the basis of $\Delta F_{f\tau} = 1$, where τ is signal length.

Then

$$(P_s/P_n)_{out} = P_s/G_p\ \tau.$$

This expression indicates that the ratio of the average signal power to the average noise power at the filter output increases in proportion to signal length. Consequently, immunity to noise increases with increase in signal transmission time.

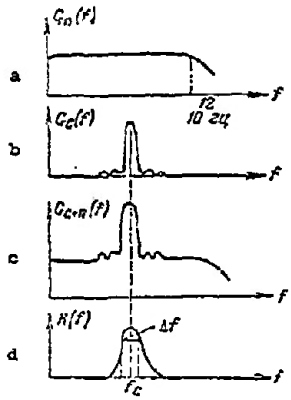


Figure 16.28. Power spectra (a - noise; b - signal; c - sum of signal and noise) and the filter frequency curve (d).

16.7. Error Correcting Codes

Error correcting codes are codes used to detect and correct errors in incoming code combinations.

Error correcting (anti-interference) codes improve the authenticity of information reception because of the use of excessiveness.

The essence of error correcting codes is contained in the fact that of the total number of code combination that can be made up from an n-element code, $N = 2^n$, for the transmission of information, there are only those N_p

combinations in which one combination would not exceed the other, given distortion of k elements.

The efficiency and effectiveness of codes with error detection can be assessed by the excessiveness factor (K_{ex}) and the error detection factor (K_{ed}).

The excessiveness factor is

$$K_{ex} = 1 - \log_2 N_p / \log_2 N, \qquad (16.54)$$

where

$N = 2^n$ is the total number of combinations that can be made up of an n-element code;

N_p is the number of combinations used.

The error detection factor is

$$K_{ed} = L/L + M, \qquad (16.55)$$

where

L is the total number of distorted combinations, errors in which can be detected;

M is the total number of distorted combinations, errors in which cannot be detected.

Error detecting codes. In constructing a single-error detecting code it is necessary to use only those combinations for which the minimum number of elements, d, by which one coded combination differs from another equals two (d = 2).

If the total number of possible code combinations is $N = 2^n$, the number of combinations that will differ from each other by two positions will equal $N_p = 2^{n-1}$.

Let there be a three-element code, $N = 2^3 = 8$. Let us select of these eight possible combinations

000, 001, 010, 100, 011, 101, 110, 111

those differing from each other by at least two elements

000, 011, 101, 110.

Now, if it is stipulated that the four combinations selected are permitted, but the others

001, 010, 100, 111

are forbidden, any single error in the permitted combinations will convert them into forbidden combinations. Incorrect recording of a character will not occur.

As we see, this code is a code with an even number of ones. Note that it can, however, be formed by adding one element (1 or 0) to the combinations of an n-element code such that the number of all the ones in the new

code [(n + 1)-digit number] will be even.

For example, we have the code combinations of a uniform five-element code

10000, 00110, 01101, 01010.

If one more character (1 or 0) is added to each combination so that the number of ones in the code combination formed is even, we obtain new six-element code combinations

100001, 001100, 011011, 010100.

A code such as this permits the detection of an odd number of errors in incoming elements, that is, the appearance, or the disappearance, of an odd number of ones.

Codes with a constant number of ones in the combinations are used to detect single, double, triple, etc., errors. A typical example is the seven-element code with a ratio of ones and zeros equal to 3:4 (each code combination contains three ones and four zeros).

The immunity to interference of codes with a constant number of ones and zeros is much higher than is that of codes with one element added.

The repeating code has the highest immunity to interference. The basis of the construction of this code is the use of methods for repeating the original combination in a direct, or in an inverted form, depending on whether the number of ones is even or odd.

Error detecting and correcting codes. What follows from coding theory is that use of combinations in which $d = 3$ will result in the construction of a single-error correcting code, or a double-error detecting code.

If the total number of code combinations is $N = 2^n$, the number of combinations, N_p, differing from each other by three positions ($d = 3$), will equal $N_p = 2^{n-2}$. Suppose we have a three-element code. Of the eight combinations in the three-element code, only two ($N_p = 2$) for which $d = 3$ can be selected. Let these combinations be 010 and 101.

Now let us suppose that during the transmission over the channel the first element in the 010 combination has been distorted and is received as 110. Comparing the combination received element by element with each of those use, we find that it differs by but one element from the first combination, and by two from the second. This will then lead to the conclusion that 010 was the combination transmitted and that the first element was the one distorted, and so must be changed to the reverse element.

The detection and correction method considered is readily realized when $N_p \leq 4$.

Other methods of constructing error detecting and correcting codes are used in the case of large N_p ($N_p = 32$, $N_p = 64$, etc.). One example is the

chain code which corrects statistically independent as well as group errors.

The principle behind the construction of the chain code is that there is a check symbol between each two informational symbols with the check symbol formed by checking the evenness of the two information transmissions. The information symbols are separated from the check symbols at the receiving end. The control symbols are formed from the information symbols received using the same rule as that applied at the transmitting end, and these control symbols are then compared with the check symbols received. If the control symbol does not match its corresponding check symbol, an error is indicated and is corrected.

Special systems can also be used to improve the immunity of communications to interference. These include systems with automatic interrogation and repetition of a signal in distorted code combinations, systems for transmission with information feedback, and others.

Two communication channels are needed for the operation of the systems names. Information is transmitted over one of them, while either an interrogation signal (in the first system) or a feedback signal (in the second system) is transmitted over the other. The decision as to the correctness of reception is made at the receiver in the first case, and at the transmitter in the second.

Thus, depending on the method used for the noise immune code, all systems transmitting binary information can be divided up into two groups; systems without feedback, in which an improvement in the authenticity of the transmission is achieved by using an error-correcting code, and feedback systems, the authenticity of transmission in which can be improved by repeating the signal in case a distorted combination is detected.

16.8. The Future of Information Transmission Systems

The foreign press has described the fundamental principles of two future information transmission systems; broad band radio communications, and synthetic telephony. The former has better secrecy and interference immunity. It provides authentic reception of signals, the level of which is considerably below the noise level. The latter provides for a significant reduction in excessiveness when transmitting speech signals.

The Broad Band Radio Communication Principle

The principle of assigning individual correspondents to predetermined narrow sections of the frequency band is the basis for present day radio communications. Improvement in immunity to interference is arrived at by transmitting narrow band signals covering a minimum band in the frequency spectrum for the specified transmission speed.

There have been recent reports that a significant improvement in immunity to interference and in secrecy can be obtained by the use of broad band signals. In these broad band communication systems each transmission (1 or 0) corresponds to the simultaneous radiation of a whole series of components, located in a predetermined manner along the frequency axis, the energies of which are summed coherently upon reception. At the same time, broad band signals from several stations that overlap each other can be transmitted simulatneously within the limits of the band, ΔF, used.

Figure 16.29 is the block schematic of one channel of a very simple broad band system for transmitting binary signals. The operating principle reduces to the following. Broad band signals $f_o(t)$ and $f_1(t)$ are formed on the transmitting sides of generators 1 and 2. These signals can be two different realizations of fluctuating noise, limited by the channel band ΔF.

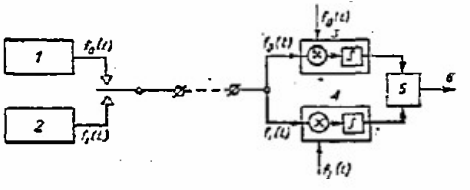


Figure 16.29. Block schematic of a broad band system for transmitting binary signals

Transmissions $f_o(t)$ and $f_1(t)$ of length t_s each are put on the radio link in accordance with the message transmitted.

The incoming transmissions are fed into the inputs of two correlators (3 and 4), each of which consists of a multiplexer and a integrator. A local reference signal, which is a copy of $f_o(t)$, acts at one of the correlators in synchronism with the incoming signal, and a reference signal which is a copy of $f_1(t)$ acts at the other.

The output voltages from both correlators, which are in proportion to the degree of correlation between the incoming signals and the reference voltages, are fed into the comparison stage, 5. The polarity of the output voltage from this stage 6, depends on which of the correlators yielded the greater degree of correlation. This latter is determined by the type of incoming transmission (1 or 0).

Multichannel communications can be provided by one link by using the same band. Signal separation can be provided by the selection of the different realizations of broad band signals, designated as 1 or 0 in each channel.

The Synthetic Telephony Principle

Human speech is made up of sonorous sounds formed by the vocal chords, and noise sounds formed by the passage of air in the gaps between the teeth, the tongue, and the palate. Sonorous sounds (all the vowels and the vocal consenants L, M, N) have a discrete spectrum and contain a fundamental tone (individual for every person) and a great many harmonics, or overtones. Man's fundamental voice tone ranges from 100 to 250 hertz, woman's from 200 to 450 hertz. Noise sounds have a continuous spectrum. Noise sounds include Sh, S, P, Shch, and there are different combination sounds (R, Kh, Z).

The significant information contained in speech can be transmitted by changing the spectrum envelope. The frequency of the fundamental tone and its change with time contains the information on the individual characteristics of the speaker's voice. Thus, rather than sending a speech signal reflecting information concerning the spectrum envelope, the frequency of the fundamental tone, and the nature of the sounds (noise or tonal). Speech can be synthesized at the receiver end by correcting the oscillations from the local generators of audio frequencies with these signals. Since the change in the spectrum envelope and in the fundamental tone takes place very slowly during the conversational process (at a frequency not in excess of 20 hertz), transmission of synthesized speech can be accomplished in a very narrow frequency band.

Devices that make reality of the synthetic telephone principle (fig. 16.30) are called vocoders ("voice coders").

Band pass semi-vocoders, in which the frequency of the fundamental tone transmits the frequency band of unconverted speech, have received the broadest practical application.

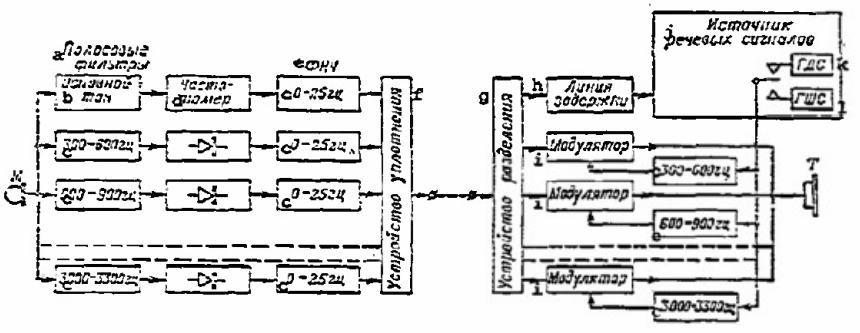


Figure 16.30. Block schematic of synthetic telephony.
a - band pass filters; b - fundamental tone; c - hertz; d - frequency meter; e - low-pass filters; f - multiplexer; g - separator; h - delay line; i - modulator; j - speech signal source; k - discrete signal generator; l - noise signal generator

A bandpass filter at the transmitting end separates and transmits the components of the low-frequency portion of the spectrum containing the fundamental tones, or their second harmonics (the fundamental band) through the multiplexer to the communication channel. The rest of the speech signal is divided up into a series of frequency bands by the band pass filters (vocoder channels). The output signals from the filter are detected, smoothed by the low-pass filters, which have a band pass of approximately 25 hertz, and fed into the communication channel through the multiplexer. The voltages across the low-pass filter outputs are proportional to the averaged smoothed levels of the speech signals in the corresponding frequency bands.

At the receiving end the separator separates the unconverted speech signals from the vocoder channel signals. The former are fed into the speech signal source, which consists of a discrete signal generator and a noise signal generator. Broad band tonal signals, containing the many harmonics of the fundamental tone, are formed as a result of the nonlinear conversion of the fundamental band signals in the discrete signal generator.

The unconverted speech signals fulfill the role of the tone/noise signal. When there is a sharp reduction in their level the noise generator is automatically cut in. The broad band signal is fed from the output of the speech signal source through the band pass filters to the modulators which are controlled by the signals from the corresponding vocoder channels and reflect the cahnge in intensity of the speech signal in the particular voice. The signals obtained as a result of the modulation are similar to the signals at the outputs of the band pass filters of the corresponding vocoder channels on the transmitter side. They are fed through the band pass filters to the output where they are summed with the unconverted speech signals.

A delay line is inserted to provide synchronism between the summed signals in the circuit carrying the unconverted speech signals.

Chapter XVII

Radar Systems and Antimissile Defense Facilities

17.1. Special Features of Ballistic Missiles and Their Principal Characteristics

Missiles today are the main weapon used by modern armed forces. The battle properties of missiles are sharply different from all the other types of weapons that have ever existed, and this has considerably increased the degree of surprise of an attack and has created great difficulty in organizing the fight against missiles.

We will consider, in brief, the principal weapon in the huge arsenal of missiles that will be used in an air-space attack, the ballistic missile.

Missiles whose trajectories (other than the active section) are those of a freely tossed body, are known as ballistic missiles. On the basis of this definition all unguided and guided missiles for a wide range of missions and with various ranges can be included among the ballistic missiles. This would be inaccurate, however, because the term "ballistic missile" has been accepted as applying only to guided missiles with a separate head and a range on the order of hundreds and thousands of kilometers.

There are in existence today ballistic missiles with a range of from 160 km to 16,000 km and more. The principal tactical and technical data on optimal trajectories of some foreign missiles are listed in Table 17.1.

The optimal trajectory of a ballistic missile is the elliptical trajectory for movement along which the missile requires a minimum expenditure of energy.

The ballistic missile, as a means of making an air-space attack, has many basic differences from modern aircraft, not only with respect to the nature of the flight itself, but with respect to design formulations. Missile flight takes place for the most part (with the exception of the initial active section) along the elliptical ballistic trajectory. The curvature of the ellipse, and the position of its plane in space, are fixed by the magnitude and direction of the missile's speed vector at the end of the active section of the trajectory. The predominant part of the trajectory of the ballistic missile occurs at tremendous altitudes, in the discharged atmosphere, where the missile head has very high speed, so the ballistic missile is less vulnerable than any modern aircraft.

The heads of ballistic missiles have comparatively small geometric dimensions and, as a result, small reflecting surfaces. This makes it very difficult to detect them with radar, or by any other system. At the same time the strength of the heads is such that it is also difficult to neutralize

Table 17.1.

Principal tactical and technical data on optimal trajectories of ballistic missiles of certain of the foreign countries.

Data Items	Types of ballistic missiles						
	Blue Water	Veronica	Pershing	Polaris A1, A2, A3	Atlas D, E, F	Titan I, II, III	Minuteman A, B, C
Maximum range, km	160	240	740*	2000; 2800; 4000*	10000; 15000; 16000	10000; 15000; 20000	9000; 10000; 11200
Maximum height of trajectory, km	-	80	320	600; 700; 750	1250; 1300; 1320	1280; 1300; 1400	1000; 1120; 1307
Maximum flight speed, km/hour	2520	6480	10800	16200; 18000; 18000	25200; 25920; 28000	25200; 25920; 28000	25200; 25560 25920
Length of missile system, meters	7.6	-	12.2	8.5; 8.8; 9.3	24; 25.1; 25.1	27.6; 30; 32	16.7; 17; 18
Reflecting surface of head, m^2	-	0.13	0.15	0.2; 0.2; 0.1	0.5; 0.5; 0.5	0.5; 0.5; 0.5	0.1; 0.1; 0.1
Takeoff weight of missile system, tons	-	-	4.5	12.6; 12.7; 12.8	115; 116; 118	110; 113; 130	29.5; 30.6; 33

* Work is in progress to increase the range by a factor of 1.5 to 2.

Table 17.1. (continued)

Principal tactical and technical data on optimal trajectories of ballistic missiles of certain of the foreign countries.

Data Items	Types of ballistic missiles						
	Blue Water	Veronica	Pershing	Polaris A1, A2, A3	Atlas D, E, F	Titan I, II, III	Minuteman A, B, C
Weight of warhead, kg	-	-	-	450; 500; 500	2500; 2000; 1300	2720	400; 400; 800
TNT equivalent, 1000 tons	10	50	1000	500; 1000; 3000	7000; 5000; 4000	4000; 5000; 10000	1000; 1500; 4000
Guidance system	Inertial	Radio Command	Inertial	Inertial and homing	Inertial	Inertial	Inertial
Type engine and number of stages	solid fuel one	liquid fuel one	solid fuel two	solid fuel two	liquid fuel two	liquid fuel two	solid fuel three
Missile launch	mobile launcher	mobile launcher	mobile launcher	submarine	open ground launcher	underground launcher	underground launcher
Part of the arsenal	Army, Great Britian	Army France	Army USA	Navy USA	Air Force USA	Air Force USA	Air Force USA

them, or to destroy them in flight.

Ballistic missiles can be used at ant time of the year, at any time of the day, in any weather, and can be launched from the ground, from underground, from submarines, for surface launchers, and from aircraft. Missile strikes can be expected from practically and direction. This makes the fight against them much more difficult and complicated.

The use of tremendously powerful atomic and thermonouclear charges on ballistic missiles requires that the defending side destroy all attacking missiles, and not just some of them. At the same time destruction should take place as far away from the targest being defended as possible.

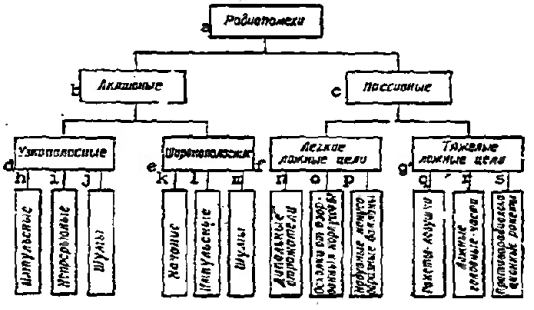


Figure 17.1. Principal types of radar jamming for antimissile defense. a - electronic jamming; b - active; c - passive; d - narrow band; e - broad band; f - small false targets; g - large false targets; h - pulse; i - continuous; j - noise; k - switching; l - pulse; m - noise; n - dipole reflectors; o - fragments from the exploded casings of ballistic missiles; p - inflated conical balloons; q - missile traps; r - false head sections; s - antiradar missiles.

The probability of ballistic missiles getting through to targets defended by air and antimissile defenses is greater than that attributable to military aircraft, the more so because missile trajectories can differ greatly once they enter the dense layers of the atmosphere, where such trajectories can be disturbed, or smooth. The heads of ballistic missiles can change direction radically in flight by using a small on-board motor, or they can maneuver in the dense layers of the atmosphere by using aerodynamic forces. In this way the heads of ballistic missiles can be aimed at a new target to destroy and negate any results that may have been obtained from tracking the head and any calculations already made in an effort to fix the anticipated point at which the warheads of attacking missiles should be intercepted. American specialists are presently intensifying their efforts to build maneuvering head sections for their ballistic missiles "Minuteman-2" and for

the ballistic missile "Poseidon" they have developed.

The heads of ballistic missiles can be fitted with special screens for protection against radiation, and their reflecting surface can be artifically reduced by ablation coatings, thus greatly reducing their vulnerability and the range at which they can be detected in particular. Nor can the fact that the heads of ballistic missiles can be fitted with gear to create a variety of types of interference (fig. 17.1), passive and active, be ignored. False targets, or special missiles designed to destroy antiaircraft and antimissile defense radar systems, can be present in the ballistic missile flight area. Missiles such as these, or the false targets, can be launched beforehand, or can separate from the head of the ballistic missile prior to the entry into the dense layers of the atmosphere.

The far from complete technical information thus far cited for the ballistic missile leads to the conclusion that ballistic missile detection along the trajectory, recognition of the true target, and finally, destruction or neutralization of the war head is an extremely complex mission, one requiring the use of tremendous numbers of men and equipment to carry out.

17.2. Antimissile Defense and its Mission

In accordance with the classification adopted abroad, the air space can be broken down into regions by altitude, as shown in Figure 17.2, and all attack weapons can be broken down into aerodynamic, ballistic, and space. On this basis then, the defense against the attack weapons can be called antiaircraft defense, antimissile defense, and antispace defense.

Antimissile and antispace defenses, as they are understood today, are complicated combat operation installations designed to provide timely warning to the military command of the initiation of an enemy nuclear missile strike and to break up nuclear strikes against important targets. The antimissile defense and antispace defense systems must have detection equipment capable of detecting ballistic missiles, and of tracking them over the entire trajectory, if this mission is to be carried out.

The process involved in coping with the ballistic missile in flight, despite all of its extreme complexity, reduces to solving the following basic problems, if one were to formulate them in brief: reliable detection of the ballistic missile; accurate determination of the trajectory; recognition of the target; interception and neutralization or destruction at acceptable distances from the targets being defended. All these problems can be resolved, but only if the antimissile defense system is in constant combat readiness, is equipped with reliable equipment for early detection, target recognition, and high-speed electronic computers for different purposes, so that the entire defense system can be fully automated.

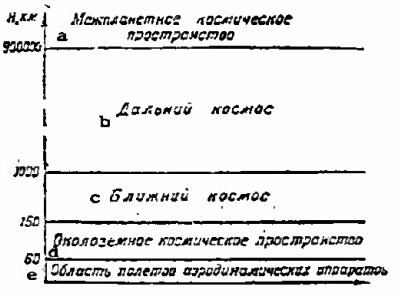


Figure 17.2. Schematic diagram of the classification of the air space by altitude.
a - region of flights of aerodynamic vehicles;
b - near-ground space; c - near space; d - far space; e - interplanetary, or outer space.

17.3. Types of Antimissile Defense

The present opinion of foreign military specialists is that it is possible to build two types of antimissile defenses, known as the duel system, and the screening system, with the duel type system further broken down on the principle of survivability of the targets being defended into an antimissile defense system for sheltered positions, and an antimissile defense system for large populated areas.

The Duel Type Antimissile Defense

The conventional duel type antimissile defense system envisages the use of interceptor missiles (antimissiles) launched from ground, or underground launchers to meet the target in accordance with a previously computed guidance program. The antimissile defense system is located around the targets being defended and is designed to destroy the heads of ballistic missiles on the final leg of their trajectories at a distance such that there will be no danger to the targets, or to the system itself, as a result of the thermonuclear explosions of the warheads carried by the ballistic missiles, and by the antimissiles.

The country's missile defense, in the case of use of the duel system, obviously will be the totality of the aggregate of defense complexes protecting all the most important administrative and political, industrial, and military targets.

The antimissile defense system should be able to perform the following basic missions: timely detection of attacking ballistic missiles such that there is enough time to prepare the active protection facilities; reliable recognition of the real targets and highly accurate determination of their trajectories so the guidance for, and the moment to launch, the antimissile

can be programmed. These missions are the passive part of the antimissile defense. The active part of antimissile defense involves solutions to the following problems: interception by the antimissile of recognized targets, and their destruction, or neutralization. Time, speed, and target maneuverability on the final leg of the trajectory are all factors of great importance in arriving at a positive solution to all the missions assigned the duel type antimissile defense system.

The Screening System for Antimissile Defense

The screening system for antimissile defense can be a wide barrier, or screen, in the discharged layers of the atmosphere, capable of destroying all missiles flying through it. This screen should be set up beforehand and maintained for extended periods of time. This antimissile defense system has many advantages.

Whereas it is desirable to orient the duel type system in the direction from which the attack can be expected beforehand, knowledge of the attack, and the nature of the movement of each of the enemy missiles, are not requirements when the screening system is used. The requirements in connection with long-range detection too can be greatly reduced when this antimissile defense system is used.

Just how is the battle-capable screen set up? That still is the problem. Foreign specialist now believe that this type of screen can be set up by using maneuvering satellites which should home on the ballistic missile, or satellites carrying antimissiles, or any other weapons capable of destroying or neutralizing the heads of ballistic missiles. A satellite screen can obviously provide antimissile defense on a local, as well as on a state, or even continental scale. According to rough estimates, this would require the use of several thousand satellites spinning in different orbits computed so that enemy ballistic misslies would come within range of possible destruction by just one satellite.

Screening systems for antimissile defense capable of destroying ballistic missiles can literally be set up over all sections of ballistic missile trajectories, and this is the great advantage of this particular system.

17.4. Detection of Ballistic Missiles in Flight

The flight of a ballistic missile can be detected by the energy radiated by the flying missile, by the energy reflected, and by the disturbed environment as it is traversed by the flying ballistic missile, or its head. Figure 17.3 is a diagram that is a visual explanation of the various phenomena occurring during the flight of a ballistic missile, and which can be used for long-range detection of the missile in flight.

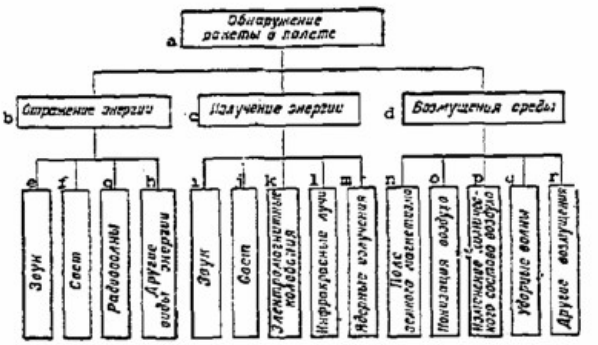


Figure 17.3. Schematic diagram of possible ways in which to solve the problem of detecting ballistic missiles in flight. a - detection of a missile in flight; b - reflection of energy; c - radiation of energy; d - disturbance of the environment; e - sound; f - light; g - radio waves; h - other types of energy; i - sound; j - light; k - electromagnetic oscillations; l - infrared rays; m - nuclear radiations; n - earth's magnetic field; o - ionization of the air; p - change in the chemical composition of the air; q - shock waves; r - other disturbances.

Foreign specialists are now giving intensive study to every possible method and way to detect ballistic missiles, but as yet most of the signs that could be used to detect a missile have been so little studied that as a practical matter there is no possible way they can be used in an anti-missile defense system.

Press reports say that work is in progress on experimental models of equipment for detecting ballistic missiles in flight that are based on the use of such phenomena as the distortion of the earth's magnetic field, change in the atmosphere's parameters, and even the propagation of sound waves, something that is generally thought to be unsuited for this purpose. These methods have been named, magnetic, barographic, and acoustic, respectively.

However, methods based on the use of radio waves, light and infrared radiation have found practical application for detecting ballistic missiles in flight in the antimissile defense system. The latter two of these methods are considered to be secondary methods because their effectiveness depends on weather conditions, and this is a significant drawback as compared with

the generally known, and well studied radar method, which has long been used successfully in air defense systems.

Thus, the current stage of development of antimissile defense uses long-range, high resolution radar as the primary means of detecting ballistic missiles in flight.

17.5. Requirements Imposed On Antimissile Defense Radar

The characteristics that establish the capabilities of combat utilization of radar in an antimissile defense system depend primarily on the specific purpose for which the radar is intended within the overall defense complex. All radars in a duel type antimissile defense system can be broken down by purpose for which intended into warning, early warning, tracking, recognition, and guidance. This breakdown is explained by the principal differences in the characteristics of the radars used to carry out concrete antimissile defense missions, something that will be discussed in detail during consideration of one of the versions of an already developed American antiballistic defense system.

Particular attention in designing the radar for use in the antimissile defense system is devoted to providing the answers to such questions as how to increase the range, improve the accuracy in determining coordinates, increase resolution and selection of the ballistic missile head against a background of false targets. The following methods are used in arriving at successful answers to these problems in the course of developing radars for use in an antiballistic defense system: increasing the sizes of the receiving and transmitting antennas; increasing the pulse and average powers; reducing the noise in the receivers; modernizing the methods used to gather and process information.

The radars in the antimissile defense system use all of the latest acheivements of radar engineering, such as molecular amplifiers, optimal filters, pulse compression methods, the Doppler method of measuring speed by the use of the correlation technique, predetection recording of information on all target parameters, phased antenna arrays, and the like.

Just what should the maximum radius of the radar detection and tracking range be for radar used in the antiballistic defense system?

As has already been reviewed (table 17.1) the maximum altitude of the optimal trajectory for the long-range ballistic missile is approximately 1300 to 1400 km. It is a known fact that conventional early warning radars operate on frequencies in the quasi-optical portion of the spectrum (about 100 MHz and higher), and that there must be line of sight between the target and the radar in order for the radar to detect any target. The formula at (2.1) can be used to find the line of sight distance, with refraction taken into consideration.

Computations reveal that at altitudes of 1300 to 1400 km the head of the ballistic missile can be detected at a range of approximately 4600 to 5000 km from the radar. This distance will be less for all points on the descending branch of the trajectory. Hence, the range of a radar used in an antimissile defense system should be at least 4800 to 5000 km.

The present day level of electronic development is such that it is possible to build radars that can detect small targets at ranges out to 5000 km, and in fact there are radars in use providing early warning of ballistic missiles with ranges of from 4500 to 5000 km.

How much warning time can an antimissile defense system expect in terms of the range at which a ballistic missile is detected?

The answer can be obtained by analyzing the ratio between warning time and ballistic missile detection range (table 17.2). As will be seen from the table, in order for an antimissile defense system to have 15 minutes warning, which is, in the opinion of foreign specialist, the minimum time needed to prepare the active defense, and to carry out certain of the measures in use to shelter the population, the system utilized by the antimissile defense for detection purposes should positively detect a target with a small radar cross section flying at 6.8 to 7.9 km/second at ranges of from 4500 to 5000 km. This is one of the principal requirements imposed on target detection and tracking radars in the antimissile defense system.

The maximum speeds for the automatic tracking of a target in range are $(dR/dt)_{max\ auto}$, and in angular coordinates $(d\beta/dt)_{max\ auto}$; $(d\varepsilon/dt)_{max\ auto}$ for the radar in the antimissile defense system should be large enough so target tracking will be stable as the target moves at entry speed on any course within radar range.

Table 17.2

Relationship between warning time and range at which the ballistic missile is detected

Warning time, minutes \ Range at which ballistic missile fired, km	50	100	500	1000	2000	5000	8000	10 000	15 000
0,5	35	45	90	100	120	170	200	270	320
1	—	50	160	190	220	320	380	500	610
2	—	—	300	320	400	600	700	900	940
5	—	—	—	800	1000	1500	1700	1800	2200
10	—	—	—	—	—	3100	3400	3600	4300
15	—	—	—	—	—	4500	4600	4800	5000
20	—	—	—	—	—	—	6200	6600	7200

Note: The computations in Table 17.2 are made for a ballistic missile with optimal trajectory.

The requirements with respect to resolution and accuracy in establishing coordinates are high. Antimissile defense radars usually find target azimuth, elevation, range, and speed.

The operating limits of a radar, in terms of the angular data, that is, the scan sector in azimuth and elevation, depend on the purpose for which the radar is intended, and on the principal characteristics of the ballistic missile. Radars in the antimissile defense system are sector, or circular scan types, and the scan period should be short enough to preclude missing a target, yet provide the most accurate attainable information on the target's position in space.

Any one of these characteristics can become decisive in assessing the combat features and capabilities of a radar, depending on the purpose for which it is intended in the antimissile defense system.

In view of the extreme complexity of the radars used in antimissile defense systems, and in view of the high order of the requirements imposed with respect to their tactical and technical characteristics, limitations as to size and weight are not considered significant. All of the radars in antimissile defense systems are usually installed in fixed, open or covered, structures.

17.6. The Role and the Place of the Radar System in the Overall Complex of Antimissile Defense Facilities

The antimissile defense radar system is the totality of radio engineering facilities dispersed over the terrain within the framework of a definite concept, and designed to provide radar support for the antimissile defense's active facilities.

The components in the antimissile defense radar system are: the radar field for the detection and recognition of air and space targets; the automatic system for collecting and processing radar information; the system for reflecting the air and space situation, and the system for controlling the operation of all the radars in the antimissile defense system.

All of these components are closely linked, so a breakdown in one can result in failure of the mission assigned the antimissile defense radar system.

The radar system is one of the important component parts of the antimissile defense complex. It is this system that must solve the following main problems: timely detection of air and space targets; establishment of the trajectories of ballistic missiles attacking, or simply flying over the territory of the country in any azimuthal direction, as well as flight speed and place of fall. All should be solved successfully in a very short period of time so the antimissile defense commander has the maximum time on his side to take active defense measures.

17.7. A Working American Antimissile Defense System

The United States mobilized a large army of scientists and specialists, as well as the latest achievements of science and engineering to solve the antimissile defense problem. Moving into this area are newer and newer projects. It is true that some have quickly been consigned to the archives, but there are those that took many years to bring about.

The American antimissile defense system developed over the last ten years includes:

a passive defense system - a ballistic missile early warning system of the BMEWS type;

an active defense system - a system for recognizing and destroying ballistic missiles of the "Nike Zeus" and "Nike X" types.

The entire process of detection, recognition, and destruction of a ballistic missile in the workable American antimissile defense system can be seen from the diagram shown in Figure 17.4.

17.8. Ballistic Missile Early Warning System

Utilizing its geographic potential, the United States has built radar stations for detecting ballistic missiles at considerable distances from its borders. The American BMEWS system has four ballistic missile early warning stations (the first of which is located in Greenland, the second in Alaska, the third in northern England, and the fourth on United States territory, in Florida), an operations center for control, and communication systems.

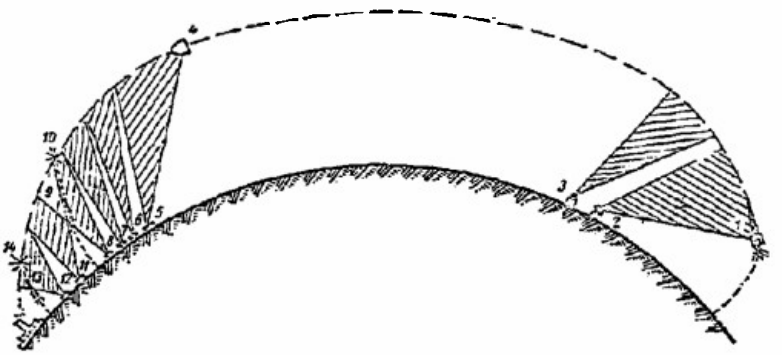


Figure 17.4. Schematic diagram of an operable American antimissile defense system.
1 - ballistic missile in flight; 2 - early warning radar in the BMEWS system; 3 - target tracking radar in the BMEWS system; 4 - ballistic missile head; 5 - detection radar in the active antimissile defense system of the "Nike Zeus" type; 6 - target recognition radar in the

"Nike Zeus" system: 7 - target tracking radar in the "Nike Zeus" system; 8 - "Nike Zeus" antimissile guidance radar; 9 - "Nike Zeus" antimissile; 10 - point at which antimissile and ballistic missile head meet; 11 - multipurpose radar in the "Nike X" antimissile defense system; 12 - launch radar in the "Nike X" system; 13 - "Sprint" antimissile in flight; 14 - point at which "Sprint" antimissile and ballistic missile head meet in the dense layers of the atmosphere.

The sectors covered by the first three of the BMEWS radar stations are oriented in a northerly direction, the fourth in a southerly direction. The BMEWS should detect long and medium range ballistic missiles in the event of an attack on United States territory from the north, or from the south, at the maximum 15 to 17 minutes prior to the missiles reaching their targets. During this time the combined antimissile defense command in the United States should take necessary measures for defensive and retaliatory actions.

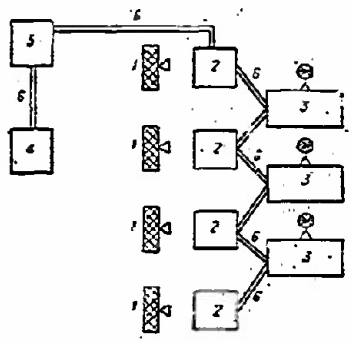


Figure 17.5. Schematic diagram of one of the first BMEWS stations.
1 - AN/FPS-50 radar antenna; 2 - building housing the AN/FPS-50 radar; 3 - AN/FPS-49 radar; 4 - IBM-7090 computer (two); 5 - buildings housing shops and personnel; 6 - concealed approaches.

The structural arrangement of the first two BMEWS radar stations differ sharply from the third, and particularly the fourth of the stations. The first and second stations have two radar systems, the third and fourth but one. The first station has four AN/FPS-50 and three AN/FPS-49 radars, while the second station has three AN/FPS-50 and two AN/FPS-49 radars. The third station only has three AN/FPS-49 radars, while the fourth station has one multipurpose radar with a phased array (AN/FPS-85). The BMEWS stations have complicated electronic computers, automated systems for collecting data, and special communication systems, in addition to their radars.

Figure 17.5 is the schematic of one of the first of the BMEWS stations.

Missions of the BMEWS radar station. The first system of radars detects the ballistic missiles making the strikes and evaluates the possible strength of these strikes. The second system of radars makes an accurate determination of the nature of the targets, that is, makes the preliminary recognition, and issued highly authentic information on the trajectory over which each of the targets is moving.

The same radar system carries out these missions in the third and fourth BMEWS stations.

The AN/FPS-50 Surveillance Radar

The AN/FPS-50 radar is used in the radar system in the first BMEWS station. Figure 17.6 is the block schematic of the radar, which is coherent-pulse radar operating in the decimeter band and searches a 38° sector in azimuth with electrical beam switching. The radar forms three fixed partial patterns in the plane of elevation, 2.5 to 3° wide. Pattern width in the azimuth plane is 1°. Consequently, when the amplitude method is used to measure the angular data of the target, the accuracy of this radar in determining the azimuth is higher than that with which the elevation is determined. Figure 17.7 is the radiation pattern for the AN/FPS-50 radar.

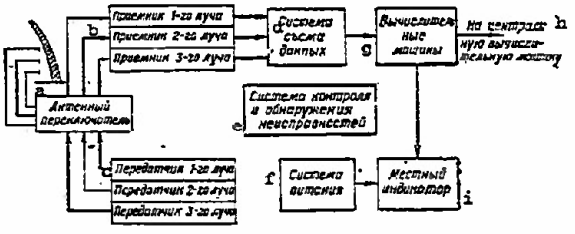


Figure 17.6. Block schematic of the AN/FPS-50 radar. a - antenna switch; b - receivers of 1st, 2nd and 3rd beams; c - transmitters of 1st, 2nd and 3rd beams; d - data collection system; e - system for monitoring and detecting faults; f - supply system; g - computers; h - to central computer; i - local indicator.

Radar pulse power is 10 megawatts, with the average power 600 kilowatts. Pulse length is 2000 microseconds, and range is some 500 km. The radar has eight operating klystron transmitters, and the same number on standby

The block schematic of the AN/FPS-50 transmitter is shown in Figure 17.8. The transmitter radiates from a fixed antenna with truncated parabolic reflectors 50 meters high and 120 meters long. Four radar antennas at the

first of the BMEWS stations scan a 150° arc in azimuth, while three antennas at the second station scan a 110° sector.

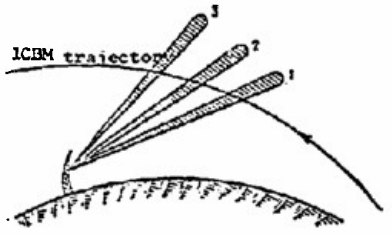


Figure 17.7. AN/FPS-50 radar radiation pattern.

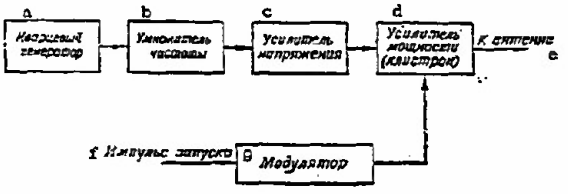


Figure 17.8. Block schematic of the AN/FPS radar transmitter.
a - quartz crystal oscillator; b - frequency multiplier; c = voltage amplifier; d - power amplifier (klystron); e - to antenna; f - trigger pulse; g - modulator.

The radar fixes target azimuth with respect to the center of gravity of the pulse packet reflected from the target at the moment the radiation pattern beam passes through, and fixes the elevation by the number of the beam the target entered. Doppler frequency shift is used to establish target speed in the direction of the station.

Special waveguide distribution switches are used to switch the AN/FPS-50 radar from the transmit mode to the receive mode.

The receiver is a superheterodyne type (fig. 17.9) consisting of a high-frequency amplifier, mixer, local oscillator, intermediate frequency amplifier, and Doppler frequency analyzer. All of the components in the functional schematic are standard, with the exception of the analyzer. The analyzer is a set of low-pass filters, the frequency curves for which partially overlap. The analyzer filters provide an overlap for the whole band of possible Doppler frequencies. The maximum value of signal amplitude is obtained at the output of this filter, the resonant frequency of which coincides with the signal frequency. All the filters

are numbered, and the number of the filter at the output of which there is a maximum signal carries the information on target radial speed.

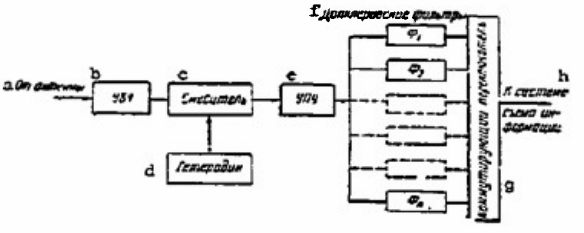


Figure 17.9. Block schematic of the AN/FPS-50 radar receiver.
a - from antenna; b - high-frequency amplifier; c - mixer; d - local oscillator; e - intermediate frequency amplifier; f - Doppler filters; g - commutating switch; h - to information removal system.

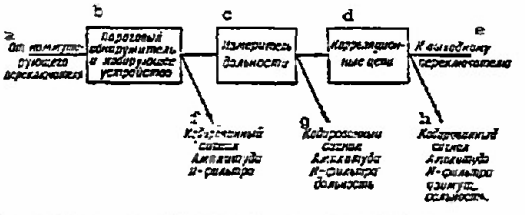


Figure 17.10. An example type schematic of the primary processing of radar information at a BMEWS station.
a - from commutating switch; b - threshold detector and coder; c - range measurer; d - correlation circuits; e - to output switch; f - coded signal, N-filter amplitude; g - coded signal, N-filter amplitude, range; h - coded signal, N-filter amplitude, azimuth, range.

Receiver sensitivity is $P_{rec} = 10^{-15}$ watt. The potential for the AN/FPS-50 radar is 240 db when the noise figure is 5 db.

The information contained in the receiver signal is extracted from it by an automatic processing system which provides for data removal, preliminary processing, and conversion of the continuous data into digital data for transmission to the station's electronic computer. This system (fig. 17.10) is a complex of high-speed computers located between the radar and the station's electronic computer.

Data on range, speed, and target azimuth are recorded in a magnetic

block in the time memory for purposes of verifying their correlation with preceding signals reflected from the same target after one scanning cycle. There is a preliminary sifting out of signals that carry no information on the ballistic targets, the result of the correlation, and only data on potentially dangerous targets is fed into the input of the electronic computer in the BMEWS station. This information is subjected to additional correlation in the electronic computer to filter out with greater reliability those signals carring no information on ballistic targets. The station's electronic computer then uses the signals on potentially dangerous targets to produce the commands for guidance to the target for the AN/FPS-49 tracking radar.

The AN/FPS-49 Tracking Radar

The second of the radar systems in the BMEWS station has AN/FPS-49 radars designed to track targets, make an accurate determination of the coordinates of targets that have been detected, make a preliminary identification, determine the points of fall of the heads of ballistic missiles, and provide target acquisition for the radar facilities for surveillance of the active antimissile defense systems. These radars can also operate in the surveillance mode, and can scan and search for targets independently, as is the case at the third BMEWS station in England. The AN/FPS-49 radar is a monopulse radar operating on two frequencies, one in the long wave portion of the decimeter band, and the other in the short wave portion of the same band. The frequency band width is from 30 to 3000 MHz.

The radar has a rotating antenna with a parabolic reflector (the scanning zone in azimuth is 360°, the elevation 90°), so either circular or sector scan can be used. The width of the antenna radiation pattern in both planes is about 2°. Range is approximately 4800 km. The pulse power of the kylstron transmitter is on the order of 10 megawatts. The parabolic antenna has a diameter of 25 meters and is installed under a radome which protects it from the harmful effects of atmospheric conditions. A special hydraulic system drives the antenna.

The AN/FPS-49 uses optimal filtration of frequency-modulated radio frequency pulses and their compression in length, thus providing long range and good resolution.

It has been reported that the AN/FPS-49 radar installed in the third of the BMEWS stations has been greatly modernized. This radar operates on frequencies of 425 and 1320 MHz in the monopulse mode with a repetition frequency on 1500 Hz. The pulse compression ratio is 50. Parametric amplifiers are used in the receiving system, so signals with 10^{-15} watt can be received.

Two radars are used for surveillance, one for target tracking. Signals reflected by targets are received by the radar are fed from the receiver output to the system that makes the pickup and does the preliminary data processing, then to the BMEWS station's electronic computers, which establish target trajectories, compare them with ballistic missile trajectories contained in the computer memories, separate the dangerous targets, determine the point of fall of the heads, and the area from which the ballistic missile was launched.

The All-purpose AN/FPS-85 Radar

The AN/FPS-85 radars comprise the basic radar system for the fourth BMEWS station and is designed for surveillance, tracking, and recognition of ballistic missiles and space vehicles crossing the equator and headed for the United States. This station is the southern early warning system. The AN/FPS-85 radar operates in the decimeter band, has a frequency band width of 400 to 500 MHz, and has a range of some 6,000 km (according to recent data the range can evidently be increased to 18,500 km). The radar simultaneously scans a 60° sector in azimuth. For the future the American specialists plan on installing six AN/FPS-85 radars capable of scanning all of the space above the equator in the fourth BMEWS station.

It is reported that the AN/FPS-85 can detect, track, and recognize a great many targets (over 100) all at the same time.

The phased transmitting antenna array is square, measuring 40 x 40 meters, the antenna array is 30 meters high, and the angle of inclination to the horizon is 45°. There are 3200 electronic switching elements forming the controlled radiation pattern, and there are the same number of transceivers. The radar's transmitting antenna array radiates a flat phase front to form the radiation pattern. The beam of the radiation pattern in this case is normal to the phase front and is controlled by the introduction of the corresponding phase delay between the elements in the antenna array, so that the phase front is tilted in the desired direction. Figure 17.11 shows the schematic for beam control over the azimuth and vertical channels.

A delay line with taps is used to rotate the beam. Each tap corresponds to one row, or column in the antenna array. One delay line is used for beam control in the plane of the azimuth, the other for control in the vertical plane. In order to turn the radiation pattern to the desired heading the programmer selects the corresponding frequency, which is produced by the generator with the help of the digital selector. The selected frequency is mixed with the fixed frequency and the intermediate frequency thus obtained is fed into the delay line with the takeoffs. The signal from the output of each takeoff is once again mixed with the selected frequency in order to

obtain the initial frequency, f_0, and f_1 with a phase shift that will depend on the frequency selected by the digital selector.

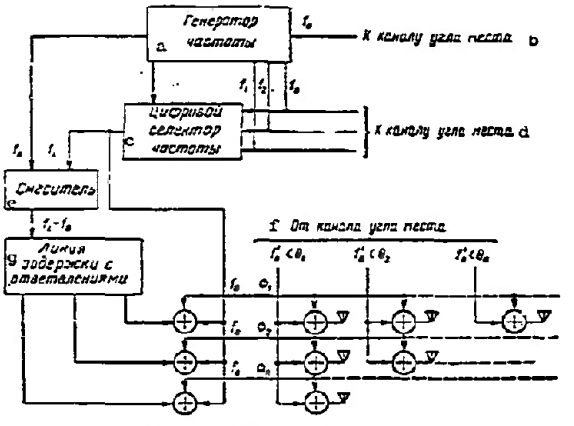


Figure 17.11. Schematic diagram of radiation pattern beam control for the AN/FPS radar.
a - frequency generator; b - to elevation channel; c - digital frequency selector; d - to elevation channel; e - mixer; f - from elevation channel; g - delay line with taps.

A similar process is carried out for each coordinate of beam control of the radiation pattern. The receiving antenna used with the AN/FPS-85 radar has a phased array in the form of a circle with a diameter of 58.5 meters. This radar can operate automatically and can be controlled remotely from the antimissile defense command post. Upon command the radar will switch from the search mode, which is optimal for ballistic missile surveillance, to the search mode for space vehicles. The radar automatically adjusts output power according to whether it is operating in the search or the surveillance modes, and exercises constant control over its parameters, such as phase, amplitude, and stability of operation of transmitters, as well as over receiver thermal noise.

BMEWS Electronic Computers

Every BMEWS station has two special electronic computers of the IBM-7090 type in addition to the radar already discussed. These computers operate in parallel on real time and monitor each other. The speed of the computers is such that 200,000 operations per second can be completed and the memory is designed for 32,768 10-digit numbers. The tasks the computers can perform were discussed above. Let us review, briefly, how the computers operate.

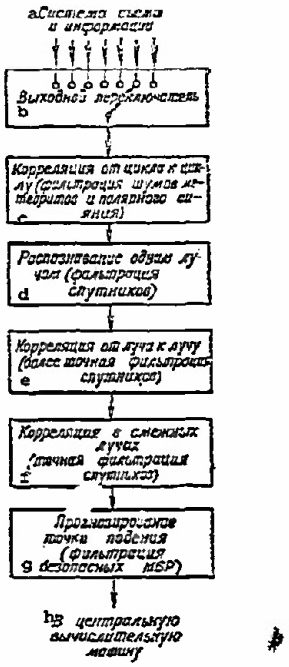


Figure 17.12. Schematic diagram of radar data processing by the BMEWS computer.
a - pickoff and information system;
b - output switch; c - correlation from cycle to cycle (filtration of meteorite and polar light noise); d - recognition by one beam (filtration of satellites);
e - correlation from beam to beam (more precise filtration of satellites);
f - correlation to adjacent beams (precise filtration of satellites);
g - prediction of point of fall (filtration of safe ICBM); h - to central computer.

Target data are brought into the computers from each of the BMEWS radars through a special pickoff and preliminary processing system. The computers made a successive correlation of the signals. Figure 17.12 is the schematic of how radar data are processed in an electronic computer. As will be seen from the schematic, the first operation performed by the computer is signal comparison over several scan cycles. This makes it possible to eliminate from further processing some of the false signals produced by polar lights and reflections from meteorites. The succession of operations

performed by the computers envisages separation of ballistic targets from artificial satellites by checking speeds, changes in range and in angular data. Dynamic characteristics of the targets are established. All of these operations are performed on the basis of the information received during the period while the target is in one of the radar beams. As a result of these operations, over 90% of the signals that are not those reflected from the ballistic missile are screened out. In order to further screen out some of the remaining false signals the station's computers compute the value of the coordinates of the assumed position of the ballistic target in the second radar beam and compare them with the measured ones, as is shown in Figure 17.13. And based on the data obtained as the target intersects all the radar beams, the station computers establish the trajectory with great accuracy, as well as the probable launch area and the area in which the missile will fall.

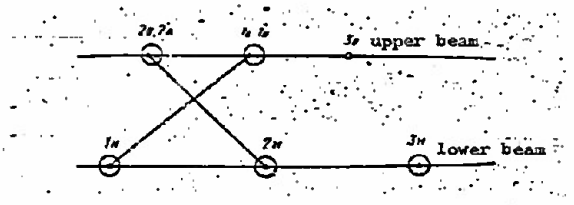


Figure 17.13. Diagram of parameter comparison in the BMEWS computer.

1u, 2u, 3u – computed values of the coordinates in the upper beam in terms of the lower beam information (1l, 2l, 3l); 1'u, 2'u – measured coordinates of upper beam; target 3 – false, because the upper beam cannot be confirmed (there is no 3'u).

The final computer operation involves the production of two types of coded messages concerning the targets for transmission to the antimissile defense command and to the Pentagon. One message contains information as to the number of threatening targets, or of the level of the threat, while the second contains information on probable areas subject to attack, time of the attack, and the areas from which the ballistic missiles were launched.

The results of the computations made by both computers are fed to a comparator. When the divergence in the results are of a magnitude in excess of permissible limits a device for automatically checking the computer is cut in and determines which of the computers is in error, finds the circuitry faults, and cuts out the particular computer. The corresponding computations are shifted to the computer that is functioning normally. It must

be pointed out that the system used to automatically check the computer also makes it possible to simulate missile attack conditions.

Operation of all BMEWS station radars and electronic computers is checked by a simulator at least once a day. In the event a real attack is picked up while the check is being made all the station's radars automatically shift to the working search and target acquisition mode.

System Operations Center

The Operations Center of the ballistic missile early warning system is in Colorado Springs. Information from all four of the BMEWS stations are fed into it. The Center has decoders, electronic computers, and special indicators.

After the information received from the BMEWS stations has been decoded and mutually correlated, it is reproduced on indicators and on a special map of the world. The potential threat thus presented is evaluated and a decision made. The Operations Center also has special equipment for simulating the flight of a ballistic missile and is used to train personnel.

The transmission of information from all four of the BMEWS stations to the Operations Center is via special communication links with very rigid specifications with respect to transmission speed, accuracy, and reliability. The communication links from each of the BMEWS stations to the Operations Center are duplicated and they run in different geographic directions. A wide variety of types of communications, including radio relay, tropospheric, and cable (land lines, underground lines, and submarine cables), are used to ensure optimal operation conditions. This considerably improves reliability in receiving information from BMEWS stations.

The foreign press, in addition to commenting on the advantages of the BMEWS, has also commented on its drawbacks, including vulnerability of the stations, the narrow sector of space being monitored, the limited range of direct visibility, and others.

The United States has decided to build an early warning system consisting of radars operating on the return angular inclination sounding principle [over the horizon] and of artificial reconnaissance satellites of the "Midas" type equipped with special infrared equipment for detecting ballistic missile launchings. This will be in addition to BMEWS.

17.9. Early Warning Facilities

The facts of successful radio communication over vast distances, many times in excess of line of sight distance, have long been known. The explanation is that there are several ionized layers in the earth's upper atmosphere capable of reflecting short radio waves.

The earth's surface, too, is capable of reflecting electromagnetic energy, as research has shown. Hence the earth's surface and the ionized layers of the atmosphere form a unique waveguide. A radio signal moving along this waveguide can circle the earth and be received by its station in weakened form after repeated reflections from the earth's surface and from the ionized layers of the air.

The Americans, using the return angular inclination sounding principle, built radars designated TEPEE and MADRE. Figure 17.14 is a schematic diagram of the principle upon which these radars function. MADRE is an improved version of TEPEE, so we will briefly review the former of the two.

MADRE, a return angular inclination sounding radar, is designed to detect ballistic missile launchings from the signal reflected from the ionized column of exhaust gases from the rocket motors, as well as nuclear explosions at ranges up to 6,000 km (according to some estimates made by American specialists, the detection range can apparently be increased to 12,000 km).

The following characteristics of the MADRE radar are known. Its range resolution is 16 km, its angular resoltuion is 1.5°. The beam width of the radiation pattern in the azimuth plane is 6 to 12°, and 12-to 24° in the vertical plane. The radiation pattern changes with the operating frequency selected. The latter is 20 to 30 MHz. The radar operates in the short wave band, the transmitter pulse power if 5 megawatts, the average power is 100 kw, the pulse length is 100 microseconds, and the repetition rate is 180 Hz.

In order to ensure reliable separation of the true signal against the noise background, the MADRE radar must illuminate a specified area for 20 seconds, after which the transition is made to scan the next space.

The main drawbacks in return angular inclination sounding radar, according to the foreign specialists, are the dependence of the quality of operation on ionospheric perturbations, the extremely poor accuracy, and resolution, and the impossibility of tracking a ballistic missile over the passive section of the trajectory.

Despite these serious drawbacks, the United States is currently working hard to design better models of these radars to supplement the ballistic missile early warning system and make it possible to detect different types of missiles immediately after launch. This will result in a greatly increased warning time.

The early warning system based on the MIDAS artificial satellite is designed to detect the launching of ballistic missiles anywhere on earth from their infrared radiation. This task can be carried out by several thousand satellites spinning in the most diverse orbits.

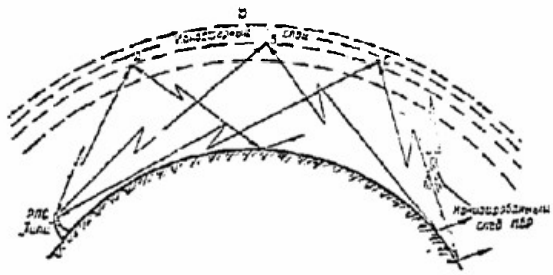


Figure 17.14. Schematic diagram of the operating principle of the return-oblique sounding radar.
a - TEPEE radar; b - ionospheric layer;
c - ionized tail of an ICBM.

The press reports that in carrying out their MIDAS program the Americans ran into difficulties, not only of an economic nature, but also because the capabilities of present day infrared equipments and recognition systems are inadequate. Separating the true signal from the ballistic missile out of the background of the different noises caused by infrared radiations from a variety of ground targets has proven to be extraordinarily complex and is still an unresolved problem. Hence, in 1962, the MIDAS launches were stopped, and in 1964, the MIDAS program was put in the experimental category. Wide ranging additional research along these lines is underway at the present time.

17.10. Active Means in the Antimissile Defense System

The early warning and long-range ballistic missile detection systems would provide the antimissile defense command with enough lead time before a missile attack to put all of the antimissile defense's active systems into action.

As a result of many years of tests of individual models of an active antimissile defense system, the United States Department of Defense arrived at a conception with respect to the building of two antimissile systems, an antimissile defense system for large populated centers of the Nike Zeus type, and an antimissile defense system for concealed combat positions of the Nike X type. The combat utilization of the two systems is conceived as in echelon; the Nike Zeus antimissile defense system for the interception and destruction of targets at relatively long distances and at relatively high altitudes, and the Nike X antimissile defense system for a follow-up attack at targets that have escaped destruction at lower altitudes. The

Americans plan on using the improved readiness antimissile defense system they have already developed, the ARPAT type, at these same low attitudes in case of a mass ballistic missile attack.

17.11. The Nike Zeus System

As soon as the ballistic missile is picked up by the BMEWS and has been preliminarily identified, data on it are transmitted not only to the antimissile defense command post, but also to the early warning radar in the Nike Zeus active antimissile system, which has not yet been accepted as part of the arsenal, but which is the only system in the United States undergoing complex range testing and is the basis for future antimissile defense modifications.

Figure 17.15 is a schematic of the operation and composition of the equipment for a Nike Zeus antimissile system station. The main equipment for a system station includes four radars, two specialized high-speed electronic computers, two antimissile batteries, and special communication links.

The station's radars are faced with the tasks of picking up the target already recognized by the BMEWS at ranges out to 1500 to 1600 km, recognizing it, and tracking it accurately, as well as controlling the flight of the antimissiles to intercept and destroy the target. The Nike Zeus antimissile system station functions as follows (fig. 17.15).

Information on the approaching target is fed from the surveillance radar to the electronic computers to process the data. The computers are located in the Operations Center of the Nike Zeus station. The output data from the surveillance electronic computer (4) is fed to the target recognition radar and to the intercept electronic computer (5). The target parameters obtained from the surveillance radar and from the recognition radar are fed into the intercept electronic computer where they are compared with the parameters of known ballistic targets contained in the computer's memeory. If the target proves to be the head of an enemy ballistic missile, data on it are fed from the intercept computer output to the target tracking radar, and when information is received from this radar an accurate computation of the trajectory being flown by the ballistic missile head and by the antimissile, as well as of the point of interception for the destruction of the recognized target, is made.

The data obtained from the intercept computer are transmitted to the antimissile guidance radar and at a predetermined moment in time this computer gives the launcher the command to launch the antimissile. The warhead carried by the antimissile is exploded at the computed point of encounter by a command transmitted over the radio link from the ground.

It takes 4 to 5 minutes from the time the ballistic target is picked up by the Nike Zeus system to intercept it. The Nike Zeus system is completely automated.

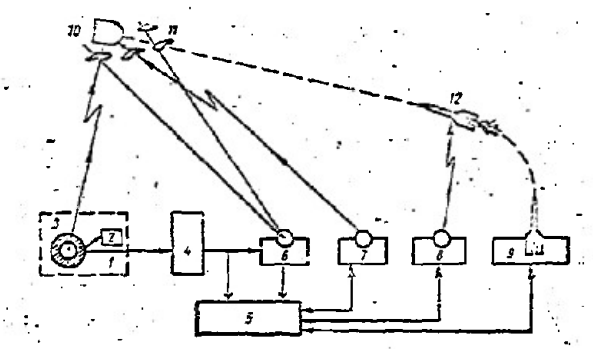


Figure 17.15. Schematic diagram of the composition and operation of the facilities in a Niki Zeus system station.
1 - ICBM surveillance radar; 2 - transmitter; 3 - receiver; 4 - surveillance electronic computer; 5 - intercept electronic computer; 6 - recognition radar; 7 - target tracking radar; 8 - antimissile tracking radar; 9 - antimissile launch site; 10 - ICBM head; 11 - false targets; 12 - antimissile in flight.

The Warning and Acquisition Radar

This radar is an intermediate link between the active defense center and the BMEWS stations. The warning radar pulses using the pulse compression technique. Radar range is 1600 km, so the Nike Zeus can, by itself, destroy ballistic targets without using the information from the early warning stations. This range is the result of high power radiation by the transmitting system and the use of methods for obtaining optimal reception of reflected signals. Radar pules power is 50 megawatts, so separate transmitting and receiving antennas are required. They are set up 300 meters apart to decouple the radar's receiving and transmitting channels.

Figure 17.16 is a block schematic of the warning and acquisition radar. The transmitting antenna consists of three radiating arrays, each of which is 24 meters long, and almost 4 meters wide. They are arranged to form what is essentially an equilateral triangle (fig. 17.16). The formed beam of the radiation pattern is narrow in the horizontal, and wide in the vertical plane. This makes it possible to use the amplitude method

to measure the azimuth. The phase method is used to measure the elevation at the radar.

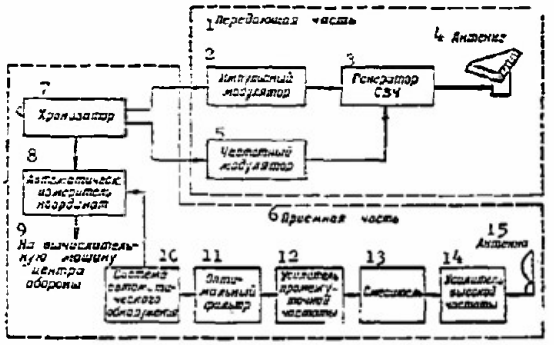


Figure 17.16. Block schematic of the warning and acquisition radar in the Nike Zeus system.
1 - transmitting section; 2 - pulse-network modulator; 3 - superhigh frequency generator; 4 - antenna; 5 - frequency modulator; 6 - receiving section; 7 - timer; 8 - automatic coordinate measurer; 9 - to defense center computer; 10 - automatic warning system; 11 - optimal filter; 12 - IF amplifier; 13 - mixer; 14 - HF amplifier; 15 - antenna.

The transmitting antenna is located on the roof of the building housing the transmitter. The building is surrounded by a metallic grid some 20 meters high. This grid eliminates reflections from nearby objects and assists in the formation of the radiation pattern. The transmitting antenna rotation rate is 10 rpm.

The receiving antenna is a Lunberg lens (fig. 17.17). It is a solid hemisphere, 24 meters in diameter, in design. It consists of foam plastic cubes, the edges of which are 45 cm long. Total antenna weight is 1000 tons. The antenna is installed under a radome 33.5 meters in diameter.

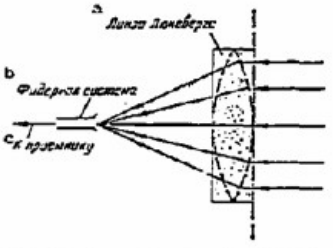


Figure 17.17. Schematic diagram of the receiving antenna for the warning and acquisition radar in the Nike Zeus system.
a - Lunberg lens; b - feeder system; c - to receiver.

The radar transmitter radiates frequency-modulated pulsed oscillations. The generators of the high-frequency oscillations are floating-drift power klystrons. The radar operates in the decimeter wave band and the pulse compression ratio is 100. A special feature of the radar is the use of high-frequency molecular amplifiers with optimum filters. This almost doubles the receiver system sensitivity as compared with conventional receiver systems used in radars.

The signal is red from the output of the optimum filters in the radar's receiver system to the information system, which is a set of devices for automatically detecting and measuring coordinates. The devices will only pass a signal when the latter exceeds some predetermined, assigned level establishing the probability of a false alarm. When the signal is passed through this circuit it is fed to the coordinate measurer which fixes the position of the ballistic missile in space and converts the position information into a form convenient for transmission to the detection electronic computer in the Operations Center of the Niki Zeus system.

Target Recognition Radar

Detection of ballistic missile and reliable tracking of them depends on the interference that will accompany the ballistic targets over the different sections of the trajectory, and these can be very different indeed. Figure 17.1 classified interferences.

It should be pointed out that recognition of a target, once it is detected, is one of the chief tasks on an antiballistic missile defense system. Target recognition is the capacity of an antiballistic missile defense system to classify individual objects in a group of targets and separate out those that have ballistic missile characteristics. All targets thus separated should then be considered as ballistic missiles and destroyed by the antiballistic missile active facilities.

Figure 17.18 is the block schematic of the target recognition radar in the Nike Zeus antimissile system. The radar comprises a transmitter, receiver, and antenna system. The transmitter uses amplitrons which, in the 10 cm band, can develop power on the order of several megawatts with 80% efficiency. The recognition radar has much better resolution, so far as angular data are concerned, than do the existing detection and acquisition radars. The high resolution is primarily the result of the special antenna system design.

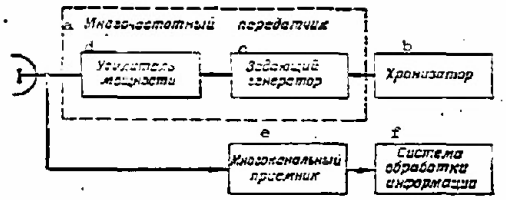


Figure 17.18. Block schematic of the target recognition radar in the Nike Zeus system.
a - multifrequency transmitter; b - timer; c - master oscillator; d - power amplifier; e - multichannel receiver; f - information processing system.

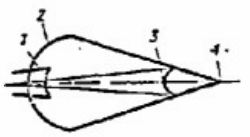


Figure 17.19. Beam forming diagram for the target recognition radar in the Nike Zeus system.
1 - real exciter; 2 - parabolic mirror: 3 - hyperbolic mirror; 4 - imaginary exciter.

The spherical waves radiated by the primary exciter strike the hyperbolic mirror, which is the secondary exciter, and as they are reflected from it form a spherical wave front, the center of which is at the point where the imaginary exciter is located (fig. 17.19).

The entire breadth of the radar's transmitting antenna can be used to radiate high frequency energy. The frequency is stepped up gradually until the radar forms 12 narrow beams. This gives the radar greater resolution than when the radiation pattern is formed of just one broad beam. Antennas such as these have certain structural advantages, the most important of which is the optimum amplitude of distribution in the aperature, and great sensitivity when measuring angular coordinates.

The radar has equipment for rapid processing of incoming information, in addition to the basic systems that are direct components of the radar's functional schematic.

The information processing schematic is shown in Figure 17.20.

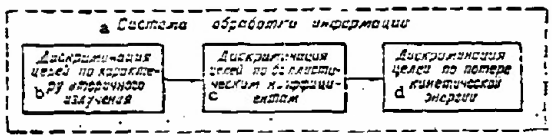


Figure 17.20. Schematic diagram of how radar information is processed by the target recognition radar in the Nike Zeus system.
a - information processing system; b - target descrimination by nature of secondary radiation; c - target descrimination by ballistic coefficient; d - target descrimination by loss of kinetic energy.

When a target is observed at long range the recognition is the result of the nature of the secondary radiation. A comparison is now made of the laws obtained when studying the reflecting properties of various bodies. The result is preliminary information on the number of real and false targets. This completes the first stage of information processing. The second stage begins with the entry of the ballistic missile warhead into the dense layers of the atmosphere (at altitudes between 130 to 100 km). This is the stage during which the second and third logical units in the rapid processing equipment begin to function. The second solves the problem of recognition by analysis of the braking laws and rejects all of the obviously false targets. The unit continues to function until the ballistic missile warhead is at an altitude of approximately 80 km. When this unit ceases operation the number of targets subject to further analysis has dropped sharply.

The third logical unit separates targets with the same mass ballistic coefficients. Here the laws of change in reflecting surface and dimensions of the jet trail with altitude are used. Information analysis by the third logical unit should result in complete separation of the ballistic missile warhead from all false targets.

Foreign specialists feel that the target recognition radar does not provide reliable recognition of ballistic missile warheads against a background of false targets as experience has demonstrated, and that the reliability imputed to it is not justified.

Target Tracking Radar

Once the ballistic missile warhead has been recognized data concerning it are fed into the target tracking radar in the Nike Zeus antimissile system. The target tracking radar is a monopulse type with a pencil-beam pattern and a range of about 1000 km. Transmitter pulse power is 20 megawatts.

The transmitter forms long pulses. Antenna gain is 35000. The antenna

is parabolic with a diameter of 7.6 meters under a special radome, thus eliminating wind loads on the antenna and at the same time increasing the accuracy with which angular coordinates are measured. Errors in measuring the coordinates of the tracked target are not in excess of 5 meters in range and 20" in angular coordinates.

The high resolution is the result of using the pulse compression method, in which the compression factor is 100.

Once the tracking radar has locked on there is continuous transmission of coordinates from it to the input of the intercept electronic computer, which establishes the trajectory over which the target is moving, the antimissile launch time, and the point of impact.

The radar can only track and provide information to the intercept computer on one target. This is considered to be the greatest of its drawbacks, one which, in the event of mass attacks, will result in saturation of the Nike Zeus system and in failure to carry out the combat mission.

The Antimissile Guidance Radar

Data from the intercept computer is fed to the Nike Zeus antimissile guidance radar continuously. When the antimissile is launched the radar locks on and guides the antimissile to the point of impact and at the proper moment in time issues the command to explode the antimissile's warhead.

A dual-beam guidance command system is used to guide the antimissile to the target. The last stage of the antimissile has a radar responder, the purpose of which is to increase the antijam properties of the channel monitoring the flight of the antimissile. The signals fed to the receiver in the guidance radar from the responder are greatly in excess of the internal receiver noise and this results in reliable control of the antimissile flight over long distances from the radar (300 to 350 km).

Target Intercept Electronic Computer

The target intercept electronic computer is the brain of the entire Nike Zeus antimissile system. The computer makes it possible to exchange information between all radars and the control point.

The high-speed computer performs 2000000 operations in a second. The machine only needs 2.8 microseconds to select data from the memory storage. Over 90% of the components are standard electronic blocks in which semiconductors are used. All of the electrical connections used in the computer are in the form of specially designed joints providing highly reliable contacts. The period of failure-free operation is 600 hours. All target data supplied the computer from the Nike Zeus radars are introduced into the computer in coded form by magnetic tape.

Thanks to the modular design and the use of the built-in facilities for detecting faults, the monitoring and technical servicing of the intercept electronic computer is virtually completely automated.

The Nike Zeus Antimissile

In design, Nike Zeus (fig. 17.21) is a three-stage missile with solid fuel engines. Missile length is 14.7 meters, launch weight is 10 tons, maximum firing range is estimated at 320 to 350 km, reaching an altitude of 150 to 200 km, and a flight speed of over Mach 4.

The first and second states of the antimissile are the launch boosters and have powerful powder jet engines with a thrust of 200 and 160 tons.

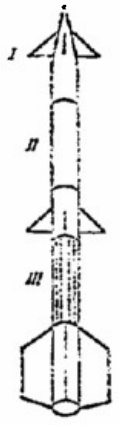


Figure 17.21. Overall view of the Nike Zeus antimissile.

The third stage of the antimissile consists of four sections. The nose section with the nuclear warhead, the engine section, the section containing the hydraulic drives for the fins, and the after instrument section with the equipment in the control system and the electrical supply.

The design of this stage is based on the appearance of the "weft" of a fabric. A gas dynamic system is used to control the antimissile because maneuvers take place outside the atmosphere's dense layers. The guidance jet engine consists of a spherical combustion chamber and four rotatable nozzles located in the body in front of the fins. The fins are tilted in the required direction as the control engine nozzles are tilted to create the controlling forces and moments in accordance with commands received from the guidance radar. The direction of the vector of forces acting on

the antimissile is changed thus causing the corresponding maneuver. The time at which the command for controlling the flight of the antimissile is issued takes the missile's maneuverability into consideration.

The electronic equipment in the control system is located in the after compartment of the antimissile's third stage and comprises the following main components: a receiver-responder; command control receivers and amplifiers; the reference gyro plate; synchronizers; supply sources; and the electrical circuits connecting all the equipment.

An important feature of the Nike Zeus is that it reaches supersonic speed even while in the relatively dense layers of the atmosphere, the result of the high acceleration. This is why the lines of the antimissile are smooth, and the outer surface is covered with a special layer of fused heat-protecting material.

The antimissile has an automatic, programmed, self-destruct system. The signal to set the self-destructor is issued at the fifth second after launch and the antimissile remains in that condition for 11 seconds. If an emergency arises during that period of time the antimissile self-destructs.

The Nike Zeus antimissile battery has 24 stationary underground launchers. The launch of any particular antimissile is by command from the intercept computer, which issues the launch command to the battery at the particular moment decided upon by the situation in the air. The antimissile is launched from its shaft vertically and by the time the boosters have burned out has reached a speed of almost 1 km/second. The antimissile's warhead is detonated in the vicinity of the meeting point upon command from the guidance radar.

Each installation in a Nike Zeus system has two batteries, controlled from one control post. The entire process of locking on to the ballistic missile warhead by Nike Zeus antimissiles is completely automated, but manual control of system operation is available in case of necessity.

The battery control post has a special scope providing a three-dimensional image of target and antimissile. This scope is used for constant visual monitoring of the progress of interception and target destruction.

Test results have revealed that Nike Zeus is comparatively inertial and not maneuverable enough to ensure interception of a ballistic target at various altitudes. Work is in progress to eliminate these drawbacks.

17.12. The Nike X System

The Nike X development in the United States began in 1963, as an extension of the program to create a reliable antimissile defense complex.

Nike X, as an antimissile system, is designed to protect underground communications stations, command posts, and underground battle stations of ICBMs against medium- and long-range ballistic missiles. It is believed that the system can fight a great many recognized ballistic targets simultaneously.

Included in the Nike X system is a control center with a multifunction array radar (MAR), and high-speed electronic computer complex, a missile site radar (MSR), for launch purposes, and a battery of SPRINT type antimissiles.

Nike X can have less rigid specifications than can the antimissile defense system for the heavily populated centers, the Nike Zeus.

This stems primarily from the small area defended. Here the attacking ballistic missiles would have to fly through a narrow corridor, the so-called attack corridor. The American specialists are assuming that because of the narrowness of the attack corridor the problem of detecting the target is simplified, and that the interception of targets at low altitudes (15 to 20 km) can simplify their recognition because false targets accompanying the warhead will differ from the warhead's characteristics for entry into the atmosphere and flight in the atmosphere because of the differences in weights and coefficients of drag.

Experiments conducted in the United States have revealed that when there is a 20:1 ratio between the warhead mass and the false target mass the difference in trajectories is enough to afford recognition of the ballistic missile warhead at altitudes of from 60 to 80 km. When the ratio is smaller, recognition is not possible until the altitude is in the 25 to 40 km range.

Moreover, at low altitudes the zone of probable target interception in the air space is smaller, and this makes for increased fire power for the installed number of antimissiles and for more reliable target destruction.

The less rigid specifications for the Nike X system have simplified it and made it cheaper than the Nike Zeus system.

The radioelectronic equipment and the SPRINT antimissile have already been developed for the Nike X system and have already been, or are in the process of, independent tests on United States ranges.

The MAR (Multifunction Array Radar)

This radar is designed to detect, recognize, and track ballistic missile warheads.

The MAR main components include the antenna with phased array, a transceiver, equipment for processing radar information, and radar control equipment. MAR operates in the centimeter band (from 19.3 to 76.9 cm), and its range is some 1200 km. Because of the rapidity of electronic control of the antenna radiation pattern one beam will be able to track several targets, and the possibility of forming several beams with individual control of each beam ensures that search and tracking functions when a great many targets are involved can be combined in one radar.

The radar antenna array consists of several thousand elements about half a wavelength apart. The sum of the powers of the individual elements in the array results in the obtaining of a high output power from the final average power amplifiers. Pulse power is 100 megawatts, average power is 10 megawatts.

Each element in the antenna array is connected to the phase controller, consisting of a large number of waveguide sections that are cut in by switching diodes, so it is possible to regulate the time of arrival of the signal from the transmitter for each of the radiating elements in the antenna. This control method reduces the level of the MAR side lobes to 35 db.

The phased antenna is not critically sensitive to failure of a particular antenna element, and the radiated power in practice is limited only by the total number of active elements in the antenna array. Direction of radiation can be changed 180°.

All of the radar equipment (with the exception of the flat screen of the antenna array) is underground in a concreted shelter. According to statements made by American specialists, the radar performs fine horizon scanning at elevation angles of 1 to 3°. This is why, in their opinion, the Nike X antimissile system should be able to fight ICBM (including missiles with low trajectories), medium range ballistic missiles, and missiles launched from submarines and underwater launchers.

The equipment for processing the radar data is such that information on azimuth, elevation, range, and recognition results can be processed on a preliminary basis, coded, and transmitted to the electronic computer complex in the Nike X system.

Electronic Computer Complex

The electronic computer complex in the Nike X system is under development. It is designed for processing all the radar information fed from the Nike Zeus and Nike X stations, for developing the command to launch antimissiles, and for controlling them is flight. The complex speed should be

such as to handle 6 billion operations a minute, and the program input frequency should be 100 MHz. The complex will use 25 processing devices, 64 program accumulation units, 64 operant accumulation units, and 4 16-channel units for data input and readout. Complex memory capacity will be 524,288 32-digit words and will provide nondestructive information readout. The operant memory too will have a capacity of 524,288 32-digit words, but the readout will be destructive.

The unusual complexity of this future computer complex for the Nike X system is creating a very serious problem, that of providing reliable operation. Whereas the electronic computers already in use in the American ballistic missile early warning system are designed to permit but one brief error in 24 hours of continuous operation, the electronic computer complex for the Nike X system would make 1000 errors in one hour of operation, given equivalent parameters. This is why the complex designers are looking for different ways in which to bring about a sharp reduction in the number of errors. A complicated system of detecting errors, for indications of parameters, and for establishing the defect site and element, is under development. Moreover, programs, and particularly important data, will be accumulated in several memories in order to avoid their complete loss in event of damage, or breakdown in one block, and thus provide greater complex reliability.

The specialists assume that as a result of these measures the breakdown in one, or in several elements, in the complex will not interrupt its operation, but will only result in a loss of speed.

The tremendous speed and reliable operation of the electronic computer complex in the Nike X system will ensure interception of ballistic missile warheads in the dense layers of the atmosphere by the highly maneuverable SPRINT antimissile in a minimally short period of time, according to the American specialists.

The MSR Launch Radar

The MSR type launch radar in the Nike X antimissile system is designed to guide the antimissile to recognized ballistic targets. The radar is under development, like the MAR and will have a phased antenna system that will enable it to simultaneously guide a great many active weapons to intercept, and will increase the effectiveness with which targets are destroyed.

Independent test are currently in progress with the MSR.

The SPRINT Antimissile

The time remaining from the time of recognition to the time of intercept of the warheads of ballistic missiles is 15 to 20 seconds for the Nike X

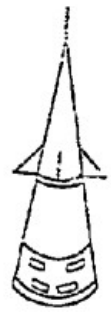


Figure 17.22. Schematic diagram of the overall appearance of the SPRINT missile.

antimissile system, thus considerably increasing the requirements imposed on the antimissile.

The SPRINT antimissile (fig. 17.22) is a single-stage missile with a maximum firing range of 150 km, and an intercept height of 30 to 50 km. The engine develops a thrust of 300 tons and burns a solid, highly economical, two-component fuel. One of the features of the SPRINT propulsion plant is the shape of the fuel charge. It must burn rapidly and be able to withstand the the forces resulting from the high initial accelerations. It is reported that SPRINT will accelerate to 80 to 100 g in 2 to 3 seconds, and that it has a maximum flying speed of over 3 km/second.

SPRINT weighs but 6 tons, thanks to the single stage and the much simplified on-board control system. The antimissile is 8.2 meters long, and the maximum diameter of the body is 1.4 meters.

SPRINT will be stowed in shafts at battle stations and will be launched by compressed air, or gas. The cruise engine cuts in at some altitude above the earth's surface. Control of the antimissile is by command from the guidance system to the aerodynamic rudders. The antimissile carries a thermonuclear warhead. It is reported that SPRINT has successfully passed independent range tests.

17.13. The ARPAT Type Improved Readiness Antimissile Defense System

The question of protection against a massive ballistic missile strike continues to disturb foreign military specialists. The United States Department of Defense has contracted for the development of the third ARPAT type antimissile defense system designed to destroy ballistic missile warheads on the final section of the trajectory in event of a massive strike when the Nike Zeus and Nike X systems cannot destroy all the targets.

According to the ARPAT project, the ballistic missile warheads will be fired at by so-called "shrapnel" antimissiles weighing some 20 to 25 kg, launched with supersonic speed from a missile carrier, when the warheads enter the dense layers of the atmosphere.

Plans call for the installation of several tens of these antimissiles on each carrier. The antimissiles will be spheroidal, or elliptical, in shape. Their in flight range will be 10 to 15 km. They will have an engine, infrared and ultraviolet systems for homing, and control jet nozzles.

Plans for the combat use of "shrapnel" antimissiles as follows. The missile carrier will be taken to an altitude of from 6 to 30 km in such a way that it is somewhat lower than the targets and several kilometers away from them. Carrier speed will now be reduced gradually by a special braking engine and when the carrier speed equals zero all of the "shrapnel" antimissiles will be launched as a salvo against the group of targets detected.

ARPAT should be highly reliable and should supplement the Nike Zeus and Nike X systems. This is why there is thinking about combining all three antimissile defense systems into one.

Naturally, the appearance of three types of antimissiles systems has placed different assessments on their advantages and disadvantages, and these assessments have considered the technical, economic, and strategic aspects of the overall problem of antimissile defense.

The decisive factors in favor of building the Nike X antimissile system are considerations that are strategic in nature. The American specialists feel that the Nike X system will be an important "deterrent" factor and will make it possible to create reliable and invulnerable forces for inflicting a "retaliatory" nuclear missile strike. This is why today the American militarists primary attention is devoted to the Nike X antimissile system, although work on improving the Nike Zeus system has not ceased, nor has the development of the ARPAT system.

Recently available press reports have made it known that a decision has been made to combine the Nike Zeus and Nike X active antimissile systems into one antimissile complex which would carry out the missions assigned both systems. It appears that the combined system will make full use of the radio-electronic equipment in the Nike Zeus system, the SPRINT antimissile, with the batteries of modernized Nike Zeus antimissiles considered supplementary.

The combined complex installation, which would, as before, be called the Nike X, would be counted on, it appears to defend an area of the United States encompassing the territroy of three to four states

The most varied of antimissile defense projects and schemes are under intensive study, but in the opinion of the Americans themselves, all that has

resulted thus far is an expansion of their own idea of how extraordinarily complex antimissile defense problems are, while at the same time bringing about even more and more difficulties that American science and engineering are in no condition to cope with at the present time.

LITERATURE

CHAPTER I

1. V. Ye. Dulevich, L. N. Korostylev, et al. Theoretical Fundamentals of Radar. "Sovetskoye Radio" Publishing House, 1964.

2. A. G. Saybel'. Radar Fundamentals. "Sovetskoye Radio" Publishing House, 1961.

3. V. V. Vasin, O. V. Vlasov, et al. Aviation Radar. Zhukovskiy VVIA Publishing House, 1964.

CHAPTER II

1. V. G. Grigor'yan. Engineering Indices for Radar Stations. Military Publishing House, 1963.

2. B. F. Stuchevskiy. Radar and Its Application. Military Publishing House, 1962.

3. V. V. Vasin, O. V. Vlasov, et al. Aviation Radar. Zhukovskiy VVIA Publishing House, 1964.

4. R. Shlezinger. Radioelectronic War. Military Publishing House, 1963.

CHAPTER III

1. F. B. Chernyy. Radio Wave Propagation. "Sovetskoye Radio" Publighing House, 1962.

2. Yu. A. Mishchenko. Detection Zones. Military Publishing House, 1963.

3. I. P. Zherebtsov. Meter Wave Engineering. DOSAAF Publishing House, 1955.

4. Yu. A. Ivanov, B. V. Tyapkin. Infrared Engineering in Military Affairs. "Sovetskoye Radio" Publishing House, 1963.

5. Ya. I. Dayenov, I. S. Krasil'nikov, "The Effect of Nuclear Explosions on Radio Communications and Radar Operation," Vestnik PVO [Air Defense Herald], No. 1, 1966.

CHAPTER IV

1. I. P. Markov. Transmission Lines. Military Publishing House, 1958.

2. I. V. Lebedev. Superhigh Frequency Techniques and Instruments. Vol. I. Gosenergoizdat, 1961.

3. V. I. Beketov. Superhigh Frequency Antennas. Military Publishing House, 1957.

4. A. L. Drabkin, V. L. Zuzenko. Antenna-Feeder Installations. "Sovetskoye Radio" Publishing House, 1961.

CHAPTER V

1. V. F. Vlasov. Electronic and Ion Instruments. Communications Publishing House, 1960.

2. N. N. Khlebnikov. Electronic Instruments. "Communication" Publishing House, 1964.

3. Ya. A. Fedotov. Fundamentals of the Physics of Semiconductor Instruments. "Sovetskoye Radio" Publishing House, 1963.

4. Semiconductor Instruments and their Application. A collection of articles edited by Ya. A. Fedotov. Issues 1-17.

5. Semiconductor Diode and Transistor Handbook. Edited by N. N. Goryunov. "Energiya" Publishing House, 1964.

6. D. S. Gurlev. Electronic Instrument Handbook. Gostekhizdat UkSSR, 1962.

CHAPTER VI

1. Ya. S. Itskhoki. Pulse Devices. "Sovetskoye Radio" Publishing House, 1959.

2. L. M. Gol'denberg. Pulse Engineering Fundamentals. Communication Publishing House, 1964.

3. B. Kh. Krivitskiy. Pulse Circuits and Devices. Gosenergoizdat, 1961.

4. L. P. Krayzmer. New Elements for Electronic Digital Computers. Gosenergoizdat, 1961.

5. V. F. Samoylov, V. G. Makoveyev. Pulse Engineering. "Communication" Publishing House, 1964.

6. G. A. Baz', et al. The Pulse Circuit Calculation. Military Publishing House, 1962.

7. V. G. Milenin, et al. Pulse Engineering Fundamentals. Military Publishing House, 1966.

8. B. V. Anisimov, V. N. Chetverikov. Fundamentals of the Theory and and Design of Digital Computers. "Machinebuilding" Publishing House, 1965.

9. N. T. Petrovich, A. V. Kozyrev. Electrical Pulse Generation and and Conversion. "Sovetskoye Radio" Publishing House, 1954.

10. S. M. Gerasimov, I. N. Migulin, V. N. Yakovlev. Fundamentals of the Theory and Calculation of Transistorized Circuits. "Sovetskoye Radio" Publishing House, 1963.

CHAPTER VII

1. M. S. Neyman. Course in Radio Transmitting Devices. "Sovetskoye Radio" Publishing House, 1965.

2. S. Ye. Temkin. Pulse Modulator Tubes. Military Publishing House, 1960.

3. S. V. Kukarin. Present Status and Trends in the Development of Superhigh Frequency Instruments. "Sovetskoye Radio" Publishing House, 1962.

4. A. Shavlov, S. Fogel', L. Dalberdzher. Lasers. "Foreign Literature" Publishing House, 1962.

5. E. A. Bernshteyn, N. K. Rudyachenko. Pulse-Type Radio Transmitting Devices. Gostekhizdat UkSSR, 1963.

6. S. A. Drobov, S. I. Bychkov. Radio Transmitting Devices. Leningrad, 1964.

7. V. S. Stal'makhov. Fundamentals of the Electronics of Superhigh Frequency Instruments with Crossed Fields. "Sovetskoye Radio" Publishing House, 1963.

8. N. A. Troshanov. Radio Equipment with Traveling Wave Tubes. Sudpromgiz, 1961.

9. "Semiconductor Diodes Used in Optical Band Generators," Foreign Radioelectronics, 1963, No. 9, pp. 93-114.

10. N. F. Remsey. "Atomic Hydrogen Laser," Foreign Radioelectronics, 1963, No. 11, pp. 88-98.

CHAPTER VIII

1. V. V. Krokhin. Elements of Superhigh Frequency Radio Receivers. "Sovetskoye Radio" Publishing Housde, 1964.

2. L. S. Gutkin, V. L. Lebedev, V. I. Siforov. Radio Receivers. "Sovetskoye Radio" Publishing House, Part I 1961, Part II 1963.

3. N. I. Chistyakov. V. M. Sidorov, V. S. Mel'nikov. Radio Receivers. Communication Publishing House, 1959.

4. V. F. Barkan, V. K. Zhdanov. Radio Receivers. Defense Publishing House, 1960.

5. L. S. Gutkin. Theory of Optimal Methods of Radio Reception in Fluctuating Interference. Gosenergoizdat, 1961.

CHAPTER IX

1. V. I. Rakov. Radar Indicators. Sudpromgiz, 1962.

2. A. S. Magdesiyev, M. M. Reznik. Scanning Radar Indicators. Military Publishing House, 1963.

CHAPTER X

1. B. Kh. Krivitskiy. Automatic Systems for Radio Engineering Installations. Gosenergoizdat, 1962.

2. A. A. Korostelev. Automatic Measurement of Coordinates. Military Publishing House, 1961.

CHAPTER XI

1. R. Shlezinger. Radioelectronic Warfare. Military Publishing House, 1963.

2. A. N. Volzhin, V. A. Yanovich. Radar Countermeasures. Military Publishing House, 1960.

3. A. I. Paliy. Radio Warfare. Military Publishing House, 1963.

4. P. A. Bakulev. Radar Methods for Selecting Moving Targets. Oborongiz, 1958.

5. K. N. Trofimov. Radar Interference. DOSAAF Publishing House, 1962.

6. Z. M. Kanevskiy, M. I. Finkel'steyn. Fluctuating Noise and the Detection of Pulse Radio Signals. Gosenergoizdat, 1963.

7. Ya. D. Shirman, V. N. Golikov. Fundamentals of the Theory of Detecting Radar Signals and Measuring Their Parameters. "Sovetskoye Radio" Publishing House, 1961.

8. S. Ye. Fal'kovich. Reception of Radar Signals in Random Noise. "Sovetskoye Radio" Publishing House, 1961.

9. Yu. G. Stepanov. Camouflage Against Radioelectronic Observation. Military Publishing House, 1963.

CHAPTER XII

1. Ye. S. Venttsel'. Theory of Probabilities. Fizmatgiz, 1964.

2. Ye. S. Venttsel'. Introduction to Operations Research. "Sovetskoye Radio" Publishing House, 1964.

3. N. V. Smirnov, I. V. Dunin-Barkovskiy. Short Course in Mathematical Statistics for Technical Applications. Fizmatgiz, 1959.

4. D. van Minter, D. Middleton. "Modern Statistical Methods in Signal Reception Theory." In the collection Signal Reception in Noise. "Foreign Literature" Publishing House, 1960.

5. M. S. Katsenbogen. Detection Characteristics. "Sovetskoye Radio" Publishing House, 1965.

6. V. V. Chuyev, et al. Fundamentals of Operations Research in Military Engineering. "Sovetskoye Radio" Publishing House, 1965.

7. M. N. Goncharenko. Cybernetics in Military Affairs. DOSAAF Publishing House, 1963.

CHAPTER XIII

1. N. A. Shishonok, V. F. Repkin, L. L. Barvinskiy. Fundamentals of the Theory of Reliability and Operation of Radioelectronic Equipment. "Sovetskoye Radio" Publishing House, 1964.

CHAPTER XIV

1. Course in Electrical Measurements. Editors V. T. Prytkov and A. V. Talitskiy. Parts I and II. Gosenergoizdat, 1960.

2. G. Ya. Mirskiy. Radioelectronic Measurements. Gosenergoizdat, 1963.

CHAPTER XV

1. Cybernetics in the Service of Communism. Collection of articles under the editorship of A. I. Berg. Gosenergoizdat, 1961.

2. A. I. Kitov, N. A. Krinitskiy. Electronic Digital Computers and Programming. Gostekhizdat, 1959.

3. Ye. A. Drozdov, V. I. Prokhorov, A. P. Pyatibratov. Computer Fundamentals. Military Publishing House, 1964.

CHAPTER XVI

1. A. A. Kharkevich. Essays on General Communication Theory. Gostekhizdat, 1955.

2. V. I. Shlyapoberskiy. Elements of Discrete Systems of Communication. Military Publishing House, 1962.

3. V. S. Gurov, G. A. Yemel'yanov, et al. Fundamentals of Data Transmission Over Wired Communication Channels. Communication Publishing House, 1964.

CHAPTER XVII

1. M. N. Goncharenko. Missiles and Antimissile Problems. DOSAAF Publishing House, 1962.

2. M. N. Nikolayev. Missile Against Missile. Military Publishing House, 1963.

3. N. F. Shibayev. The Struggle with Missiles. Military Publishing House, 1965.

4. Missiles and Antimissile Defense. A collection of translated articles. Military Publishing House, 1962.

5. Radioelectronics Abroad, 1964-1965, Nos. 1-50.

6. Air Defense Herald, 1963, Nos. 5, 10, 12; 1964, Nos. 4, 10, 12; 1965, Nos. 1, 3, 9.

7. Electronics, 1964, No. 17; 1965, No. 12. "Mir" Publishing House.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Foreign Science and Technology Center US Army Materiel Command Department of the Army	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

HANDBOOK OF RADAR ENGINEERING FUNDAMENTALS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Translation

5. AUTHOR(S) (First name, middle initial, last name)

V. Ya. Tsylov, et. al.

6. REPORT DATE 25 Jun 70	7a. TOTAL NO. OF PAGES 690	7b. NO. OF REFS N/A
8a. CONTRACT OR GRANT NO. b. PROJECT NO. c. T702301 2301 d. Requester: Booth	9a. ORIGINATOR'S REPORT NUMBER(S) FSTC-HT-23-1114-70 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) ACSI Control Number: None	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY US Army Foreign Science and Technology Center

13. ABSTRACT

This translation is a reprint of RA-015-68 produced by Redstone Arsenal.

The handbook contains information on various questions concerned with radar theory and engineering, control systems, communications, and instruments.

Physical fundamentals, operating principles, circuit construction, technical data, indices and criteria for evaluating radio frequency engineering installations, their stages, and elements, are discussed in a practical way.

Formulas for evaluations and computing certain of the parameters of radio frequency engineering installations are also cited.

The handbook is intended for officers with average and higher qualifications in the radio frequency engineering specialty. It can also be useful to the general reader interested in expanding and systematizing his knowledge in the field of radar.

DD FORM 1 NOV 65 1473 REPLACES DD FORM 1473, 1 JAN 64, WHICH IS OBSOLETE FOR ARMY USE.

UNCLASSIFIED
Security Classification

UNCLASSIFIED
Security Classification

KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Radar theory						
Radar engineering						
Radar operating principles						
Radar control systems						
Radar communications						
Radar instruments						
Radar circuit constructions						
Radio frequency evaluation						

UNCLASSIFIED
Security Classification