FTD-HC-23-722-71

AD 741156

FOREIGN TECHNOLOGY DIVISION

AIR FORCE SYSTEMS COMMAND

THEORY AND PRACTICE OF RADAR SYSTEM OPERATION

by

S. M. Latinskiy, V. I. Sharapov,
S. P. Ksenz, S. S. Afanas'yev

FOREIGN TECHNOLOGY DIVISION

Approved for public release;
Distribution unlimited.

470

R

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Radar Stations						
Radar Range Capability						
Radar Parameter						
Radar Operating Mode						
Radar Target						
Radar System Efficiency						
Radar Technology						
Radar Design Parameter						

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 Technology Division Air Force Systems Command U. S. Air Force	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

THEORY AND PRACTICE OF RADAR SYSTEM OPERATION

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Translation

5. AUTHOR(S) (First name, middle initial, last name)

Latinskiy, S. M.; Sharapov, V. I.; Ksenz, S. P.; Afanas'yev, S. S.

6. REPORT DATE 1970	7a. TOTAL NO. OF PAGES 463	7b. NO. OF REFS 69

8a. CONTRACT OR GRANT NO. F33657-71-D-0057	9a. ORIGINATOR'S REPORT NUMBER(S) FTD-HC-23-722-71
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d. DIA Task No. T71-05-10E	SL: 3262

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY Foreign Technology Division Wright-Patterson AFB, Ohio

13. ABSTRACT

In this book certain problems of the theory and practice of the operation of radar systems are discussed. The methods are directed toward achieving the maximum effective range of a system. Considerable attention is devoted to increasing the reliability of radio electronic equipment. It is one of the important technical problems. This book discusses a wide circle of problems connected with the design, manufacture, and operation of the equipment. Among these problems, increasing the equipment reliability in the design stage has a very important place. Equipment reliability may be increased if it is controlled in the design stage.

DD FORM 1 NOV 65 1473

UNCLASSIFIED
Security Classification

FTD-HC-23-722-71

EDITED TRANSLATION

THEORY AND PRACTICE OF RADAR SYSTEM OPERATION

By: S. M. Latinskiy, V. I. Sharapov, S. P. Ksenz, S. S. Afanas'yev

English pages: 463

Source: Teoriya i Praktika Ekspluatatsii Radiolokatsionnykh Sistem 1970 pp. 1-342

Translated under: F33657-71-D-0057

UR/0000-70-000-000

THIS TRANSLATION IS A RENDITION OF THE ORIGINAL FOREIGN TEXT WITHOUT ANY ANALYTICAL OR EDITORIAL COMMENT. STATEMENTS OR THEORIES ADVOCATED OR IMPLIED ARE THOSE OF THE SOURCE AND DO NOT NECESSARILY REFLECT THE POSITION OR OPINION OF THE FOREIGN TECHNOLOGY DIVISION.	PREPARED BY: TRANSLATION DIVISION FOREIGN TECHNOLOGY DIVISION WP-AFB, OHIO.

FTD-HC-23-722-71 Date 1 Feb 1972

TABLE OF CONTENTS

PAGE

PAGE

PREFACE

One of the characteristic features of modern technology is the strong role played by radio electronics. Radar systems occupy a special place among radio electronic devices. They are, as a rule, complicated radio electronic complexes — "large scale systems". Their normal functioning is unthinkable at present without a scientifically organized operation. In connection with this, the necessity arises for developing a theory for the operation of radio electronic devices, and in particular of radar systems. In recent years, many works devoted to problems of operation theory have appeared. The works of the Soviet scientists N. A. Shishonk, G. I. Vladimirovich, A. I. Shirokov, Ye. Yu. Barzilovich and B. V. Vasil'yev may be considered the main ones.

Reliability and technical diagnostics are clearly connected with the various aspects of operation. The reliability problems were most thoroughly worked out by A. M. Polovko, G. V. Druzhinin, N. M. Sedyakin, A. S. Grunichev, Ya. B. Shor, I. M. Malikov, N. P. Buslenko, A. D. Solov'yev, I. A. Ushakov, and others. A major contribution to the development of the theory of technical diagnostics was made by L. S. Timonen, V. I. Rabinovich, P. P. Parkhomenko, V. V. Karibskiy, I. M. Sindeyev, A. V. Mozgalevskiy, F. Ye. Temnikov, Ye. S. Sogomonyan, and V. I. Perov.

In this book certain problems of the theory and practice of the operation of radar systems are discussed. Besides the systemization of information taken from the literature, original works of the authors and material from lectures delivered by them at a number of scientific institutions are included in the book.

Considerable attention is devoted to investigating methods for objective control, making it possible to detect, in plenty of time, a reduction of the effective range of a radar system. Permissible limits are determined for varying the technical parameters which are responsible for the effective range of a radar system (Chapter II). In Chapter III, problems connected with maintaining the measurement precision of the target coordinates within specified limits are discussed.

Several variations for constructing an efficient procedure for searching out malfunctions are investigated in Chapter IV. These variations take into account both the failure probability and also the work involved in checking the elements. The use of this material makes it possible for the equipment builders and the specialists who operate these radar systems to determine the plan of attack which must be used when a failure appears. To this end, the instructions for the equipment operations must contain, not a catalogue of probable malfunctions and the methods for eliminating them as is the case at present, but instructions for finding the malfunctions. These instructions are search programs that have been developed beforehand and are compiled on the basis of a specific method. The method, applicable to the equipment being used, is designed so that each specialist determines his problems while still studying the radar system, before the appearance of a failure, and not at the moment when it arises. Only a previously well-thought-out sequence of actions eliminates unjustified assumptions and the useless expenditure of labor before establishing such assumptions as false. The method is illustrated by examples of formulating search programs.

In Chapter V the more general problem of constructing various kinds of malfunction search programs with an arbitrarily assigned set of checks is investigated. Such a problem arises when the simultaneous failure of several elements in the system is possible, and also when the failures are caused by abrupt cutoff, short circuits, and parasitic pickups. The various states of the system in these cases are described by tables of state. The system's model represented by an oriented graph is a particular case of this more general description of the system's state. A modern method of mathematical programming is used for the solution of this problem — the so-called method of branches and boundaries. It enables one to construct an optimum program for searching out the malfunctions in the system. However, the solution of practical problems of large dimensionality by this method is possible only on an electronic computer.

One of the fields of application of the method discussed in Chapter V is the construction of malfunction search programs for automatic monitoring devices.

When operating the equipment, it is very important to monitor its technical state. The equipment's state may be evaluated on the basis of the output parameters. In Chapter VI, the problem of selecting a minimum number of parameters characterizing the system operating efficiency is investigated. The checks of the selected parameters in a specified sequence represent a program for checking the system operating efficiency. An algorithm based on applying the general ideas of the method of branches and boundaries is investigated. This algorithm enables one to construct an optimum program for monitoring the operating efficiency.

Another important problem is determining the monitoring periodicity and, on the basis of the monitoring data, making a judgment as to the possible subsequent utilization of the equipment. This problem also is investigated in Chapter VI.

Increasing the reliability of radio electronic equipment is at present one of the important technical problems. This encompasses a wide circle of problems connected with the design, manufacture, and operation of the equipment. Among these problems, increasing the equipment reliability in the design stage occupies a very important place. The equipment reliability may be increased if it is controlled in the design stage.

One of the effective methods for increasing the reliability of the operational process is predicting equipment failures and conducting, on the basis of the resulting information, preventive measures for replacing and restoring elements which have deteriorated parameters. The problem of predicting a change in the equipment during the design and operation is investigated in Chapter VI of this book.

The adjustment of complicated devices represents a sequence of operations for adjusting and tuning the equipment elements. The adjustment is complicated in those cases where there are complex couplings between the equipment elements. One of the problems connected with this is determining an efficient sequence for the adjusting operations.

The process of constructing such programs is investigated in Chapter VII. The method for constructing action programs connected with the adjustment is illustrated by a specific example.

One of the most important and unstable tactical parameters of a radar system is its effective range. The complex tuning of a radar system's high-frequency equipment, which determines this parameter, presents the greatest difficulties. In Chapter VIII, general information concerning the elements of a radar system's high-frequency circuits is given. In connection with this, special attention is paid to the units for adjusting and tuning the high-frequency elements. In the same chapter, methods for tuning the high-frequency elements and the system as a whole are discussed. The methods are directed toward achieving the maximum effective range of a system.

The book was written by a group of authors: the preface, Chapters II and III and §1.1, 1.3, 6.5, and 6.6 were written by S. M. Latinskiy, Chpaters V and VII and §1.2, 1.4, 6.1 - 6.4, and 6.7 were written by V. I. Sharapov, Chapter IV by S. P. Ksenz, and Chapter VIII by S. S. Afanas'yev.

The authors consider it their duty to express their profound thanks to Professor I. V. Brenev for his valuable advice which was used when writing the book, to the candidates of technical sciences K. K. Lyapin, V. I. Yedidovich, V. V. Cherkasov, and Ye. G. Chubarov who participated in the discussion of the manuscripts, and also to I. P. Kuz'min and V. I. Anosov who reviewed the manuscripts.

The authors will appreciate remarks and suggestions concerning the contents of the book from the readers.

1. GENERAL PROBLEMS OF RADAR SYSTEM OPERATION

1.1. The Main Concepts and Definitions

The theory of operation, as is the case of all theories, must be a system of fundamental ideas which generalize experience and reflect the objective relationships governing the evolution of a given field of knowledge. The theory of operation is a new field of science. It arose on the basis of reliability theory, the theory of mass servicing, technical diagnostics, and cybernetics.

The theory of operation is based on a number of fundamental concepts and definitions. By operation, we mean the collection of organized actions for putting the equipment into the operating state, maintaining it in this state, and using it for its assigned purpose with the required efficiency.

Depending on whether the radar equipment is restored during use or not, the equipment may be assigned to one of two groups:

1) equipment for a single use;

2) equipment to be used many times.

The equipment for a single use is used for its assigned purpose only once (for example, the onboard equipment of a rocket). A second use of the equipment is impossible because of its destruction, or it is inadvisable for some other reason. Single-use equipment is not restored during use.

The main tasks performed by the personnel servicing the repeated use equipment — that is, the equipment restored during the operation process — include the following:

— storage of the equipment;
— preparation of it for use;
— using equipment for its assigned purpose;
— technical servicing;
— restoration.

For the unrestorable equipment, the main tasks are: storage of the equipment, preparing it for use, and using it for its assigned purpose.

Let us briefly discuss each type of task which is performed in connection with the operation.

Storage — the set of measures directed toward maintaining the equipment in a prescribed state and guaranteeing its subsequent use for the assigned purpose. In connection with this, the equipment may be located in storerooms, in working areas, or directly at the spots where it is used. This set of measures, on the one hand, consists of creating and maintaining conditions which preserve the equipment, and, on the other hand, of carrying out the necessary monitoring, preventive, and restorative measures.

Preparation — the set of measures which ensure the transition of the equipment from one state to another — for example, from being stored to being used for an assigned purpose, or vice versa.

With this step, putting in and taking equipment out of protective storage, checking and tuning of the equipment, and when necessary its preventive inspection and repair are performed.

Utilization — operation of the equipment in which it fulfills its aims. In military matters, this process carries the designation of combat operation.

Technical servicing — the set of measures directed at maintaining the equipment in an operating state.

For this technical servicing, the following types of tasks are characteristic:

— monitoring-inspection tasks which include measurement and monitoring of the equipment technical parameters and its operating conditions for determining its readiness for use;

— tuning and adjusting tasks, the purpose of which is maintenance of the system's main parameters at the specified level;

— preventive measures which guarantee trouble-free operation of the equipment during a prescribed time interval.

An important factor for increasing the effectiveness of preventive inspections is forecasting (predicting) the equipment failures. The effectiveness of these measures depends on both the equipment construction, and also on the man servicing it.

Serviceability — is the characteristic of the equipment adaptability to servicing. Serviceability is a parameter of the equipment which characterizes it in terms of servicing convenience.

By restoration of the equipment is understood the set of measures whose main object is:

— elimination of the failures which arise in the equipment during its operation (current maintenance);

— elimination of potential failures (preventive maintenance);

— restoration completely or partially of the equipment consumed resources (intermediate or capital maintenance).

Restoration of the equipment lost properties in the most general case comprises the following operations: detecting the failure, searching out the malfunction, replacing the failed elements, tuning and adjusting the equipment, and monitoring the output parameters for checking the quality of repair.

An important property of each piece of equipment is <u>repair suitability</u> which includes the equipment adaptability to prevention, detection, and elimination of failures and malfunctions by technical servicing and maintenance [68]. This property must be built into the equipment during its design and production stage, and be realized in the operational period.

The concepts of "maintenance" and "adjustment" are interconnected. They determine a certain sequence of tasks (operations) which are directed toward eliminating the consequences of a failure.

It is advisable to connect the definition of the maintenance and adjustment concepts with two possible states of the system elements:

— unrestorable,

— restorable.

Thus the interelectrode breakdown of a tube is an unrestorable state of this element. When the tube is in this state, it must be replaced. As an example, the detuning of a receiver's klystron resonator which leads to a decrease of the system effective range must be referred to as a restorable state of this element. By changing the resonator cavity or the voltage on the klystron repeller plate, one is able to restore the operating efficiency of the system. Thus the set of tasks connected with replacing the individual elements of a system and the adjustment of it, which is necessitated by this replacement to bring the system parameters to those values corresponding to the requirements of the technical specifications, may be called the maintenance.

By adjustment we understand the tasks connected with the coupling of the separate subassemblies and assemblies of the system, and with bringing the parameters which characterize their performance to the values which satisfy the requirements of the technical specifications. Sometimes the term "adjustment" is replaced by the term "tuning", which designates the tuning of the various resonance systems.

The operational process is a complicated interrelationship of man and equipment. Therefore, the operation is a concept which refers to the "man - machine" system. Out of the entire set of problems arising with the operation of the system, we will investigate in this book certain problems involved in the technical servicing and serviceability of radar equipment.

The generality of the formal characteristics of complicated systems and the processes which occur in them enables one to use quantitative methods for solving the main operation problems — such as the optimization of the malfunction search process, planning the preventive, tuning, and maintenance tasks, and others. When investigating the operational processes of complicated technical systems, it is convenient to deal not with actually existing systems and processes, but with certain abstract analogs or models of them. In this case, mathematical equipment and a unified approach to the problem may be used for describing and studying the processes and systems which are diversified in their structure, state, and physical nature.

A convenient and graphic method for representing complex systems and the discrete processes taking place in them is the oriented graph. Therefore, in this chapter (§1.4) the main concepts concerning oriented graphs and operations with them are introduced. Without these, it would be impossible to proceed with the investigations of the material in the following sections.

1.2. Systems and Processes

The investigation of various operational problems of complicated technical devices is based on a series of concepts and definitions which characterize the systems and the processes occurring in them. In the first place, "system" and "element" refer to such concepts.

System — this is a collection of elements, whose interconnected functions are coordinated for the purpose of fulfilling some common purpose. In this way, the concept of "system" characterizes the device as a whole. Thus, for example, a radar station is a system. In the operational process, a man, as a rule, is found to have an intimate interrelation with the system. A characteristic example of the relations of "man - machine" is the servicing of some technical device by a man.

Another concept connected with the definition of the separate parts or components of a system is "element".

Element — a part of a system which fulfils an assigned function. The division of a system into an arbitrary number of parts, each of which fulfills a specific function, is nonunique, because the designation "element" cannot be rigidly attached to a specific circuit or part (for example, to an assembly, a stage, a resistor, etc.). Each time it is necessary to stipulate the level at which the division of the system into elements is to take place. In particular, the transmitting, receiving, indicating, and other devices of a radar system may be regarded as its elements which fulfill assigned functions. On the other hand, the various parts (tubes, capacitors, transformers, etc.) which go into the construction of a transmitting device are its elements and at the same time are elements of the radar station. In the "man - machine" system, the elements will be qualitatively different components — the man and the technical device.

When servicing complicated electronic systems, it is important to know their structure and the functioning characteristics. A system consists of interconnected and interrelated elements, whose properties and coupling characteristics are of vital importance for the normal operating efficiency of the system. A system's functioning characteristics are revealed in its interaction with its surroundings and other systems. The structure and functioning characteristics of a system are clearly connected. This connection is due to cause - effect relationships. The presence of these relationships leads us to another important concept — "state".

By the state of a system, we understand the precisely determined behavior of a system which is characterized by a combination of certain of its properties which may be recognized if it returns again. Depending on the specific requirements, this concept may be applied either to the system as a whole or to an element of it. From the effect of external influences or as a result of processes taking place inside the system itself, the system may change from one state to another. The number of various states of a system may be very large.

External influences applied to a system will be called input quantities, and the elements to which they are applied will be called inputs or input connections of the system. In precisely the same way, the effect of the system on its surroundings and on the systems connected with it may be characterized by its output quantities. The system elements from which the output quantities are taken are called the outputs or output connections of the system.

Each element of the system or group of them being investigated as an isolated unit has inputs and outputs.

The evaluation of the state of a technical system is an important aspect of many operational procedures. One must attempt to obtain a precise evaluation of the state, because the correctness of the solution accepted for the subsequent method and form of servicing the system depends on this. In practice, as a rule, the state of any system is evaluated on the basis of a set of input and output quantities. These quantities emerge as the characteristics of the various properties of a system (of an element) which is found to be in this given state, and they are called the parameters.

The following quantities may be the parameters of a system or element: voltage, power, inductance of a circuit, etc. For example, the state of a radar transmitter may be characterized by the radiated power and an electronic tube may be characterized by the characteristic curve.

In order to obtain a precise evaluation of a system's state, a range of acceptable values of the parameter is assigned. The system is called defective if even one of its parameters goes beyond the range of permissible values.

Various forms for describing the state of a system exist. One may, for example, express the state of the system by listing the states of all its elements, or one may indicate the values of all its output quantities. In this case, each of the possible states of the system, as will be shown in Chapter V, may be represented by a state vector. This is especially important in diagnosis, when it is necessary to have a complete inventory of the system's possible states and the quantities characterizing these states.

One may define an event by using the concept of "state". A change of a system's state at a specific moment of time will be called an event. There are various examples of events: the failure of a system element, the disturbance of an adjustment, the deviation of a parameter from the permissible values, etc. The possibility of the appearance of one or another event may be evaluated quantitatively by using, as the numerical measure, the degree of objective possibility of the event's appearance — the probability measure. An event should be investigated as a component of a certain process taking place in the system. In the operation of a system, one may find many examples illustrating this concept. In particular, adjustment, maintenance, diagnosis — these are types of various processes. The majority of operational processes are processes directed toward achieving a specific purpose, and they may be called control processes. Thus, the ultimate purpose of the diagnostic process is to determine that state, from among all the states of the technical system, in which the system is found at a given moment.

In the majority of cases, the operational processes have a discrete character and may be represented in the form of abstract mathematical models. Very frequently the various stages of a process are called steps, and the process itself is called a <u>multistep process</u>.

Depending on the results obtained in each step, the discrete operational process may be determinative or selective (alternative). We call this process determinative if it always leads to one definite result when it is carried out. An example of such a process is the adjustment of some subassembly of radar equipment or the replacement of an element of it. A selective or alternative process is one which leads to one of several results. Among selective processes we can list any of the processes for selecting one or a certain number of results from a large number of possible results. In particular, the malfunction search process in a complicated technical system is a selective process.

1.3. The Parameters of Radar Equipment

The parameters of radar equipment may be subdivided into tactical, technical, and operational-technical.

Tactical parameters directly characterize the effectiveness with which the equipment fulfills its purpose. The tactical parameters of a radar system are: the range for detecting a target (the effective range), the errors in measuring the target coordinates and motion parameters, the resolution, the minimum effective range (the dead zone), and others. These parameters may be monitored as a rule in conjunction with using the radar system. However, in the overwhelming majority of cases it is necessary to know the state of the equipment before it is used. An estimate of its working effectiveness in this case may be made by monitoring the technical parameters which determine in their totality the equipment's tactical parameters.

The technical parameters which characterize the state of several elements of the system will also be called the composite or unified parameters.

Among the composite technical parameters of a radar system are the transmitter power, the receiver sensitivity, the standing-wave ratio, etc.

The operational-technical parameters characterize both the equipment's technical state and also the optimum nature of the methods for organizing the operational process. Such parameters are, for example, the equipment's reliability and longevity. As practice has shown, deterioration of the system's tactical parameters takes place during operation due to breakdown of the tuning and adjustment of the individual devices of the equipment, a change in the parameters of the vacuum-tube and semiconductor instruments, and for other reasons. This is especially strongly expressed on board a ship where the equipment operates under conditions of increased humidity, vibration, jarring, and abrupt temperature changes.

Depending on the purpose of the radar system, one or another of its tactical parameters is the decisive value. However, for the majority of radar types, the foremost value is the effective range.

The problem of maintaining the radar's effective range at the prescribed level arises in technical servicing, since this parameter is very strongly subject to change. Figures 1.1 and 1.2 present histograms of a decreased power of shipboard and airborne radar systems, which illustrate this point. The histograms were compiled on the basis of observations carried out in the U.S.A. during 1943 - 1945 [42]. This period was characterized, as the American specialists state, by the fact that measuring equipment was not used in the everyday routine for monitoring the radar's technical state.

As follows from the histograms (Figures 1.1 and 1.2), the reduction of the radar's power in respect to the rated value reached on the average 14 - 15 dB. If one considers that a reduction of the power by 12 dB decreases the radar system's effective range for

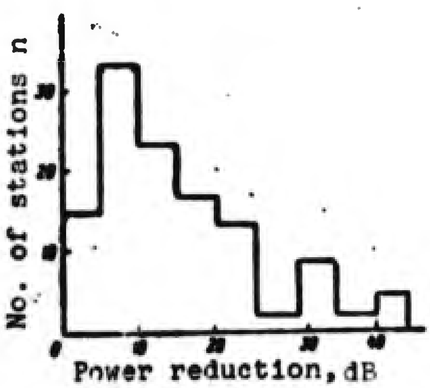


Figure 1.1. Histogram of the decrease of power drop of shipboard and airborne radar systems of the U.S.A. (1943 - 1944)

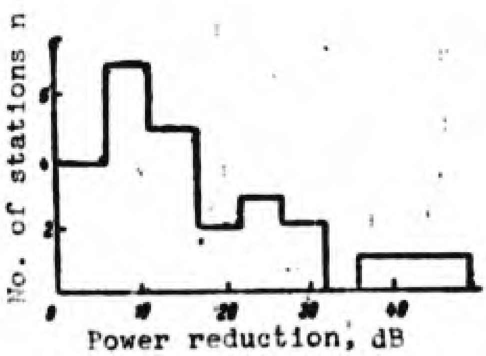


Figure 1.2. Histogram of the decrease of the power drop of shipboard radar systems of the U.S.A. (1945)

detecting a plane by one-half, then it becomes obvious that unsatisfactorily effective operation may nullify the tremendous efforts expended by the designers for achieving optimum tactical parameters.

The histograms shown in Figure 1.1 and 1.2 graphically illustrate the necessity of monitoring and maintaining the radar effective range at the prescribed level during technical servicing. A radar system is an instrument which indirectly measures the target coordinates. The measurements are unavoidably accompanied by errors which quantitatively characterize the measurement precision.

In addition to the detection range, the precision of the target coordinate measurement is one of the main parameters of a radar system. In connection with this, it is important that the errors in measuring the target coordinates do not increase while the radar equipment is being used. However, an operational test of radar systems shows that the coordinate measurement errors increase with time, and if special measures are not taken they may reach intolerable values. Consequently, it is necessary to eliminate, during technical servicing of a radar system, the possibility of an increase in the measurement errors of the target coordinates.

One of the main operational-technical parameters of equipment is reliability. <u>Reliability</u> is the property of a manufactured article to fulfill its assigned functions, preserving its operational characteristics within prescribed limits during a required interval of time or a required service period before failure. The reliability of a product is dependent on its "trouble-free nature", "maintenance suitability", "preservation quality", and also the "longevity" of its parts [68].

Thus, the high reliability of equipment must be achieved when it is created, and it must be maintained in the operational process.

It is possible and proper to influence the reliability characteristics of complicated radio-electronics systems which are designed for repeated use, not only in the planning stage but also during the operational process. Since equipment failures, as a consequence of element imperfections and other factors not taken into consideration, will always occur when the main task of the operator-specialist is conducting a set of measures which will enable the equipment, despite this, to continue to operate. With an efficient servicing system for the equipment, its actual reliability will approach the maximum value calculated in the planning process. But, on the other hand, if a technical servicing system is absent or is constructed on irrational principles, any equipment, even the most reliable, becomes unreliable. Moveover, operational practice shows that more than one-third of the failures arise because of incorrect operation of the equipment [43]. It is necessary to keep in mind that, no matter how reliable the equipment is and how perfect its technical servicing system, it is practically impossible to avoid failures. One of the main problems in the operation of equipment for repeated use is, consequently, maintaining reliability when failures originate.

For this equipment, it is very important to detect the malfunction rapidly and to eliminate it. For radio-electronic equipment, and in particular radar systems, the most complicated problem connected with this is finding the malfunctions.

In order to use the radar equipment effectively, its output parameters must be within the prescribed limits. As was already mentioned, the deterioration of the radar output parameters takes place during the operational process because of the disturbance of the tuning and adjusting of individual devices and elements of the equipment. Therefore, the problem of maintaining the equipment parameters at the required level arises during the technical servicing process. This problem is met by periodic tuning and adjustment of the equipment. An important problem connected with this is determining the optimum sequence of the adjustment operations — that is, determining an efficient program of action.

1.4. Mathematical Formulation

As was indicated, for many practical cases the most convenient and graphic method for representing discrete processes in complicated systems is an orientated (directed) graph.

The concept of a graph, which by itself is very simple, proves to be very fruitful when studying such complicated operations as diagnosis, monitoring, tuning, and adjustment of systems. Before giving a definition of this concept, it is necessary to introduce a number of definitions related to set theory.

Every finite or infinite collection of objects of any nature is a <u>set</u>. For example, the collection of all the assemblies of a radar station forms a set of assemblies; the collection of all the indications of various malfunctions forms the set of these indications, etc. Any of the objects belonging to the set is called and <u>element</u> of it.

A set customarily is designated by capital Latin letters A, B, C, ..., X, and the elements of the set — by small Latin letters a, b, c, ..., x y. The membership of element x to set A is designated by $x \in A$. If x is not an element of set A, then this is written in the form $x \notin A$ or $x \bar{\in} A$. Set C whose elements are a, b, c, d is designated by the symbol C = {a, b, c, d}. Other notation forms for sets also exist. For example, all the elements of set X may be written in the following way:

$$X = \{x_i / i = 1, 2, \ldots, n\}.$$

Such a notation form means that only those elements whose indexes run through the values from 1 to n belong to the set X.

Sets consisting of a finite number of elements are called <u>finite sets</u>. Among the finite sets are the <u>unit set</u> and <u>empty set</u>.

The unit set contains only one element. The empty set does not contain any element, and is designated by $A = \emptyset$.

If A and B are finite sets, then $|A|$ and $|B|$ designate the number of elements contained in each of these sets.

In that case when all the elements of set A belong to set X and certain elements of set X do not belong to set A, this is designated by

$$A \subset X$$

and we say that A is a subset of X, or A is contained in X. If sets A and B match — that is, every element of set A also belongs to set B and vice versa — then this fact is written in the form

$$A = B.$$

In a number of cases when investigating models of systems and processes, the necessity arises for using certain operations with sets. Among the basic operations with sets are: the union, intersection, and difference of sets.

The sets, whose elements belong to set A or set B or to both of them, is called the <u>union</u> (sum) of the two sets A and B. The union of two sets A and B is written in the form

$$A \cup B.$$

If it is necessary to write the union of several sets — for example, the sets A_i the number of which is n — the the following notation is used:

$$\bigcup_{i=1}^{n} A_i.$$

The intersection of two sets is formed by the elements which are common to both sets. The fact that the sets A and B have common elements may be written in the form

$$A\cap B \text{ or } B\cap A.$$

The intersection of n sets A_i is designated by the symbol

$$\bigcap_{i=1}^{n} A_i.$$

If the sets A and B have no common elements, then their intersection will be the empty set, that is

$$A\cap B=\varnothing.$$

The difference of sets — this is the set of elements belonging to set A but not belonging to set B. The difference of sets A and B is designated by $A \setminus B$. The operations of union, intersection, and difference may be given a graphic, geometric interpretation, For example, Figure 1.3 illustrates the operation of the union of the sets

$$A_1=\{a, b\},\ A_2=\{b, c, d\},\ A_3=\{e, g\}.$$

The other operations for the sets B_1 = {a, b, c} and B_2 = {b, c, d, e} are shown in Figure 1.4 and 1.5. The main problem when constructing a mathematical model of a system is to correctly describe the connections and interaction processes between the system elements. In the language of mathematics, this means that it is necessary to establish the relation between the elements of the set.

Among the various kinds of relations the relation of equivalence and relation of order play an especially important role.

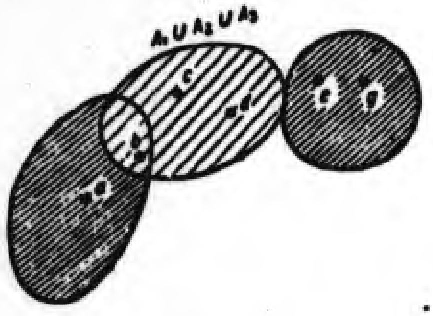


Figure 1.3. Union of sets

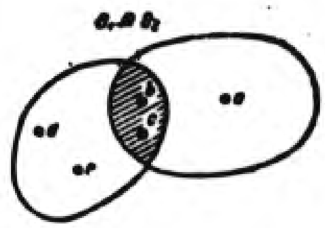


Figure 1.4. Intersection of sets

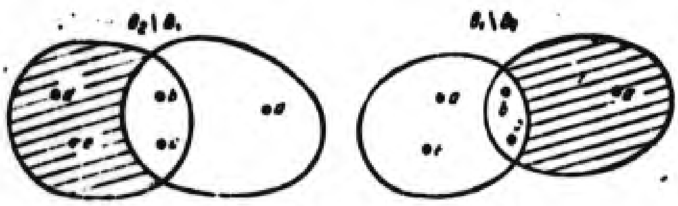


Figure 1.5. Difference of two sets

Very frequently one is obliged to assume certain elements of a set are equivalent elements, since their properties are such that one element may be replaced by another. For example, if in a system there are two elements such that, with the failure of either of them the same malfunction indications appear, then these two elements are equivalent.

The equivalence relation is usually designated by the symbol ≡. If a, b, and c are three arbitrary elements of set A, then an equivalence relation exists between them when the following three properties occur:

1. Each element is equivalent to itself — that is, $a \equiv a$ (the reflexive property).

2. If two elements are equivalent, then it is immaterial which of the elements is considered the first and which the second — that is, if $a \equiv b$, then $b \equiv a$ (the symmetric property).

3. Two elements equivalent to a third are equivalent to each other — that is, if $a \equiv b$ and $b \equiv c$ then $a \equiv c$ (the transitive property).

An equivalence relation between elements of set A enables one to divide this set into subsets. These subsets are disjoint, and each element of set A belongs to one and only one of them.

In practice, one is frequently obliged to deal with the subdivision of a set. For example, when diagnosing a system, the entire set of its elements is divided into two subsets, one of which contains the malfunctioning element and the other subset contains all the elements in good working order.

Another form of relation, as was already indicated, is the relation of order. We encounter this relation every time when we use such terms as: "earlier", "later" or "preceding", "following" or "smaller", "larger". Various symbols may be used in order to designate the various nuances of this relation. Thus, for example, the symbol $>$ designates a strict partial ordering, and the symbol $\geqslant$ designates a partial ordering.

The relation of strict partial ordering is characterized by the following properties:

1. The relation $a > a$ is impossible.

2. If $a > b$ and $b > c$, then $a > c$.

If an equivalence relation is defined between the elements of a certain set, then we frequently use the relation of partial ordering which has the following properties:

1. a $\geqslant$ a.

2. If a $\geqslant$ b and b $\geqslant$ c then a $\geqslant$ c;

3. If a $\geqslant$ b and b $\geqslant$ a then a = b.

The concept of a graph is clearly connected with the relations on a set. The graph is only another method of describing the relations, and enables one to graphically illustrate them. An illustration of an order relation by a graph will be presented below.

The term <u>graph</u> designates a set of points connected by continuous lines. The points which are connected by the lines are called the <u>vertices of the graph</u>, and the connecting lines are called its <u>edges</u>. If the edges of a graph have arrows on them, then they are called <u>arcs</u>, and such a graph is called an <u>oriented graph</u>. A graph is designated by the special symbol G = (X, V) which corresponds to the statement: the graph is a pair of sets consisting of the set of vertices X and a set of arcs V.

It is more convenient to represent a graph geometrically. Thus, for example, Figure 1.6 shows the graph determined by the sets

$$X=\{x_1, x_2, x_3, x_4, x_5\},$$
$$V=\{v_1, v_2, v_3, v_4, v_5, v_6\}.$$

Very frequently the symbol (x_i, x_j) is used for designating the arcs of a graph. This symbol means that an arc runs from vertex x_i to vertex x_j adjacent to it. Therefore, the set of arcs of the graph shown in Figure 1.6 may be written as follows:

$$V=\{(x_1, x_2), (x_1, x_3), (x_2, x_4), (x_2, x_5), (x_3, x_5), (x_4, x_5)\}.$$

By listing the arcs of a graph in this way, we have established a rule for connecting the vertices of the graph by arcs. A special term "<u>mapping</u>" exists in set theory for this rule. The mapping of a graph usually is designated by the symbol Γ, and the graph is written in the form $G = (X, \Gamma)$. The latter notation for a graph is equivalent to the notation $G = (X, V)$. Let us define, for example, the mapping of the graph shown in Figure 1.6. The mapping Γ of this graph is characterized by the following equalities:

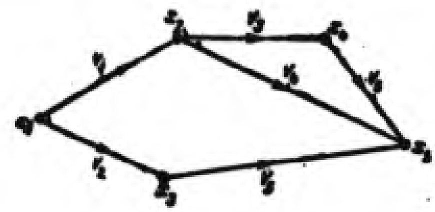


Figure 1.6. An oriented graph

$$\Gamma x_1 = \{x_2, x_3\}; \ \Gamma x_2 = \{x_4, x_5\}; \ \Gamma x_3 = \{x_5\};$$
$$\Gamma x_4 = \{x_5\}; \ \Gamma x_5 = \varnothing.$$

The last equality $\Gamma x_5 = \varnothing$ means that vertex x_5 is an exit vertex of the graph, and it does not have a mapping to any other vertex. In precisely the same way, one may say that no other vertex is mapped to the vertex x_1, or saying it another way: the inverse mapping Γ^{-1} of this vertex is an empty set — that is, $\Gamma^{-1}x_1 = \varnothing$. In talking about such a vertex, we refer to it as an entrance of the graph. A graph may have one or several entrance and exit vertices.

By investigating the graph (Figure 1.7), we see that from an entrance vertex x_1 one may reach an exit vertex x_5 by traversing in succession certain of its vertices and arcs.

The sequence of arcs in which the end of one arc is the beginning of another arc is called a <u>path</u>. Thus, a path is defined either by the sequence of its arcs, or by the sequence of the vertices of these arcs. A path is designated by the symbol $\mu = [x_1, x_2, \ldots, x_k]$.

In particular, from vertex x_1 of the graph shown in Figure 1.6 there are three paths to vertex x_5, more precisely:

$$\mu_1 = [x_1, x_2, x_4, x_5],$$
$$\mu_2 = [x_1, x_2, x_5],$$
$$\mu_3 = [x_1, x_3, x_5].$$

If in the graph there is a path which passes through all the vertices, then it carries the designation "Hamiltonian Path". In particular, the graph of Figure 1.7 contains a Hamiltonian Path which is marked on the drawing by the heavy lines.

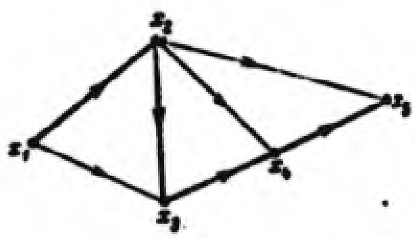


Figure 1.7. An oriented graph having a Hamiltonian Path

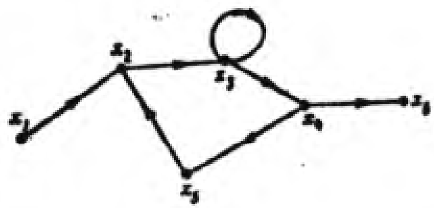


Figure 1.8. An oriented graph having a contour and a loop

When the starting and ending vertices of the path coincide, a contour is formed. In an unoriented graph, this concept corresponds to cycle. A contour formed by an arc of the type (x, x) is called a loop, because a loop connects a vertex with itself. Figure 1.8 shows a graph in which the path $[x_2, x_3, x_4, x_5, x_2]$ forms a contour, and the arc (x_3, x_3) is a loop.

Earlier it was mentioned that the concepts of "relation" and "graph" are clearly interconnected. Actually each oriented graph defines a certain relation on the set of its vertices. For example, the relations $\leqslant$ and $<$ are order relations which are defined by the graph G. When $a \leqslant b$, we say that vertex a precedes vertex b.

This relation may also be written in the form b $\geqslant$ a, and we say that vertex b follows vertex a. If one has in mind the relation of strict ordering, then with the notation a < b we say that the vertex a strictly precedes vertex b. In exactly the same way, when we write b > a we say that vertex b strictly follows vertex a.

For the purpose of illustrating these relations, graphs are shown in Figure 1.9 which correspond to the relations of partial and strict partial ordering. As is seen from the figure, the difference between the concepts of partial and strict partial ordering is expressed in the fact that in the first graph (Figure 1.9a) there are

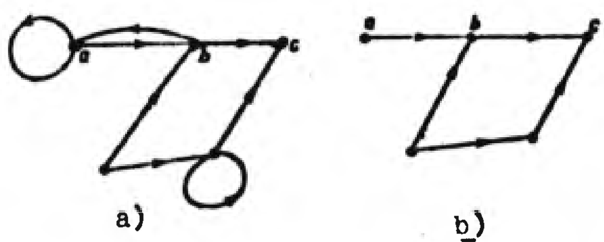


Figure 1.9. Relation of order defined by a graph: a — a graph of partial ordering; b — a graph of strict partial ordering

loops which correspond to the relation a $\geqslant$ a and a contour defining the relations a $\geqslant$ b and b $\geqslant$ a, and in the second graph (Figure 1.9b) there are no loops or contours.

The order relation defined by a graph may be evaluated quantitatively. For this purpose, we introduce precedence numbers on the graph in the following way. Let us assign to each vertex x of the graph a number equal to the number of vertices which precede it. Let us note that precedence numbers may be introduced only on a graph of strict partial ordering which does not have contours and loops. In such a graph — for example, the one shown in Figure 1.10

— each of its vertices is marked with a precedence number corresponding to it. Vertices x_1 and x_2 are entrance vertices and no other vertex precedes them; therefore, each of them must be assigned zero. Furthermore, only one vertex x_1 precedes the vertices x_3 and x_4, and consequently each of them is assigned the number 1. Vertex x_5 is preceded by three vertices x_1, x_2, and x_4, and vertex x_6 is preceded by five vertices x_1, x_2, x_3, x_4, and x_5.

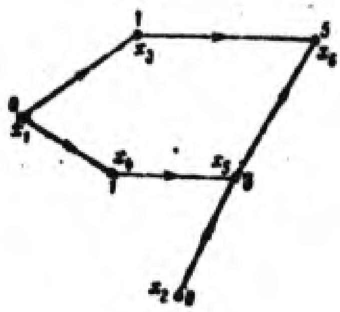


Figure 1.10. Precedence numbers on an oriented graph

The precedence numbers to a certain degree characterize the structure of an oriented graph, and henceforth will be used by us when investigating certain diagnostic problems. In connection with this, it is more convenient to use precedence numbers in a slightly altered form, by increasing each precedence number by one. The numbers obtained in this way will be called reduced precedence numbers or precedence indexes.

The numbering of a large number of vertices is a laborious process. This difficulty may be avoided if one uses the transitive property which establishes the connection between the precedence numbers and the arrangement of the arcs in a graph. Let us introduce for this purpose the concept of the semi-degree of emergence and semi-degree of approach. The number of arcs which emerge from a certain vertex x is called the semi-degree of emergence of this vertex x; the number of arcs approaching the vertex x is called the semi-degree of approach of vertex x. For example, in the graph shown in Figure 1.10, vertex x_5 has an approach semi-degree equal to two, and an emergence semi-degree equal to one.

Now let us use the transitive property in order to construct a new graph for graph G. This graph we will call the <u>transitive closure</u> of graph G, and we will designate it by G^T. The graph G^T has the property that, for each pair of arcs (a, b) and (b, c), there is a closing arc (a, c).

In every case, an oriented graph may be made a transitive graph by adding to each pair of adjacent arcs a closing arc. Thus, for example, Figure 1.11 shows the transitive closure G^T constructed for the graph of Figure 1.10.

Another important property of the graph G^T, which establishes the connection between the precedence number and the arrangement of the arcs in it, is that the approach semi-degree of each vertex equals the precedence number. It is not difficult to verify this by investigating the graph G^T shown in Figure 1.11.

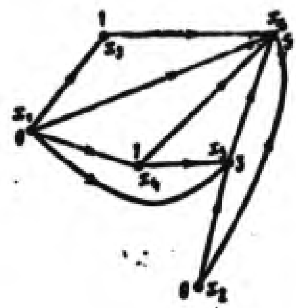


Figure 1.11. Transitive closure of the graph shown in Figure 1.10

When studying certain problems connected with the application of graphs, it is convenient to use the formal properties of matrices. A square matrix $A = (a_{ij})$ with n rows and n columns (that is, the dimensions $n \times n$) is called a <u>matrix of adjacency</u> of the graph G consisting of n vertices x_i ($i = 1, 2, \ldots, n$) if its general element $a_{ij} = 1$ when there is an arc in the graph connecting the vertices x_i and x_j.

Let us illustrate the definition of an adjacency matrix with an example. Figure 1.12 shows a graph which corresponds to the adjacency matrix

$$
\begin{array}{cc}
\begin{array}{cccccc} j=1 & 2 & 3 & 4 & 5 & 6 \end{array} & \\
A=\begin{bmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} & \begin{array}{c} i=1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\end{array}
$$

Let us note that, in the adjacency matrix of the graph, the sum of the elements of the i^{th} row equals the emergence semi-degree of the vertex, and the sum of the elements of the j^{th} column equals the approach semi-degree of the vertex x_j.

In the future when investigating a number of problems, a special kind of graph called a <u>tree</u> will frequently be used. By tree, we understand a graph, not containing contours, in which not a single arc approaches the vertex x_1 called the <u>root</u>, and just one arc approaches every other vertex.

The graph shown in Figure 1.13 is a tree. In this graph from each vertex, except for vertices x_5, x_6, x_7, x_8, and x_9, two arcs emerge. The vertices x_5 and x_6, x_7, x_8, and x_9 from which not a single arc emerges are called <u>suspended vertices</u>. The remaining

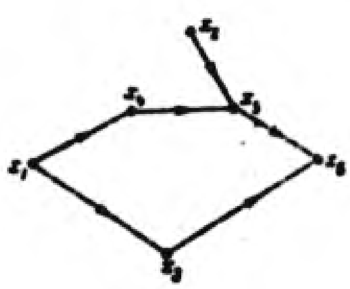


Figure 1.12. An oriented graph for the construction of an adjacency matrix

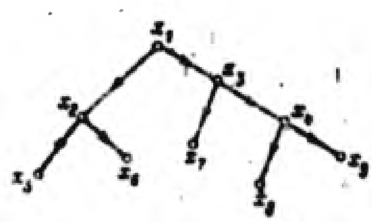


Figure 1.13. A binary tree

vertices which are not extended vertices we will call <u>internal vertices</u>. Trees, in which each internal vertex has an emergence semi-degree of two, frequently are called <u>binary trees</u>. A tree is a distinctive geometric way of representing logical possibilities which may appear as a result of conducting a test. Therefore, such graphs have found broad application in diagnostics for representing the various programs for discovering malfunctions and monitoring the operating efficiency.

2. MAINTAINING THE EFFECTIVE RANGE OF RADAR STATIONS

One of the main problems in the operation of radar stations is maintaining the station's effective range at a level which does not drop below the prescribed level. It is a necessary prerequisite to obtain information concerning the change of this tactical parameter. As a consequence, let us begin the investigation of the problem with an analysis of the possible methods for monitoring a radar station's effective range.

2.1. Methods for Monitoring a Radar Station's Effective Range

Up the present, there is still no single and established method which enables one to monitor a radar's effective range with sufficient objectivity and reliability. The effective range of radar stations may be monitored by the following basic methods:

— by direct measurement of the range to an object (a target);
— based on the strength of the signal reflected from a local object;
— by means of a standard reflector;
— by the separate measurement of the technical parameters responsible for the radar's effective range;
— by means of a monitoring resonator (an echo chamber);
— by means of a short-circuit delay line (the self-calibration method).

Direct Measurements of the Radar's Effective Range

Direct measurements of a radar's effective range may be made both from especially assigned (specified) objects and also from random targets. This method has a number of substantial shortcomings. The systematic checking of stations by this method entails a large expenditure of time and diverts the test objects from their main assignments.

Because a radar's effective range depends on the radio wave transmission conditions, its checking must be carried out under conditions of normal radar observation powers. The target's effective reflecting surface, which, in turn, depends on the character and aspect of the target, also affect a station's effective range. Therefore, in order to obtain comparable results, it is necessary to measure the range of the same targets observed with the same aspects. This requirement greatly complicates the measurements.

Sometimes it is impossible to conduct direct measurements for organizational reasons because even before starting to use a station, it must be verified that it has the required effective range.

The results of the range measurements to a target depend on many factors which vary within considerable limits and are independent of the station's technical condition. Moreover, the direct measurement method does not enable one to continuously monitor a radar station's effective range or to clarify the reasons leading to its reduction during operation. Thus, this method is not an optimum one for monitoring the constancy of a station's effective range during operation.

The Method of Measuring Signal Strength from a Local Object

Stationary local objects may be found within the radar's effective zone. Signals from the objects are clearly observed on the screen of the station's indicator. Chimneys, lighthouses, etc., may serve as such objects.

Sometimes it is assumed that one is able to make a judgment as to a station's effective range from the extent to which the amplitude of the signal reflected from a local object exceeds the receiver noise level. However, this is not true, because the signal amplitude depends not only on the station parameters, but also on the radar observation conditions and on interference phenomena (local objects, the reflections from which are not subjected to interference, are very rarely encountered). Moreover, measurement of the signal level can be performed only very approximately.

Frequently, deterioration of radio wave propagation conditions leads to decreased signal amplitude of the object, and creates the false impression of a reduction in the radar effective range. In this case, an attempt to fine-tune the station leads to the opposite effect — detuning it. Consequently, this method cannot be considered acceptable for monitoring the radar effective range during operation.

The Method of Measuring with the Help of a Standard Reflector

The essence of this method consists of the fact that the radar effective range may be determined from the extent to which a signal exceeds the noise from a nearby "standard" target. A standard reflector may be used as such a target. The reflector is set up near the station. Because of this, radio wave propagation conditions have practically no influence on the signal strength. In order to decrease the signal fluctuation and its masking as a consequence of reflection from waves or from local objects, it is necessary that the reflector signal be considerably larger than the receiver noise. However, in this case with a linear sweep on the indicator, it is impossible to determine precisely the amount by which the reflected signal exceeds the noise level, since the signal will be limited because of the receiver saturation. It is necessary to reduce the power of the radiated and received signals by introducing attenuation (a high power attenuator) into the radar circuit. The power of the signal sent to the antenna from the transmitter and the power of the reflected signal at the receiver input is reduced when this is done. By regulating the attentuation magnitude, one is able to reduce the station power potential artificially, and the reflected signal is brought to the receiver noise level. The attenuation introduced by the attenuator will characterize the station's performance efficiency.

One of the advantages of such a method for monitoring the effective range is that the measurements are practically independent of the radio wave transmission conditions. One of the defects is the necessity of inserting a special high-power attenuator in the station's high-frequency circuit. For shipboard radars, such a monitoring method must be used while the ship is at anchor or at a base which naturally limits its value.

The Method of the Separate Measurement of the Technical Parameters Determining the Radar's Effective Range

The essence of this method is that the technical parameters which determine the radar's effective range are measured separately with instruments. The availability of a group of measurement instruments whose measurement errors are found to be within acceptable limits and the availability of connection elements between the radar and these instruments are necessary conditions for conducting such measurements.

The main advantages of this method are:

1. Monitoring of the radar's effective range does not require diverting test objects from their assignments.

2. The measurement results do not depend on the radar observation conditions when the measurements are made.

3. A radar's technical parameters may be monitored while the radar is operating without radiation into space.

4. It is possible to detect, during operation, a deterioration of the radar technical parameters and, consequently, to carry out the necessary preventive measures for maintaining the parameters at the prescribed level.

5. It is possible to monitor the technical state and tuning of all the elements which determine the station effective range and, consequently, to locate the elements leading to a reduction of the range.

The shortcomings of this method are:

1. The necessity of having a large number of transferable measuring instruments.

2. The considerable time required for performing the measurements.

The use of built-in measuring equipment to a considerable extent eliminates these shortcomings.

Separate measurement of the technical parameters is one of the main methods for monitoring a radar's effective range, especially when carrying out preventive inspection. It is investigated in more detail in § 2.2.

Monitoring a Station's Effective Range by a Monitoring Resonator

Radar effective range may be monitored with the help of a monitoring resonator which is either an integral part of the station or is attached to it. A monitoring resonator enables one to approximately evaluate the quality of radar performance. Simplicity of use, the possibility of monitoring the station's main parameters, and the small overall size and weight make a monitoring resonator an irreplaceable instrument for daily monitoring of station performance. The device and principle of operation of a monitoring resonator and also the method for monitoring the effective range by means of a monitoring resonator are investigated in § 2.7.

The Self-Calibration Method (A Short-Circuit Delay Line) [51]

A stable short-circuit delay line is connected to a radar waveguide channel through a directional coupler. Multiple reflection of part of the main pulse energy takes place in the line. Each time the pulse reaches the entrance of the short-circuit line, part of the energy enters the receiver channel. When this happens, calibration pulses are observed on the screen of the indicator with the linear sweep. The number of pulses indirectly characterizes the station effective range. This method has common features with the method of monitoring a radar's effective range by means of a monitoring resonator.

In both cases, a monitoring signal is introduced into the station circuit. The parameters of this signal correspond to the parameters of the transmitter high-frequency pulse. The radar effective range is monitored on the basis of the ratio of the monitoring signal voltage to the noise at the receiver output.

The self-calibration method is investigated in more detail in § 2.7.

2.2. Technical Parameters Responsible for the Radar Effective Range

As was mentioned above, separate measurement of the technical parameters most objectively and completely ensures that the radar effective range is constant. The question arises as to which measurements must be performed systematically in order to monitor a station's effective range.

Let us turn to the equation for the maximum range of radar detection:

$$D_{\text{макс}} = \sqrt[4]{\frac{1}{m_{\text{c}}}\frac{P_{\text{и}}}{P_{\text{пр}}}\eta^{2}\frac{G^{2}\lambda^{2}}{(4\pi)^{2}}\frac{\sigma_{\text{ц}}}{4\pi}}\,V = D_{0}V\text{ м}, \tag{2.1}$$

where m_s is the coefficient which determines the extent to which the received signal exceeds the set noise necessary to achieve a given detection probability;

P_p is the transmitter pulse power in watts;

P_{rec} is the actual sensitivity of the receiving device in watts;

G is the antenna directive gain with respect to an isotropic radiator;

σ_t is the object's effective reflecting area in m^2;

λ is the wavelength in m;

η is the efficiency of the antenna-waveguide channel in transmission (reception);

V is the field attenuation function;

D_0 is the station effective range for propagation of radio waves into free space and for no absorption of energy in the atmosphere.

In order to take into account the actual radio wave propagation conditions, the effective range in free space D_0 is multiplied by the field attenuation function V. The field attenuation function depends on the heights of the antenna and the target, on the distance to the target, the wavelength, the kind of polarization, the soil, and the extent of refraction and absorption of the radio waves in the troposphere.

In a number of cases, in order to determine the sensitivity of the receiving devices, instruments which measure the receiver noise factor are used. If the receiver noise factor is known, its actual sensitivity may be determined in the following way:

$$P_{пр} = kT_0 F\Delta f v, \tag{2.2}$$

where F is the receiver noise factor;

k is the Boltzmann constant which equals 1.38×10^{-23} joule · · degree^{-1};

T_0 is the absolute temperature in °K;

Δj is the receiver's pass band;

v is the discrimination coefficient.

For the terminal device — that is, the station indicator — the discrimination coefficient may be determined from the expression

$$v = 3\frac{\tau\Delta f}{\sqrt{F_{ш}}}\left(1+\frac{1}{\tau\Delta f}\right)^{2}. \tag{2.3}$$

Substituting the value for P_{rec} from Expressions (2.2) and (2.3) into Expression (2.1), we obtain

$$D_{max} = K\sqrt[4]{\frac{P_p \tau_p^2 G^2 \lambda^2 F_p^{1/2} \sigma_{ts}}{P(\tau_p \Delta f + 1)^2}}, \tag{2.4}$$

where τ_p is the pulse duration;

F_p is the pulse repetition frequency;

K is the coefficient which takes into account the effect on the station effective range of the parameters which cannot be changed during station operation.

Let us evaluate the stability of radar technical parameters in Expressions (2.1) and (2.4) in order to determine the parameters to be monitored.

The Power Drop of a Radar Station

The quantity $\Pi_p = \frac{P_p}{P_{rec}} \cdot \tau_p^2 \frac{G^2 \lambda^2}{(4\pi)^3}$ is called the power potential of a radar station.

The constancy of this quantity during operation indicates that the effective range of the radar is the same for the same target and for unchanged radio wave propagation conditions.

A deviation of the quantities λ and G from their normal value during the operational process can be neglected, because with normal functioning of the transmitter, the relative change of its frequency is small. Since the quantity $\frac{G^2 \lambda^2}{(4\pi)^3}$ remains constant for a station during operation, it is sufficient to monitor only the stability of the quantity $\frac{P_p}{P_{rec}} \cdot \tau_p^2$. This means that it is necessary to measure the transmitter power and receiver sensitivity in such a way that, when making the measurements, the electromagnetic field passes through the antenna-waveguide channel. In other words, the antenna-waveguide channel efficiency would be considered in both transmission and reception.

The coupling of the radar with the measurement instrument must be accomplished by means of an antenna attached to the instrument. Such a measurement method is complicated, because it requires precise mounting of the measurement instrument antenna with respect to the radar antenna. Therefore, in the operational process only measurements of the transmitter power P_p and the receiver sensitivity P_{rec} can be performed, and the antenna-waveguide channel efficiency must be monitored by another method.

The ratio

$$H = \frac{P_p}{P_{rec}} \tag{2.5}$$

is called the <u>power drop of the radar station</u>.

It is convenient to express the value of a station power drop in logarithmic values — decibels. The power in relative values is determined by the expression

$$P_{dB} = 10\lg\frac{P}{P_1},$$

where P is the absolute power value;

P_1 is the initial power level;

P_{dB} is the power expressed in decibels above or below the selected initial power level.

The initial level may be selected arbitrarily; all the calculations are conducted with respect to it. In radar, initial levels of 10 microwatts, 1 milliwatt and 1 watt are usually used. Thus, for example, in the combination test instrument for monitoring radar parameters (a radar-tester) GK4-14, a power of 1 milliwatt is taken as the initial level. In the instrument GK4-3, a power of 10 microwatts was chosen.

It is also convenient to express the power drops of radar stations in these units. Relationship (2.5) thus takes the following form:

$$H = P_p - P_{rec.}$$

Here, all the quantities are expressed in decibels.

Since the receiver sensitivity in relative units is expressed by a negative number, the power drop in relative units equals the sum of the absolute values of the transmitter power and the receiver sensitivity. The power drop is a very unstable quantity. This follows especially from the histogram presented in Figure 1.1.

For propagation of radio waves in free space and for no attenuation ($V = 1$), the effective range and the power drop of a station are connected by a simple relationship as follows from Expression (2.1). This relationship is shown graphically in Figure 2.1,a. As is seen from the figure, a reduction of a radar's power drop by 2 dB corresponds to a 10% decrease of its effective range for airborne targets. For a reduction of the power drop by 12 dB, the range decreases by 50%. When detecting surface targets, the effective range and the power drop of the station are connected by a complicated relationship (Figure 2.1,b). The effective range corresponding to the normal value of the station power drop is assumed to be 100% [42]. The character of the change of the radar effective range (Figure 2.1,b) depends on the size of the object. This is explained by the fact that in one case detection of the object takes place in the half-shadow region and in the other case in the illuminated region. The reduction of the power drop, as follows from Figure 2.1,b, also substantially affects the detection range of surface objects.

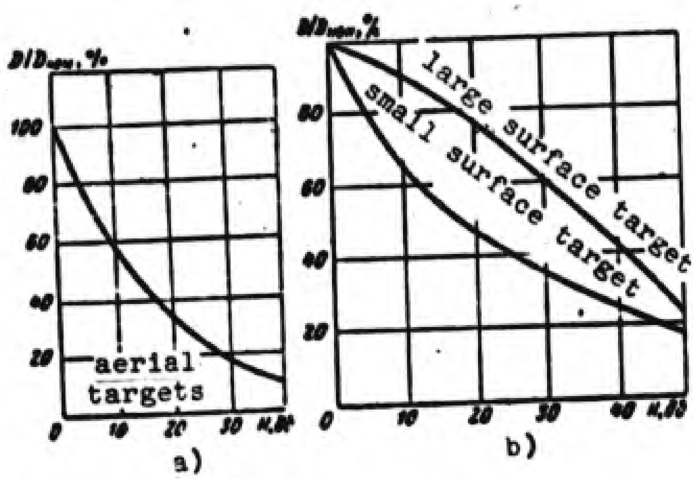


Figure 2.1. A radar's effective range as a function of the power drop:

a - aerial targets ($V = 1$); b - targets on the water's surface ($V \neq 1$).

Consequently, checking the power drop is a necessary condition for monitoring the constancy of the radar effective range. It is necessary to monitor both the receiver sensitivity and the transmitter power.

Practice shows that the change of the receiver sensitivity without proper monitoring and preventive measures may reach 10 - 12 dB and the change of the transmitter power by 3 dB.

Receiver Sensitivity

The receiver sensitivity is defined by the smallest power of the received signals at which a definite ratio between the received signal power and the internal noise is obtained at the output. By convention the smallest input signal power taken for the sensitivity value is that at which the presence of a signal is barely distinguishable over the internal noise background at the output.

The receiving device sensitivity, as is known, depends on the level of the noise in the receiving device and the level of the external noise at its input. In the centimeter, decimeter and meter wave ranges, the level of external noise is relatively small and the sensitivity is determined basically by the receiver internal noise. Thus, the receiver internal noise level in these ranges is the determining measure of sensitivity.

The magnitude of a receiver internal noise is characterized by the noise factor.

A receiver noise factor F is a number which shows how many times larger the ratio of the signal power to the noise power at the receiver input is than the ratio of these powers at the output of the receiver linear section:

$$F = \frac{S_1 N_2}{S_2 N_1}, \tag{2.6}$$

where S_1 and N_1 are the signal power and noise power at the receiver input;

S_2 and N_2 are the signal power and noise power at the receiver output.

The expression for the noise factor may be written also in another form:

$$F = \frac{P_{sog}}{kT_0 \Delta f}, \tag{2.7}$$

where P_{sog} is the power generated at the receiver input by an equivalent noise source.

From this expression it follows that the receiver noise factor is a number which characterizes the noise power at the input of the receiver connected to the equivalent of the antenna. It is expressed in units of kT_0 per hertz of the frequency band.

If the receiver noise factor is known, the actual sensitivity may be determined from Equation (2.2). Since the sensitivity also depends on the receiver pass band, let us briefly analyze the stability of this parameter.

A Receiver Pass Band

One of the parameters affecting a station effective range is the receiver pass band, as follows from (2.4). The instability of a receiver pass band leads to a change of the signal/noise ratio since for each pulse duration there corresponds an optimum value of the receiver pass band.

Let us determine to what degree the instability of a receiver pass band is able to affect the station's effective range. The connection between these quantities may be obtained if one investigates the reaction of the receiver to a change of its pass band. In order to simplify the problem let us assume that the pulses supplied to the receiver input have an ideal rectangular shape and the rectangular pass band of the receiver corresponds to it.

For oscillations which have the same frequency during a pulse, the frequency spectrum of an individual radio pulse with the duration τ_p is described by the expression

$$P(\omega)=\left[\frac{\sin(\omega-\omega_0)\frac{\tau_p}{2}}{(\omega-\omega_0)\frac{\tau_p}{2}}\right]^2,$$

where $\omega_0=2\pi f_0$; f_0 is the oscillation frequency of the superhigh frequency oscillator.

Figure 2.2 shows the main lobe of the frequency spectrum of an individual pulse with duration τ_p and the optimum pass band of the receiver expressed in dimensionless units.

For the case when the receiver pass band has an optimum value, the signal power at the receiver output will equal, according to Figure 2.2

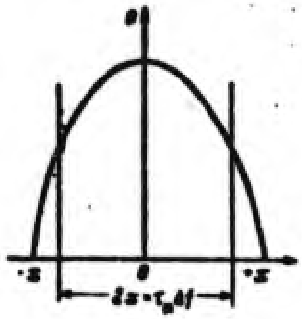


Figure 2.2. Determining the signal's power at the receiver's output with an optimum pass band of the receiver.

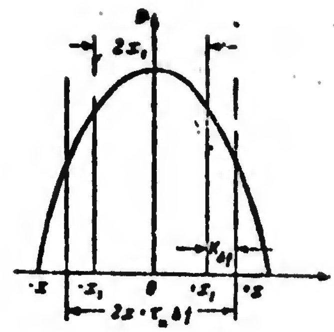


Figure 2.3. Determining the decrease of the signal's power at the receiver's output because of a change of the receiver's pass band.

$$P_0 = \int_{-x}^{x} \frac{\sin^2 \pi u}{(\pi u)^2} du = 2\int_{0}^{x} \frac{\sin^2 \pi u}{(\pi u)^2} du = 2P_x,$$

where $2x = \tau_p \Delta f$ is a dimensionless quantity;

τ_p is the pulse duration in microsecs;

Δf is the receiver pass band in Mc.

Integrating this expression, we obtain

$$P_0 = -\frac{1}{\pi}\frac{\sin^2 \pi x}{\pi x} + \frac{1}{\pi}\int_{0}^{2\pi x} \frac{\sin t}{t} dt,$$

where $\int_{0}^{2\pi x} \frac{\sin t}{t} dt = si(2\pi x)$ is from a table of integrals.

With a deviation of the receiver pass band from the normal value by an amount $d\Delta f$, the signal power at the receiver output will equal, according to Figure 2.3,

$$2P_{x_1} = 2\int_0^{x_1} \frac{\sin^2 \pi u}{(\pi u)^2}\, du, \qquad (2.8)$$

where

$$x_1 = x - K_{\Delta f};\ K_{\Delta f} = x\frac{d\Delta f}{\Delta f}.$$

hence $x_1 = x\left(1 - \frac{d\Delta f}{\Delta f}\right)$.

The decrease of the signal power at the receiver output as a consequence of a change of the receiver pass band is

$$\frac{\Delta P}{P_x} = \frac{P_x - P_{x_1}}{P_x} = 1 - \frac{P_{x_1}}{P_x}. \qquad (2.9)$$

Figure 2.4 shows a graph of P_x as a function of x. This same graph represents P_{x_1} as a function of x_1.

The graphs of Figure 2.5 characterize the decrease of the signal power at the receiver output due to a possible change of the receiver pass band during operation.

Intermediate frequency amplifiers change their pass band by several percent during operation. As follows from Figure 2.5, the decrease of the signal power at the receiver output in connection with this is small and does not exceed 0.3 dB.

Thus, in order to check the receiver sensitivity it is sufficient to monitor its noise factor, because the receiver pass band and the discrimination coefficient remain practically constant during the operational process.

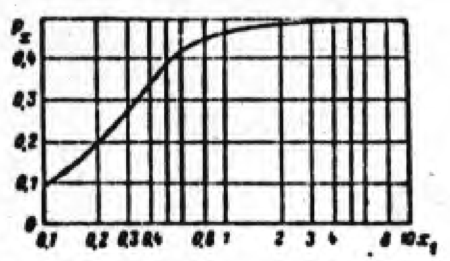


Figure 2.4. A graph of P_x as a function of x.

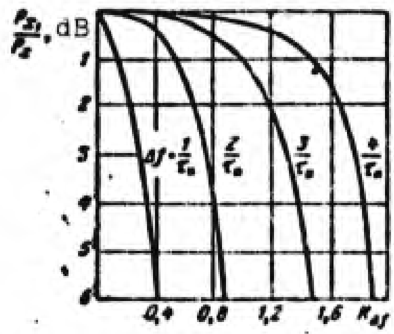


Figure 2.5. Graphs of $\frac{P_{x_1}}{P_x}$ as a function of $K_{\Delta f}$ for $\Delta f = 1/\tau_p$, $2/\tau_p$, $3/\tau_p$ and $4/\tau_p$.

The Frequency Energy Spectrum of Magnetron Oscillator Pulses

The relative energy distribution of the radiated pulses over the frequency is called the energy spectrum of frequencies of an oscillator pulse.

With pulse modulation of high-frequency oscillations, the frequency spectrum of the pulse being generated depends on the pulse duration and the character of the change of the oscillation amplitude and frequency during the pulse.

A periodic sequence of radio pulses has a line spectrum (Figure 2.6). The width of the spectrum main lobe and the ratio of the amplitude of the side lobes to the amplitude of the main lobe are the most important characteristics of the spectrum.

The function P(ω) = 0 when $(\omega_0-\omega)\frac{\tau_p}{2}=n\pi$, where n = ± 1, 2, 3, ...

Hence, it is possible to obtain an expression for the width of the spectrum's main lobe

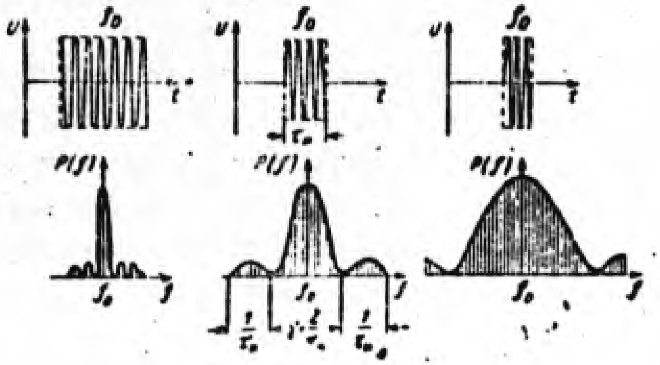


Figure 2.6. The energy spectra of frequencies of a periodic sequence of rectangular pulses of various duration with a fixed oscillation frequency during the pulse.

$$\gamma = 2(f_0 - f_1) = \frac{2}{\tau_и},$$

where f_1 is the spectrum frequency corresponding to the first minimum of the function $P(\omega)$.

Thus, for example, with a pulse duration of $\tau_p = 1$ microsec the width of the spectrum's main lobe equals 2 Mc. The main energy of the pulse also is concentrated in this frequency band. All this is correct if the high-frequency pulse has an exact rectangular shape, if there is no parasitic modulation frequency, and if the oscillator operates with a precisely matched load.

In order to obtain the station's optimum effective range and a high-quality reproduction of the pulse shape, the optimum value of the receiver pass band must correspond to each spectrum width.

Expressions (2.1) and (2.4) in their explicit form do not reflect the effects of the frequency energy spectrum of the oscillator pulses on the radar effective range. However, it is impossible to ignore the affect of this parameter on the effective range of radar having a magnetron oscillator as the superhigh frequency oscillator. The spectrum of a magnetron oscillator does not remain constant during the operational process. The amplitude, shape, and duration of the modulating pulse, the magnetic field intensity, and the character of the lobe have an influence on it.

If the main power of the pulse being generated fits into a frequency band equal to the receiver pass band, then the station effective range will be a maximum. If part of the power of the generated pulse lies outside this band, the power of the reflected pulse will be wasted, and consequently, the radar effective range will be reduced. Similarly a distortion of the spectrum, in which the power of the pulse belonging to the receiver's pass band is reduced, leads to a reduction of the station effective range.

The decrease of the signal power at the receiver output due to broadening of the energy spectrum of the frequencies is also described by Expression (2.9).

In this case, we have

$$x_1 = x - K_\gamma,$$

where $K_\gamma = \Delta f \frac{\tau_p}{2} \frac{d\tau_p}{\tau_p} = x \frac{d\tau_p}{\tau_p}$;

$\frac{d\tau_p}{\tau_p}$ is the relative change of the width of the spectrum's main lobe which corresponds to the relative change of the pulse duration.

Since the width of the spectrum's main lobe is $\gamma = \frac{2}{\tau_p}$, then $\pm\frac{d\gamma}{\gamma} = \mp\frac{d\tau_p}{\tau_p}$. Consequently, $x_1 = x\left(1 \pm \frac{d\gamma}{\gamma}\right)$.

In Figure 2.7 the graphs of P_{x_1}/P_x as a function of K_γ are presented for $\tau_p = \frac{1}{\Delta f}, \frac{2}{\Delta f}, \frac{3}{\Delta f}, \frac{4}{\Delta f}$. These graphs are similar to the graphs of Figure 2.5.

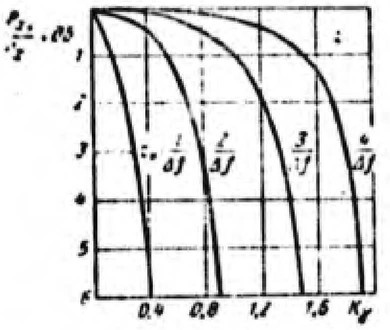


Figure 2.7. Graphs of $\frac{P_{s_1}}{P_s}$ as a function of K_γ for τ_p = $= 1/\Delta f,\ 2/\Delta f,\ 3/\Delta f,\ 4/\Delta f.$

From Figure 2.7 it follows that the effect of a change of the spectrum width is more significant than a narrower pass band of the receiver. Thus, for example, for a station having a main pulse duration of $\tau_p = 1$ microsec and if the width of the spectrum main lobe is 2 Mc, a change of the spectrum width $d\gamma/\gamma$ by 20% may cause a decrease of the signal power by 1.4 dB if $\Delta f = \frac{1}{\tau}$. When $\Delta f = \frac{2}{\tau}$, the signal power is decreased by only 0.2 dB. Consequently, the possible decrease of the signal power because of a change of the spectrum width must be considered only when the receiver pass band is relatively narrow.

Traveling Wave Ratio

Let us investigate a system consisting of an oscillator, waveguide, and load (Figure 2.8). Applied to a radar station, the antenna is the load. Such a load has a complex impedance. The matching of it with the waveguide can be achieved at one specific frequency. A partial reflection of energy in the direction toward the oscillator will occur in the frequency range of such a load.

If the waveguide is not matched with the load, then the field at an arbitrary point of the waveguide may be expressed in terms of the field of the forward wave. E_f and reflected wave $\dot{E}_r$:

$$\dot{E}=\dot{E}_{\text{п}}+\dot{E}_{\text{о}}.$$

The matching of the transmission line is characterized by values of the traveling wave ratio and the standing wave ratio. For brevity, we will designate them as TWR and SWR.

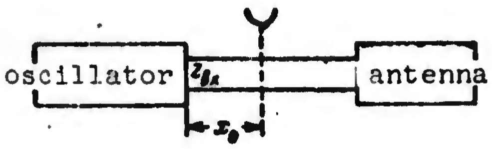


Figure 2.8. A waveguide transmission line.

The traveling wave ratio based on the voltage is defined as the ratio of the minimum electric field intensity in the waveguide $|\dot{E}|_{min}$ to the maximum $|\dot{E}|_{max}$:

$$s = \frac{|\dot{E}|_{мин}}{|\dot{E}|_{макс}}.$$

Since

$$|\dot{E}|_{мин} = |\dot{E}|_{п} - |\dot{E}|_{о},$$
$$|\dot{E}|_{макс} = |\dot{E}|_{п} + |\dot{E}|_{о},$$

and the modulus of the reflection coefficient is $|\dot{\rho}| = \frac{|\dot{E}|_{о}}{|\dot{E}|_{п}}$, then

$$s = \frac{1}{K} = \frac{1 - |\dot{\rho}|}{1 + |\dot{\rho}|}, \qquad (2.10)$$

where K is the standing wave ratio based on the voltage.

Modern radars have, as a rule, complicated antenna-waveguide channels. Rotary junctions, scanning heads, high-frequency switches and waveguides which have couplings and bends are used. All these elements cause a disruption of the matching. The use of radar domes also interferes with the matching conditions, because the amplitude and phase of the signal reflected from the radar dome are changed with a change of the antenna beam position.

A ship's superstructure and masts also have an influence on the matching of radar transmission lines. When the antenna is rotated, both the modulus and also the phase of the reflection coefficient are changed. If special measures are not taken, mismatching may lead to significant changes of the radar operating conditions.

A mismatch of the load with a waveguide transmission line of considerable length transmitting a large power causes a number of undesirable effects:

— it increases the losses in the waveguide;

— it increases the electric field intensity which can reduce the waveguide's electric strength (breakdown);

— it changes the power transmission conditions with a change in frequency.

Let us determine the relationship between the maximum electric field intensity in the waveguide and the standing wave ratio.

The power flux density in a mismatched waveguide is

$$p=\frac{|\dot{E}|_{\text{макс}}|\dot{E}|_{\text{мин}}}{Z_0}, \tag{2.11}$$

where Z_0 is the wave impedance of the waveguide.

The power flux density in a matched wave guide consequently will equal

$$p=\frac{|E|_{\text{п}}^2}{Z_0}.$$

Expression (2.11) may also be written in another form:

$$p=\frac{s|\dot{E}|_{\text{макс}}^2}{Z_0}.$$

Hence, it follows that

$$\frac{|\dot{E}|_{\text{макс}}}{|\dot{E}|_{\text{п}}}=\frac{1}{\sqrt{s}}.$$

The dependence of the TWR on the ratio $\frac{|E|_{max}}{|E|_p}$ is shown in Figure 2.9. It follows from the figure that the maximum value of field intensity in a waveguide which is not matched with the load may considerably exceed the field intensity in a matched waveguide.

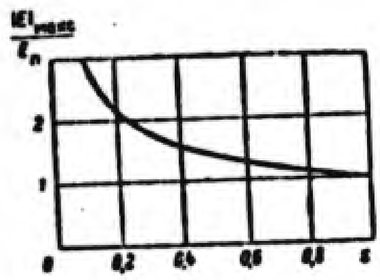


Figure 2.9. The maximum electric field in a mismatched waveguide as a function of the TWR.

This is especially dangerous in those cases when the transmitted power is large and the waveguide dimensions are small.

Let us proceed to an investigation of the third factor — the change of the power transmission conditions with a change in the frequency. Let us return again to Figure 2.8.

The input impedance of a waveguide which is the oscillator load for no losses in the waveguide is determined by the relationship

$$Z_{in}=Z_0\frac{1+\dot{\rho}}{1-\dot{\rho}}, \qquad (2.12)$$

where $\dot{\rho}=|\dot{\rho}|e^{j2kl}$ is the complex reflection coefficient;

$k=\frac{2\pi}{\lambda_w}$ is the wave number;

λ_w is the wavelength in the waveguide;

l is the length of the waveguide.

Since the wave impedance of the line is fixed, its input impedance is completely determined by the reflection coefficient ρ.

In superhigh frequency technology, the input impedance of a line is determined indirectly by measuring the traveling wave ratio and determining on this basis the modulus of the reflection coefficient $|\dot{\rho}|$ and by measuring the phase of the reflection coefficient.

From Expression (2.10), it follows that

$$|\dot{\rho}| = \frac{1-s}{1+s}.$$

The phase of the reflection coefficient is determined by measuring the distance x_0 from a point corresponding to the minimum value of the field intensity to the beginning of the waveguide (the flange) (Figure 2.8).

During operation, the frequency of the magnetron may vary within small limits. This may occur, for example, because of a change of the anode voltage or of the magnetic field intensity of the magnetron oscillator. This can be seen from the operating characteristics of a magnetron oscillator.

For small changes of the frequency, the quantities Z_0 and $|\dot{\rho}|$ remain practically unchanged. As regards $e^{j2\beta l}$, with a long waveguide it varies substantially even with small frequency changes. This implies abrupt changes of the magnetron load resistance and, in turn, a changed frequency of the oscillations being generated and of the power. All this has an adverse effect on the radar performance because the majority of centimeter wave oscillators are sensitive to a change of the load resistance.

If one assumes that within the limits of a limited frequency band the waveguide is sufficiently well matched with the load and consequently, the reflection coefficient $|\rho_2|$ is small, then one is able to determine relatively simply the change of the load resistance with small frequency changes.

When $\rho_2 \approx 0$, Expression (2.12) may be written in the following way:

$$Z_{\text{н}} = Z_0 (1 + 2 |\dot{\rho}_2| e^{j2\beta l}).$$

By differentiating, we obtain

$$\frac{dZ_{in}}{dk} = jlZ_0|\rho_2|le^{j2kl}.$$

Hence

$$dZ_{in} = jlZ_0|\rho_2|le^{j2kl}dk.$$

But since

$$k = \frac{2\pi}{\lambda_B} = \frac{v_\phi}{c^2}\omega,$$

where v_{phase} is the phase velocity in the wave guide;

c is the velocity in an unbounded dielectric medium, then

$$dZ_{in} \approx jl\frac{v_\phi}{c}Z_0|\rho_2|le^{j2kl}d\omega. \qquad (2.13)$$

From Expression (2.13) it follows that, if the waveguide is matched with the load $|\rho_2| = 0$, then a change of the frequency within fixed limits does not produce a change of the load on the oscillator ($dZ_{in} \approx 0$). The input impedance and, consequently, the power at the input and output of a waveguide transmission line will, in this case, depend slightly on small changes of both the frequencies and also the length of the waveguide.

Consequently, in order to reduce the influence of the load on the amount of the power and on the frequency of the oscillations being generated, it is necessary to match the magnetron oscillator and the load as well as possible. The antenna-waveguide channel, as we indicated above, consists of a large number of elements which worsen the matching conditions. It is impossible to achieve ideal matching ($s = 1$) in such a line, but a value of $s \geq 0.67$ is considered acceptable. In order to weaken the effect on the magnetron of the wave reflected from the load, a ferrite rectifier-switch is inserted in the antenna-waveguide channel. It is located as close as possible to the magnetron. The reflected waves from the elements located beyond the ferrite rectifier switch in this case are shunted off and sent to an absorber. In this way conditions are created which correspond to operating with a practically matched load.

As we showed above, the magnetron imposes the most rigorous requirements on the channel matching. The magnetron frequency and its oscillation stability, as is the case for any oscillator with self-excitation, depend on the character of the load connected to it. When the radar is operating, the load on the magnetron transmitter, and therefore the traveling wave ratio, do not remain constant. The reasons for such an instability were investigated above. A change of the channel's TWR entails a frequency change, as follows from the magnetron's load characteristics. A frequency change of the generated oscillations associated with a load change is called <u>pulling of the frequency</u>.

The pulling of the frequency is customarily characterized by the frequency pulling coefficient, that is, by the maximum change of the frequency of the oscillations being generated when $|\dot{p}| = 0.2$ ($s = 0.67$). As a consequence of the pulling of the frequency, a reduction of the received signal power takes place, and consequently there is a reduction of the station's effective range.

Slow changes of the magnetron frequency which arise, for example, with a change of the magnetron temperature may be compensated by manual tuning of the heterodyne. Changes of the magnetron frequency because of a change in the conditions of matching with the load take place rapidly. Fine tuning in this case may be done only with an automatic high-speed system for frequency fine tuning (automatic frequency control - AFC).

Let us find the relationship between the pulling of the frequency and the change of the received signal power. The connection between these quantities may be obtained if one investigates the reaction of a receiver having a prescribed pass band to its detuning, as a consequence of the frequency pulling which corresponds to a TWR change.

The receiver output power with precise tuning equals $P_0 = 2P_x$. The output power with detuning, as follows from Figure 2.10, equals

$$P = P_{(x+K_f)} + P_{(x-K_f)},$$

where $K_f = d_f \tau_p$;

d_f is the amount of receiver detuning which corresponds to the channel TWR change in Mc.

A graph of P_x as a function of x is shown in Figure 2.4.

From this same graph, it is possible to determine the magnitude of the receiver sensitivity decrease associated with the detunings $P_{(x+K_f)}$ and $P_{(x-K_f)}$. Having determined the frequency detuning from the magnetron load characteristic, one is able to determine the decrease of the signal power at the receiver output with detuning from the expression

$$\frac{\Delta P}{P_0} = \frac{P_0 - P}{P_0} = 1 - \frac{P_{(x+K_f)} + P_{(x-K_f)}}{2P_x}.$$

The dependence of P/P_0 on K_f* for various values of Δf is shown in Figure 2.11.

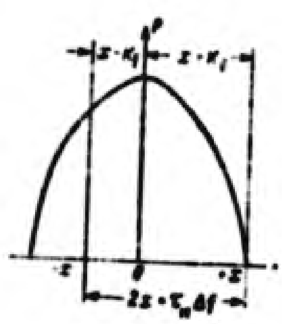


Figure 2.10. Determining the decrease of the signal's power at the receiver's output with a detuning of the frequency.

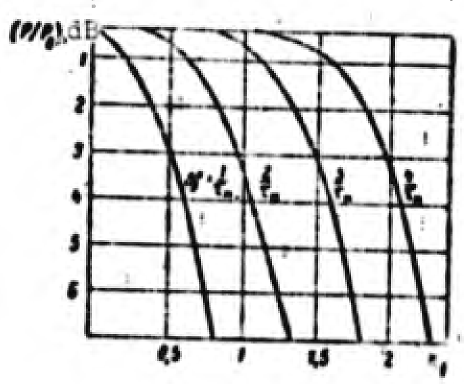


Figure 2.11. The dependence of P/P_0 on K_f for $\Delta f = 1/\tau_p$; $2/\tau_p$; $3/\tau_p$ and $4/\tau_p$.

*Translator's Note: This is incorrectly given as K_j in foreign text.

With a normally operating AFC circuit, which compensates for all the transient frequency changes, the decrease of the signal power at the receiver output due to this detuning is small and may be ignored.

During the operation of the radar one must, consequently, monitor the precision with which the automatic frequency control system is operating.

Mismatch may also lead to the appearance of parasitic frequency modulation which distorts the energy spectrum of the magnetron pulse frequencies. Parasitic frequency modulation produces either a broadening of the frequency band in which the spectrum energy is concentrated or leads to the appearance in the spectrum of maximums with considerable spacing between them. In the first case, this reduces the radar effective range, since part of the high-frequency pulse energy may appear beyond the receiver pass band limits. In the second case, it hinders and occasionally makes impossible the stable operation of the AFC circuit. Thus the presence of mismatch may produce substantial changes in the radar operating conditions as a whole.

During operation, monitoring must be periodically performed to see that the antenna-waveguide channel's TWR is constant. When a ferrite rectifier switch is used for this purpose, this monitoring indicates whether the switch is operating normally.

The Efficiency of an Antenna-Wave Guide Channel

The efficiency of an antenna-waveguide channel equals the ratio of the power at the output P_2 to the power at the input P_1 of the channel (Figure 2.12)

$$h=\frac{P_2}{P_1}.$$

For constant attenuation, the efficiency is at a maximum when the waveguide is matched with the load, since the presence of a reflected wave reduces the power entering the load. Let us determine the dependence of the efficiency on the traveling wave ratio.

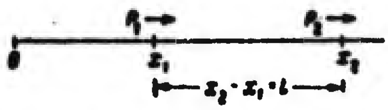


Figure 2.12. Determining the efficiency of an antenna-wave guide channel.

The power flux through the waveguide cross section is

$$P = P_f - P_r,$$

where P_f is the power flux through the waveguide cross section from the oscillator to the load;

P_r is the power flux through the waveguide cross section from the load to the oscillator.

The power flux through the waveguide input cross section is

$$P_1 = P_{f_1} - P_{r_1}$$

and through the waveguide output cross section it is

$$P_2 = P_{f_2} - P_{r_2}.$$

The expression for the power entering the load may also be written differently:

$$P_2 = P_{п_2}(1 - |\rho_2|^2).$$

where $|\rho_2|^2$ is the power reflection coefficient at the point x_2.

In turn, we have

$$P_{п_2} = P_{п_1} e^{-2\alpha l},$$

where 2α is the power attenuation constant in the waveguide in neper/m;

l is the waveguide length.

Consequently, we have

$$P_2 = P_{п_1} e^{-2\alpha l}(1-|\rho_2|)^2. \qquad (2.14)$$

Let us find the relationship between the reflection coefficients at the points x_1 and x_2.

Since the field intensities of the forward and reflected waves equal, respectively,

$$E_{п_2} = E_{п_1} e^{-\alpha l}, \qquad (2.15)$$
$$E_{о_1} = E_{о_2} e^{-\alpha l},$$

and the reflection coefficient at the point x_1 is $|\rho_1| = \frac{E_{r_1}}{E_{p_1}}$ and at the point x_2 is $|\rho_2| = \frac{E_{r_2}}{E_{p_2}}$, then

$$|\rho_1| = |\rho_2| e^{-2\alpha l}.$$

The power reflection coefficient $|\rho_1|^2$ consequently equals

$$|\rho_1|^2 = |\rho_2|^2 e^{-4\alpha l}.$$

The power flux passing through the waveguide entrance cross section is

$$P_1 = P_{п_1} - P_{о_1} = P_{п_1}(1-|\rho_1|^2)$$

or

$$P_1 = P_{п_1}(1-|\rho_2|^2 e^{-4\alpha l}). \qquad (2.16)$$

The expression for the efficiency may be obtained from Equations (2.14) and (2.16):

$$\eta = \frac{P_2}{P_1} = \frac{e^{-2\alpha l}(1-|\rho_2|^2)}{1-|\rho_2|^2 e^{-4\alpha l}}. \qquad (2.17)$$

The efficiency reaches its maximum value when $|\rho_2| = 0$;

$$\eta_{макс} = e^{-2\alpha l}. \qquad (2.18)$$

It is convenient to express the antenna-waveguide channel efficiency in terms of the TWR. Since

$$|\rho_0| = \frac{1-s}{1+s}, \tag{2.19}$$

then by substituting into expression (2.17) the values η_{max} and $|\dot{\rho}_0|$ from Equations (2.18) and (2.19), we finally obtain

$$\eta = \eta_{\text{макс}} \frac{4s}{(1+s)^2 - (1-s)^2 \eta^2_{\text{макс}}}. \tag{2.20}$$

From the last equation, it follows that the antenna-waveguide channel efficiency depends on two factors: the matching of the antenna with the channel and the attenuation in the channel.

The efficiency basically is determined by the losses in the antenna-waveguide channel, and if they are small then $\eta \approx \eta_{max}$ in a considerable interval of values of the traveling wave ratio (Figure 2.13). Thus, for example, when $\eta_{max} = 0.9$ and $s = 0.67$ the decrease of the efficiency is 1%, and when $\eta_{max} = 0.9$ and $s = 0.5$ the decrease of the efficiency equals 2.3%.

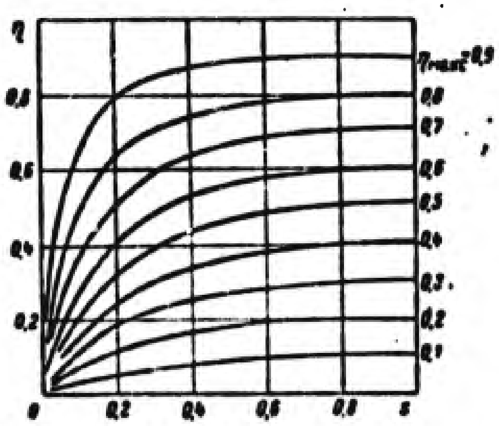


Figure 2.13. The antenna-waveguide channel's efficiency as a function of the traveling wave ratio, s, for various values of η_{max}.

<u>Attenuation in the Antenna-Wave Guide Channel</u>

The following factors affect the attenuation in an antenna-waveguide channel:

- the power losses in the waveguide walls,
- radiation through the flanges and other leaky high-frequency connections,

- losses as a consequence of oxidation of the inner surfaces of waveguides and rotary couplings, the formation of salts on the waveguide walls and introduction of moisture into the waveguide.

The effect of attenuation due to power losses in the waveguide walls is expressed by an exponential decrease of the amplitude of the field components in proportion to the passage of the wave along the waveguide (2.15). The relationship between the antenna-waveguide channel efficiency and the attenuation may be obtained from expression (2.20). By expanding η_{max} into a series in powers of αl and limiting ourselves to two terms of the expansion, we obtain

$$\eta_{max} = e^{-2\alpha l} \approx 1 - 2\alpha l.$$

Substituting this value into expression (2.20) and performing simple transformations, we find

$$\eta = \frac{1}{1+\left(s+\frac{1}{s}\right)\frac{\alpha l}{1-2\alpha l}}.$$

Considering that $\eta_{max} \approx 1 - 2\alpha l \approx 1$, we obtain

$$\eta = \frac{1}{1+\left(s+\frac{1}{s}\right)(\alpha l)}, \qquad (2.21)$$

where (αl) is the attenuation constant in nepers (np).

Since

$$(\alpha l)_{np} = \frac{1}{2}\ln\frac{P}{P_1}, \text{ and } (\alpha l)_{dB} = 10\lg\frac{P}{P_1}, \text{ then}$$

$$(\alpha l)_{np} = \frac{1}{2}\ln 10^{0.1(\alpha l)_{dB}},$$

consequently

$$(\alpha l)_{np} = 0,115(\alpha l)_{dB}.$$

Finally, we obtain

$$\eta = \frac{1}{1+\left(s+\frac{1}{s}\right)0,115(\alpha l)}, \qquad (2.22)$$

where (αl) is the attenuation constant in decibels.

In Figure 2.14 the antenna-waveguide channel efficiency as a function of the attenuation is shown for s = 0.5 and 1, from which it also follows that the efficiency basically is determined by the attenuation in the channel.

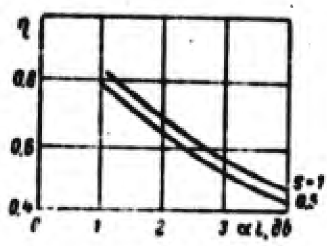


Figure 2.14. Curves of the antenna-waveguide channel's efficiency as a function of the attenuation for values of s = 0.5 and 1.

The attenuation depends on the conductivity of the waveguide wall material, on the dimensions of the waveguide, and on the frequency (Table 2.1). The data of Table 2.1 [7] refer to a rectangular waveguide with a cross section of 10 X 22.5 mm^2 operating at a frequency of 10 Mc. The relationships characterizing the attenuation as a function of the frequency for copper waveguides of rectangular and circular cross sections are shown in Figure 2.15 [3].

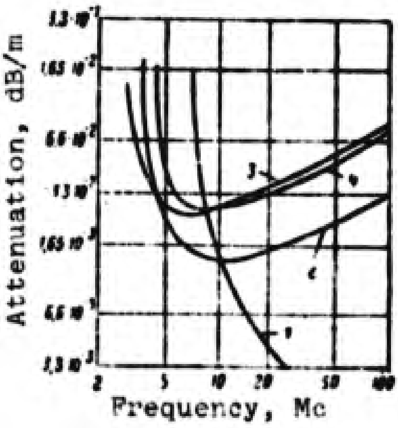


Figure 2.15. Curves of the attenuation of oscillations of various kinds as a function of the frequency for copper waveguides of rectangular cross section (50.8 X 25.4 mm) and circular cross sections (a diameter of 50.8 mm):

1 - circular waveguide, the H_{01} wave;
2 - circular waveguide, the H_{11} wave;
3 - circular waveguide, the E_{01} wave;
4 - rectangular waveguide, the H_{01} wave.

The conductivity of copper is $6 \cdot 10^7$ mho/m^3.

The power losses due to radiation through the flanges and other leaky high-frequency connections are small. The increase of the attenuation with the use of equipment aboard ships is connected mainly

TABLE 2.1. VALUE OF THE ATTENUATION AS A FUNCTION OF THE WAVEGUIDE'S MATERIAL

Material of the Waveguide's Walls	Conductivity, mho/m^3	Relative Attenuation	Attenuation, dB/m
Pure copper	$5,8 \cdot 10^7$	1,00	0,117
Brass annealed with Cu 90%	$2,52 \cdot 10^7$	1,52	0,178
Brass annealed with Cu 60%	$1,65 \cdot 10^7$	1,88	0,220
Pure aluminum	$3,47 \cdot 10^7$	1,29	0,151
Pure silver	$6,27 \cdot 10^7$	0,96	0,113

with the presence of moisture in the waveguide, with oxidation and with the formation of films of salt on the inner surfaces of waveguides. The effect of external conditions and the surrounding environment on the operation of a radar is investigated in reference [2].

Shipboard radars usually have long antenna-waveguide channels, the attenuation in which may vary during the period when the equipment is being used. In this case, if only insufficient drying methods are available, it is advisable to monitor the attenuation in the waveguide. In waveguides of short length one may neglect the affect of attenuation.

From what has been said it follows that, from the point of view of the power drop, good matching is necessary only with small η_{max}, that is with large losses in the channel. Nevertheless, definite attempts must be made to increase the TWR because the effectiveness of the radar performance to a significant degree is determined by the matching of the magnetron oscillator and the elements of the antenna-waveguide channel.

On the basis of what has been discussed, one may come to the following conclusion. Among the basic measurements which must be systematically performed in order to monitor the constancy of a radar effective range are the following:

- measurement of the receiving device sensitivity or its noise factor;
- measurement of the transmitter power;
- determining the energy spectrum of the frequencies of the magnetron oscillator pulses;
- measurement of the attenuation in the antenna-waveguide channel;
- measurement of the antenna-waveguide channel TWR.

It is also necessary to monitor the precision with which the automatic frequency control circuit operates.

2.3 Permissible Limits for the Deviation of the Technical Parameters

A radar station is a complex system, whose normal functioning is characterized by specific tactical parameters. These parameters, as was indicated above, are: the effective range, the measurement precision of the object coordinates, the station resolution, the dead zone, etc.

During the operational process, deterioration of these parameters occurs as a consequence of breakdown of the tuning and adjustment of the individual radar devices, a change of the parameters of the vacuum-tube and semiconductor instruments, and for a number of other reasons. Therefore, the problem of technical servicing includes the measurement and monitoring of the system technical parameters and its operating conditions, and tuning and adjustment tasks for the purpose of maintaining its tactical parameters at the prescribed level.

In order to arrive at a decision as to the state of a system by monitoring the technical parameters, one must establish requirements for the permissible limits of the change in these parameters. They, naturally, depend on the permissible boundaries of the change of the tactical parameters, which in each separate case are determined based on the requirements imposed to solve a specific problem.

If the boundaries of the change in the tactical parameters are large, then the equipment's possibilities which were built in when it was designed are not being used to the proper degree. If these boundaries are small, then the measures directed toward maintaining these parameters in the limits of the norm must be performed more frequently, which leads to additional expenditure of manpower and facilities.

During the equipment operation, one is obliged to deal with two problems:

1. On the basis of known deviations of the technical parameters from the normal values, it is necessary to determine the possible boundaries of the change of the tactical parameters.

2. Knowing the boundaries of the permissible changes of the tactical parameters, it is necessary to determine the permissible deviations of the technical parameters controlling these boundaries.

Let us assume that a station tactical parameters are a function of several independent variables — technical parameters. The first problem consists of determining the error of the function if the errors of the arguments and the form of the functional relationship are known.

As is known from the general theory of errors, the limiting relative error of a function of several independent variables $(\delta y)_{lim}$ equals the differential of the natural logarithm of this function. The sum of the absolute values of all the terms of the expression should be considered, that is, if in the general case $y = f(x_1, x_2, x_3, \ldots, x_n)$ and the function satisfies the continuity conditions, then

$$(\delta y)_{np} = \pm d[\ln f(x_1, x_2, x_3, \ldots, x_n)], \tag{2.23}$$

where $x_1, x_2, x_3, \ldots, x_n$ are independent variables.

Expression (2.23) may also be written in another form:

$$(\delta y)_{np} = \pm(a_1\delta x_1 + a_2\delta x_2 + a_3\delta x_3 + \dots + a_n\delta x_n) = \\ = a_1\frac{dx_1}{x_1} + a_2\frac{dx_2}{x_2} + a_3\frac{dx_3}{x_3} + \dots + a_n\frac{dx_n}{x_n}. \qquad (2.24)$$

where x_1, x_2, x_3, ..., x_n are the relative errors of the arguments; a_1, a_2, a_3, ..., a_n are constants which are the coefficients of the corresponding arguments.

The coefficients a_1, a_2, ..., a_n determine the degree of influence of the argument errors on the error of the function. Therefore, they are called the weighting factors or influence coefficients.

Expression (2.24) enables one to determine the function error limit if the argument errors are known. In this case, it is assumed that the least favorable case occurs — that is, the argument errors all have the same sign.

The inverse problem of the theory of errors is to determine the argument errors if the function errors and the form of the functional relationship are known. From Equation (2.24), it is seen that the problem of finding the argument errors associated with known errors of the function is indeterminate. In order to eliminate the indeterminacy, we proceed in the following way: we assume that all the terms of polynomial (2.24) have the same effect on the error of the function. This assumption is called the principle of equal effects.

With such an assumption, the following equalities hold

$$a_1\delta x_1 = a_2\delta x_2 = a_3\delta x_3, \dots, a_n\delta x_n = \frac{(\delta y)_{np}}{n}.$$

From the equalities, one may determine the relative errors of all the arguments if the limiting error of the function is known

$$\delta x_1 = \pm\frac{(\delta y)_{np}}{na_1};\ \delta x_2 = \pm\frac{(\delta y)_{np}}{na_2};\dots;\ \delta x_n = \pm\frac{(\delta y)_{np}}{na_n}. \qquad (2.25)$$

Let us investigate these problems as applied to monitoring a radar station effective range. It is known, for example, that during the operational process changes of the technical parameters responsible for the station's effective range take place. Therefore it is necessary to determine the limiting deviation of the station's effective range from the normal value. In order to do this, naturally, it is necessary to know the relationship between the effective range and the technical parameters of the station.

Let us begin with a simpler case of detecting airborne targets, which is the case when radio wave propagation takes place in free space. Since only the dependence of the effective range on the station technical parameters interests us, let us write expression (2.4) in the form

$$D_{\text{макс}}=k_1\sqrt[4]{\frac{P_{\text{ср}}}{P_{\text{пр}}}\frac{1}{F_{\text{ш}}\tau_{\text{и}}}\eta^2G^2\lambda^4}, \quad (2.26)$$

where k_1 is the proportionality constant;

P_{av} is the transmitter average power.

The limiting relative deviation of the station effective range as applied to the case being investigated by us will equal

$$(\delta D)_{\text{пр}}=\frac{dD_{\text{макс}}}{D_{\text{макс}}}=d\left[\frac{1}{4}\left(\ln P_{\text{ср}}\eta^2G^2\lambda^4-\ln P_{\text{пр}}F_{\text{ш}}\tau_{\text{и}}\right)\right]. \quad (2.27)$$

As was already mentioned above, one may neglect the deviation of λ and $G=\frac{1}{\lambda^2}$ from the normal value.

Consequently, expression (2.27) may be written as

$$(\delta D)_{\text{пр}}=\frac{1}{4}\frac{dP_{\text{ср}}}{P_{\text{ср}}}+\frac{1}{4}\frac{dP_{\text{пр}}}{P_{\text{пр}}}+\frac{1}{2}\frac{d\eta}{\eta}+\frac{1}{4}\frac{d\tau_{\text{и}}}{\tau_{\text{и}}}+\frac{1}{4}\frac{dF_{\text{ш}}}{F_{\text{ш}}}, \quad (2.28)$$

where

$$\frac{dP_{\text{ср}}}{P_{\text{ср}}},\ \frac{dP_{\text{пр}}}{P_{\text{пр}}},\ \frac{d\eta}{\eta},\ \frac{d\tau_{\text{и}}}{\tau_{\text{и}}},\ \frac{dF_{\text{ш}}}{F_{\text{ш}}}$$

are the relative deviations from the normal value of the station corresponding technical parameters.

The pulse duration and pulse repetition frequency during the operational process vary insignificantly, so that the deviations of τ_p and F_p from their normal values do not have to be considered.

The effect of changed width of the frequency energy spectrum of the magnetron oscillator pulses on the station effective range may be taken into account as was shown in § 2.2, by changing the pulse duration by an amount corresponding to the relative change of the width of the spectrum main lobe.

Expression (2.28), with consideration of what has been said, takes the form

$$(\delta D)_{\text{пр}}=\frac{1}{4}\frac{dP_{\text{ср}}}{P_{\text{ср}}}+\frac{1}{4}\frac{dP_{\text{пр}}}{P_{\text{пр}}}+\frac{1}{2}\frac{d\eta}{\eta}+\frac{1}{4}\frac{d\gamma}{\gamma}, \qquad (2.29)$$

where $\frac{d\gamma}{\gamma}=\frac{d\tau_\gamma}{\tau_\gamma}$ is the relative deviation from the normal value of the width of the spectrum main lobe.

Expression (2.29) enables one to determine the limiting deviation of the station effective range from the normal value with a change of the station corresponding technical parameters.

The solution of the inverse problem is considerably more interesting. The inverse problem may be formulated in the following way. Let us determine the permissible simultaneous deviations from the norm of the technical parameters which are responsible for the station effective range with consideration of the fact that the limiting deviation of the station effective range must not exceed the value $(\delta D)_{lim}$.

Let us assume that all the terms of polynomial (2.29) have the same effect on the function. Using the principle of equal effects, one is able to determine the permissible deviations from the norm of the technical parameters from the following expressions:

$$\frac{dP_{\text{ср}}}{P_{\text{ср}}}=\pm(\delta D)_{\text{пр}},\quad \frac{dP_{\text{пр}}}{P_{\text{пр}}}=\pm(\delta D)_{\text{пр}},\quad \frac{d\eta}{\eta}=\pm 0,5(\delta D)_{\text{пр}},$$

$$\frac{d\gamma}{\gamma}=\pm(\delta D)_{\text{пр}}. \qquad (2.30)$$

Because under actual conditions there is only a small probability for the simultaneous deterioration of all the parameters, the relative errors of the components may be taken as 2 - 3 times larger than those obtained from expressions (2.30).

The maximum permissible deviation of the effective range from the normal value must be determined by starting from the conditions for the resolution of a specific problem by a station. Thus, for example, as such a quantity one may select the scatter of the measurement results of the station maximum effective range, since this quantity may be neglected when solving a number of problems.

For the case when the sensitivity instead of the noise factor of the receiver is measured, the problems may be solved in a similar way.

It is convenient to write expression (2.4) for the station effective range in this case in the form

$$D_{\text{макс}}=k_2\sqrt{\frac{P_{\text{ср}}}{F}\eta^2\frac{G^2\lambda^2}{(\tau_{\text{и}}\Delta f+1)^2F_{\text{ш}}^{1/2}}}, \qquad (2.31)$$

where k_2 is the proportionality constant.

Using expression (2.31), one may find the dependence of the function error on the argument errors:

$$(\delta D)_{\text{пр}}=\frac{\Delta D}{D_{\text{макс}}}=d\left\{\frac{1}{4}\left[\ln P_{\text{ср}}\eta^2\right]-\ln FF_{\text{ш}}^{1/2}(\tau_{\text{и}}\Delta f+1)^2\right]\right\}.$$

In this expression, as in the previous case, we neglect the deviation of λ and G from the normal value.

By performing simple transformations, we obtain

$$(\delta D)_{\text{пр}}=\frac{1}{4}\frac{dP_{\text{ср}}}{P_{\text{ср}}}+\frac{1}{4}\frac{dF}{F}+\frac{1}{2}\frac{d\eta}{\eta}+\frac{1}{4}\frac{d\tau_{\text{и}}}{\tau_{\text{и}}}+\frac{1}{8}\frac{dF_{\text{ш}}}{F_{\text{ш}}}+\frac{1}{4}\frac{d\Delta f}{\Delta f}.$$

When deriving this expression, it was assumed that $\tau_p \Delta f = 1$, which usually is true for radar receivers.

Neglecting, as in the previous case, the deviation of τ_p and F_p from the normal values and taking into account the fact that during operation the receiver pass band remains practically constant, we obtain

$$(\delta D)_{np}=\frac{1}{4}\frac{dP_{cp}}{P_{cp}}+\frac{1}{4}\frac{dF}{F}+\frac{1}{2}\frac{d\eta}{\eta}+\frac{1}{4}\frac{d\gamma}{\gamma}. \tag{2.32}$$

Let us assume, as before, that all the terms of polynominal (2.32) must have the same effect on the error of the function. Then we have

$$\frac{dP_{cp}}{P_{cp}}=\pm(\delta D)_{np}, \quad \frac{dF}{F}=\pm(\delta D)_{np},$$
$$\frac{d\eta}{\eta}=\pm 0{,}5(\delta D)_{np}, \quad \frac{d\gamma}{\gamma}=\pm(\delta D)_{np}. \tag{2.33}$$

In this case, there also is only a slight probability of the simultaneous deterioration of all parameters. Therefore, the relative errors of the components may be taken as 2 - 3 times larger than the values obtained from expressions (2.33).

Since during operation one monitors not the antenna-waveguide channel efficiency but the attenuation in it, let us determine the relationship between the relative errors of these quantities.

From expressions (2.22) and (2.23), it follows that

$$\frac{d\eta}{\eta}=\pm d\ln\frac{1}{1+aL}=\frac{dL_1}{L_1},$$

where

$$L_1=L+\frac{1}{a}; \quad a=\left(s+\frac{1}{s}\right)0{,}115; \quad L=al.$$

Let us proceed to detection of targets on the water surface. When detecting surface targets, the effective range and the station technical parameters are connected, as we already mentioned above, by a complicated relationship. This follows, for example, from expression (2.1). In the case being investigated, it is very important to know whether the detection of the target takes place in the illuminated region or in the semi-shadow region. Thus, for example, calculations show that a decrease of a station effective range by 15% for a surface object may be brought about by a decrease of the power drop by 3 - 4 dB for detection in the illuminated region and by 6 - 8 dB for detection in the semi-shadow region. In the overwhelming number of cases, the surface targets are detected in the semi-shadow region. The limiting deviation of the station effective range from the normal value associated with a change of the technical parameters may be determined in this case in the following way.

Let us find the maximum possible deviation of the station power potential from the normal value by using the expression

$$\Delta\Pi_{\text{э}} \approx \Delta P_{\text{и}} + \Delta P_{\text{пр}} + \Delta P_{L} + \Delta P_{\gamma}\ \text{дб}, \qquad (2.34)$$

where ΔP_p is the deviation of the transmitter pulse power from the normal value, in dB;

ΔP_{red} is the deviation of the receiver sensitivity from the normal value, in dB;

ΔP_L is the power loss in the antenna-waveguide channel as the result of the deviation of the attenuation from the normal value, in dB;

ΔP_γ is the power loss of the arriving signal due to the change of the spectrum main lobe width, in dB.

In contrast to the premises which were made when deriving expressions (2.29) and (2.32), it is customary to assume when deriving expression (2.34) that the changes of the components have the same effect on the station's power potential — that is, that the weighting factors are equal.

Since the equation relating the station power potential and its effective range when detecting a surface target is very complicated, to solve it it is desirable to introduce an acceptable approximation.

We will assume a linear relationship between these quantities as the approximation. Such an approximation is particularly justified for changes in the range due to relatively small deviations of the station power potential during operation.

The limiting deviation of the station effective range from the normal value because of a change of the station power potential consequently may be determined from the equation

$$\Delta D_{\text{пр}} = k_D \Delta \Pi_{\text{э}}, \qquad (2.35)$$

where k_D is the proportionality constant.

The proportionality constant k_D corresponds to the magnitude of the decrease of the station effective range when detecting surface targets in the semi-shadow region and a change of the power drop by one decibel. This value may be determined by calculations or by experiments.

As an example of the evaluation of the changed effective range for surface targets, one may use the following constants. For stations in the centimeter wave range, a change of the station's power potential $\Delta\pi_e$ by 1 dB corresponds approximately to a change of the effective range of 1.5 - 2.5 cables [1] (k_D = 1.5 - 2.5). The large value of the constant corresponds to the upper part of the range and smaller targets, and the smaller constant corresponds to the lower part of the range and larger targets.

When it is necessary to solve the inverse problem — that is, to determine the permissible deviations of the technical parameters associated with a given change of the station effective range — it is necessary to assume as before that the principle of equal effects is valid.

Footnote (1) appears on page 151a.

Then we have

$$\Delta P = \frac{\Delta \Pi_{\text{и}}}{4} = \frac{\Delta D_{\text{пр}}}{4k_D} \text{ dB.} \tag{2.36}$$

and

$$\Delta P = \Delta P_{\text{и}} = \Delta P_{\text{пр}} = \Delta P_L = \Delta P_{\gamma}. \tag{2.37}$$

Because the simultaneous deterioration of all the parameters is unlikely, the permissible deviations of the components may be taken to be 2 - 3 times larger than those obtained from expression (2.36).

<u>Example</u>. The maximum permissible deviation of the effective range of a centimeter range station from the normal value, when the station is being used for detecting a surface target, equals 12 cab (1 cab = 185.2 m). The receiver pass band is $\Delta f = \frac{1}{\tau_p}$.

It is necessary to determine the permissible deviations from the normal value of the transmitter pulse power, the receiver sensitivity, the attenuation in the antenna-waveguide channel, and the width of the spectrum main lobe.

Let us assume that all the technical parameters have the same effect on the deviation of the effective range from the normal value. Using expression (2.36), we obtain

$$\Delta P = \frac{12}{4 \cdot 2} = 1.5 \text{ dB.}$$

Because there is only a small probability that the errors of the components of expression (2.34) will have the same sign, let us take them to be two times larger than those which follow from expressions (2.36). Then

$$\Delta P_{\text{и}} = \Delta P_{\text{пр}} = \Delta P_L = \Delta P_{\gamma} = 3 \text{ dB.}$$

Let us determine the deviation of the width of the spectrum main lobe from the normal value.

Since the power loss of the arriving signal which corresponds to this deviation is $\Delta P_{\tau} = 3$ dB, we find from the graphs of Figure 2.7 for $\tau_p = \frac{1}{\Delta f}$ the value $K_y = 0.3$.

Consequently, $\frac{d\gamma}{\gamma} = \pm 30\%$.

2.4. Permissible Errors in the Measurement of the Technical Parameters

While evaluating the state of the system by monitoring the technical parameters, we did not consider the errors of the measurement instruments. This may lead to two types of situations:

1. Missing the Malfunction when the value of one or several parameters goes beyond the permissible limits, but the monitoring cannot detect it because of measurement instrument errors.

2. A False Alarm when the system parameters are found to be within the permissible limits, but because of measurement instrument errors a decision will be made that the system has malfunctioned.

From what has been said, we may arrive at the conclusion that in each separate case one must make a sound choice of the measurement instruments, the range of the field of tolerances, and the acceptance boundaries for making a decision as to whether a parameter has exceeded the permissible limits. This all must be done in order that the probability of a false alarm does not exceed a specified value.

In connection with this, during operation one may encounter the following problems:

1. The measurement instrument errors are given, and it is necessary to determine the possible deviations of the output parameters which may not be observed during the measurement process as a consequence of instrument errors.

This problem arises when the measurement instruments are already specified — for example they are built into the equipment — and it is necessary to determine the possible deviations of the output parameters which may possibly not be noticed because of the errors of these instruments.

2. The deviations of the output parameters from the normal values are given, and it is necessary to determine the permissible measurement errors of the technical parameters which will still enable one to detect these deviations in order to make a sound choice of measurement instruments.

The first problem consists of determining the function error if the argument errors and the form of the functional relationship are known. Consequently, for this case, expression (2.29) is valid. If one assumes in it that $\frac{dP_{av}}{P_{av}}$, $\frac{dP_{lim}}{P_{lim}}$, $\frac{d\eta}{\eta}$, and $\frac{d\gamma}{\gamma}$ are the relative errors of the measurement instruments, then the expression characterizes the dependence of the limiting deviation of the station effective range from the normal value, which may not be noticed because of the errors of these instruments.

Expression (2.34) is also valid. If one assumes in it that ΔP_p, ΔP_{rec}, ΔP_L and ΔP_γ are the measurement instrument errors, then $\Delta\pi_e$ will be the limiting deviation of the station power potential which will not be detected because of the measurement instrument error.

The second problem consists of determining the argument errors if the function error is known.

In this case, expression (2.30) is valid. As in the previous case, $\frac{dP_{av}}{P_{av}}$, $\frac{dP_{lim}}{P_{lim}}$, $\frac{d\eta}{\eta}$, and $\frac{d\gamma}{\gamma}$ are the relative errors of the measurement instruments.

Expressions (2.36) and (2.37) which enable one to determine the measurement instrument errors ΔP_p, ΔP_{rec}, ΔP_L and ΔP_γ are also valid. When the measurement errors obtained from conditions (2.34) and (2.36) are so small that the measurements may not be made by the measurement instruments which are available, it is necessary either to use instruments with smaller measurement errors or to reduce the requirement for the limiting value of the function.

Example 1. The measurement instruments enable one to measure the station technical parameters with the following errors:

- the transmitter pulse power ± 0.8 dB;
- the receiver noise factor ± 1.2 dB;
- the antenna-waveguide channel attenuation ± 0.5 dB;
- the spectrum width ± 20%.

The receiver pass band is $f = \frac{1}{\tau_p}$. It is necessary to determine the decrease of the maximum effective range of a centimeter wave station for a surface target, which may be unnoticed because of measurement instrument errors.

Solution. Let us determine the arriving signal power loss corresponding to an error in the measurement of the spectrum main lobe width.

From the graphs of Figure 2.7, when $\tau_p = \frac{1}{\Delta f}$ and $k_\gamma = \frac{d\gamma}{\gamma} = 0.2$ then $\Delta P_\gamma = 1.4$ dB.

Using expression (2.34), let us find the maximum measurement error of the station power potential:

$$\Delta \Pi_s = 0.8 + 1.2 + 0.5 + 1.4 = 3.9 \text{ dB}.$$

Let us determine the limiting change of the station effective range for a surface object which may not be noted because of the measurement error of the station power potential by using Equation (2.35):

$$\Delta D_{эф} = k_D \Delta П_э = 2\cdot 3{,}9 = 7{,}8 \text{ cab}.$$

Example 2. The maximum permissible deviation of the effective range from the normal value when the station is used for detecting an aerial target is $(\delta D)_{lim} = 10\%$.

It is necessary to determine the errors which may occur in the instruments provided for measuring the transmitter average power, the receiver sensitivity, the antenna-waveguide channel efficiency, and the width of the spectrum main lobe.

Solution. From expressions (2.33), it follows that

$$\frac{dP_{ср}}{P_{ср}} = \pm 10\%, \quad \frac{dP_{пр}}{P_{пр}} = \pm 10\%,$$

$$\frac{d\eta}{\eta} = \pm 5\%, \quad \frac{d\gamma}{\gamma} = \pm 10\%.$$

Because the probability that all the errors will have the same sign is small, let us assume that the errors of the components will be two times larger than those obtained from expressions (2.33). Consequently,

$$\frac{dP_{ср}}{P_{ср}} = \pm 20\%, \quad \frac{dP_{пр}}{P_{пр}} = \pm 20\%, \quad \frac{d\eta}{\eta} = \pm 10\%, \quad \frac{d\gamma}{\gamma} = \pm 20\%.$$

If the specified error in the measurement of one of the parameters — for example, the width of the spectrum main lobe — is small and cannot be measured with the measurement instruments at our disposal, then it is necessary to reduce the requirement for the maximum permissible deviation of the station effective range. Let us assume, for example, that $(\delta D)_{lim} = 12.5\%$. Then

$$\frac{dP_{ср}}{P_{ср}} = \pm 25\%, \quad \frac{dP_{пр}}{P_{пр}} = \pm 25\%,$$

$$\frac{d\eta}{\eta} = \pm 12{,}5\%, \quad \frac{d\gamma}{\gamma} = \pm 25\%.$$

2.5. High-Frequency Elements for Connecting Measurement Instruments with Radar Stations

In operation, in order to measure electronic quantities in the superhigh frequency range both portable measurement instruments and also measurement instruments built into the equipment are widely used. For monitoring the technical state of radar equipment while it is being used and for operational tuning, measurement instruments built into the equipment are more and more widely used. The field of application of portable measurement instruments is shrinking. They are used mainly for measuring the equipment parameters in factories and in repair shops and in operation when performing preventive measures. Assuming that the reader is familiar with commonly used measurement equipment, let us briefly discuss certain special portable measurement equipment and high-frequency elements for connecting them with radar.

When measuring the main technical parameters of radar, portable combined test instruments — that is, radar-testers — and in recent years radar test instruments have been widely used. The advantages of these instruments are that they enable one to check and measure a number of radar technical parameters, and they are compact and relatively light weight. As a consequence, it is convenient to use them for checking and tuning radars under field conditions and aboard ships.

Radar-testers are universal instruments designed for measuring the technical parameters and for tuning radar stations. They enable one to measure the operating frequency and power of transmitters and also the sensitivity of receiver devices.

Radar test instruments enable one to perform all-around tests of radar receivers (to measure the sensitivity and to determine the receiver pass band, to measure the recovery time of the receiver sensitivity, to check the operation of the automatic frequency circuit), to measure the transmitter power, to measure the frequency energy spectrum of the transmitter pulses, and to check the antenna-waveguide channel matching. A radar test instrument may also be used as an artificial radar target.

In order to measure the technical parameters, it is necessary to achieve a coupling of the measurement instruments with the radar being tested (Figure 2.16). In order to connect the instruments with the radar, the following elements are used:

- measurement probes;
- unidirectional or bidirectional couplers;
- antenna of measurement instruments.

Let us investigate the advantages and shortcomings of the various coupling elements and methods for connecting measurement instruments with radars.

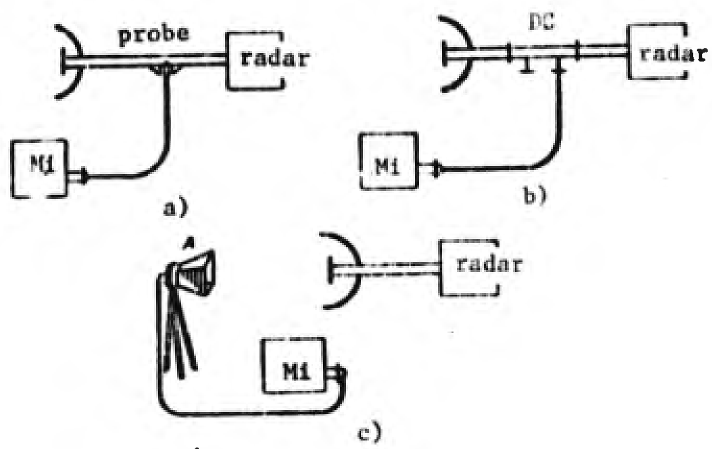


Figure 2.16. Schemes for the connection of a measurement instrument (Mi) with a radar: a - by means of a measurement probe and feeder; b - by means of a unidirectional or bidirectional coupler (DC) and a set of connecting elements attached to the instrument; c - by means of an antenna (A) located in the radar's radiation field.

Measurement Probe

A measurement probe is the simplest high-frequency coupling element. It is a joint mounted in a waveguide (Figure 2.17). The probe pin passes through a hole in the wide wall of the waveguide and extends into the waveguide approximately 1-2 mm. The probe coupling attenuation varies as a function of the pin insertion depth. The probe is mounted in such a way that the pin insertion depth would be rigidly fixed. When this is done the coupling attenuation is a constant. The value of the coupling attenuation also depends on the displacement of the probe from the waveguide center line. After the probe is mounted, it must be calibrated — that is, one must determine with an acceptable degree of decision its coupling attenuation. The amount of the coupling attenuation is taken into account when making measurements.

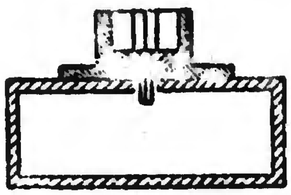


Figure 2.17. Measurement probe.

A measurement probe is the simplest but least ideal form of connection. This is explained by the following reasons. The voltage induced in the probe pin is proportional to the field intensity at the point where the probe is located.

When measuring the traveling wave ratio (TWR) of an antenna channel, the distribution of the electric field along the waveguide varies and consequently the power drawn off by the probe also varies. Actually, the power drawn off by the probe equals

$$P_{\text{з}} = kP_{\text{в}}\frac{(1+|\dot{\rho}|)^2}{1-|\dot{\rho}|^2}, \qquad (2.38)$$

where k is the proportionality constant;

P_p is the power drawn off by the probe;

$|\dot{\rho}|$ is the modulus of the reflection coefficient for the voltage;

P_w is the power being transmitted along the waveguide.

Thus the dependence of the power drawn off by the probe on the reflection coefficient (frequency) leads to errors in the measurement of the transmitter power and the receiver sensitivity.

The power drawn off by the probe will be practically independent of the reflection coefficient if one neglects, in expression (2.38), the term $|\dot{\rho}|$ of the first order. Since usually $|\dot{\rho}| = 0.2$, which corresponds to $s = 0.67$, such an omission is correct only with a certain assumption.

With normal matching of the waveguide channel (normal TWR is no less than 0.7) the probe coupling attenuation varies insignificantly.

Directional Coupler

A directional coupler is a device which achieves the directional coupling of a small part of the power passing along a transmission line.

It is the most ideal device for a high-frequency connection, since it enables one to most effectively provide a connection of the radar with measurement instruments. Unidirectional and bidirectional couplers are manufactured commercially (Figures 2.18 and 2.19). It is more advisable to use the latter since they enable one to simply and rapidly measure the TWR.

A unidirectional coupler (Figure 2.19) used as the connection element when making measurements in the centimeter wave range consists of two waveguides which abut one another along the wide wall. There is a common coupling hole between these wide walls. The lower waveguide is the main waveguide and the upper waveguide is an auxiliary waveguide. One end of the upper waveguide terminates in an absorber and the other end terminates in a high-frequency joint. In such a type coupler both magnetic and electric coupling exists between the main and auxiliary waveguides. In the auxiliary waveguide the electric field at the coupling hole is similar to the field created by

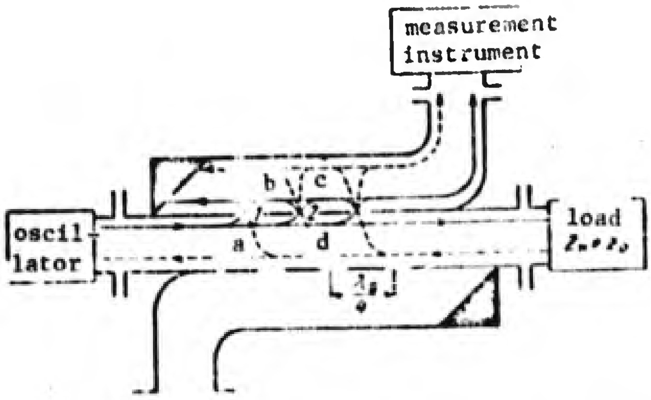


Figure 2.18. The arrangement of a waveguide bidirectional coupler.

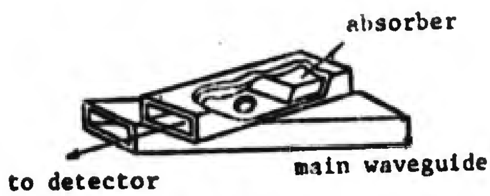


Figure 2.19. A unidirectional coupler.

an electric dipole whose moment is parallel to the electric field of the incident wave (Figure 2.20). The magnetic field at the coupling hole is similar to the field created by a magnetic dipole whose moment is parallel to the transverse field in the main waveguide but is directed in the opposite direction (Figure 2.21).

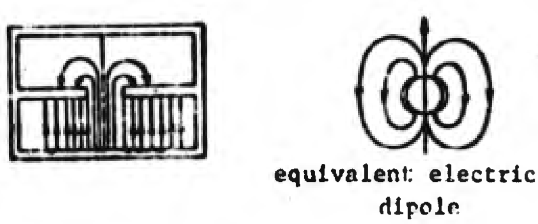


Figure 2.20. Electric coupling through a small hole.

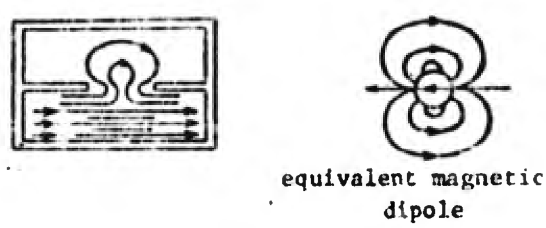


Figure 2.21. Magnetic coupling through a small hole.

The corresponding wave is established in accordance with the distribution of the electric and magnetic fields in the auxiliary waveguide. The propagation direction of this wave is opposite to the propagation direction of the wave in the main waveguide.

The waves excited as a consequence of the presence of the electric and magnetic coupling must be equal in amplitude in order that they would cancel out in one of the directions. In order to achieve this equivalence, one of the waveguides is rotated in respect to the other to a specific angle. When this is done, the electric coupling remains constant but the magnetic coupling is decreased.

For equal couplings, the phase relations between the electric and magnetic waves are such that in the auxiliary waveguide energy will be propagated only in one direction. The broad-band capacity of such a coupler is around 20%.

A shortcoming of such a directional coupler is that it does not allow one to make measurements of the antenna-waveguide channel TWR. This defect may be eliminated if there is an adaption which enables one to rotate an additional waveguide section around an axis which coincides with the coupling hole axis.

In this case one may determine the power of both the incident wave and the reflected wave at two fixed positions of the additional waveguide section, and consequently, determine the antenna-waveguide channel TWR. The coupling elements in directional couplers may also be made in the form of crossed slits.

In this case the coupler is a section of a waveguide in which two additional waveguide sections are abutted to it from both sides so as to be perpendicular to the main waveguide (Figure 2.22). The main and additional waveguide sections are interconnected by means of two slits. The slits are located in such a way that one of them is cut in the direction of the wave propagation and the other is cut perpendicular to it. Radiation of energy from one waveguide to another occurs through the slits.

The main advantages of these couplers is the broad-band capacity (around 20%) and the small size. When making measurements in the

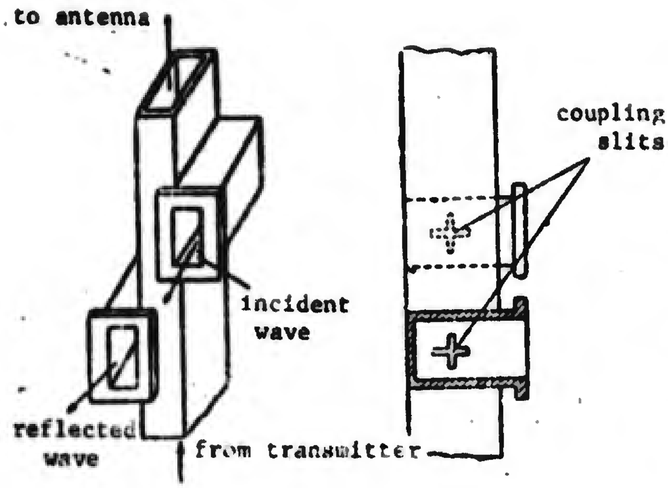


Figure 2.22. A directional coupler with coupling elements in the form of crossed slits.

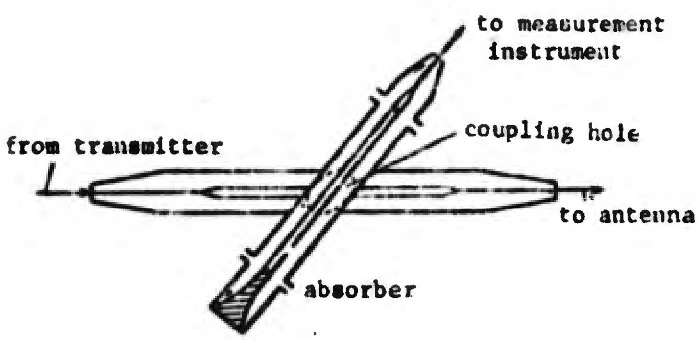


Figure 2.23. A coaxial directional coupler.

decimeter wave range, mainly coaxial directional couplers are used.

A coaxial directional coupler is an auxiliary feeder section which is positioned at a specific angle to the main feeder of the radar and has a common coupling hole with it. The additional feeder section is terminated on one end by an absorber and on the other by a high-frequency joint (Figure 2.23). Bidirectivity in such a coupler is achieved by mutually exchanging the points where the measurement instrument and the absorber are connected.

The main technical parameters of a directional coupler are:

- the coupling attenuation;
- the directivity;
- the broad-band capacity (frequency range);
- the permissible power.

The ratio of the power in the main waveguide to the power in the auxiliary waveguide is called the coupling attenuation of a directional coupler. The coupling attenuation is expressed in relative values and equals

$$\zeta = 10 \lg \frac{P_{a}}{P_{m}} \text{ dB},$$

where P_w is the power entering the main waveguide;

P_{ax} is the power shunted to the auxiliary waveguide and entering the power meter.

The coupling attenuation of directional couplers used for making measurements lies in the limits 15-55 dB; consequently an insignificant part of the power passing through the main waveguide is shunted to the power meter. Directional couplers when connected to antenna-waveguide channels introduce practically no mismatch. The magnitude of the coupling attenuation is set by the directional coupler and must be taken into account when making measurements.

The coupler directivity is characterized by the ratio of the power of the waves being propagated in the auxiliary waveguide in opposite directions in association with a traveling wave in the main waveguide. This quantity also is expressed in decibels and equals

$$v = 10 \lg \frac{P'_{ax}}{P''_{ax}},$$

where P'_{ax} and P''_{ax} are the powers of the waves being propagated in the auxiliary waveguide in opposite directions.

Thus the directivity is a measure of the coupler quality. The directivity depends on the frequency, the distances between the coupling elements and their shape. A coupler directivity varies within broad limits with a change of the frequency.

An important technical parameter of a directional coupler is its broad-band capacity.

Providing for the broad-band capacity of a coupler is especially important for radars which operate not at a fixed frequency, but in a range of frequencies. In order to increase the broad-band capacity, couplers with a large number of coupling elements are used. Moreover, usually a system of coupling elements is used in which the amplitude of the excited wave varies in accordance with the law for the coefficients of a binomial expansion.

Thus, for example, when the number of holes equals eight, the coupling values should vary in accordance with the series of coefficients 1, 7, 21, 35, 35, 21, 7, 1.

The permissible power of a directional coupler depends on the limiting power of the absorber dissipation (the absorbing impedence) and on the coupling attenuation of the coupler.

The possibility of obtaining directional coupling, broad-band capacity, good matching with the antenna-waveguide channel and independence of the parameters from external conditions make the directional coupler the best element for high-frequency connection of a station with measurement instruments.

The Antenna of a Measurement Instrument

As was indicated above, the connection between the measurement instrument and a radar may be achieved through space. When this is done, in order not to introduce an additional error in the measurement, the antenna of the measurement instrument must be located in the far zone of the field. It is possible to correct the antenna directive gain of the near zone for the purpose of obtaining the directive gain for the far zone. However, in order to do this, additional data are necessary. It is desirable to perform the measurements at the minimum distances which satisfy the conditions of the far zone. The minimum distance between the antennas of the radar and the measurement instrument which satisfies this condition is determined by the expression

$$R_{\text{мин}} = \frac{(2a_1 + 2a_{\text{и}})^2}{\lambda}, \qquad (2.39)$$

where $2a_1$ is the aperture of the radar antenna;
$2a_1$ is the aperture of the measurement instrument antenna.

With large antenna apertures, the quantity R_{min} may be several tens of meters which in a number of cases complicates the measurements. With this coupling method the electric axes of the antennas must coincide.

When calculating the overall attenuation from the instrument antenna to the radar antenna, it is necessary to take into account the antenna gain of the radar, the attenuation in space and the antenna gain of the measurement instrument.

The attenuation in space varies in proportion to the square of the distance between the transmitting antenna and the receiver antenna.

As was indicated above, a coupling between the test instrument and a station through space has the advantage that the efficiency of the station antenna waveguide channel is taken into account when making measurements. However, one should not always use this measurement method during the radar operation process. It has the following disadvantages:

- it allows one to measure only the receiver sensitivity;
- the error of the sensitivity measurement is large because it depends on the precision of setting up and orienting the measurement instrument antenna with respect to the radar antenna;
- the antennas of the majority of radars are mounted at the highest points and it is difficult to set up the test instrument antenna in front of them.

From what has been said above, one may draw the conclusion that it is advisable to use such a measurement method only for radars having antennas which are situated relatively low, and in those cases when a precise measurement of the receiver sensitivity is not required.

2.6. The Method of Measuring Radar Technical Parameters

Let us discuss certain features of the measurement of technical parameters which determine effective range of à radar.

Measurement of the Receiver Sensitivity

The sensitivity of radar receivers may be measured by means of combined test instruments (radar-testers) GK4-14 and GK4-3 and by the radar test instruments GK4-19, GK4-21 and others.

Two methods for measuring receiver sensitivity are possible:

1. in the continuous oscillation mode; with this method the ratio of the signal to the noise is established by means of a pointer-type instrument which is included in the circuit of a second detector of a radar receiver;
2. in the pulsed oscillation mode; with this method the ratio of the signal to the noise is established on the basis of the station indicator or an oscillograph.

The first measurement method is more precise than the second. Thus, for example, the maximum random error with radar-testers in the continuous oscillation mode equals $\pm$ (1 - 1.5) dB whereas in the pulse oscillation mode it is $\pm$ 2.5 dB. However, the second measurement method corresponds more closely to the operation of a radar. A measurement using this method gives a sensitivity value several decibels higher than a measurement in the continuous oscillation mode. This error may be determined for each station and taken into account when making measurements.

The relative monitoring of a receiver sensitivity, especially if it is performed by the same person using the same measurement instrument, enables one to determine the reduction of the receiver sensitivity with a high precision with either measurement method. It is also necessary to keep in mind that the sensitivity measurement of a receiver during switching on and switching off of the transmitting device may give different results. This is explained by the fact that a temporary deterioration of the conversion properties of the detectors of the mixer (worsening of the detectors), which entails a certain reduction of the receiver sensitivity, occurs under the action of the transmitter pulses. This is shown by the fact that, after the transmitter is switched on, the current of the crystal decreases to a specified value and then remains constant. After the transmitter is switched off, the crystal current and the receiver sensitivity are restored to the initial value after 5-20 minutes. Most significantly, the crystal current of the diodes DK-S4 and DK-S3 is decreased (60-80%), which corresponds to a decrease of the receiver sensitivity by 4-6 dB.

An insignificant decrease of the crystal current (by 5-10%) which causes practically no change of the receiver sensitivity is observed in the germanium diodes D403B and D403V.

Thus, in general, it is necessary to measure the receiver sensitivity when the transmitting device is switched on, that is, under conditions approximating the actual conditions.

It is necessary to point out that the sensitivity measurement in the continuous oscillation mode has a number of advantages over the measurement in the pulsed oscillation mode and is the preferred method.

Measurement of the Transmitter Power

Measurement of the transmitter power usually is preceded by measurement of the frequency of the oscillations being generated.

The frequency of the oscillations generated by a magnetron with stable temperature conditions of the transmitter remains practically constant; therefore it is not necessary to systematically monitor this parameter during operation.

Measurement of the transmitter power may be performed by means of a calorimetric average power meter, for example M3-4. It enables one to measure the radar average power in the limits from 5 to 2000 watts with an error of ± 7%. The power meter in this case is connected in place of the antenna and absorbs all of the transmitter power. Thus a calorimetric power meter does not allow one to measure the transmitter power while the transmitter is operating with a real load, that is, while the station is operating. Moreover the use of a calorimetric power meter for regular measurements is made difficult by the heavy weight and large overall size of the instrument.

Measurement of the transmitter power by a radar-tester through a directional coupler gives a measurement error amounting to 2 dB which exceeds, in a number of cases, the permissible measurement error. In order to reduce the measurement error it is advisable to use a calorimetric power meter as the more precise instrument for determining the systematic error which arises when measuring the power with the radar-tester.

Figure 2.24 shows the arrangement for measuring a transmitter power with consideration of what has been said above. The simultaneous measurement of the transmitter power by a calorimetric meter and a radar-tester is first performed (Figure 2.24, a). Then, during the operational process, power measurements are performed using the radar-tester but the systematic measurement error is taken into account (Figure 2.24, b). When this is done the measurement error does not exceed ± 10%. Measurement of the transmitter power may be achieved by means of radar test instruments. The measurement error of the average power when this is done does not exceed ± 0.8 dB, as a consequence of which one need not determine and take into account the systematic error of the instrument.

Measurement of an Antenna-Waveguide Channel TWR

In stations which have bidirectional couplers, the TWR may be determined by making measurements of the incident wave power and the

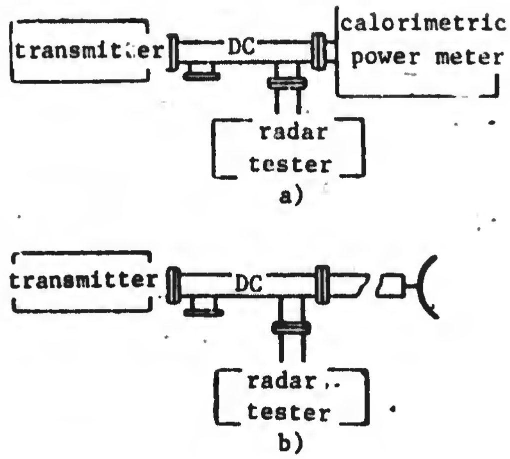


Figure 2.24. Arrangement for measuring a transmitter's power.

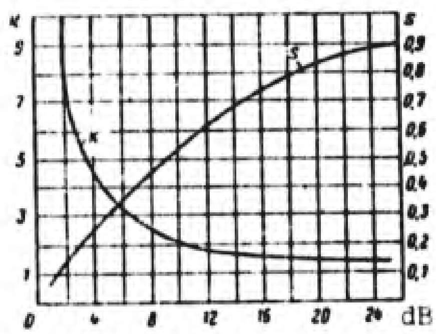


Figure 2.25. Graphs of TWR (s) and SWR (K) as a function of the difference of the forward wave power and the reflected wave power.

reflected wave power by means of a radar test instrument or a radar-tester.

The traveling wave ratio in connection with this may be calculated from the equation (2.10):

$$s=\frac{1-|\dot{\rho}|}{1+|\dot{\rho}|},$$

where $|\dot{\rho}|=\sqrt{\frac{P_r}{P_f}}$—is the modulus of the reflection coefficient;

P_r is the power of the reflected wave;

P_f is the power of the forward wave.

The precision of TWR measurements by this method is low and decreases with an increase of the absolute value of the TWR. Figure 2.25 presents graphs of the TWR and SWR as a function of the difference of the forward wave power and the reflected wave power expressed in decibels.

Determining the Frequency Energy Spectrum of Magnetron Oscillator Pulses

As was indicated above, a very important technical parameter which characterizes radar performence, is the frequency energy spectrum of the magnetron oscillator pulses.

Monitoring the frequency energy spectrum of the radiated pulses may be carried out with the help of the following portable instruments:

- a radar test instrument;
- a spectrum analyzer.

Under conditions aboard ship, a radar test instrument is usually used for determining the energy spectrum of the frequencies. Spectrum analyzers, because of their complexity and awkwardness, are used for making measurements principally in workshops. With normal operating conditions of a magnetron on a matched load, the energy spectrum of the frequencies approximates the ideal (design) spectrum.

In Figure 2.26 envelopes of the frequency spectrum of magnetron pulses which may be observed when a radar is operating are shown. Figure 2.26, a shows a good magnetron spectrum. The spectrum envelope is slightly blurred because of the limited pass band of the spectrum analyzer. The magnetron spectrum in Figure 2.26, b can be considered satisfactory. A spectrum of magnetron pulses which has the form of a blurred symmetric curve without expressed minimums is shown in Figure 2.26, c. A spectrum of such a type attests to a change in the shape of the modulating pulse. In this case it is necessary to check the shape of the modulator pulse and to establish the reason for its distortion. The spectrum of magnetron pulses shown in Figure 2.26, d has two space maximums. Skipping of the magnetron frequency within the pulse limits may be one of the reasons for the formation of two maximums. The frequency skipping is a consequence of a change in the magnetron operating conditions.

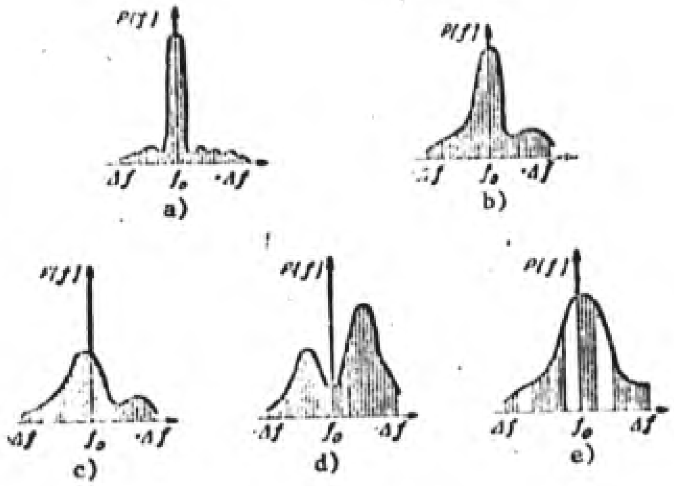


Figure 2.26. The frequency spectra of magnetron oscillator pulses.

This phenomenon is explained in the following way. The frequency of a magnetron oscillator, as follows from the load characteristics, depends both on the modulus of the reflection coefficient and also on its phase. Let us assume that the magnetron operates on a load which is not matched with the transmission line. In this case from the moment when generation begins and up to the moment when the electromagnetic wave being propagated along the line reaches the load and is sent back, the magnetron will be loaded with a load equal to the line wave impedence. The arrival of a reflected wave having a phase delay in respect to the wave being generated may be regarded as a change of the phase of the reflection coefficient and this must imply a change of the magnetron frequency.

A change of the magnetron frequency in turn leads to a change of the phase of the reflection coefficient. As a consequence of this, the magnetron high-frequency oscillations are found to be frequency modulated and several spaced maximums arise in the spectrum. Parasitic frequency modulation and noticeable distortion of the shape of the modulating pulse may also entail the appearance of a series of secondary maximums in the magnetron spectrum.

Figure 2.26, e shows a magnetron spectrum having lapses of particular spectral lines. The reasons for such lapses usually are: an increased steepness of the modulating pulse front, an insufficient amplitude or a misfiring of the submodulator pulses. Figure 2.27 illustrates the effect of the modulating pulse shape and of a change of its duty cycle frequency on the shape of the envelope of the frequency energy spectrum. As follows from the figure, a decrease in the steepness of the modulating pulse front leads to a decrease of the spectrum side lobes. In connection with this, the spectrum of a bell-shaped pulse has practically no side lobes.

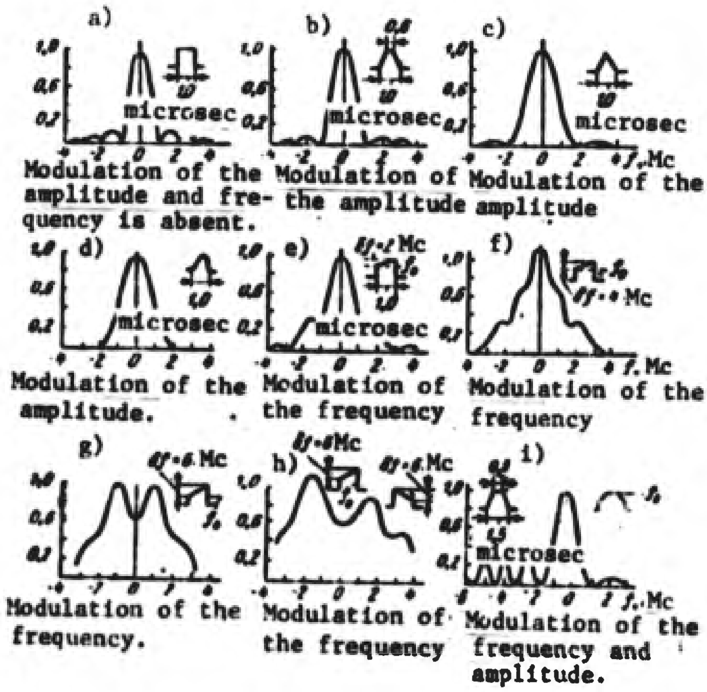


Figure 2.27. The shape of modulating pulses and the spectra corresponding to them.

A change of the pulse duty cycle frequency distorts the spectrum shape. With small frequency changes the spectrum has the form of a blurred curve without expressed minimums. With large frequency changes, several spaced maximums arise in the spectrum. When a distortion of the shape of the spectrum envelope is detected or gaps of the spectral lines are observed, it is necessary to check the modulating pulse shape and the magnetron operating conditions and to improve the matching of the transmitter with the antenna-waveguide channel.

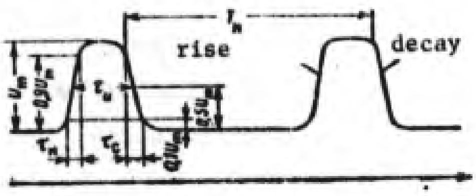


Figure 2.28. The main characteristics of a modulator pulse.

The main characteristics of modulating pulses are (Figure 2.28):

- the pulse amplitude U_m;
- the pulse duration τ_p;
- the rise time (front) of a pulse τ_r;
- the decay time of a pulse τ_d;
- the period T_p or pulse repetition frequency F_p.

The pulse duration τ_p (for pulses whose shape is close to rectangular) is determined at a level equal to half of the amplitude. The rise time τ_r and the decay time τ_d of a pulse are determined as the time during which the pulse changes in amplitude from 0.1 U_m to 0.9 U_m and from 0.9 U_m to 0.1 U_m, respectively.

The amplitude, pulse duration, rise time and decay time of a pulse are measured by means of an oscillograph. The pulse repetition frequency may be measured at the input of the submodulator or by measuring the pulse frequency of the magnetron current or the frequency of the pulses radiated by the transmitter.

2.7. Built-In Monitoring and Measuring Instruments

In the operation process, it is not always possible, and in a number of cases it is difficult, to measure radar technical parameters with portable instruments. Such measurements require a relatively long time and the availability of measuring instruments, and they do not allow one to carry out a continuous monitoring of the technical parameters of the equipment.

Frequently additional difficulties arise when using portable instruments which are connected with the inaccessible location of the equipment in stations. Therefore the necessity arose for creating built-in measuring instruments which constitute an integral part of the radar. These built-in instruments enable one to rapidly measure the main parameters characterizing the equipment normal operation with a sufficient degree of precision. Built-in monitoring-measuring equipment is developing in the direction of automatic devices which will carry out continuous monitoring of the technical parameters of a station, of its sub-assemblies and of its most important elements without disturbing the normal operation of the station.

The development of built-in measuring equipment and combination of it with computers opens broad possibilities for the automatization of radar checking. That is, the creation of devices will enable one to check the working order of all assemblies and instruments of the radar, generate a signal for the failure of a unit or element, and even give recommendations for the elimination of a malfunction.

Besides measuring instruments which enable one to determine the numerical values of the measured quantities expressed in acceptable units, monitoring instruments also have acquired great importance.

Monitoring instruments enable one to perform qualitative and rough quantitative evaluations of the process being studied or of a parameter. These instruments do not replace measuring instruments but supplement them. The main advantages of monitoring instruments are their continuous readiness, and the speed and convenience of monitoring station performance.

The main requirements which built-in measuring equipment must satisfy are the following:

1. When making measurements the normal operation of the radar should not be disrupted.

2. Continuous recording of the measurement results with signaling at the moment the technical parameter deviates from the limits of the norm is very desirable.

3. The reliability of built-in instruments should be greater than the reliability of the station being monitored; failures in the operation of the instruments should not affect the operation of the equipment.

4. The built-in instruments should have a specified precision for monitoring the parameters.

5. The instruments should couple simply with a station.

6. The overall size and weight of the instruments must be a minimum.

Let us investigate the principles of operation of a built-in instrument and the measurement method for controlling, with these instruments, the technical parameters which determine effective range of a radar.

Measurement of Transmitter Power

One of the main requirements which a built-in power meter must satisfy is the possibility of conducting measurements without disrupting the normal operation of a radar, that is, while the transmitter is operating with the antenna. Power flow meters satisfy this requirement only if the power consumed by them is small.

The power flow, that is the power which enters the antenna, equals

$$P_{\text{пр}}=P_{\text{п}}-P_{\text{о}},$$

where P_i is the incident wave power;
P_r is the reflected wave power.

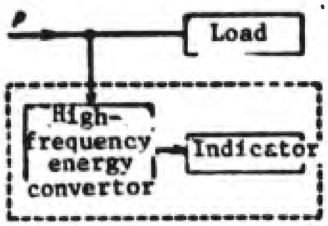


Figure 2.29. Arrangement for measuring the passing power.

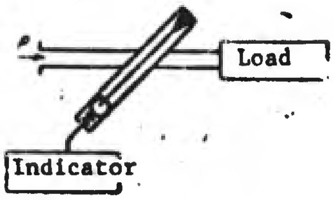


Figure 2.30. Arrangement for measuring the power of the incident (reflected) wave.

The arrangements for measuring power flow, incident power and reflected power are shown in Figures 2.29 and 2.30.

Recently diode power meters, power meters with an absorbing wall and power meters with a gas-discharge tube have been used in built-in instruments for measuring the power flow. In order to measure the incident (reflected) power, diode power meters are used.

A diode power meter has a directional coupler, a detector head and a peak voltmeter. The amplitude of the detected high-frequency pulses is measured with a diode meter. For making the measurements, unidirectional or bidirectional couplers having a coupling attenuation of no less than 30 dB are used. With a directional coupler, a part of the power proportional to the power of the incident or reflected wave is sent to the input of a high-frequency diode detector. Usually in such power meters, diodes of the type D605 are used.

Video pulses whose amplitude at a low power level is proportional to the power of the high-frequency pulses are sent to the peak detector from the output of the diode. In self-contained systems, the DC voltage from the output of the peak detector which is proportional to the high-frequency oscillation power is sent to a pointer-type indicator calibrated in units of pulse power. In automatic monitoring systems, this voltage is sent to processing and indicating units. The circuit of the simplest diode power meter is shown in Figure 2.31.

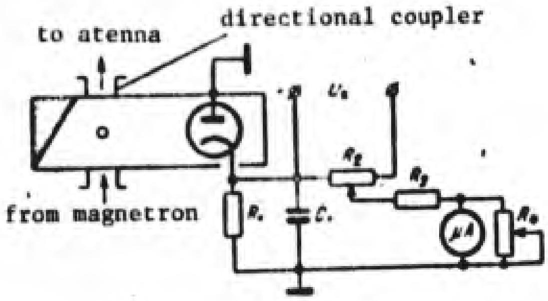


Figure 2.31. The circuit of a diode power meter.

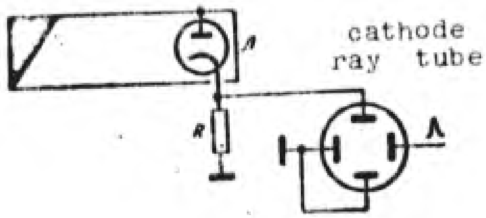


Figure 2.32. The circuit for turning on a detector head for the purpose of observing the envelope of high-frequency pulses.

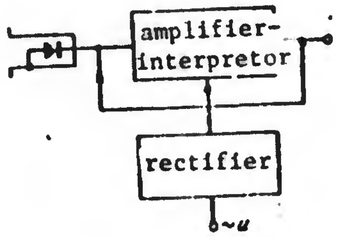


Figure 2.33. The block diagram of a diode power meter with a self-compensating peak voltmeter.

A pulse which is the envelope of the magnetron high-frequency pulse is taken from the diode cathode. Resistor R_1 and capacitor C_1 average the pulses entering from the detector head. The average component of these pulses is proportional to the magnetron average power.

The measuring instrument (MI) is calibrated in values of average power. During the intervals between pulses, the diode is cut off by a voltage U_k, which is supplied from a secondary power supply source. Since the power meter is a narrow-band meter, in order to obtain the power value when the frequencies are different from the frequency at which the instrument was calibrated, it is necessary to use a graph. With such a measurement method, the error reaches 40% and consists of the errors of the power meter and the directional coupler calibration. The instrument's calibration must be checked when a diode is replaced or with aging of the diodes.

The power meter detector head enables one to inspect the envelope of the high-frequency pulses. Switching on the detector head when observing the high-frequency pulse envelope is shown in Figure 2.32.

Recently, diode power meters with a self-compensating peak voltmeter (Figure 2.33) have begun to find broad application [4]. In the intervals between superhigh frequency pulses, the diode is cut off by a feedback voltage which is fed from the output of the integrating circuit of an amplifier. In that case and when the amplitude of the pulses fed to the diode exceeds the feedback voltage, the uncompensated part of the video pulse is amplified, integrated, and sent to the input of the voltmeter-converter and to the feedback circuit.

Such instruments enable one to measure the power of an incident or reflected wave. When it is necessary to measure the power flow, simultaneous measurement of the power of the incident and reflected waves is achieved and the DC voltages from the output of the two peak voltmeters are sent to a subtraction circuit. A voltage proportional to the value of the power flow is obtained at the output of the subtraction circuit.

A more modern instrument is a power flow meter with an absorbing wall.

<u>A power meter with an absorbing wall</u>. The power measurement in these instruments is based on the phenomenon of power absorption by the inner surfaces of waveguides and coaxial transmission lines when high-frequency energy is channeled through them. The instrument is a waveguide section, a small part of the wall of which is replaced by semiconductor thermoelements.

The thermoelements are mounted in the waveguide in such a way that they constitute part of the waveguide wall. As a consequence of this, there are no significant reflections in the transmission line. A second terminal of the thermoelements is in thermal contact with the waveguide wall and its temperature equals the waveguide temperature.

Figure 2.34 shows the construction of semiconductor thermoelements. Semiconductor thermoelements are made from alloys of zinc-antimony

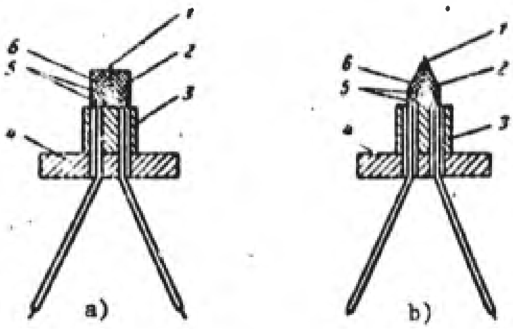


Figure 2.34. Constructions of semiconductor thermoelements: 1 - thermojunction; 2 - SbZn thermoelement; 3 - copper base; 4 - copper platform; 5 - Mica spacers; 6 - SbCd thermoelement.

(Zn-Sb) and cadmium-antimony (Cd-Sb). Thermoelements made of the zinc-antimony alloy have n-type conductivity and those of the cadmium-antimony have p-type conductivity. The use of thermoelements with different conductivity simplifies their connection circuit.

With the passage of high-frequency energy through the waveguide, the surface of the semiconductor thermoelement which is located in the waveguide section is heated. A thermoelectromotive force arises as a consequence of the temperature difference at the ends of the element. A linear relationship exists between the value of the thermo-emf and the temperature difference of the thermoelement ends. This relationship is

$$e=\alpha_{\tau}(t_1^{\circ}-t_2^{\circ}),$$

where α_T is the thermoelement sensitivity. A thermoelement made from the zinc-antimony alloy has a sensitivity of 200-250 microvolt/degree, and one made from the cadmium-antimony alloy has a sensitivity of 300-400 microvolt/degree.

When one thermoelement is mounted in a high-frequency transmission line, its emf will depend on the magnitude and phase of the reflection coefficient in the line and the wavelength. This means that the position of the nodes and antinodes of the standing wave will change depending on these factors. In order to eliminate the dependence of the emf on the phase and reflection coefficient, two thermoelements usually are used. These thermoelements are mounted so that the distance between them is $\lambda_W/4$ (Figures 2.35 and 2.36).

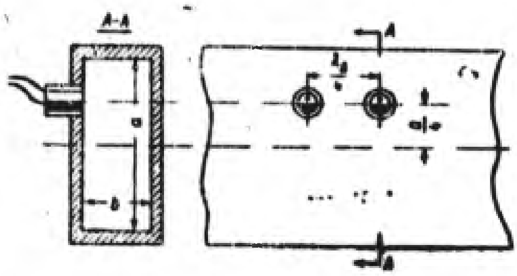


Figure 2.35. The position of the semiconductor thermoelements in the waveguide section.

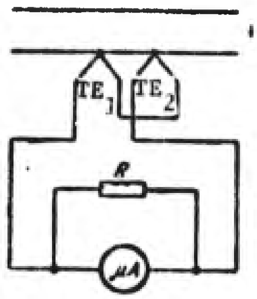


Figure 2.36. The circuit of the power meter with semiconductor thermoelements.

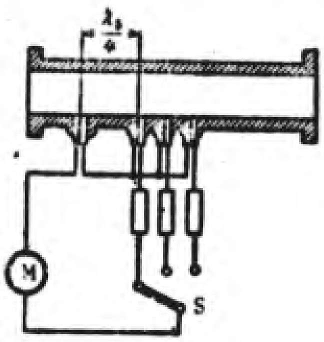


Figure 2.37. The switching circuit for semiconductor thermoelements.

In order to eliminate the dependence of the thermoelement emf on the wavelength, several additional thermoelements are also mounted in the waveguide section in addition to the two main thermoelements. One of the thermoelements is connected directly and the other thermoelements are connected through a switch S (Figure 2.37). With a change in the wavelength, switching of the additional thermoelements occurs in order that the distance between the main thermoelements and the additional thermoelement remain approximately equal to $\lambda_w/4$. Naturally such a device is expedient and may be realized only when the wavelengths are very small. A microammeter calibrated in values of the average power is used as the indicator for a self-contained measurement system. Furthermore in automatic monitoring systems, a DC amplifier is used in order to obtain the required value of the output signal.

Among the advantages of a built-in power meter with an absorbing wall, one must list the following:

- the possibility of achieving continuous monitoring of the power;

- the small measurement error (less than ± 15%); the instrument error is made up of errors connected with the change of the transmission line TWR, losses in the thermoelements and temperature changes;

- the broad range of measureable powers;
- the broad-band capacity;
- the operating stability and simplicity of the device.

The advantages of a power flow meter with sensors made from semiconductor thermoelements make such a meter very promising as a built-in instrument.

A power meter with a gas-discharge tube. In order to determine transmitter power, as a rule, the average power is measured and the pulse power is calculated from the equation

$$P_{и}=\frac{P_{ср}}{\tau F_{и}}.$$

Naturally when this is done one obtains the average power in the pulse. Because of the inertia of an average power meter, such a measurement method does not enable one to observe such phenomena as arcing in the magnetron and lapses of the pulses. These phenomena lead to unstable radar operation since they arise sporadically.

A power meter with a gas-discharge tube is a rapid-response pulse power meter. It is a glass tube filled with an inert gas (neon) at a pressure of 5-30 mm Hg. The tube is inserted in a waveguide (Figure 2.38). The gas-discharge tube is mounted along the axis of the wide wall of the waveguide. The action of the high-frequency

Figure 2.38. A gas-discharge tube of the type 1N524.

field causes an electrodeless high-frequency discharge in the tube. The height of the ionized glowing column in the tube associated with this discharge is proportional to the field intensity and consequently to the value of the power flow. In order to obtain a stable discharge in the tube, preliminary ionization of the gas is created. In one of the methods for creating the preliminary ionization an auxiliary pre-ionizing is used. For this purpose two auxiliary electrodes are mounted in the tube. An ignition voltage is fed to the auxiliary electrodes. Table 2.2 presents the basic data of the power meters with a gas-discharge tube which are manufactured by the firm CSF (France) [36].

The measurement error for the power meters presented in Table 2.2 equals ± 5% and the range of frequencies is 1-11 Ghertz. The measurement error to a significant degree depends on the matching of the power meter with the waveguide.

TABLE 2.2. DATA ON POWER METERS WITH A GAS-DISCHARGE TUBE

Tube Type	Level of Power Being Measured, milliwatt	Pulse duration τ_p, microsec
1 N 663	2	5
1 N 10	1	4
1 N 524	0.05	1

The height of the glowing column, and this means also the instrument sensitivity, increases with an increase of the insertion depth of the tube into the waveguide. However, with an increase of this depth the matching of the power meter with the waveguide deteriorates and the measurement error increases. Therefore, the tube insertion depth should be established in such a way that a reasonable compromise is obtained between the sensitivity and the error of the instrument.

Measurement of Antenna-Waveguide Channel TWR

The periodic monitoring of a high-frequency oscillation transmission line by means of portable instruments is no longer satisfactory with the increasing requirements for the monitoring of these lines. Because of this, continuous monitoring of transmission lines with built-in instruments has acquired even greater importance. Two methods for measuring the TWR are used in built-in instruments: by means of a bidirectional coupler and by means of semiconductor elements (the method of four probes) [4].

Measurement of the TWR by means of a bidirectional coupler. The method is based on the separate measurement of voltages which are proportional to the powers of the incident and reflected waves.

A self-contained built-in TWR meter consists of a bidirectional coupler and a ratiometer which measures the ratio of the voltages obtained. This ratio uniquely determines the transmission line TWR.

When it is necessary to send a signal from the TWR meter to the input of an automatic monitoring system, a logarithmic type indicator is unacceptable. In this case a measurement device is used which first generates the logarithm of the voltages which are proportional to the powers of the incident and reflected waves and then subsequently subtracts them. The meter (Figure 2.39) includes two identical devices, each of which consists of a directional coupler 2, 5, a self-compensating voltmeter-converter 3, 6, and a logarithmic device 4, 7.

Part of the power of the incident (reflected) wave is sent to a detector. The video pulses from the detector output are sent to the self-compensating voltmeter-converter. The DC voltage from the voltmeter-converter output, which is proportional to the power of the incident (reflected) wave, is sent to the logarithmic device. The voltages from the outputs of the logarithmic devices are sent to a subtraction circuit 8. The voltage difference is proportional to the TWR. Figure 2.25 presents graphs of an antenna-waveguide channel TWR and SWR as a function of the difference of the powers of the forward and reflected waves. The measurement error of the TWR with the instrument is larger, the smaller the value of the TWR being measured. The

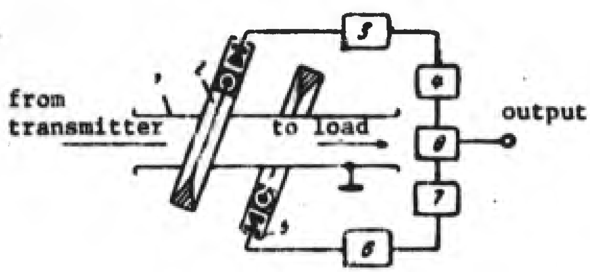


Figure 2.39. The block diagram for a TWR measurement using a bidirectional coupler.

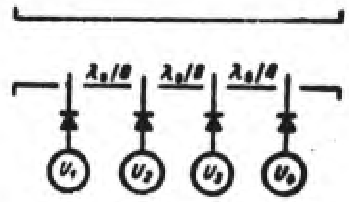


Figure 2.40. Measurement of the TWR by the four probe method.

error arises because the coupling attenuation of the directional couplers is not identical and because of errors of the self-compensating voltmeter-converters, of the logarithmic devices and of the subtraction device.

The total error of the built-in TWR meter equals approximately 15% when the value of the coupler directivity is no less than 35 dB

The method of four probes. Four probes with detectors are mounted in a waveguide section. The distance between the probes equals $\lambda_w/8$ (Figure 2.40). All the detectors have a square-law characteristic.

The voltage of the incident wave at the first detector will equal $U_1 \cos \omega t$ and the voltage of the reflected wave will equal $U_r \cos (\omega t + \theta)$.

The voltage at the output of the first detector is

$$U_1 = [U_n \cos \omega t + |\dot{\rho}| U_n \cos(\omega t + \theta)]^2;$$

If one disregards terms containing the high-frequency components of the signal then

$$U_1 = \frac{U_n^2}{2}(1 + |\dot{\rho}|^2 + 2|\dot{\rho}|\cos\theta).$$

Correspondingly the voltages taken from the remaining detectors will equal

$$U_2 = \frac{U_{\text{п}}^2}{2}(1 + |\dot{\rho}|^2 + 2|\dot{\rho}|\sin\theta).$$

$$U_3 = \frac{U_{\text{п}}^2}{2}(1 + |\dot{\rho}|^2 - 2|\dot{\rho}|\cos\theta).$$

$$U_4 = \frac{U_{\text{п}}^2}{2}(1 + |\dot{\rho}|^2 - 2|\dot{\rho}|\sin\theta).$$

From these expressions one is able to determine the modulus of the reflection coefficient and consequently also the transmission line TWR.

The outputs of the probes are connected so that the following emf differences and sums are obtained:

$$U_1 - U_3 = 2|\dot{\rho}|U_{\text{п}}^2\cos\theta,$$

$$U_2 - U_4 = 2|\dot{\rho}|U_{\text{п}}^2\sin\theta,$$

$$U_1 + U_3 = U_2 + U_4 = U_{\text{п}}^2(1 + |\dot{\rho}|^2).$$

If the voltage differences are fed to the deflection plates of a cathode-ray tube and the amplitude of the incident wave is maintained constant, then the position of the bright spot on the tube screen will be determined by the coordinates

$$x = k|\dot{\rho}|\sin\theta \text{ and } y = k|\dot{\rho}|\cos\theta.$$

The radius connecting the center of the tube with the spot will graphically represent the reflection coefficient modulus which equals

$$|\dot{\rho}| = \frac{\sqrt{(U_1 - U_3)^2 + (U_2 - U_4)^2}}{2U_{\text{п}}}.$$

Knowing the modulus of the reflection coefficient, one is able to determine the TWR from equation (2.10).

In built-in instruments instead of probes with detectors one may find semiconductor thermoelements being used [4]. When measuring TWR's which lie in the limits from 0.5 to 0.95, the error of such instruments does not exceed 10%.

Measurement of Receiver Noise Factor

In order to measure receiver sensitivity, noise generators are used in built-in instruments. The main defects of standard signal generators which hinder their use as built-in instruments for monitoring receiver sensitivity are:

- the high measurement error which reaches 2 dB;
- in the continuous oscillation mode, the indicator device is not checked; in the pulsed mode the measurement errors increase (by 1.5-1 dB) because of the subjective error of the operator when determining the signal/noise ratio on the indicator screen;
- the narrow range;
- the servicing complexity;
- the difficulty of monitoring automation;
- the considerable size and weight of the instrument.

In accordance with equation (2.7) the noise factor equals

$$F = \frac{P_{\text{сош}}}{kT_0 \Delta f}.$$

Consequently if a calibrated power P_c equal in magnitude to the equivalent power of the receiver internal noise P_{sog} is sent to the receiver input, the noise factor may be found from the expression

$$F = \frac{P_c}{kT_0 \Delta f}.$$

The noise factor of actual receivers of the meter range equals 2-10, of the decimeter range 10-16, and of the centimeter range 16-160.

If the receiver noise factor is known, its actual sensitivity may be determined from equation (2.2). Thus, in order to monitor the sensitivity, it is sufficient to monitor the receiver noise factor since during operation the receiver pass band and distinguishability coefficient remains practically constant.

From the point of view of making measurements, the direct determination of the noise factor has a number of advantages over the measurement of the receiver sensitivity using a standard signal generator, more precisely:

- the noise generator power has the necessary value which is required for checking a receiver device and therefore it does not require significant attenuation of the noise signal. Consequently, the use of attenuators is not necessary;
- the output power level of a noise generator may be determined with a high accuracy by means of calculation which simplifies generator calibration;
- the generator noise has the same character as the receiver internal noise and therefore the reaction of the indicator instrument to this noise corresponds to the reaction of the instrument to the receiver noise;
- the high stability of the power being supplied;
- the simplicity of the device.

A noise generator is a source of noise, the level of which is known with a sufficient degree of accuracy. Noise diodes and gas-discharge tubes are used as such a source.

Figure 2.41 shows the construction of a noise diode in a coaxial arrangement with an axial cathode. The cathode and anode of such a diode form the inner and outer part of the coaxial line. The diode is matched with the line by means of transforming elements.

The thermal motion of electrons which are found in the positive column of the gas discharge is the source of noise in gas-discharge tubes. The output power of gas-discharge tube noise exceeds the power

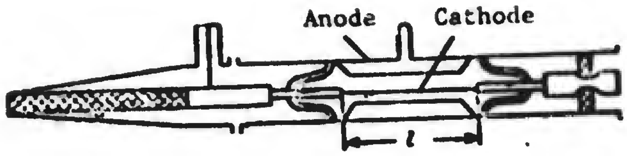


Figure 2.41. The construction of a noise diode in a coaxial arrangement.

of the thermal noise by 15-18 dB, that is, the noise power generated by the same resistance at the temperature 290°K. The noise power depends on the kind of gas, its pressure and the diameter of the gas-discharge tube. It varies only slightly with a change of the frequency, of the discharge current and of the ambient temperature. This makes a gas-discharge tube a very stable noise source.

Noise generators are waveguide generators or coaxial generators. In the range 2.6-10 Ghertz waveguide noise generators are used, and in the range 1-4 Ghertz, coaxial noise generators are used. In the waveguide noise generator, the gas-discharge tube is mounted inside the waveguide at an angle to its axis. This ensures the matching of the gas discharge with the waveguide over a broad range of frequencies (Figure 2.42).

The ends of the tube which extend beyond the waveguide are terminated in metal cylinders which are waveguide sections with a diameter smaller than the critical diameter. In this way, the region of the gas discharge creating the noise is limited. The frequency range in which the generator operates is determined by the construction of the waveguide.

In a coaxial noise generator, the gas-discharge tube is placed inside a ribbon spiral. The spiral is the inner conductor of the coaxial line; the cylindrical sleeve in which the gas-discharge tube is placed (Figure 2.43) serves as the outer conductor. The step of the spiral, the size of the gap between adjacent turns and the diameter of the spiral determine the wave impedence of the line and the

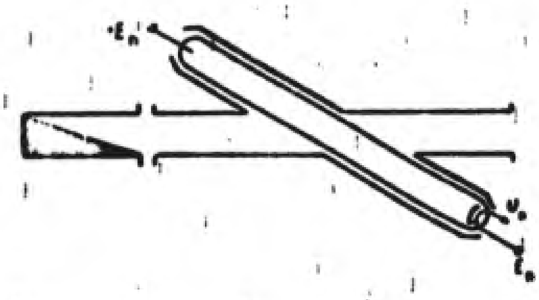


Figure 2.42. The construction of a waveguide noise generator.

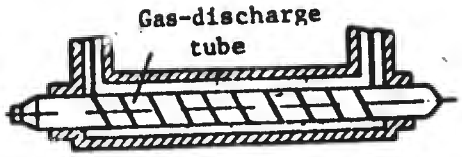


Figure 2.43. The construction of a coaxial noise generator.

attenuation which the ignited tube introduces into the waveguide channel.

The advantages of the noise generator in comparison with other noise sources are the following:

- the high stability of the noise parameters;
- the sufficient noise power;
- the large TWR (0.83 - 0.9) over a broad frequency range;
- simple construction.

Built-in instruments for measuring the noise factor may be subdivided into instruments with an unmodulated noise signal and with a modulated noise signal.

In the case of an unmodulated noise signal, all the power of the noise signal is sent to the radar high-frequency circuit and a comparison is made of it with the power of the receiver internal noise. A diode noise generator or a gas-discharge noise generator may be used as the signal source. When a diode noise generator is used, its operating conditions may be varied by changing the diode current. Gas-discharge noise generators are used with two operating conditions: the noise generator is switched on and the noise generator is switched off. One of the possible arrangements for such a measurement of a receiver noise factor is shown in Figure 2.44. A gas-discharge tube

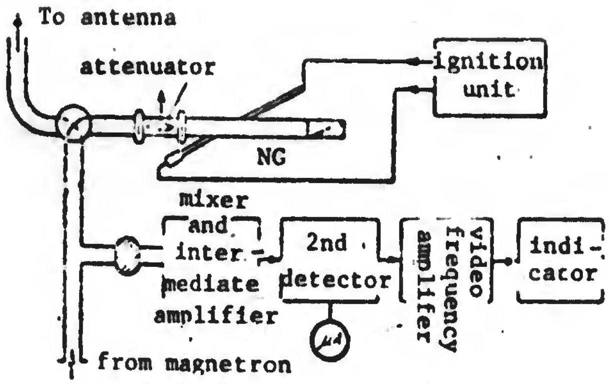


Figure 2.44. The block diagram for measuring a receiver noise factor by feeding an unmodulated noise signal to its input.

serves as the noise generator. An attenuator which enables one to vary the noise power is connected to the output of the generator. An instrument connected to the circuit of the receiver second detector serves as the indicator. Before making the measurement, the automatic volume control (AVC) and automatic frequency control (AFC) circuits are switched off. When making the measurement, the value of the current of the second detector is recorded when the noise generator is not operating and the attenuator is fully inserted. Then the noise generator is switched on and the attenuator is withdrawn to a position for which the microammeter readings increase by a factor of 2 for a linear characteristic of the detector and by a factor of 1.41 with a square-law detector. This corresponds to an equivalence of the noise signal power supplied by the generator and the receiver noise power. The receiver noise factor is determined from the expression

$$F = P_{\text{гш}} - \alpha_{\text{атт}},$$

where P_{ng} is the power of the noise generator, in dB;

α_{att} is the attenuator attenuation, in dB.

Such a method for measuring a receiver device noise factor has been designated the attenuator method. In this instrument, the attenuator must have a small error when making relative measurements and a constant output impedance, because a change of the output impedance will lead to a change of the readings of the indicator instrument.

The instrument may not have an attenuator. In this case auxiliary graphs or tables are used. The quantity equal to the ratio of the instrument current, with the noise generator switched on, to the instrument current, with the noise generator switched off, is used as the input data of the table. This measurement method is called the two -reading method.

Measurements with an unmodulated noise signal fed to the receiver input are distinguished by their simplicity; however, they do not enable one to determine the noise factor while the radar is operating. Moreover, since during the measurement process, the antenna is disconnected from the receiver channel and the transmitter is switched off, the quantity actually measured differs slightly from the true value of the receiver sensitivity. Therefore, the measurement error of the noise factor may reach 2.5 dB.

The significant error when measuring large values of the noise factor also is a defect of this method. These defects do not occur in instruments which operate on the basis of the modulated method for measuring the noise factor. With this measurement method, the noise signal is fed to the receiver input through a directional coupler. In this case, the measuring noise signal is smaller by several orders of magnitude than the receiver internal noise and the measurement process does not disturb the radar normal functioning.

In order that the noise signal may be distinguished from the receiver background noise, it is modulated according to some law. This modulation is achieved by periodically switching the noise generator on and off. During the modulation half-cycle when the noise generator is not switched on, a noise signal from the antenna, $N_1 = kT_0\Delta f$, equal to the noise power of the output resistance of the noise source at the absolute temperature T_0, is sent to the receiver input. During the modulation half-cycle when the noise generator is switched on, the noise signal $S_1 = k(T_1-T_0)\Delta f$ is fed to the receiver input.

Substituting the expressions for N_1 and S_1 into equation (2.6), we obtain

$$F=\frac{\frac{T_1}{T_0}-1}{\frac{S_2}{N_2}},$$

where $\frac{T_1-T_0}{T_0}$ is the relative noise temperature of the noise source;

T_1 is the absolute temperature of the electron gas in the gas-discharge tube;

T_0 is the absolute temperature of the surroundings.

Since the values of T_1 and T_0 are known the noise factor measurement is reduced to a measurement of the ratio

$$\frac{S_2}{N_2}, \text{ i.e. } F=R\frac{N_2}{S_2},$$

where R is the proportionality constant.

All the known modulated noise factor meters differ in principle only in the method of measuring the ratio S_2/N_2.

Let us investigate one of the possible arrangements for a built-in modulated noise factor meter (Figure 2.45). A gas-discharge tube whose noise signal power is close to the power of the receiver internal noise is used as the calibrated power generator in the circuit. The calibration error of gas-discharge tubes equals approximately ± 0.5 dB. The signal from the gas-discharge tube is modulated by rectangular pulses supplied by a pulse generator. The noise signal is fed through a directional coupler to the radar receiver.

The directional coupler coupling attenuation is selected on the basis of the condition for ensuring the normal function of the radar

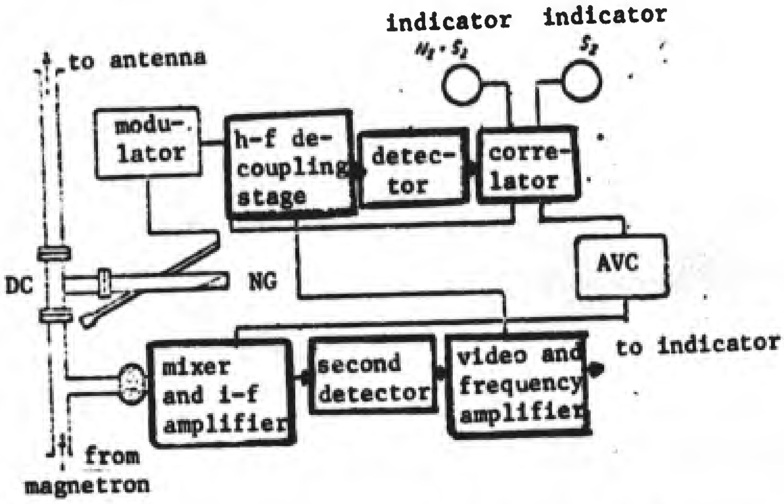


Figure 2.45. The block diagram of a modulated noise factor meter.

in which the noise factor is built in. The noise signal from the gas-discharge tube should not interfere with the normal observation of signals from targets on the indicator screens. In addition, the noise signal should have a power which is sufficient for the normal operation of the instrument. Usually the power of the noise measuring signal is smaller by one to three orders of magnitude than the receiver internal noise power. The modulation of a standard noise generator connected to the input of a receiver device is common to all modulated noise meters.

During the half cycle of the modulation when the tube is not ignited, a noise signal with the power N_1 is fed to the receiver. During the other half of the cycle, the tube is ignited and a noise signal with the power $N_1 + S_1$ is sent to the receiver input.

Figure 2.46 shows the form of the signals in the circuit of a modulated meter. AT the output of the intermediate frequency amplifier the signal is the sum of the receiver internal noise and the noise measuring signal.

The noise voltages in the various cycles will respectively equal $\sqrt{N_1}$ and $\sqrt{N_1+S_1}$. The receiver internal noise superimposed on the

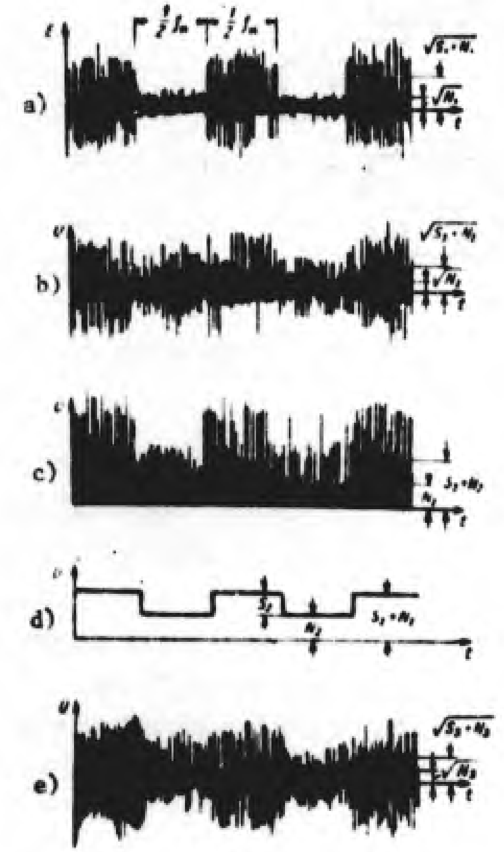


Figure 2.46. Diagrams of the voltages at the output of the main stages of the circuit of a modulated noise factor meter: a - the signal at the input of the receiver; b - the signal at the output of the intermediate frequency amplifier; c - the signal at the output of the receiver second detector; d - the low-frequency component at the output of the second detector; e - the component of the intermediate frequency at the output of the second detector.

noise from the noise generator decreases the modulation intensity. The noise signal from the output of the intermediate frequency amplifier is sent to the second detector which has a linear or a square-law characteristic. With a square-law characteristic the intensity of the video noise at the output of the receiver second detector will be N_2 and $S_2 + N_2$.

In addition to the modulated noise voltage, the voltage taken from the output of the second detector contains noisy signals and therefore it is first sent to a decoupling stage.

The amplified modulated video-frequency noise is sent through the decoupling stage to the detector and from the output of the detector to a correlator. A simplified arrangement of the mechanical correlator and commutator of the modulator is shown in Figure 2.47. The correlator is used for separating out voltages proportional to $N_2 + S_2$ and N_2. These voltages are sent to the noise factor indicator and the fixed level indicator. The modulator

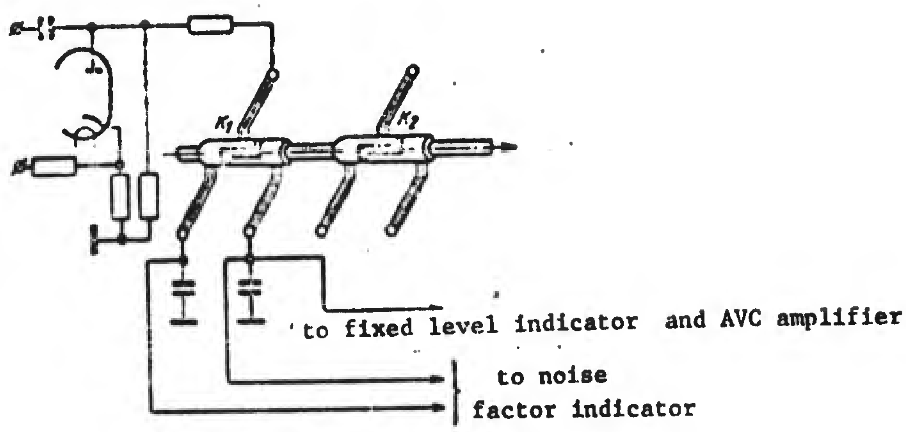


Figure 2.47. The arrangement of the mechanical correlator and commutator of the modulator.

switches the noise generator on and off.

The mechanical correlator and commutator of the modulator consists of two four-section collectors K_1 and K_2 which are mechanically driven by a common motor.

The voltage from the detector is fed to the middle brush of collector K_1. The following are taken from the brushes which slide along the rings: a voltage proportional to $N_2 + S_2$ which is sent to the fixed level indicator and a voltage proportional to the difference of $N_2 + S_2$ and N_2 which is sent to the noise factor indicator. A voltage from the auxiliary collector K_2 is supplied for triggering the modulator of the noise generator. The voltage going to the fixed level indicator also is sent to the amplifier of the receiver automatic volume control (AVC) by means of which the $N_2 + S_2$ level is maintained constant. With an increase of the receiver internal noise, the AVC circuit reduces the amplification of the receiver in order to maintain the $N_2 + S_2$ level unchanged which leads to an increase of the role of the noise signal S_2 in the composition of the signal $N_2 + S_2$. With this the S_2 indicator (Figure 2.45) shows an increase of the receiver noise factor.

When measuring a receiving device noise factor, it is necessary that the amplitude of the noise generator signal not pass beyond the limits of the linear part of the receiver amplitude characteristic. Nonobservance of this condition leads to an increase of the measurement error.

In superheterodyne type receiver devices, the noise signal may pass both along the main channel and also along the image channel. The noise factor in this case may be determined from the expression

$$F=F_0\left(1+\frac{1}{d}\right),$$

where F_0 is the noise factor obtained when making the measurement; d is the image channel selectivity of the receiver device.

The modulation method for measuring the noise factor has a number of significant advantages:

- a continuous indication of the receiver noise factor is possible which enables one to opportunely note a decrease of the receiver sensitivity;

- the noise factor is determined when the radar connected to the antenna and the radar transmitter is operating;

- there is a high measurement accuracy.

The modulation measurement method also is used in portable noise factor meters (NFM) which are designed for measuring the noise factor of radar receiver-amplifier devices under laboratory and field conditions. It is necessary to keep in mind that, since the noise generator of the noise factor meter generates a broad frequency spectrum, it is impossible to check the receiver tuning with it. In order to check the tuning it is necessary to also have an echo chamber (monitoring resonator) in the radar.

Monitoring the Magnetron Current and the Mixer Current

Besides built-in instruments which enable one to measure the transmitter power and the receiver noise factor and consequently to determine the power drop in radar stations, radars have, as a rule, simple instruments which enable one to monitor the technical state of the individual devices of the radar. Among such instruments, built-in pointer-type instruments which measure the magnetron current and the mixer current ("the crystal current") are very useful.

The magnetron current to a specific degree characterizes magnetron operation. Figure 2.48 presents the performance characteristics of a magnetron oscillator which reflect the dependence of the magnetron parameters (the power, efficiency and frequency) on the voltage, current and magnetic induction with a constant load. From these characteristics, with given values of the magnetic induction and anode voltage, one is able to uniquely determine the anode current, power and efficiency of a magnetron. Therefore with a constant value of the induction the anode current characterizes the power supplied by the magnetron. Monitoring the magnetron operating conditions may also be performed, in principle, on the basis of the voltage. However, the small slope of the volt-ampere characteristics indicates a strong dependence of the anode current on the voltage, as a consequence of which it is more advisable to monitor the magnetron operating conditions on the basis of the current.

It is also very important to monitor the values of the mixer current, since the receiver maximum sensitivity corresponds to the optimum value of this current. This follows from the following considerations:

The conversion losses and the noise temperature are the main parameters of a mixer on which the receiver sensitivity depends.

The conversion losses L are defined as the ratio of the high-frequency signal power P_{hf} to the power of the converted intermediate-frequency signal P_{if}:

$$L = \frac{P_{вч}}{P_{пч}}$$

or

$$L = 10 \lg \frac{P_{вч}}{P_{пч}} \text{ dB}.$$

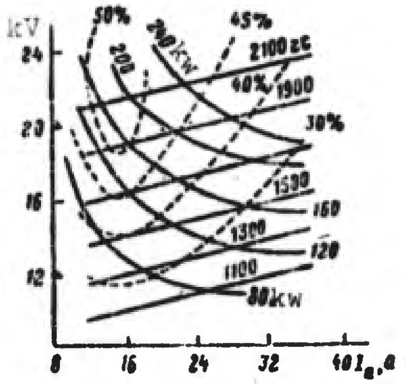


Figure 2.48. The performance characteristics of a magnetron oscillator.

From this expression it follows that with a given value of the converted signal power the required power of the high-frequency signal is smaller, the smaller the conversion losses; consequently the smaller the conversion losses, the higher the receiver sensitivity will be.

The conversion losses depend on the amount of power supplied to the mixer from the heterodyne. With an increase of the supplied power, the conversion losses decrease. The reciprocal of the conversion loss $k_p = \frac{1}{L}$ is called the power transmission coefficient.

The noise temperature t_m characterizes the mixer noise properties and is defined as the ratio of the nominal noise power P_m in the diode when a current in the frequency interval Δf passes through it to the nominal power of the thermal noises of an equivalent resistance at room temperature T_0 in this same frequency interval. A relationship exists between the mixer noise factor F_m and the noise temperature t_m which is

$$F_c = \frac{t_c}{k_p}.$$

The noise temperature also depends on the heterodyne power which is supplied to the mixer. With an increase of the heterodyne power the noise temperature and consequently also the noise factor of the mixer increase.

Thus, in order for the station mixer to operate normally, it is necessary to feed a sufficient power to it from the heterodyne.

However, this does not mean that the larger this power the higher the receiver sensitivity. Figure 2.49 presents graphs of the power transmission coefficient k_p and the noise temperature t_m of a silicon mixer as a function of the constant component of the mixer current. Since the mixer current constant component is determined by the heterodyne power, the graphs have the same form as if they were a function of the heterodyne power. As follows from Figure 2.49, optimum values of the power transmission coefficient and the noise temperature occur with a specified value of the heterodyne power. These values correspond to the receiver maximum sensitivity. The necessity for monitoring a mixer diode current follows from this. When this current deviates from the norm, its rated value is reestablished by adjusting the coupling between the heterodyne and the mixer. The constant component of the mixer current ("of the crystal current") usually equals 0.4 - 0.5 mA.

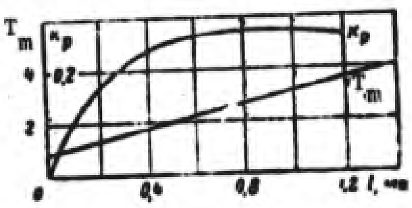


Figure 2.49. Graphs of the power transmission coefficient and noise temperature of a silicon mixer as a function of the magnitude of the constant component of the diode current.

Monitoring the Attenuation in an Antenna-Waveguide Channel

Monitoring the attenuation in an antenna-waveguide channel may be achieved by comparing the values of the incident power and the reflected power when the channel is short-circuited by a special waveguide switch. For this purpose a built-in power meter must be connected by turns to the outputs of a bidirectional coupler.

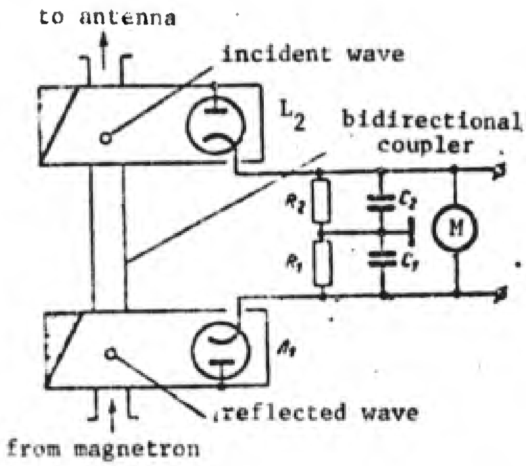


Figure 2.50. The circuit for connecting the detector heads for monitoring the attenuation in a waveguide channel.

With the connection of two detector heads in the instrument, one is able to directly monitor the attenuation in the channel (Figure 2.50).

Monitoring of the antenna-waveguide channel's attenuation may also be achieved by monitoring the high-frequency power at the beginning and the end of the channel by means of power meters with sensors of semiconductor thermoelements. In this case one is able to judge as to the constancy of the attenuation value on the basis of the constancy of the power difference at the beginning and end of the waveguide channel. An attenuation measurement based on the second method is somewhat more precise.

Monitoring the Energy Spectrum of the Frequencies of Magnetron Oscillator Pulses

In order to monitor the frequency energy spectrum of magnetron oscillator pulses, built-in spectrum analyzers have begun to be used in radar equipment [45].

They are assembled on the basis of the circuit of a superheterodyne tunable receiver with a cathode-ray indicator. In built-in spectrum analyzers certain assemblies and units of the radar are used, for example, the heterodyne, mixer, cathode-ray tube and power supply. In this case one can succeed in considerably reducing the overall size and weight of the instrument.

Monitoring the Operation of the Automatic Frequency Control Circuit

Receiver devices of modern radar stations would be unthinkable without a number of automatic adjustment systems. The most important of these systems is the automatic frequency control (AFC) circuit. During the station operation, because of changes of the transmitter frequency and the receiver adjustment, the frequency difference may differ to a considerable degree from the intermediate frequency to which the receiver was tuned. This may lead to a reduction of the station effective range.

The necessity for AFC is explained basically by two reasons:

- the presence of asymmetry in the rotary waveguide junctions and scanning heads and also the presence of variable reflections, for example from the ship superstructure and masts with the rotation of the antenna, which change the phase of the reflection coefficient and the TWR at the input of the antenna-waveguide channel. This leads to a change of the magnetron frequency;

- a change of the dimensions of the oscillatory systems of the magnetron and klystron because of mechanical events and changes of the temperature and pressure, and also a change of the power supply conditions. All these lead to a detuning of the receiver and transmitter.

Manual adjustment of the frequency in these cases does not achieve a fixed difference of the frequency since, first, its rate is low, and second, with frequency drift it is impossible to notice at once a decrease of the receiver sensitivity. Consequently normal operation of the automatic frequency control circuit is of very great importance. In principle, frequency tuning may be achieved by changing the frequency of either the transmitter or the heterodyne..

Two arrangements are used for connecting the AFC with the transmitter in stations: through a discharger or through a special attenuator. In modern stations the circuit for connecting the AFC with the transmitter through a special attenuator is being used. Based on the character of the operation, the AFC circuits are made as scanning circuits or nonscanning circuits. Scanning devices with a high tuning rate and ensuring a frequency tuning within broad limits have found the broadest application.

The block diagram of a receiver with automatic frequency control of the heterodyne is shown in Figure 2.51.

The AFC channel includes the following elements:
- a power divider (attenuator);
- an AFC mixer;
- an intermediate-frequency amplifier;
- a discriminator;
- adjusting devices.

As was indicated above, among the main operations which must be systematically conducted for monitoring the constancy of an effective range of the station, one must list the monitoring of the operation of the AFC circuit. A tuned and well operating AFC circuit is characterized by the following properties:

1. The receiver sensitivity in the AFC mode should be no lower than when operating in the manual mode for frequency adjustment. In order to check this, a distant target is selected and manual tuning of the receiver to a signal maximum is performed. In switching from

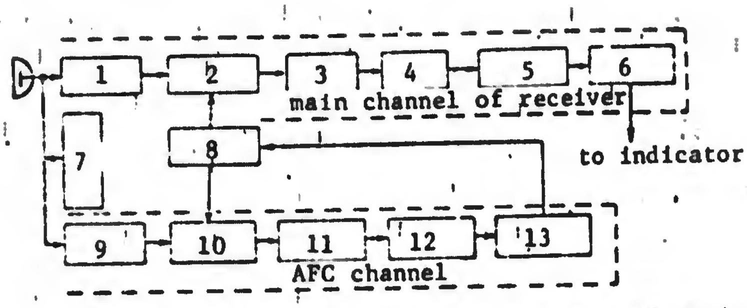


Figure 2.51. Block diagram of a receiver with automatic frequency control.

1 - discharger; 2 - mixer; 3 - i-f pre-amplifier; 4 - i-f amplifier; 5 - detector; 6 - video frequency amplifier; 7 - transmitter; 8 - heterodyne; 9 - power divider; 10 - AFC mixer; 11 - AFC amplifier; 12 - discriminator; 13 - adjustment device.

manual adjustment of the frequency to AFC, the value of the signal should not decrease below the maximum obtained in the "manual adjustment" situation. When it is impossible to carry out a checking of the distant target, the monitoring of the receiver sensitivity in the AFC mode may be performed by means of a monitoring resonator.

2. The search system must operate normally. This may be checked in the following way. When the AFC circuit is switched on and the transmitter is switched off, the pointer of the instrument which shows the mixer diode current ("crystal current") must fluctuate ("sweep") with the frequency of the adjustment device search.

3. The holding band of the AFC should correspond to the technical conditions of the radar.

The holding band is determined by the extreme frequencies at which the AFC circuit guarantees a holding of the frequency. In order to check the holding band, the klystron is mechanically detuned to both sides of the optimum tuning until the AFC passes from a following mode to a search mode. The klystron frequencies corresponding to the transition of the circuit to the search mode are measured.

For normal operation of the radar, it is necessary that the frequency difference of the transmitter and heterodyne be maintained

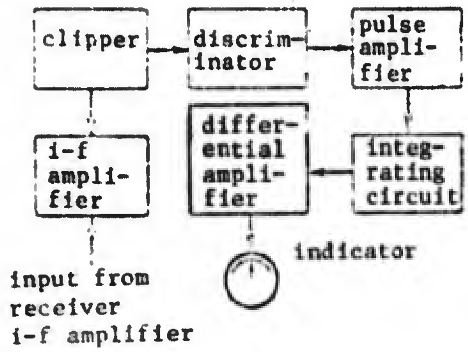


Figure 2.52. Block diagram for monitoring the relative tuning of a receiver.

equal to the intermediate frequency to which the receiver was tuned.

A built-in instrument for monitoring the receiver relative tuning must show with what precision the AFC system maintains the difference frequency. Figure 2.52 shows one of the possible block diagrams of such an instrument.

The tuning indicator shows the difference between the nominal value of the intermediate frequency and the actual frequency of the signal arriving from the i-f amplifier.

Monitoring Radar Power Drop

Monitoring the constancy of the radar effective range, as was shown above, may be performed by the separate measurement of certain technical parameters.

In connection with everyday operation, the method of evaluating the radar power potential or power drop, which determines its effective range, also is very promising.

In order to monitor the station power drop, it is necessary to introduce a monitoring signal into the station circuit. The parameters of this monitoring signal must correspond to the parameters of the transmitter high-frequency pulse. In the general case, the monitoring signal may be introduced into the station antenna-waveguide channel, into the i-f amplifier circuit or into the antenna (Figure 2.53). The voltage ratio of the monitoring signal to the noise at the receiver output when its amplitude characteristic is known will also characterize the radar power drop or power potential.

In principle two methods of forming the monitoring signal are possible. The first method is based on the shaping of an imitation monitoring signal, and the second method is based on forming a monitoring signal from the main pulse.

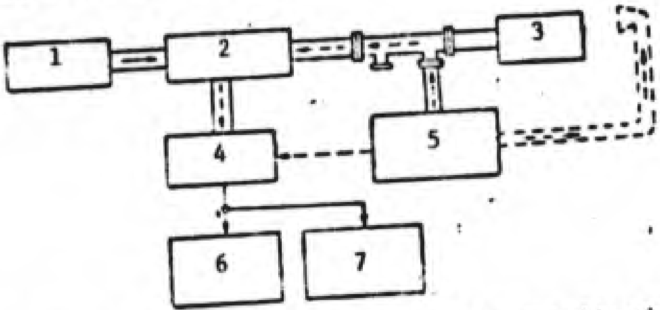


Figure 2.53. Methods for feeding the monitoring signal when determining radar power drop.

1 - transmitter; 2 - antenna switch; 3 - antenna;
4 - receiver amplifier; 5 - monitoring signal
generator; 6 - indicator; 7 - signal/noise ratio meter

The first method provides for the use of a high-frequency monitoring pulse imitator. The imitator is a high or intermediate-frequency monitoring pulse oscillator; the parameters of the monitoring pulse sequence must correspond to the parameters of the main pulses.

Since the main pulse level may vary during the operation process, the correlation between the parameters of the monitoring pulses and main pulses should be stipulated. The monitoring signal should vary according to the law governing the change of the main signal, that is, a system should be provided which would ensure the automatic adjustment of the monitoring signal level to the main pulse level.

When the main pulse level varies insignificantly during the operation process, the construction of the monitoring signal generator may be simplified since there is no need to provide an automatic adjustment system. In connection with this, not the change of the station power drop will be monitored, but the change of the receiver sensitivity.

The advantages of the method for monitoring the station power drop by using an imitation monitoring signal are:

- the small measurement error not exceeding 3 dB;
- the wide-band capacity;
- the possibility for the continuous measurement and monitoring within broad limits of the power drop.

Among the defects are the complexity of the instrument and its great weight and overall size.

The second method provides for the direct attenuation of the transmitter pulse and the introduction of it into the station circuit as the monitoring signal.

Monitoring of the station power drop by such a method may be performed with the help of a monitoring resonator (echo chamber) or a short-circuit delay line.

Let us discuss in more detail the devices and technical possibilities of these instruments.

Monitoring a Station Technical Performance by Means of a Monitoring Resonator

A monitoring resonator, or an echo chamber, is a high-quality cavity resonator which is connected through a coupling element with the radar (Figure 2.54). It consists of a metal cylinder in which a tuning piston moves. The retuning mechanism enables one to change the position of the piston and to tune the cavity resonator to the frequency of the oscillations being supplied to it. The scale of the tuning instrument is calibrated in relative units. A detector section is connected with the resonator through a second coupling element. The detector is loaded by a microammeter which is used as the tuning indicator of the resonator.

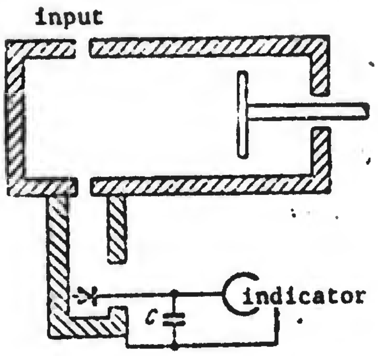


Figure 2.54. Schematic diagram of a monitoring resonator (echo chamber).

The instrument operation is based on the following principle: high-frequency signals from the radar are sent through the coupling element to the resonator and excite oscillations in it. The oscillations arising in the resonator do not die out instantaneously after the transmitter pulse ceases, but they are maintained (as a consequence of the resonator's high quality) for a certain time which considerably exceeds the pulse duration. High quality of the resonator is achieved by the selection of the type H_{01} working wave, by the good conductivity of th inner surface, by the correct selection of the resonator dimensions and by precise manufacturing.

Selection of the type H_{01} wave is explained by the fact that, for oscillations of this mode, the resonator quality is several times larger than for other modes. The quality of a loaded resonator for waves of the 10-cm range equals approximately 10,000, and for waves of the 3-cm range, 20,000. As a consequence of this, upon determination of the transmitter pulse, the resonator itself becomes the source of oscillations. These oscillations are sent through a coupling element to the station receiver circuit. An image of the signal reemitted by the resonator is displayed on the station screen (Figure 2.55).

Figure 2.56 shows time diagrams illustrating the principle of operation of a monitoring resonator. The transmitter high-frequency pulses are shown in Figure 2.56a. Oscillations in the resonator increase according to an exponential law during the generating time of the high-frequency pulse by the transmitter. After termination of the pulse, the oscillations also will die out according to an exponential law as is shown in Figure 2.56b. As a consequence of saturation

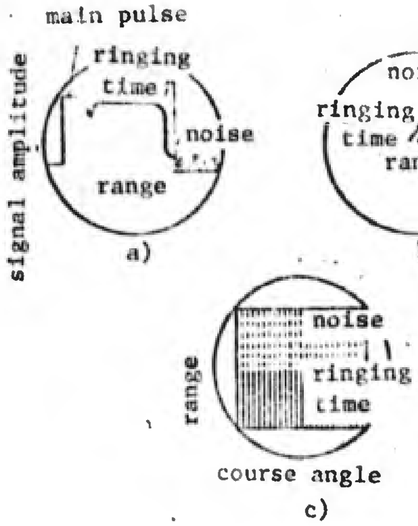


Figure 2.55. Signal of the monitoring resonator in various types of indicators: a - linear sweep; b - radial-circular sweep; c - line sweep.

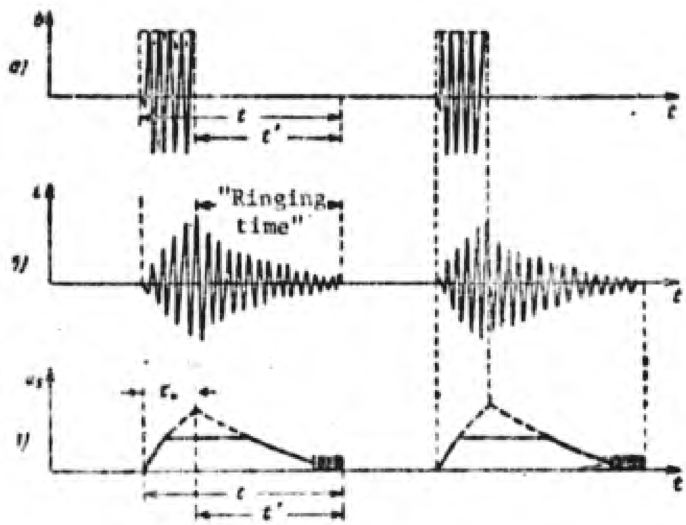


Figure 2.56. A time diagram illustrating the operating principles of a monitoring resonator.

which occurs in a receiver with a relatively high level of the re-emitted pulse, the signal from the resonator to the indicator having linear sweep will have the form shown in Figure 2.56c.

The time from the start to the end of the re-emission of oscillations is called the "ringing time." The moment when the amplitude of the signal from the monitoring resonator becomes equal to the noise amplitude on the radar indicator screen (Figure 2.56c) is taken as the end of re-emission.

The relationship between the ringing time and the station parameters may be determined from the following considerations. The power of the re-emitted pulse equals

$$P_{\text{с}}(t')=P_{\text{п}}Ke^{-\frac{t'}{T_{\text{зв}}}}, \tag{2.40}$$

where P_p is the power of the transmitter;

K is a constant which depends on the coupling coefficient of the high-frequency channel with the resonator, and on the pulse duration and amplitude;

T_{rin} is a time constant which depends on the resonator quality and on the coupling coefficient of the high-frequency channel with the resonator.

Expression (2.40) provides the relation between the resonator ringing time, the transmitter power and the receiver sensitivity.

If one designates by P_{rec}, as before, the minimum power necessary for the observation of a signal on the indicator, which occurs when $t' = t_{rin}$, then from the expression (2.40) we obtain

$$P_{\text{пр}}=P_{\text{п}}Ke^{-\frac{t_{\text{зв}}}{T_{\text{зв}}}}$$

Consequently the ringing time equals

$$t_{\text{зв}}=T_{\text{зв}}\ln\left(K\frac{P_{\text{п}}}{P_{\text{пр}}}\right). \tag{2.41}$$

From equation (2.41) it follows that the ringing time is proportional to the radar power drop.

The monitoring resonator signal which is observed on the radar indicator is longer than the resonator ringing time by the duration of the main pulse, $t = t_{rin} + \tau_p$. However, the quantity t also characterizes the radar power drop. Thus a monitoring resonator enables one to monitor the main quantity characterizing a radar station effective range — the power drop.

It is more convenient to characterize the signal from the monitoring resonator not by the ringing time but by the "persistence" of the ringing. Therefore in the future we will use this term.

The ringing persistence is the main parameter of a monitoring resonator and is determined by it. It has a magnitude of the order of 5-20 dab. In spite of the simplicity of the device, a monitoring resonator, whose scale is calibrated according to the frequency and which has a tuning indicator, enables one to monitor the following in addition to determining the radar power drop:

- the transmitter power;
- the receiver sensitivity and its tuning;
- the frequency of the transmitter and of the receiver heterodyne;
- the frequency energy spectrum of the transmitter pulses;
- the restoration time of the receiver sensitivity;
- the frequency pulling of the magnetron.

<u>Methods of connecting with a radar</u>. Figure 2.57 shows the more frequently used methods for connecting a monitoring resonator with a radar.

The connection of a monitoring resonator in the circuit of an equivalent antenna has the advantage that it enables one to check the radar and its tuning without radiation into space. A defect of this connection arrangement is the impossibility of determining the frequency energy spectrum of the transmitter pulses, since the transmitter spectrum when operating on the dummy antenna is different from the spectrum when operating on the antenna.

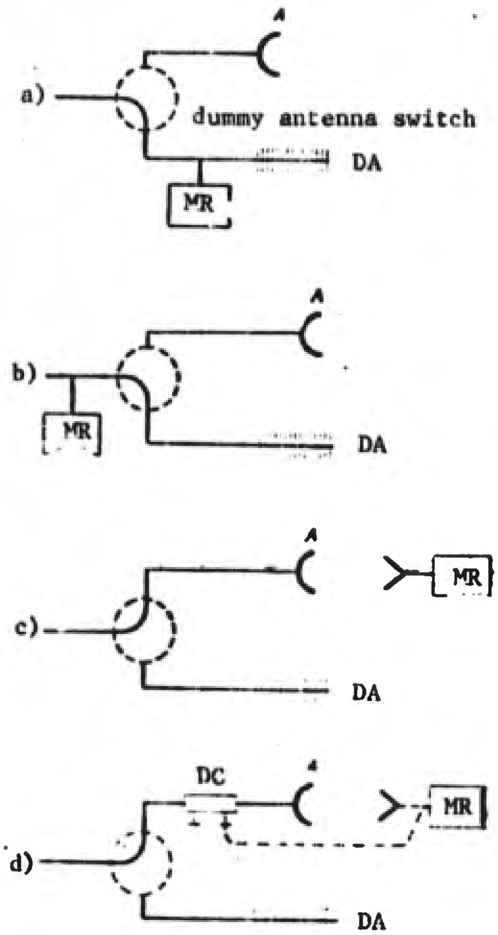


Figure 2.57. Methods of connecting a monitoring resonator (MR) for the purpose of coupling it with a radar: a - in the waveguide channel of a dummy antenna (DA); b - in the wave guide channel before the switch of the dummy antenna; c - through the antenna (A); d - either through a directional coupler (DC) or through the antenna (A).

Connecting the monitoring resonator before the dummy antenna switch enables one to perform the checking and tuning of the station both without radiation and also with radiation into space. Moreover with such a connection method, one is able to check the frequency pulling of the magnetron Among the defects of such a connection method one must list the fact that when the checking of the radar is finished, the resonator must be detuned in order that a signal from it does not interfere with the observation of signals from targets. Moreover, with such a connection, it is impossible to make a judgment as to the change of the attenuation in the antenna-waveguide channel.

The coupling with the monitoring resonator through the antenna eliminates the possibility of checking and tuning the radar without radiation into space, but has the valuable advantage that it enables one to also monitor the antenna-waveguide channel, that is, essentially to monitor the station power potential.

The possibility of connecting a monitoring resonator both to a radar directional coupler and also to the antenna enables one to use the advantages of both methods of connecting a monitoring resonator to the radar.

The advantages of a monitoring resonator are that is is contantly connected to the radar and always ready for use, that it has a small overall size and weight, that it is simple to use and that it monitors a number of the main radar technical parameters. Among its disadvantages, which sharply decrease the effectiveness of its use, one must list the low sensitivity to a change of the radar power drop and the variability of its quality factor during operation.

Therefore our next subject of discussion will not be the absolute measurement of the radar power drop, but the monitoring of the change of the power drop. Consequently a calibration must precede the monitoring, that is, one must determine the radar power drop or effective range by precise methods (by the separate measurement of the parameters which determine the station effective range or by direct measurements of the effective range).

Despite the indicated defects, monitoring resonators are widely used for tuning radar stations, for an approximate estimate of the change of a station power drop, for measurement of the transmitter frequency and sometimes for plotting the frequency energy spectrum of a transmitter pulse.

Monitoring the power drop. In order to use a monitoring resonator, it is necessary to calibrate it. First it is necessary to verify the fact that the station parameters correspond to the certified data. Measurement of the following parameters of the radar must precede the calibration: the receiver sensitivity, the transmitter frequency and power, and the frequency energy spectrum of the transmitter pulses. This enables one to determine whether the radar is optimally tuned, that is, to establish the initial level in respect to which one is able to make a judgment as to a change of the radar tuning during operation.

The calibration process starts with tuning of the monitoring resonator to the transmitter frequency. The optimum tuning value corresponds to the maximum reading on the resonator indicator. Then the radar receiver is tuned to the monitoring resonator signals. With optimum tuning of the receiver, the signal from the monitoring resonator remains the longest on the radar indicator. The noise on the indicators should have the form shown in Figure 2.55. Subsequent measurements are performed with the amplifier control knob in this same position. The resulting value of the ringing persistence corresponds to a tuned radar and is taken as the base when monitoring the radar by means of the monitoring resonator.

It is necessary to remember that a cavity resonator has spurious tunings, that is, tuning to resonance with oscillations of non-working modes. Unually these false tunings correspond to a ringing persistence which is considerably smaller than the normal persistence and a reduced reading of the indicator instrument. This enables one to distinguish false tuning from fundamental tuning.

The ringing persistence is proportional to the relative value of the station power drop. Consequently, a comparison of the ringing persistence obtained during the operation process and the ringing persistence with calibration enables one to make a judgment as to the change in the power drop, and this means a change of the radar station effective range. However, it is necessary to keep in mind that the sensitivity of a monitoring resonator is relatively low.

The ringing time is determined by the empirical equation

$$t_{зв} = \frac{Q \ln \frac{P_{п}}{P_{пр}}}{2.73 f},$$

where Q is the quality factor of the cavity resonator;

f is the frequency.

A monitoring resonator sensitivity is characterized by the value by which the ringing persistence is changed with a change of the station power drop of 1 dB. For good monitoring resonators, this equals 0.6-0.7 cab/dB (110-130 m/dB). Depending on the quality of the cavity resonator and on the possible error in determining the ringing persistence,one is able to observe a change of the station power drop only when it decreases by an amount larger than 2.0-2.5 dB.

Moreover, a change of the resonator quality may lead to the emergence of an additional error in monitoring a station power drop. For the majority of practical cases, a reduction of the station power drop up to 3 dB may be considered acceptable.

A reduction of the power drop attests to a reduction of either the transmitter power or of the receiver sensitivity. The amount of the decrease of the radar power drop which is due to each of these factors may be determined in the following way.

If the transmitter power changes from the value P_p to P'_p, then the decrease of the power drop will equal the following as follows from expression (2.41)

$$\Delta t_{зн} = t_{зн} - t'_{зн} = T_{зн}\left[\ln\left(K\frac{P_и}{P_{пр}}\right) - \ln\left(K\frac{P'_и}{P_{пр}}\right)\right],$$

hence

$$\Delta t_{зн} = T_{зн}\ln\left(\frac{P_и}{P'_и}\right).$$

In a similar way if the receiver sensitivity changes from P_{rec} to P'_{rec}, then

$$\Delta t_{зн} = T_{зн}\ln\left(\frac{P'_{пр}}{P_{пр}}\right).$$

During operation it is important to know which of the factors is the reason for the decrease of the radar power drop. In order to know this, it is necessary to monitor the transmitter power.

Monitoring the transmitter power. The monitoring resonator must be tuned to the transmitter frequency which corresponds to the maximum deflection of the pointer of the indicator instrument. Since the deflection of the indicator instrument pointer is proportional to the transmitter power, a comparison of the indicator instrument readings obtained in the calibration process and during the operation enables one to evaluate the change of the transmitter power.

Naturally a change of the coupling of the cavity resonator with the indicator device and also a replacement of the detector require that the calibration of the monitoring resonator be repeated. Moreover, it is necessary to take into account the fact that changes of the detector parameters in time will lead to significant errors in the evaluation of the transmitter power. However, even an approximate evaluation of the change of the transmitter power during the operation process in the period between routine checks of the radar technical parameters is of great importance.

Measuring the transmitter frequency. The monitoring resonator is tuned to the transmitter frequency on the basis of the maximum reflection of the indicator instrument pointer. A value corresponding to the tuning to resonance is read from the scale and the transmitter frequency is determined by means of a calibration curve. The error of the relative measurements of the frequency does not exceed $\pm$ 2 Mc.

Monitoring the frequency energy spectrum of the magnetron oscillator pulses. In order to monitor the energy spectrum of the frequencies of the magnetron oscillator, a monitoring resonator is included in the antenna channel and is tuned to the oscillator frequency on the basis of the maximum reflection of the indicator pointer. The reading of the indicator and the value of the tuning from the scale are noted. By changing the tuning of the monitoring resonator first to one side of the carrier frequency and then to the other side, readings of the indicator and values of the tuning based on the scale which correspond to them are recorded. The readings of the indicator instrument are plotted as a function of the frequency of the monitoring resonator tuning.

It is necessary to keep in mind that, when using a monitoring resonator, a spectrogram of low quality is obtained. It is also necessary to consider that a cavity resonator has non-working ranges, that is frequencies at which (with a given position of the piston) the simultaneous resonance of two oscillation modes is observed. When this happens, the resonator quality falls off sharply. The distribution of the points of the quality drop along the frequency range usually is listed in the logbook of the monitoring resonator. If there are non-working ranges in the limits of the spectrum, then it naturally is impossible to check the spectrum with a monitoring resonator.

Checking the magnetron frequency pull. In radar stations the operating conditions of the magnetron oscillator do not remain constant because of the presence of rotary joints, scanning heads, etc. This leads to a change of the frequency being generated by the magnetron, that is, to a pulling of the frequency. A monitoring resonator enables one to detect a frequency pulling of the magnetron and to approximately determine the amount of the pulling.

In order to do this one should tune the monitoring resonator to the transmitter frequency on the basis of a maximum reading of the indicator instrument pointer and determine the ringing persistence. Then one should set the antenna into circular rotation and observe the signal from the resonator on the radar indicator. If the ringing persistence does not change or changes insignificantly, then there is practically no pulling of the frequency.

In radars with scanning of the antenna beam, a change of the ringing persistence while the scanning head is switched on and off is also of interest. This enables one to determine the source of the frequency pulling. The amount of the frequency pulling may be determined very approximately if one stops the antenna in the position in which the ringing persistence is reduced, and then retunes the monitoring resonator so that it again has the maximum ringing persistence. The difference of the tunings in units of the frequency will correspond to the amount of the magnetron frequency pulling. The amount of the pulling must not exceed the limits of the AFC retuning, since if it does, this will lead to a decrease of the station effective range. An increase of the frequency pulling above the norm attests to the malfunctioning of rotary joints or of the scanning head of the antenna.

Careful utilization of a monitoring resonator enables one to obtain additional information characterizing radar operation. Thus, for example, the appearance of noise on the sweep line under the signal of the monitoring resonator (Figure 2.58) attests to lapses of the pulses or to operation of the magnetron in two oscillation modes.

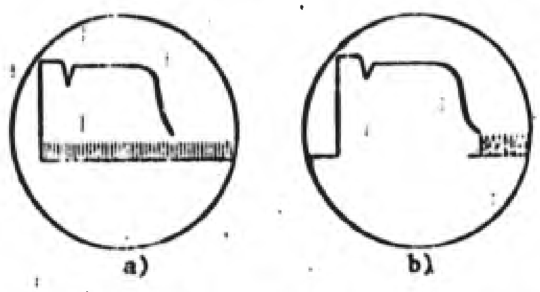


Figure 2.58. The form of the monitoring signal on an indicator with linear sweep: a - with operation at two oscillation modes; b - with normal operation.

As is seen from Figure 2.55, the monitoring resonator signal on an indicator with a linear sweep has a characteristic trough at the beginning. The presence of the trough is a consequence of the finite restoration time of the receiver sensitivity after the radiation of the transmitter pulse.

<u>The sensitivity restoration time</u> of the receiver designates the time during which deionization of the receiver discharger takes place after passage of the transmitter pulse. During this time the receiver input is shunted by the small resistance of the discharger, as a consequence of which the receiver sensitivity is gradually restored to the maximum value. This time determines the dead zone of the radar. The restoration time of a receiver sensitivity usually equals 4-8 microsec and depends on the quality of the discharger.

An increase and broadening of the trough at the beginning of the signal from the echo chamber attests to slow restoration of the receiver sensitivity. When this happens it is necessary to replace the discharger or to fine tune the cavity resonator of the receiver antenna switch.

Overall working order of a radar when monitoring its operation with a monitoring resonator is characterized by the following data:

- the deviation of the ringing persistence to the lower side of the value obtained when calibrating the monitoring resonator is no more than 2-2.5 cab;
- the value of the magnetron frequency corresponds to the certification;
- the signal shape satisfies the established requirements;
- on the indicator screen there is no noise under the image of the signal from the monitoring resonator.
- the trough at the beginning of the signal from the monitoring resonator ends at a distance which does not exceed a specified value.

Thus a monitoring resonator enables one to check a number of technical parameters which are responsible for the radar station effective range. The effectiveness of using a monitoring resonator is higher, the greater its sensitivity. The simplicity of use, the possibility of monitoring a large number of radar main parameters, and the small overall size and weight make a monitoring resonator an irreplaceable instrument for the daily operation of a radar.

Monitoring Radar Technical Performance by Means of a Short-Circuit Delay Line (The Method of Self-Calibration)

The self-calibration method for a radar was developed for measuring the effective reflecting surface of targets [51]. In order to solve this problem, it is necessary to precisely calibrate the receiver output in units of P_{rec}/P_p. This method may also be used for monitoring the power drop of a radar, that is, $H = P_p/P_{rec}$.

The essence of the method consists of the fact that the transmitter high-frequency pulse is delayed and attenuated and then it is introduced into the radar receiver channel. The ratio of the signal voltage to the noise at the receiver output when its amplitide characteristic is known characterizes the power drop or the power potential of the radar.

Figure 2.59 shows a block diagram which illustrates the method for monitoring radar operation by means of a short-circuit delay line. Part of the main pulse energy is sent through a directional coupler to a long delay line which is short circuited on both ends. A signal must be shunted to the delay line with such a level that it does not affect the radar effective range. The energy entering the line is propagated along it and reflected from the short-circuited end, reaching the line input. When this happens, part of the pulse energy enters the receiver and part is reflected in the opposite direction, and then it again is reflected, etc.

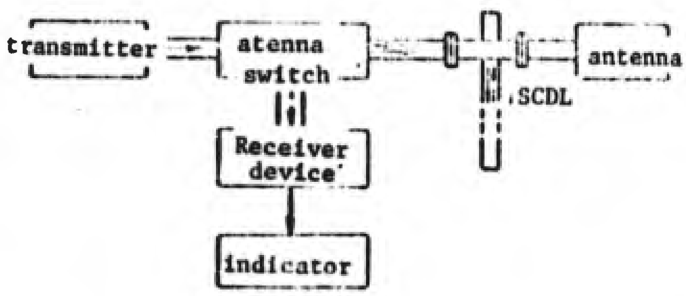


Figure 2.59. The block diagram for turning on a short-circuit delay line (SCDL) when monitoring a radar power drop.

Thus a sequence of pulses, delayed in time and decreased in amplitude, is sent to the receiver. On the linear sweep indicator connected to the receiver output, one is able to observe the transmitter forward pulse and a series of calibration pulses which are located on the decreasing part of the oscillogram and which melt into the receiver noise spikes (Figure 2.60). The larger the number of calibration pulses fitted onto the falling part of the oscillogram, the larger the radar power drop. In order to obtain an acceptable precision for the measurement of the power drop, the number of pulses should be no less than eight.

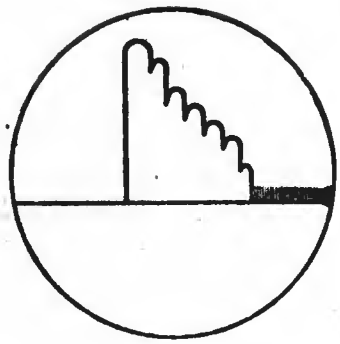


Figure 2.60. The form of the signal of a short-circuit delay line on an indicator with linear sweep.

The time displacement between adjacent pulses is determined by the delay time of the pulse in the line. The minimum length of the line l_{min} for the resolution of two adjacent pulses must equal

$$l_{\text{мин}} = \frac{c\tau_{\text{и}}}{2}.$$

When this condition is not fulfilled, the pulses following each other along are superimposed on one another, and the resulting signal depends on the magnitude and relative position of the pulses. Moreover, the time displacement between adjacent pulses must be larger than the restoration time of the receiver's sensitivity in order that the character of the change of the receiver sensitivity during its restoration time need not be taken into account.

If one assumes that for the clear fixation of pulses the attenuation in one cycle of delay is α_0, then the attenuation per unit length of the line must equal

$$\alpha = \frac{\alpha_0}{c\tau_{\text{и}}}$$

Since, when making the measurement, the oscillogram on the radar indicator screen appears at the start of the sweep, this naturally disrupts the radar normal operation. In this case the measurements must be performed randomly. For the continuous monitoring of the power drop one may use the non-working part of the sweep. In order to do this it is required to achieve a delay of the high-frequency signal by an amount almost equal to the transmitter pulse repetition period. This is connected with great technical difficulties.

Initially the oscillogram (Figure 2.60) is plotted with an optimally tuned radar, that is, the meter is calibrated. Having determined the number of calibration pulses and knowing the coupling attenuation of the directional coupler and the attenuation of the delay line, the nominal value of the radar power drop is determined. Using the oscillogram one is able to plot the dependence of the beam deflection as a function of the ratio P_p/P_{rec} (Figure 2.61). The nominal value of the radar power drop, H_n, will correspond to the value of P_p/P_{rec} associated with a beam deflection equal to the noise amplitude.

The total error in measuring the radar power drop by means of a short-circuit delay line is composed of:

- the calibration errors of the directional couplers and errors connected with the instability of the directional coupler coupling attenuation during operation;
- errors because of the presence of mismatch in the antenna-waveguide channel;
- errors in reading the measurement result.

The self-calibration method enables one to perform a measurement of the radar power drop with a high degree of precision. The measurement error of the relative value of the power drop does not exceed 0.5-0.9 dB.

A short-circuit delay line may be used as a built-in instrument. Its small measurement error favorably distinguishes it from a monitoring resonator. A short-circuit delay line is inferior to a monitoring resonator in respect to its overall size and weight.

Application of the self-calibration method when using a delay line in the high-frequency channel of the radar still encounters a number of difficulties, the main one of which is the fact that the line must have a very long length and it must delay a superhigh frequency pulse up to several microseconds.

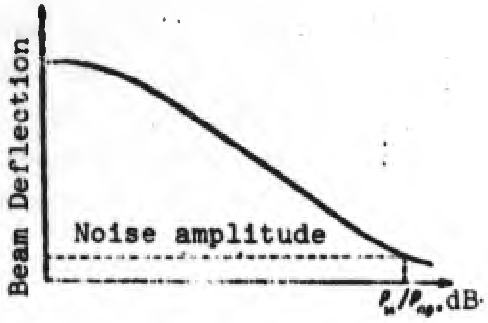


Figure 2.61. The amount of the beam deflection in an indicator with linear sweep as a function of the radar power drop.

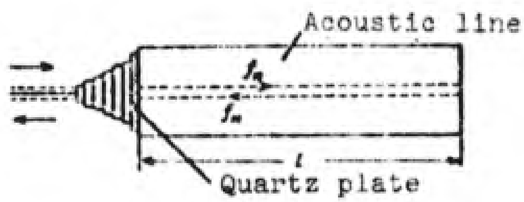


Figure 2.62. The schematic diagram of an ultrasonic delay line.

Monitoring Station Operation by Means of an Ultrasonic Delay Line

This method is based on the direct attenuation and delay of the transmitter pulse. It is used as the monitoring signal which is introduced into the intermediate frequency channel. An ultrasonic line is used and is the most suitable delay line from the point of view of the frequency and amplitude characteristics. The schematic digram of an ultrasonic delay line with an integrated input and output is shown in Figure 2.62. The following main elements go into the makeup of the ultrasonic delay line:

- an electromechanical converter which converts the electromagnetic oscillations to ultrasonic oscillations and vice versa;
- an acoustic line in which propagation of the ultrasonic oscillations take place;
- integrated input and output devices.

The electromechanical converter is a plate made from crystalline quartz with a cut perpendicular to the x axis (a x-cut). Metal films which are the electrodes of the converter are deposited on the lateral surfaces of the plate. The quartz possesses a forward and a reverse piezoelectric effect. The forward piezoelectric effect consists of the generation of charges on the surfaces of the quartz plate with mechanical excitation of it. The reverse piezoelectric effect consists of a change of the dimensions of the quartz plate under the action of a potential difference applied to the lateral surfaces of the quartz plate. The forward and reverse piezoelectric effects of the quartz enable one to use it as an electromechanical converter. The acoustic line is a metal or quartz rod. The integrated input and output device is used for feeding and removing the high-frequency signal. Moreover it is the acoustic load for the acoustic line.

Several methods for connecting an ultrasonic delay line with a radar for the purpose of checking its power potential are possible. The ultrasonic delay line may be connected in the following ways:

- in the circuit of the receiver intermediate frequency amplifier;
- in the automatic tuning control circuit;
- in the high-frequency circuit based on the active responder principle.

When the delay line is connected in the intermediate-frequency amplifier channel of the receiver, it may be connected either to the mixer output or to one of the stages of the i-f amplifier (Figure 2.63).

In the first case, the delay and attenuation of the monitoring signal is carried out after the mixer, which enables one to conduct the measurement with a high level of the input signal. Naturally when this is done the parasitic signal effect is lessened, and the measurement error is reduced.

The coupling is achieved in the following way. During the period of measuring the radar power potential, the transmitter is connected by means of the equivalent antenna switch to the equivalent antenna. The signal from the transmitter is sent to the receiver input along two channels: along the main channel through the directional coupler

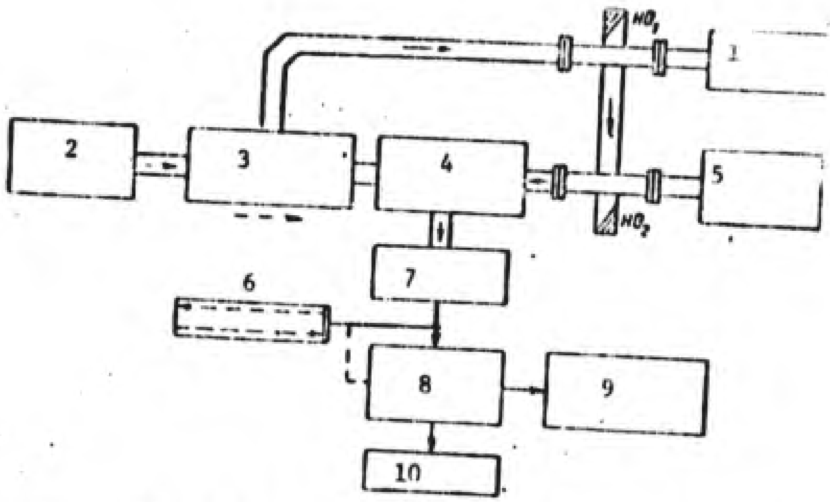


Figure 2.63. The block diagram for connecting a short-circuit delay line after the receiver mixer.

1 - Dummy antenna; 2 - transmitter; 3 - "Dummy antenna" switch; 4 - Antenna switch 5 - Antenna; 6 - Delay line; 7 - Mixer; 8 - i-f amplifier, detector; 9 - Indicator part of meter; 10 - Indicators.

and along the parasitic channel through the antenna switch.

For normal functioning of the device, sufficient decoupling must be provided for the parasitic channel, that is, the total attenuation along this channel must be significantly larger than the attenuation along the main channel. This condition imposes specific requirements on the attenuation which the switch must have when it is in the "dummy antenna" position, and it determines the upper limits of the radar power for which such a method may be used.

The monitoring signal is sent through the directional coupler to the receiver mixer. The coupling attenuation of the directional coupler must be selected in such a way that the magnitude of the intermediate-frequency signal would be sufficient for the purpose of exciting the ultrasonic delay line. The monitoring signal converted to intermediate-frequency oscillations acts on the exciter giving rise to mechanical oscillations of the quartz.

The pulse of mechanical oscillations of the quartz exciter is propagated with specific velocity along the delay line; it is reflected from the end of the line and is returned to its beginning. Having lost part of the energy in exciting the quartz, the pulse again is reflected, etc. Thus the process of exciting the quartz is periodically repeated until the mechanical oscillations in the delay line die down to such a level that the quartz ceases to be excited. When pulses of mechanical vibrations act on the quartz exciter, electric oscillations with the frequency f_{rec} arise on the coatings of the quartz. These electric oscillations with the frequency f_{rec} are sent to the receiver i-f amplifier.

The signals from the ultrasonic delay line form a sequence of pulses with a decreaseing amplitude on the indicator with a linear sweep. The same signals form a sequence of pulses with decreasing brightness on an indicator with a circular sweep. The number of pulses observed on the indicator and their brightness depend on the station power drop. An ultrasonic delay line enables one, consequently, to monitor the working order of the transmitter-receiver channel and to evaluate the radar power drop.

The measurement result may be indicated in several ways:

1. The determination of the value of the station power drop may be made by measuring the ratio of the monitoring signal to the noise at the receiver output. Such an indication method is used if the amplitude characteristic of the receiver is linear.

2. A reading of the deviation of the radar power drop from the threshold value is obtained by means of a special circuit. Such a circuit should automatically signal a reduction of the

radar power drop below the acceptable value.

3. In a radar with automatic range tracking of a target, a tolerable monitoring of the power drop may be achieved on the basis of the threshold value of the monitoring signal.

In the latter case, in the circuit for exciting the ultrasonic delay line it is advisable to include an attenuator which enables one to set the threshold value of the signal entering the receiver and then sent to the circuit for range tracking the target. One may judge as to the correspondence of the station power drop to the rated value on the basis of the holding of the signal from the ultrasonic delay line by the automatic tracking circuit.

The rated value of the threshold signal is set with the factory adjustment of the radar or after conducting preventive maintenance tasks. This setting should take into account a certain permissible decrease of the value of the threshold signal which may be observed during the operation process for a given station. The absence of signal holding attests to a reduction of the circuit sensitivity. In connection with this the attenuator, having been calibrated, enables one to make a quantitative evaluation of the change of the station power drop during the operation process.

The total measurement error with this device is made up of errors which are due to the following factors:

- the instability of the attenuation introduced by the high-frequency channel;
- the error introduced by the indicator part of the meter;
- the difference of the power drop values when operating on the antenna and on the dummy antenna;
- the instability of the ultrasonic delay line parameters.

A determination of the absolute value of the station power drop leads to unacceptably large measurement errors. Therefore such a device is suitable for measuring the relative value of a station power drop. The maximum measurement error when this is done does not exceed 2.2-2.5 dB. As in the case of a monitoring resonator, a calibration must precede the measurements, that is, one must determine the radar power drop of effective range by precise methods.

Thus we have investigated the operation principles and we have evaluated the technical possibilities of monitoring and measuring instruments which are installed in a station.

It is necessary to point out that the method of individual measurement of the technical parameters which determine station effective range enables one to most objectively and fully determine the station technical performance, as a consequence of which it is used when conducting preventive measures. When making measurements of the technical parameters, besides portable instruments, built-in measuring instruments have begun to find ever-increasing use.

With the daily operation of a radar, the method of individual measurement of the technical parameters by means of built-in monitoring and measuring instruments and the procedure of evaluating a station power potential which determines the radar effective range in a complicated way is the most effective approach. Radar power potential may be monitored by means of a monitoring resonator or a short-circuit delay line.

FOOTNOTES

Footnote (1) on page 74. One cable (1 cab) equals 185.2 m.

3. MAINTAINING CONSTANT MEASUREMENT PRECISION OF TARGET COORDINATES

3.1. General Statements About the Theory of Errors

A radar station, as any other measuring instrument, makes measurements with a specific degree of precision — that is, errors accompany the measurements.

By the error and correction of a measurement we mean the following. If a is the precise value of a certain quantity and x is the value of the same quantity containing a certain error, then $x - a = \Delta x$ is the measurement error and $a - x = \nabla x$ is the correction to the quantity x.

Consequently

$$a = x - \Delta x,$$

$$a = x + \nabla x$$

that is, in order to correct an erroneous result it is necessary to algebraically add the correction to it or to algebraically subtract the error from it. The causes, character and nature of errors are very diverse. Measurement errors may be classified, on the basis of the sources giving rise to them, as equipment errors, personnel errors and external errors.

Errors which arise as a consequence of the presence of noise accompanying the measurement and the limited precision of the instruments making the measurements are called equipment errors. In the case being investigated by us, these errors are due to the limited precision in determining the target coordinates by the radar station.

Personnel errors arise because of the limited resolution of man's sense organs. As applied to a radar station this means the limited resolution of man's eyes.

External errors arise because of the influence of the surroundings. These include, for example, angle of elevation measurement errors which are connected with refraction of radio waves (errors of approach) and the measurement errors of small angles of elevation of a target.

Measurement errors may also be classified on the basis of the law governing their generation — that is, random errors and systematic errors.

Random errors — these are the unavoidable small errors which arise when making any measurement. When measuring target coordinates, the random errors arise as a consequence of the fact that during the measurement process — because of a change of the factors characterizing the station, the radio wave propagation conditions and the object — the results of measurements of the same quantity unavoidably differ from one another by small amounts. As a consequence, compensation of random errors is impossible.

Random errors as a rule are subject to the law of normal statistical distribution (Gauss law) which is described by the function

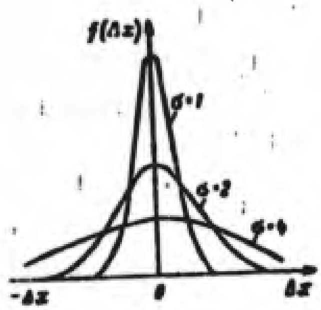


Figure 3.1. Gauss curves.

$$f(\Delta x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\Delta x)^2}{2\sigma^2}}. \quad (3.1)$$

where σ^2 is the standard deviation;
Δx is the measurement error.

The equation establishes the connection between the standard deviation, the amount of the error, and the probability of its appearance. The curve of this relationship for three values of σ ($\sigma = 1$, $\sigma = 2$ and $\sigma = 4$) is shown in Figure 3.1.

Expression (3.1) may be written in another form:

$$f(\Delta x)=\frac{h}{\sqrt{\pi}}e^{-h^2(\Delta x)^2},$$

where h is called the modulus of accuracy; $h=\pm\frac{1}{\sqrt{2}\sigma}$.

Systematic errors are caused by factors which act in the same way when the same measurements are repeated over and over again. They enter into the measurement results, Depending on the source which gives rise to them, their change may be different: they are either constant, or reoccur regularly, or vary according to a specific law. The effect of these errors on the measurement results may be taken into account by introducing a correction, or it may be eliminated by adjusting the equipment.

For the purpose of characterizing measuring instruments, among them radars, a number of measurement precision criteria are used:

- the mean square error;
- the arithmetical mean error;
- the mean or probable error;
- the maximum error.

The quantity

$$s_n = \pm\sqrt{\frac{\Delta x_1^2 + \Delta x_2^2 + \ldots + \Delta x_n^2}{n}} = \pm\sqrt{\frac{\sum_1^n \Delta x_i^2}{n}},$$

where Δx_1, Δx_2, ..., Δx_n are the measurement errors, is called the mean square error.

If the number of measurements is large, then s_n tends toward its statistical limit.

$$s = \lim_{n\to\infty} s_n.$$

This limit also is essentially the mean square error. However, since n is finite, in the measurement process the quantity s_n is determined which will be closer to σ, the larger the number of measurements. The quantity σ^2 enters into the Gauss equation and is called, as was already indicated, the standard deviation.

The mean square error is connected with the accuracy modulus by the relationship

$$s = \pm\frac{1}{h\sqrt{2}}.$$

The arithmetical mean error is calculated from the equation

$$r_n = \pm\frac{|\Delta x_1| + |\Delta x_2| + \ldots |\Delta x_n|}{n} = \pm\frac{\sum_1^n |\Delta x_i|}{n}.$$

As can be seen from this expression for r_n, the absolute values of the errors are summed when determining the arithmetical mean error.

As in the previous case the true value of the arithmetical mean error equals

$$\rho = \lim_{n\to\infty} r_n.$$

The arithmetical mean error is connected with the accuracy modulus by the relationship

$$\rho = \pm \frac{1}{h\sqrt{\pi}}.$$

The value of the error only gives a one-sided characterization of the measurement result. In order to complete the picture it is also necessary to know the degree of reliability of the results obtained.

If the arithmetical mean value obtained as a result of the measurement is $\bar{x}$ and the measurement error is Δx, then the probability that the measurement result differs from the true value by an amount larger than Δx equals

$$P[(\bar{x} - \Delta x) < a < (\bar{x} + \Delta x)] = P_{\alpha}.$$

In this expression, the probability P_{α} is called the confidence coefficient and the interval between $\bar{x} - \Delta x$ and $\bar{x} + \Delta x$ is called the confidence interval.

From the expression it follows that the larger the confidence interval, the higher the probability with which one may state that the measurement result does not go beyond its limits.

Thus a random error is characterized by two parameters: the magnitude of the error and the value of the confidence coefficient.

Depending on the required degree of reliability of the results, the measurements are given with one or another value of the confidence coefficient; usually it is taken as equal to 0.9 or 0.95.

The mean square error σ corresponds to a confidence coefficient of 0.68; twice the mean square error, 2σ, corresponds to a confidence coefficient of 0.95, and three times the mean square error, 3σ, corresponds to 0.997. For other error values the confidence coefficient may be determined from special tables.

That error which divides all the random errors of the series of measurements into two parts — each of which has an equal number of errors such that one part contains 50% of the errors which are larger than this error and the other part contains 50% of the errors which are smaller than this error — is called the mean error

The mean error is calculated from the equation

$$r = \pm 0.675 \sqrt{\frac{\Delta x_1^2 + \Delta x_2^2 + \ldots + \Delta x_n^2}{n}} = \pm 0.675 \sqrt{\frac{\sum_1^n \Delta x_i^2}{n}}.$$

The connection between r and h may be found from the relationship

$$r = \pm \frac{0.477}{h}.$$

In that case when the errors Δx_1, Δx_2, ..., Δx_n remain unknown as a consequence of the fact that the precise value of the quantity is unknown, by using a theorem from the theory of probability, the arithmetical mean value calculated from the series of measured values is taken as the most probable value of the measured quantity. In this case, when determining the mean square error and the mean error the deviation of the results of the individual measurements from their arithmetic mean $\bar{x} - x_i$ will be used in the numerator of the expressions, and in the denominator we will use the number of measurements n reduced by one.

The error that the probability of exceeding its absolute value practically approaches zero is called the maximum error. Usually four times the value of the mean error is taken as the maximum error p.

This error corresponds to a confidence coefficient of 0.99, that is, the probability of exceeding such an error equals 1.0%.

The connection between the various error criteria is expressed by the relationship

$$r = 0{,}25p = 0{,}477\sqrt{2}\,\sigma = 0{,}477\sqrt{\pi}\,\rho,$$

or

$$r = 0{,}25p = 0{,}68\sigma = 0{,}83\rho.$$

With a sufficient number of measurements one may use the arithmetical mean error or the mean square error for evaluating the precision with the same success. But with a small number of measurements the mean square error reflects a great deal better the effect of large errors which are most strongly expressed in the measurement result.

When checking a radar station it is necessary to determine both the random errors and also the systematic errors. Random errors which characterize the precision in measuring the coordinates of objects are recorded in the station logbook. The maximum error is taken as the accuracy criterion for stations and systems used for the precise determination of target coordinates.

Among all the errors of a radar station, the systematic errors are of special interest since they change while the station is functioning. The systematic errors in measuring the range and angular coordinates of radar objects must be comprehensively studied and eliminated with the maximum possible degree of accuracy from the measurement results.

Let us investigate the question as to the sources which give rise to systematic errors in the measurement of the range and angular coordinates of objects, and the methods of determining and eliminating them.

3.2. The Systematic Error in Measuring the Range

As is known, the determination of the range by a radar station is based on measuring the time required for the passage of electromagnetic energy from the radar to the reflecting object and back again. Two properties of radio waves form the basis of this principle of measurement: the constancy of the wave speed and rectilinear propagation of the wave. The errors when making the range measurement are divided into external errors, equipment errors and personnel errors.

As is indicated above, errors arising as a consequence of the instability of the radio wave propagation conditions and errors introduced by the object are referred to as external errors. Since these errors, which have both a random component and also a systematic component, do not depend on the circuit and structural features of a radar, we will not discuss them.

Equipment errors are caused by noises accompanying the measuring process and by instrument errors due to the imperfect nature of the station as a measuring instrument.

The nature of systematic instrument errors is as follows. In order to measure the time for the passage of electromagnetic energy from the station to the reflecting object and back again, it is necessary to fix two moments of time: the moment of the start of the radiation of the main pulse into space and the moment at which the reflected pulse reaches the antenna. Actually the determination of the range is made by measuring the time delay of the video pulse from the target with reference to the forward pulse of the transmitter on the radar's indicator. The difference between these time intervals also leads to the generation of a systematic error in the range. The delay of the signal connected with this is caused by the spreading of the fronts, by the delay of the pulses in the stages which contain reactive elements, and by the delay associated with the transit of the antenna-waveguide channel. The spreading of the pulse fronts leads to a delay in triggering the submodulator trigger circuits. The reactive elements cause significant delays in the receiver circuits.

This error, consequently, is caused by a delay of the pulse signals with their passage through the transmitter-receiver and waveguide channels of the radar.

The second component of the systematic error in determining the range is connected with the fact that both systematic errors and random errors also accompany the measurement of the range to the object.

First let us discuss the nature of the origination of the systematic error because of time delays in the radar station channels. Let us trace the path for the passage of pulses along the transmitting channel and the receiving channel of the radar.

Depending on its purpose and circuit, a radar may be synchronized in various ways. The type of circuit and the location of the synchronizing voltage source are determined by the requirements imposed on the radar in respect to the precision of the range measurement. From this point of view, the synchronizing sources for the range may be divided into two groups.

In the first group are synchronizing sources for radars on which rigorous requirements in respect to the precision of the range measurement are not imposed. The station power supply generator, the transmitter or a special synchronizer is used as the synchronizing voltage source in these radars.

In the second group there are synchronizing sources on which rigorous requirements are imposed for the range measurement accuracy. Such radars have a special synchronizer which ensures the strict time synchronization of the various radar devices.

A sinusoidal signal oscillator with quartz stabilization of the frequency and a frequency divider are the component units of the synchronizer in this case.

An analysis of the systematic error of a radar station when measuring the range naturally is of interest for radars on which rigorous requirements for the accuracy of the range measurement are imposed. Therefore, let us assume that the radar is synchronized by a sinusoidal oscillator with quartz stabilization (a quartz-crystal oscillator). In this case, the frequency of the main pulse is determined by the frequency of the quartz-crystal oscillator and by the division factor of the frequency divider, and the moment of triggering the transmitter is rigidly tied to a specific phase of the voltage of the quartz-crystal oscillator.

Figure 3.2 shows the main channels for the passage of the main and reflected pulses in a radar having two indicators (a coarse range indicator [CRI] and a fine range indicator [FRI]). In this circuit, a sinusoidal signal oscillator with quartz stabilization of the frequency serves as the synchronizer.

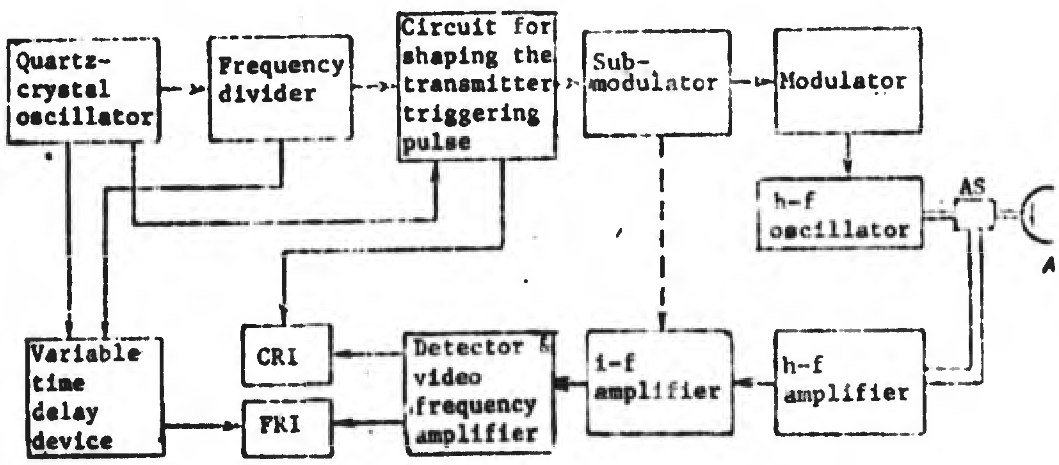


Figure 3.2. The main channels for the passage of the main pulses and reflected pulses for synchronization of a radar by a quartz-crystal oscillator.

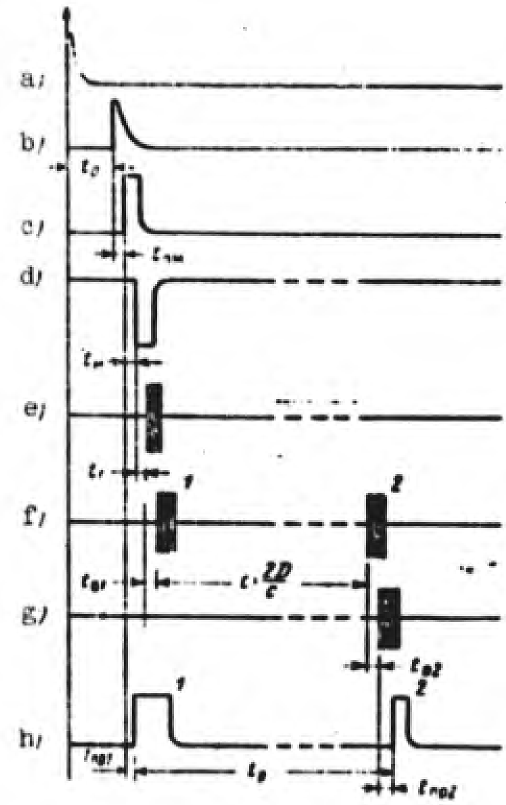


Figure 3.3. A diagram of the time delays in the circuits of a radar: a - the sweep triggering pulse; b - the transmitter triggering pulse; c - the voltage pulse at the submodulator output; d - the voltage pulse at the modulator output; e - h-f oscillator pulse; f - the radiated h-f pulse (1) and reflected pulse from the target (2) (at the antenna); g - the pulse at the receiver input; h - the transmitter forward pulse (1) and the video pulse of the target (2).

Figure 3.3 presents the diagram of the time delays in a radar station circuit for the case when the triggering pulse for the sweep precedes in time the transmitter triggering pulse by the time t_0. In this case the front of the transmitter forward pulse, which also is the zero range reading, will be observed on the linear sweep indicator. The nature of the formation of this pulse is as follows. The transmitter triggering pulses which are synchronized by the quartz-crystal oscillator are sent to the submodulator input. A pulse from one of the stages of the submodulator creates an impact excitation in one of the circuits of the intermediate-frequency amplifier. Then this pulse of intermediate frequency is amplified. It also is the source for the formation of the transmitter forward pulse on the radar indicator.

Thus the transmitter forward pulse appears on the indicator with a time delay t_{f1} with respect to the leakage pulse from the submodulator.

Figure 3.3, b shows the case when the impact excitation in the i-f amplifier circuit is created by the output pulse of the submodulator.

Besides this the transmitter triggering pulse, having been delayed in the submodulator circuits by a time t_{sm}, gives rise to a voltage pulse at the submodulator output. In turn the submodulator pulse gives rise to a voltage pulse in the modulator output with a time delay t_m.

The voltage pulse of the modulator, acting on the oscillator, initiates after an interval of time t_g a high-frequency pulse at the oscillator output (Figure 3.3, b). The radiated high-frequency pulse will have an additional time delay t_{w1} equal to the delay time of the pulse in the station waveguide. The radiated high-frequency pulse and the high-frequency pulse reflected from the object at the reception point are shown in Figure 3.3, f. The video pulse from the target which is observed on the radar indicator will have a time delay with respect to the high-frequency pulse equal to the sum of the transit time of the pulse in the waveguide, receiver and indicator of the station ($t_{w2} + t_{rec2}$).

As follows from Figure 3.2 the time interval between the transmitter forward pulse and the video pulse from the target, which determines the measured distance, will equal

$$t_{\text{р}} = t_{\text{м}} + t_{\text{г}} + t_{\text{в1}} + t + t_{\text{в2}} + t_{\text{пр2}} - t_{\text{пр1}}.$$

where t_m is the time of the pulse delay in the circuits of the modulator;

t_g is the time interval between the modulator pulse and the oscillator pulse;

t_{w1}, t_{w2} are the delay times of the forward and reflected pulses in the waveguide channel;

t is the transit time of the high-frequency pulse from the radar antenna to the target and back again;

$t_{rec\ 2}$ is the delay time of the reflected pulse in the receiver and indicator circuits;

t_{f1} is the delay time of the submodulator leakage pulse (forward pulse) in the receiver and indicator circuits.

The range corresponding to this time interval will equal

$$D_{\text{р}} = D_0 + \frac{c}{2}[(t_{\text{м}} + t_{\text{г}}) + (t_{\text{в1}} + t_{\text{в2}}) + (t_{\text{пр2}} - t_{\text{пр1}})].$$

The systematic error in determining the range which arises when making the measurement will correspondingly equal

$$\Delta D = \frac{c}{2}(t_{\text{мг}} + 2t_{\text{в}} + t_{\text{пр2}} - t_{\text{пр1}}) = \frac{c}{2} t_{\text{c}}, \tag{3.2}$$

where

$$t_{\text{мг}} = t_{\text{м}} + t_{\text{г}};$$
$$2t_{\text{в}} = t_{\text{в1}} + t_{\text{в2}}.$$

From expression (3.2) and the diagrams of Figure 3.3 it is seen that the systematic error in the measurement of the range is determined by the time delays of the signal in the modulator stage, the oscillator stage, the receiver device and the antenna-waveguide channel. From an investigation of the diagram it also follows that with a change of the time delays t_0, t_{sm} and t_{f1}, the systematic error is not changed because the time t_p remains constant. The delays of the submodulator leakage pulse and of the reflected pulse in the video-frequency amplifier stages are equal, as a consequence of which, as follows from expression (3.2), they compensate one another. Consequently a change of the time delays in the video signal channel also does not show up in the size of the radar systematic error. In both cases, a displacement of the range marker line in respect to the front of the transmitter forward pulse takes place. This is illustrated by Figure 3.4 from which it follows that a change of the time delays t_0, t_{sm} and t_{f1} leads (with a fixed range to the object) to a displacement of the video pulse from the target in respect to the marker line, and consequently to an error in the range measurement.

The error ΔD_0, as we will show below, may be compensated. It is not necessary to determine the systematic error of the range measurement in order to do this.

Let us discuss in somewhat more detail the components which determine the systematic error in a range measurement.

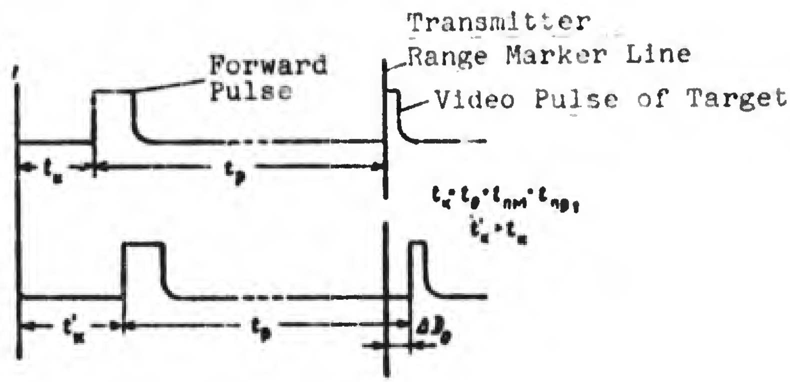


Figure 3.4. The effect of the time delays t_0, t_{sm} and t_{fl} on the error in the range measurement.

3.3. The Components of the Systematic Error in the Range Measurement

The systematic error in the range measurement of a radar results from two main causes:

- the presence of time delays in the transmitter-receiver channel and the waveguide channel of the station;
- the emergence of an error when measuring the time delay of the pulse from the target with reference to the transmitter forward pulse. In connection with this the time delays in the circuits for the shaping and reception of the pulse signals are the main source giving rise to the systematic error.

Let us investigate the components of the systematic error in the range measurement.

The Pulse Delay in the Modulator and Oscillator

The passage of the signal in the transmitter is connected with the greatest transformation of its shape (amplitude and duration). The total delay is made up of delays in the submodulator, modulator and oscillator. It is legitimate to sum the delays because the pulse moves along the channel sequentially from stage to stage. Obtaining an analytical relation for the delay of the pulse in the modulator and oscillator is a complicated problem. The time delays basically are determined by the conditions of the tubes in the trigger circuits and by the delay of the pulse in the artificial delay lines, in the pulse transformer and in the magnetron oscillator.

The finite steepness of the pulse fronts leads to a delay in the triggering of the submodulator blocking oscillators.

Thyratrons, performing the role of commutator, create a time delay resulting from the ionization time. Thus, for example, the time delay of a pulse in a hydrogen thyratron does not exceed 0.1 microsec, and the fluctuations of the delay time which are caused by a change of the anode voltage are no larger than 0.07 microsec.

The delay in a long artificial line is commensurate with the duration of the pulse being shaped. A considerable delay is caused by the rise time of the modulating voltage pulse to the critical value at which generation of the high-frequency pulse by the magnetron beings.

The total time delay in the modulator and oscillator usually does not exceed 0.3-0.6 microsec.

The Delay of Radio Pulses in a Radar Waveguide

The group velocity of a radio pulse with its propagation in a waveguide equals

$$v = \frac{\lambda}{\lambda_w} c,$$

where λ_W is the wavelength in the waveguide;
c is the speed of light.

In turn the wavelength in the waveguide is

$$\lambda_s = \frac{\lambda}{\sqrt{1-\left(\frac{\lambda}{2a}\right)^2}},$$

where a is the dimension of the long side of the waveguide cross section.

Consequently, the delay time of the forward pulse and the reflected pulse in the waveguide is

$$2t_s = \frac{2l_s}{c\sqrt{1-\left(\frac{\lambda}{2a}\right)^2}}.$$

This delay is different depending on the construction of the radar and the position of its instruments.

The delay time of the radio pulses in the waveguide may vary as a consequence of a change in the length of the waveguide or of the transmitter frequency. These changes are so insignificant while the radar is being used that they need not be taken into account. The delay of the radio pulses in the radar waveguide may be considered practically constant.

Delay of the Reflected and Forward Pulses in the Receiver Circuits

As was indicated above, the delays of the reflected and forward pulses in the video-frequency amplifier stages are equal, and thus they have no effect on the radar systematic error when measuring the range. Therefore let us evaluate the delay of the reflected pulse only in the circuits of the intermediate-frequency amplifier (i-f amplifier). As is known [19], when a resonance frequency voltage jump is fed to the input of the resonance amplifier, the middle of the front of the output voltage build-up will correspond to the group delay time. From this it follows that the middle of the leading front of the reflected pulse at the amplifier output also will correspond to the group delay time in the case being investigated. It is convenient to express the group delay associated with a given pass band of the receiver Δf by the dimensionless quantity $\Delta f t_0$.

For an i-f resonance amplifier (Figure 3.5) having a sufficiently large number of stages the approximation equation of [19] is valid

$$\Delta f t_0 = \frac{1}{\pi}\sqrt{2n \ln S_\Delta}.$$

where t_0 is the group delay time;

n is the number of amplifier stages;

S_Δ is the disparity of the amplifier amplification factor.

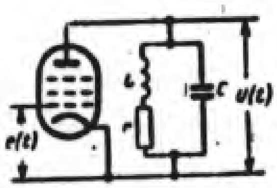


Figure 3.5. The circuit of a resonance amplifier stage.

If the receiver pass band is related to a disparity of the amplification factor of $S_\Delta = \sqrt{2}$, which corresponds to a drop of the frequency characteristic of 3 dB, then

$$\Delta f t_0 = 0.265\sqrt{n}. \qquad (3.3)$$

This equation shows that with an increase in the number of i-f amplifier stages, the delay increases in proportion to the square root of the number of stages. Figure 3.6 presents a graph which illustrates the character of the increase of the delay time as a function of the number of stages with a given pass band of the receiver.

For a band-pass amplifier with two coupled circuits (Figure 3.7), having a large number of stages, the following approximation equation is valid

$$\Delta f t_0 = \frac{\sqrt{2}}{\pi}(2 \ln S_\Delta)^{1/4} n^{3/4}.$$

If one assumes, as before, that $S_\Delta = \sqrt{2}$, then

$$\Delta f t_0 = 0.411 n^{3/4}. \qquad (3.4)$$

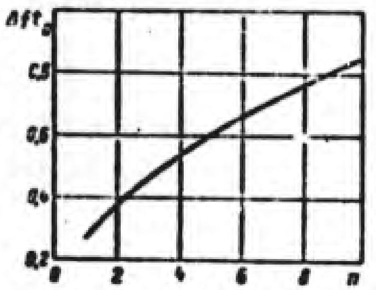


Figure 3.6. The delay time as a function of the number of stages n of a resonance amplifier.

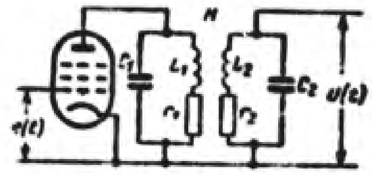


Figure 3.7. The circuit of a band-pass amplifier with two coupled circuits.

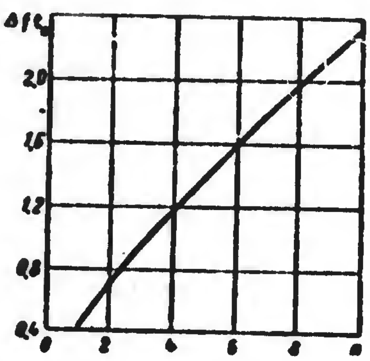


Figure 3.8. The delay time as a function of the number of stages n of a band-pass amplifier with coupled circuits.

The graph of Figure 3.8 illustrates the character of the increase of the delay time with a given pass band of the receiver as a function of the number of stages for a band-pass amplifier with coupled circuits.

From a comparison of expressions (3.3) and (3.4) and also of Figures 3.6 and 3.8, one may come to the conclusion that for a band-pass amplifier the delay time increases with an increase in the number of stages more rapidly than for a resonance amplifier, and for the same number of stages the delay time of a band-pass amplifier is larger. Thus, for example, when $n = 8$ for a resonance amplifier, $\Delta f t_0 = 0.75$, and for a band-pass amplifier, $\Delta f t_0 = 1.95$. In this case the delay time for a band-pass amplifier exceeds the delay time for a resonance amplifier by a factor of 2.6.

The stability of the delay value of the reflected pulse in the receiver circuits depends on the stability of the receiver pass band. As was shown above, it is high and consequently the constancy of the delay value of the reflected pulse in the i-f amplifier circuits also is great.

The time delay of a transmitter forward pulse t_{f1} may be determined from the following considerations. The transmitter forward pulse appears on the range indicator screen as a consequence of the fact that the submodulator pulse acts on one of the stages of the intermediate-frequency amplifier. This action may be investigated as a single jump of constant voltage, under the influence of which natural oscillations arise in the circuit of one of the stages of the intermediate-frequency amplifier. These oscillations are the inputs for the next stage. The delay time of the transmitter forward pulse t_{f1} will, consequently, equal the group delay time in (m - 1) stages of the intermediate-frequency amplifier since the stage in which the impact excitation circuit is introduced does not result in a delay. The path of the action of the submodulator's pulse in the intermediate-frequency amplifier naturally depends on the structural features of the submodulator and amplifier and their relative arrangement, and therefore it may differ not only in radars for various purposes but even in radars of the same type. As a consequence of this, the quantity (m - 1) — the number of stages participating in the amplification process of the transmitter forward pulse — must be determined experimentally. As in the previous case, there is no basis for considering that this time may vary while the radar is being used.

Let us analyze the components of the systematic error connected with the process of measuring the time delay of the pulse from the target with reference to the transmitter forward pulse.

The systematic errors which arise in this case depend on the type of device which carries out this measurement. Such devices have acquired the name of variable time delay devices. Wide application is being found for variable time delay devices which use the method of changing the phase of a standard frequency sinusoidal voltage.

When reading the distance, the time from the beginning of the transmitter forward pulse to the beginning of the pulse from the target is compared with the number of cycles of this standard frequency [39]. These devices are characterized by a small error in measuring the time delay, by an extremely smooth variation of the delay, and by the possibility of obtaining high rates for the delay change. A quartz-crystal oscillator as a rule is used as the standard frequency sinusoidal voltage oscillator in such circuits (Figure 3.9). This same quartz-crystal oscillator is simultaneously used as the source of the synchronizing voltage. The high accuracy of the range measurement is achieved because of the high stability of the quartz-crystal oscillator frequency. The sinusoidal voltage is sent to the phase splitter, at the output of which voltages are taken with phase shifts of 90, 180, 270 and 360° with respect to the input voltage. The constant amplitude voltage with a smoothly adjustable phase taken from the phase shifter is used for shaping the delayed pulses. The fine range sweeps and the marker line pulses are formed from these pulses.

The functional circuit of a station range measurement channel has two indicators: a coarse range indicator (CRI) and a fine range indicator (FRI). Such a functional circuit, which also has a variable time delay device of this type, is shown in Figure 3.10. This circuit consists of five main channels which are designed for shaping:

- the transmitter triggering pulse;
- the sweep voltage for the "coarse range";
- the selector pulse;
- the sweep voltage for the "fine range";
- the marker line pulse.

In this circuit the moment for triggering the transmitter is rigidly tied in time to a specific phase of the sinusoidal voltage of the quartz-crystal oscillator. In order to increase the precision of the adjustable delay time, the frequency of the quartz-crystal oscillator exceeds the repetition frequency of the main pulses F_p by a factor of n; therefore, frequency division is carried out in the channel for shaping the transmitter triggering pulse. When this is done, the frequency of the main pulses turns out to be n times smaller — that is, equal to F_p. When removing the systematic error, it is necessary to observe the leading front of the main pulse on the fine range indicator, FRI. Consequently the transmitter must be triggered with a time delay with respect to the leading front of the frequency divider pulse. This delay is achieved by means of a selector (Figure 3.10).

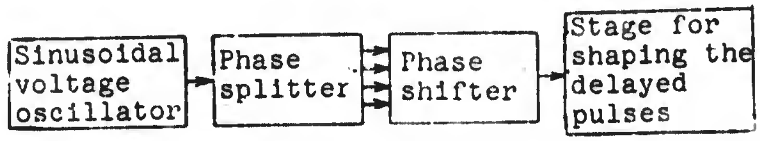


Figure 3.9. The functional circuit of a variable time delay device.

The output pulses of the frequency divider trigger the sweep and brightening circuit of the coarse range indicator, CRI. All the distances are simultaneously scanned on this indicator, whereas on the fine range indicator, a limited part of the distance is scanned on a large scale. The selection of the distance on the CRI sweep which is to be scanned on the FRI and the brightening of the working part of the range sweep of the FRI is achieved by a signal selector (Figure 3.10). In order to have the brightening pulse coincide in time with the working part of the FRI range sweep, the signal selector also is triggered by the standard frequency pulses which are shaped from the phase shifter voltage. The repetition frequency of the selector pulses equals the repetition frequency of the radar main pulse.

Shaping the sweep voltage for the fine range and the marker line pulse is accomplished, as we already showed above, from the constant amplitude voltage with the smoothly adjustable phase taken from the phase shifter. The phase shifter shifts in time the sweep and range marker line of the FRI by the distance being read off from the distance data unit (Figure 3.10).

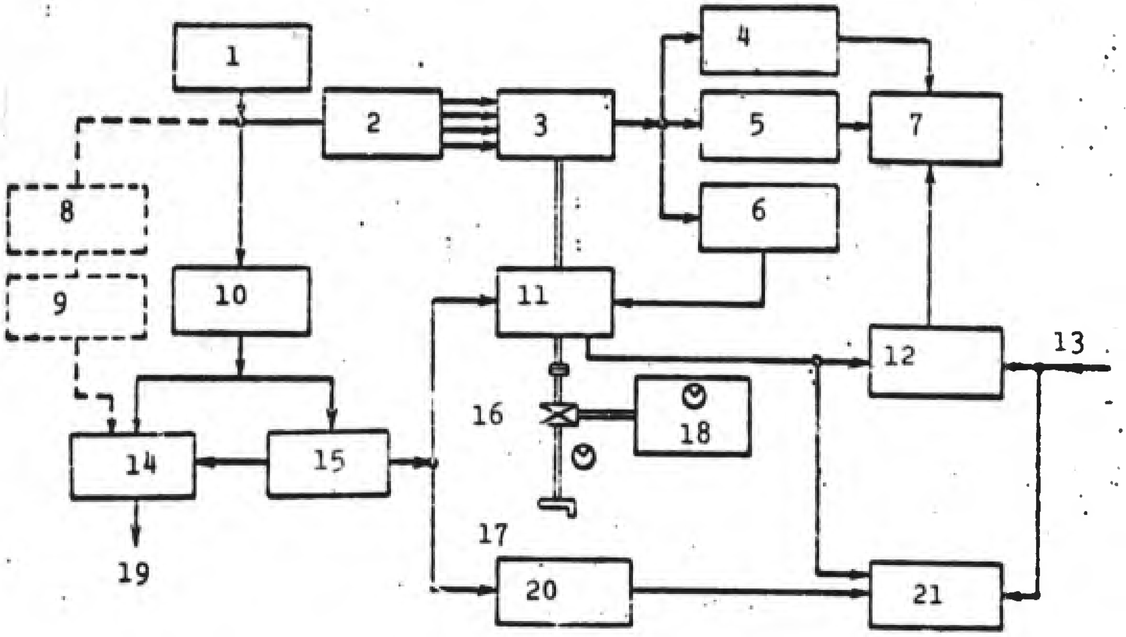


Figure 3.10. The functional circuit of a device for measuring the range in a station having CRI and FRI indicators.

1 - Quartz-crystal oscillator; 2 - Phase shifter; 3 - Phase shifter; 4 - Range sweep circuit for the FRI; 5 - Range marker line circuit for the FRI; 6 - Sharpener 1; 7 - FRI; 8 - Phase shifting circuit; 9 - Sharpener 3; 10 - Sharpener 2; 11 - Signal selector; 12 - Wave of the target and brightening signals; 13 - From receiver output; 14 - Selector; 15 - Frequency divider; 16 - Mechanical device; 17 - Control wheel; 18 - Distance data unit; 19 - Transmitter triggering; 20 - CRI sweep and brightening circuit; 21 - CRI.

The sources which give rise to systematic errors in such circuits are:

- instability of the quartz-crystal oscillator frequency;
- noncorrespondence between the rotor rotation angle of the phase shifter and the phase of the voltage at its output;
- the error in matching the range marker line with the front of the transmitter forward pulse;
- the error in removing the systematic error because of an inaccurate setting of the quantity ΔD on the range scale;
- errors in determining the range which arise with automatic tracking of a target and the transmission of data along the synchronous transmission line.

Let us analyze these errors in somewhat more detail.

<u>The instability of the quartz-crystal oscillator frequency</u>. A deviation of the quartz-crystal oscillator frequency from the rated value leads to a systematic error in measuring the time delay between the transmitter forward pulse and the pulse reflected from the target:

$$\Delta D_f = \pm \frac{\Delta f_{qo}}{f_{qo}} D_s,$$

where Δf_{qo} is the deviation of the quartz-crystal oscillator frequency from the rated value;

f_{qo} is the rated value of the quartz-crystal oscillator frequency;

D_s is the range from the scale.

The relative instability of the quartz-crystal oscillator frequency is characterized by the quantity $\frac{\Delta f_{qo}}{f_{qo}} = (1-2)10^{-6}$. From this it follows that the systematic error of the time delay measurement even with maximum range to the target does not exceed several meters.

<u>Noncorrespondence between the rotor rotation angle of the phase shifter and the phase of the voltage at its output</u>. In order to change

the phase of the sinusoidal voltage, one may use the following types of phase shifters: bridge circuits consisting of active and reactive resistances (elements) potentiometric inductive or capacitive phase shifters. The capacitive and inductive phase shifters are the most modern ones.

The phase of the sinusoidal voltage at the phase shifter output during the distance measuring process must vary in strict correspondence with the rotor rotation angle of the phase shifter. However, under actual conditions, noncorrespondence occurs between the rotor rotation angle of the phase shifter and the phase of the voltage at its output. The reasons for this noncorrespondence are the inaccuracies in manufacturing the phase shifter and the changes of the phase and amplitude of the voltages sent from the phase splitter during operation. The measured range to the target is connected with the phase shifter rotor rotation angle ϕ by the relationship

$$D_{\text{изм}}=\frac{cT_{\text{кг}}}{2}\left(m+\frac{\varphi^{\circ}}{360^{\circ}}\right),$$

where m is the number of whole cycles of the alternating voltage;

T_{qo} is the period of the voltage from the quartz-crystal oscillator.

The voltage period of the quartz-crystal oscillator is

$$T_{\text{кг}}=\frac{1}{nF_{\text{к}}},$$

where n is the frequency division factor.

The true distance to the target corresponds to the phase shift angle ϕ_p. Consequently

$$D_{\text{ист}}=\frac{cT_{\text{кг}}}{2}\left(m+\frac{\varphi_{\text{и}}^{\circ}}{360^{\circ}}\right).$$

The systematic error in determining the range which arises because of the errors of the phase shifter equals

$$\Delta D_{\phi} = \pm \frac{cT_{\text{кв}}}{2} \frac{\Delta\varphi^{\circ}}{360^{\circ}},$$

where $\Delta\phi = \phi - \phi_p$ is the error in the amount of the phase shift. This phase shift error may reach values of 0.25 - 0.5°.

The error ΔD_{ph} may be reduced, as follows from the last expression, if one increases the frequency of the quartz-crystal oscillator. However, this is not always possible, since the fine range sweep is shaped, as a rule, from the same voltage (Figure 3.10). An increase of the quartz-crystal oscillator frequency leads to a proportional decrease of the range being simultaneously observed on the fine indicator, because the latter range is connected with the period of the sinusoidal voltage by the relationship

$$d_{\phi} = c \frac{T_{\text{кв}}}{4}.$$

Let us estimate the error in determining the distance which arises because of phase shifter errors. As an example, let us assume that $d_p = 3$ km, and $\Delta\phi° = 0.5°$, then $\Delta D_{ph} = \pm 8.3$ m.

Figure 3.11 shows typical graphs characterizing the errors in the phase shift value as a function of the rotor rotation angle of a capacitive phase shifter. As follows from the figure, the error has a periodic character. The systematic error in determining the distance because of phase shifter errors consequently may be reduced. In order to do this, it is necessary to determine the systematic error several times at distances which are within the limits of the range corresponding to the same period of the quartz-crystal oscillator. Thus, for example, if — when measuring the range to three objects which satisfy this condition — systematic errors equal to ΔD_1, ΔD_2 and Δ_3 are obtained, the resulting value of the systematic error should be taken as equal to the arithmetic mean

$$\Delta D = \frac{\Delta D_1 + \Delta D_2 + \Delta D_3}{3}.$$

In this case, the error introduced by the phase shifter is averaged.

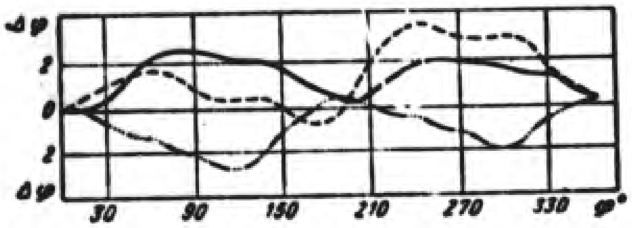


Figure 3.11. The error in the magnitude of the phase shift as a function of the rotor rotation angle of the phase shifter.

The error in matching the range marker line with the front of the transmitter forward pulse. When setting the conditional zero of the range, about which we will have more to say later on, in order to eliminate the systematic error in determining the range from the measurement results, the range marker line is matched with the leading front of the transmitter pulse.

Several factors affect the size of the error in matching the range marker line with the leading front of the transmitter pulse:

1. The transmitter forward pulse marking and the electronic marker line do not have clearly expressed matching points.

2. The electron beam which shapes the transmitter forward pulse and the marker line have a finite diameter.

3. The noise voltage at the receiver output changes the point at which the marker line is matched with the transmitter forward pulse. Thus the following error occurs

$$\Delta D_c = \pm d_p \frac{\Delta l}{l},$$

where d_p is the range corresponding to the length of the fine range sweep;

Δl is the error in matching the range marker line with the leading front of the transmitter forward pulse;

l is the sweep length of the fine range indicator.

In the absence of noise, the error in matching the range marker line with the leading front of the transmitter forward pulse Δl is determined basically by the rise time of this pulse, and is approximately 0.1 of the pulse front duration.

The magnitude of the error depends on the sweep scale of the range indicator, and the subjective characteristics of the operator. In order to reduce the error, it is advisable to have a large sweep scale and to use a fine range indicator with an electronic marker line. It is advisable to establish the zero range during preventive inspections with a fixed amplification of the receiver.

<u>The error in removing the systematic error because of inaccurate setting of the quantity ΔD on the range scale</u>. This error lies in the limits of a scale division of the fine range scale. All the values of this error have the same probability — that is, they are subject to the law of uniform distribution. The mathematical expectation of the quantity ΔD consequently equals half of a scale division of the fine range scale.

<u>The error in transmitting data along the synchronous transmission line</u>. This error depends on the direction and speed of the target. Consequently, it cannot be eliminated.

Of all the listed components of the systematic error, the error resulting from the noncorrespondence between the phase shifter rotor rotation angle and the phase of the voltage at its output is the greatest, but even it is not large as was shown above.

The analysis conducted shows that the main source of systematic error in the range measurement of radars making a fine measurement of the target coordinates is the pulse delays in the transmitter-receiver channel and the wave-guide channel. The size of the radar systematic error usually is within 40 - 80 m.

The systematic error in determining the range also depends on the length of the waveguide, as a consequence of which it must be determined and eliminated to the maximum possible accuracy when initially installing the radar.

However, while the equipment is being used a change in the time delays occurs, and consequently the systematic error of the range measurements change. Slow changes take place because of gradual changes of the component parameters, and discrete changes take place when parts are replaced. Since it is not possible to continuously monitor and remove the systematic error while the station is being used, an accumulation of errors occurs. The stability of the systematic range error in time and also under various operational conditions is a very important parameter. The stability of the systematic range error is determined as frequently as necessary. This is done while the equipment is being used to monitor the error. When the frequency of the preventive inspection is specified, it is necessary to determine what value this error may reach during the period. It is necessary to consider the fact that, since the systematic error in determining the range is determined by measuring this quantity with the radar station itself, the error will change when the measurements are repeated. The systematic errors will have differences which are of a random character.

3.4. Determining and Removing the Systematic Error in the Range Measurement

In order to find the systematic error, it is necessary to know the range measured by the radar and to compare this with the range to the target determined by another more precise instrument. This value of the range is taken as the true value.

Let us investigate several possible methods for determining the systematic error in the range measurement — the geodesic method, the method of two radar stations, and the instrumental method.

The Geodesic Method

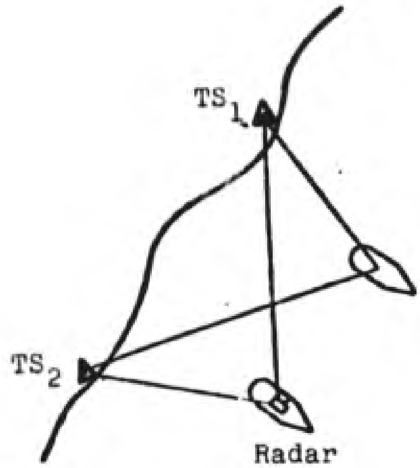


Figure 3.12. Determining the systematic error of a station by the geodesic method.

Determining the systematic error by the geodesic method is done by comparing the distances to the object as measured by the radar and by the geodesic method. The distance measured by the geodesic method is taken as the true distance.

The geodesic distance may be determined, for example, from the measurement data of theodolite stations. In this case, theodolite stations which determine the bearing to the radar-equipped ship and an ancillary ship (Figure 3.12) are established on the coast. The cross bearings using the theodolite stations TS_1 and TS_2 must be performed simultaneously with the distance measurements to the auxiliary ship, using the radar on the ship being checked. The systematic error in the range measurements may be determined from the equation

$$\Delta D = \frac{\sum_{i=1}^{n} (D_{r_i} - D_{g_i})}{n},$$

where D_{r_i} is the distance between the radar-equipped ship and the auxiliary ship as measured by the radar;

D_{g_i} is the distance between the ships determined on the basis of the geodesic measurements;

n is the number of measurements.

The complexity of organizing the measurements and the dependence of the error on meteorological conditions are basic defects of this method for determining the systematic error. The systematic error may be determined more simply by measuring the range to a stationary local object. The following may be used as such an object:

- an object which can clearly be seen by the radar station which is practically a point object (a chimney, lighthouse, etc.);
- a reflector especially set up on the shore or on a breakwater;
- a ship.

The signal from the object whose range is being measured naturally should not be masked by signals reflected from the vicinity, or some other objects.

The systematic error in the range measurement may be determined with a sufficiently large number of measurements as the distance between the arithmetic mean of the range measurements and its actual value as found by the geodesic method;

$$\Delta D = \frac{\sum_{i=1}^{n} D_{r_i}}{n} - D_0$$

where D_{r_i} is the distance as measured by the radar;
D_0 is the distance determined by the geodesic method.

Let us use this method for determining the systematic error of both shore-based radars and also shipboard radars. The ship with the radar being checked must lie near a mole or a sea-wall. The precision in determining the systematic error will be larger, the smaller the error of the radar distance measurement, and the larger the number of measurements.

The Method of Two Radar Stations

We use this method for determining the systematic error of shipboard radars. Three ships participate in the measurements, one of which carries the radar being checked (C). On another, there is a monitoring radar (M). The third ship is an auxiliary ship (A) (Figure 3.13). When making the measurements, the possibility of drift and swaying of the ships must be eliminated. The systematic error is measured in the following way. Measurements of the distance to the auxiliary

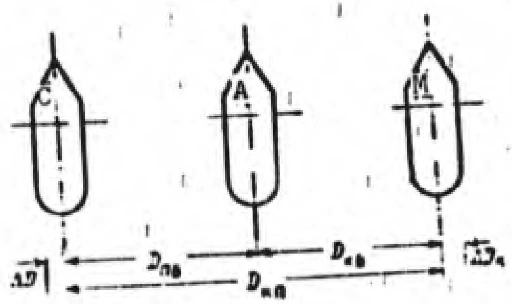


Figure 3.13. Determining the systematic error of a station by the method of two radars.

ship are made by the radar being tested and measurements of the distances to the auxiliary ship and to the ship carrying the radar being checked are made by the monitoring radar.

The systematic error of the range measurement is determined from the equation

$$\Delta D=\frac{\sum_{i=1}^{n}[D_{i(CA)}-(D_{i(MC)}-D_{i(MA)})]}{n}, \tag{3.5}$$

where $D_i(CA)$ is the distance from the ship with the station being checked to the auxiliary ship, as measured by the radar being checked;

$D_i(MC)$ is the distance between the ships which have the monitoring radar the the ship with the radar being checked, as measured by the monitoring radar;

$D_i(MA)$ is the distance between the ship with the monitoring radar and the auxiliary ship, as measured by the monitoring radar.

As follows from equation (3.5) and Figure 3.13, with this measurement method, the systematic error of the monitoring radar ΔD_M is eliminated from the calculation and the difference of the three distances gives the value of the systematic error of the radar being checked. Naturally, the precision in determining the systematic error is better, the higher the measurement accuracy of the monitoring radar station. However, a relatively low measuring accuracy of the monitoring radar may be compensated by increasing the number of measurements. In order to eliminate the errors introduced by the phase shifter from the measurement results, it is necessary to perform the measurements at several distances which differ from one another by an amount which is found to be within the limits of the range corresponding to one voltage cycle of the quartz-crystal oscillator.

The necessity of making special provisions, the complexity of the organization for performing the measurements and the effect of meteorological conditions on the precision of the results constitute the main disadvantages of this method for determining the systematic error.

The Instrumental Measurement Method

For the instrumental method of determining the systematic errors of radar stations, a systematic error meter may be used.

Figure 3.14 shows the block diagram of the instrument. It consists of three main parts: a high-frequency unit, a marker generator and an oscillograph. It may operate both with radiation into space and also without radiation into space. In the first case, it is connected with the output of the radar directional coupler, and in the second case it is coupled with the radar by means of an antenna.

The radar high-frequency pulses are sent either through the directional coupler or through the horn antenna to the input of the instrument high-frequency unit where attenuation and rectification take place. The video pulses are sent to an oscillograph for triggering the sweep circuit and to the marker generator. The marker generator generates a series of video pulses with a rigidly fixed interval between them. These video pulses are sent to a klystron oscillator and produce a modulation of its high-frequency oscillations. The re-emitted high-frequency pulses are sent through the directional coupler (or through the horn antenna) to the receiver input and then to the radar indicator (Figure 3.15). The high-frequency pulses in addition are rectified and sent to an oscillograph with a fast sweep.

A special sinusoidal-signal calibration oscillator installed in the oscillograph enables one to determine the time delay in the

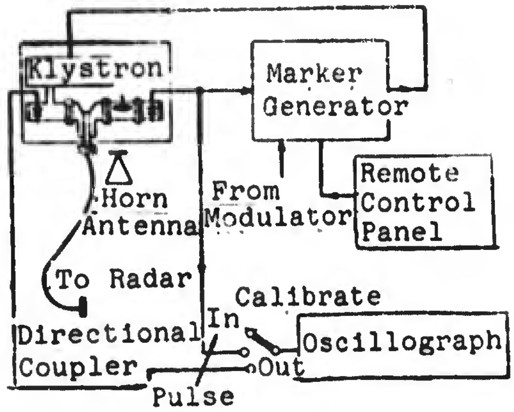


Figure 3.14. Block diagram of a systematic error meter for a radar.

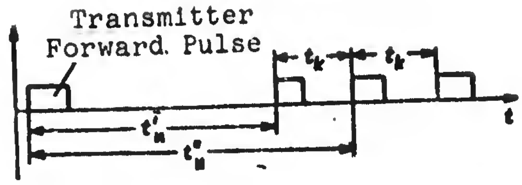


Figure 3.15. Pulses from the meter on a radar indicator screen.

instrument itself t_{del}. The time interval between the beginning of the transmitter forward pulse and the beginning of the first marker pulse is

$$t'_{\text{и}} = t_c + \frac{2D_0}{c} + t_{\text{зад}}.$$

where t_c is the time of the pulse delay when it passes through the transmitter-receiver channel and waveguide channel of the radar in the forward and reverse directions, which causes systematic errors in the range determination.

Consequently,

$$t_c = t'_{\text{и}} - \frac{2D_0}{c} - t_{\text{зад}}.$$

from this the systematic error for the range can be determined:

$$\Delta D = D'_{\text{и}} - D_0 - \frac{t_{\text{зад}} c}{2},$$

where D_0 is the distance between the radar antenna and the horn antenna of the instrument h-f unit;

D'_p is the distance measured by the radar to the first marker pulse.

If the systematic error of the measurement is performed for the k-th marker pulse, then

$$\Delta D = D_{\text{и}}^{k} - D_0 - \frac{t_{\text{зад}}c}{2} - \frac{(k-1)t_{\text{м}}c}{2}.$$

In order to eliminate the error introduced into the measurement results by the radar phase shifter, the systematic error measurement is performed for five markers.

The advantages of this measurement method are:

- the possibility of measuring the systematic error both in determining the distance and also in determining the angular coordinates;
- the high accuracy in determining the systematic errors;
- the possibility of determining the systematic error without radiation into space;
- no special provisions are required; such as establishing a geodesic link between local prominances or objects;
- the possibility of eliminating the errors introduced by the phase shifter unit from the measurement.

The disadvantages of the measurement method being investigated are:

- the necessity of closely linking the instrument to the radar being checked which, for example, under the conditions aboard ship, is not always convenient;
- the complexity of preparing and tuning the instrument when making measurements.

In order to check radar stations for distance measurements, one may also use ultrasonic range calibrators in principle.

Figure 3.16 shows the schematic diagram of an ultrasonic range calibrator. It consists of a number of ultrasonic delay lines (UDL),

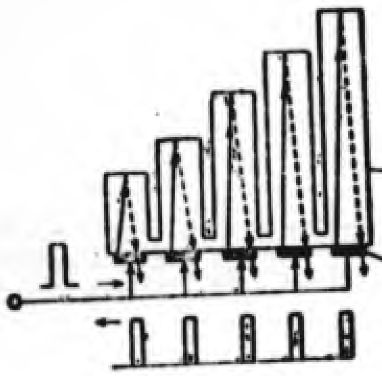


Figure 3.16. Schematic diagram of an ultrasonic range calibrator.

each of which delays the pulse for a specific time interval. The signal delay time in the acoustic line t_d depends on its length l and on the propagation velocity of the ultrasonic oscillations in it v, that is,

$$t_{з}=\frac{l}{v}.$$

The range calibrator is connected to the intermediate-frequency amplifier circuit (Figure 3.17). When an impulsive excitation pulse, which is a consequence of the submodulator leakage pulse, is sent to the calibrator input, the calibrator reproduces a series of pulses which are shifted with respect to one another by fixed time intervals. A set of calibration pulses will be observed on the radar indicator together with the transmitter pulse. Knowing the distance between the calibration pulses, we can determine the systematic error of the range determination. Ultrasonic range calibrators may be made both in the form of portable instruments and also instruments which are built into the radar.

The small overall size and weight, the simplicity of the coupling with the radar and the absence of a power supply source are factors which recommend the use of such a measuring device.

The station systematic error in the range measurement must be eliminated from the measurement results with the maximum possible degree of accuracy after it has been determined. In order to do this, it is necessary that when the leading edge of the transmitter pulse coincides with the range marker line, the reading on the range scale corresponds to the negative value of the systematic error (Figure 3.18). This reading is designated the <u>conditional zero of the range</u>.

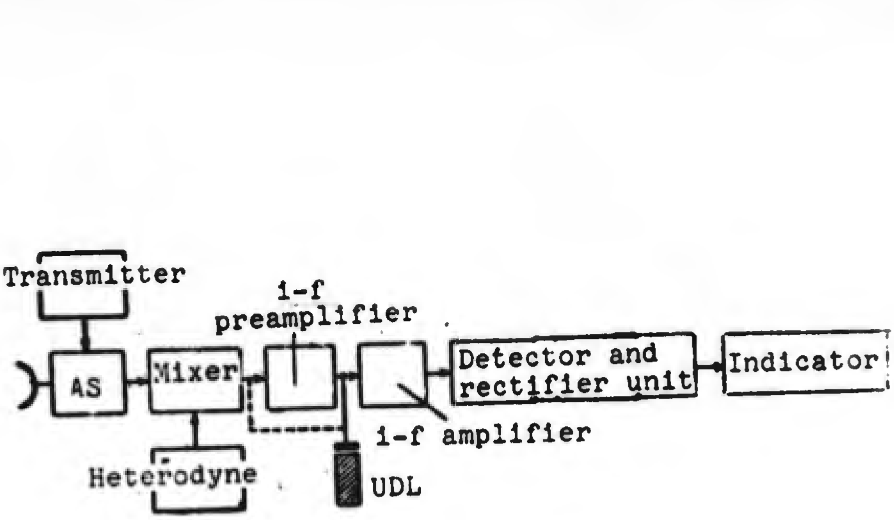


Figure 3.17. The circuit for connecting an ultrasonic range calibrator to a i-f amplifier channel.

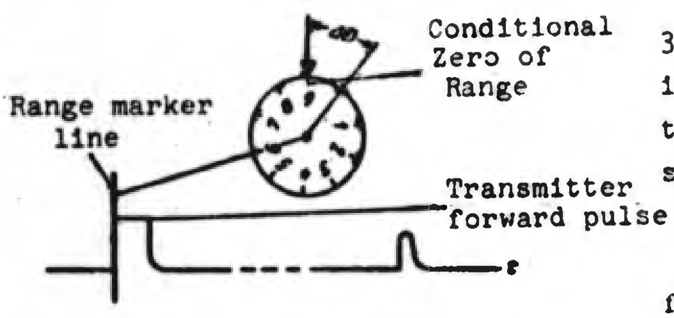


Figure 3.18. The diagram illustrating the elimination of the systematic error in a range determination.

As follows from Figure 3.10, three methods for removing the systematic error from the range measurement are possible in principle.

1. On the FRI, the leading front of the transmitter forward pulse is super-imposed on the range marker line. After this, by means of a special mechanical device, the distance data unit is disconnected from the phase shifter. By turning the control wheel, a reading corresponding to the conditional zero of the range is established on the range scale.

2. A reading corresponding to the range conditional zero is set on the range scale. By changing the sinusoidal voltage phase of the quartz-crystal oscillator by means of a phase-shifting circuit, delay of the transmitter triggering pulse is achieved with respect to the pulse for triggering the fine range indicator sweep. Thus, the leading front of the transmitter forward pulse is matched with the range marker line. In this case, the peak voltage of the quartz-crystal oscillator

must enter the selector, not from the sharpener 2, but through a phase-shifting circuit, and the auxiliary sharpener 3, as is shown in Figure 3.10 by the dashed line.

3. The leading edge of the transmitter forward pulse is matched with the range marker line on the fine range indicator. By rotating the scale of the fine reading selsyn, a reading corresponding to the range conditional zero is established.

It is advisable to measure the systematic error in the range at the manufacturing plant and after mounting the radar on the object.

As was already established, the stability of the systematic error in time and under various operational conditions is great. Consequently, it is inadvisable to measure the systematic error of the range determination during operation.

As was shown in § 3.2, a change of the time delays in the stages of the synchronization unit, modulator and receiver device, leads to a displacement of the range marker line with respect to the front of the transmitter forward pulse and to the emergence of an error when measuring the range (Figure 3.4). Therefore, during preventive inspections, it is necessary to monitor the removal of the systematic error regularly, that is, to check that — when the leading front of the transmitter forward pulse coincides with the marker line of the fine range indicator — the reading on the range scale corresponds to the conditional zero of the range.

3.5. Systematic Errors When Determining the Direction to an Object

On a ship, the radar stations are both interconnected and also connected with numerous shipboard systems which use, in one form or another, data sent from these radars. In addition, data on the angular coordinates of objects also enter these systems from optical facilities of the ship. As a consequence, the ship radar must be matched with the optical devices — that is, the electric axes of the radar antennas must match the axes of the optical devices. Only in this case will the radars and optical devices give the same readings when measuring the angular coordinates of objects. A preliminary condition for doing this is the matching of the optical-mechanical axes and electric axes of antenna devices. This is done in the radar manufacturer's plant. In order to do this, an auxiliary optical device is used — a cold adjustment tube (CAT) which is mounted on the radar antenna. During the complete tuning of the equipment at the plant, the axis of this device must be aligned with the electrical axis of the antenna. This is accomplished by repeatedly directing the radar and the auxiliary optical device to a distant object with small angular dimensions, and removing the misalignment between the optical axis of the device and the antenna electrical axis with the maximum possible accuracy. After this, the position of the optical device holder is rigidly fixed, and the residual systematic error is recorded in the radar logbook.

Subsequently, when mounting a radar on a ship, the alignment of the ship's optical facilities and the auxiliary optical devices mounted on the antennas is achieved by ordinary methods. When this is done, the residual systematic errors not removed at the plant are taken into account.

Instrumental systematic errors in measuring the direction are due to two main causes, which depend on the parameters of the receiver-indicator device, and the mechanical features of the antenna devices and the devices for horizontal levelling and orientation of the antennas.

Thus, for example, the systematic errors which arise when measuring the direction by bisecting a signal arise as a consequence of:

- the asymmetry of the main lobe of the radar antenna directional pattern;
- the nonlinearity of the time sweep within the limits of the marking from the object;

- the error in the time of shaping the electronic marker line: the moment of shaping the electronic marker line must coincide with the moment when the axis of the antenna directional pattern coincides with the antenna optical-mechanical axis;
- the error in matching the electronic marker line with the middle of the marking from the object.

When the optical-mechanical axis is aligned with the electrical axis of the antenna, part of these errors are eliminated.

The nature of the instrumental errors arising when determining the direction, which depend on the mechanical properties of the antenna systems and the systems for the orientation and stabilization of them, is extremely diverse and peculiar to each station.

One of the sources of systematic errors in determining the direction is the elastic deformations of the support structure of the radar antenna-rotation device due to the effects of variable wind, inertial loads and weight loads. For example, the cause of systematic errors when determining the direction may be a change of the angle of inclination of the antenna rotation axis because of deflections or imprecise centering. Thus, an antenna deflection of 1 mm or an imprecise centering of it with respect to the rotation axis by the same amount causes a systematic error amounting to values of 1.5 thousandths of a degree, which certainly must be considered in a number of systems.

Imprecise horizontal levelling of antenna systems also leads to systematic errors when measuring direction. When mounting a radar on a ship, the antenna systems are aligned, which includes the following steps:

- checking the precision of the horizontal levelling of the antenna system base;
- determining the systematic error in the bearing angle and angle of elevation;

- removing the systematic errors in the angles.

During the radar operation, the alignment must be periodically checked. Therefore, the systematic errors in the angles must be determined and eliminated with the maximum possible accuracy.

The following methods are possible for determining the systematic error in the bearing angle:

- the method of measuring a distant object;
- the instrumental measurement method.

<u>The Method of Measuring a Distant Object</u>

The radar systematic error in the bearing angle on the basis of a distant coastal object is determined by comparing the bearing angle values measured by the radar and an optical sighting device. The arithmetic mean of the bearing angle, obtained when making measurements with the optical sighting device, is taken as the actual bearing angle.

A single coastal reference point or a special reflector observable on the radar indicator screen and visible to the optical sighting device is usually taken as the distant object. The object must be far enough away from the radar that its angular dimensions are such that it can be considered a point target. In that case, when the radar antenna and the optical sighting device coincide, the systematic error may be determined from the expression

$$\Delta q = \frac{\sum_{i=1}^{n} \Delta q_{b_i}}{n},$$

where Δq_{b_i} is the deviation of the target from the sighting device axis when the radar is precisely directed at the object.

If the radar antenna and the optical sighting device are offset

$$\Delta q = \frac{\sum_{i=1}^{n} (q_{r_i} - q_{b_i})}{n} \pm a, \tag{3.6}$$

where q_{r_i} is the bearing angle measured by the radar;

q_{b_i} is the bearing angle measured by the optical sighting device;

a is the parallax error.

In this case, the parallax error connected with the fact that the radar antenna and the optical sighting device are offset must be eliminated from the measurement errors.

Instrumental Measurement Method

Measuring the systematic error by the instrumental method is performed in the following way.

The remote high-frequency part of the instrument is set up in such a way that the instrument horn antenna is directed toward the radar antenna. In that case when the systematic error of a shipboard radar is being determined, care must be taken that there are no masts, superstructures, or other obstructions between the station antenna and the instrument horn antenna, as such obstacles will distort the antenna directional pattern. As a consequence of this, they will introduce an error into the measurement. The effect of the ship superstructure and masts on the operation of radar stations was investigated in [28].

The remote part of the instrument must be located at a distance from the radar antenna which will satisfy expression (2.39). For synchronizing the instrument, triggering pulses are sent from the station along a special cable. The bearing angle to the instrument is measured by means of the radar and the optical sighting device (Figure 3.19, a). The value of the angle measured by the optical sighting device is taken as the true value.

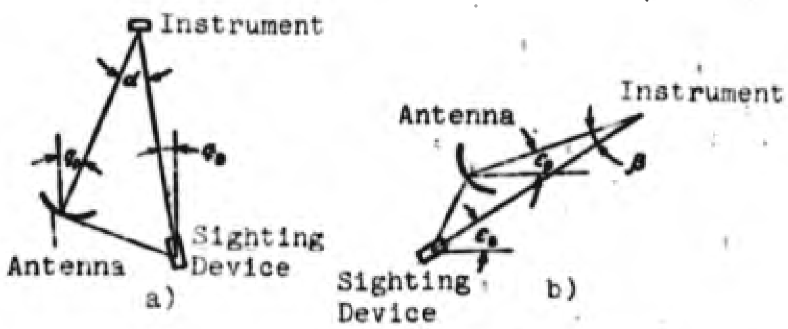


Figure 3.19. Relative position of the radar antenna, sighting device and instrument when determining the systematic error in the bearing angle (a) and in the elevation angle (b).

The systematic error in the bearing angle is determined from Equation (3.6). After the systematic error of the bearing angle is determined, it must be eliminated with the maximum possible accuracy. The elimination of the systematic error is done by rotating the stator of the bearing angle selsyn-data unit by an amount equal to the value of the error. When doing this, one must take into account the sign of the error.

Determining and Removing the Systematic Error in the Angle of Elevation

Two methods are possible when determining the systematic error in the angle of elevation, the same as when determining the systematic error in the bearing angle: on the basis of a remote object, and by the instrumental method. The remote object may be a reflector suspended from pilot balloons which are in free flight, an airplane, or a helicopter. The reflector or aerial target is tracked by the radar and the optical sighting device, and on the basis of this, the radar systematic error in the elevation angle is determined.

The systematic error in the angle of elevation may also be determined by the instrumental method (Figure 3.19, b).

The systematic error in the angle of elevation is determined from the equation

$$\Delta\varepsilon = \frac{\sum_{i=1}^{n} (\varepsilon_{r_i} - \varepsilon_{b_i})}{n} \pm \beta,$$

where ε_{r_i} is the elevation angle measured by the radar;

ε_{b_i} is the elevation angle measured by the sighting device;

β is the parallax error.

When determining the systematic error in the elevation angle, it is necessary that:

1 - The ship superstructure and masts, as much as possible, have no effect on the measurement process;

2 - The effect of the Earth's surface has no influence on the measurement results.

The first requirement is similar to the requirement imposed when measuring the bearing angle. Let us discuss in somewhat more detail the second requirement.

The determination of a target elevation angle under conditions when the influence of the Earth's surface has practically no effect, from the point of view of the measurement errors, in no way differs from the determination of the bearing angle. The determination of a target elevation angles, when affected by the Earth's surface, leads to additional measurement errors, as shown by special studies [28].

The additional errors when the radar is directed on targets which are at small angles of elevation arise as a consequence of the distortion by the Earth's surface of the antenna directional pattern in the vertical plane. Errors arising with this are called <u>measurement errors of small elevation angles of the target</u>.

Figure 3.20 shows a typical dependence of the measurement error on the angle when determining the direction to an object on the basis

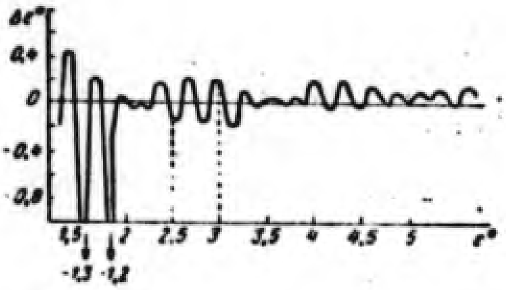


Figure 3.20. The error in the measurement of a target elevation angle as a function of the size of this angle. The width of the radar directional pattern in the vertical plane as measured at the zero level equals $2\phi_0 = 8°$.

of the center of the bundle of reflected pulses [28]. As is seen from the Figure, the effect of the water surface leads to the appearance of errors in the measurement of small angles of elevation. As follows from Figure 3.20, determination of target elevation angles less than $\phi_0/2$ is practically impossible, and when determining elevation angles larger than $\phi_0/2$, the error varies and decreases periodically with an increase of the elevation angle.

A similar phenomenon also occurs when determining the elevation angle on the basis of the equisignal zone method.

The measurement errors of small angles of elevation are systematic external errors. Naturally, it is necessary to determine the systematic error in the angle of elevation under conditions which eliminate the distortion of the measurement results because of the coincidence of external systematic errors which are not eliminated in the measurement process — that is, with angles for which there is "breaking away" of the main lobe of the radar antenna directional pattern from the water or Earth's surface in the vertical plane.

What has been said above concerning determining small angles of elevation refers both to measurement of the systematic error on the basis of a distant object, and also to the instrumental method of determining the systematic error. In order that the instrument horn antenna is located at the required height, its high-frequency unit is mounted on a mast, on a ship superstructure, or on a special tower on shore.

After determining the systematic error in the elevation angle, it must be eliminated with the maximum possible accuracy. Methods for removing the systematic error in the angle of elevation and in the bearing angle are the same.

4. MAINTAINING THE RELIABILITY WHEN FAILURES ARISE

4.1. Evaluating the States of Radar Elements

The reliability depends on the quality of the elements used in the equipment, their operating conditions, and the construction of the equipment. It is characterized by the absence of sudden and gradual failures and failures of the "faltering" type. The division of failures into sudden and gradual failures, to a considerable degree, is arbitrary. The difference between them is that with a gradual failure the parameter of the system or element changes slowly, but with a sudden failure the change is practically instantaneous (Figure 4.1).

The methods for detecting gradual failures basically reduce to establishing the tendency of the elements' parameters to change. As distinct from a gradual failure, a sudden failure may not be detected by these methods, since changes of the parameters when a sudden failure occurs take place very rapidly. Second, the moments when the failures arise are distributed randomly in time. The reason for a sudden failure usually is a hidden defect which is not detected before the beginning of the equipment's use.

Failures of the "faltering" kind are caused by occasional reversible changes of an element's parameters as a consequence of a change of the equipment's operating conditions. Failures of this kind have practically no effect on the operation of radars.

In actual systems, various laws govern the distribution of the time between failures. Consideration of this fact is extremely difficult. Therefore, in order to simplify the reliability calculations, one usually assumes that the failures of all the elements are subject to an exponential distribution law, the parameter of which is the constant rate of the failures

$$\lambda=\lambda(t)=\text{const.}$$

The failure rate is connected with the frequency of the failures by the following relationship:

$$\lambda(t)=\frac{f(t)}{P(t)},$$

where f(t) is the frequency (probability density function) of the failures;

P(t) is the probability of trouble-free operation of the element.

In practice, the rate of failures λ is the main index of the reliability of an element or of a system as a whole. The failure rate makes it possible to evaluate an element from the point of view of its reliability, independently of the other elements of the system, but it does not allow one to determine the effect of the state of each element on the state of the entire system. This effect is evaluated by the coefficient (or percent) of the failures. Let us assume that in a certain system the simplest flow of failures is realized. Then the number of failures B_j of the various elements of the system constitutes in the sum the total number B of system failures, that is

$$B=\sum_{j=1}^{n}B_j,\quad \sum_{j=1}^{n}\frac{B_j}{B}=1.$$

where n is the number of elements into which the system is arbitrarily divided ($j = 1, 2, \ldots, n$).

The ratio $\frac{B_j}{B} = k_{o_j}$ is called the failure coefficient. In connection with this

$$0 < \frac{B_j}{B} < 1.$$

The failure coefficients are close in their values to the relative probabilities of the system's failures which arise as a consequence of breakdowns of its various elements. This may be written in the following way:

$$P(A_j/X) = \lim_{B \to \infty} \frac{B_j}{B}.$$

From this, it follows that the failure coefficients of a system which has the simplest flow of failures converge (in terms of probability) to the relative probabilities of the failures of the system's various elements. When making practical calculations, one may assume that

$$P(A_j/X) \approx \frac{\lambda_j}{\sum_{j=1}^{n} \lambda_j}.$$

In [47] the quantities $P(A_j/X)$ are investigated as components of a certain additive probability measure which is distributed over the set of the system's elements. It can thus be assumed that the set of values of $P(A_j/X)$ is the differential series of the distribution of a random point of break in the connection between the system elements. Thus, the ordered set of the system elements, the relation between which may be represented by an oriented graph, is the region of possible locations of this point.

Knowing the structure of the system and the probabilities $P(A_j/X)$ makes it possible to find the probability that the index of the break point does not exceed a certain specified index ξ, that is:

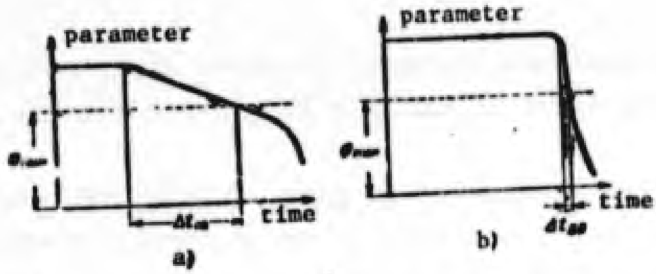


Figure 4.1. Change of an element's parameter with a gradual (a) failure and a sudden (b) failure.

Δt_g designates the time in which the parameter changes with a gradual failure; and Δt_s designates the same time with a sudden failure.

$$P(\xi) = \text{Вер}\{j \leqslant \xi\} = \sum_{j=1} P(A_j/X),$$

or the probability that the index of the break point will be less than ξ:

$$W(\xi) = 1 - P(\xi) = \text{Вер}\{\xi < j\}.$$

Having determined the values of $P(\xi)$ for each element of the system, we obtain the integral function of the distribution of the random break point on the set of system elements. Knowing the quantities $P(\xi)$, as will be shown below, enables us to select an efficient method for searching out malfunctions.

4.2. Methods for Evaluationg the Test Results of Radar Elements

In order to determine the state of an element, it is necessary to agree on what characteristic must be taken for this evaluation.

First of all, the output parameter of the element (or of a system as a whole) is a characteristic of its state. Let us assume that there is an abstract element, to whose input connections inputs are applied in the form of supply voltages and signals.

The element is considered in good working order if its output parameters are within the limits of the assigned tolerances. An element's output parameters go beyond these tolerance limits in the following cases:

— with an internal malfunction of the element;
— with the absence or deviation from the norm of one (or more) of the inputs (the signal or the power supply voltage).

One may introduce the concept of a composite output parameter for any radar element. The composite output parameter means the totality of all the output signals of a given element. The following is known about the composite output parameter of any element:

1. Limits always exist for the change in this parameter, within which the system remains operational.

2. The element's state may be evaluated on the basis of the degree to which the composite output parameter deviates from the norm.

It is necessary to distinguish between two cases of the deviation of an element's output parameter from the norm:

— the deviation is caused by an internal malfunction of the element (all the inputs are within the limits of the norm);
— the deviation is caused by a distortion or by the absence of one or more of the inputs.

In the first case monitoring the input parameter leads to a definite conclusion as to the necessity for the repair (replacement) of the element being monitored. In the second case, the element being monitored is in good working order, but is not able to function normally as a consequence of external causes. Monitoring of the states of the elements preceding the element being checked is required.

Monitoring the states of elements presupposes the presence of certain mutually exclusive results (alternatives). The mutual exclusiveness of the results is the property of testing (checking), which indicates that it is impossible for two different results to appear simultaneously. For example, the output parameter is not able to simultaneously be within the tolerance limits and also beyond these limits.

The number of mutually exclusive results d of a certain test depends on many factors of a technical character and on the training of the servicing personnel. Here it is necessary to point out that, with the existence of a failure probability for the element being monitored, the test may not have less than two mutually exclusive results ($d \geq 2$). On the other hand, an excessive increase of d leads to an unavoidable convergence of the mutually exclusive results, so that it is difficult to isolate each one from the others. Several kinds of test result sets with various values of d are shown in Figure 4.2.

Let us represent geometrically the tests together with their possible results in the form of trees (Figure 4.3). Such a representation enables one to subsequently construct geometric models for searching out malfunctions.

The majority of failures of elements (systems) are irreversible failures of a small number of the simplest parts. Therefore, generally speaking, the number of mutually exclusive results of the tests of an element may be increased to the number of the simplest parts in it. However, such an increase is not always advisable, since when there is such an increase the obviousness of the test evaluation decreases. A binary evaluation of test results ($d = 2$) is the simplest and most commonly used in practice. It reduces to the following: when

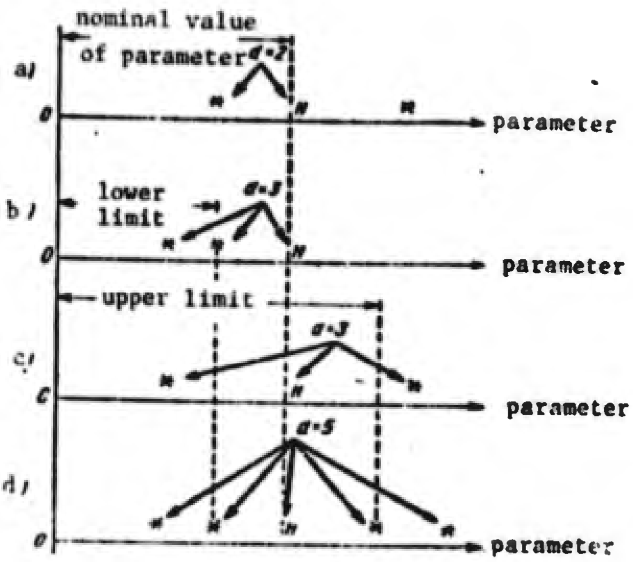


Figure 4.2. Examples of test gradation according to the number of results:

a) d = 2 ("in the norm — not in the norm");
b) d = 3 ("in the norm — at the limit — below the norm");
c) d = 3 ("below the norm — in the norm — above the norm");
d) d = 5 ("below the norm — on the lower limit — in the norm — on the upper limit — above the norm").

monitoring the state of an output parameter, the question arises as to whether the output parameter is within the limits of the established tolerances or not. The favorable characteristics of an evaluation "in the norm — not in the norm" are its simplicity and universality. Moreover, all the more complicated evaluations (d > 2) in the final analysis may be reduced to a binary evaluation (d = 2).

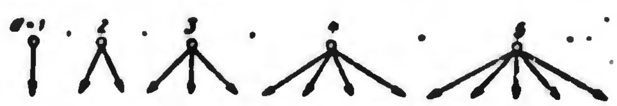


Figure 4.3. Graphical representation of tes[illegible]

As an example, let us cite the possibility of reducing the evaluation shown in Figure 4.4 ($d = 5$) to a binary evaluation. The reduction of complicated evaluations ($d > 2$) of test results to binary evaluations ($d = 2$) makes it necessary to divide one test with a complicated evaluation into several tests with binary evaluations.

It will be shown further on that a binary evaluation has still another favorable characteristic: when it is carried out, the set of system elements is divided into two nonintersecting subsets. Such a division can be used in everyday operation. Thus, a binary evaluation of test results ("in the norm — not in the norm") may serve as the basis for constructing any procedure for searching out a malfunction. The transition to more complicated evaluations leads to the necessity of increasing the qualifications of the servicing personnel. An increase of the complexity of a test evaluation also is inadvisable with automatic monitoring and malfunction searching, since it considerably complicates the monitoring equipment.

4.3. The Work Involved in Tests

Each test (check), depending on its specific contents, entails a definite amount of work.

The work involved in a test may be expressed in units of the working time required for conducting it. In economic calculations, it is customary to measure the work in man hours.

Let us determine the work as the expenditures of time which are necessary for one man to make in order to carry out the check. The time expenditures in conducting tests differ for specialists of various qualifications. However, one is always able to find a work value calculated for a specialist of average qualifications.

The sequence of tasks constituting the test may be planned, for example, by making use of a network graph. When this is done, it

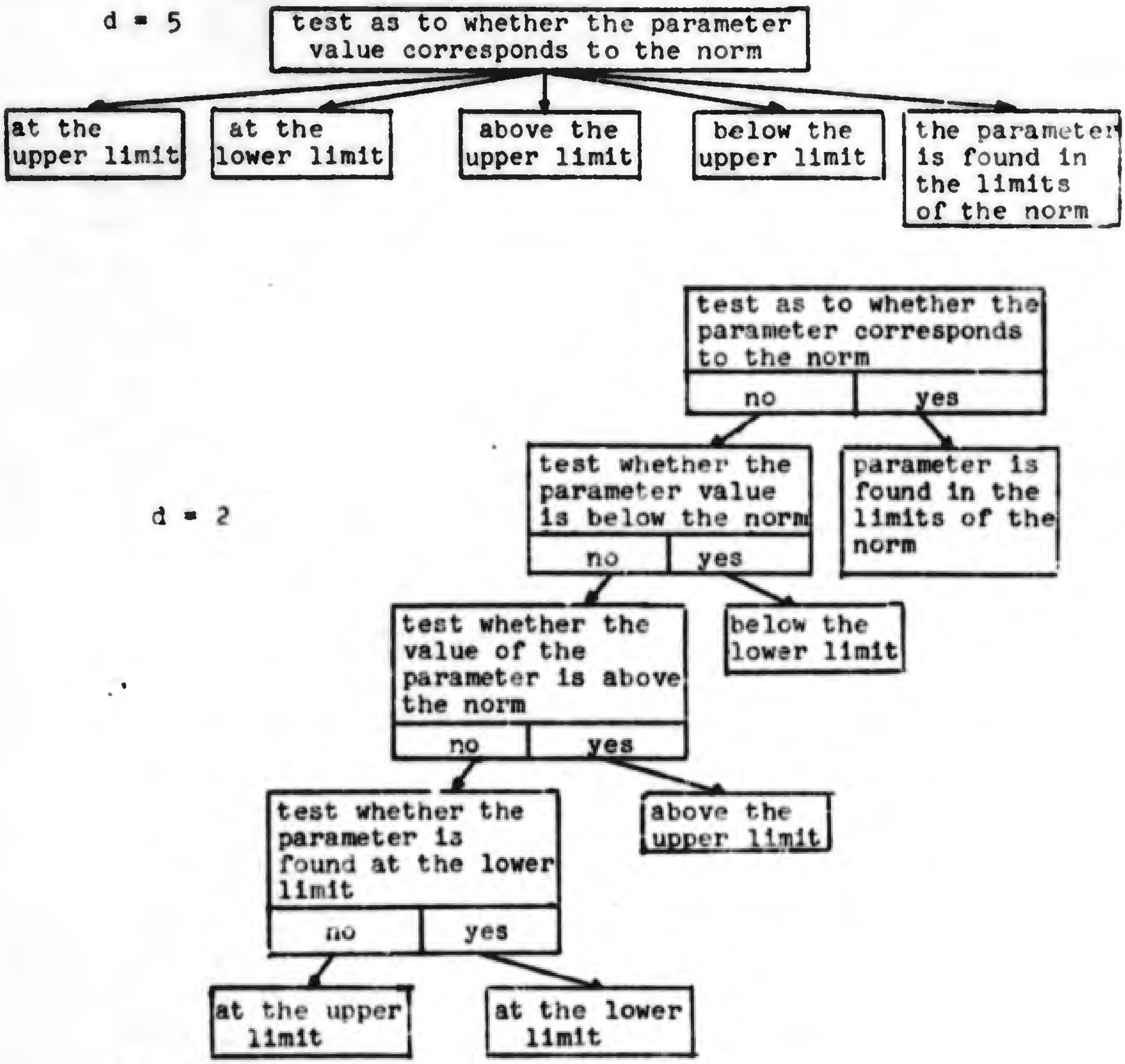


Figure 4.4. The reduction of a complicated evaluation (d = 5) to a binary evaluation (d = 2).

becomes possible to select the most efficient order for conducting the tests. Let us designate by t_j the work involved in testing the j^{th} element in a certain system. This work consists of two components:

$$t_j = t_{p\,j} + t_{i\,j}$$

where $t_{p\,j}$ is the work of the secondary tasks (for example, access to the j^{th} element or to its monitoring point);

$t_{i\,j}$ is the work of the main tasks (tests).

The first component of testing work depends on the relative position of the system devices and on the convenience of access to the element. The magnitude of the first component depends on the construction of the system and its arrangement on the object.

The second component includes the operation of readying the measuring equipment for connecting to the circuit, and conducting the measurements. The magnitude of this component depends on the construction of the system, the presence of monitoring-measuring equipment, and the quality of the technical documentation.

The test complexity is determined by the number of results d. When d = 2 ("in the norm — not in the norm"), the check has a lesser complexity than, for example, when d = 3. An increase of d leads to the fact that the time expended in determining the test results is increased.

On the other hand, as follows from Figure 4.4, the use of only a binary evaluation (d = 2) causes an increase in the number of tests. Thus, the selection of the number of results d of a test substantially affects the work. Tests may be divided into various groups:

— measurements using portable monitoring-measuring instruments;
— measurements using measuring instruments which are built into the system;

— a check of the element's operation based on built-in signallers which operate on the principle "yes — no";

— a check of the element's operation on the basis of external signs;

— test replacements of elements.

Measurements using portable monitoring-measuring instruments and test replacements require great expenditures of labor in comparison with tests of other types. Table 4.1 presents certain values for the expenditures of labor in carrying out various tests when they are performed by a single man and each test has only two results.

All the tests listed in Table 4.1 are not of equal value. Thus, for example, when monitoring the state of an ordinary tube stage, one may use the test replacement of a tube, external inspection of the stage, or measurement of its output signal. However, tests of the first two types do not always enable one to correctly make a judgement as to the state of the stage, and only a test of the third kind leads to a complete evaluation of its state.

TABLE 4.1. THE LABORIOUSNESS OF CERTAIN TESTS IN MAN HOURS

External inspection	up to 0.5
Opening of blocks	0.16 - 0.25
Measurement with a built-in instrument	0.017
Readying the measuring instrument (synchroscope, vacuum-tube voltmeter)	0.2 - 0.25
Measurement using an ammeter-voltmeter	0.017
Measurement using a synchroscope or vacuum-tube voltmeter	0.05 - 0.10
Test replacements	0.017

Cases are possible when a laborious test requiring the use of portable instruments is successfully replaced by a check using a signal monitoring device. In order to determine whether a voltage is present at a rectifier's output signal, bulbs are frequently used. One may determine whether the rectifier is operating from the lighting up of the signal bulbs. This points to the fact that, among the various tests of an element, one may select checks which have minimum work with the same value of d.

A reduction of the first component, $t_{p\ j}$ is possible only as a result of the most efficient arrangement of the instruments and units inside the system. A reduction of the second component, $t_{i\ j}$, is achieved by using fast measuring techniques.

One test is insufficient for determining the state of an element, particularly for determining the state of a system. Several tests must always be conducted in order to establish the good working order or failure of a certain part of the system. The number of such tests and the expenditures of labor on conducting them may be very different. The selection of the order in which a test is conducted depends on many factors, among them the work involved in the tests. In particular, monitoring a system's state sometimes does not begin with tests having the greatest work, but a gradual transition is carried out from the easiest tests to more laborious tests. A method for considering the work involved will be discussed below.

4.4. Sequences of Malfunction Searches

Sudden failures of electron equipment as a rule are a consequence of the breakdown of a small number of the simplest, closely interconnected parts (resistors, capacitors, tubes, etc.). Since the flow of failures in radar which is not in a standby condition is assumed to be an ordinary flow, the defective parts usually are found in a single, sharply limited section of the system and are in one of its elements. In connection with this, the problem of detecting a malfunction reduces to searching out the element containing the defective part. Obviously the accuracy of detecting the defective part (circuit) depends on the level at which the system is divided into elements.

The degree of accuracy is increased when the system is divided into a large number of elements of lesser complexity.

With the appearance of a sudden failure, it is assumed that the cause of the failure is found in one of the elements of the system — but where, precisely, is unknown. Also at what point of the system one should start checking for it is unknown.

One rarely succeeds in isolating the defective element with one test. Usually, several tests at various points of the system are required in order to do this. It is necessary to conduct these tests in such an order that, after conducting each of them, the system is divided into two (or more) subsets of elements, in only one of which is it necessary to conduct the following tests.

Let us assume there is a circuit of series-connected stages (for example, an amplifier) at the output of which there is no signal (Figure 4.5). Let us assume that as the first step a check is made for the presence of a signal at the output of the stage x_4. This check divides the system into two subsets of elements: $\{x_1, \ldots, x_4\}$ and $\{x_5, \ldots, x_{11}\}$. Let the result of this check be negative. Then one must assume that the cause of the failure is found in the subset $\{x_1, \ldots, x_4\}$, and the following check must be performed in this subset. With a positive result of the first test, one must consider the elements $\{x_1, \ldots, x_4\}$ to be in good working order, and the following check must be conducted in the subset $\{x_5, \ldots, x_{11}\}$.

Repeating this procedure for subdividing the set of the system's elements into nonintersecting subsets enables one to isolate only the one element about which the following is known:

— its output parameter deviates from the norm or the prescribed tolerances;
— the inputs of the isolated element are within the limits of the norm.

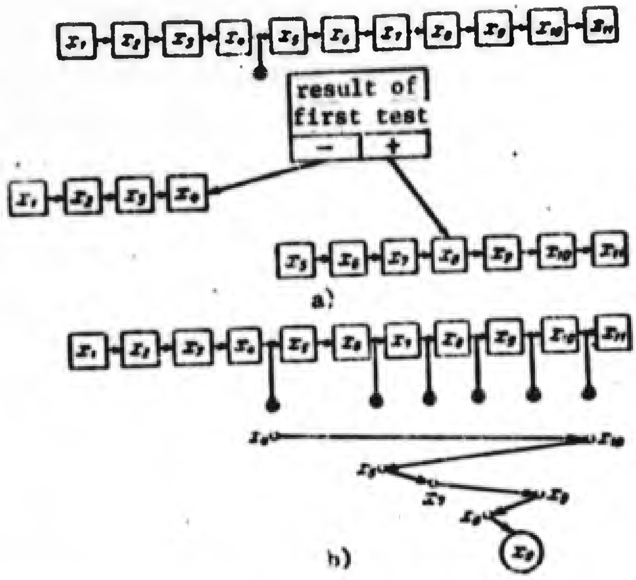


Figure 4.5. One of the variations of the sequence of tests for isolating an element with the index j when it fails (with a negative result, move to the left; with a positive result, move to the right).

As an example, one of the possible testing sequences with the breakdown of elements x_9 (Figure 4.5, b) may be written in the following way:

Index of the element being monitored	x_4	x_{10}	x_6	x_7	x_9	x_8
Result of the test (check)	yes	no	yes	yes	no	yes

The last test of element x_8 indicates that there is a signal at the input of element x_9, and the tests conducted previously show that the output parameter of element x_9 goes beyond the prescribed limits. As a result, it may be concluded that a failure of element x_9 has occurred.

The search sequence presented in Figure 4.5 as an example is not the only one possible. Other sequences leading to the same results may be found. Moreover, one may construct all the possible sequences applicable to the case when the cause of the failure is found in any of the system elements. The totality of the checking sequences, which have as their purpose the isolation of any failing element from the set of system elements, is called the malfunction search program.

The malfunction search program has two characteristics:

1. The number of checks k_j leading from the first check selected to the check giving the result. In designating the number of checks, j is the index of the defective element which is isolated when the given search sequence is carried out.

2. The total time (work involved) of the search sequence is

$$T_j=\sum_{k=1}^{k_j} t_k,$$

where t_k is the work of the test;
k is the serial number of the test in the search sequence counting from its beginning.

When $t_1 = t_2 = \ldots = t_{k_j} = t$ $T_j = k_j t$.

Above it was indicated that each test may have several mutually exclusive results, depending on the state of the system's part being monitored. This enables one to construct all the variations of the malfunction search sequence which, in totality, form a tree of logical possibilities (Figure 4.6 a). Each of the branches of such a tree leads to only one of the system elements. In Figure 4.6, the dashed line shows the path for isolating the element x_9 which was shown in Figure 4.5.

In Figure 4.6 b,c other variations of the logical possibility tree are shown. When using the various logical possibility trees,

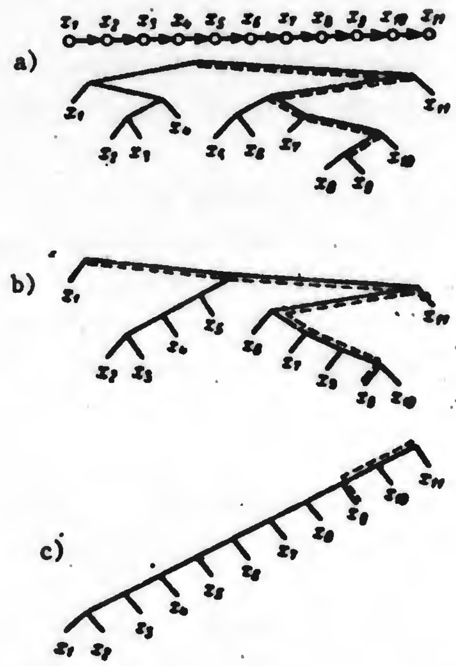


Figure 4.6. Certain variations of the tree of logical possibilities for the example presented in Figure 4.5.

the length of the search sequence k_j leading to the elements x_9 will differ. For the variations a, b and c in Figure 4.6, the lengths of the sequences equal 6, 7 and 3, respectively.

An analysis of Figure 4.6 shows that a change in the shape of the tree — more precisely, a decrease in the length of one of the search sequences — leads to an increase in the lengths of the other sequences. One may assume that certain average values of k_j exist which ensure an approximate or precise equivalence of all the search sequences in the logical possibility tree.

Thus, from the examples which have been cited it follows that a malfunction search program has many forks, and may be represented in the form of a logical possibility tree. The trees may have different shapes, as a result of which the number of tests in the search sequences will vary. Consequently, one is able to find a tree having a minimum number of tests in each of the sequences.

When constructing a search program, as will be shown below, a subdivision of the set of the system elements into nonintersecting subsets is performed. The number of elements in each of the subsets depends on the structure of the system. Therefore, it is necessary to investigate the effect of the system's structure on the compilation of the search program.

4.5. The Effect of a System's Structure on the Malfunction Search Program

The number of malfunction search programs sharply increases with an increase in the number of search results which may be found. As will be shown below, for a system consisting of four elements ($n = 4$) the number of search programs (in order to have two mutually exclusive test results) equals 5, and for a system where $n = 8$ the number of these programs reaches 429. Therefore, it is necessary to find from among all the possible search programs those programs which have minimum average expenditures of labor (or, in a particular case, a minimum average number of tests) in arriving at a single result of the search.

Let us investigate an arbitrary tree in which each test has $d = 2$, and let us evaluate the average length of its branches:

$$k_{cp} = \frac{1}{n} \sum_{j=1}^{n} k_j.$$

If the times involved in checking the various elements of the system being monitored are not the same, then an expression is found for the average work involved in the search sequence

$$T_{cp} = \frac{1}{n} \sum_{j=1}^{n} T_j = \frac{1}{n} \sum_{j=1}^{n} \sum_{k=1}^{k_j} t_k.$$

For a start, let us assume that $t_j = t$ ($j = 1, 2, \ldots, n$)

$$T_{cp} = \frac{t}{n} \sum_{j=1}^{n} k_j.$$

By changing the shape of a tree by various alterations in the tests of the elements, we find that binary trees have three main varieties:

1. Trees representing various variations of a consecutive train of tests. The number of these variations with an arbitrary ordering of the elements in the system being monitored is

$$R(n)_{\text{макс}} = 2^{n-2}.$$

2. Trees in which each test leads to a subdivision of the set of results into subsets with equal numbers of elements in subsets or numbers of elements which differ by one. The number of such trees may be determined in the following way:

$$R(n)_{\text{мин}} = \binom{2^k}{n-2^k}.$$

where k = $[\log_2 n]$ is the whole part of the quantity $\log_2 n$.

Such trees are variations of the subdivision of the test set into equally large subsets.

3. Trees which have various combinations of a consecutive train of tests and a subdivision into equally large subsets. They differ substantially in form from the trees of the first and second kinds.

Figure 4.7 shows all the variations of the trees for n = 2, 3, 4 and 5.

The total number of trees having d = 2 may not be larger than the value

$$R(n) = \sum_{l=1}^{n} R(l)R(n-l).$$

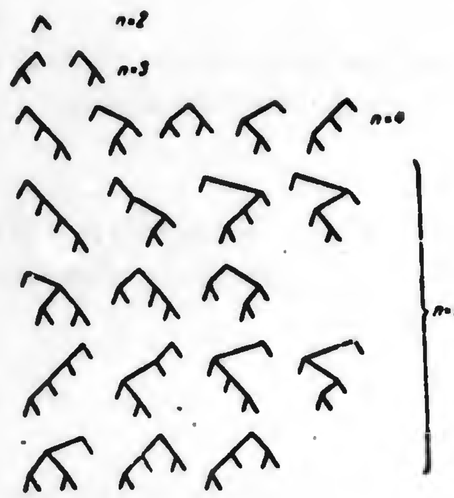


Figure 4.7. All the binary trees for n = 2, 3, 4 and 5.

where l is the index ranging from 1 to n;

$$R(1)=R(2)=1;\quad R(3)=2.$$

Some values of R(n) when n = 2 - 8 are presented in the following table:

$\tilde{n}$	2	3	4	5	6	7	8
$R(\tilde{n})$	1	2	5	14	42	132	429

Trees differ not only in their shape, but also in the average number of tests in the sequence.

Let us designate

$$\sum_{j=1}^{n} k_j = K_\Sigma(2, n-1).$$

This sum is the variable set in the expression for k_{av}. Calculation of $K_\Sigma(2, n-1)$ for trees of various shapes shows that for a consecutive train of tests

$$K_\Sigma(2, n-1)_{\text{макс}} = \binom{n-1}{2} + 2(n-1);$$

and for trees reflecting subdivision into equally large subsets

$$K_\Sigma(2, n-1)_{\text{мин}} = \sum_{i=1}^{n-1} [\log_2 i] + 2(n-1)$$

and for trees of the third kind

$$K_\Sigma(2, n-1)_{\text{мин}} < K'_\Sigma(2, n-1) < K_\Sigma(2, n-1)_{\text{макс}}.$$

Thus, for each finite value of n, there is some finite set M of binary ($d = 2$) trees of logical possibilities. To each tree of M there corresponds a symmetric tree in the same set. Certain trees have internal symmetry, and they do not have paired trees in M (Figure 4.8).

Let us arrange the trees of set M on a single horizontal axis in such a way that they follow after one another in the order of decreasing $K_\Sigma(2, n-1)$. When this is done, symmetric trees are not repeated. In this way trees of the first kind having $K_\Sigma(2, n-1)_{max}$ appear at the origin of the coordinates, then trees of the second kind for which $K_\Sigma(2, n-1) < K_\Sigma(2, n-1)_{max}$, and after this trees of the third kind [$K_\Sigma(2, n-1)_{min}$]. Such an arrangement for n = 5 is shown in Figure 4.9.

Let us add to these trees the trees which are symmetrical with them in order of increasing $K_\Sigma(2, n-1)$ (Figure 4.10). By joining the graphs shown in Figure 4.9 and 4.10, we obtain a step curve of the chain of k_{av} as a function of the shape and symmetry of the trees of M (Figure 4.11). This step curve always has a minimum corresponding to trees of a minimum shape. From this it follows that, by making a rational selection of the test points, one is able to construct for a system a tree having as small as possible

$$k_{\text{ср}_{\text{мин}}} = \frac{1}{n} K_\Sigma(2, n-1)_{\text{мин}}.$$

A tree of such a kind appears when the tests subdivide a system into two nonintersecting subsets. One of the subsets contains the elements preceding the element being monitored, and the other subset contains all those elements which are not in the first subset. Each of the subsets contains approximately an equal number of elements.

When the structure of a system does not allow one to form a tree of minimum shape, the subdivision of the system must be carried out in such a way that the numbers of elements in each of the subsets are as close as possible to one another.

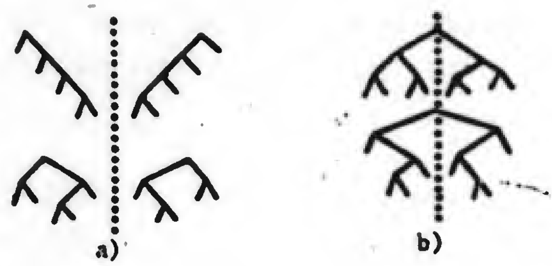
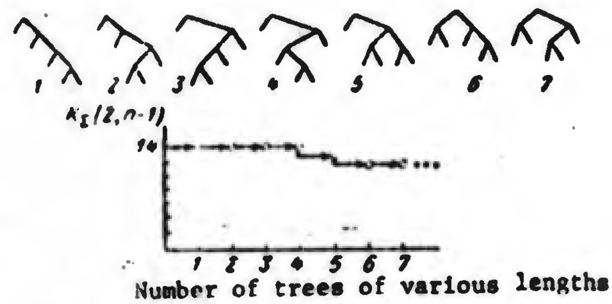
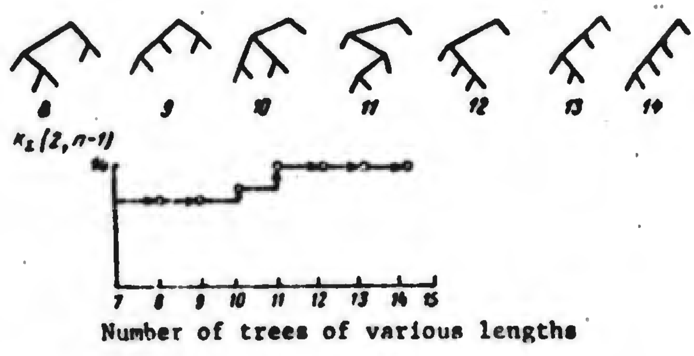


Figure 4.8. The symmetry of binary trees.



Number of trees of various length

Figure 4.9. The decrease of the total length of the branches of a tree with a change of its shape (n = 5).



Number of trees of various length

Figure 4.10. The increase of the total length of the branches of a tree with a change of its shape (n = 5).

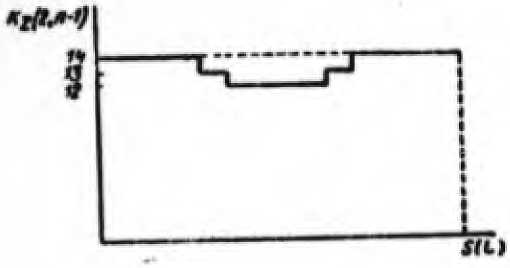


Figure 4.11. A step curve of the change of k_{av} as a function of the shape and symmetry of the trees.

Let us present several examples of the subdivision of various structures in accordance with the requirement advanced. Figure 4.12 shows a graph containing a Hamiltonian path. The subdivision of this graph leads to the formation of a tree of minimum shape.

Figure 4.13 shows a graph in which the number of entrances is equal (n - 1), and it has a single exit vertex. In this case, each subdivision separates only one vertex from the set of vertexes of the graph, and the possibility of shortening the test train does not exist. As a result, a tree of maximum shape is formed. Figure 4.14 shows a graph in which there is no Hamiltonian path. The subdivision of the graph of Figure 4.14 leads to the formation of a tree occupying an intermediate position between trees of minimum and maximum shapes. This tree is a tree of the minimum possible shape for the graph in Figure 4.14.

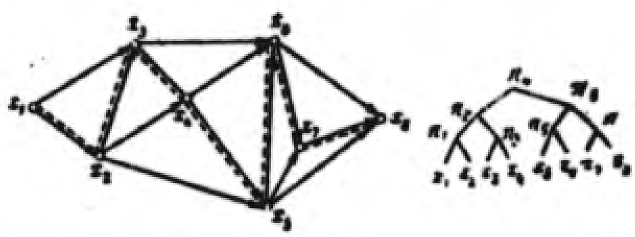


Figure 4.12. Subdivision of a graph containing a Hamiltonian path.

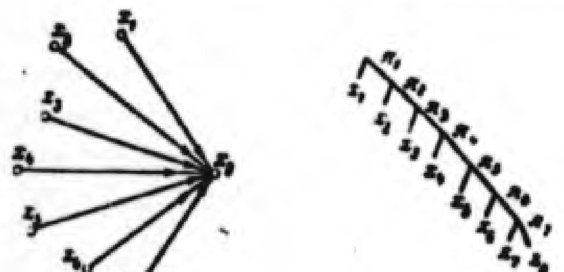


Figure 4.13. Subdivision of a graph having (n - 1) entrance vertexes and one exit vertex.

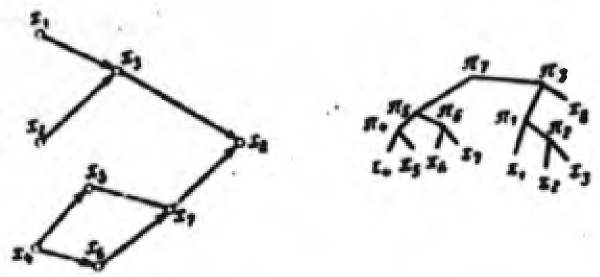


Figure 4.14. Subdivision of a graph not having a Hamiltonian path.

4.6. Failures Caused by Breaks and Overloads

Sudden failures of electronic systems may be divided into two kinds:

1. Breaks of the functional circuits. A characteristic feature of failures of this kind is that they do not influence the working order of the power supply sources and the protective circuits.

2. Overloads, which cause an increase in the power consumption by the system and lead to a response of the protective circuits. Short circuits are referred to as the limiting case of overloads.

Let us assume that the sources of failures are the system elements, and the connections between them are perfect.

Depending on the character of the element's failure, two mutually exclusive results of the search may occur: a break in the element or an overload. But if this is so, then the malfunction search program must be constructed in such a way that failures of the "break" kind and of the "overload" kind are always unequivocally distinguished.

A break of a functional circuit shows up in the form of the disappearance or distortion of the output signals of the system or element.

An overload has somewhat different symptoms: besides the distortion or disappearance of the output signals of certain elements, a response of the protective circuits takes place, and part of the system (or the system as a whole) is disconnected from the power supply source. It must be pointed out that disconnection with overload occurs only when the protective circuits are correctly adjusted. In the opposite case, destruction of the individual elements is possible.

Thus, a system may be represented in the form of two geometric models — oriented graphs — corresponding to the functional circuits of the system for the cases of break and overload. An example of the interrelationship of the elements for breaks and overloads in a simple circuit is shown in Figure 4.15. The oriented graph of this circuit, taking into account the mutual influences of the elements with breaks, is shown in Figure 4.16. When it is necessary to consider the effect of overloads, it is necessary to construct an oriented graph for the case of overloads. In this case, the directions in which the elements are connected must be changed to the opposite directions (Figure 4.17). However, the separate use of two graphs is inadvisable from the point of view of shaping a tree. Therefore, it is better to join them into one graph (Figure 4.18) in which the same elements precede the protective circuit for the overloads and follow the protective circuit for the breaks. With a break or a short of the protective circuit, the power supply is disconnected, as a consequence of which it is not repeated twice in the unified graph.

The probabilities of breaks and short circuits of the elements in general are different. We will show below that consideration of the failure probabilities affects the malfunction search procedure. Therefore, it makes sense to discuss the question of the failure probabilities.

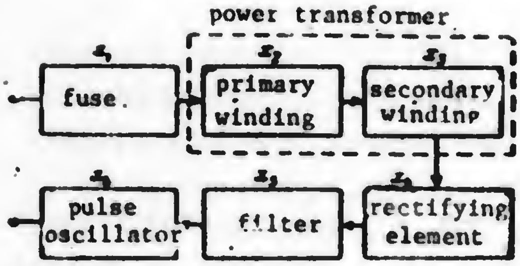


Figure 4.15. A simple circuit illustrating the interrelationship of elements with respect to breaks and overloads.

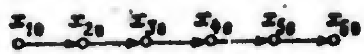


Figure 4.16. A graph of the interrelationship of the system elements in Figure 4.15 in terms of breaks.

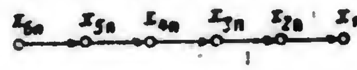


Figure 4.17. A graph of the interrelationship of the system elements in Figure 4.15 with respect to overloads.

Figure 4.18. A unified graph for the system of Figure 4.15.

In Table 4.2 statistical data characterizing the rates of breaks λ_b and of short circuits λ_s in various parts of electronic circuits are presented.

The total rate of failures is $\lambda \cong \lambda_b + \lambda_s$. Consequently, we can approximately assume that in an element of any complexity it is possible to determine λ_b and λ_s, that is

$$\lambda = \sum_{j=1}^{r} \lambda_j = \sum_{j=1}^{r} \lambda_{\text{о}j} + \sum_{j=1}^{r} \lambda_{\text{з}j} = \lambda_{\text{о}} + \lambda_{\text{з}}.$$

where r is the number of the simplest parts in the element.

Knowing λ_b and λ_s enables one to calculate the fraction of probabilities $P(A_j/X)$ for each element as applied to short circuits and breaks by using the following simple relationships:

TABLE 4.12. DISTRIBUTION OF THE FAILURES OF ELECTRONIC CIRCUIT PARTS CAUSED BY BREAKS AND OVERLOADS IN %

Parts of a circuit	Kind of failure	
	Break	Overload
Tubes	80	20
Resistors	44	56
Capacitors	23	77
Inductance coils	82	18
Semiconductor devices	32	68
Transformers:		
Power	31	69
Pulse	75	25
High-voltage	-	100

$$P(A_j/X)_b = \frac{\lambda_b}{\lambda} P(A_j/X);$$

$$P(A_j/X)_s = \frac{\lambda_s}{\lambda} P(A_j/X),$$

in connection with which

$$P(A_j/X)_b + P(A_j/X)_s = P(A_j/X).$$

Thus the relationship

$$\sum_{j=1}^{n} P(A_j/X) = 1$$

may be written in the form

$$\sum_{j=1}^{n} P(A_j/X)_b + \sum_{j=1}^{n} P(A_j/X)_s = 1.$$

The possibility of dividing the relative failure probabilities into $P(A_j/X)_b$ and $P(A_j/X)_s$ enables one to compare these probabilities with the two mutually exclusive states of the element: a break and a short circuit.

Let there be a set with a definite form of ordered elements of a certain system X_n. The system contains γ protective circuits against overloading. The power supply for the system's elements located in the subsets $E_1, E_2, \ldots, E_\gamma$ is provided through each of the corresponding protective circuits.

Each i^{th} protective circuit will agree with its subsets E_i ($i = 1, 2, \ldots, \gamma$). These subsets are connected by the following relationships:

$$\bigcup_{1\leq i\leq\gamma} E_i = X_{nb} \text{ and } \sum_{i=1}^{\gamma} n_i = n.$$

For each subset E_i, let us construct a graph of the short circuits of the elements of E_i in the i^{th} protective circuit, and let us combine these graphs with the graph of the set X_{nb} constructed for the breaks.

As a result of this operation, a new finite set M is formed whose elements will be all the various states corresponding to the short circuits and breaks in the elements of set X_n. The number of elements in set M depends on n and the number of protective circuits γ.

Let us eliminate from each E_i the single element of the set which corresponds to the protective circuit. Then we may write

$$M = X_{nb} \cup \left(\bigcup_{i\leq\gamma} E_i\right).$$

and the number of elements of this set equals

$$n + \sum_{i=1}^{\gamma} (n_i - 1) = 2n - \gamma.$$

The set M is shown in Figure 4.19. Knowing the structure of set M and the failure probabilities of its elements, we can construct the probability additive measure or failures probability distribution in set M.

Let us discuss in somewhat more detail the symptoms accompanying the appearance of a break or a short circuit. In order to do this, let us investigate an arbitrary set E_i in which the elements are ordered in relation to their effect in the i^{th} protective circuit. Let us select from the set E_i a certain element $x_{\xi_i} \in E_i$, such that one part of the elements of E_i precede it, and the other part of the elements follow it.

Let us assume that the element x_{ξ_i} is the cause of the overload. Then, when searching for this element in the set E_i disconnecting all the elements preceding x_{ξ_i}, does not lead to the disappearance of the overload. The overload disappears when the element x_{ξ_i} is disconnected from all the elements following it. This is equivalent to a subdivision of the set E_i into a subset of elements preceding x_{ξ_i} and a subset of elements which do not precede x_{ξ_i}. The phenomenon described, in this sense, is different from the symptoms of a break in the element (Table 4.3).

TABLE 4.3. SYMPTOMS OF A BREAK AND AN OVERLOAD IN AN ELEMENT

Break	Overload
The input signals are normal	An increased power consumption which turns off the protective circuit
The output signals are distorted or absent	The output signals disappear

From this it follows that in both cases it is necessary to carry out two tests in order to be certain of an element's failure. These tests have a different character for breaks and short circuits (Table 4.4).

TABLE 4.4. METHODS FOR ISOLATING A BREAK AND AN OVERLOAD

Kind of failure	Monitor the inputs	Monitor the output
Break	Connect an external measuring instrument or monitor the readings of the built-in instruments	Connect an external measuring instrument or check the readings of the built-in instruments
Overload	Disconnect the element being monitored from all the power circuits and monitor the response of the protective circuit	Disconnect the element being monitored from all the elements following it and monitor the response of the protective circuit

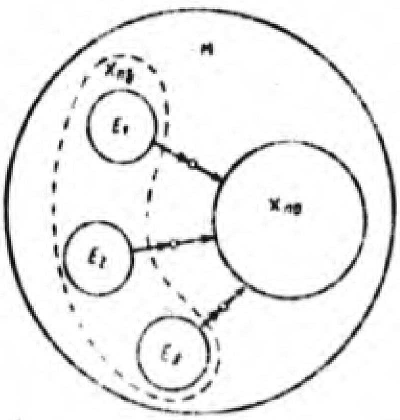


Figure 4.19. An overall view of set M.

The subsets are designated E_1, E_2, ..., E_γ, and the number of elements in them equals, respectively, n_1, n_2, ..., n_γ.

However, these differences are unimportant, since any test may be reduced to a formal calculation of the number of its mutually exclusive results.

Thus, in order to construct a search program which considers the possibility of breaks and overloads, it is necessary first of all to construct an oriented graph of set M. In addition, it is necessary to calculate the probabilities of the search results and the work involved in the tests of the elements in M when they are monitored in the presence of a break or a short circuit.

As a more complicated example of the construction of set M for an actual system, let us investigate the antenna rotation drive of the marine navigation radar "Donets-2" (Figure 4.20). For simplicity, we will assume that the drive is a completely independent system. The functional circuit of this drive, made at the level of the elements of unit I-9 which are independent in the sense of failures, is shown in Figure 4.21. All the contact plates and plug joints which frequently are the causes of breaks and short circuits are included as elements in this circuit. When this is done $n = 20$, in connection with which two of them (x_{15} and x_{34}) are protective circuits ($\gamma = 2$) (the fuse Pr-6 and the automatic protective device AP-50, respectively). The elements preceding x_{15} and x_{34} form the sets E_1 and E_2 in which the elements are interconnected with respect to overloads.

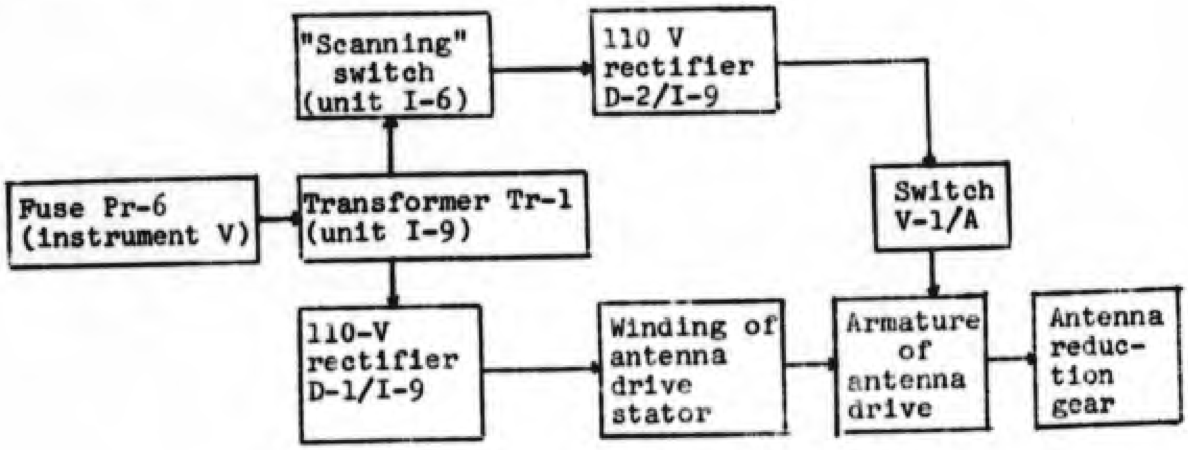


Figure 4.20. The functional circuit of the antenna rotation drive of the radar "Donets-2".

By finding the graphs of sets E_1 and E_2 and combining them with the drive's functional circuit, we obtain a diagram which corresponds to set M. The oriented graph of the set with the number of elements $2n - \gamma = 38$ is shown in Figure 4.21. On the same figure the precedence indexes are indicated near each vertex of the graph.

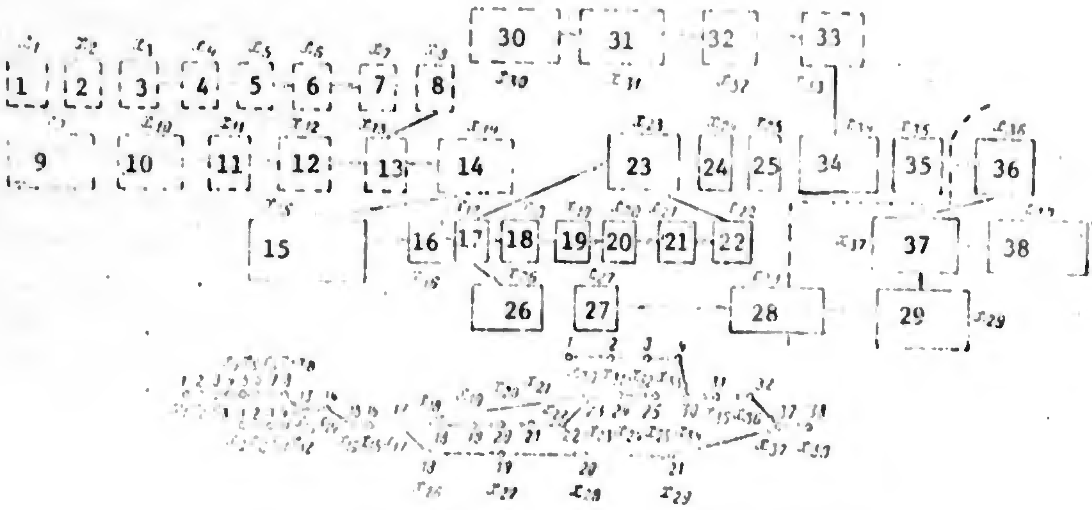


Figure 4.21. Expanded functional circuit and oriented graph of the antenna rotation drive of the radar "Donets-2".

1 - I-10 P-5; 2 - S-h-8 14 (plug 8, contact 14); 3 - D-2 I-9; 4 - Sh-8 16; 5 - Sh-5 19; 6 - I-6 B-7; 7 - Sh-5 18; 8 - Sh-8 10; 9 - tator of antenna motor; 10 - plate P-1/A; 11 - Sh-8 35; 12 - recti D-2 I-9 (rectifier); 13 - Tr-1 I-9 (transformer); 14 - plate P-7/I-10; 15 - Fuse Pr-6 (instrument B); 16 - P-7 I-10; 17 - Tr-1 I-9; 18 - Sh-8 10; 19 - Sh-5 16; 20 - N-6 B-7 21 - Sh-5 19; 22 - Sh-8 16; 23 - recti D-2/I-9; 24 - Sh-8 14; 25 - P-5 I-9; 26 - recti D-1/I-9; 27 - Sh-8 3; 28 - antenna P-1, 74, 66; 29 - stator of antenna motor; 30 - antenna reduction gear; 31 - armature of antenna motor; 32 - switch B-1/A; 33 - I-10 P-5, 7, 8; 34 - protective circuit AP-50; 35 - I-10 P-5, 6, 7; 36 - switch B-1/A; 37 - armature of antenna motor; 38 - antenna reduction gear.

4.7. The Construction of a Search Program for a Specific System

Let us investigate the construction of a search program as applicable to the antenna rotation drive of the radar "Donets-2" (Figure 4.21).

Let us assume that the failure probability distribution is uniform and the work of monitoring the various elements of this drive is the same. Let us suppose it is discovered that the radar's antenna is not rotating, and the other elements of the radar are operating normally. It is necessary to construct a search program such that the average length of a branch of the tree would be a minimum. In order to do this, let us find the elements whose precedence index equals $l=\frac{2n-\gamma}{2}=19$. Then the set of the drive elements will be divided by the tests into two subsets with equal numbers of elements. Either contact 16 of plug Sh-5 or contact 3 of plug Sh-8 may be such an element. Let us select contact 16 for monitoring. As a result of checking for the presence of an alternating voltage between this contact and the casing, the set M is divided into two subsets M_1 and M_2: set M_1 corresponds to a negative result for the check π_{19}, and set M_2 — to a positive result (Figure 4.22, a and b). Let us number the elements in each of these sets separately. In order to subdivide these subsets, it is necessary to find the indexes equal to or as close as possible to $|M_1|/2=9$ and $|M_2|/2=9$. For M_1, index 8 is closest to 9. Index 8 corresponds to contact 10 of plug Sh-8. For M_2, the automatic protective device AP-50 (l_{22} = 11) satisfies this requirement. Repeating the subdivision procedure, let us find in M_1 and M_2 such subsets as will contain approximately equal numbers of elements.

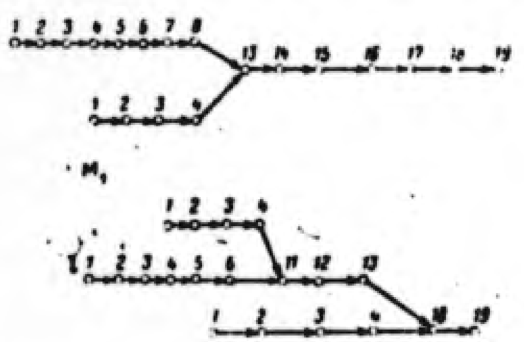


Figure 4.22. The subdivision of the oriented graph into two subgraphs for the antenna rotation drive of the radar "Donets-2" with the first testing.

Continuing the subdividing leads finally to the isolation of each element and to the completion of the search. The tree reflecting the subdivision sequence is shown in Figure 4.23. The number of tests in it, based on arriving at a single result, equals

$$k_{cp}=\frac{1}{38}\sum_{i=1}^{38}k_i=\frac{202}{38}\approx 5.31,$$

which corresponds to the quantity

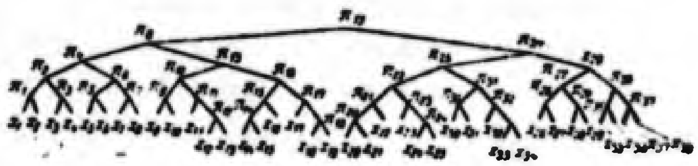


Figure 4.23. The search tree for the antenna rotation drive of the radar "Donets-2" with the condition that $P(A_j/X)$ and t_j are equivalent.

$$K_{\Sigma}(2, 2n - \gamma - 1)_{\text{мин}},$$

where

$$n = 20; \quad \gamma = 2,$$

that is

$$K_{\Sigma}(2,37)_{\text{мин}} = \sum_{i=1}^{37} [\log_2 i] + 74 = 202.$$

This means that the tree which has been constructed is one of the variations of the subdivision into equally large subsets. Each decrease of the length, even if by only one sequence of tests, gives rise to an increase of k_{av} in comparison with $k_{av_{min}} = 5.31$.

Let us compare the resulting variation of the search program with the variation of a consecutive train of tests which is recommended in the majority of handbooks for repairing equipment. In this case

$$K_{\Sigma}(2,37)_{\text{посл}} = \left(\frac{37}{2}\right) + 74 = 740.$$

The advantage of the search procedure selected by us in comparison with the train of tests is

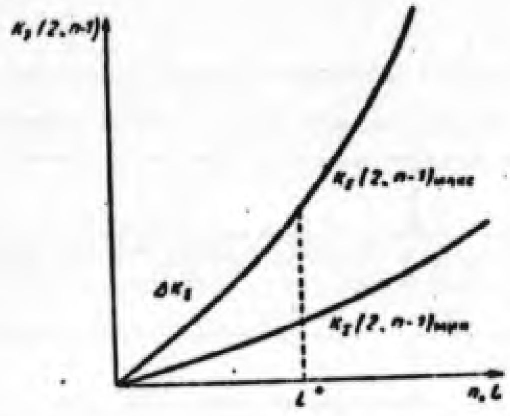


Figure 4.24. The increase of the values of $K_\Sigma(2, n - 1)_{max}$ and $K_\Sigma(2, n - 1)_{min}$.

$$\eta = \frac{K_\Sigma(2,37)_{\text{макс}}}{K_\Sigma(2,37)_{\text{мин}}} \approx 3{,}66.$$

An increase of n leads to an increase of η, since the difference between $K_\Sigma(2, n - 1)_{max}$ and $K_\Sigma(2, n - 1)_{min}$ is an increasing function (Figure 4.24). This example graphically illustrates the operation of the algorithm shown in Figure 4.25 for constructing a malfunction search program for the case when the distributions of $P(A_j/X)$ and t_j are uniform.

4.8. Construction of a Malfunction Search Program with Consideration of the Failure Probabilities

From the preceding section, it follows that it is possible to minimize the arithmetic mean of the number of tests in a malfunction search sequence. However, the method investigated above may not be sufficiently effective when the elements of the system being monitored have different failure probabilities. It happens, in the practice of operating radar stations, that it is difficult to divide a system into elements which are equally reliable and independent in the sense of failures. Therefore, a malfunction search program constructed only with consideration of the system's structural characteristics may

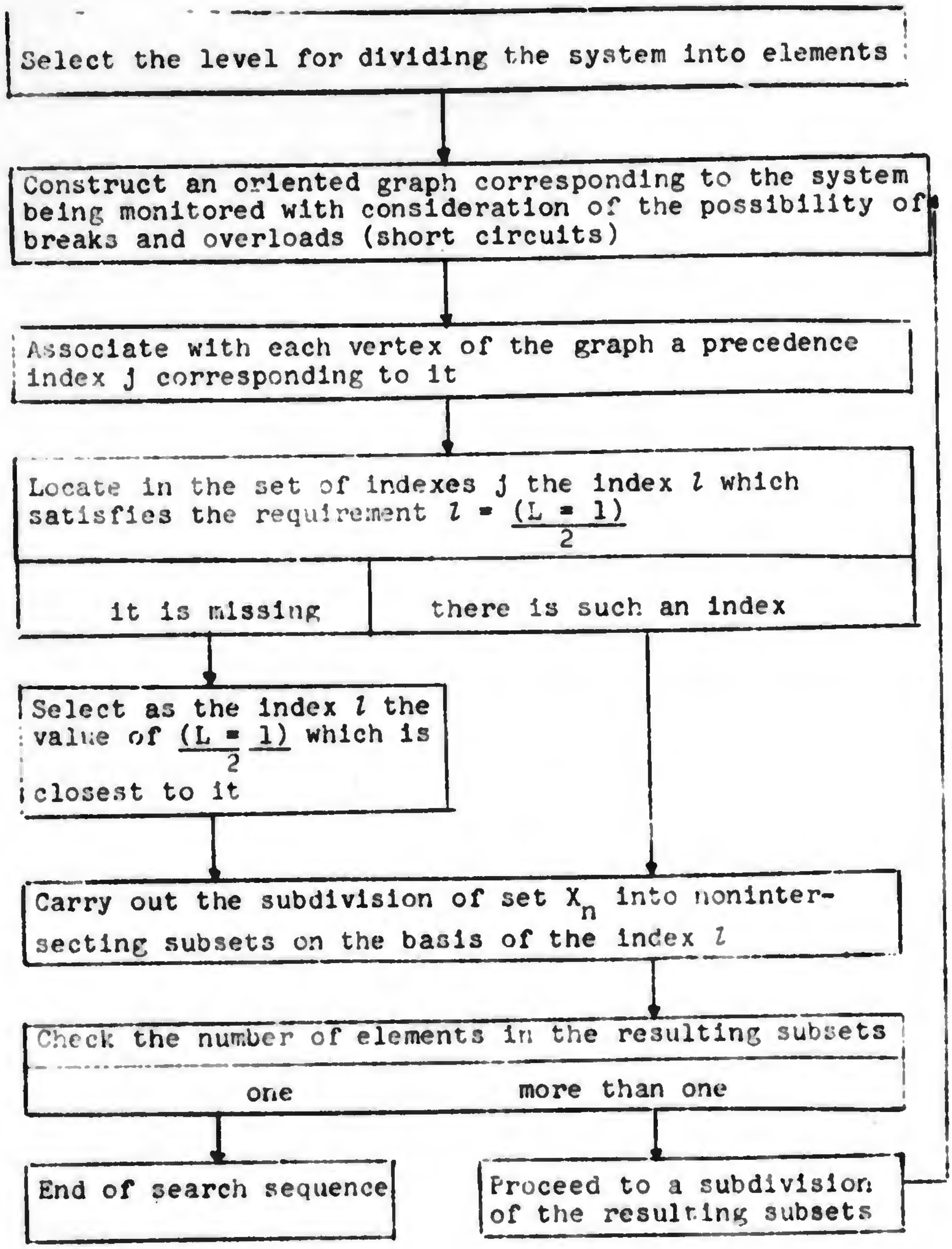


Figure 4.25. The block diagram of the algorithm for constructing a malfunction search program based on a structural criterion.

turn out to be ineffective. Actually, when there are failures, the test sequences leading to the least reliable elements will be repeated more frequently, and the sequences leading to the more reliable elements will be repeated less frequently, but the number of tests in them will be equal. A better search program will be a program in which the sequence length is shorter, the more frequently the element to which this sequence leads breaks down. Let us designate such a length by $k_j(P)$. Let us write in a general form its dependence on $P(A_j/X)$:

$$k_j(P)=f\left(\frac{1}{P(A_j/X)}\right), \quad j=1, 2, \ldots, n.$$

Let us find the total probability that a failure would occur as a consequence of a breakdown of an element whose precedence index does not exceed j, that is

$$\text{Вер}\{i\leq j\}=\sum_{i\leq j} P(A_i/X).$$

This distribution may be constructed for the set of elements of system X_n.

It is possible to find the values of $k_j(P)$ as a function of $P(A_j/X)$ on the basis of coding theory. In particular, when constructing optimum code sequences, the following relationship is valid [64]:

$$k_j(P)\approx-\log_2 P(A_j/X).$$

The mathematical expectation of the length of the coded sequence will be a minimum:

$$\text{MO}\, k_j(P)=\sum_{j=1}^{n} P(A_j/X)\, k_j(P)=k_0(P)_{\min}.$$

The logical possibility trees of a complete code system and a malfunction search program match; therefore, the last expression is also valid for a malfunction search program.

In order to achieve the values of MO $k_j(P)$, it is necessary to carry out subdivisions of the set of the system's elements into subsets having equal total failure probabilities or total failure probabilities, which are as close as possible to one another. This means that one should take the index of the median of the distribution of $P(A_j/X)$ as the subdivision index l, that is, one must satisfy the condition

$$\sum_{j \leq l} P(A_j/X) \approx \sum_{l < j} P(A_j/X),$$

where l is the index of the median of the distribution of $P(A_j/X)$.

When the index of the median l is selected as the monitoring point, the test being conducted gives the maximum information concerning the system's state.

When there is no Hamiltonian path in the system being monitored, it may turn out that the integral function of the distribution gives several indexes whose total probabilities are close to 0.5. Then one selects from among these values the value closest to 0.5, or the various subdivisions are sorted for the purpose of selecting the tree having the least value of $k_B(P)$.

Thus the shape of the distribution of $P(A_j/X)$ plays an important role when forming the malfunction search program and may be considered together with the system's structure.

The block diagram of an algorithm for the construction of a malfunction search program considering the failure probabilities is presented in Figure 4.26.

With a uniform distribution of $P(A_j/X)$ only the system's structure will be considered, and the block diagram of Figure 4.26 leads to the formation of a tree of minimum shape or to a shape which is as close as possible to the minimum. When this is done, the

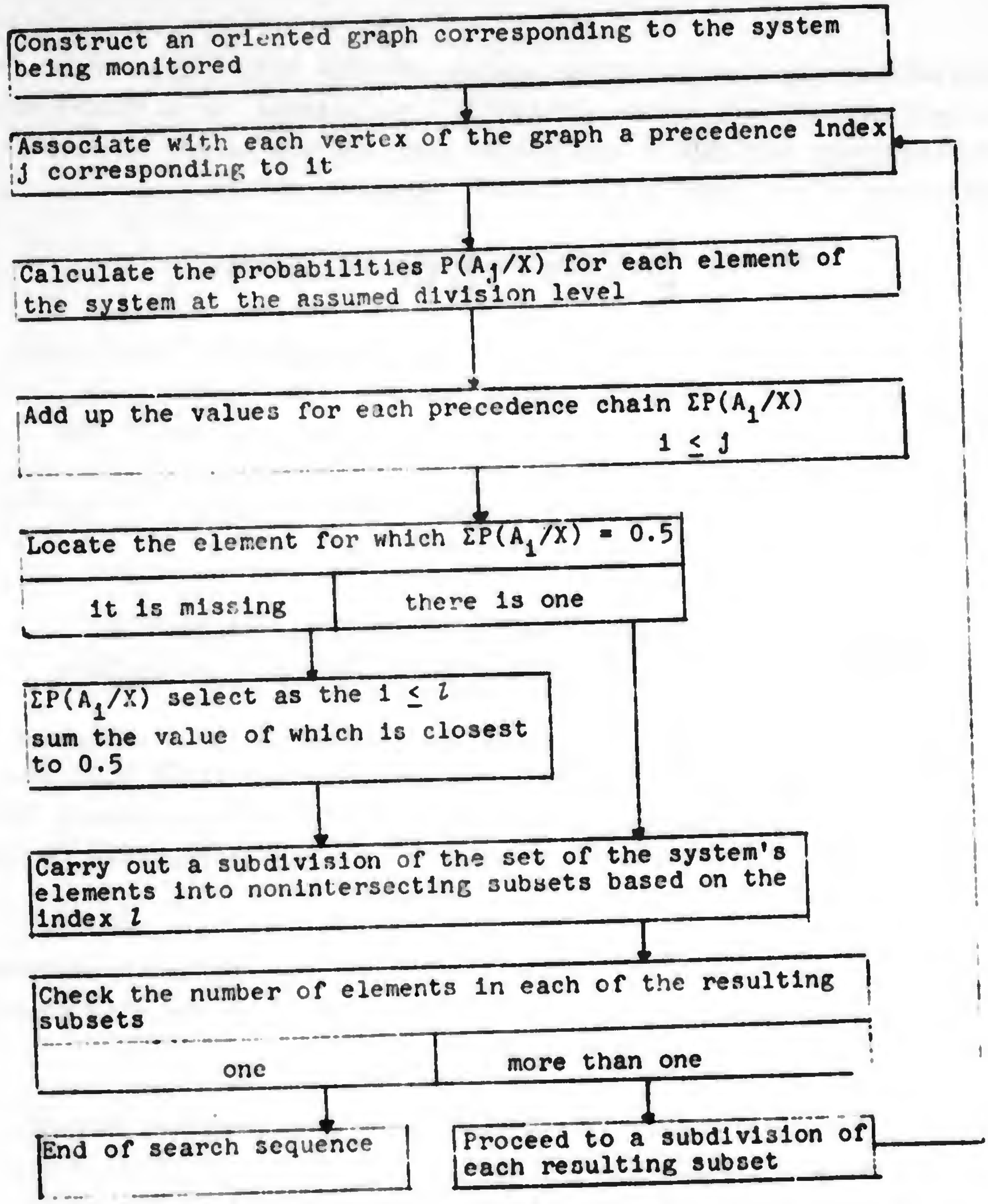


Figure 4.26. The block diagram of the algorithm for constructing a malfunction search program using the probability criterion.

mathematical expectation of the search sequence length degenerates to the arithmetical mean number of tests. For example, in a system whose oriented graph contains a Hamiltonian path and the number of elements multiplied by two, when $P(A_j/X) = 1/n$ holds, we obtain

$$MO\,k_j(P)=\sum_{j=1}^{n}(A_j/X)\,k_j(P)=\sum_{j=1}^{n}\frac{1}{n}\left(-\log_2\frac{1}{n}\right)=\log_2 n.$$

From this expression it follows that the number of tests which must be conducted in order to detect the defective elements with the selected level of division equals the log to the base two of the number of elements of the system. Consequently, with a uniform distribution of $P(A_j/X)$, the construction of the program may be carried out on the basis of the structural criterion.

4.9. Construction of a Malfunction Search Program with Consideration of the Work Involved in Tests of the Elements

When searching for malfunctions in a number of cases which are important for practical purposes, a reduction of the total work involved in the test sequence is required. With this requirement, the search program may differ from the search programs constructed on the basis of structure and probability criteria.

Let us investigate the possible methods for reducing the duration (work) of the malfunction search procedure, first for the case when the distribution of $P(A_j/X)$ is uniform.

An experienced specialist in the repair of electronic systems, when searching for malfunctions, usually selects as the first tests the tests having the minimum work in order to isolate in the least possible time, to the extent possible, a more limited section of the system. The malfunction search sequence is such that the work of each succeeding test is greater than or equal to the work of the preceding test. When this is done, a situation is created in which the greater the work of the tests, the smaller the section of the system in which it is carried out. Figure 4.27 shows the general order for the subdivision of a system using tests with minimum work.

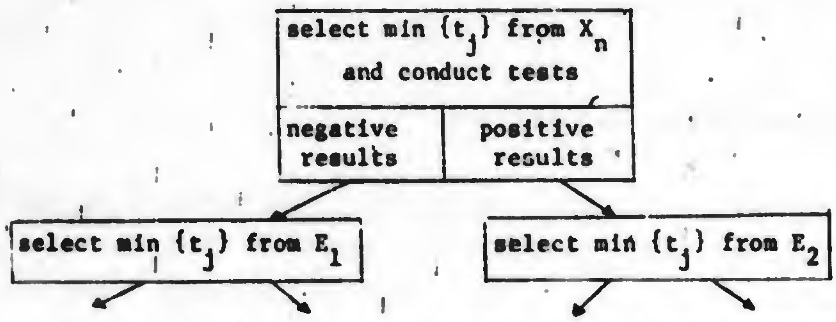


Figure 4.27. The order of subdividing a system by checks having minimum laboriousness.

In practice, one may meet a situation where the monitoring point having the minimum work is found at the input or output of the system. Then the search program leads to a sequential testing series. Moreover, the case is possible where the distribution of the work of the tests over the set of the system's elements is uniform, and the selection of min $\{t_j\}$ loses its meaning.

Therefore, it is necessary to develop a general method for approaching the construction of a malfunction search program with consideration of the work involved in the tests.

For this purpose, let us note that the likelihood of selecting a test from a certain number of tests will be larger, the smaller its work. Thus there is an inverse relationship between this likelihood and the work involved in tests of these elements. Let us designate this "likelihood" by the symbol τ_j ($j = 1, 2, \ldots, n$).

If the distribution of the failure probabilities of the system's elements is uniform, then the likelihood of selecting τ_j depends only on the distribution of the work t_j. The simplest relationship which shows the connection between τ_j and t_j is the expression

$$\tau_j = \alpha \frac{1}{t_j},$$

where α is a normalizing coefficient.

The coefficient α in this expression is determined on the basis of the fact that, with the selection of one test from the set containing n possible tests, one result of n is realized. But if this is so, then the likelihoods τ_j must satisfy the requirement

$$\sum_{j=1}^{n} \tau_j = 1.$$

Hence the coefficient

$$\alpha = \frac{1}{\sum_{j=1}^{n} \frac{1}{t_j}},$$

and the expression for τ_j may be written in the following form

$$\tau_j = \frac{1}{t_j \sum_{j=1}^{n} \frac{1}{t_j}}.$$

Thus, the likelihoods of the preferred selection of the j^{th} test from the set X_n are distributed over this set. The shape of the distribution of τ_j is such that the work maximum corresponds to the minimum of the likelihood τ_j, and vice versa. The distribution of τ_j, based on its quantitative characteristics, is no different from the distribution of $P(A_1/X)$, and it may be used for forming the search program. Let us turn our attention to the fact that, when the set X_n contains several elements having the same work equal to min $\{t_j\}$, the selection of one of them is difficult, if one does not take into account the locations of these minimums on the graph of the system. But consideration of the locations of the minimums of t_j leads to the necessity of taking into account the system's structure.

Sorting out the various distributions of τ_j and comparing all the permissible logical possibility trees with them enables one to draw the conclusion that, when there is a nonuniform distribution of τ_j, the set X_n should be divided into subsets at the point having the distribution median index of τ_j. When this is done, the lengths of the search sequence are

$$k_j(\tau) \approx -\log_2 \tau_j.$$

The mathematical expectation of the total duration or work of the search sequence may be calculated on the basis of the criterion of τ_j:

$$MOT_j(\tau) = T_B(\tau) = \sum_{j=1}^{n} \tau_j T_j(\tau) = \sum_{j=1}^{n} \tau_j \sum_{k=1}^{k_j(\tau)} t_k.$$

The method of constructing a program having a laboriousness $T_B(\tau)$ is presented in the form of a block diagram in Figure 4.28. When using this method, one should keep in mind that:

1. The points of dividing the system by tests are grouped in the regions of minimum work.

2. The individual peaks on the distribution of τ_j (the work maximums) do not affect the location of the division points.

3. When there are several minimums in the distribution, the division point is located between the indexes of the minimums, approximating the least work.

The use of this method is most effective when there is a considerable difference in the work of the tests. With a uniform distribution of the work, the distribution of τ_j also is uniform, and the average work of the malfunction search procedure equals

$$T_B(\tau) = \frac{t}{n} \sum_{j=1}^{n} k_j = T_{cp}(\tau).$$

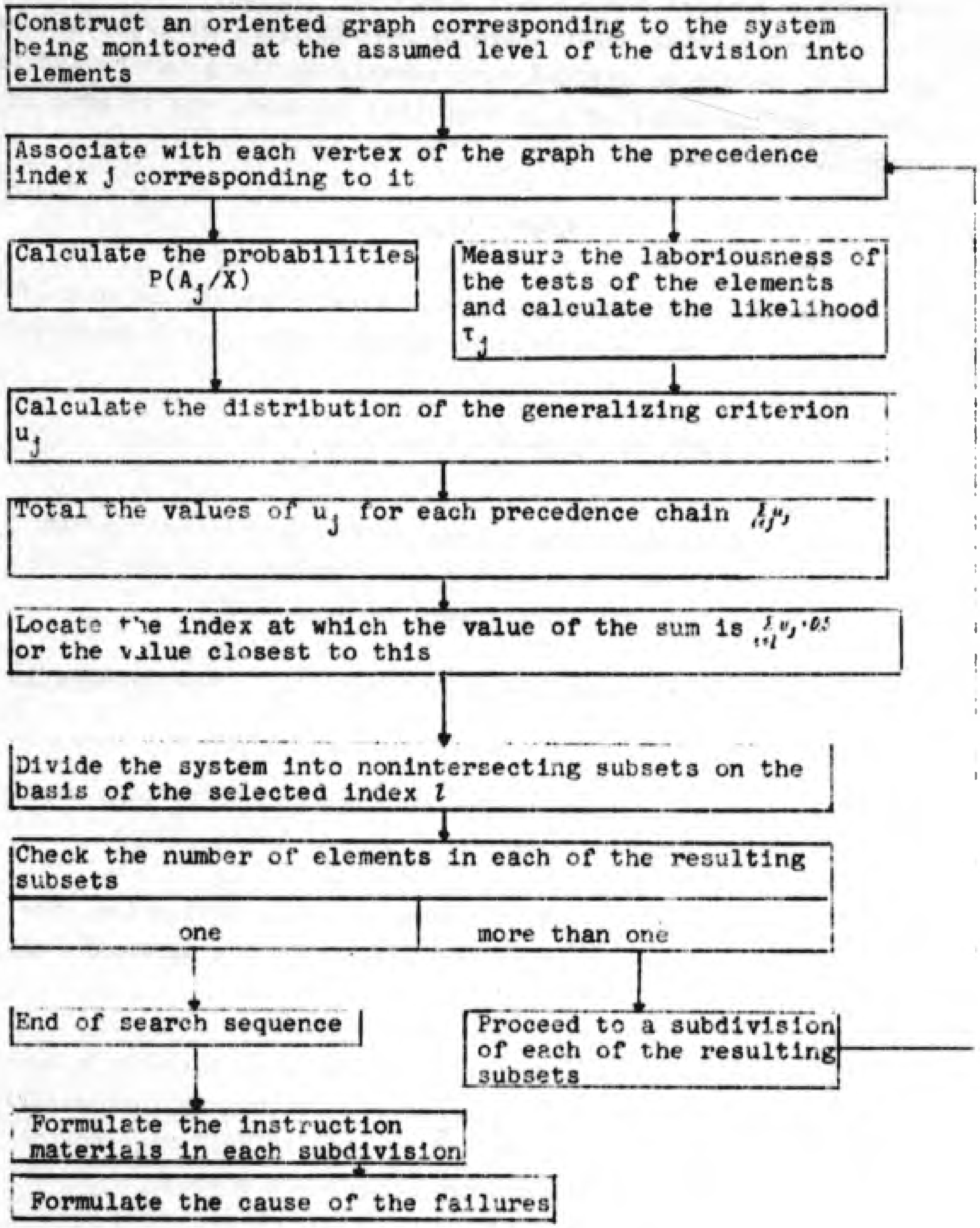


Figure 4.28. Block diagram of the algorithm for constructing a malfunction search program using the composite criterion.

where $t_k = t_j = t$ and $k_j \simeq [\log_2 n]$.

This is legitimate for the case when the distribution of $P(A_j/X)$ is uniform, and the system's structure allows it to be divided into equally large subsets of elements.

Consequently, with a uniform distribtuion of τ_j the malfunction search program should be constructed only with consideration of the system's structure (see Section 4.7).

4.10. Construction of a Malfunction Search Program with Consideration of the Failure Probabilities and the Work Involved in Tests of the Elements

Let us proceed to a more general case, when uniform distributions of $P(A_j/X)$ and τ_j occur simultaneously in the system. When designing radar, the distinctive operational features of the circuit are usually known beforehand. At the same time, new models of stations have different circuit solutions from the old models, and sometimes also different parts. Therefore, the designer, based on his experience, to a considerable extent arbitrarily selects the monitoring points. In connection with this, the question as to whether the selection of the monitoring jacks and the built-in monitoring instruments was correctly made and whether there is a sufficient number of them requires a more precise definition. The selection may be made so that the elements which rarely break down will be most completely encompassed by the monitoring facilities, and the least reliable elements in general will not be monitored. From this it follows that the dependence between the distributions of $P(A_j/X)$ and τ_j is fairly weak, and when constructing the malfunction search programs these distributions may be considered independent.

In this case, the criterion for selecting the test point may be the conditional probability of the selection

$$W_j=\tau_j P(A_j/X),$$

whose distribution over the totality of the system's elements simultaneously takes into account the laboriousness of the tests and the failure probabilities of the elements. If one normalizes the distribution of W_j, considering that the events comprising the selection of one test from n possible tests are a complete system, then we obtain the normalized distribution

$$U_j=\delta W_j,$$

where

$$\delta=\frac{1}{\sum_{j=1}^{n} W_j}.$$

The normalized distribution U_j depends on the values of $P(A_j/X)$ and τ_j, so that an increase of $P(A_j/X)$ leads to an increase of the value of U_j, and an increase of t_j leads to a decrease of U_j. A connection is revealed in the distribution of U_j with the selection criterion proposed in Section 4.8.

The order of constructing a search program using the distribution U_j is no different from the method discussed in the previous sections. The sequence of the calculations is presented in the block diagram of the algorithm in Figure 4.28.

The malfunction search program constructed with consideration of the distribution of U_j has the following value of the weighted mean work of the malfunction search:

$$T_a(U)=MOT_j(U)=\sum_{j=1}^{n} U_j T_j(U).$$

where

$$T_j(U) = \sum_{k=1}^{k_j(U)} t_k, \quad \text{и} \quad k_j(U) \approx -\log_2 U_j;$$

$T_j(U)$ — is the total work of the test sequence leading to the j^{th} element;

$k_j(U)$ — is the number of tests in the branch of the tree leading to the j^{th} element.

In the general case, with distributions of $P(A_j/X)$ and τ_j which are nonuniform and differ in shape, the following inequality is valid for $T_B(U)$:

$$T_в(\tau) \le T_в(U) \le T_в(P).$$

In practice, several particular cases which depend on the relationship of the shapes of the distribution of τ_j and $P(A_j/X)$ may occur:

1. The distribution of τ_j is uniform. Thus

$$T_в(U) = T_в(P) = t\sum_{j=1}^{n} P(A_j/X)\,k_j(P) = t_{ср}(P).$$

The design of the search program is carried out with consideration of the structural criterion and probability criterion (see Section 4.8).

2. The distribution of $P(A_j/X)$ is uniform. Thus

$$T_в(U) = T_в(P) = \sum_{j=1}^{n} \tau_j T_j(\tau)$$

and the design of the search program is carried out on the basis of the criteria of structure and work (see Section 4.9).

3. The distributions of $P(A_j/X)$ and τ_j are uniform. Then

$$T_в(U) = T_в(P) = T_в(\tau) = tk_{ср}.$$

and when designing the program one takes into account only the structural criterion (see Section 4.7).

4. The shapes of the distributions of $P(A_j/X)$ and τ_j coincide. In this case, one proceeds as if there were an increase of the maximums of U_j, but

$$T_s(U) - T_s(P) - T_s(\tau),$$

as in the previous case.

Thus, the method discussed in this section for constructing a malfunction search program enables one to average the effect of the distributions of τ_j and $P(A_j/X)$ and the structure factor on the work involved in the search. When this is done, the maximum values of U_j correspond to the test sequences which are very close to the minimum in respect to length. The total work of these sequences is minimized, first in terms of magnitude due to taking into account the distribution of τ_j, and secondly in terms of frequency (in the case of repeated cases of failures) due to taking into account $P(A_j/X)$.

The work involved in checks when the productivity of the servicing personnel is a constant is directly related to the cost of performing the work. Therefore, one may consider that the method which takes into account the work involved in the tests will also be valid when, instead of work, the test costs are used.

4.11. Compilation of Instructions for Malfunction Search

The malfunction search programs may be used for developing instructions for malfunction search.

The compilation of instructions for malfunction search begins with dividing the system into elements.

Above it was pointed out that the levels for dividing a system into elements may be different. However, it is customary to divide a system into elements at the following levels:

1. The instruments.
2. The blocks of instruments.
3. The units and modules in the blocks.
4. The stages.
5. The elementary circuits and devices.
6. The simplest parts.

By assigning definite levels for the division of the system, one is able to determine beforehand the intensity of the malfunction search procedure. Thus, if the modules consisting of several stages are selected as the elements of a system, then it would not be possible to construct a search program in which one could succeed in isolating an individual defective stage.

Henceforth, in order to simplify the discussion, we will limit ourselves to divisions approaching the level of a tube stage in the most detailed case.

After division of the system into elements, it is necessary to construct the oriented graph of a system which must take into account the connections of the elements with one another (in the order in which they are followed) from all the inputs to the outputs.

The number of the radar inputs and outputs depends on its purpose. The order in which a search program is determined is by the number of the system outputs. When one output element is present in the radar, monitoring of its operating efficiency constitutes the testing of this element. A deviation of the output signal of this element from the norm is the reason for starting the malfunction search.

For each element of the radar, there is a vertex of the graph. It is necessary to compute the precedence index of these vertices (see Section 1.4).

Furthermore, it is necessary to resolve the question as to the methods of monitoring each element. The following requirements are imposed on the monitoring method:

1. A minimum amount of labor involved in the monitoring.

2. The possibility of monitoring an element without removing it from the system.

3. The most complete representation of the element state.

Then the selection criteria for the order of the tests should be calculated:

1. Calculate the failure probabilities $P(A_j/X)$.

2. Calculate the quantities τ_j based on the known work involved in the selected monitoring methods t_j.

3. Calculate the conditional probabilities of selecting U_j.

The resulting values of U_j are summed in respect to the precedence chains of the elements and a search tree is constructed in accordance with the block diagram of Section 4.10. The tree which is constructed serves as the formal circuit or skeleton of the program (instruction) for the malfunction search. The internal vertices of the tree are the places for locating instructive materials concerning the kinds of tests and the places for conducting them. Instructive materials are distributed along the sequences which have been determined by the shape of the tree. Each sequence is completed by descriptions of the reasons for the failures and recommendations for eliminating them.

As a result of such an arrangement of the checks, an instruction designed for servicing personnel is formulated for searching out malfunctions.

When several independent output elements are present in the radar, during the monitoring of the operating efficiency it is necessary to check in turn the output signals of all the elements. Here the order of monitoring the operating efficiency is connected with the malfunction search program. By testing the output elements, in certain cases, one can succeed in dividing the system into subsets of elements which are not interconnected and which have a single output element. After this is done, search programs can be constructed for each of the subsets in the same way as was done for a system having a single output element.

5. OPTIMUM PROGRAMS FOR DIAGNOSING THE STATE OF A SYSTEM

In the preceding Chapter certain situations which are encountered when searching out system malfunctions were investigated. An oriented graph was used as a possible model which would enable one to study these situations and to organize the malfunction search process.

In practice, one is frequently obliged to deal with considerably more complicated situations which are a result of a set of factors acting on the equipment and its elements. In certain cases, these factors give rise to the slow aging of elements which leads to a change of their parameters. In other cases, breaks or short circuits in a number of circuits or the simultaneous failure of a group of elements appear. In such cases it is customary to talk, not about a search for the defective elements of the system, but about diagnosing its state.

Diagnosing a system states is a more general problem than searching out malfunctions and the solution of this problem using the methods of the theory of graphs is less effective, and sometimes it is impossible in principle.

The problem of constructing optimum diagnostic programs is a problem of mathematical programming and in order to solve it special methods are used. In [48] one of the possible methods for solving this problem is described. This method is based on the application of the concepts of dynamic programming.

In this Chapter we plan to use the method of branches and boundaries for the solution of this same problem [67]. The application of this method in a number of cases enables one to reduce the size of the calculations without reducing the solution precision.

5.1. A Model for the State of a System

As was already indicated, when solving diagnostic problems it is convenient to deal not with actual systems, but with abstract models of these systems. When constructing such models, one must isolate the most significant characteristics of the actual system and discard the secondary characteristics.

One such model when the system is represented by an oriented graph was described in Chapter 4.

Here let us investigate one of the most general models which enables one to describe the various states of an extensive class of actual systems, the so-called table of states [15]. This model gives not only a complete description of the set of the system states, but also indicates the relations between this set and the set of possible checks (tests) with a given number of results of each check.

When constructing the state table, the following assumptions are made. One may divide the system into a certain number of elements which are functionally interconnected. Each such element which represents part of the system may be found in one of two mutually exclusive states: in good working order or out of order. The required input designated by ϕ is fed to each input of the system. In a good working state, the element x_1 responds with the required reaction ψ_1 to the assigned totality of inputs and reactions of other elements. The reaction of an element in good working order is called acceptable. If the element is defective, then its reaction is called unacceptable. In exactly the same way, the input which is applied to an element in good working order in order to obtain an acceptable reaction is called an acceptable action. It is also assumed that an acceptable reaction of any element is obtained only when all the inputs which are applied to this element are acceptable inputs. Moreover, the reaction of a defective element should not depend on the acceptable inputs applied to it. An element's failure is noted in the table of states by a zero, and its good working order is denoted by a one.

Earlier (Chapter 4) the problem was investigated in which each state corresponded to the failure of one and only one element. The number of such states, which equals the number of the system elements, was designated by n. When solving more general diagnostic problems, the number of one system states will not correspond to the number of its elements. The number of the system elements will be designated by N, and then the number of all the logically possible states of the system is $n = 2^N$.

In the general case, each possible state of a system consisting of N elements may be represented by an N-dimensional vector of the states s. The ν^{th} $(1 \leq \nu \leq N)$ component of each of these vectors equals 1 if the ν^{th} element of the system is in good working order, and it equals 0 if the ν^{th} element is defective. Thus, for example, for the system shown in Figure 5.1 the state vector s = (00111) corresponds to the simultaneous failure of elements x_1 and x_2. The vector s = (11111) represents the system state in good working order.

The representation of each state of the system by an N-dimensional vector is very convenient when making a diagnosis. It enables one, after determining the system state, to establish which of its elements is defective.

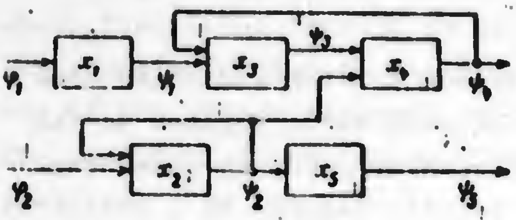


Figure 5.1. Functional model of a system.

The set of all possible states of the system will be designated by

$$S=\{s_i/i=1, 2, \ldots, n\}.$$

For the diagnosis of a system, it is insufficient to know only the vectors of the states. It is also necessary to describe all the possible checks which will be performed in order to determine one or another state of the system. In the given model, in order to monitor the reaction of the state of one or another of the system elements, a check is prescribed, which we will designate by π. Each check may have two results: positive, which is designated by a one if the reaction of the elements being monitored is acceptable, and negative — designated by a zero if the reaction of this element is unacceptable.

In the case when each check has more than two results, other numbers of the series 0, 1, 2, 3, ..., n-1 are also used for their designations. This question was investigated in more detail in Chapter 4.

In order to diagnose an actual system, a set of checks $\Pi = \{\pi_i/i = 1, 2, \ldots, m\}$ is prescribed in each specific case. Each check $\pi_i \in \Pi$ enables one to establish to which result of this check (negative or positive) the subset of states consisting of $k(1 \leq k \leq n)$ elements pertains. When this is done, the remaining n - k states are still unverified. Saying it another way, each check π_i subdivides the set of states into two non-empty subsets. Since from the n elements of set S it is possible to form $2^n - 1$ different non-empty subsets, then the number from all possible checks for the system equals $2^n - 1$. In many practical cases the number of checks assigned is significantly smaller. This is explained by the fact that on certain checks various limitations are imposed which come about because of the inadmissibility of structural changes in the system, because of the absence of necessary monitoring-measuring instruments, etc. Further on, other reasons leading to a reduction in the number of checks being used also will be indicated. Let us proceed to an investigation of the methods for constructing a table of states.

At present one may indicate the following methods for constructing a state table for any specific system:

1. The method of logical analysis of the functional model of a system.

2. The method of artificially introducing malfunctions into the system.

3. The method based on utilizing extrapolated data concerning the states of a system and its elements.

Let us investigate in more detail each of these methods for constructing a table of states.

The Method of Logical Analysis of the Functional Model of a System

The foundation of this method is based on the following [9]. An assumption is made that the system is in a certain state s. By means of a logical analysis of the circuit, its possible behavior is evaluated and the response reactions are determined for each element of the system for the given state. This analysis is carried out for each state $s_i \in S$, and on the basis of the response reactions the results of each check $\pi_i \in \Pi$ are established. The results of the check are entered in a table in which the rows correspond to the states, and the columns correspond to the checks.

As an example, let us construct the state table for the system whose functional circuit is shown in Figure 5.1. Let us assume for simplicity that in the system the failure of any one and only one of the system elements is possible, and each check monitoring the reaction of one of the elements has two results: negative and positive. The negative result occurs when the element's reaction is unacceptable, and the positive result occurs in the opposite case, under the condition that acceptable external actions ϕ_1 and ϕ_2 are fed to the system inputs.

Since each check monitors the response reaction of only one of the elements, the set Π will be composed of five checks — that is, $\Pi=\{\pi_1, \pi_2, \pi_3, \pi_4, \pi_5\}$.

Let us assume that the system is in the state s_1 = (01111) which corresponds to the failure of element x_1. Consequently, its reaction will be unacceptable, and therefore the element listed at the intersection of the first row and the first column of the table will equal zero (Table 5.1). Moreover, the element x_1 precedes elements x_3 and x_4, and this means that the reactions of these elements also will be unacceptable. At the intersection of the first row of the table with the third and fourth columns, zeros will appear. The remaining elements of the first row will be filled with ones.

TABLE 5.1. TABLE OF STATES FOR THE SYSTEM

	π_1	π_2	π_3	π_4	π_5
s_1	0	1	0	0	1
s_2	1	0	0	0	0
s_3	1	1	0	0	1
s_4	1	1	0	0	1
s_5	1	1	1	1	0
s_6	1	1	1	1	1

An analysis of the system circuit is performed in a similar way for the remaining states. For the case when the system state is in efficient working order — that is, s_6 = (11111) — the results of all the checks are positive. The state table obtained as a result of the system analysis is the starting material for constructing diagnostic programs.

The state tables obtained as a result of the analysis of a functional model of a system have a number of properties which we should discuss:

1. With the presence of feedbacks in the system, identical rows and columns appear in the table. Thus, for example, in Table 5.1 the rows corresponding to the states s_3 and s_4, and the columns which represent the checks π_3 and π_4, are such rows and columns.

Henceforth, the states which are represented in the table by identical rows will be called <u>inditinguishable states</u>.

The checks π_3 and π_4 are called equivalent checks. The question of equivalent checks is investigated in more detail in § 5.4 where a precise definition of this concept will be given. The check, which in the table of states is represented by the same ones or the same zeros, will be called an <u>indistinguishable check</u>. Indistinguishable checks do not contain any information concerning the system states, and should be eliminated from the table of states.

In order to eliminate indistinguishable states, it is necessary to introduce into the table additional checks or to change the system structure. One way to change the system structure, for example, is to break the system feedbacks. In particular, in the example being investigated, a break of the feedback leading from the element x_4 to the element x_3 and a feeding of an external action equivalent to the action of x_4 to the input of element x_3 enables one to eliminate the indistinguishable states in Table 5.1. The prescribed set of checks with the absence of feedbacks enables one to determine any state of the system.

2. If the failure of one and only one of the system elements is assumed to exist in the system, then the system may be represented in the form of an oriented graph, as was shown in Chapter 4. In this case, the set of possible system states will be equivalent to the set of its elements.

The Method of Artificially Introducing Malfunctions into the System

This method is based on simulating the states of an actual system by artificially introducing malfunctions into it. The method may be used both in designing systems, and also in their operation.

The construction of the table of states may be performed in the following order.

A list of the possible states of a system is compiled. When compiling this list, one must consider: the degree of detail of the states, the possibilities of simultaneous failure of several elements, the presence in the system of failures of various types (breaks, short circuits, failures of the "faltering" type, stray couplings, pickups, etc.).

First, the output parameters of the system and its elements which will be monitored when the experiment is being performed are selected. Subsequently, the list of selected parameters is refined on the basis of data from the experiment. The list may be expanded or shortened. Each of the selected parameters or a combination of them will correspond to a certain check in the state table.

The system is switched on and readied for the experiment to be carried out. The preparation for the experiment includes the following. The inputs being fed to the system are checked; they must be acceptable inputs. Deviations of the inputs from the acceptable values must not exceed the amounts indicated in the technical specifications for the system. The outputs (reactions) of the system elements are checked; they also must be acceptable. If this is not the case, adjustment of the system or replacement of the defective elements is carried out.

The experiment with the system includes the artificial introduction of malfunctions into the system according to the earlier compiled list of states, and measurement of the parameters selected for monitoring. The measured values of the parameters for the given state are listed in the table of states in accordance with the principle: with a deviation of the parameter from the accepted values a 0 is listed in the table; in the other case, a 1 is listed.

The essence of this method lies in the fact that, when conducting the experiment, it is necessary to know the tolerances for each parameter being monitored, since the subsequent precision and reliability of diagnosing the system states depends on this. In the majority of cases, the tolerances for the parameters are given by the technical specifications for the system. However, with the development of a new system the tolerances of many parameters are unknown, and the problem of determining them is raised. This problem may be solved either by calculation or by experiment. The question of calculating the acceptable values of the parameters is an independent problem [17]. Deterermining the tolerances by experimental means may be performed by boundary tests, which are investigated in Chapter 6.

In order to increase the reliability of the experiment and to reveal the parameters which are most sensitive to changes in the states of the system elements, one may use factor analysis [53].

The described method for constucting a state table entails a number of difficulties. The main difficulty is that the number of all possible states of the system may reach a huge value. Thus, for example, for a system consisting of 20 elements the number of all possible states equals $2^{20} = 1,048,576$. In order to avoid these difficulties, simplified models of the system usually are investigated. In particular, when constructing such models it is assumed that only an ordinary flow of failures may exist in the system. Moreover, the degree of detail of the system states is considerably reduced by combining several elements into a single unit. The experiment for compiling the table of states is carried out, not for the entire system at once, but only for its individual units. For example, if in the system consisting of 20 elements there are five different stages, each of which consists of, on the average, four elements, then for each stage altogether only $2^4 = 16$ states must be investigated. Then considering each stage to be a functional element of the system having an ordinary flow of failures, altogether $5 \times 16 = 80$ states will be investigated, instead of 1,048,576.

It is necessary to point out that this method may be widely used when designing systems consisting of unified units.

Method Based on the Use of Extrapolated Data Concerning the States of the System and Its Elements

During the prolonged use of various systems, abundant data on failures are accumulated. For a certain state of a system in the case of a failure, these data usually contain all the necessary characteristics: the reasons for the failure, the signs or symptoms of the failure, and the rate of failures. By using these data, one is able to compile the system's table of states. One should keep in mind that the completeness of the information concerning all possible states of the system, and consequently the extent of details on the states, depends on the length of time the system has been operated [58].

Let us investigate the application of this method with an example. Let us construct a state table for a tube amplifier using resistors (Figure 5.2).

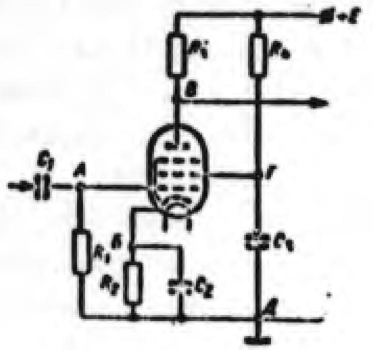


Figure 5.2. A rheostat amplifier.

The state table for this circuit will be compiled on the basis of the following assumptions. A normal input (input signal) is fed to the circuit input from the preceding state; the voltages of the anode power supply and of the tube heater are within acceptable limits. There is no signal at the circuit output. Let us assume that the circuit may be found in the following states:

s_1 - a break of the tube cathode L;

s_2 - a short circuit in the circuit of the leakage resistance R_1;

s_3 - a short circuit in the automatic bias circuit;

s_4 - a break in the circuit of resistor R_4;

s_5 - a breakdown of the capacitor C_3 or a short of its leads;

s_6 - a break in the circuit of resistor R_3;

s_7 - a break in the circuit of the capacitor C_1.

The possible monitoring points A, B, C, D and E at which one may check the circuit parameters are shown on the circuit (Figure 5.2). For these monitoring points, let us assign the following set of checks:

π_1 - monitor the voltage of the input signal (the points A and E);

π_2 - monitor the automatic bias voltage (the points B and E);

π_3 - monitor the anode voltage (the points C and E);

π_4 - monitor the screen grid voltage (the points D and E);

π_5 - monitor the tube parameters using a tube tester;

π_6 - measure the value of the resistance R_1;

π_7 - measure the value of the resistance R_2;

π_8 - monitor the screen grid voltage with capacitor C_3 sealed off.

Using the data concerning malfunctions of a similar type of circuit [57], one may construct a table of states (Table 5.2).

TABLE 5.2. TABLE OF STATES FOR A RHEOSTAT AMPLIFIER

	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8
s_1	1	0	0	0	0	1	1	1
s_2	0	1	1	1	1	0	1	1
s_3	1	0	0	1	1	1	0	1
s_4	1	0	0	0	1	1	1	0
s_5	1	0	0	0	1	1	1	1
s_6	1	0	0	1	1	1	1	1
s_7	0	1	1	1	1	1	1	1

Subsequently Table 5.2 may be used for formulating various diagnostic programs.

The process of constructing diagnostic programs is an independent problem which will be investigated in the following sections.

5.2. Formulating a Diagnostic Program for the Case of an Arbitrary Set of Checks

Before studying the method for constructing optimum diagnostic programs for the state of systems, let us formulate the problem which we have chosen to solve.

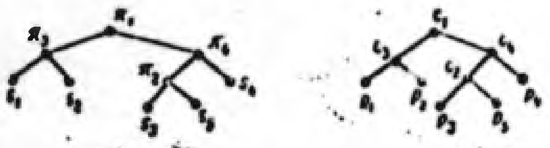
In the preceding section, a model of the system states was investigated. This model enabled one to describe the various states of the system. An arbitrary set of checks was assigned for determining these states. When using this model for the diagnostic procedure certain quantitative characteristics must be specified. One such characteristic is the probability that the system will be found in a given state. In connection with this, the probabilities of the simultaneous failures of several elements are calculated on the basis of the general theorems of reliability theory. Another quantitative characteristic reflecting the expenditures of labor and time, and taking into account the value of the measuring equipment, is the cost of performing the check.

For diagnosing a system state, checks are preformed which must be carried out in a certain sequence. The prescribed sequence of checks is called the diagnostic program. Depending on the system structure, the construction of the automatic monitoring device and other conditions, the system diagnosis may be performed with a program of one of two types — conditional or sequential.

A diagnostic program in which each successive check is based on the result of the previous check is called a conditional program. Thus, for example, a conditional program for a certain system is shown in Figure 5.3, a. The program starts with the check π_1. The checks π_3 and π_4 which follow this are selected on the basis of the results of check π_1.

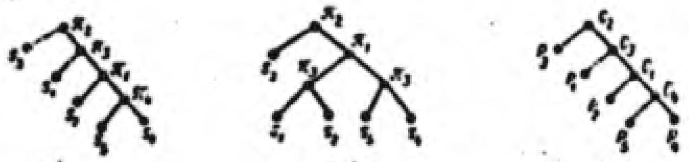
The program for diagnosing a state in which the checks are performed in a definite, previously assigned order independent of their results is called a sequential program.

A sequential diagnostic program is presented in Figure 5.4. This program was constructed for the same data as the conditional program investigated earlier (Figure 5.3, a).



a) b)

Figure 5.3. A sample diagnostic program:
a - a conditional program; b - tree of a conditional program after changes have been made in it.



a) b) c)

Figure 5.4. Types of sequential programs:
a - a program without repetitions; b - a program with repetitions; c - the tree of a sequential program without repetitions after changes have been made in it.

One should note that a sequential program is a particular case of a conditional program. Sometimes it is very difficult to clearly differentiate between the kinds of programs. Thus, for example, in Figure 5.4,b a sequential program is represented whose shape resembles a conditional program. However, this is not a conditional program since for this program, the same as for the program in Figure 5.4,a, the sequence of checks is fixed and does not depend on their results. The difference between these two sequential program is that in the first program (Figure 5.4,a) there are no repeated checks, but in the second program there are. Therefore, programs of the kind shown in Figure 5.4,a will be called <u>sequential programs without repetitions</u>, and programs of the kind shown in Figure 5.4,b will be called <u>sequential programs with repetitions</u>.

Sequential programs have found broad applications in connection with automatic searches and also for the diagnosis of systems in which only elementary checks are possible. A program of any kind is characterized by the average cost for diagnosing a state. If, with specified costs of the checks and given probabilities of the states, a program is constructed for which the average cost of the diagnosis is the least, then such a program is called an <u>optimum program</u>.

For convenience in solving the problem, we will calculate the average cost for diagnosing a state in the following way. Each check π_i which goes into the program corresponds to the beginning of a certain part of the program (subprogram) for diagnosing the states which make up the subset $S_i \subseteq S$. For example, the check π_4 shown in Figure 5.3 of the program is the beginning of the subprogram for the set $S_4 = \{s_3, s_4, s_5\}$. The check π_2 is the beginning of the subprogram for the set $S_2 = \{s_3, s_5\}$, etc. In the program, let us replace each check with its cost c, and the state being diagnosed — with the probability p (Figures 5.3,b and 5.4,c). For each subprogram, the number of which each equals the number of all the checks of the program, the sum is found of the probabilities of those states which are diagnosed by the given subprogram. The resulting sum is multiplied by the cost of the check which began the given subprogram. Finally, the sum of all the products obtained in this way gives the desired result — the average cost for diagnosing the states.

Thus, for example, the average cost of diagnosing the states for the program shown in Figure 5.3,a will equal

$$c_1(p_1+p_2+p_3+p_4+p_5)+c_2(p_3+p_5)+$$
$$+c_3(p_1+p_2)+c_4(p_3+p_4+p_5).$$

The average diagnostic cost for the sequential program (Figure 5.4,a) is calculated in a similar way, that is

$$c_1(p_1+p_2+p_3+p_4+p_5)+$$
$$+c_2(p_2+p_3+p_4+p_5)+c_3(p_3+p_4+p_5)+c_4(p_4+p_5).$$

In order to diagnose the states of a system, an arbitrary set of checks usually is assigned. The number of these checks may be larger than the number of states. When forming the diagnostic program, from the entire set of checks a certain subset of checks which make up the program must be selected in such a way that the average diagnostic cost would be a minimum.

Thus, we arrive at the more general problem of constructing optimum diagnostic programs.

There is a system consisting of N arbitrarily connected elements each of which may be found in one of two mutually exclusive states: in good working order or out of order. The set of possible states of the system $S\{s_t/t=1,2,\ldots,n;\ (n\leqslant 2^N)\}$ is given. In this set, each state is represented by an N-dimensional vector, the ν^{th} component of which equals 1 if the ν^{th} element of the system is in good working order, and equals 0 if the ν^{th} element of the system is defective. The system may be in the state s_t with the probability $p_t\left(0<p_t<1,\ \sum_{t=1}^{n}p_t=1\right)$.

For the given system, a finite set of possible checks $\Pi=\{\pi_i/i=1,2,\ldots,m;\ (m<2^n-2)\}$ is assigned. Each check has two results: positive or negative. The set of checks is assigned by a table of states in which for each check $\pi_i\in\Pi$ its result for each state $s_t\in S$ is indicated. The cost of performing the check π_i equals $c_i\,(c_i>0)$.

Diagnosing a system consists of determining its state by performing a certain sequence of checks. It is assumed that during the diagnosis the system does not change from the given state to another state, and by carrying out at the very most all the checks of Π one is able to determine any state of the system (that is, the states of the system are distinguishable). The cost c_i does not depend on the order of performing the checks. Moreover, it is assumed that the binary tree $H(\Pi_h \cup S, V)$ is made to correspond with the diagnostic program. Here S is the set of suspended vertexes in the tree which correspond to the final results of the checks. Π_h is the set of vertexes, each of which is not a suspended vertex, and we will call these vertexes internal vertexes; V is the set of arcs of the tree. The internal vertexes correspond to the checks being conducted, and the semi-degree of emergence of each internal vertex equals two. The number of internal vertexes of the tree always is one less than the number of suspended vertexes — that is, it equals $h = n - 1$.

The average diagnostic cost for any state of the system may be designated either by C (H) or by $C(\pi_{i_1}, \pi_{i_2}, \ldots, \pi_{i_h}; S)$, with equal validity when it is necessary to indicate the order of performing the checks. In connection with this, the accepted line form of the notation for a given sequence of checks corresponds to the following order of examining the internal vertexes of the tree H. The examination begins with the root vertex, and then each time one descends one level and the internal vertexes of this level are examined from left to right. Thus, for example, for the programs of Figures 5.3,a and 5.4,a, respectively, the sequences of performing the checks will be written in the form

$$\pi_1, \pi_3, \pi_4, \pi_2 \quad \text{and} \quad \pi_3, \pi_2, \pi_1, \pi_4.$$

For the given program H, the average diagnostic cost of a state is determined from the expression

$$C(H) = \sum_{i=1}^{h} c_i \left(\sum_{s_l \in S_i} p_l \right). \tag{5.1}$$

where $S_i \subseteq S$ is the set of states, whose diagnosis is achieved by the subprogram represented by the tree $H_i \subseteq H$ with the initial vertex π_i.

The problem consists of constructing the optimum diagnostic program which will minimize the average cost of determining any state of the system.

The mathematical formulation of this problem and its solution by the methods of dynamic programming are attributed to L. S. Timonen.

In the following sections of this Chapter, another possible method for solving this problem is investigated, which is based on the method of branches and boundaries.

5.3. The Method of Branches and Boundaries

In order to develop algorithms for constructing optimum diagnostic programs or diagnostic programs which are close to the optimum program, one may use the method of branches and boundaries.

The method of branches and boundaries [67], as dynamic programming, is the judicious structure of a search in the set of permissible solutions. For this purpose, the region of all permissible solutions is successively subdivided into ever smaller and smaller subsets. When this is done, the lower bound of the function being minimized is determined for each subset in the subdivision process. The subsets whose lower bound exceeds the permissible solution are eliminated from subsequent investigations. The subdivision process is continued until a permissible solution is found whose value does not exceed the value of the lower bound for any of the subsets.

For constructing diagnostic programs, the solution search process may be represented in the following way. The state table specified for the system contains m checks. Any sequence of checks from among the m checks may be included in the diagnostic program. At least one such sequence will correspond to the optimum program. Nothing is known beforehand, however, about this sequence, and therefore it is reasonable to assume that the program may begin with any of the checks π_e ($e = 1, 2, \ldots, m$). This makes it possible to define the location of the first check in the program; it will correspond to the root vertex of the tree H. The subsequent check, whose location is precisely determined at a certain internal vertex of the tree H, we will call a <u>fixed check</u>.

It is not difficult to determine, on the basis of the table of states, that the check π_e subdivides the entire set of states S into two, non-empty subsets S_e^0 and S_e^1 which correspond to the following: the first subset S_e^0 — to a negative result, and the second subset to a positive result of the check.

Since the order of performing the remaining checks for diagnosing the states included in the subsets S_e^0 and S_e^1 is not determined, the amounts of the average costs $C(H_e^0)$ and $C(H_e^1)$ are unknown. Let us replace the desired solutions for the subsets S_e^0 and S_e^1 by their lower bounds $CM(S_e^0)$ and $CM(S_e^1)$, respectively. Then the lower bound of the average cost of the diagnostic program which began with a certain fixed check π_e may be expressed by the following relationship:

$$CM(\pi_e;\ S)=c_e\sum_{s_i\in S}p_i+CM(S_e^0)+CM(S_e^1). \tag{5.2}$$

where c_e is the cost of performing the fixed check.

Thus, in the first step of the algorithm, the lower bound of the average cost $CM(\pi_e,\ S)$ of each possible diagnostic program, starting with the fixed check π_e ($e = 1, 2, \ldots, m$), is determined. From all the m possible diagnostic programs of the first step of the algorithm, that program is selected which has the least lower bound.

In the second step, for each of the sets S_e^0 and S_e^1 which were separated by the program selected in the first step, the lower bound also is determined for the set of the various pairs of fixed checks which we postulate by π_f for S_e^0, and by π_g for S_e^1, that is

$$CM(\pi_f;\ S_e^0)=c_f\sum_{s_i\in S_e^0}p_i+CM(S_f^{00})+CM(S_f^{01}). \tag{5.3}$$

$$CM(\pi_g;\ S_e^1)=c_g\sum_{s_i\in S_e^1}p_i+CM(S_g^{10})+CM(S_g^{11}). \tag{5.4}$$

where the indexes 00, 01, 10 and 11 designate respectively the subdivisions $S^{0}=S^{00}\cup S^{01}$ and $S^{1}=S^{10}\cup S^{11}$.

Then the lower bound of the average cost of the diagnostic program $CM(\pi_e, \pi_j, \pi_g; S)$ for the set of distinct fixed checks selected in the first and second steps of the algorithm is determined from (5.2), (5.3) and (5.4) as

$$CM(\pi_e, \pi_j; \pi_g; S) = c_e \sum_{s_i \in S} p_i + CM(\pi_j; S^{0}) + CM(\pi_g; S^{1}). \qquad (5.5)$$

This process is carried out in a similar way in the third, fourth and remaining steps of the algorithm until all the subsets in S containing more than two states are separated, and a permissible sequence of checks is found. This sequence of checks corresponds to one of the possible diagnostic programs.

If the average cost of the resulting diagnostic program exceeds the lower bound of the average cost of any of the possible programs of the first and all other steps of the algorithm, then the process is repeated until the optimum solution is obtained.

When solving various problems of optimization by the method of branches and boundaries, the solution process customarily is represented in the form of a so-called <u>solution tree</u>. Thus, for example, the solution tree for a certain problem for which m = 4 and n = 5, which is shown in Figure 5.5, enables one to graphically represent this method.

Each vertex of this tree is associated with a certain permissible solution and its lower bound. The resulting solution is recorded inside, and the value of its lower bound is recorded near the vertex. The vertexes of the same series correspond to a given step. The vertexes in each series are arranged in order of decreasing values of the lower bound from left to right. In Figure 5.5 the four vertexes of the upper series correspond to the first step of the algorithm, and the vertex with solution π_3 has the least lower bound $CM(\pi_3; S)$.

The root vertex of the solution tree corresponds to the set of all permissible solutions which generally may be obtained with the given state probabilities and the assigned costs of the checks. The lower bound of this vertex is CM (S). A calculation of the value of CM (S) for the solution of the problem is of no vital importance.

The branches leading from one vertex to another show the direction of motion toward that permissible solution which follows from the previous solution. Thus, for example, in the solution tree of Figure 5.5 the solutions π_1, π_2; π_1, π_4 and π_1, π_3 follow from the solution π_1. From the solution π_1, π_3, the solutions π_1, π_3, π_2; π_1 π_3, π_4 are obtained, and the complete solution of the problem π_1, π_3, π_4, π_2 follows. The value of the lower bound of the permissible solution for a given vertex nowhere exceeds the values of the lower bounds of those vertexes which follow it. In particular, for the solution tree being investigated (Figure 5.5), we have

$$CM(\pi_1; S) < CM(\pi_1, \pi_3; S) < CM(\pi_1, \pi_3, \pi_4; S) <$$

$$< C(\pi_1, \pi_3, \pi_4, \pi_2; S).$$

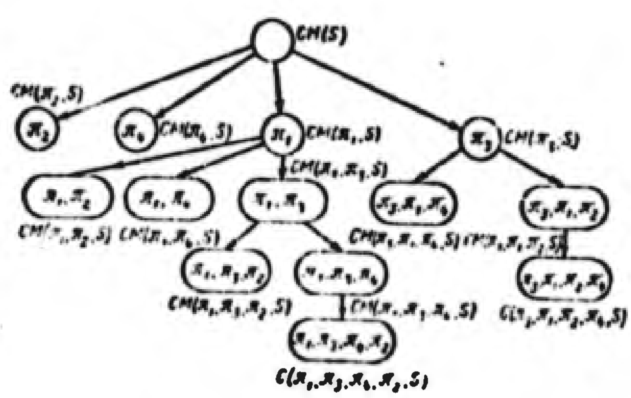


Figure 5.5. A solution tree.

In the solution tree, besides the vertexes with branches, there are suspended vertexes — that is, vertexes from which not a single where the indexes 00, 01, 10 and 11 designate respectively the subdivisions $S^0 = S^{00} \cup S^{01}$ and $S^1 = S^{10} \cup S^{11}$.

Then the lower bound of the average cost of the diagnostic program $CM(\pi_e, \pi_j, \pi_g; S)$ for the set of distinct fixed checks selected in the first and second steps of the algorithm is determined from (5.2), (5.3) and (5.4) as

$$CM(\pi_e, \pi_j; \pi_g; S) = c_e \sum_{s_i \in S} p_i + CM(\pi_j; S^0_*) + CM(\pi_g; S^1_*). \quad (5.5)$$

This process is carried out in a similar way in the third, fourth and remaining steps of the algorithm until all the subsets in S containing more than two states are separated, and a permissible sequence of checks is found. This sequence of checks corresponds to one of the possible diagnostic programs.

If the average cost of the resulting diagnostic program exceeds the lower bound of the average cost of any of the possible programs of the first and all other steps of the algorithm, then the process is repeated until the optimum solution is obtained.

When solving various problems of optimization by the method of branches and boundaries, the solution process customarily is represented in the form of a so-called solution tree. Thus, for example, the solution tree for a certain problem for which m = 4 and n = 5, which is shown in Figure 5.5, enables one to graphically represent this method.

Each vertex of this tree is associated with a certain permissible solution and its lower bound. The resulting solution is recorded inside, and the value of its lower bound is recorded near the vertex. The vertexes of the same series correspond to a given step. The vertexes in each series are arranged in order of decreasing values of the lower bound from left to right. In Figure 5.5 the four vertexes of the upper series correspond to the first step of the algorithm, and the vertex with solution π_3 has the least lower bound $CM(\pi_3; S)$.

branch emerges. Among such vertexes are the following: vertexes giving a complete (not necessarily optimum) solution and vertexes whose subsequent branchings certainly do not lead to an optimum solution. If the value of the solution which has been found does not exceed the value of any permissible solution which is represented by a suspended vertex, then this solution is the optimum solution. This also means that the solution tree has finally been constructed.

The investigated method for constructing optimum diagnostic programs may also be used for a number of other diagnostic problems. However, when determining the solution, each time the problem is raised anew as to the construction of the lower bound. Therefore, in the following two sections, we will investigate the procedures for constructing lower bounds for various types of programs.

5.4. The Lower Bound of the Average Cost of a Conditional Diagnostic Program

The essential feature of the method of branches and boundaries is that a specific requirement is imposed on the lower bound. This requirement is reduced to the fact that the value of the lower bound must not exceed the average cost of any permissible solution.

Let us investigate a certain intermediate step of an algorithm in which it is necessary to determine the lower bound of the average cost of a conditional diagnostic program.

Let S_k ($S_k \subseteq S$) designate the set containing k ($1 \leqslant k \leqslant n$) states which was separated in the preceding step of the algorithm by a certain check. Let us designate by Π_r ($\Pi_r \subseteq \Pi$) the subset of checks such that any check from Π, which subdivides S_k into two nonempty subsets, is included in Π_r. In the general case, Π_r contains r ($0 \leqslant r \leqslant m$) checks. Let us note that when $k = n$ $S_n = S$, $\Pi_r = \Pi_m$ and when $k = 1$ $\Pi_r = \Pi_0 = \varnothing$.

Let the conditional diagnostic program begin with a fixed check π_e ($e = 1, 2, \ldots, r$). The check π_e subdivides S_k into two nonempty subsets $S^0_{k_e}$ and $S^1_{k_e}$, such that $|S^0_{k_e}| = l$ and $|S^1_{k_e}| = k - l$. Moreover, by means of the check π_e two such subsets $\Pi^0_{r_e}$ and $\Pi^1_{r_e}$ are separated from Π_r, such that in each of them only those checks from Π_r are included which subdivide $S^0_{k_e}$ and $S^1_{k_e}$, respectively, into two nonempty subsets.

The calculation of the lower bound for the sets S_k and Π_r begins with calculations of the lower bounds CM $(S^0_{k_e})$ and CM $(S^1_{k_e})$, respectively, for the subsets $S^0_{k_e}$ and $S^1_{k_e}$, and then the value $CM(\pi_e; S_k)$ is determined from Equation (5.2). Since this procedure is the same for any of the subsets, it is sufficient to investigate the process for calculating the lower bound for any one of them — for example, for $S^0_{k_e}$.

Let us investigate the process of calculating the lower bound on the assumption that the costs of all the checks of Π_r are the same, and equal c.

In this case, the procedure for determining the lower bound is equivalent to the procedure for constructing optimum codes, which was proposed by D. A. Kaufman [54]. This procedure proposes that the following operations be performed:

1°. Construct a collection of probabilities of those states which are included in $S^0_{k_e}$. This collection should be ordered in the order of increasing probabilities.

2°. Find the probability $\check{p}_1$ equal to the sum of the two smallest probabilities of the constructed collection.

3°. Repeat 1 and 2 successively, constructing each time a new collection of probabilities. However, as distinct from the procedure indicated in 1, in this collection one must include the probability

$p_\lambda (1 < \lambda < l-1)$ and those probabilities from the preceding collection which were not included in the sum $\check{p}_\lambda$.

4. The process is terminated when a collection consisting of a single component is formed.

For the given probabilities of the set of states $S^0_{k_e}$, this process leads to the construction of a tree of a certain shape. As a result of performing 1°, 2°, 3° and 4° on each suspended vertex of the tree, a certain state $s_i \in S^0_{k_e}$ is possible, and on each internal vertex the binary probability $\check{p}_\lambda$ is possible. The procedure being investigated is graphically illustrated in Figure 5.6. Let us write the resulting binary probabilities $\check{p}_\lambda$ in increasing order, that is

$$\check{p}_1 < \check{p}_2 < \ldots < \check{p}_{l-1}. \tag{5.6}$$

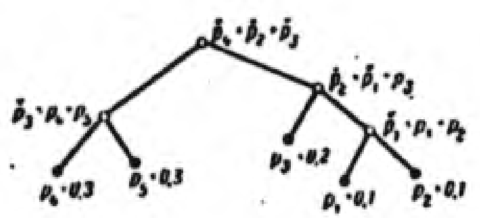


Figure 5.6. An example of forming a tree when calculating binary probabilities.

Henceforth, the collection consisting of $\check{p}_\lambda$ and ordered in accordance with (5.6) will be called the binary collection of probabilities. The number of components in this collection always equals the number of internal vertexes in the tree which may be compared to the diagnostic program for $S^0_{k_e}$. Consequently, by multiplying each component of the collection by the cost of the check c, we obtain

$$CM(S^0_{k_e}) = c \sum_{\lambda=1}^{l-1} \check{p}_\lambda. \tag{5.7}$$

The quantity CM $(S^1_{k_e})$ for the subset $S^1_{k_e}$ is determined in a similar way, that is

$$CM(S^1_{k_e}) = c \sum_{s=1}^{k-l-1} \check{p}_s. \tag{5.8}$$

It is necessary to point out that if $|S^0_{k_e}| = |S^1_{k_e}| = 1$ then $CM(S^0_{k_e}) = CM(S^1_{k_e}) = 0$, and if $|S^0_{k_e}| = |S^1_{k_e}| = 2$ then $CM(S^0_{k_e}) = C(H^0_{k_e})$ and $CM(S^1_{k_e}) = C(H^1_{k_e})$.

Finally, in accordance with (5.2), (5.7) and (5.8), the lower bound of the average cost of the conditional diagnostic program, starting with the fixed check π_e when the costs of the checks are equal, is determined from the relationship

$$CM(\pi_e; S_k) = c\left[\sum_{\pi_i \in S_k} p_i + \sum_{\lambda=1}^{l-1} \tilde{p}_\lambda + \sum_{\nu=1}^{k-l-1} \tilde{p}_\nu\right]. \tag{5.9}$$

As was already indicated, the lower bound of the average cost of the conditional diagnostic program starting with the fixed check π_e must not exceed the average cost of any permissible solution obtained in the following steps of the algorithm. In other words, the following inequality must be fulfilled

$$CM(S_k) \leqslant CM(H_k). \tag{5.10}$$

The proof of this inequality results from the following considerations.

Kaufman's method always leads to an optimum set of code words, in the sense that no other set has a smaller average number of symbols in a message. If one carries out an analogy between the concepts of a "code" and a "program" [9], then the correctness of this method becomes clear. Then the proof for inequality (5.10) follows from the proof that Kaufman's code is the optimum code [52]. Inequality (5.10) changes into an equality when, and only when, the fixed locations of the checks in the tree H_k coincide with their unfixed locations in the tree which determine the quantity CM (S_k).

Now let us investigate the process of calculating the lower bound when the costs of the checks are not equal. As before, let us take $S^0_{k_e}$ as the separated subset.

The process of calculating the lower bound in this case is performed in two independent stages. In the first stage, the probabilities included in the binary collection (5.6) are calculated. In the second stage, the checks are selected, and the value CM ($S^0_{k_e}$) is calculated.

The average cost of the lower bound, when the costs of the checks are not equal, depends essentially on the costs of the selected checks and in which order are they selected. Let us discuss certain specific features of conditional diagnostic programs.

The first distinctive feature is that an equivalence relation exists in the set between certain checks, that is, if

$$\pi_i \equiv \pi_j \to S^0_{k_i} = S^0_{k_j} \text{ or } S^0_{k_i} = S^1_{k_j}, \qquad (5.11)$$

holds, then the checks π_i and π_j are called equivalent checks. The equivalence relation leads to the fact that the set $\Pi^0_{r_e}$ will be partitioned into equivalence classes. In each class, there may be either one or several checks. Therefore, if the checks π_i and π_j are equivalent, when $c_i < c_j$ the check π_j is not included in the program, and when $c_i = c_j$ the check π_i always may be replaced by π_j.

Another distinctive feature is that, in order to distinguish the states of the system, it is sufficient to have a minimum of 1 + + [$\log_2$ (n - 1)] checks, where [a] is the smallest whole part of the number a. Since the number of internal vertexes in the tree equals n - 1, certain of the checks which are included in the program will be repeated. Thus, Figure 5.7 shows a program which enables one to determine seven states of a certain system with three checks. It should be emphasized that this feature does not occur when, as the object being diagnosed, a system is selected which consists of n series connected elements, and only one element may be defective in the system. Then all the checks in the conditional diagnostic program are different.

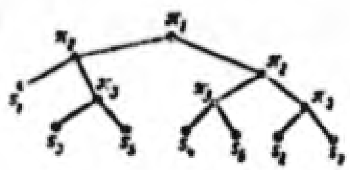


Figure 5.7. A sequential program for seven states and three checks.

Based on the features of conditional diagnostic programs, let us formulate the following laws for the selection of checks, whose costs are included in the lower bound:

1°. Select one check of the least cost from each equivalence class in Π_r^0, and form from them the set of working checks. The set of working checks is used in the following step of the algorithm for calculating the lower bound $CM(\pi_\varepsilon; S^{n}_{k_e})$. One should note that Π_r also is a set of working checks, which was formed in the preceding step of the algorithm.

2°. From the set of working checks, depending on the structure of the object being diagnosed, select:

a) the $(l - 1)$ checks of least cost if the system consists of n series connected elements, and only the failure of one element is possible;
b) the $1 + [\log_2 (l - 1)]$ checks of least cost if the elements of the system are connected in an arbitrary way, or the simultaneous failure of several elements in the system is possible.

After selecting the checks, the procedure for calculating the lower bound is reduced to the following.

Let us form from the costs of the selected checks a collection of costs ordered according to increasing costs, in which $\hat{c}_\lambda (1 < \lambda < l - 1)$ designates the k^{th} component, that is

$$\hat{c}_1 < \hat{c}_2 < \ldots < \hat{c}_{l-1}. \qquad (5.12)$$

The rules for selecting the checks provide for two variations of the system structure. If the system consists of series connected elements, then in accordance with 2°, a, $(l - 1)$ checks will be selected. The cost of each selected check is included exactly one time in collection (5.12). If the elements of the system are connected in an arbitrary way (see 2°, b), then the number of selected checks is smaller than the length of collection (5.12). In this case, let us proceed in the following way. Let the numbering of the costs of the selected checks be made so that

$$c_1 < \ldots < c_j < \ldots < c_z;\ z = 1 + [\log_2 (l - 1)].$$

Let us insert the cost c_1 as the first component of collection (5.12). Let us introduce the cost c_2 as the second and third components, the cost c_3 as the fourth, fifth, sixth and seventh components of collection (5.12), etc.

In the general case, each c_j is included in collection (5.12) 2^{j-1} times except for c_z. The number of repetitions of costs c_z will be smaller in that case when the length of the collection is $l-1 < 2^z-1$. Then multiplying term by term the components of collection (5.6) by the components of collection (5.12), arranged in the opposite order, and summing the resulting products, we calculate the lower bound:

$$CM(S^0_{h_e}) = \sum_{\lambda=1}^{l-1} \breve{p}_\lambda \hat{c}_{l-\lambda}. \qquad (5.13)$$

Precisely in the same way we find that

$$CM(S^1_{h_e}) = \sum_{\kappa=1}^{h-l-1} \breve{p}_\kappa \hat{c}_{h-l-\kappa}. \qquad (5.14)$$

Thus, the general expression for calculating the lower bound of a conditional diagnostic program which begins with the fixed check π_e in the case of unequal costs of the checks has the form

$$CM(\pi_e; S_h) = c_e \sum_{e_i \in S_h} p_i + \sum_{\lambda=1}^{l-1} {}^0\breve{p}_\lambda \hat{c}_{l-\lambda} + \sum_{\kappa=1}^{h-l-1} {}^1\breve{p}_\kappa \hat{c}_{h-l-\kappa}. \qquad (5.15)$$

The lower bound determined by Expression (5.15) also satisfies condition (5.10). This follows from the fact that collections of the kind (5.6) and (5.12) are formed independently, and the methods of obtaining the components of these collections ensure a minimum possible value for each of them. Therefore, in accordance with [55 (theorem 368)] statement (5.10) also remains valid.

5.5. The Lower Bound of the Average Cost of a Sequential Diagnostic Program

The procedure for calculating the lower bound for sequential diagnostic programs without repetitions is somewhat different from the procedure investigated for calculating the bounds of conditional programs. The structure of the tree H which corresponds to the sequential program without repetitions determines this difference. The procedure for forming the binary collection of probabilities will be somewhat different in this case. Moreover, any check included in a permissible sequence always separates two subsets corresponding to its results, such that one of them contains only one state (the final result of the check) and the other contains all the remaining unverified states.

Let us assume that the set $S_{k-n}^{0\vee 1} \subseteq S \setminus \bigcup_{l=1}^{k} s_{t_l}$ is such a subset, which is separated by the fixed check π_{i_k} and which contains all the remaining unverified states. s_{t_l} is the state separated by the check π_{i_l} and the index $0\vee 1$ means that the subset S_{k-n} may be separated by either a negative or positive result of the check π_{i_k}. Then the lower bound of the average cost of the sequential diagnostic program without repetitions, which we will designate by CL, may be written in the form

$$CL(\pi_{i_1}, \ldots, \pi_{i_k}; S) = \sum_{l=1}^{k} c_{i_l} \left(\sum_{s_t \in S_{i_l}} p_t \right) + CL(S_{n-k}^{0\vee 1}). \tag{5.16}$$

where S_{i_l} is the set of states whose diagnosis is achieved by the subprogram represented by the tree with the initial vertex π_{i_l}.

Let us investigate the procedure for calculating the quantity $Cl(S^{ovl}_{n-k})$. This procedure must take into account the structure of the sequential program without repetitions and include the construction of the collection of probabilities and the selection of checks.

Let us form a binary collection of the state probabilities which are included in the set S^{ovl}_{n-k}. In order to do this, it is necessary to carry out the following operations.

1°. Construct a collection of state probabilities ordered according to increasing probability for those states which are included in S^{ovl}_{n-k}, that is

$$p_{i_1} < \ldots < p_{i_j} < \ldots < p_{i_{n-k}}.$$

2°. Find the probability $\hat{p}_1 = p_{i_1} + p_{i_2}$.

3°. Eliminate from the collection 1° the probabilities included in the sum $\hat{p}_l (1 < l < n-k-2)$ and find the probability $\hat{p}_{l+1} = \hat{p}_l + \hat{p}_{i_j}$ $(3 < j < n-k)$.

4°. Successively repeat 3° until a binary collection of probabilities is constructed consisting of n - k - 1 components, that is

$$\hat{p}_1 < \hat{p}_2 < \ldots < \hat{p}_{n-k-1}. \tag{5.17}$$

The checks which are included in the lower bound are selected in accordance with the rules which were presented in Section 5.4 for 1° and 2° and for the selection of the checks. The costs of the selected checks are included in the collection

$$\hat{c}_1 < \hat{c}_2 < \ldots < \hat{c}_{n-k-1}. \tag{5.18}$$

By multiplying the components of collections (5.17) and (5.18) and summing the products, we calculate the lower bound

$$CL(S^{ovl}_{n-k}) = \sum_{i=1}^{n-k-1} \hat{p}_i \hat{c}_{n-k-i}. \qquad (5.19)$$

In the case of equal costs of all the checks, no selection of them is carried out and relationship (5.19) acquires the form

$$CL(S^{ovl}_{n-k}) = c \sum_{i=1}^{n-k-1} \hat{p}_i.$$

The essential feature of the application of the method for constructing sequential programs without repetitions is that in each step of the algorithm the lower bound is not determined for all the fixed checks. It is determined only for those fixed checks which separate one state and the set S^{ovl}_{n-k} from the set of states being investigated in this step of the algorithm.

When constructing sequential programs with repetitions, the procedure for calculating the lower bound in no way differs from the procedure investigated in Section 5.4 for conditional programs. The selection of the checks for the lower bound when this is done is performed in accordance with selection rules 1° and 2°, b. However, the solution tree includes only those checks and their sequences which will either separate a single state (as in the case of sequential programs without repetitions) or subdivide the set S_k into two equally large subsets — that is, $|S^0_{k_i}| = |S^1_{k_i}|$.

When this is done, the condition must be satisfied that for $S^0_{k_i}$ and $S^1_{k_i}$ there are checks which will subdivide each of these two subsets again into equally large subsets, etc. Only sets consisting of 2^z or $3 \cdot 2^z$ ($z = 0, 1, 2, \ldots$) elements make such subdivisions possible. The number of checks which are included in the program for S_k must equal z if $|S_k| = 2^z$, and $z + 2$ if $|S_k| = 3 \cdot 2^z$.

For the selection of these checks, one may use the following principle [48]. Let Π_r be a subset of checks such that after performing them all the states included in S_k will be determined. In order to find these states for the assigned subset Π_r, it is necessary to delete from the table of states the columns corresponding to the checks which are not included in Π_r. Then those states are written out which correspond to the rows of the resulting table which are not encountered twice. Thus, for example, if in S_k there are six states which are represented by the table:

	π_1	π_2	π_3	π_4	π_5
s_1	0	1	0	0	1
s_2	1	0	0	1	0
s_3	1	0	0	1	1
s_4	1	1	1	1	1
s_5	1	1	1	0	0
s_6	1	1	1	0	1

then in order to form a sequential program without repetitions it is sufficient to select the checks π_3, π_4, π_5, since after deleting the columns of π_1 and π_2 all six rows of the table will be different.

5.6 Examples of Constructing Programs for Diagnosing States

Example 1.

Let us investigate the process of constructing a program for diagnosing the states in a tube amplifier, whose circuit was presented in Figure 5.2. As was established, the table of states of this amplifier has the form

	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8
s_1	1	0	0	0	0	1	1	1
s_2	0	1	1	1	1	0	1	1
s_3	1	0	0	1	1	1	0	1
s_4	1	0	0	0	1	1	1	0
s_5	1	0	0	0	1	1	1	1
s_6	1	0	0	1	1	1	1	1
s_7	0	1	1	1	1	1	1	1

(5.20)

Let us assume that, on the basis of the extrapolated data for the amplifier circuit, the following state probabilities are assigned: $p_1 = 0.21$, $p_2 = 0.12$, $p_3 = 0.16$, $p_4 = 0.09$, $p_5 = 0.1$, $p_6 = 0.18$ and $p_7 = 0.14$. The cost of each check from the set $\Pi = \{\pi_i / i = 1, \ldots, 8\}$ includes the work involved in the monitoring and the cost of the monitoring equipment, and equals $c_1 = c_2 = \ldots = c_8 = c = 40$.

Let us use the method investigated earlier for constructing an optimum conditional diagnostic diagram for the case of equal costs.

First of all, let us turn our attention to the fact that in the table of states (5.20) each of the checks π_1, π_2, π_3 subdivides the set of states into the subsets

$$\pi_1 : S_1^1 = \{s_1, s_3, s_4, s_5, s_6\} \text{ and } S_1^0 = \{s_2, s_7\};$$
$$\pi_2 : S_2^0 = \{s_1, s_3, s_4, s_5, s_6\} \text{ and } S_2^1 = \{s_2, s_7\};$$
$$\pi_3 : S_3^0 = \{s_1, s_3, s_4, s_5, s_6\} \text{ and } S_3^1 = \{s_2, s_7\}.$$

In accordance with the definition (5.11), the checks π_1, π_2 and π_3 are equivalent, and, since their costs are equal, any of them may be included in the diagnostic program. Let us select, for example, the check π_1. Consequently, in the first step of the algorithm, it is necessary to calculate the lower bound only for the checks π_1, π_4, π_5, π_6, π_7, π_8. For these checks, the lower bounds are indicated near the corresponding vertexes of the solution tree which is shown in Figure 5.8.

Let us calculate, for example, the lower bound for the check π_4. The check π_4 subdivides set S into two subsets $S_4^0 = \{s_1, s_4, s_5\}$ and $S_4^1 = \{s_2, s_3, s_7\}$.

From the probabilities which correspond to the states of S_4^0, let us form a collection which is ordered in order of increasing probability — that is, (0.09, 0.1, 0.21).

By using the rules for the construction of a binary collection of probabilities, we find that the components of this collection equal 0.19 and 0.40. In accordance with Expression (5.7), we find

$$CM(S_4^0) = 40(0.19 + 0.40) = 23.6.$$

In a similar way let us calculate the lower bound for the subset S_4^1. It is convenient to determine the probabilities of the binary collection on the basis of the scheme

x	$\breve{p}_x$	p	p	p	p
1	0,26	0,12	0,14	0,16	0,18
2	0,34		0,16	0,13	0,26
3	0,60			0,26	0,34

in which the looping line connects the probabilities included in the sum $\check{p}_x$. By using (5.8), we obtain

$$CM(S_4^1) = 40(0.26 + 0.34 + 0.6) = 48.$$

Finally we obtain the lower bound of the average cost of the conditional diagnostic program which begins with the check π_4 from Expression (5.9)

$$CM(\pi_4; S) = 40 + 23.6 + 48 = 111.6.$$

As is seen from Figure 5.8, the diagnostic program beginning with the check π_4 has the least lower bound. Therefore, in the

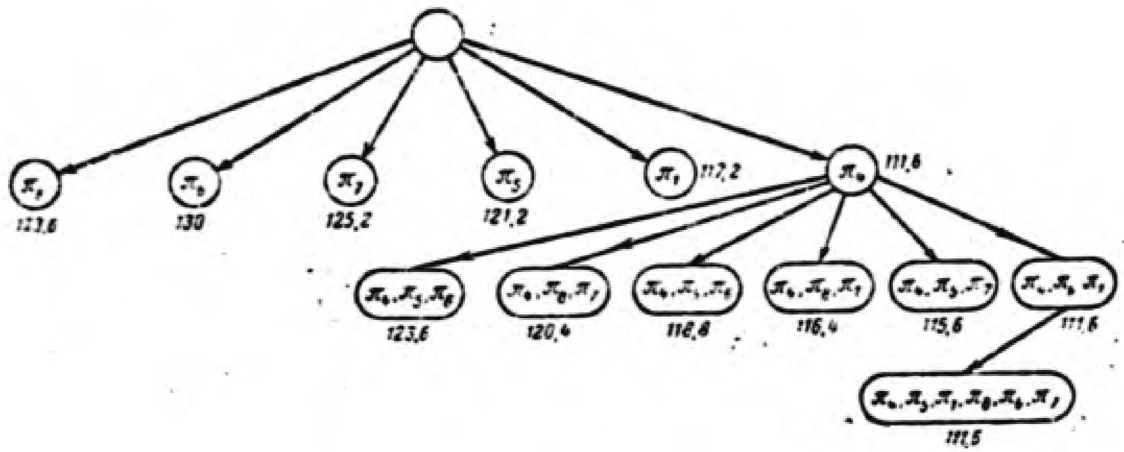


Figure 5.8. The solution tree for Example 1.

second step of the algorithm, it is necessary to calculate the lower bounds for all possible programs which begin with the check π_4 and all the possible pairs of fixed checks following the check π_4.

By rearrangement of the columns and rows, let us represent the table of states (5.20) in the form

	π_4	π_1	π_5	π_6	π_7	π_8	Set
s_1	0	1	0	1	1	1	S_4^0
s_4	0	1	1	1	1	0	
s_5	0	1	1	1	1	1	
s_2	1	0	1	0	1	1	S_4^1
s_3	1	1	1	1	0	1	
s_6	1	1	1	1	1	1	
s_7	1	0	1	1	1	1	

(5.21)

After reorganizing the state table, it is not difficult to see that the checks π_1, π_6, π_7 for the set S_4^0, and the checks π_5 and π_8 S_4^1 are indistinguishable. Therefore, the lower bounds need to be calculated only for the checks π_5 and π_8 of the set S_4^0 and for the checks π_1, π_6 and π_7 of the set S_4^1. Thus, for example, the lower bound for the sequence π_4, π_5, π_1 will equal

$$CM(\pi_4, \pi_5, \pi_1; S) = c\sum_{i=1}^{7} p_i + CM(\pi_5; S_4^0) + CM(\pi_1; S_4^1) =$$
$$= 40\cdot 1 + 40(0.40 + 0.19) + 40(0.6 + 0.26 + 0.34) = 111.6.$$

In the third step, we obtain the final solution of the probelm C (π_4, π_5, π_1, π_8, π_6, π_7; S) = 111.6 for which the conditional diagnostic program is represented in Figure 5.9.

Example 2.

Let us investigate the construction of a conditional diagnostic program for the system whose model is an oriented graph (Figure 5.10).

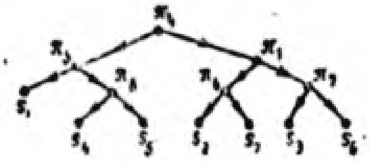


Figure 5.9. The optimum conditional diagnostic program for the rheostat amplifier shown in Figure 5.2.

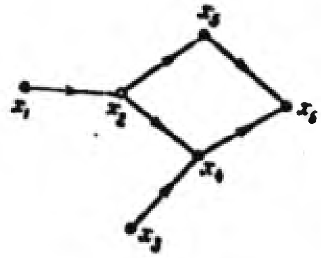


Figure 5.10. The oriented graph of the system.

Each state of the system corresponds to the failure of one and only one element of the model from the set $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, and each check assumes the monitoring of the response reaction of this element. Then the state table of the system will have the form

	x_1	x_2	x_3	x_4	x_5
s_1	0	0	1	0	0
s_2	1	0	1	0	0
s_3	1	1	0	0	1
s_4	1	1	1	0	1
s_5	1	1	1	1	0
s_6	1	1	1	1	1

(5.22)

The costs of the checks and the state probabilities are presented in the following table:

i	c_i	P_i
1	47	0,2
2	71	0,05
3	95	0,24
4	34	0,15
5	32	0,06
6	—	0,3

The entire process of searching for the optimum solution is shown in the form of a solution tree in Figure 5.11. In the first step of the algorithm, the lower bounds for all the checks which are included in $\Pi = \{\pi_i / i = 1, \ldots, 5\}$ are determined, since there are no equivalent checks.

The procedure for calculating the lower bound in this example is somewhat different from the preceding example, since the checks have unequal costs, and the elements in the system are not connected in series.

Let us calculate, for example, the lower bound for the fixed check π_4. The check π_4 subdivides set S into two subsets S_4^0 and S_4^1 in accordance with the structure of the state table (5.22), that is

[illegible]	π_1	π_2	π_3	π_4	π_5	
s_1	0	0	1	0	0	
s_2	1	0	1	0	0	
s_3	1	1	0	0	1	S_4^0
s_4	1	1	1	0	1	
s_5	1	1	1	1	0	S_4^1
s_6	1	1	1	1	1	

(5.23)

From table (5.23) it follows that the checks π_2 and π_5 in S_4^0 are equivalent. Since $c_2 > c_5$, in order to form the lower bound, one should select the check π_5. In the set S_4^1 only the single check π_5 may be included in the lower bound, since the remaining checks are indistinguishable. Thus, in accordance with the rules for selecting checks, one should take the checks π_1, π_3, π_5 for S_4^0 and π_5 for S_4^1. When this is done, two ordered collections of costs will be obtained: for S_4^0 32, 47, 47 and for S_4^1 32.

In the collection of costs for S_4^0, the cost of check π_1 is included twice, since S_4^0 contains four states and consequently there

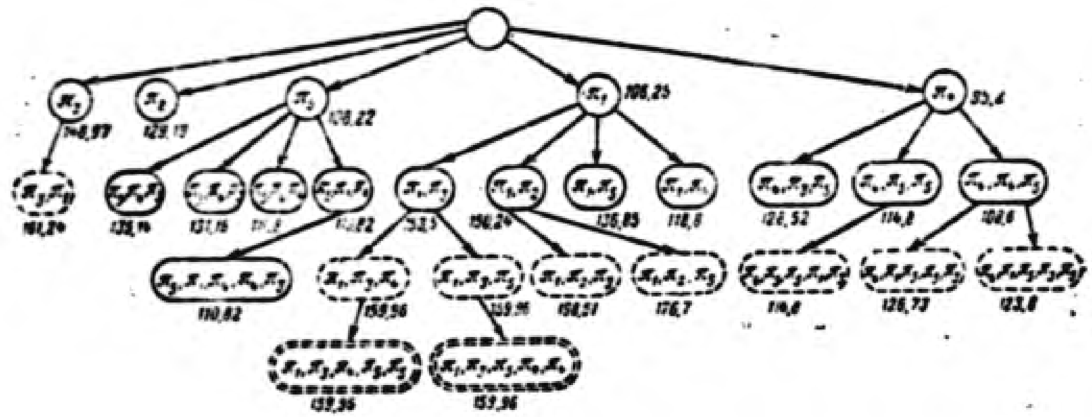


Figure 5.11. The solution tree for Example 2.

must be three internal vertexes in the subprogram tree. The binary collection of probabilities for S_4^0 and S_4^1 is formed in the same way as in the previous example. As a result we find, from (5.15), the lower bound for the fixed check π_4:

$$CM(\pi_4; S) = c_4 \sum_{i=1}^{6} p_i + CM(S_4^0) + CM(S_4^1) =$$

$$= 31 \cdot 1 + (0{,}20 \cdot 47 + 0{,}10 \cdot 47 + 0{,}61 \cdot 32) + 0{,}36 \cdot 32 = 95{,}2.$$

In a similar way, the lower bounds are found for any sequence of checks which are indicated in the solution tree in Figure 5.11. The optimum solution corresponds to the sequence π_5, π_1, π_4, π_3 which is represented in Figure 5.12 by the tree of a conditional diagnostic program. The average cost of locating any defective element of the system for this solution equals C (π_5, π_1, π_4, π_4, π_3; S) = 110.82.

Example 3.

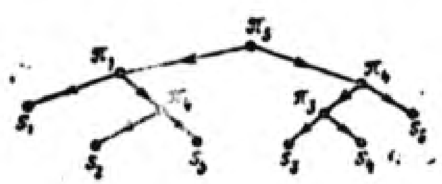


Figure 5.12. The optimum conditional diagnostic program for the system whose model is shown in Figure 5.10.

Let us investigate the process of constructing sequential diagnostic programs for the same initial data which were used in Example 2.

Let us construct a sequential program without repetitions.

In the first step, we may select only the two checks π_1 and π_3 as the checks for which the lower bound must be calculated, since only these checks separate a single state. The lower bounds for these checks are indicated in the solution tree in Figure 5.13.

In order to determine which of the checks should be selected in the following step, it is necessary to delete from the table of states those rows which correspond to the states which have been separated out by the preceding checks.

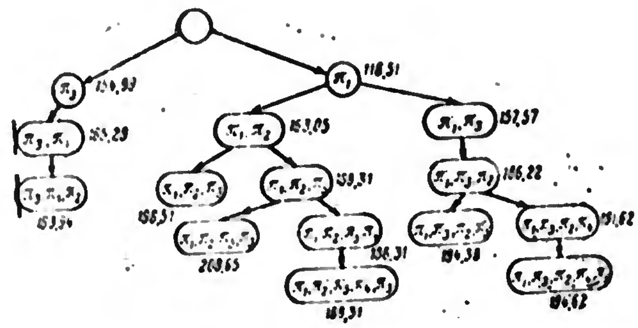


Figure 5.13. The solution tree for Example 3.

Thus, if we hope to calculate in the third step the lower bounds, then it is necessary to delete from the state table (5.22) the rows for the states s_1 and s_3. When this is done, the table takes the form

	π_1	π_2	π_3	π_4	π_5
s_2	1	0	1	0	0
s_4	1	1	1	0	1
s_5	1	1	1	1	0
s_6	1	1	1	1	1

(5.24)

In table (5.24) only the check π_2 separates out a single state. Therefore, in the third step a single sequence of checks π_1, π_3, π_2 is possible. Let us calculate the lower bound for this sequence of checks. The check π_2 subdivides the set $\{s_2, s_4, s_5, s_6\}$ into the subsets S_2 and $S_3^{IVI} = (s_4, s_5, s_6)$.

Let us determine for the set S_3^{IVI} the collections of probabilities (0.21, 0.51) and costs (32.34).

Let us calculate in accordance with (5.19) the value of the lower bound:

$$CL(S_3^{(VI)}) = 0{,}21 \cdot 31 + 0{,}51 \cdot 32 = 23{,}46.$$

Let us determine from Expression (5.16) the lower bound for the sequence of checks π_1, π_3, π_2:

$$CL(\pi_1, \pi_3, \pi_2; S) = 47 \cdot 1 + 95 \cdot 0{,}80 + 71 \cdot 0{,}56 + 23{,}46 = 186{,}22.$$

After performing all the calculations we obtain the final solution:

$$C_*(\pi_1, \pi_2, \pi_3, \pi_4, \pi_3; S) = 188{,}31.$$

The sequential program without repetitions for the resulting solution is shown in Figure 5.14.

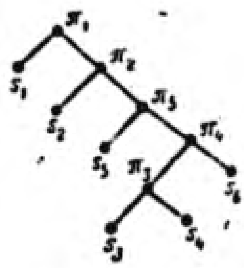


Figure 5.14. The optimum sequential program without repetitions for the system whose model is shown in Figure 5.10.

If it is necessary to construct a sequential program with repetitions, the process of calculating the lower bound and constructing the solution tree would be different, and it would correspond to the procedure for constructing conditional programs.

In the solution tree for the conditional program in Figure 5.11, additional vertexes are shown by the dashed lines which correspond to the search for the optimum sequential program with repetitions. The sequential programs with repetitions for the given case are presented in Figure 5.15. The cost of locating the defective element for any of these programs is

$$C(\pi_1, \pi_3, \pi_4, \pi_3, \pi_2; S) = C(\pi_1, \pi_2, \pi_3, \pi_4, \pi_6; S)_0 = 159{,}96.$$

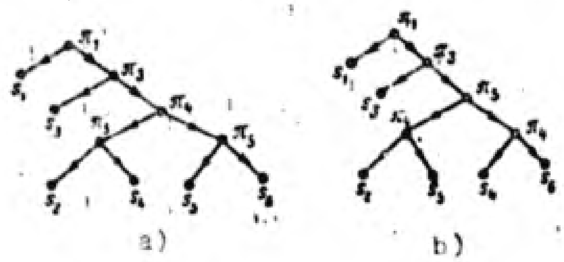


Figure 5.15. Two optimum sequential programs with repetitions for the system investigated in Example 2.

This value is considerably smaller than the cost of the search based on the sequential program without repetitions.

As the examples which have been presented show, the method of branches and boundaries is occasionally connected for practical purposes with a large number of calculations, which increase considerably with an increase of m and n. Therefore, execution of the investigated algorithm is possible only when an electronic computer performs all the calculations.

One of the important features of the practical use of the method of branches and boundaries is the possibility of obtaining a suboptimum (approximate) solution which differs from the optimum solution by no more than a specified amount.

Let us assume that it is necessary to obtain a permissible solution which differs from the optimum solution by no more than 10%. Thus, for example, if in Example 2 the permissible solution CM (π_4, π_5, π_5; S) with the cost 114.8 is found in the second step (Figure 5.11), then one may not investigate vertexes with a lower bound of 105 or more (1.1 X 105 = 115.5 > 114.8).

Then the solution which corresponds to the program shown in Figure 5.16 will be a suboptimum program, since its value

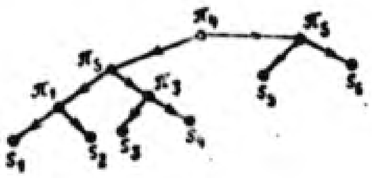


Figure 5.16. A suboptimum conditional diagnostic program for the system investigated in Example 2.

$C(\pi_4, \pi_5, \pi_5, \pi_1, \pi_2; S) = 114.8$ is larger than the optimum solution found in Example 2. The resulting suboptimum solution differs from the optimum solution by only 4%, and the number of calculations is reduced by approximately a factor of 2.2, since the solution tree associated with the search of this suboptimum solution will consist of nine vertexes instead of twenty.

If, when solving the problem, the time for its solution is limited, then in order to find the best solution after a specific interval of time one may use various "adaptive" computational schemes. One such scheme for searching out the best solution during an interval of time t is the following:

1) attempt to find the optimum solution of the problem during the time t/2. If the optimum solution is not found, then

2) attempt to find the suboptimum solution which differs from the optimum solution by no more than 5%; if such a solution is not found after a time equal to t/4, then

3) attempt to find the suboptimum solution which differs from the optimum solution by no more than 10% during the interval of time t/8, etc.

Using the principles of suboptimization when solving problems of large dimensionality enables one to significantly shorten the number of calculations and to reduce the requirements for the storage capacity of the computer.

6. MONITORING RADAR EQUIPMENT

Monitoring may be performed at three stages: when designing the equipment, when manufacturing it, and when operating it. When operating the equipment, the monitoring is carried out to evaluate the operational state or to determine the future behavior of the system.

Monitoring the current state of a system — that is, its operating efficiency — consists of checking the correspondence of the main parameters with the requirements of the technical specifications.

A necessary prerequisite for this is:

— the selection of the minimum number of parameters to be monitored;

— the construction of a program for monitoring the operating efficiency,

— determining the time interval between monitoring checks;

— selecting the necessary set of instruments for performing the monitoring.

Monitoring complicated equipment during its operation may be regarded as one of the effective means for maintaining the reliability.

Since with the passage of time gradual changes of the system parameters take place, it is important to know the tendencies or laws governing such changes. Knowing the laws governing the change of a system state enables one to predict its behavior in the future.

Forecasting a change of a system's state in the future is called a prediction. A necessary condition for predicting is monitoring the change of a system's state during the period preceding the prediction.

The prediction of a change in the equipment's state during its operation enables one to take preventive measures for maintaining the reliability. The prediction of a change in the equipment's state when designing the equipment makes it possible to select the operating conditions of the equipment elements which ensure maximum reliability.

Monitoring technical systems is one of the most important and at the same time least solved problems.

6.1. Selection of the Minimum Number of Parameters for Monitoring a System's Operating Efficiency

When creating an effective system for monitoring an equipment's operating efficiency, the problem arises of rationally selecting the parameters to be monitored. The selected parameters must fully characterize the ability of the equipment to fulfill its tasks.

Monitoring the operating efficiency of equipment calls for checking the state of the equipment or its individual devices — that is, when monitoring equipment, whether it is in good working order or out of order is recorded. In connection with this, the number of parameters selected for monitoring should be at a minimum. The selection of the parameters to be monitored includes the simultaneous selection of the circuit points to be monitored. Since the concept of a "monitoring point" characterizes the specific place in the equipment where a certain parameter is checked, we will henceforth not make too great a distinction between these concepts.

The problem of selecting the minimum number of points for monitoring a system's operating efficiency may be formulated in the following way.

There is a system consisting of functionally interconnected units (elements). Let us assume that one monitoring point for the external output exists in the system for checking the state of each unit (element). Let us designate the set of all external outputs of the system by $X=\{x_i/i=1, 2, \ldots, n\}$.

In those cases when we talk about the set of elements of the system, in order to simplify the designations we will also designate it by X. Obviously this does not lead to any ambiguity, since — according to the formulation of the problem — each unit (element) has one and only one output.

There is a set of external inputs $\Phi=\{\varphi_\nu/\nu=1,2,\ldots,\omega\}$ in the system. It is assumed that each external input of the system is monitored. It is necessary to find a certain set of external outputs $K\subseteq X$ which it is necessary to monitor in order to evaluate the system operating efficiency. Naturally, when this is done it is necessary that the number of elements in set K be a minimum, and the elements of this set must be selected in such a way that, when the monitoring is carried out, complete information as to the system's state is obtained.

Let us use the method discussed in [59] for solving the problem. This method is based on finding the least externally stable set in the graph of the system's outputs.

Let us represent each external output of the system $x_i \in X$ by a vertex of a certain oriented graph $G = (X, U)$. This graph will be called the graph of outputs. In order to construct this graph, let us connect each pair of vertexes x_i and x_j by an arc (x_i, x_j), if the external output x_i is the input of the unit having the external output x_j, or if x_i is the input of some series of units with the external output x_j.

Let us investigate, as an example, the block diagram of a system consisting of eight units (elements) each of which has an external output (Figure 6.1). Consequently, the output graph must have eight

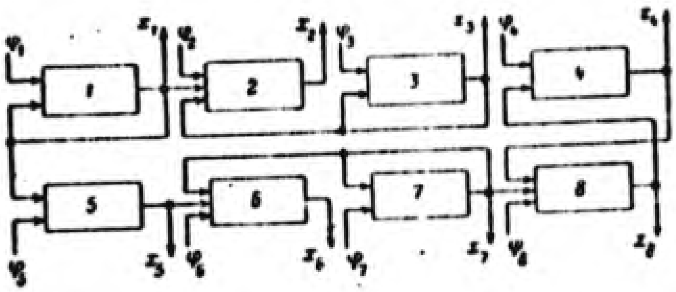


Figure 6.1. The functional circuit of the system

vertexes which may be arranged in an arbitrary way (Figure 6.2a). Let us find the set of arcs of this graph. In order to do this, let us proceed in the following way. Take, for example, unit 1 with output x_1. The output of this unit is connected directly to the inputs of units 2 and 5. Consequently, in the output graph, from the vertex x_1 one must run arcs (x_1, x_2 and x_1, x_5) to the vertexes x_2 and x_5 (Figure 6.2b). Moreover, vertex x_1 must be connected by an arc (x_1, x_6) with the vertex x_6, since the output of unit 1 is connected to the input of the series of units consisting of units 5 and 6.

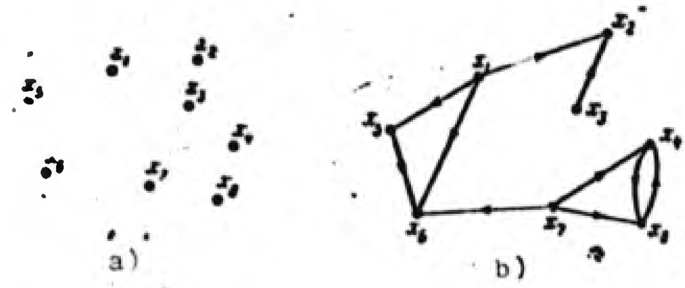


Figure 6.2. Construction of the output graph for the system shown in Figure 6.1

The method of connecting the remaining vertexes of the graph by arcs follows directly from Figure 6.2b.

Let us discuss for a moment one of the distinctive features of an output graph. In any real system, there are functional circuits such that the output of a certain unit is connected to the input of this same unit (feedback). Therefore, in the graph for such outputs loops appear — that is, arcs of the form (x, x). In the given example, outputs x_1, x_3 and x_7 are such outputs. It is not difficult to see that all loops may be eliminated without damaging the solution of the problem, since the state of the elements with feedback is taken into account when monitoring the output of the unit encompassed by the feedback.

Let us use the output graph for solving the problem which has been formulated — namely, for finding a certain least set of outputs by means of which all the remaining outputs are monitored. This problem reduces to finding the least externally stable set in the output graph.

An externally stable set is a certain set of vertexes of the graph in which arcs approach from all the rest of the vertexes which do not belong to K. For the given problem, it is necessary to find that externally stable set of the graph which contains the least number of elements.

Let us investigate the output graph G = (X, U) which is shown in Figure 6.2b. Let us define for it the mapping Δ of the set $X=\{x_i/i=1, 2, \ldots, 8\}$ into a new set $\bar{X}=\{\bar{x}_i/i=1, 2, \ldots, 8\}$ in accordance with the following rule:

1. Each vertex x_i is mapped into the vertex $\bar{x}_i$ (for example, $\bar{x}_1$ into x_1, x_2 into $\bar{x}_2$, etc.).

2. If in graph G there is an arc (x_i, x_j) then for it the arc $(x_j, \bar{x}_i)$ must be formed.

In this way, a so-called simple graph $\bar{G}=(X, \bar{X}, \Delta)$ will be constructed (Figure 6.3a).

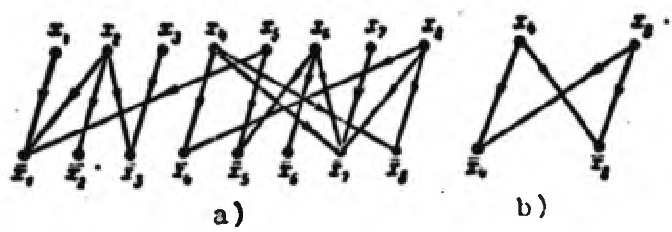


Figure 6.3. A simple graph before simplification (a) and after simplification (b)

The subsequent solution of the problem includes determining, in the output graph, the minimum number of vertexes corresponding to the parameters subject to monitoring. In order to do this, it is necessary to use the algorithm for finding the least externally stable set [6]. As applied to the given problem, this algorithm consists of the following:

1. Remove from the simple graph each such vertex x_i as may be completely replaced by the vertex x_j, such that $\Delta x_i \subset \Delta x_j$ and $x_i \neq x_j$. In our example, let us remove from the graph $\bar{G}$ the vertexes x_1, x_3 and x_7.

2. If in graph $\bar{G}$ there is suspended arc $(x_i, \bar{x}_i)$, then $x_i \in K$. Therefore, vertexes x_2 and x_6 certainly belong to set K.

3. Eliminate from the simple graph the vertexes x_2 and x_6 which are included in K and the sets $\Delta x_2 = \{\bar{x}_1, \bar{x}_2, \bar{x}_3\}$, $\Delta x_6 = \{\bar{x}_1, \bar{x}_5, \bar{x}_6, \bar{x}_7\}$. We obtain the graph shown in Figure 6.3b)

4. Repeat operations 1 and 2. If it is impossible to simplify the simple graph (as in the given example), then we call it an irreducible graph.

5. Furthermore, let us assign to set K that vertex x_i which is mapped into the largest number of vertexes of the set $B \subseteq X$. In the given example, both vertexes x_4 and x_8 are equivalent in this sense, and either of them may be put into K (Figure 6.3b).

As a result of the algorithm, we obtain two solutions which satisfy the formulated problem:

$$K_1 = \{x_2, x_4, x_6\}$$

and

$$K_2 = \{x_2, x_6, x_8\}.$$

Thus monitoring the operating efficiency of the system shown in Figure 6.1 may be accomplished with equal success by checking the outputs of either x_2, x_4 and x_6, or x_2, x_6 and x_8.

In a number of cases, when constructing an output graph, one may make certain simplifications. For example, it is satisfactory to represent a system consisting of series connected units (elements) (Figure 6.4) by a single vertex x_n which corresponds to the output of unit n.

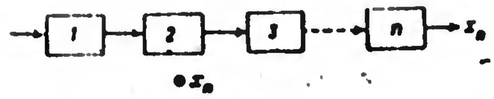


Figure 6.4. The functional model of a system consisting of series connected elements and its output graph represented by the vertex x_n

For a system spanned by feedback (Figure 6.5a) the graph will be symmetric — that is, in this graph two adjacent vertexes x_i and x_j must always be connected by two oppositely oriented arcs (Figure 6.5b). Therefore, one may select any of the outputs when monitoring such a system.

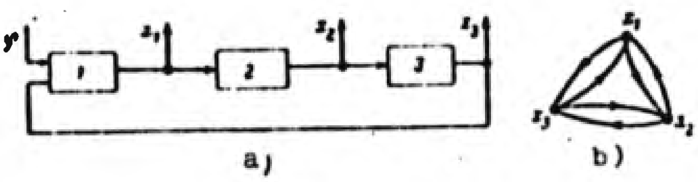


Figure 6.5. The effect of feedbacks in the system on forming the output graph

The method which has been discussed enables one to select the minimum number of outputs (parameters) in a system, the monitoring of which furnishes information on the state of the system and its elements.

Determining the minimum number of parameters for monitoring a system's operating efficiency is one part of the problem. The other part of the problem is constructing the monitoring program.

6.2. Constructing an Optimum Program for Monitoring a System's Operating Efficiency

In order to effectively monitor a system's operating efficiency, it is very important to select an optimum sequence of checks of its external outputs.

Any ordered sequence of checks of a system's external outputs is called a program for monitoring the operating efficiency. By the the optimum program for monitoring the operating efficiency, we understand that program for which the time for carrying out the monitoring will be the least.

Selecting the optimum monitoring sequence is a rather complicated mathematical problem, for whose solution methods of dynamic programming may be used [49].

Here we will investigate an algorithm which enables one to construct, for specified criteria, an optimum monitoring program. This algorithm is based on the method of branches and boundaries [67].

First, let us investigate certain concepts connected with the general principle of constructing programs for monitoring the operating efficiency. Let us assume that, in order to monitor a system's operating efficiency as a whole, it is sufficient to check the total number of external outputs as was shown in §6.1. Monitoring each output of this group consists of measuring the value of a certain parameter. In a specific piece of equipment, such parameters may be: the shape or amplitude of a pulse, the power being generated, the current or voltage in a circuit, etc.

Henceforth, if in a given output x_j the parameter ψ_j is monitored, we will say that the check of this parameter π_j is performed.

Any check has two and only two results: negative when the value of the parameter being monitored lies outside the field of tolerance, and positive when the value of the parameter being monitored is found to be within the tolerance field. When a certain check π_j has a negative result which corresponds to an inoperable state of the system, it separates out a certain set $X_j \subseteq X$ of the system's elements, among which there is one or several defective elements. On the other hand, the defectiveness of any element $x \in X_j$ always leads to a negative result of the check π_j.

Thus, the elements which belong to the set X_j are combined into the set on the basis of their common properties — the properties which give rise to a negative result of the check π_j when there is a failure of any of them. In the theory of sets, this property is called the <u>relation of equivalence</u>. The given equivalence relation leads to the subdivision of set X into nonintersecting and nonempty subsets (see Chapter I). For example, for the system which was investigated earlier (Figure 6.1) if the checks are performed in the sequence π_2, π_6, π_8, then set X will be subsivided by these checks into the subsets:

$$X_1 = \{x_1, x_2, x_3\}, X_2 = \{x_5, x_6, x_7\}, X_3 = \{x_4, x_8\}$$

(Figure 6.6). Let us note that any other sequence consisting of these same three checks and applied to the given system leads to a subdivision of set X into three other subsets which are distinct from X_1, X_2, and X_3. If all the checks which are performed have a positive result, then the system is considered to be operating efficiently, and this monitoring result is designated by Y.

It is convenient to represent the sequence of checks being performed — that is, the program for monitoring the operating efficiency — in the form of a tree $H = (\Pi_k \cup Z, V)$. In tree H, we designate: Π_h is the set of internal vertexes of the tree which corresponds to the checks being conducted; h is the number of internal vertexes in the tree; $Z = Y \cup \{X_j/j = 1, 2, \ldots, h\}$ is the set of suspended vertexes which correspond to the monitoring results, and V is the set of arcs in the tree.

From each internal vertex of the tree two arcs emerge, one of which corresponds to the negative result of the check, and the other to the positive result. In particular, for the previously investigated sequence of checks, the tree is shown in Figure 6.6b. In this tree the internal vertexes designate the sequence of checks π_2, π_6, π_8 and the monitoring results are designated by the suspended vertexes X_1, X_2, X_3, and Y.

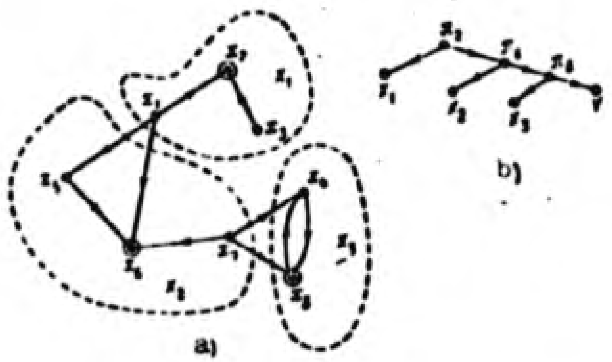


Figure 6.6. The process of constructing a program for monitoring the operating efficiency:
a — subdividing the system into subsets; b — the program for monitoring the operating efficiency

Let us assume that in order to solve the problem we have the following initial data on the time characteristics of the monitoring and the probabilities of the system's states:

t_j — is the time required for performing the check Π_j;

P_i ($i = 1, 2, \ldots, n$) — is the probability of the defective state of the system associated with the failure of the element $x_i \in X$;

$P(t)$ is the probability that the system is operating efficiently.

In order to determine these quantities, one may use the fundamental relationships from the theory of reliability which, as applied to the case being investigated, may be written in the form

$$P_i = \frac{\lambda_i}{\sum_{x_i \in X} \lambda_i}(1 - P(t)) = P(A_i/X)(1 - P(t)). \tag{6.1}$$

$$P(t) + \sum_{x_i \in X} P_i = 1. \tag{6.2}$$

Let us designate by $T(\pi_{j_1}, \pi_{j_2}, \ldots, \pi_{j_k}; Z)$ the average time for monitoring the system's operating efficiency with the given sequence of checks $\pi_{j_1}, \pi_{j_2}, \ldots, \pi_{j_k}$. For notational brevity, we will also designate the average monitoring time by $T(\Pi_k; Z)$. The average time for monitoring the operating efficiency is determined from the following relationship:

$$T(\Pi_k; Z) = \sum_{r=1}^{k} t_{j_r}\left[P(t) + \sum_{l=r}^{k} P(X_l)\right], \tag{6.3}$$

where $P(X_l) = \sum_{x_i \in X_l} P_i$, and the sum is taken over all x_i which belong to the set X_l which was isolated with a negative result of check π_{j_l}.

The problem is to construct the optimum monitoring program which minimizes the average time for monitoring the system operating efficiency.

Let us investigate the process of constructing such a program which, as was indicated above, is based on the general concepts of the method of branches and boundaries.

The entire process is divided into a series of steps. Since the program may begin with any of the checks which may be used for the monitoring, in the first step, the lower bounds are determined for the average monitoring time of each such program. The number of these programs equals the number of checks. From all the possible programs of the first step, that program is selected for which the lower bound of the average monitoring time is the least. Since the check in the first step is determined, in the second step the lower bounds are found for programs with two checks. As the second check, any of the checks selected for the monitoring are taken, except for the check which was already included in the program selected in the first step. Again from all the programs of the second step, that program is selected with the least lower bound. In the third step, the process is repeated, but of the three checks two have already been determined in the two preceding steps. The lower bounds of all possible programs are found in respect to the third check, and the program with the least lower bound is selected.

The process is continued until, in the h step, one of the possible programs is completely constructed. If the lower bound obtained in step h of the program is smaller than the lower bounds of all possible programs of the preceding steps, then it corresponds to the desired solution and is the optimum solution. If the lower bound of the program obtained is larger than the lower bound of even one of the programs of the first step, then the process is repeated in the same order as was described but for the other possible program of the first step.

Thus, the process of searching for the optimum solution is a branching process, and it is conveniently represented in the form of a tree. This tree will be called the solution tree for constructing the program for monitoring the operating efficiency. The solution tree for the example presented at the end of the section is shown in Figure 6.7.

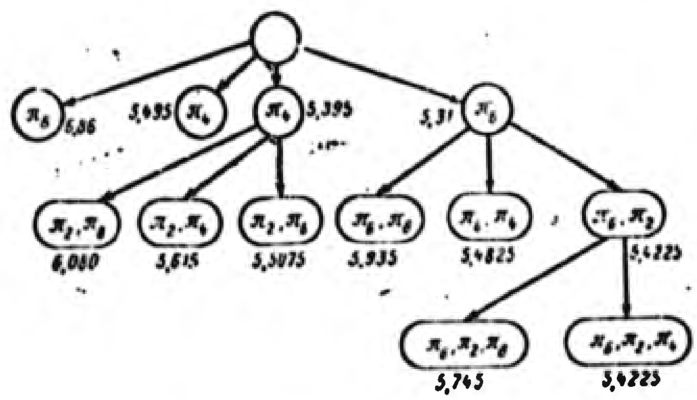


Figure 6.7. The solution tree when searching for the optimum program for monitoring the operating efficiency

In order to construct an optimum program by the method of branches and boundaries, it is necessary to derive a relationship for the lower bound of the average time for monitoring the operating efficiency. The lower bound of the average time for monitoring the operating efficiency for a certain monitoring program will be designated by $TR(\Pi_k; Z)$. Let us assume that, in a certain step of the branching algorithm, a sequence of k checks $(0 \leqslant k \leqslant h-2)$ has already been found which begins with the check π_0 — that is, the following sequence occurs: $\pi_0, \pi_{j_1}, \pi_{j_2}, \ldots, \pi_{j_k}$. The check π_0 is called a fictitious check. This check has no physical meaning, but it is convenient to introduce it for solving the problem.

Let us designate by $\Pi_k = \bigcup_{r=0}^{k} \pi_{j_r}$ the set of those checks which are included in the sequence which has been found. In precisely the same way, $Z_k = \bigcup_{r=0}^{k} X_r$ designates the set of the system elements which are monitored by this sequence. For the check π_0, we will assume $\Pi_0 = \pi_0 = \varnothing, Z_0 = X_0 = \varnothing$. Moreover, $Z_1 = X_1$.

Let $T(\Pi_k; Z_k)$ be the average time for monitoring the elements of set Z_k by means of the checks from Π_k. Then for the fixed check $\pi_e \in \Pi \setminus \Pi_k (0 \leq k \leq m)$ the lower bound may be written in the form of the following relationship:

$$TR(\Pi_k, \pi_e; Z) = T(\Pi_k, \pi_e; Z_{k+1}) + \dot{T}R(\Pi \setminus \Pi_{k+1}; Z \setminus Z_{k+1}) \qquad (6.4)$$

In this relationship, the first term is the average monitoring time for the program constructed in accordance with the method after the first k steps. The second term is the lower bound for that part of the monitoring program which is constructed during the remaining h - k steps.

Let us represent this relationship in a form which is more convenient for making calculations. From Expression (6.3), we have

$$T(\Pi_{k+1}; Z_{k+1}) = \sum_{r=0}^{k} t_{s_{r+1}} \left[P(t) + \sum_{l=r+1}^{k} P(X_l) \right]$$

or considering that

$$P(t) + \sum_{l=r}^{k} P(X_l) = 1 - \sum_{l=0}^{r-1} P(X_l),$$

we obtain the following relations for $k = 0$ and $k > 1$.

When $k = 0$:

$$T(\Pi_k, \pi_e; Z_{k+1}) = t_e$$

for any $e (1 < e < m)$;

when $k \geqslant 1$:

$$T(\Pi_k, \pi_e; Z_{k+1}) = \sum_{r=0}^{k-1} t_{j_{r+1}} \left[1 - \sum_{i=0}^{r} P(X_i)\right] + $$

$$+ t_e \left[1 - \sum_{i=0}^{k} P(X_i)\right]. \qquad (6.5)$$

The lower bound $TR(\Pi \setminus \Pi_{k+1}; Z \setminus Z_{k+1})$ may be determined from the following considerations. Since inequality $TR(\Pi; Z) \leqslant T(\Pi; Z)$ always must hold, we formally assume that after the fixed check π_e a check follows which makes it possible to monitor the remaining set of elements $Z \setminus Z_{k+1}$.

Assuming that the time for performing the check which follows check π_e equals $\min_{\pi_j \in \Pi \setminus \Pi_{k+1}} \{t_j\}$, we find the expression for the lower bound:

$$TR(\Pi \setminus \Pi_{k+1}; Z \setminus Z_{k+1}) = [\min_{\pi_j \in \Pi \setminus \Pi_{k+1}} \{t_j\}] \times$$

$$\times \left[1 - \sum_{i=0}^{k+1} P(X_i)\right] + P(1) \sum_{\pi_j \in \Pi \setminus \Pi_{k+1}} t_j. \qquad (6.6)$$

Let us illustrate the proposed method of constructing an optimum program for monitoring a system's operating efficiency with an example.

Example. Let us investigate the system whose functional circuit was presented in Figure 6.1. Let the following failure rates of its elements be assigned for this system

$$\lambda_1 = \lambda_2 = 2 \cdot 10^{-6}, \quad \lambda_3 = 10 \cdot 10^{-6}, \quad \lambda_4 = 2 \cdot 10^{-6},$$
$$\lambda_5 = 7 \cdot 10^{-6}, \quad \lambda_6 = \lambda_7 = 5 \cdot 10^{-6}, \quad \lambda_8 = 7 \cdot 10^{-6}.$$

Let us assume that the system operates efficiently with a probability equal to 0.8. The system's external outputs which are subject to monitoring were determined by us earlier, and they correspond to the following two sets of parameters: $\{\pi_2, \pi_4, \pi_6\}$ and $\{\pi_2, \pi_4, \pi_8\}$.

Let us assume that, in order to monitor each parameter, the following times are required: $t_2 = 2$ min, $t_4 = 2.5$ min, $t_6 = 1.5$ min, $t_8 = 3.0$ min. Let us construct the optimum program for monitoring the system operating efficiency for the given case.

In accordance with Expressions (6.1) and (6.2) let us find the values of the probabilities $P(A_i/X)$ and P_i. The data are listed in the following table:

i	$P(A_i/X)$	P_i
1	0,050	0,010
2	0,050	0,010
3	0,250	0,050
4	0,050	0,010
5	0,175	0,035
6	0,125	0,025
7	0,125	0,025
8	0,175	0,035

<u>Step 1</u>. Let us determine the lower bounds of the possible monitoring programs which may begin with one of the checks π_2, π_4, π_6, and π_8. For example, let us determine the lower bound of the program beginning with check π_6. Check π_6 in the case of a negative result separates out the set $X_1 = \{x_1, x_5, x_6, x_7\}$. Since check π_6 is the first one of the sequence, from Expressions (6.5) and (6.6) it follows:

$$TR(\pi_6; Z) = t_6 + [\min(t_2, t_4)][1 - (P_1 + P_5 + P_6 + P_7)] +$$
$$+ P(l)\,t_6 = 1{,}5 + 2(1 - 0{,}095) + 0{,}8 \cdot 2{,}5 = 5{,}31.$$

For the remaining checks, the lower bounds are indicated in Figure 6.7 near the corresponding vertexes of the solution tree. Since the lower bounds were calculated for the monitoring programs using the two sets of parameters (π_2, π_4, π_6) and (π_7, π_6, π_8), it is necessary to investigate only the permissible sequences of checks. Any sequence of checks is called permissible if all the checks of this sequence belong to one and only one of the sets of checks which are selected for monitoring. For example, the sequence π_2, π_4, π_8 is inadmissible, since in this sequence checks π_2 and π_4 are included in one set of checks, and check π_8 is included in another set.

From Figure 6.7 it is seen that the monitoring program beginning with check π_6 has the least lower bound; therefore, it is selected for the subsequent construction.

<u>Step 2</u>. In the given case this step enables one to construct only a single monitoring program from all the possible programs. In order to do this, let us calculate the lower bounds for the various permissible pairs of checks which begin with check π_6 — that is, π_6, π_8; π_6, π_4; π_6, π_7. The lower bounds for the permissible pairs of checks are indicated in Figure 6.7. Let us determine, for example, the lower bounds for the sequences of checks π_6, π_4 and π_6, π_7. The first check π_6 in the sequence π_6, π_4 separates out the set $X_1 = (x_1, x_5, x_6, x_7)$; the check π_4 which follows it separates the set $X_2 = (x_4, x_5)$. Since only the check π_2 may follow the check π_4, we obtain from Expressions (6.5) and (6.6) the lower bound for the sequence of checks:

$$TR(\pi_6, \pi_4; Z) = t_6 + t_4[1-(P_2+P_5+P_6+P_7)] +$$
$$+ t_8[1-(P_1+P_5+P_6+P_7+P_4+P_8)] =$$
$$= 1.5 + 2.5(1-0.095) + 2(1-0.14) = 5.4825.$$

In precisely the same way, we determine the lower bound for the sequence of checks π_6, π_2 in which the check π_2 separates out the set $X_2 = \{x_2, x_3\}$. Moreover, it is necessary to take into account the fact that either check π_4 or check π_8 may follow check π_2. Therefore, we have

$$TR(\pi_6, \pi_2; Z) = t_6 + t_2[1-(P_1+P_5+P_6+P_7)] +$$
$$+ [\min(t_4, t_8)] \times [1-(P_1+P_5+P_6+P_7+P_2+P_8)] =$$
$$= 1.5 + 2(1-0.095) + 2.5(1-0.155) = 5.4225.$$

Among the various sequences of checks beginning with check π_6, the sequence π_6, π_2 has the least lower bound, but at the same time the lower bound TR (π_6, π_2; Z) is larger than the lower bound π_2, π_4; π_2, π_6; π_2, π_8. Therefore, in keeping with the concept of this method, it is necessary to determine the lower bounds of all the permissible sequences of checks beginning with the check π_2.

The calculated lower bounds for the sequences π_2, π_4; π_2, π_6; π_2, $_8$ are indicated in Figure 6.7. However, as follows from Figure 6.7, the selection of any of the monitoring programs which begin with check π_2 is no improvement over the results obtained for sequence π_6, π_2, π_4.

Thus, we obtain the solution

$$T(\pi_6, \pi_2, \pi_4; Z) = 5.4225,$$

which is the optimum solution.

The optimum program for monitoring the operating efficiency of the system investigated in the example is shown in Figure 6.8.

The optimum program for monitoring the operating efficiency of the system investigated in the example is shown in Figure 6.8. The method which has been discussed enables one to construct the optimum program for monitoring the system operating efficiency. However, this method has the shortcoming that, when there is a large number of parameters which determine the system operating efficiency, a large number of calculations must be made.

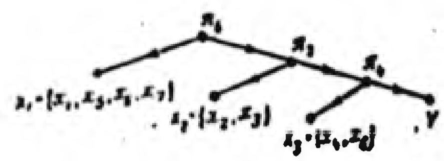


Figure 6.8. The optimum program for monitoring the operating efficiency of the system shown in Figure 6.1

A reduction in the extent of calculations when solving engineering problems is very desirable. Therefore, in the following section a method will be investigated which enables one to find the approximate solution with a relatively small number of calculations.

6.3. Construction of a Program for Monitoring the Operating Efficiency Based on the Method of Preference

The construction of a program for monitoring the operating efficiency using the method of preference is based on the following. The order of performing the checks is selected in such a way that each preceding check gives information on the state of the greatest number of elements in the system. In order to determine the sequence of checks, a certain preference function and preference rules are used.

The preference function F makes the j^{th} $(1 \leqslant j \leqslant m)$ parameter of the system agree with a certain number F_j, which in the general case depends on the system's structure. It also depends on which parameters are included in the monitoring program. The preference rule determines the system's parameter which must be introduced into the program after performing the earlier specified checks.

Depending on the specific problem, the initial data on which the program is constructed, and the system's structure, various preference functions may be introduced.

For many practical purposes, it is convenient to use as the preference function the function $F_p(j_\xi)$ which takes into account the relation of precedence of the system's elements.

When the system has a number of outputs, each of which must be monitored, one is able to formulate and solve the problem by the preference method in the following way.

Let there be a system of n elements containing $m > 1$ outputs which are not interrelated by a precedence relation. Let us assume that there is no feedback in the system. The probability of a defective state of the system associated with the failure of element x_i equals $P(A_i/X)=\frac{1}{n}$. The same time t is required for checking any of the m outputs of the system. A certain precedence index j is made to agree with each element of the system according to the rule indicated in Chapter I. It is clear that the precedence indexes of the output elements are always smaller in magnitude than the total number of elements of the system — that is, the inequality holds

$$j_\xi < n,$$

where $j_\xi (\xi = 1, 2, \ldots, m)$ — is the precedence index of the system's output element with the order number ξ.

Since checking the state of an output element with an arbitrary index j_ξ enables one to obtain information about the state of the subset of only those elements which precede this output, the preference function may be represented in the form of the relation

$$F_p(j_\xi)=\frac{k_\xi}{n}. \tag{6.7}$$

Obviously the closer $F_p(j_\xi)$ is to one, the larger the number of the system's elements that are monitored by a single check and, consequently, the larger the "share" this check has when monitoring the operating efficiency.

Let us number the system outputs in decreasing order of the quantities $F_p(j_\xi)$:

$$F_p(j_1)\geqslant F_p(j_2)\geqslant \ldots \geqslant F_p(j_m). \tag{6.8}$$

and we will perform the checks when monitoring the operating efficiency in order of increasing index ξ. Thus Relation (6.8) indicates the rule for constructing the program for monitoring the operating efficiency. However, this rule is valid only for systems whose structure is such that there are no intersections between the subsets of the elements being monitored.

In any actual system, the most diverse intersections exist between the subsets of elements, each of which is monitored at its own output element. This leads to the fact that the initial numbering of the system's output elements which was assumed in (6.8) subsequently gives an incorrect result when constructing the program for monitoring the operating efficiency. Let us illustrate what we have said with an example. For this, let us investigate the system represented in Figure 6.9 in the form of a graph. The system has three output elements which are designated in the figure by x_6, x_7, and x_8. Each element of the system agrees with a precedence index, as is shown in Figure 6.9. If Rule (6.8) is followed, then the system's operating efficiency must be checked in the following order.

In accordance with the accepted numbering output, x_6 which monitors the largest number of elements is checked first, then output

Figure 6.9. An oriented graph which is a model of a system

Figure 6.10. An example of constructing a monitoring program by the method of preference

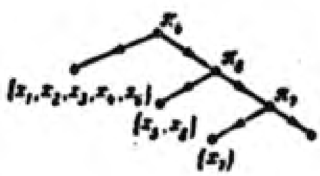


Figure 6.11. An example illustrating the reduction of the monitoring time with a change in the sequence order of the checks

x_7, and finally output x_8. The program for this case is shown in Figure 6.10. Near each suspended vertex of the tree which is shown in this figure the subsets are shown which are separated out by each of the checks in the given sequence.

In the figure it is seen that check π_7, which precedes check π_8, separates the single element x_7, and the check π_8 separates two elements x_5, x_8. Therefore, it is necessary to rearrange checks π_7 and π_8. Then the program for monitoring the operating efficiency of the given system will appear as shown in Figure 6.11.

Comparing both these programs, one may conclude that the monitoring time for the system in the first program (Figure 6.10) will be somewhat longer than in the second program (Figure 6.11).

When constructing the program for monitoring the operating efficiency for a system with a small number of elements, this defect is unimportant. However, such errors for large systems lead to a considerable increase in the monitoring time.

Taking into account these features of the method, we arrive at the following rule. In each step of constructing the program, in order to determine the order number of the following check, it is necessary to:

— calculate new precedence indexes for the elements belonging to that part of the system which was not checked by the preceding checks;

— determine the number $k (0<k\leqslant n)$ equal to the number of elements in the unverified part of the system;

— calculate the preference function for each output element whose check has still not been included in the program using the equation

$$F_p(j_k)=\frac{l_k}{k} \quad (0<k\leqslant m); \tag{6.9}$$

— acting according to the rule indicated earlier (6.8), determine the order number of the following check.

If a certain check divides a system into disjoint subsets, each of which is monitored at its output, there is no need to make further calculations of the function F_p.

Let us investigate this method of constructing a program for monitoring the operating efficiency with a specific example. Let us assume that, for a system consisting of 16 elements (Figure 6.12), a program must be constructed for monitoring the operating efficiency using the method discussed above. The initial data for constructing the program are as follows:

— the probability that the system is operating efficiently is $P(t) = 0.8$;

— the probability of a defective state of the system because of the failure of the i^{th} element equals

$$P_i = \frac{1}{16}(1 - 0.8) = \frac{1}{80};$$

— in order to check any output of the system, the time $t = 2$ min is required.

The system operating efficiency is monitored by checking the outputs of the elements 4, 8, 12, and 16. It is assumed that when this is done the inputs of elements 1, 5, and 13 also are monitored.

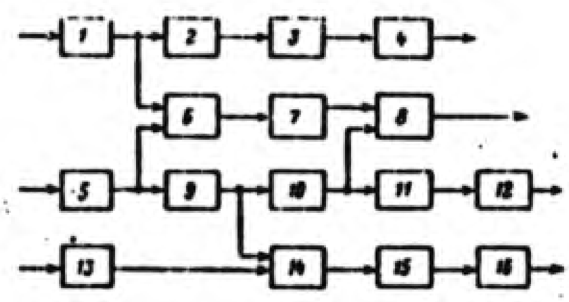


Figure 6.12. The functional circuit of the system

For convenience in solving the problem, this system is represented in the form of a graph in Figure 6.13a. Each vertex of the graph is marked with the precedence index corresponding to it. The set of output vertexes of the graph $(x_4, x_8, x_{12}, x_{16})$ corresponds to the system outputs being monitored.

In accordance with Expression (6.7), let us determine the preference function for each output of the system:

$$F_{в}(4)\approx 0{,}25;\quad F_{в}(8)\approx 0{,}43;$$
$$F_{в}(12)\approx 0{,}31;\quad F_{в}(16)\approx 0{,}37.$$

Since

$$F_{в}(8)>F_{в}(16)>F_{в}(12)>F_{в}(4),$$

then the check of output x_8 must be performed first in the sequence of checks — that is, check π_8 is first (Figure 6.14). Check $\pi 8$ separates out the subset $X_1=\{x_1,\ x_5,\ x_6,\ x_7,\ x_8,\ x_9,\ x_{10}\}$, which is monitored by this check from the entire set of the system elements.

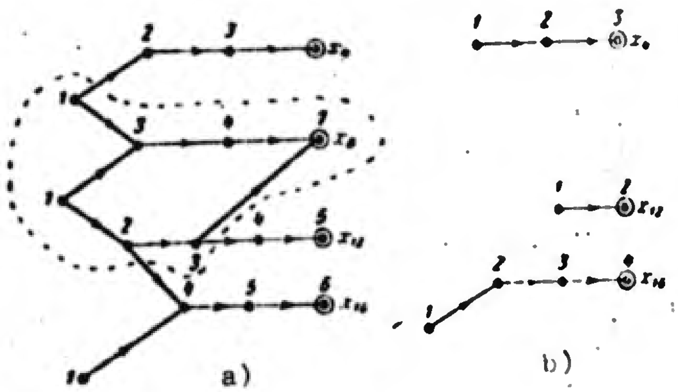


Figure 6.13. An oriented graph of the system shown in Figure 6.12 (a) and its subdivision into non-intersecting subsets (b)

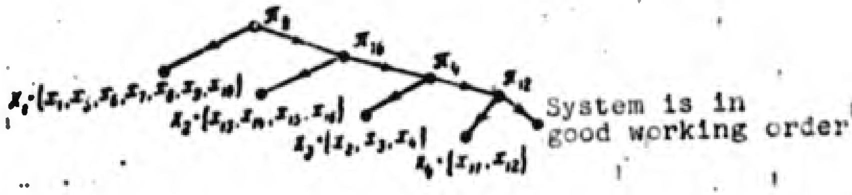


Figure 6.14. The program of monitoring the operating efficiency for the system shown in Figure 6.12

The subset of elements monitored by check π_8 is indicated in Figure 6.13a by the dashed line. The remaining part of the system elements which is not monitored by this check is shown in Figure 6.13b. New precedence indexes for the remaining part of the system are shown in the same figure. Since, in the remaining part of the system, the subsets of elements whose states are monitored at the outputs x_4, x_{12}, and x_{16} are disjoint, further calculations of the preference function are not performed. For the outputs x_4, x_{12}, and x_{16}, we have

$$F_\pi(16)>F_\pi(4)>F_\pi(12).$$

As a result, we obtain the program for monitoring the operating efficiency of this system which is shown in Figure 6.14. The average time for monitoring the operating efficiency on the basis of this program is determined from Expression (6.3). For the given program, this time equals

$$T(\pi_8, \pi_{14}, \pi_4, \pi_{12}) = t\left[4P(l)+\sum_{i=1}^{4} P(X_i)+\sum_{i=2}^{4} P(X_i)+\right.$$
$$\left.+\sum_{i=3}^{4} P(X_i)+P(X_4)\right]=7.2 \text{ min.}$$

The elements included in the set $X_i (i=1, 2, 3, 4)$, are indicated in Figure 6.14.

This method may also be used when constructing programs for monitoring the operating efficiency when the failure probabilities of the system elements and the times for performing each check are different.

As the preference function in this case, it is convenient to use the ratio of the total probability $\sum_{i\in h} P(A_i/X)$ to the time $t_k (k=1,2,\ldots,m)$, which is necessary for monitoring the output ξ, that is

$$F_{n}(l_{\xi})=\frac{\sum\limits_{j\leq l_{\xi}} P(A_{j}/X)}{l_{\xi}}. \tag{6.10}$$

In Expression (6.10), the sum $\sum\limits_{j\leq l_{\xi}} P(A_{j}/X)$ means that the probabilities $P(A_j/X)$ of those elements which are monitored at the output ξ are summed.

In view of its simplicity, the preference method may find broad applications in practice. However, a program constructed on the basis of this method may differ considerably from the optimum program.

6.4. Periodicity of Monitoring a System Operating Efficiency

When operating radar equipment, correct monitoring of its operating efficiency is very important. It is especially important when the system operates for a long period of time — that is, when the time required for performing its tasks is long.

As a rule, the majority of systems or certain of their devices may not be used while the monitoring procedure is being performed. For example, when monitoring a receiver noise factor by means of an ordinary noise generator, the high voltage to the radar station transmitter must be removed. This means that too frequent monitoring leads to a reduction of the time during which the system is ready for action.

On the other hand, if the system is rarely monitored, there will be no assurance that it will operate efficiently in the future. Moreover, frequent monitoring leads to an increase of the operational expenditures connected with the wear of the monitoring equipment and with an increase of its failure rate.

Consequently there are a number of factors which are at variance with the factors affecting the reliability and other operational characteristics of the system. These factors also determine to a considerable degree the effectiveness of the selected monitoring prodedure. The problem of selecting an effective monitoring procedure is investigated in the literature on reliability [40].

When solving this problem, it is assumed that the time for performing the monitoring procedure is constant. Here we will investigate one of the characteristic cases of planning the monitoring procedure, when the cost of performing it is a variable quantity and depends on the monitoring periodicity. It is assumed that, when the monitoring is carried out, absolute reliability of each check is achieved and failures of the monitoring equipment may be disregarded.

With a constant monitoring period T_{pm}, the system is checked during the time $T_B \ll T_{pm}$. The probability that the system is operating efficiently decreases exponentially. After the system has been monitored and it is established that it is in good working order (or it has been repaired if a malfunction were detected), the probability that the system is efficiently operating again increases to one (Figure 6.15). The time T_B is the time for monitoring and restoring the system, and has the following components:

$$T_{\text{в}} = T_{\text{к}} + T_{\text{п}} + T_{\text{у}}. \qquad (6.11)$$

where $T_{\text{к}} = T(\Pi;\ Z)$ — is the average time for monitoring the system with the assigned monitoring program;

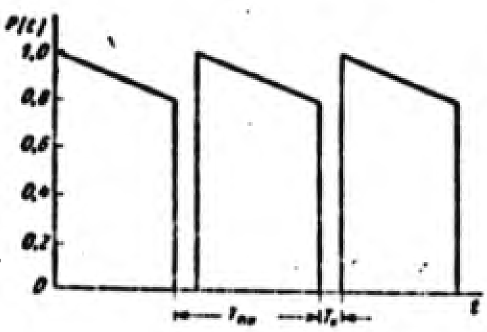


Figure 6.15. The probability of trouble-free operation as a function of the periodicity of monitoring a system's operating efficiency

T_s — is the average time for searching out a malfunction in the system with the given search program;

T_y — is the average time for eliminating the malfunction.

Let us assume that, when the monitoring procedure is performed, monitoring instruments and outfits are used whose operating costs per unit of time can be takes as c_1 rubles. In the case of monitoring and restoring a system on the basis of the assigned program, the average cost of each monitoring and restoration procedure will equal

$$C_1 = c_1 T_{\text{в}}. \tag{6.12}$$

Let us illustrate how the average monitoring time of a system changes with a change of the monitoring periodicity. In order to do this, let us use Expression (6.2) which was presented in §6.2. In accordance with the description of the monitoring process which was investigated earlier (see Figure 6.15), the probability that the system is operating efficiently at the moment when the monitoring procedure is performed will equal $P(t) = e^{-\lambda T_{pm}}$. Let us substitute the value of $P(t)$ into Expression (6.3) taking the fact into account that the sequence of performing the checks is determined, and after this they are numbered in the order they are performed, that is

$$T_{\text{к}} = T(\pi_1, \pi_2, \ldots, \pi_h; Z) = \sum_{j=1}^{h} t_j [e^{-\lambda T_{pm}} + P(X_j/X) \times \\ \times (1 - e^{-\lambda T_{pm}})]. \tag{6.13}$$

where

$$P(X_j/X) = \sum_{\pi_i \in X_j} P(A_i/X).$$

It is convenient to write Expression (6.13) in the following way:

$$T_{\text{к}}=\sum_{j=1}^{h} t_j P(X_j/X)+e^{-\lambda T_{\text{пм}}}\left[\sum_{j=1}^{h} t_j-\sum_{j=1}^{h} t_j P(X_j/X)\right]. \tag{6.14}$$

One may determine the limits in which the quantity T_K changes from Expression (6.14). Actually, if the system in continuously monitored out, then $T_{pm} = 0$ and

$$T_{\text{к}}=T_{\text{к макс}}=\sum_{j=1}^{h} t_j, \tag{6.15}$$

that is, the monitoring time is a maximum, and does not depend on the sequence of performing the checks, since the probability of the system's failure is Q (t) = 0, and it is necessary to conduct all h checks. When the system is monitored after too long an interval of time, the limiting case $T_{pm} = \infty$, then

$$T_{\text{к}}=T_{\text{к мин}}=\sum_{j=1}^{h} t_j P(X_j/X). \tag{6.16}$$

Thus considering (6.15) and (6.16), we arrive at the following form of writing expression (6.14):

$$T_{\text{к}}=T_{\text{к мин}}+e^{-\lambda T_{\text{пм}}}(T_{\text{к макс}}-T_{\text{к мин}}). \tag{6.17}$$

Let us designate by C_2 the cost of the losses from an undetected malfunction which arises between two consecutive monitoring procedures.

Let us investigate how the monitoring periodicity affects the cost C_2. This may be understood from the following considerations. The time of the defective state in the interval between two consecutive monitoring procedures depends on the moment of the system's failure. When the system fails at the moment of time $t(0 \leq t \leq T_{pm})$, the time during which the system is out of order equals $(T_{pm} - t)$.

The probability of the appearance of a malfunction at the moment of time t equals f (t) dt and, consequently, the average time of the defective state may be determined from the expression

$$T_{\text{нс}} = \int_0^{T_{\text{пм}}} (T_{\text{пм}} - t) f(t)\, dt = \int_0^{T_{\text{пм}}} Q(t)\, dt. \tag{6.18}$$

Assuming that the probability density function of the system failures is subject to an exponential law — that is $Q(t) = 1 - e^{-\lambda t}$ — we have from (6.18)

$$T_{\text{нс}} = T_{\text{пм}} - \frac{1}{\lambda}\left(1 - e^{-\lambda T_{\text{пм}}}\right). \tag{6.19}$$

Considering that, if a system is in a defective state for one hour its cost to us is c_2 rubles, we find

$$C_2 = c_2 T_{\text{нс}}. \tag{6.20}$$

Figure 6.16 presents the quantities C_1 and C_2 as a function of the monitoring period T_{pm}. In this figure, the value of the monitoring period $T_{pm\ eff}$ corresponds to a balance of the cost of the monitoring and restoration procedure (C_1) and the cost of the losses (C_2). It is clear that the value of the quantity $T_{pm\ eff}$ corresponds, for the given conditions, to the most efficient monitoring periodicity.

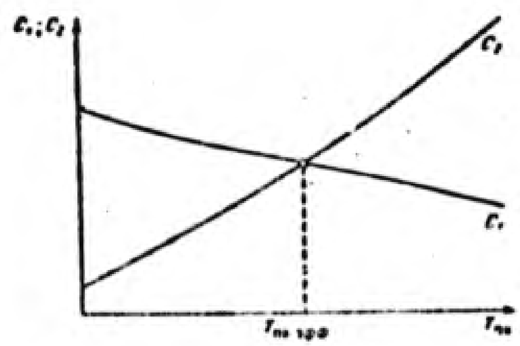


Figure 6.16. The costs C_1 and C_2 as a function of the monitoring periodicity

As an example, let us determine the effective monitoring periodicity of the system investigated in §6.1. Let us assume that a system remaining in a defective state leads to a loss of $c_2 = 6{,}000$ rubles/hour. The failure rate of the system is $\lambda = \Sigma\lambda_i = 4 \cdot 10^{-3}$ hour^{-1}. In order to monitor the system, the program constructed in §6.2 is used. With this program, the average time for locating and eliminating system malfunctions, which are detected when it is monitored, equals 0.5 — that is $T_s + T_y = 0.5$.

From Relationships (6.15) and (6.16), we obtain

$$T_{K\ min} \approx 3 \text{ min } (0.05 \text{ hour}) \text{ and } T_{K\ max} \approx 6 \text{ min } (0.1 \text{ hour}).$$

By assigning various values to T_{pm}, let us graphically solve the equations for C_1 and C_2. When this is done, we obtain $T_{pm\ eff} \approx$ 1 hour. Thus the system's monitoring should be carried out periodically after every hour of its operation.

6.5. Monitoring the Change of the Equipment's State

Reliability, as was already indicated before, is characterized by the absence of sudden or gradual failures. The problem of maintaining the reliability of radar systems with sudden failures was investigated in the preceding chapters. Let us discuss the problem of maintaining the reliability with gradual failures.

Gradual failures are caused by relatively slow change of the physical-chemical structure of elements as a consequence of their wear and aging, and the change of their properties from the effects of heat, cold, humidity, and other factors. As a result of these conditions, the system may cease to function or its parameters may exceed the limits of the norm.

Under actual conditions, both sudden and gradual failures of a system occur.

The probability of trouble-free operation of a system under the assumption that these failures are independent is determined by the expression

$$P(t) = P_{вн}(t) P_{п}(t),$$

where $P_{su}(t)$ — is the probability of trouble-free operation of the equipment in respect to sudden failures;

$P_{ga}(t)$ — is the probability of trouble-free operation of the equipment in respect to gradual failures.

The failure rate is the main parameter determining reliability in respect to sudden failures.

The probability of trouble-free operation in respect to gradual failures is defined as the probability of finding its output parameters within the given limits:

$$P_{п}(t) = \prod_{i=1}^{m} P_i(v_{i\,мин} < v_i(t) < v_{i\,макс}),$$

where m — is the number of output parameters which characterize the equipment operation;

$v_{i\,min}$, $v_{i\,max}$ — are the lower and upper limits of the i^{th} parameter.

The relationship between the probability of the appearance of gradual and sudden failures for certain elements is presented in Table 6.1. As follows from the table, for a number of elements gradual failures are a considerable part of all the failures. Moreover, it is necessary to keep in mind that failures which occur because of the absence of information concerning gradual changes in the state of the system elements very frequently are referred to as sudden failures. One should expect that the fraction specified as gradual failures will increase in proportion to the increase of warning information concerning the state of the system. When designing and operating equipment, one must take measures for increasing

TABLE 6.1. PROBABILITIES OF THE APPEARANCE OF GRADUAL AND SUDDEN FAILURES

Elements	Failures, in %	
	Gradual	Sudden
Electric vacuum devices	70 - 80	20 - 30
Semiconductor instruments	70 - 80	20 - 30
Transformers, relay	50 - 60	40 - 50
Selenium rectifiers	70 - 80	20 - 30
Resistors	20 - 30	70 - 80
Capacitors	7 - 10	90 - 93

the probability of trouble-free operation by reducing both sudden and gradual failures.

One of the effective methods for increasing the reliability is predicting the equipment failures, and — on the basis of the resulting information — carrying out preventive maintenance measures for replacing and restoring elements which have deteriorated parameters.

In principle, two methods of prediction are possible: the instrumental method and the statistical method. The instrumental prediction method may be applied in those cases when the gradual changes of the physical-chemical structure of elements may in some way be monitored. When this is impossible, statistical prediction methods are used. The instrumental method of predicting failures is based on the assumption that a period of gradual change in the parameters of the system or element precedes the emergence of failures, and that by recognizing the signs of an approaching failure one may predict the moment when it will occur. As a consequence of this, the instrumental prediction method may be used only for gradual failures, when the parameter used for the prediction is known. By a failure in this case, we mean that a system ceases to function or its parameters exceed the limits of the norm.

However, gradual failures of elements for which no predicting parameter is known and sudden failures in a number of cases also may be prevented if the statistical law for the distribution of the malfunction appearance time is known. Instrumental predicting may be accomplished when the equipment is operated, and also during its design. In the first case, the prediction reveals the projected failures and makes it possible to replace the elements with deteriorating parameters in time. The prediction is done in this case to reduce the number of equipment failures and, consequently, to increase the probability of its trouble-free operation during a given interval of time. In the second case, the prediction is made to select equipment operating conditions which will ensure a specific reliability margin.

Forecasting a change in the equipment's state in the future — predictions — is based on the theory of extrapolation — that is, extending conclusions obtained from observations of one part of the phenomenon to a future time. When this is done, naturally, it is assumed that the future state of the equipment depends on its state in the preceding operational period.

In the majority of cases, this assumption is valid, since — despite all the varied laws governing the change of an element's parameters with time — a characteristic feature of them is the monotonic nature of the change and the absence, in the overwhelming majority of cases, of bending points in the prediction region (Figures 6.17 and 6.18).

Obviously the prediction will be more precise, the more complete is the information on the equipment's preceding operational period. Either statistical data accumulated during the equipment's preceding operational period or data obtained as a result of specially conducted measurements may be used as such information.

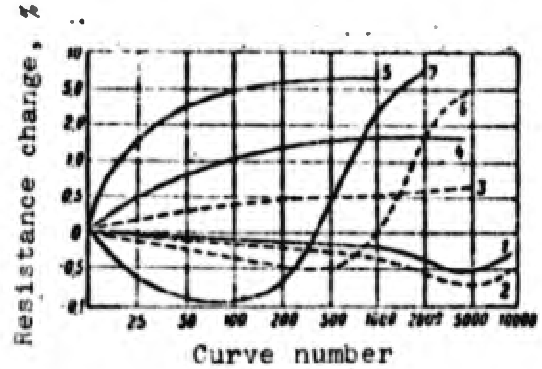


Figure 6.17. Graphs of the aging of type VS-0.5 resistors when stored in a warehouse (curves 1 and 2) and when operated under field conditions (curves 3 - 7):

Curve number	1	2	3	4	5	6	7
Rated resistance, megohm	0.5	3.0	0.5	10.0	10.0	0.5	0.5

Curves 6 and 7 characterize the aging of resistors with a load which is 0.1 and 2.5 of the rated load and a temperature of 150 and 70° C, respectively

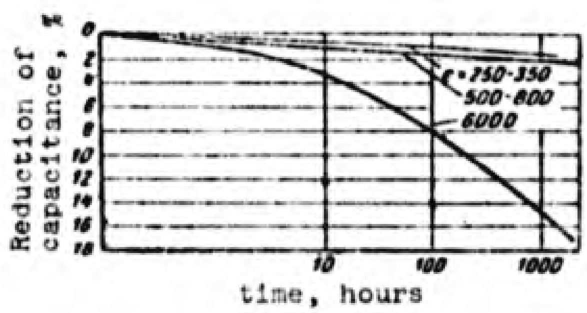


Figure 6.18. Change of the capacitance of ceramic capacitors as a function of time

The prediction problem may be formulated in the following way. Let the parameter being monitored be a function of time $\psi(t)$. Figure 6.19 shows the change in the monitored parameter with time. It is known that the function $\psi(t)$ during a certain length of time T_1 takes the values

$\psi(t_0), \psi(t_1), \ldots, \psi(t_{n-1})$

at the moments of time $t_0, t_1, \ldots, t_n$. It is necessary, on the basis of the values of $\psi(t_n), \psi(t_{n-1}), \ldots, \psi(t_0)$ for the function $\psi(t)$, to forecast the values $\psi(t_{n+m})$ with $m > 0$ for the times $t_{n+1}, t_{n+2}, \ldots, t_{n+m}$ which belong to the length of time T_2.

Various methods may be used for solving the prediction problem, which differ in the mathematical procedure used. The methods for predicting a change in the state of complicated systems may be divided into two groups. The first group uses the mathematical device of numerical analysis, and the second uses the device of the theory of random functions [31].

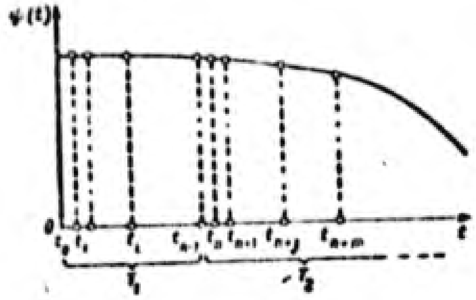


Figure 6.19. Curve showing the change of the parameter being monitored as a function of time

The essence of predicting failures consists of the following. Let us assume that for an element or system the failure level is known, as well as the dependence of the parameter change on time.

The parameter change as a function of the operating time may be obtained, for example, from the results of equipment tests during its service life, from the results of observing the parameter changes during a specified time period, or from the results of extrapolating, by one or another method, the results obtained to the following period. The parameter value at which the device, when this value is reached, is not able to perform its function is taken as the failure level. For an element (tube, resistor, etc.), the failure level may be the tolerance limit based on the technical specifications, for the limiting value of the parameter which causes the device containing this element to stop functioning.

For a system, the failure level should be the allowance for output parameter changes. Figure 6.20 shows the parameter ν as a function of time and also the failure level ν_{fail} which corresponds to the time of failure t_{fail}. The problem is to determine the maximum period of preventive maintenance checks. This period must be selected on the condition that the system, which is in good working order at the time of the preventive check, will not, with a specified probability, break down before the time of the next check.

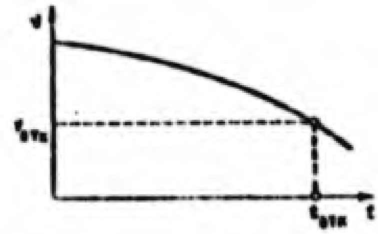


Figure 6.20. Change of the output parameter as a function of time

This period is designated the forecast period T_{for}, and the value of the parameter corresponding to this moment of time is called the forecast level ν_{for}. Thus the prediction problem is to determine — on the basis of the information obtained concerning the change of the system's parameter — the maximum period of preventive checks which will guarantee the specified forecast reliability.

One of the main problems which arises from a prediction is the optimum selection of the parameters on the basis of which it is advisable to perform the monitoring. Modern electronic systems are so complicated that predicting the state of individual elements (resistors, capacitors, etc.) is inadvisable, and sometimes impossible, because when this is done the complexity of the monitoring system increases considerably. It is advisable, for forecasting failures, to select the minimum number of parameters which will enable one to obtain maximum information on the state of the entire system or even of an individual stage or unit. It is desirable that these parameters most completely reflect the changes taking place in these systems.

Instrumental prediction of failures may be done on the basis of changes in the output parameter, the controlling parameter, or a specially selected auxiliary (forecasting) parameter. Let us discuss in somewhat more detail this classification of parameters.

An output parameter most comprehensively characterizes the state of a circuit or stage from the point of view of performing its assigned functions, and it conveys the maximum information on its state. The voltage at the circuit output is such a parameter, for example.

A controlling parameter most completely characterizes the state of an element which is mounted in a specific circuit. Thus, for example, either the characteristic curve or the value of the anode current of an electronic tube may be the controlling parameter of an electronic tube, depending on the circuit in which it is mounted.

An auxiliary parameter indirectly characterizes the qualitative state of an element. The functional relations between an auxiliary parameter and the controlling parameter enable one to use the former for predicting an element's state. Thus, for example, an increase in the internal noise of certain types of electronic tubes precedes a change in their characteristic curve. The tube internal noise in this case is an auxiliary (predicting) parameter.

Two methods for predicting the moment when a failure will arise have found the broadest application:

1) based on a change of the controlling (predicting, output) parameter with periodic measurement of its value during operation;

2) by changing the system operating conditions.

Since the largest number of gradual equipment failures occur as a result of failures of electric vacuum devices, prediction methods are developed most frequently for them.

Let us investigate the methods for predicting the moment when a failure will arise.

Predicting Failures from a Change in the Controlling Parameter

Predicting failures from a change in the controlling parameter is used mainly for checking electric vacuum devices. A necessary condition for making predictions in this way is knowing the failure levels for all stages of the equipment, and the laws governing the change of the electron tube parameters. The forecast level is selected from the condition that the tube — which is in good working order at the time of the preventive check -- will not break down before the next check.

Figure 6.21 shows the failure level and the forecast level and also the dependence of the controlling parameter on the operation time.

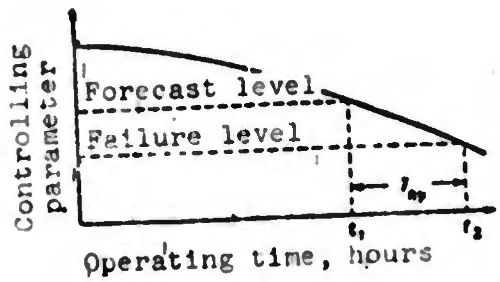


Figure 6.21. Predicting failures on the basis of the controlling parameter

In order to decrease the number of checks, it is advisable to increase the forecast period, but when this is done the tube service life is shortened. Usually the periodicity of preventive checks has been specified, and it determines the forecast level.

A high forecast reliability without significant shortening of the service life is the main advantage of this method. The defects of the method include the difficulties of determining the failure levels of tubes, and the laws governing the change of the parameters. This is because in order to measure the controlling parameters, in the majority of cases, the tubes must be removed from the equipment. Moreover, the failure level for tubes of one type as a rule is different when the tubes are mounted in different circuits. Therefore, tests must be made for each circuit, which naturally is complicated. Therefore, this prediction method usually amounts to periodically checking the extent to which the tubes comply with the norms of the particular technical specifications, using for this tube testers. However, such a checking method has a number of significant disadvantages, namely:

— the measurement of the electron tube parameters is made under conditions which differ from the actual conditions, since the tubes are removed from the circuit;

— in circuits there are stages which are more critical, and others which are less critical, to a change in the tube parameters. Checking tubes on tube testers does not allow one to take this into account. As a result, tubes are removed from stages which are not critical to a change in the parameters, and these tubes may still be used in the equipment, and vice versa. Operational experience has proven that from 40 to 80% of the electron tubes — which checking indicated did not satisfy the particular technical specifications — actually provide normal functioning of the equipment;

— removing tubes from the equipment and then again re-installing them leads to mechanical damage of the tubes and circuit boards, and as a consequence to equipment failures;

— removing tubes from the equipment for the purpose of monitoring reduces the operational ability of the equipment;

— mass replacement of tubes may lead to the opposite result — a reduction of the equipment reliability because the probability of sudden tube failures, which take a specified time, is lower than for the tubes which have been re-installed.

As follows from operational experience, such a method of preventive maintenance is advisable for monitoring electron tubes which have been stored in warehouses and are included in the reserve stock.

Predicting Failures on the Basis of Output Parameter Changes

The monitoring is conducted, not with respect to the controlling parameter of the element (tube, transistor, etc.), but on the basis of the output parameter of the stage, assembly, or equipment. A necessary condition for making a prediction in this case is knowing the correlation between the equipment output parameter and the quality index of the elements. The prediction technique is similar to the procedure discussed above.

An advantage of such a prediction method is the possibility of checking the equipment as a whole. It opens up broad possibilities for automating the monitoring operations by using systems for automatically monitoring the parameters.

A defect of this method is the complexity and great amout of work required in obtaining the dependencies of the output parameter changes on time and in determinging the failure levels. It is also necessary to know the dependence of the output parameter on the controlling parameters. It is advisable to use this method for making predictions only for relatively uncomplicated equipment.

Predicting Failures on the Basis of Auxiliary Parameter Changes

For each element, in particular for an electron tube, there is an auxiliary parameter, in which a noticeable change indicates a change in the basic parameters.

In a number of cases, making a prediction on the basis of an auxiliary parameter may turn out to be more convenient than on the basis of a basic parameter. Thus, for example, the qualitative state of electron tubes may be determined from their noise parameters. From theory, it is known that current fluctuations increase with a decrease of the emission current. As follows from experimental data, an increase of the tube noise usually precedes the deterioration of parameters such as the characteristic curve and the anode current. Therefore, unavoidable deterioration of the tube cathode emission properties may be determined approximately 100 hours in advance.

Figure 6.22 shows the change of the controlling parameter and auxiliary parameter as a function of the operating time which illustrates this prediction method. If, at the time t_1, an abrupt change of the auxiliary parameter is detected, this indicates that the controlling parameter will pass beyond the acceptable limits at the time t_2.

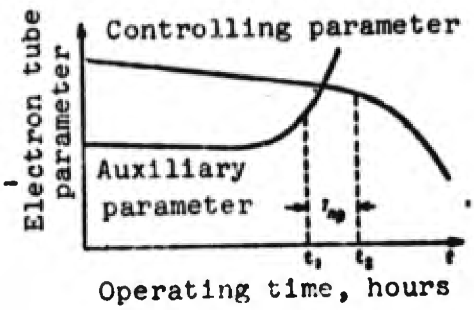


Figure 6.22. Predicting failures on the basis of an auxiliary parameter

Predicting Failures by Changing the Element Operating Conditions

This method consists of creating overloaded or eased operating conditions of elements in order to simulate their aging or wearing out. This enables one to approximate the element failure.

Such operating conditions may be most simply created by changing the voltage on the electrodes of tubes and semiconductor devices. In general, the operating conditions are changed in such a way that the rate of change of the element parameter is increased, and it reaches the failure level earlier than would occur under normal operating conditions (Figure 6.23).

Predicting failures by such a method gives acceptable results for electric vacuum devices. When this method is used, changes of the voltages on the electrodes must reproduce the physical phenomena occurring when the electron tubes age.

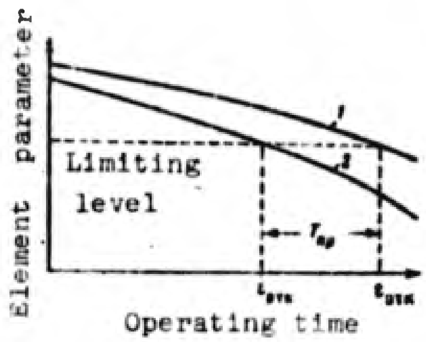


Figure 6.23. Predicting failures by changing the operating conditions of elements. Curve 1 characterizes the operation of an element at nominal voltage; curve 2 is with changed voltage

For receiver-amplifier tubes, a reduction of the filament voltage most nearly satisfies these requirements. This is explained by the following reasons.

1. The nature of the change in filament characteristics of electron tubes as a consequence of aging is the same as with underheating. It is connected with deterioration of the cathode emission properties The filament characteristics of tubes when they are operated for a long time are shifted in the direction of smaller values of the anode current (Figure 6.24). From this it follows that, with a reduction of the filament voltage, the change of the anode current of new tubes takes place considerably more slowly than for old tubes. Thus for example, the function $I_a = f(U_H)$ for old and new tubes of the 6Zh4 type has the form shown in Figure 6.25. Thus a decrease of the filament voltage enables one to evaluate the cathode emission properties which have an influence on the electron tube reliability and to distinguish, with a sufficient degree of precision for practical purposes, an old tube from a new tube.

2. A change of the filament voltage leads to a sharp change of the anode current and characteristic curve, and therefore the filament voltage is the most convenient predicting parameter.

3. The filament voltage may be reduced by very simple technical means.

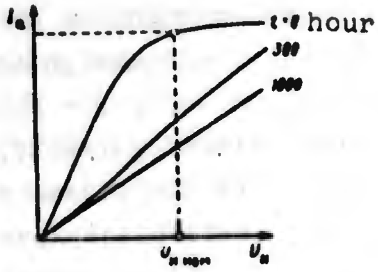


Figure 6.24. Change of a tube filament characteristics during operation

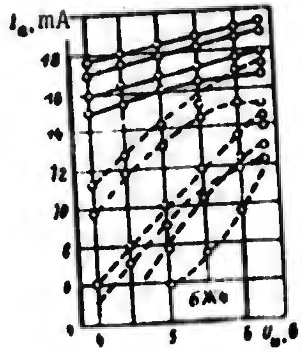


Figure 6.25. The anode current as a function of the filament voltage for old (dashed lines) and new (continuous lines) tubes of the 6Zh4 type

The prediction is accomplished in the following way. A reduced filament voltage is fed to the tube or group of tubes, and the anode current or characteristic curve of the tubes is measured.

Measurement of the anode current is the preferred method, since its rate of change with a change of the filament voltage is larger than the rate of change of the characteristic curve. The magnitude of the filament voltage reduction is determined by the character of the circuit for which the prediction is being made, and is found to be in the limits of 5 - 20% of the rated value of the tube filament voltage. The conditions may be selected in such a way that, when the system is underheated, the failure level is reached earlier by a length of time equal to the period of the preventive maintenance checks (Figure 6.23). In this case, if at the time of checking, the value of the parameter is higher than the failure level, then one may assume that the system will not break down during the time between two checks. If the parameter value is below the indicated level, then failure is possible, and measures must be taken to detect and eliminate the elements having limiting parameters. The period of the preventive maintenance checks usually is selected as equal to 100 - 150 hours.

The possibility of predicting tube failures by changing the filament voltage must be provided for in the equipment design stage. Since the individual stages, groups of stages, and units for which the prediction is being made may have different limiting levels, they must be combined on the basis of the limits of the filament voltage change (which are necessary for making a reliability prediction) which they have in common. In this case if the predicting was not provided when the equipment was being designed, it may be performed (as a consequence of the simplicity of the method) in the operating equipment. Reducting of the filament voltage may be achieved by inserting a resistance in the primary or secondary circuit of the filament transformer (Figure 6.26). Including a resistance in the primary circuit of the filament transformer naturally

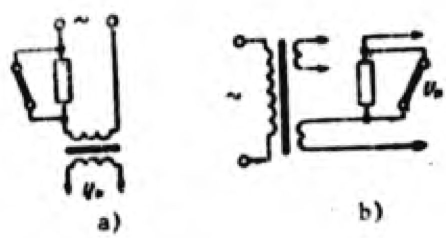


Figure 6.26. Circuits for supplying the filament voltage when predicting failures by changing the operating conditions of electric vacuum devices

may be used in the case when the filament transformer and the anode transformer are not structurally combined. The circuit shown in Figure 6.26,a may be used if the prediction is specified when the equipment is designed, and the circuit of Figure 6.26,b is convenient for achieving the prediction in operating equipment.

The prediction for pulsed circuits has certain distinctive features which are caused by the fact that, in the majority of such circuits, the amplitude and shape of the output pulses as a rule does not depend on the input signals. The purpose of a preventive maintenance check in such circuits is not to determine the tube with the worst parameter, but to recognize the tube which, with a reduction of the filament voltage, does not ensure the operation of the following stage.

It is impossible to use this prediction method for checking kenetrons, since the dependence of the rectified voltage on the filament voltage change is very weakly expressed for both old and new kenetrons.

The advantage of the method for predicting failures by changing the element operating conditions is its simplicity. The defects of this method are that it does not allow one to conduct preventive maintenance with normal operation of the equipment and, moreover, it creates the danger of irreversible changes in the structure of elements because of the deviation of their operating conditions from the rated conditions during the test process.

It is necessary to point out that predicting failures enables one, in a number of cases, to obtain information on the state of a system or element. On the basis of this information, one may determine the period of preventive maintenance tasks for which the system should have the maximum possibility of being in an operating state at any arbitrarily selected moment of time. However, the extent, periodicity, and order of conducting the preventive maintenance tasks depend on many factors which are not monitored. It is not often that one is able to completely take into account these unmonitored factors. An increase in the extent of the preventive maintenance tasks leads to an increase in the time expended in performing them which is economically extremely unprofitable. Moreover, the performance of the preventive maintenance task itself frequently causes the appearance of failures in the equipment, and may not only not increase the equipment reliability, but on the contrary may decrease it.

Thus, with an increase in the extent of the preventive maintenance tasks, the number of induced failures increases, but with a reduction of the extent of checks and a decrease of their number, the number of gradual failures increases. Consequently, when resolving the problem of the extent and periodicity of preventive maintenance tasks, a reasonable compromise is necessary which enables one to ensure a sufficiently high reliability of equipment with a minimum of standby time.

6.6. Application of the Boundary Test Method for Monitoring Equipment

Increasing the reliability of electronic equipment is an important engineering problem. It encompasses a broad range of problems connected with the design, manufacture and operation of equipment. Among them, increasing the equipment reliability in the design stage is of very great importance.

Reliability is a unique reserve of durability which must be provided for when designing the equipment.

In the equipment design process, the designer must deal with elements which are mass-produced by industry and have definite parameters, a guaranteed lifetime, and a precision class. The parameters of all elements, especially electron tubes, change with time. The aging of other elements also occurs (see Figure 5.17 and 6.18). Since — when a station is mass-produced — the elements from which it is assembled have various deviations of their parameters from the rated value within the limits of the precision class, the circuit must be developed in such a way that the element parameter scatter in the radar does not disturb its operation. Aging of the elements is a slowly-changing irreversible process which occurs in the material of elements as a consequence of complex physical-chemical phenomena. The parameters of elements and their reliability also change when they operate under conditions which differ from normal.

As a consequence of this, a large scatter in the failure rates is observed even for elements of the same type (Table 6.2).

TABLE 6.2. FAILURE RATES OF CERTAIN ELEMENTS

Kind of element	Failure rate, $\lambda_i \cdot 10^5$ hours
Electric vacuum devices	5.9 - 20.2
Resistors	0.25 - 2.2
Capacitors	0.11 - 0.41
Transformers	0.83 - 5.6
Relays	2.0 - 8.3
Selenium rectifiers	3.8 - 27.8
Electric motors	4.2 - 12.5

The reason for significant deviations in the failure rates mainly is the difference in the thermal and electrical operating conditions of the equipment elements. The circuit must be designed so that aging of the elements and changes of their parameters, because of the effects of their surroundings, does not disturb its operation — that is, the circuit must be stable toward changes of the element parameters which occur during the equipment operation. Efficiently developed equipment must have circuit operating conditions which are selected in accordance with the requirements imposed on its reliability. The specified technical parameters of the equipment may be obtained with various operating conditions of the circuit's elements. A different reliability reserve for the equipment corresponds to each of these operating conditions.

In the design process one must find a reasonable compromise between the requirements imposed on the equipment technical parameters and those imposed on its reliability.

It is necessary to point out that very frequently, before the problem of reliability is brought up, the equipment specified technical parameters are arrived at when making the design without consideration of its reliability, and sometimes even to the detriment of the reliability. The complex interrelationship of phenomena which occurs even in simple equipment does not allow one to demand of the designer that optimum operating conditions be achieved in the equipment in terms of reliability. Therefore, the equipment reliability may be increased if it is monitored in the design stage.

The use of analytical methods for studying the equipment reliability causes considerable difficulties. The statistical methods for monitoring the reliability which are being widely used enable one to determine only whether the reliability corresponds or does not correspond to the imposed requirements. When these methods are used, the factors determining the equipment reliability are not discovered, and consequently the possibility is eliminated of making purposeful efforts to increase the reliability.

The use of simulation methods for studying reliability has considerably greater possibilities: these simulation methods include statistical, boundary and matrix tests [13]. However, the complexity of the equipment and the difficulty of processing the results when conducting statistical tests, and the extremely time consuming nature of matrix tests even for relatively simple circuits, make it impossible to use these methods when the equipment is operating, and they hinder their use during the equipment design process. Therefore, let us discuss only the method of boundary tests.

Boundary Tests of Equipment

The method of boundary tests is based on imitating a change in the element parameters and other input parameters for the purpose of determining the changes of the parameters at the circuit output.

In principle, analytical, experimental, and graphical-analytical solution of this problem is possible.

The analytical method requires knowing the dependence of the system output parameters on the input parameters. Let us assume that the system output parameters are a monotonic func ion of many variables (input parameters).

In the general case, we have

$$\nu_i = f(\varphi_1, \varphi_2, \ldots, \varphi_j, \ldots, \varphi_n),$$

where ν_i — is an output parameter of the equipment;

ϕ_j — is an input parameter (the parameter of an element in the equipment; an input signal; a factor affecting the operation of an element).

Thus, for example, for a blocking generator (Figure 6.27) the output parameters are the amplitude, duration, and repetition frequency of the pulses, and the input parameters are the resistance in the anode circuit R_a, the capacitance of capacitor C_1, the resistance of R_1 and R_k, the voltage of the power supply source E_a, the ambient temperature, etc. Let us assume that the device operates trouble-free if the output parameter ν_i is within acceptable limits — that is, the following inequality is fulfilled

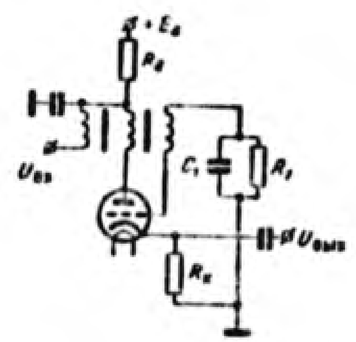


Figure 6.27. Circuit of a blocking generator

$$a_i \leq \nu_i \leq b_i.$$

The change of the output parameter $\Delta\nu_i$ may be determined if one expands the quantity ν_i into a Taylor series in powers of ϕ_j:

$$\Delta\nu_i = \sum_{k=1}^{s} \sum_{j=1}^{n} \frac{1}{k!} \left(\frac{\partial^k \nu_i}{\partial \varphi_j^k} \right) \Delta\varphi_j^k + Q_k, \tag{6.21}$$

where s — is the number of terms in the series,

Q_k — is the remainder term of the series.

For the case when the changes of the input parameters are small and one may limit oneself to the first term of the expansion, Expression (6.21) has the form

$$\Delta\nu_i = \sum_{j=1}^{n} \left(\frac{\partial \nu_i}{\partial \varphi_j} \right) \Delta\varphi_j. \tag{6.22}$$

One may find, on the basis of Equation (6.22), the change of the system output parameters with a change of the input parameters, and consequently one may determine the acceptable values of the input parameters.

For complex electronic systems, it is impossible in practice to obtain Relationship (6.22). As a consequence of this, the analytical method of boundary tests has very limited applications.

The essence of the experimental method of performing boundary tests is to simulate failures on a mock-up of the equipment by artificially changing to specified limits the parameters of the elements, the power supply voltages, the currents, the temperature, etc. As with the analytical method, it is assumed that the system output parameters are a monotonic function of the input parameters. The acceptable change of the input parameter is determined as a function of one of the parameters of the device. This parameter is called the boundary test parameter. The voltage of the power supply source is most frequently taken as the boundary test parameter.

Such an investigation enables one to establish the region within the limits of which the circuit operates trouble-free.

The technique of performing boundary tests may be illustrated with the following example. Let us establish the rated values ϕ_{jH} of the input parameter ϕ_j and of the boundary best parameter U_H. In the graph of Figure 6.28 the point corresponding to the rated values of the parameter is designated H. It is assumed that the values of the remaining input parameters also are within the norm. Let us decrease the boundary test parameter. Failure of the circuit begins with a certain value of this parameter (point 1). Failure of the circuit means either a breakdown of the circuit elements or that state of the circuit when the output parameter goes beyond acceptable limits.

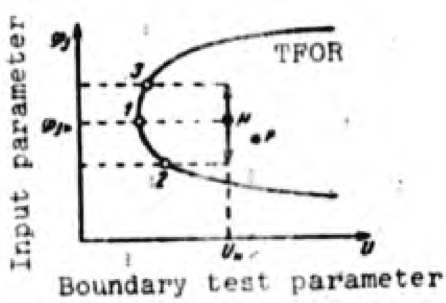


Figure 6.28. Construction of the trouble-free operating region (TFOR)

For example, for a submodulator this may be a distortion of the pulse shape greater than the acceptable value. For a trigger, this may be the situation in which the tirgger pulse is not able to trigger the circuit into a new state; for a holding diode this is the condition when it ceases to hold, etc. Now let us reduce the input parameter and again change the boundary test parameter until we get a new circuit failure. When this is done, we obtain point 2. Then let us increase the value of the input parameter and, by changing the value of the boundary test parameter, we obtain point 3. Continuing the tests in this way, one is able to find the boundary dividing the region of trouble-free operation of the circuit from the region of failures. The circuit trouble-free operating region may be a closed or an open curve, depending on the limits within which the element parameter change at which failure occurs. The operating point P because of manufacturing tolerances and possible deviations of the input parameter from the rated value will not correspond to point H representing the input parameter rated value.

The location of the working point inside the circuit trouble-free operating region characterizes the device reliability.

Actually, if the operating point is at the center of the trouble-free operating region, then the device will respond in a like manner to any change of the element parameter.

Consequently, the location of the operating point in respect to the trouble-free operating region characterizes the reaction of the circuit to a change in the parameter of the elements being studied. Strictly speaking, it is necessary to investigate not the circuit operating point for the parameter being studied, but the operating region.

Let us introduce the concept of an operating region. The operating region designates the probable location of the circuit operating points in respect to the input parameters. This may be, for example, the region formed by deviation of the input parameters due to manufacturing errors of the elements and a change of the power supply voltage, or it may be a region formed by deviation of the triggering pulse voltage and the power supply voltage, etc.

Figure 6.29 shows the trouble-free operating region of a circuit in the coordinate system Φ_i, U and it also shows the operating region. In order to determine the operating region, it is necessary to know the laws governing the distribution of the element production errors for the parameter being studied.

Usually it is assumed that, when manufacturing a batch of elements, the following conditions giving rise to production errors occur:

1. The production error is the sum of components, each of which is the result of the action of a large number of random factors and a certain number of systematic factors.

2. All the random factors are mutually independent and have approximately the same effect on the total error — that is, they are quantities of the same order of smallness.

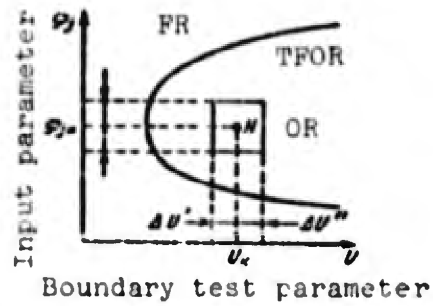


Figure 6.29. Failure region (FR), trouble-free operating region (TFOR), and operating region (OR)

3. All the components of the total error are not correlated. This may occur with an automatically operating plant.

If these conditions are fulfilled, then the distribution of the manufacturing errors is subject to the normal distribution law — that is, to the Gauss law (Figure 3.1). The distribution according to the Gauss law is described by Expression (3.1). In the case when there are systematic dominating factors which cause a fixed displacement of the deviations, the distribution also is subject to the Gauss law, but the center of the clustering is shifted. If the rated value of the part parameter is taken as the origin of the coordinates, and the coordinate of the center of the deviation clustering is designated by r_n, then the distribution curve will be described by the following expression:

$$f(\Delta x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\Delta x-r_n)^2}{2\sigma^2}},$$

where r_n is the mean value of the parameter; σ is the root-mean square deviation.

Under actual conditions, the production errors may be caused by dominating factors which have a variable character. Cases occur when the distribution in general is not subject to the normal distribution law. However, because of the fact that manufacturing of elements is a specialized and mass-produced operation, the effect of subjective factors and dominating factors is smoothed out. Therefore, the distribution of production errors of element (electron tubes, resistors, capacitors) in the majority of cases may be considered, with a sufficient degree of precision, to be subject to a normal law (Figures 6.30 and 6.31).

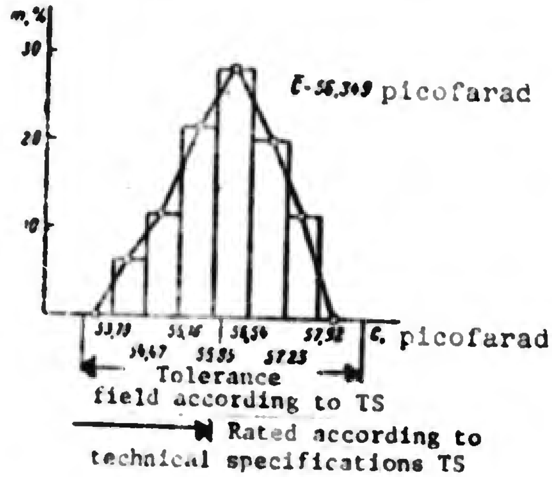


Figure 6.30. Histogram and distribution polygon of the production errors of type KTK capacitors

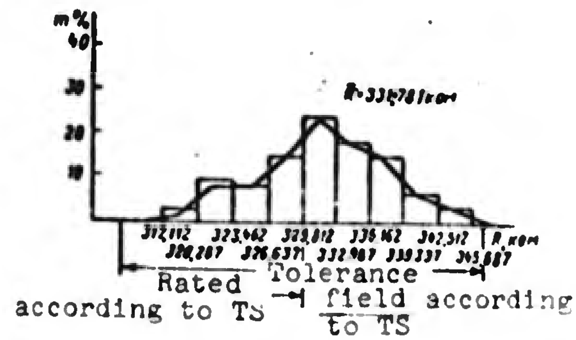


Figure 6.31. Histogram and distribution polygon of the production errors of resistors ULM-0.12 with a resistance of 330 kilohm

Elements of electron equipment are characterized by two parameters: the rated value and the tolerance field. The tolerance determines the maximum acceptable deviation of the given parameter value from the rated value. It must always be larger than the field of scatter of the production errors. The cross section of the operating region — that is, its width (in respect to the parameter being studied — will equal, consequently, the element tolerance field.

During the equipment operation, a drift of the operating region occurs as a consequence of aging of the elements and a change of their parameters because of the effect of temperature, on the one hand, and a change of the power supply voltages, on the other. Figure 6.32 shows the drift of the operating region with a change of the parameters ϕ_j and U.

Approach of the operating region to the boundary separating the circuit trouble-free operating region from the failure region is undesirable, since this may be accompanied by a reduction of the equipment reliability. Passage of the operating region beyond the limits of the trouble-free operating region leads to failure of the equipment. With an equally probable drift direction of the parameters ϕ_j and U, the operating region must be located in the center of the trouble-free operating region. If the drift of the parameters has a dominant direction, then the operating region must be shifted away from the center in a direction opposite to the drift of the parameters ϕ_j and U. The amount of this shift may be determined if the change of the parameters in time is known. The main problem consequently is to select, when the equipment is being designed, those operating conditions at which the equipment is not critical to possible changes of the input parameters.

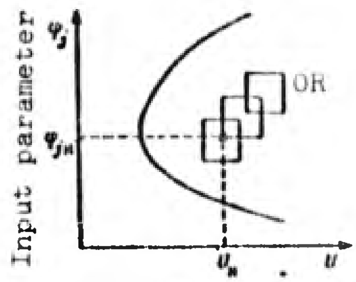


Figure 6.32. Drift of the operating region when input parameters change

The location of the average point of the operating region in respect to the trouble-free operating region must be selected on the basis of reliability conditions and is determined as the sum of three tolerances: the production tolerance, the temperature tolerance, and tolerance toward aging. Since all these quantities have a random character, the summing must be performed in accordance with the summation rules for random quantities. The total tolerance — that is, the operational tolerance — may be determined from the expression

$$\delta\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\Sigma}=\sqrt{\delta^2\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\text{пр}}+\delta^2\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\text{т}}+\delta^2\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\text{ст}}}\,.$$

where $\delta\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\Sigma}$ — is half the operational tolerance field of parameter ϕ_j;

$\delta\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\text{pro}}$ — is half the production tolerance field of parameter ϕ_j;

$\delta\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\text{temp}}$ — is half the temperature tolerance field of parameter ϕ_j;

$\delta\left(\frac{\Delta\varphi_j}{\varphi_j}\right)_{\text{age}}$ — is half the tolerance field of parameter ϕ_j toward aging.

When making the calculation, the time interval during which the circuit element parameter changes by an amount equal to half the tolerance field toward aging, is assumed approximately to be the equipment operating period from the moment of its production up to the first average overhaul, or the time between average overhauls.

The configuration of the trouble-free operating region depends on the device circuit. When the parameters of the circuit elements may vary within broad limits and a change of their operating conditions does not cause damage to these elements, the contour of the circuit trouble-free operating region is a closed curve. The area encompassed by this curve characterizes the circuit operating reliability. Other conditions being equal, that circuit is more reliable for which the area occupied by the trouble-free operating region is larger.

As follows from what has been said, the boundary test method enables one to effectively and clearly examine the reliability of circuits.

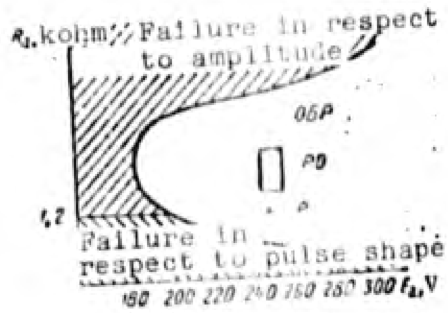


Figure 6.33. Trouble-free operating region (TFOR) of a blocking generator when input parameters R_a and E_a change

As an example, let us investigate the boundary test graph of a blocking generator circuit (Figure 6.27). The power supply voltage is taken as the boundary test parameter, as is usually the case. Figure 6.33 shows the trouble-free operating region of the blocking generator connected with a change of the anode resistance R_a and the voltage of the power supply source E_a. The circuit output parameters are the amplitude of the output pulse U_{out} = 10 V, and its shape. The boundary of the circuit trouble-free operating region corresponds to a change of the output pulse amplitude of ± 20%. This value was selected due to the necessity of obtaining a specific amplitude of the output pulse. Figure 6.33 shows the initial position of the operating point P which corresponds to the value of the input parameters R_{ap} = 1.2 kohm and E_{ap} = 250 V. Type VS-2 resistor R_a of the second precision class has a tolerance of ± 10%. The deviation of the power supply voltage from the rated value does not exceed ± 3%. The operating region consequently is bounded by the limits ± 10% in respect to the anode resistance, and ± 3% in respect to the power supply voltage.

It follows from Figure 6.33 that the selected resistance value of the resistor from the point of view of reliability is not the optimum one. It is advisable to increase the resistance value of the resistor R_a by 20 - 30%. The trouble-free operating region for the same circuit with a change in the amplitude of the triggering (input) pulse and the voltage of the power supply source E_a is shown in Figure 6.34. The boundary of the circuit trouble-free operating region also is obtained from the acceptable limits of the change in the output pulse amplitude of ± 20%.

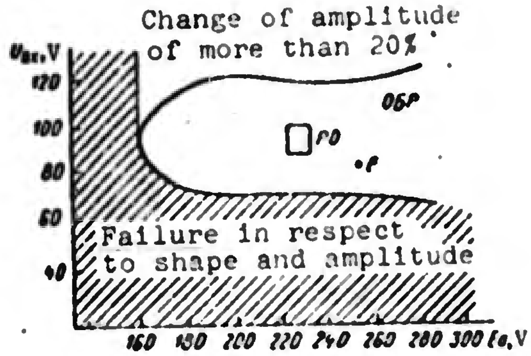


Figure 6.34. Trouble-free operating region (TFOR) of a blocking generator when input parameters U_{in} and U_a change

The operating point P corresponds to an input pulse amplitude $U_{in\ p}$ = 80 V and a power supply voltage source $E_{a\ p}$ = 250 V. The operating region is bounded by the limits ± 5% in respect to the input voltage and ± 3% in respect to the power supply voltage. As is seen from Figure 6.34, in order to increase the reliability it is necessary to shift the circuit operating point P upward in respect to the parameter being studied — that is, it is necessary to increase the amplitude of the trigger pulse to 90 - 100 V.

Figure 6.35 shows the circuit trouble-free operating region with a change of the capacitor C_1 capacitance and the voltage of the power supply source E_a. Capacitor C_1 is of SGM-1 type with a capacitance of 400 picofarad and of precision Class 1. The operating region is bounded by the limits ± 2% in respect to the capacitance and ± 3% in respect to the power supply voltage. The cross section of the trouble-free operating region in respect to the parameter C_1 is small, and a change of the capacitor value with time can lead to failure.

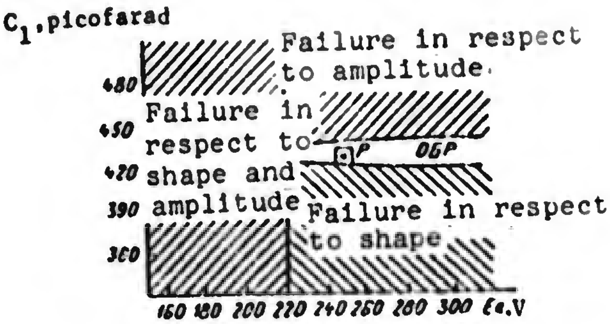


Figure 6.35. Trouble-free operating region (TFOR) of a blocking generator with a change of the input parameters — capacitance of capacitor C_1 and voltage of the power supply source E_a

Based on the boundary test results, in this case one may determine the boundaries of acceptable changes and the optimum values of the quantities R_a, U_{in}, and C_1 of the power supply voltage of the tube anode circuit E_a.

Above, a technique was investigated for conducting boundary tests on the assumption that the output parameters depend on a single input parameter. In actual cases, the output parameter depends on several input parameters. The trouble-free operating region in this case may be obtained in the following way. One must determine the circuit trouble-free operating region associated with all the values of the input parameters. Then these regions are plotted in the relative coordinates ϕ_j/ϕ_{jH} and U/U_H and, by superimposing them, one finds the combined trouble-free operating region (CTFOR). This region also will be the trouble-free operating region in respect to the i^{th} output parameter with allowance for the influence of all the input parameters. Figure 6.36 shows the boundary monitoring graph of the i^{th} output parameter for n input parameters. The combined trouble-free operating region in Figure 6.36 is cross hatched.

The boundaries of the operating regions may be determined from the expressions:

$$\frac{\varphi_{j\,мин}}{\varphi_{j\,н}} = 1 - \frac{\Delta\varphi'_j}{\varphi_{j\,н}},$$

$$\frac{\varphi_{j\,макс}}{\varphi_{j\,н}} = 1 + \frac{\Delta\varphi''_j}{\varphi_{j\,н}}; \quad \frac{U_{мин}}{U_н} = 1 - \frac{\Delta U'}{U_н};$$

$$\frac{U_{макс}}{U_н} = 1 + \frac{\Delta U''}{U_н};$$

where $\Delta\varphi'_j$ and $\Delta\varphi''_j$ — are the negative deviations and positive deviations of the input parameters from the rated value.

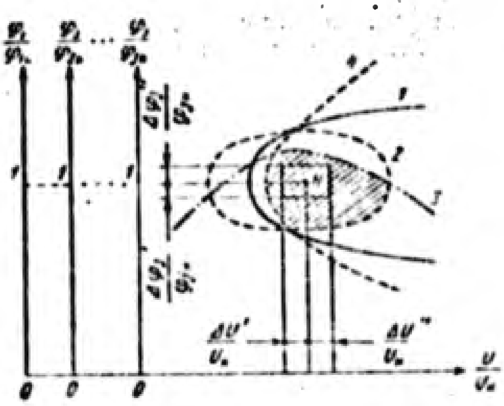


Figure 6.36. Construction of the combined trouble-free operating region (CTFOR)

It is seen from Figure 6.36 that the circuit is less critical toward a change of certain input parameters, and more critical toward a change of others.

In boundary tests, it is necessary to simulate a change of the parameters of various elements: resistors, capacitors, inductance coils, semiconductor devices, electron tubes, etc. A change of the parameters of resistors, capacitors, and inductance coils may be simulated by a series or parallel connection of appropriate elements. A change in the parameters of resistors may also be simulated by changing the voltage being supplied to the separate parts of the circuit.

A change in the parameters of semiconductor diodes also may be produced by series or parallel connection of resistances. It is considerably more complicated to simulate a change of the parameters of electric vacuum devices and transistors, since their characteristics depend on many parameters. In these cases, a tube or transistor is selected all of whose parameters are the same as in the test element, except for the parameter being simulated and the boundary tests are conducted with it. A number of artificial simulation methods are also used.

Let us proceed to the graphical-analytical method of performing boundary tests. The essence of this method is that the phenomena occurring in actual circuits may be reproduced graphically when making the boundary tests by using the characteristics of electron tubes. We do not use such a method on all circuits, but in a number of cases it may be very convenient. As an example for illustrating the graphical-analytical method for calculating boundary test graphs, let us investigate a rheostat amplifier (Figure 6.37). Figure 6.38

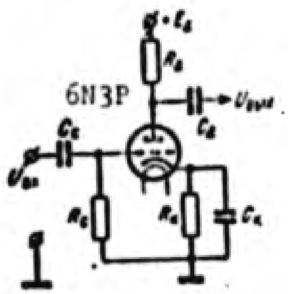


Figure 6.37. Circuit of a video pulse amplifier

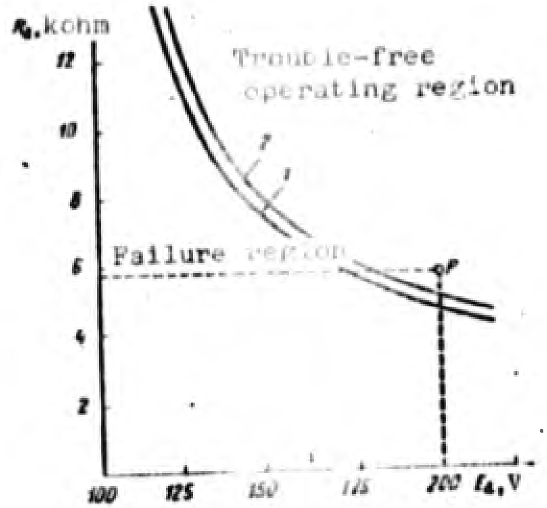


Figure 6.38. Boundary test graphs of a video pulse amplifier

shows the experimental graph (1) and the calculated graph (2) of the amplifier boundary tests in the R_a and E_a coordinate system. The output pulse amplitude $U_{out} = 45$ V is taken as the output parameter.

The boundary of the trouble-free operating region corresponds to a change of the pulse output parameter of 10%. The operating point P for the values of the input parameters $R_{a\,p} = 5.8$ kohm, and $E_{a\,p} = 200$ V is shown in the same figure.

Figure 6.39 shows similar graphs, but in the R_k, E_a coordinate system. The operating point corresponds to the input parameters $R_{k\,p} = 260$ ohm and $E_{a\,p} = 200$ V. The error in the calculation for both cases does not exceed 10%.

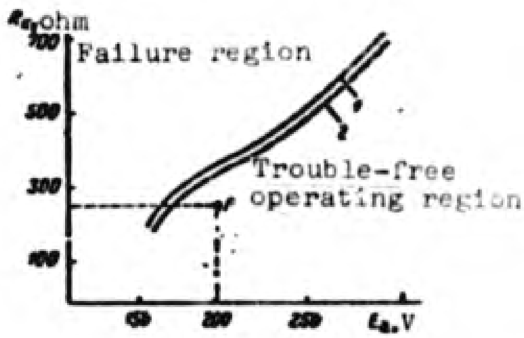


Figure 6.39. Boundary test graphs of a video pulse amplifier in the R_k, E_a coordinate system (1 — experimental curve; 2 — calculated curve)

Boundary Monitoring of Equipment

Boundary tests also provide useful information during the equipment operation. In this case, it is more correct to talk about boundary monitoring of the equipment.

During the operating process, boundary monitoring reveals the elements whose parameters are approaching the limiting value — that is, it determines the equipment reliability reserve. In order to do this, it is necessary to have graphs of the equipment boundary tests which were plotted before beginning its operation.

Boundary monitoring may be performed during preventive maintenance checks. This reduces to a minimum the equipment standby time.

In a number of cases, the boundary monitoring may be automated [31]. The solution of this problem is simplified if the power supply voltage is the boundary monitoring parameter. In this case, in the radar it must be possible to change the power supply voltage. It is simplest to change the voltage of the common power supply source. However, such a monitoring technique has one substantial drawback, which is that a change of the voltage necessary for checking one device may turn out to be unsatisfactory for checking another device — or, what is even worse, it may cause its failure. As a consequence, it is advisable to divide the radar circuit into groups of stages and to feed the power supply to them along separate lines — boundary monitoring lines.

Groups of circuits or stages must be combined based on the power supply voltages and the limits of their change (which are necessary for predicting the reliability reserve) which they have in common. Naturally, during the monitoring one must not cause the equipment to fail. Th minimum acceptable reliability reserve must be established, and the limiting power supply voltage corresponding to this state must be determined. Such monitoring is very helpful, but it has limited possibilities, since a change of all the input parameters which affect the equipment reliability is not simulated.

The experimental boundary monitoring method may also be used when tuning and adjusting equipment. Tuning and adjusting electronic equipment at present is done by the method of successive approximations. The complicated interrelationship of the phenomena occurring in circuits, in a number of cases, makes it impossible to establish that the circuit optimum tuning conditions have been selected with respect to reliability. The boundary monitoring method makes it possible to solve this problem. Let us demonstrate this with an example.

Let us assume that a device has two adjustment units, whose element parameters are ϕ_j and ϕ_k. Let us also assume that the purpose of the adjustment is to achieve optimum operating conditions for the device, and that these conditions are characterized by two output parameters whose values must equal v_n and v_m. If one plots the boundary monitoring graphs for both output parameters in the

ϕ_j and ϕ_k coordinate system and superimposes these graphs as is shown in Figure 6.36, we obtain the combined trouble-free operating region -- that is, the region which is common for both output parameters. This region is cross hatched in Figure 6.36. Having obtained the combined trouble-free operating region, we are able to soundly select the operating point corresponding to the circuit's optimum tuning conditions. With equally probable drift directions of the parameters ϕ_j and ϕ_k, the operating point must be selected in the center of the combined zone, but with unequally probable drift directions the operating point must be shifted away from the center, in the direction opposite to the drift of the parameters ϕ_j and ϕ_k.

Boundary monitoring graphs also indicate which adjustments must be made in the tuning process and equipment operation, and within what limits. When designing equipment, the question usually arises as to the advisable number of adjustment units. Adjustments and tunings take place both in the factory and during operation. Factory adjustment and tuning is performed by using all the units which are provided for this purpose. Operational adjustment and tuning as a rule is accomplished using those units which have been especially brought out to the front panels of the instruments. This adjustment has the purpose of maintaining constant operating conditions and constant equipment parameters with changing operational conditions.

A small number of adjustment and tuning units does not always guarantee the possibility of selecting the equipment optimum operating conditions. A large number of adjustment units leads to an increase of labor, both in the manufacturing process and also during operation. Practice shows that approximately 30 - 50% of the labor is expended in monitoring and adjusting electronic equipment during the manufacturing process. Moreover, a large number of adjustment units complicates the adjusting process, especially when there is coupling between the circuits in which these processes are being carried out. The presence of adjustment elements in equipment also reduces its reliability. This is explained by the fact that the reliability of elements with variable parameters, which are used for tuning and adjusting equipment, is lower than the reliability of elements with constant parameters. Moreover, during operation disturbance of the setting for the adjustment units takes place, which has an adverse effect on the equipment operation. If this is supplemented by the fact that servicing equipment with specially developed adjustment units requires technical personnel of high qualifications and, frequently, complicated built-in or portable monitoring-measuring equipment, it is not difficult to come to the conclusion that it is advisable to reduce in any way possible the number of adjustment and tuning units in equipment.

The configuration of the combined trouble-free operating region indicates the parameters on the basis of which the adjustment must be accomplished. If the cross section of the trouble-free operating region in respect to the input parameter is smaller than the limiting value of the parameter drift, it is necessary to have an adjustment unit for this parameter. On the other hand, if the cross section of the trouble-free operating region is large, and as a consequence of this the possibility of the operating region passing beyond the limits of the trouble-free operating region is eliminated, there is no necessity of having an adjustment.

Figure 6.40 shows the trouble-free operating region in the ϕ_j, ϕ_k coordinate system (TFOR) and the operating region (OR). The limits of the drift in the parameters ϕ_j and ϕ_k are shown in this same figure by dashed lines. It follows from the figure that the equipment must have an adjustment unit for the parameter ϕ_j, and it does not require adjustment for the parameter ϕ_k. The drift rate of parameter ϕ_j will characterize how frequently it is necessary

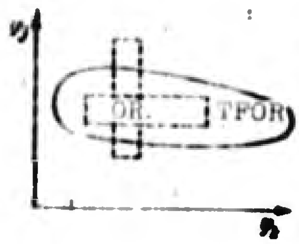


Figure 6.40. Displacement of the operating region with drift of the parameters ϕ_j and ϕ_k

to perform such an adjustment when servicing the equipment. Knowing the limits of the parameter drift and the cross section of the trouble-free operating region, we may also determine the advisable adjustment limits in respect to this parameter.

As an example, let us investigate the boundary monitoring graph of a superhigh frequency transmitter (Figure 6.41). The transmitter generates a series of pulses which are amplitude modulated.

The output parameters of the oscillator are:

— the magnitude of the peak voltage on the anode of the superhigh frequency oscillator tube;

— the radiated power in a pulse;

— the percentile modulation of the radiated pulses.

Figure 6.42 shows boundary monitoring graphs in the α, l coordinate system, where α is the rotor rotation angle of the trimmer capacitor C_{c} which changes the feedback capacitance; l is the position of the clearance which changes the coupling of a superhigh frequency oscillator with the antenna. The combined trouble-free operating region is cross hatched in Figure 6.42. If one assumes that the drift directions of the parameters α and l are equally probable, then the operating point must be selected in the center of the cross hatched region at $\alpha_p = 5°$ and $l_p = 25$ mm.

Passing beyond the limits of the cross hatched region, which for the given circuit is the combined trouble-free operating region, reduces the radiated power or leads to an abrupt increase of the peak voltage on the anode of the superhigh frequency oscillator tube, which may lead to transmitter failure.

Boundary graphs are one of the ways of predicting gradual failures. They enable one to solve the following problems:

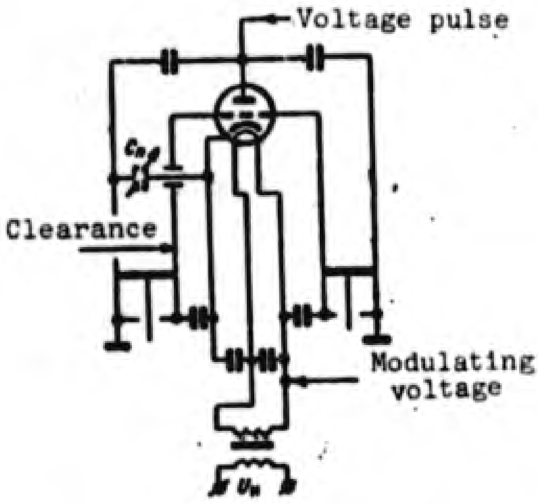


Figure 6.41. Circuit of a superhigh frequency transmitter

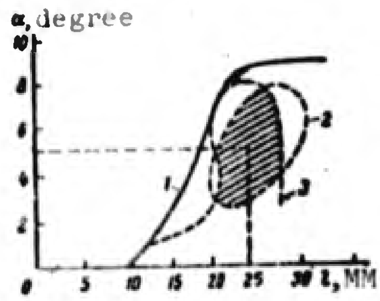


Figure 6.42. Determining the optimum parameters of the adjustable elements of the superhigh frequency transmitter:

1 — change of the peak voltage on the tube anode; 2 — radiated power per pulse; 3 — percentile modulation of the radiated pulse

— to make a selection of the rated values of the input parameters which ensure maximum reliability of the device when the parameters change during operation;

— to determine the region of acceptable deviations of the parameters of the elements in the circuit — that is, to determine the necessary precision class of the elements;

— to select the most reliable circuit from a number of possible circuits;

— to determine whether in a certain state a circuit performs its function with the most unfavorable operating conditions, when several factors leading to failures act simultaneously and have maximum values;

— to determine the circuit reliability reserve in respect to gradual failures;

— to determine the number of adjustment and tuning units for the equipment and the feasible limits of their use.

Inherent in the method of boundary tests are a number of defects. Among them, we must list:

— the considerable amount of labor in conducting the tests;

— the necessity of having measuring equipment of a high precision class;

— the complicated nature of simulating a change in the parameters of electric vacuum devices and transistors;

— the complexity of conducting tests, due to the necessity of taking into account the effect of all the destabilizing factors.

However, unification of the separate circuits enables one to carry out compound boundary tests, and subsequently to use the circuits whose reliability has been verified.

6.7. Acceptance of the Solution When Monitoring

During operation, monitoring is accomplished to check the system operating efficiency and to predict and search out failures. The content and extent of the monitoring depends on the scope of problems being determined by the system, its complexity and state. Therefore, the monitoring problems may be solved slightly differently for each specific system. However, there is a certain common approach for the most efficient monitoring method. This common approach was developed from practical experience in operating systems, and is confirmed by theory.

The models created recently in the theory of reliability and technical diagnosis for servicing complicated technical systems enable one, in a number of cases, to optimize the operational process. In operational practice, in addition to the scientific approach to solving various problems, "volitional" solutions still are frequently used. When this is done, incorrect solutions inevitably are taken and, as a consequence of this, normal operational conditions of systems are disrupted. Therefore, it is necessary to establish, even if only in the first approximation, those connections which exist between the various kinds of monitoring and the other operational processes.

Monitoring of a system is not an end in itself. Acceptance of a solution directed toward achieving normal use of a system should be the result of monitoring.

In operational practice, monitoring of the operating efficiency usually dominates the other kinds of monitoring. This is due to the following. When preparing complicated systems for use, it is necessary first of all to know in what state they are found. Then monitoring the operating efficiency enables one to take the proper solution, depending on the system state. If the system is defective, then the decision is reached to repair it or replace it with a standby. If the system is in good working order, then the decision is taken as to its subsequent utilization.

Monitoring the operating efficiency clearly is connected with another kind of monitoring — predicting the system state. This connection appears both in the common nature of their purpose, and also in the selection of common monitoring means. Actually, when monitoring operating efficiency, the current state of a system is determined, and when making a prediction its possible states in the future are forecast. Therefore, the monitoring device used for monitoring the operating efficiency may simultaneously serve the function of a predicting device. Such combined monitoring devices at present find broad use in equipment being designed and operated. The development, operational principle, and operation of devices for monitoring and searching out malfunctions are very special problems, and we will not investigate them here. Special literature [4, 31] is devoted to these questions.

With the investigation in §6.2 of the procedure for monitoring a system operating efficiency, it was considered that the monitoring conditions are ideal — that is, the errors of the measuring devices are small, and the monitoring-measuring equipment has a high reliability. However, for actual conditions these assumptions are not warranted. Therefore, the process of accepting a solution must be regarded as a random process. This process may be evaluated quantitatively. When this is done, in order to describe the solution acceptance process, various mathematical models are used. One such model, which characterizes the solution acceptance process associated with monitoring a system operating efficiency, was proposed by B. S. Abramenko. This model enables one to evaluate the degree of confidence in the fact that the system, which is authorized for use after its monitoring, actually is in good working order. We will take as the measure of such confidence the probability that a positive monitoring result obtained as a result of checking the system corresponds to actuality.

When constructing the model, it is assumed that the system is monitored before use. In the case of positive results of all the monitoring checks, the system is accepted for use. In the case of failure of the monitoring device, the use of the system is delayed until the monitoring device is restored. The possibilities of failure of the system or its monitoring device, and the appearance of errors in the monitoring process and in the solution acceptance process, are evaluated by introducing the following probabilities:

P_1 — the probability that the system is in good working order before the monitoring $(P_1 = P(t) = e^{-\lambda t})$;

P_2 — the probability that the monitoring device is in good working order before the monitoring;

P_3 — the probability of the appearance of malfunctions in the system during the monitoring process;

P_4 — the probability that the system, which prior to the check was in good working order, refuses to begin to operate;

P_5 — the probability that the monitoring device is in good working order during the monitoring process;

P_6 — the probability of acceptance of a system in good working order as a defective system;

P_7 — the probability of acceptance of a defective system as a system in good working order;

P_8 — the probability of repairing a defective monitoring system.

It is assumed that the events whose probabilities are presented above are independent.

The solution acceptance process is an alternative process, and it is convenient to represent it in the form of a graph. The graph for the events taken in this solution acceptance model is presented in Figure 6.43. From each vertex of the graph, except for vertexes x_8, x_{14}, and x_{15}, two arcs emerge which correspond to the two alternatives. For example, arc (x_1, x_2) emerging from vertex x_1 corresponds to the event which occurs before the beginning of the monitoring — "the monitoring

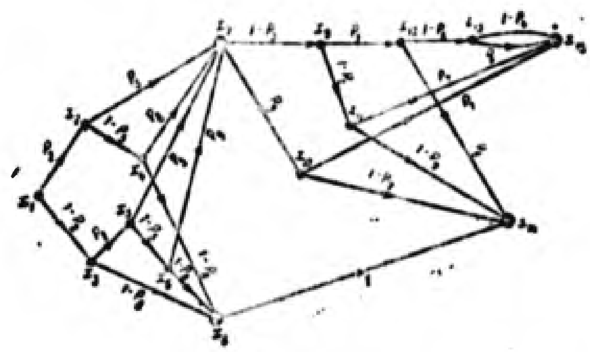


Figure 6.43. Graph of acceptance of a solution

device is in good working order". The other arc (x_1, x_3) shows the event "the monitoring device is defective". The remaining arcs of the graph correspond to other assumed events. Each arc of the graph is marked with the probability characterizing the possible appearance of this event.

Thus vertexes $\{x_7, x_8, x_{14}, x_{15}\}$ show the acceptance of a final solution:

x_7 — the monitoring device is in good working order;

x_8 — the monitoring device is defective;

x_{14} — the system is not authorized for use;

x_{15} — the system is authorized for use.

Let us use this model for calculating the probabilities of the acceptable solutions. One may do this in a general form in the following way. Starting from vertex x_1 and moving along a certain path μ to vertex x_j, let us find the product of the probabilities which are marked on the arcs belonging to this path. Thus, for example, if from vertex x_1 to vertex x_7 one follows along the path $\mu [x_1, x_2, x_1, x_7]$, then we obtain the product

$$P_2(1-P_5)P_6.$$

Several arcs approach each vertex which shows a final solution. Therefore, it is necessary to proceed along all paths which lead to this vertex, to find all the products, and to sum them. Let us find, for example, the probability that the monitoring device is in good working order, having designated it by P_{md}. In order to find this probability, it is necessary to proceed from vertex x_1 to vertex x_7 along four paths:

$$\mu[x_1, x_2, x_7]; \quad \mu[x_1, x_3, x_6, x_7];$$
$$\mu[x_1, x_2, x_6, x_7]; \quad \mu[x_1, x_3, x_5, x_6, x_7].$$

As a result, we obtain

$$P_{ey} = P_2P_5 + P_2(1 - P_5)P_6 + (1 - P_2)P_6P_5 + (1 - P_2)P_6(1 - P_5)P_6$$

or after reorganization

$$P_{ey} = [P_2 + (1 - P_2)P_6][P_5 + (1 - P_5)P_6];$$

The remaining probabilities are determined just as simply. In particular, the probabilities that the system will be authorized for use P_{sd}, or that it will not be authorized for use P_{sn}, equal respectively

$$P_{\text{сд}} = [P_2 + (1 - P_2)P_6][P_5 + (1 - P_5)P_6]\{(1 - P_3)\times$$
$$\times[P_1(1 - P_6) + (1 - P_1)P_7] + P_3P_7\};$$
$$P_{\text{сн}} = [P_2 + (1 - P_2)P_6][P_5 + (1 - P_5)P_6]\{(1 - P_3)\times$$
$$\times[(1 - P_1)(1 - P_7) + P_1P_6] + P_3(1 - P_7)\}.$$

The model being investigated clearly represents monitoring the operating efficiency as concerns the solution acceptance process, and it may be used for obtaining certain objective monitoring indexes. The monitoring advisability coefficient K_{ad} is such an index, for example. The monitoring advisability coefficient characterizes how many times one's confidence is increased that the system is in good working order when the system is monitored, in comparison with the case when the system is not monitored.

The coefficient K_{ad} may be determined by using the following considerations. Let us assume that the monitoring device has absolute reliability in comparison with the system being monitored — that is, $P_2 = P_5 = 1$. Then the probability that the system after monitoring will be authorized for use equals

$$P'_{сд} = (1-P_3)[P_1(1-P_6)+(1-P_1)P_7] + P_6P_7.$$

The most dangerous disturbances in the monitoring process are: the appearance of malfunctions in the system being monitored owing to the monitoring devices and monitoring errors of the type "accepting a system in good working order as a defective system". The probability of the combined nonappearance of these events equals

$$P_н = (1-P_3)(1-P_6).$$

Then the monitoring advisability coefficient may be represented as the ratio of the quantity P_n to P'_{sd}, that is

$$K_ц = \frac{(1-P_3)(1-P_6)}{(1-P_3)[P_1(1-P_6)+(1-P_1)P_7]+P_6P_7}.$$

The decision as to the advisability of monitoring should be affirmative only when the quantity $K_{ad} > 1$.

If, when monitoring the operating efficiency, the decision is reached that the system is not authorized for use, then the next decision must be to change over to a restoration procedure. The restoration process is an alternative process, and it also is conveniently represented in the form of an oriented graph. The main features and characteristics of the process for restoring a system were investigated by us earlier (see Chapters 4 and 5).

Let us discuss the problem as to how a solution is reached concerning the restoration of a system.

When monitoring the operating efficiency (see §6.2) on the basis of a prescribed program, if there is a negative result from any check, subsequent monitoring ceases, and the following assumption is made. One or several of those elements in the subset of the system elements causing a negative result of the check are defective. In accordance with the accepted technique of constructing a program for monitoring the operating efficiency, this subset of elements is known beforehand. Consequently, by using the technique for constructing a malfunction search program, one may construct beforehand a malfunction search program for each such subset.

By combining the malfunction search programs constructed for each such subset with the program for monitoring the operating efficiency, we obtain a single diagnostic program for the given system. Let us find, for example, the single diagnostic program for the example investigated in §6.2.

Let us assume that the time for checking the state of each element is the same, and equals t = 2 min. The data for the probabilities of the defective states of the elements are taken from the table of §6.2 (p. 340).

Let us investigate the program for monitoring the system operating efficiency which was presented in Figure 6.8. In the case of a negative result of check π_6, the subset $X_1 = \{x_1, x_5, x_6, x_7\}$ is separated out. The malfunction search program for this subset is represented in Figure 6.44,a. The average time for searching out a malfunction according to this program, with the condition that the

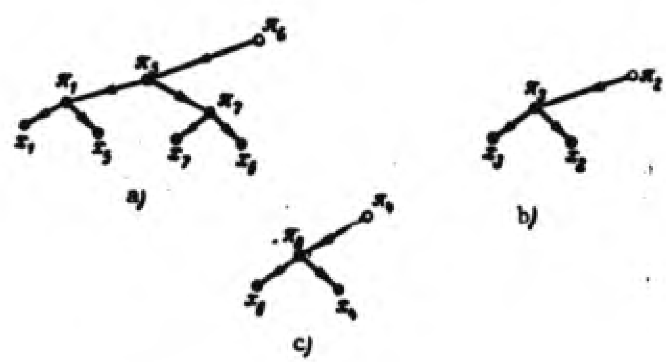


Figure 6.44. Malfunction search programs for the system shown in Figure 6.1

sum of the probabilities of the elements included in X_1 is normalized to one, equals

$$\frac{1}{P_1+P_5+P_6+P_7}[(P_1+P_5)+$$
$$+(P_6+P_7)+(P_1+P_5+P_6+P_7)]=4 \text{ min.}$$

For subsets $X_2 = \{x_2, x_3\}$ and $X_3 = \{x_4, x_8\}$, the programs corresponding to them are shown in Figures 6.44b and c. The average time for locating a defective element for the program of Figure 6.44b equals 2 minutes, and for the program of Figure 6.44c it is 3 minutes.

Let us note that the program for subset $X_3 = \{x_4, x_8\}$ was constructed with the condition that the feedback between these two elements may be broken when the check is being made.

The single diagnostic program is shown in Figure 6.45.

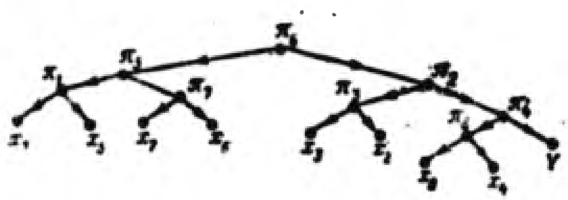


Figure 6.45. Diagnostic program for the system shown in Figure 6.1

7. ADJUSTMENT AND REPAIR PROCESSES

Just as in diagnostics, repair and adjustment are component parts of a single process for restoring the operating efficiency of a system and its elements. Despite the fact that rapair and adjustment as a rule are less time consuming and considerably simpler than the diagnosis, the struggle for economy in the overall time for restoring the operating efficiency of complicated systems requires a comprehensive study of these processes and an increase in their efficiency.

7.1. Methods of Carrying Out Adjustment Tasks

The necessity for adjustment operations is caused by the unavoidable excursions of the parameters of individual elements, units and assemblies of a system beyond the acceptable limits. When making an adjustment, the excursions of parameters beyond the acceptable limits are compensated by using adjustable and tuneable elements. Among the adjustable and tuneable elements are: variable resistors and trimmer capacitors (trimmers), inductance coils with adjustable cores, resonance systems of the elements of a radar high-frequency circuit, etc.

On the average, in radar stations there is one adjustable element for each 8 - 12 elements. The main criterion for the correctness of the repair or adjustment which has been performed is checking for the agreement between certain parameters of the system and the technical specifications. One may check either the parameter whose value directly depends on the repair-adjustment tasks which have been performed, or the composite parameter which is a function of the parameters in which changes show up as a direct result of these tasks. A check using the composite parameter is characteristic for repair-adjustment tasks, since it enables one to avoid elementary checks and consequently to shorten the monitoring time. Thus, for example, when tuning the high-frequency section of a radar station, the individual parameters of the tunable elements are not measured, but the overall parameter for them (the power drop, the receiving device sensitivity, the noise factor, the transmitter power).

When making a check on the basis of the composite parameter, one should take into account the degree to which each adjustable element influences this parameter, since the quality and order of the adjustment depend on this. As operational experience has shown, the degree of this influence differs and may vary within considerable limits. Therefore, when making an adjustment one must first turn his attention to those elements whose degree of influence on the composite parameter is greatest.

Using the composite parameter for checking the adjustment quality does not always lead to the desired results. In a number of cases, the use of this checking method may lead to a redistribution of the conditions in the circuits, and as a consequence of this to the failure of individual elements. Let us illustrate this with a simple example. Let us assume that it is necessary to adjust the value of an equivalent resistance R_e in the electric circuit based on the circuit shown in Figure 7.1. It is not difficult to see that, if the adjustment is conducted by directly measuring the resistance R_e and not its individual components $(R_1 + R_3)$ and $(R_2 + R_4)$, then the value of resistance R_e does not uniquely depend on the resistances of resistors R_3 and R_4. When this is done, the power dissipated in the resistors will not be the same, and consequently the failure probabilities of the individual elements increase.

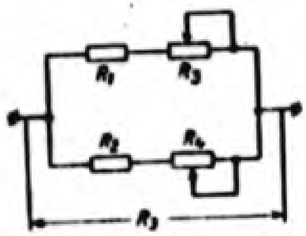


Figure 7.1. Circuit for adjustment and measurement of an equivalent resistance.

When carrying out adjustment tasks, it is always necessary to maintain the necessary precision in obtaining the given values of the individual equipment parameters. The adjustment precision depends on the adjustment methods being used and the precision of the monitoring-measuring equipment. In operational practice, two adjustment methods have found use:

- the adjustment method based on measuring instruments;
- the comparison method.

Adjustment based on measuring instruments reduces to the following. The required inputs are fed to the inputs of the unit (assembly) being adjusted, and measuring instruments are connected to the external outputs. By means of the adjustment elements, the outputs of the adjustable units are made to correspond to acceptable values. The essence of this method is illustrated in the block diagram shown in Figure 7.2. With this method, absolute values of the input and output quantities are measured. An objective evaluation of the adjustment quality is made on the basis of these quantities. Therefore, very rigid requirements must be imposed on the precision of the measuring instruments which are used for monitoring the parameters when making the adjustment. Adjustment on the basis of measuring instruments finds wide use in operational practice. In particular, when tuning a radar to its maximum effective range, this method should be considered basic. The question of the necessary precision of the measuring instruments being used when carrying out tuning was investigated in Chapter 2.

The essence of the comparison method is that the values of the output quantities of the unit being adjusted are compared with the values of the output quantities of a unit which is of the same type

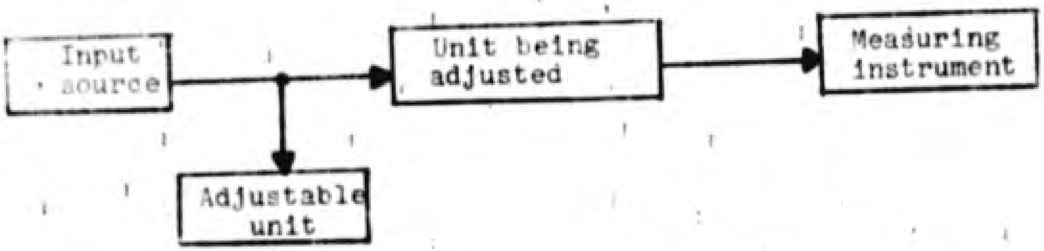


Figure 7.2. Arrangement for connecting measuring instruments when adjusting a unit of equipment.

as the unit being adjusted. When this is done it is not necessary to know the precise value of the output parameter of the unit being adjusted, but it is sufficient to determine whether this parameter is within the tolerance limits.

The comparison method finds wide use in mass production and repair of equipment. Under operational conditions, the use of this method is limited, because of the absence of a sufficient number of units of the same type. However, this method is not neglected when there are no precise data on the values of the output quantities of the unit being adjusted, or when the precision of the measuring instruments is low.

The question of the precision of these methods is investigated in more detail in [32].

The repair-adjustment process consists of individual elementary operations or tasks. An <u>elementary operation</u> when making repairs and adjustments means an operation which is performed without changing the measuring equipment or the places of its connection, and at the same time the very same set of instruments is used. Examples of elementary operations are: replacement of a tube, mechanical tuning of a klystron (changing the cavity of the klystron's resonator), replacement of the transformer in a power supply unit, installing a selsyn stator when matching synchronous transmission lines, etc.

In turn, each elementary operation may be a combination of such universal operations as: disassembly, assembly, mounting, dismantling, soldering, fine tuning, lubricating, cleaning, etc.

When performing repair-adjustment tasks, elementary operations must be performed in a specific order. The sequence of performing the elementary operations is called the program of the repair-adjustment tasks.

Let us assume that the repair and adjustment processes being investigated are determinative processes (see Chapter 1). The defective elements which are detected in the diagnostic process are replaced by elements which are knowingly in good working order, and each elementary operation and its check have only one result. It is also assumed that the interval of time between the end of the preceding operation and the beginning of the following operation equals zero. Definite relations exist between elementary operations. These relations determine the sequence of performing the individual operations. The multiple character of these relations leads to a large number of variations in the repair and adjustment sequences for equipment. Among the different variations of the sequences for performing the operation, one may select that sequence which ensures the most efficient program of action when making repairs and adjustments.

7.2. Formulation of Programs for Repair-Adjustment Tasks

Let us designate by y an elementary operation associated with the adjustment and repair of equipment and let us investigate, for example, the set $Y=\{y_i/i=1, \ldots, 8\}$, consisting of eight such elementary operations. Let the following relations between the elementary operations of this set be specified:

1. Operation y_1 is performed before operation y_2. This necessary condition may be designated by $y_1 < y_2$.

2. The sequence of performing operations y_1 and y_3 has no importance — that is, one may first perform operational y_1 and then operation y_3, or conversely first y_3 and then y_1. This condition should be written in the form $y_1 < y_3$ and $y_3 < y_1$, or more concisely $y_1 >< y_3$.

3. Operation y_6 must be performed immediately after performing operation y_4. This relation determines that y_4 directly precedes y_6, and it is designated by $y_4 |< y_6$.

Let us assume that the relations between the elements of set Y are assigned in the following way:

$$\begin{gathered} y_1 < y_2,\ y_3,\ y_5;\ y_2 < y_3; \\ y_1 >< y_3,\ y_4;\ y_2 < y_6;\ y_3 < y_4,\ y_5,\ y_6; \\ y_4 |< y_6;\ y_5 < y_3,\ y_6;\ y_6 >< y_8. \end{gathered} \tag{7.1}$$

These relations may be represented by means of an oriented graph, which for case (7.1) is shown in Figure 7.3. The vertexes of this graph correspond to the elementary operations, and the arcs characterize the relations between these operations.

The process of directly constructing the graph from the given system of relations (7.1) presents certain difficulties. It is more convenient to proceed at once to (0.1)-matrices.

Such an n X n-dimensional matrix consisting of zero's and one's may be understood as the distribution of n elements over n sets: the one's in the i^{th} row indicate the elements included in the i^{th} set, and the one's in the j^{th} column indicate the sets containing the j^{th} element. For a given graph G, such a matrix is called the adjacency matrix (see Chapter 1).

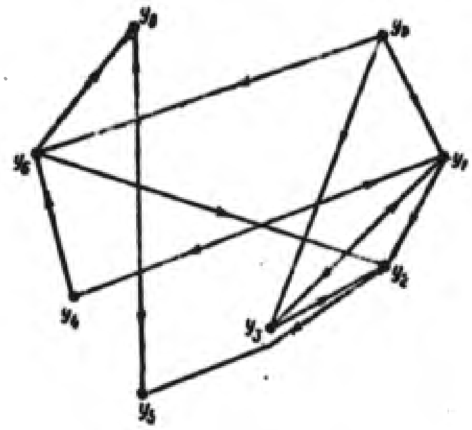


Figure 7.3. Oriented graph of the relations of elementary operations.

Let us designate the adjacency matrix of graph $G = (Y, U)$ by $A = (a_{ij})$. Element a_{ij} of the adjacency matrix which is at the intersection of the i^{th} row with the j^{th} column equals 1 or 0, under the following conditions:

$$a_{ij}=\begin{cases}1, \text{ if } & y_i|<y_j \text{ for } y_i<y_j;\\ 0, \text{ if not, one of the following conditions occurs} & \end{cases} \quad (7.2)$$

$$y_i<y_j,\ y_i|<y_j,\ y_i><y_j.$$

For the given system of relations (7.1) of the elements of set Y, the adjacency matrix has the form

$$A=\begin{array}{c|c} j=1\ 2\ 3\ 4\ 5\ 6\ 7\ 8 & \\ \begin{bmatrix} 0&1&1&1&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 1&1&0&0&0&0&0&0\\ 1&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1\\ 0&1&0&0&0&0&0&1\\ 1&0&1&0&0&0&1&0\\ 0&0&0&0&1&0&0&0 \end{bmatrix} & \begin{array}{c} i=1\\2\\3\\4\\5\\6\\7\\8\end{array} \end{array} \quad (7.3)$$

Let us check whether conditions (7.2) were fulfilled when constructing matrix (7.3). Let us investigate, for example, the relations $y_1 < y_2$ and $y_1 >< y_3, y_4$. Relation $y_1 < y_2$ indicates that in the graph (Figure 7.3) between vertexes y_1 and y_2 there should be an arc running from vertex y_1 to the vertex y_2, and in the matrix (7.3) element a_{12} must take the value 1. In matrix (7.3) this condition is fulfilled. For the relation $y_1 >< y_3, y_4$ condition (7.2) also is fulfilled since we have $a_{13} = 1$, $a_{14} = 1$ $a_{31} = 1$ and $a_{41} = 1$. In precisely the same way, the fulfillment of conditions (7.2) is checked for the remaining relations (7.1) of the example being investigated.

The next problem is to determine whether some certain path exists in the graph (Figure 7.3) between a certain vertex which is an entrance (start of the tasks) and any vertex which is an exit (end of the tasks). As is known, this path — which contains all the operations between which relations are specified by system (7.1) — is called a Hamiltonian path in the graph.

In order to find the Hamiltonian path in the graph, one may use the method which was indicated in [26]. According to this method, in order to find the Hamiltonian path in graph G its transitive closure $G^T = (Y, V)$ is constructed. As is known, the graph G^T has the following property: for each pair of arcs (a, b) and (b, c), there is a closing arc (a, c). By using this property of graph G^T, let us transform the matrix A to an adjacency matrix B of the transitive closure graph. This may be done by using the following algorithm [65]:

1°. Let B_i be a row vector of dimensionality n. Let us assume that the initial value of $B_i^{(0)}$ equals A_i — that is, the i^{th} row of matrix A. The row vector A_i corresponds to a certain vertex y_i of graph G.

2°. Let us investigate the elements of the row vector $B_i^{(0)} = (b_{i1}^{(0)}, \ldots, b_{in}^{(0)})$. Let b_{ik} which is different from zero be an element of $B_i^{(0)}$ which was not checked earlier. Let us transform $B_i^{(0)}$ by combining it with the k^{th} row of matrix A using elementary logic — that is, $B_i^{(1)} = B_i^{(0)} \cup A_k$.

The operation of logical summation is defined as follows:

$$1 \cup 1 = 1, \quad 1 \cup 0 = 1, \quad 0 \cup 0 = 0.$$

3°. Let us repeat the operation of logical summation as was done in 2 for all $b_{1k} \neq 0$ until the vector B_1 stops changing.

As a result of applying rules 1°, 2° and 3° to all the rows of matrix A, we obtain matrix B of the transitive closure graph.

Let us apply rules 1°, 2° and 3° to our example. Let us investigate the row vector $B_1^{(0)} = A_1 = (01110000)$ of matrix (7.3). In this row vector, the elements b_{12}, b_{13} and b_{14} are different from zero. Consequently, row vector A_1 must be logically combined successively with row vectors A_2, A_3 and A_4 — that is,

$$B_1^{(1)} = A_1^{(0)} \cup A_2 = (01110000) \cup (00001000) = (01111000),$$

$$B_1^{(2)} = B_1^{(1)} \cup A_3 = (01111000) \cup (11000000) = (11111000),$$

$$B_1^{(3)} = B_1^{(2)} \cup A_4 = (11111000) \cup (10000100) = (11111100).$$

After the operations have been performed on row vector B_1, new elements which are distinct from zero appear: b_{11}, b_{15} and b_{16}. Therefore, $B_1^{(3)}$ must be logically combined with A_1, A_5 and A_6. The final value of B_1 will have the form

$$B_1 = (11111101).$$

After applying rules 1°, 2° and 3° to all the remaining rows of matrix A, we obtain

$$\begin{array}{cc} & \begin{array}{cccccccc} j=1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ B = & \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} \\ i=1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \qquad (7.4)$$

In matrix (7.4) let us replace the elements along the main diagonal which equals zero with one's. Elements b_{22}, b_{66} and b_{77} are such elements. After this is done, let us rearrange the rows and columns of matrix (7.4) in such a way that the zero's would be located below the main diagonal, and the one's above it:

y_7	y_3	y_1	y_4	y_6	y_2	y_5	y_8	
1	1	1	1	1	1	1	1	y_7
0	1	1	1	1	1	1	1	y_3
0	1	1	1	1	1	1	1	y_1
0	1	1	1	1	1	1	1	y_4
0	0	0	0	1	1	1	1	y_6
0	0	0	0	0	1	1	1	y_2
0	0	0	0	0	0	1	1	y_5
0	0	0	0	0	0	1	1	y_8

(7.5)

In matrix (7.5), let us separate out the square matrices which consist entirely of one's and which rest on the main diagonal. These square matrices form <u>equivalence classes</u> with reference to the principle: vertexes y_i and y_j are interconnected in both directions. Therefore, one may simplify the initial graph (Figure 7.3) by dividing it into equivalence classes. In the given example, we have a division of the graph into five ordered subgraphs G_1, G_2, G_3, G_4 and G_5, (Figure 7.4). In particular, the set of vertexes $\{y_1, y_3, y_4\}$ forms an equivalence class generating subgraph G_2, and the set $\{y_5, y_8\}$ generates the subgraph G_5.

Dividing the initial graph G into equivalence classes also enables us to considerably simplify it by eliminating from it all the transitive closure arcs (Figure 7.5). Now determining the Hamiltonian path becomes a simple problem. This path, starting with vertex y_7 and terminating at y_8, successively passes through the vertexes y_7, y_3, y_1, y_4, y_6, y_2, y_5, y_8. Consequently, this path also determines the only possible (in the given example) order of performing the elementary tuning operations.

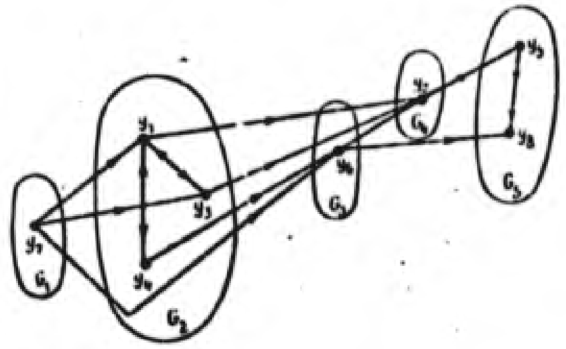


Figure 7.4. Division of the initial graph into equivalence classes.

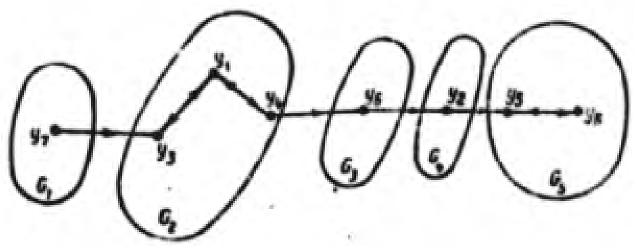


Figure 7.5. Selection of the Hamiltonian path in the graph.

Certain similar problems may lead to several solutions: thus, if in the example which was presented, there the relation $y_3 >< y_4$ occurred, then the additional solution y_7, y_1, y_3, y_4, y_6, y_2, y_5, y_8 would exist.

Let us investigate the application of this method for constructing a program of repair-adjustment tasks with a specific example. Let us assume that tuning of a radar high-frequency section is performed with the simultaneous replacement of the magnetron and klystron. The tuning is done with the help of a monitoring resonator. The magnetron is not tunable. When performing the tuning, the AFC circuit is switched off. In this case, the elementary operations will be:

y_1 - replacement of the magnetron;

y_2 - replacement of the klystron;

y_3 - switching on the station (without the high voltage);

y_4 - switching on the high voltage;

y_5 - setting the voltage on the repeller electrode of the klystron;

y_6 - tuning the monitoring resonator (echo chamber);

y_7 - tuning the klystron's cavity resonator;

y_8 - adjusting the mixer.

Based on operational experience, one may assign the following relations between the elements of the set:

$$y_1 >\!< y_2;\ y_1 < y_3;\ y_2 < y_3;\ y_3 < y_4,\ y_5,\ y_6,\ y_7,\ y_8;$$
$$y_4 >\!< y_5;\ y_4 >\!< y_6;\ y_5 < y_6;\ y_6 | < y_7;\ y_7 < y_8.$$

The graph of these relations is shown in Figure 7.6.

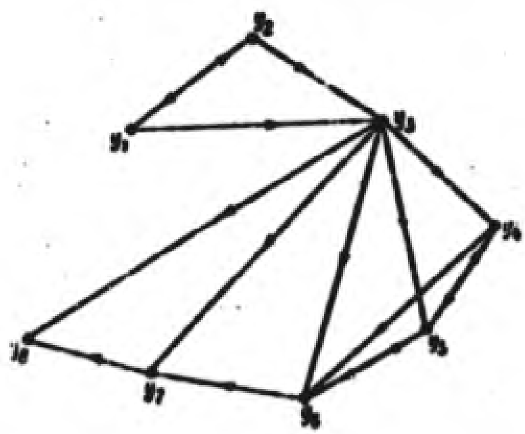


Figure 7.6. An oriented graph of the relations of the elementary operations when tuning a radar high-frequency section.

Using the algorithm for finding the Hamiltonian path in the given graph, we find that four Hamiltonian paths exist in it. These

Hamiltonian paths correspond to the following four tuning programs:

$$y_1,\ y_2,\ y_3,\ y_4,\ y_5,\ y_6,\ y_7,\ y_8;$$
$$y_2,\ y_1,\ y_3,\ y_4,\ y_5,\ y_6,\ y_7,\ y_8;$$
$$y_1,\ y_2,\ y_3,\ y_5,\ y_4,\ y_6,\ y_7,\ y_8;$$
$$y_2,\ y_1,\ y_3,\ y_5,\ y_4,\ y_6,\ y_7,\ y_8.$$

Any of these programs may be used for tuning the radar high-frequency section.

Naturally, in operational practice more complicated situations with a larger number of elementary operations may be encountered. Similar programs may also be constructed for them by using this method.

Selection of any other arbitrary sequence which is different from a program leads to a considerable loss of time when performing the tuning. For example, if one first tunes the klystron's cavity resonator and then sets the voltage on its repeller, the klystron frequency and the power being generated by it change. It would be necessary to repeat the tuning.

Thus, the method which has been discussed enables one to select the most efficient sequence of performing the operations and to shorten the overall time for repair and adjustment.

The repair and adjustment of each type of radar and its individual units and assemblies have their own peculiarities. These peculiarities result from the physical principles of their operation, the construction, the arrangement of the equipment, the monitoring-measuring instruments being used, the instruments being used when performing the tuning, etc. Therefore, in the next chapter the fundamental principles and techniques of tuning certain radar devices will be investigated.

8. TUNING A RADAR TO MAXIMUM EFFECTIVE RANGE

A radar effective range is its most important tactical parameter, frequently determining the effectiveness and use of a radar system. In Chapter 2 the effect of various factors on a radar effective range was investigated in some detail, and it was shown that — with preassigned conditions for the propagation of radio waves and the target parameters — a station's effective range is determined by its power potential.

In a vague way a station's power potential may be regarded as a function of the parameters of the superhigh frequency instruments which are used in the station for generating, switching, amplifying and transforming the electromagnetic oscillations. The superhigh frequency instruments are structurally combined in the radar high-frequency section. The composition and construction of the high-frequency section are determined by the purpose of the station, and the range at which it operates. Therefore, the high-frequency sections of radars which have different purposes and different ranges differ substantially. As an example, the block diagram of the high-frequency spectrum of the "Don" radar is shown in Figure 8.1.

A radar high-frequency section is the self-contained functional unit combining the high-frequency elements of the station's transmitter, receiver, and antenna-wave guide devices. The operational

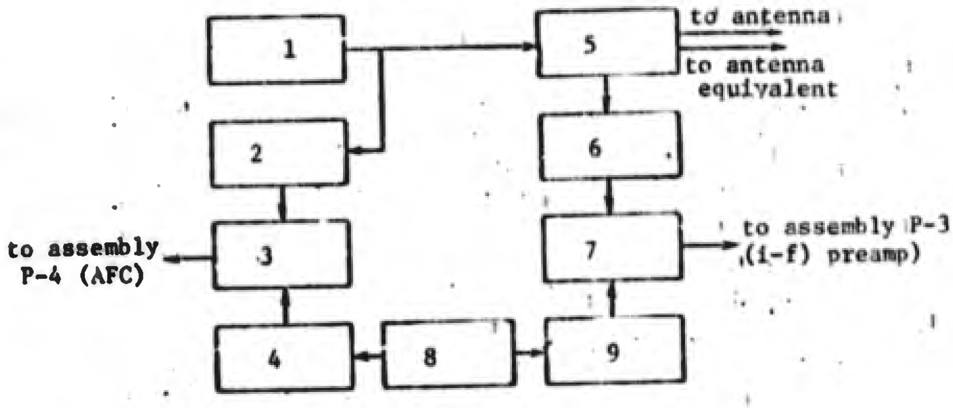


Figure 8.1. The block diagram of the high-frequency section of the radar "Don".

1 - magnetron oscillator; 2 - wave guide-coaxial junction; 3 - AFC balanced mixer; 4 - attenuator A-3; 5 - antenna switch; 6 - TR tube; 7 - balanced mixer; 8 - heterodyne; 9 - attenuator A-2.

quality of this unit to a considerable degree determines the station power potential, and consequently also its effective range.

Under the influence of various destabilizing factors, the parameters of superhigh frequency instruments which are included in the high-frequency section change their values in respect to the rated values. This reduces the radar power potential. In the technical servicing process, the operator by tuning, adjusting and replacing the elements and instruments of the station high-frequency section compensates for a change in the parameters of the superhigh frequency instruments, and achieves in this way the maximum possible effective range of the radar.

The superhigh frequency instruments in themselves are rather complicated devices. Their tuning and operation have a number of characteristic features. The tuning and adjustment of a set of superhigh frequency instruments and elements is further complicated by the fact that they operate in a close interrelationship: a change in the parameters of one of the elements causes a disturbance in the operation, and in certain cases it also causes failure of other superhigh frequency elements.

Therefore, prior to investigating the composite change in radar high-frequency section as a whole, let us discuss the tuning and operation of its individual instruments and elements. The main attention will be devoted to the dependence of their output parameters on the operating conditions, and to the methods of monitoring their functioning, as well as the requirement for combining the parameters of the superhigh frequency instruments in connection with their combined operation in the radar high-frequency section.

8.1. Superhigh Frequency Oscillators

In the transmitting devices of centimeter-wave range radar, magnetron type instruments are widely used as superhigh frequency oscillators.

Magnetron type instruments, which are sometimes called instruments with crossed fields, are an extensive class of electric vacuum devices which use the principle of synchronous interaction of electrons with electromagnetic waves in a phase velocity decelerating system.

By now a considerable number of various kinds of magnetron type instruments have been developed. Multicavity magnetrons, platinotrons, traveling-wave tubes and backward-wave tubes of the magnetron type (BWT of type M) have found the most practical use.

The superhigh frequency oscillator is the main electric vacuum device of a radar. One of the main requirements imposed on it is operating stability. By operating stability in this case, we understand invariable values of the frequency and power of the oscillations being generated. Operating stability of these devices may be achieved only when definite conditions for their use are fulfilled.

Assuming that the reader is familiar with the operating principle of magnetron type instruments, let us investigate in more detail the questions connected with the operation of certain instruments of this class.

Parameters and Characteristics of a Multicavity Magnetron

Multicavity magnetrons continue to remain the most widely used generating instruments for the transmitting devices of centimeter-wavelength range radars. This is explained by the fact that magnetrons enable one to generate high power oscillations in a pulsed mode, they have a compact construction, they operate with a comparatively low power supply voltage, and they combine in themselves an electron tube and an oscillatory system. However, there are defects which are inherent in a multicavity magnetron. Among them one should first of all mention the criticality of the frequency and power of the oscillations being generated to changes of the power supply voltage and load, and also the narrow range of electronic frequency retuning (5 - 10 Mc).

The main parameters by which the operational properties of a magnetron are evaluated are the following:

- the amount of the useful power P delivered to the load;
- the efficiency η;
- the frequency of the oscillations being generated f;
- the pulse duration τ_p and pulse repetition frequency F_p;
- the electron frequency drift (EFD) Δf_e;
- the frequency pulling δ.

The magnetron parameters which have been listed depend on its operating conditions which are determined by the values of the constant magnetic field H, the anode voltage U_0, the anode current I_{a0}, the filament voltage U_H, and the load impedance Z_H.

The relationships between these quantities are determined experimentally for each type of magnetron and are clearly reflected by its operating characteristics and load characteristics. Typical operating characteristics of a pulsed magnetron are shown in Figure 8.2. Each

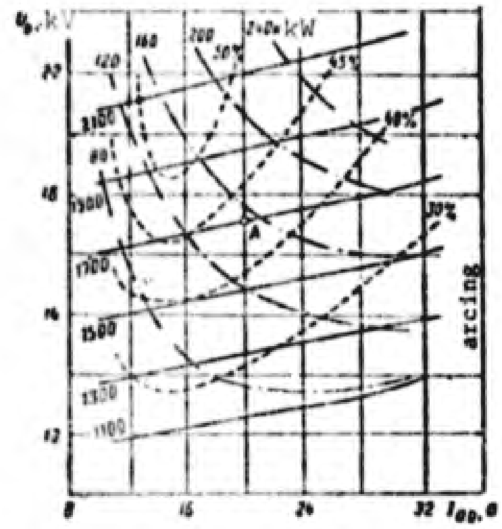


Figure 8.2. The operating characteristics of a magnetron. The lines of constant value of the magnetic field (continuous lines), constant power (dashed-dot lines) and constant overall efficiency (dashed lines) are plotted on the coordinate plane U_0, I_{a0}.

point of the plane U_0, I_{a0} shows the amount of power supplied to the magnetron $P_0 = U_0 I_{a0}$, and the slope of the characteristic is its static resistance $R_M = U_0/I_{a0}$. The operating characteristics are plotted with a fixed matched external load and the rated value of the filament voltage. They have great practical importance, since they enable one to select those magnetron operating conditions at which greatest stability of operation is ensured under operational conditions. By investigating the characteristics, one may come to the conclusion that the selected operating conditions uniquely characterize the magnetron parameters. Thus, for example, if the magnetron operates with a magnetic field intensity of 1700 oersted and a voltage of 17 kV, then the current equals 20 A, the output power — 150 kW, and the efficiency — 43%.

Consequently, the stability of its parameters may be determined on the basis of the results derived from monitoring the magnetron performance. Continuous monitoring of a magnetron established operating conditions is possible in principle, as was already indicated above, both on the basis of the current and also on the basis of the voltage. However, as follows from the operating characteristics, a deviation of the magnetron operating conditions from the established conditions gives rise to a considerably larger change of the anode current in comparison with the anode voltage. Thus, for example, with a magnetic field intensity of H = 1700 oersted, an increase of the anode voltage U_0 from 17 to 18 kV — that is, by less than 6% — gives rise to an increase of the current I_{a0} from 20 to 28 A, that is by 40%. For this reason, monitoring a magnetron operating conditions in a radar is not achieved on the basis of the anode voltage, but on the basis of the anode current (Figure 8.3).

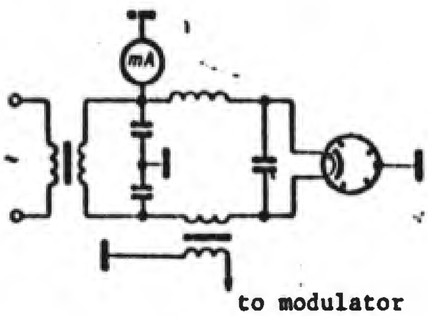


Figure 8.3. The circuit for turning on the instrument when measuring the average magnetron current.

In certain cases, constant frequency lines are plotted in addition to the operating characteristics. However, in order to evaluate the dependence of the frequency of the oscillations generated by the magnetron on its operating conditions, it is more convenient to use curves of the electron frequency drift (EFD) (Figure 8.4). Electron frequency drift is characterized by the amount of the frequency deviation associated with a change of the anode current by 1 A. Depending on the type of magnetron, the amount of the EFD is 0.5 - 2 Mc/A in the region of low currents, and 0.1 - 0.4 Mc/A at the rated conditions. The EFD value is an important parameter of a magnetron. This parameter imposes rigid requirements on the voltage stability of the modulating pulse flat section.

With pulsed modulation, changes of the anode voltage during the generation of the main part of the pulse cause parasitic frequency modulation, and consequently a broadening and distortion of the shape of the magnetron high-frequency pulse spectrum.

The dashed line in Figure 8.5 shows the shape of the spectrum of a high-frequency pulse with a duration of 1 microsecond when there is no parasitic frequency modulation. The continuous line shows the spectrum shape associated with a linear change of the frequency within the limits of the pulse of ± 3 Mc. As is seen from Figure 8.5, with a distortion of the spectrum the energy is distributed in a considerably

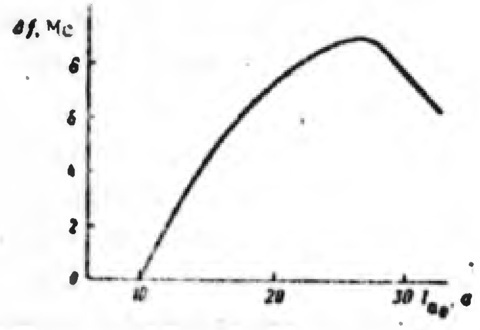


Figure 8.4. Curve of magnetron electron frequency drift.

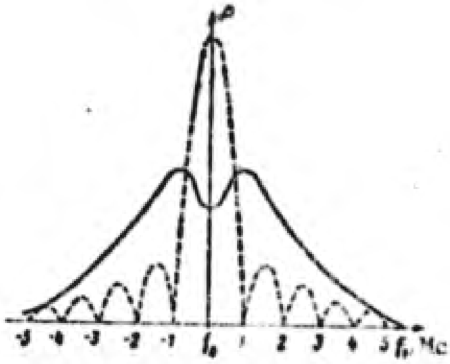


Figure 8.5. Broadening of a magnetron high-frequency pulse spectrum due to parasitic frequency modulation.

larger band of frequencies in comparison with a rectangular pulse. Since the receiver pass band is designed for operating with a normal shape of the spectrum, with distortion of the spectrum, part of the energy does not fall within the receiver pass band. This causes a reduction of the radar effective range. Moreover, the presence in the spectrum of several maxima of considerable size may cause unstable operation of the receiver AFC circuit.

Consequently, in the case of an unsatisfactory shape of the magnetron high-frequency pulse spectrum, one should check the shape of the modulating pulse. The technique for this check is discussed below.

A magnetron operating characteristics describe the behavior of a magnetron oscillator when operating on a fixed matched load. Under actual conditions, as was shown in Chapter 2, the load constantly changes. The effect of the load on the output power and frequency of the oscillations being generated by the magnetron is evaluated by means of the load characteristics. As is seen from Figure 8.6, the load characteristics express the dependence of the power and frequency of the generated oscillations on the modulus $|\rho|$ and phase ϕ of the

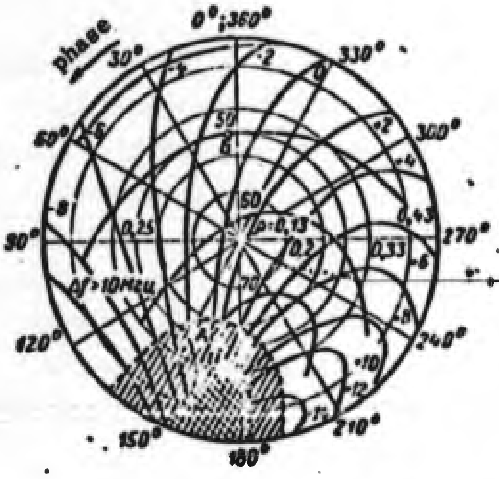


Figure 8.6. Load characteristics of a magnetron.

reflection coefficient by which the load is characterized. It is seen from the load characteristics that, with the same TWR value (on the characteristic curve, points A and B), the magnetron operation will be less stable in the region of larger powers. In this region, the power and frequency of the generated oscillations change sharply with insignificant changes of the load's phase. The change of the magnetron frequency, which is due to a change of the load, as was mentioned before, is called the frequency pulling. It is characterized by the pulling coefficient.

The <u>pulling coefficient</u> is the maximum change of the frequency of the generated oscillations when $\rho = 0.2$ (TWR = 0.67).

In order to determine the frequency pulling with any TWR value, we use the following expression:

$$\delta_f = 1{,}2\,\delta_{f_0}(1 - s). \qquad (8.1)$$

where δ_f is the amount of frequency pulling with the actual TWR value;

δf_0 is the magnetron frequency pulling coefficient.

Obviously the effect of the load on the stability of the frequency and power of the oscillations generated by the magnetron may be reduced by increasing the antenna-wave guide channel TWR. However, in practice it is impossible to completely match the magnetron with the antenna-wave guide channel. Matching, when a tuneable magnetron is used in the radar, presents special complications. The insertion of a ferrite rectifier in the wave guide channel is a radical means of eliminating the effect of the load on the magnetron operation. The operation of the ferrite rectifier is explained in §8.5. Since the ferrite rectifier prevents the passage of the reflected wave toward the magnetron and eliminates multiple reflections which occur between the oscillator and an unmatched load, the TWR of the antenna-wave guide channel in front of the ferrite rectifier is noticeably increased. The graphs presented in Figure 8.7 illustrate the TWR increase associated with the presence, between the magnetron and the antenna, of a ferrite rectifier with attenuation in the reverse direction equal to 5, 10 and 15 dB.

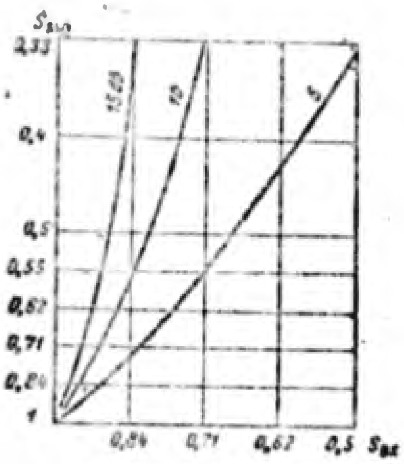


Figure 8.7. Graph for determining the TWR at the input s_{in} of a ferrite rectifier based on the TWR value at its output s_{out} with various values of the rectifier attenuation in the reverse direction.

The next important characteristic of a magnetron is the filament characteristic which expresses the relationship of the filament voltage U_H to the value of the average power P_{av} being supplied to the magnetron (Figure 8.8).

Stable and lasting operation of the magnetron is ensured when the temperature of its cathode is constant and equals the value indicated in the certificate. A temperature increase of an oxide-coated cathode of only 4% shortens the magnetron lifetime by a factor of three. Moreover, an increase in the cathode current density due to an increase of the cathode temperature leads to arcing in the magnetron. Not only is overheating of the cathode unacceptable, but also underheating of the cathode is unacceptable. With a low cathode temperature, as a consequence of insufficient emission, the electric field strength near the cathode is increased, which also leads to arcing in the magnetron and to destruction of its cathode.

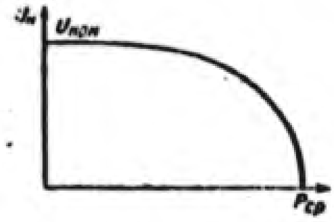


Figure 8.8. A typical filament characteristic of a multicavity magnetron.

During the operation of a magnetron, its cathode is additionally heated as a result of intense reverse bombardment by electrons. Heating of the cathode as a result of this is considerable, and increases with an increase of the average power supplied to the magnetron. For certain types of magnetrons, the amount of power consumed in additional heating is commensurate with the power consumed by the filament circuit. With such a power, the cathode may overheat and break down.

Having investigated the effect of the cathode temperature on the stability of the magnetron operation and its lifetime, one may draw the following conclusions:

- a high anode voltage must be supplied to the magnetron over the period of time required for heating the cathode to the rated temperature value;

- in the dynamic mode, the filament voltage must be reduced to the value indicated by the magnetron filament characteristic.

Checking a Magnetron Operating Conditions

In a radar, the optimum operating conditions of a magnetron are specified when it is designed, and they are ensured by the selected electric conditions and the operational conditions.

Trouble-free operation of a magnetron and the invariability of its parameters will occur when the electric conditions do not deviate from the rated values. However, in the operational process, under the influence of various destabilizing factors, the magnetron electric conditions change. Therefore, the necessity arises of monitoring the magnetron conditions and, on the basis of the monitoring data, of performing the appropriate tunings and adjustments.

When checking the operating conditions of a pulsed magnetron, it is necessary to measure:

- the magnetron average current;
- the filament voltage;
- the pulse repetition frequency;
- the amplitude and shape of the voltage of the modulating pulse;
- the amplitude and duration of the magnetron anode current pulse.

Measuring the magnetron average current. In radar, the magnetron average current I_{av} is monitored by means of a built-in milliammeter. With stable magnetron operation, the instrument readings should be steady and correspond to the rated value of the magnetron current under the operating conditions selected for it. In a radar where there is no ferrite rectifier, insignificant fluctuations of the instrument pointer relative to the position which corresponds to the rated value of the magnetron current are acceptable. These fluctuations may be caused by a periodic change of the load phase when the antenna rotates. Abrupt deflections of the instrument pointer from the position corresponding to the current rated value indicate unstable magnetron operation, which is produced by arcing in the magnetron or breakdowns in the modulator and load.

Arcing, as was indicated before, arises mainly as a consequence of overheating or underheating of the cathode. Moreover, arcing occurs in a magnetron when the pulse current exceeds the maximum acceptable value. This is possible when there is a noticeable reduction of the modulating pulse repetition frequency F_p, since $I_{av} = I_{a0}\tau_p F_p$, where I_{a0} is the magnetron pulse current.

A reduction of F_p produces an increase of the current I_{a0}, since the magnetron average current I_{av}, fixed by the instrument, is maintained at the rated value.

Let us assume that the established magnetron conditions correspond to point A on the operating characteristic (Figure 8.2) and the pulse repetition frequency has been reduced by a factor of 1.5. In order that the readings of the instrument monitoring the magnetron average current correspond to the rated value, the operator must set a value of the anode voltage U_0 such that the current I_{a0} increases by a factor of 1.5, in comparison with the initial value. As is seen from the operating characteristics, this causes the magnetron to operate in a region of large currents where arcing may arise.

Consequently, when instability of the magnetron current occurs, it is necessary first of all to check the magnetron in respect to the filament and the pulse repetition frequency.

Measurement of the filament voltage. The filament voltage is measured in the static (with the high voltage switched off) and in the dynamic (with the high voltage switched on) modes of the magnetron. In order to measure the filament voltage, one should use a voltmeter of precision class no lower than I.

When measuring the filament voltage in the static mode, the voltmeter should be connected directly to the magnetron filament terminals (Figure 8.9).

In the dynamic mode, the filament voltage is reduced in accordance with the filament characteristic. In the majority of radars, the filament voltage is reduced by automatically switching on an additional resistance in the primary winding of the filament transformer, when a specified value of the magnetron average current is reached. Since the magnetron heater has a high voltage applied to it in the dynamic mode, it is not advisable to measure the filament voltage directly at the magnetron terminals.

In this case, the filament voltage is measured in the following way.

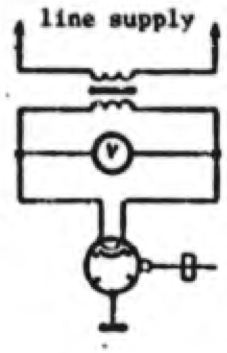


Figure 8.9. The arrangement for connecting a voltmeter when measuring the filament voltage of a magnetron in the static mode.

First the transformer ratio of the filament transformer is determined. After determining the transformer ratio, the voltage on the transformer primary winding is measured with the high voltage switching on and with the magnetron's average current at the rated value (Figure 8.10).

The filament voltage is calculated from the equation

$$U_{\text{н}} = \frac{U'_1}{k},$$

where U'_1 is the filament voltage on the filament transformer primary winding with the high voltage switched on;
k is the transformer ratio.

<u>Measurement of the pulse repetition frequency</u>. The pulse repetition frequency in the majority of radars is determined either by the frequency of the power supply voltage, or by the frequency of a quartz-crystal oscillator and its frequency division factor. Since the frequency of the power supply voltage — to say nothing of the frequency of the quartz-crystal oscillator — is quite stable, significant changes of the pulse repetition frequency may be produced by altering the operation of the frequency divider. The pulse repetition frequency may be checked, with sufficient precision, at the modulator output by means of an oscillograph.

The pulse repetition period $T_p = 1/F_p$ is determined from the number of calibration markers on the oscillograph between two adjacent pulses.

A more precise measurement of the pulse repetition frequency is made by means of an oscillograph and an audio-frequency oscillator

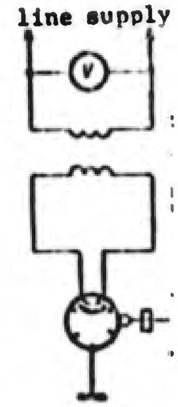


Figure 8.10. Arrangement for connecting the voltmeter when measuring the filament voltage of a magnetron in the dynamic mode.

on the basis of Lissajou Figures.*

Measurement of the amplitude and shape of the modulating voltage pulse. The modulating voltage pulse (Figure 8.11) is characterized by the amplitude, duration and shape.

The shape of the modulating voltage pulse is determined by:

- the average rise time of the front of the modulating pulse, $\tau_{r\ av}$;
- the amount of the ripple on the flat part of the pulse, ΔU_{a_2};
- the slope of the pulse flat part, ΔU_{a_1};
- the decay time of the pulse, τ_d;
- the amount of the positive overshoot after the main pulse has ended, ΔU_{a_0}.

The voltage amplitude of the modulating pulse is measured with a pulse voltmeter which is connected according to the arrangement shown in Figure 8.12 or by means of a high-voltage oscillograph. When these instruments are not available, one may use a static voltmeter which is connected to the magnetron's cathode by means of a special adaptation (Figure 8.13). It is impossible to directly connect the instrument in view of the fact that the static voltmeter registers the average effective value of the voltage which is smaller by a factor of $\sqrt{q}$ than the pulse voltage, where q is the off-duty factor.

*Translator's Note: When a point undergoes two periodic motions, which are at right angles to each other, the resultant movement of the point traces a curve which is called a Lissajou Figure.

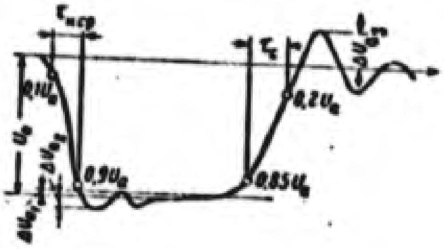


Figure 8.11. Modulating voltage pulse.

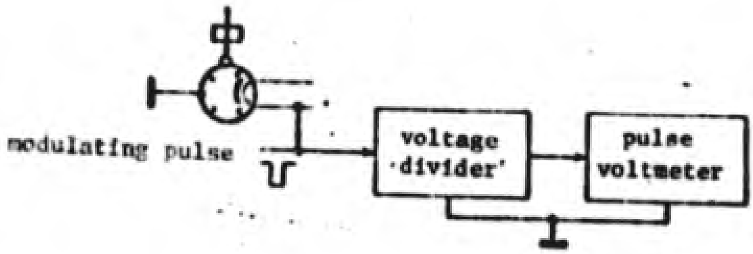


Figure 8.12. Arrangement for turning on a pulse voltmeter for measuring the amplitude of a modulating voltage pulse.

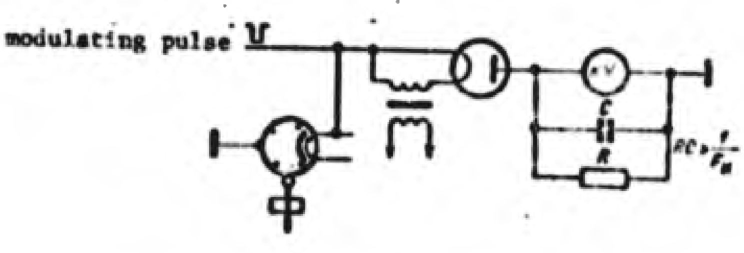


Figure 8.13. Arrangement for turning on a static voltmeter with an attachment for measuring the amplitude of a modulating voltage pulse.

When a kenotron is used for the attachment it must be of such a type that the value of the allowable reverse voltage in it exceeds the amplitude of the voltage being measured by no less than a factor of 1.5.

Measuring the duration and monitoring the shape of the modulating pulse must be performed by means of an oscillograph when the magnetron anode current is at the rated value. In order to eliminate distortions in the oscillograph amplifier circuits, the pulse being monitored must be fed directly to the plates of the oscillograph cathode-ray tube. The tube plates and the magnetron cathode are connected by a capacitive scaling circuit (Figure 8.14) which may be formed from the interelectrode capacitance of a high-voltage kenotron of type V1-01/30 and the capacitance of the tube plates. In order to reduce the voltage on the tube plates, a capacitor is connected in parallel to them. The capacitance of this capacitor is selected on the basis of a calculation of the value which is needed in order to obtain the necessary pulse amplitude and it lies in the limits of 100 - 300 picofarad. In order to observe the front of the modulating pulse on the oscillograph screen, it is necessary to use an external synchronization mode for the oscillograph. The synchronizing pulse must lead the modulating pulse by an interval of time equal to the duration of its leading front. In order to evaluate the quality of the pulse, the image obtained on the screen should be photographed or copied on tracing paper and the oscillograph calibration markers plotted. The calibration markers enable one to determine the pulse duration which is measured at the level of 0.85 of the amplitude value of the pulse.

The average slope of the pulse front is determined from the equation

$$S_{\text{ср}}=\frac{0.8U_a}{\tau_{\text{ср}}},$$

where $\tau_{r\ av}$ is the length of the pulse front measured in the voltage interval from 0.1 to 0.9 of the amplitude value.

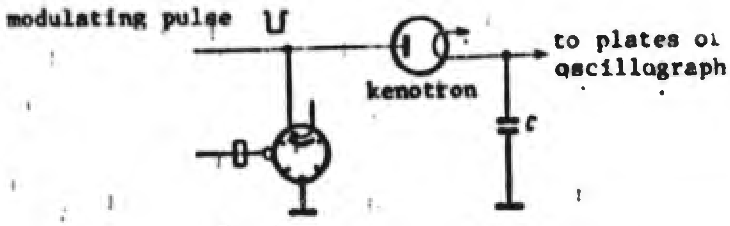


Figure 8.14. Arrangement for measuring the shape of the modulating voltage pulse.

It was experimentally established that in order to excite oscillations of the opposite phase, the average slope of the pulse front must be 100 - 150 kv/microsec.

With a small slope of the modulating pulse front, significant changes of the magnetron frequency occur and also oscillations of low-voltage modes may be excited. This produces broadening and a distortion of the frequency spectrum.

When the slope of the front is larger than the indicated value an unstable condition arises in the magnetron. This condition shows up as a fluctuation in the front of the pulses being generated and in misfirings. The sloping-off of the pulse flat part is characterized by the ratio $\frac{\Delta U_{a1}}{U_a} 100\%$.

The pulsation of the flat part of the pulse is also characterized by a similar ratio $\frac{\Delta U_{a2}}{U_a} 100\%$.

The values of the sloping-off and pulsation of the flat part of the pulse must not exceed a specified value. The acceptable deviation of the voltage on the pulse peak is determined by the magnetron electron frequency drift which causes an undesirable broadening of the spectrum of the radiated pulse. Since the broadening of the main part of the spectrum approximately equals the maximum frequency change

Δf_{max} and the width of the main part of the high-frequency spectrum of a pulse having a rectangular shaped envelope has a value equal to $2/\tau_p$, the broadening of the spectrum will not turn out to have a significant influence on the radar effective range when the following condition is fulfilled

$$\Delta f_{макс} < \frac{2}{\tau_{и}}.$$

In practice it is considered acceptable when the amount of the sloping-off and the pulsations of the flat part of the pulse do not exceed 3% of the amplitude value of the voltage.

The decay time of the modulating pulse is measured in the voltage interval from 0.2 to 0.85 of the amplitude value and must be no more than 20 - 30% of the pulse duration. In modulators with partial discharge of the reservoir capacitor, a considerable increase of the modulating pulse decay time (up to 10 or more microseconds) may be caused by a malfunction of the choke circuit shunting the magnetron. In connection with this, the magnetron in the pulse decay section will generate noise and parasitic oscillations of low-voltage modes, which interfere with the reception and display of the reflected pulses from targets. Moreover a substantial trailing-off of the pulse decay increases the radar minimum effective range.

Measurement of the amplitude and duration of a magnetron anode current pulse. Since significant changes of the anode current correspond to small changes of the anode voltage, one is able to judge more reliably as to the shape of the modulating pulse peak on the basis of the shape of the current pulse. The monitoring is achieved by means of an oscillograph. A voltage pulse taken from a noninductive resistor R_1 connected between the magnetron anode and the chassis (Figure 8.15) is fed to the vertical deflection plates of the oscillograph. The shape of the voltage pulse in this resistance duplicates the shape of the magnetron anode current pulse. The size of resistance R_1 is selected in such a way that a voltage pulse with

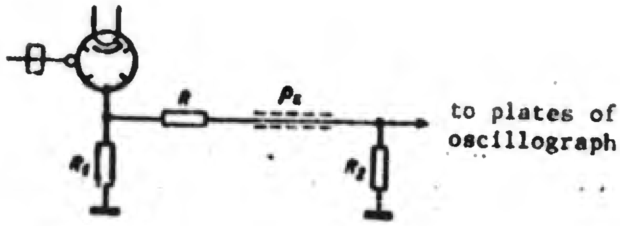


Figure 8.15. Arrangement for measuring the shape of the magnetron current pulse.

an amplitude of 50 - 150 V is obtained at the oscillograph input.

In order to eliminate distortions of the pulse being monitored it is necessary to match the cable with the load. This may be done by connecting a resistance equal to the cable wave impedance to the vertical deflection plates of the oscillograph in parallel. It may also be done by connecting a resistance $R = \rho_K - R_1$ in series with the cable, where ρ_K is the cable wave impedance.

The duration of the magnetron anode current pulse is measured at the level of 0.5 of the amplitude value.

The pulse amplitude is determined from the expression

$$I_a = \frac{dl}{R_1},$$

where d is the sensitivity of the oscillograph cathode-ray tube in V/mm;

l is the height of the pulse on the oscillograph screen in mm.

Figure 8.16 shows current pulses of a magnetron. With normal operation of the magnetron, the current pulse has a sharp outline (Figure 8.16,a). The slope of the flat part of the pulse and the pulsations does not exceed 10% of the amplitude value. Splitting of the pulse peak (Figure 8.16,b) or the presence of steps on it (Figure 8.16,c) attest to a change of the oscillation mode from pulse to pulse and during the duration of a single pulse. Moreover such a

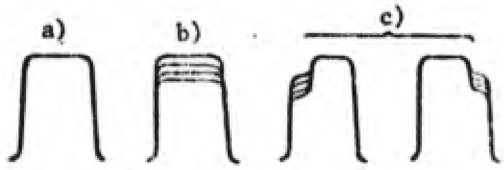


Figure 8.16. Shape of a magnetron current pulse:

a - with normal operation of the magnetron; b - with generation of an undesirable oscillation mode; c - with transition to excitation of an undesirable mode during the generation of individual pulses.

shape of the magnetron current pulse will occur when there is mismatching between the magnetron and the load.

Checking a magnetron before installing it in a station. Installing a magnetron in a radar and preparing it for operation is a rather laborious process requiring definite skills and observance of a number of conditions and precautions. Nonfulfillment of these conditions may entail the breakdown of the magnetron and the nonproductive expenditure of time.

Before arriving at a decision to replace the magnetron, it is necessary to reliably establish that the cause of the radar failure is the magnetron. A change of the magnetron output parameters which is caused by its natural aging may be detected on the basis of such signs as a reduction of the generated power, a reduction of the anode current while the anode voltage remains at the rated value, and a gradual decrease of the frequency being generated. A reduction of the anode current is connected with exhaustion of the cathode, and a change of the frequency is caused by a deposition of cathode material on the lamellae of the anode unit.

A magnetron which has been installed in a radar must be checked by an external inspection and simple checks. The purpose of the external inspection is to make sure there are no mechanical defects. When doing this, the main attention should be paid to the condition of the window for extracting the energy and to the smoothness with which the device for retuning the frequency operates. A magnetron having mechanical defects or cracks and scratches on the energy extraction window is unsuitable for operation.

The simplified checks enable one to establish such malfunctions as a breakdown of the vacuum, the presence of a short circuit in the heater, shorting of the cathode to the anode, and a break in the heater circuit without switching the magnetron into the generating mode.

Checking the vacuum may be done on the basis of the filament current of the heater. An increase of the filament current when the filament voltage is at the rated value of more than 15% with respect to the value indicated in the certificate attests to a breakdown of the vacuum and to the unsuitability of the magnetron for operation. The increase of the filament current associated with the breakdown of the vacuum is explained by the intensification of the cathode cooling as a result of the gas liberated by the metal of the assembly. This leads to a reduction of the heater resistance.

A break or a short circuit in the heater is detected rather simply by a change in the value of its resistance.

After the indicated checks, a magnetron not having defects may be mounted in the station. When mounting it, all the magnetron fastening operations must be done with an instrument made of non-magnetic materials in order to avoid demagnetizing the magnetic system. The magnetron mounting must avoid mechanical stresses in the glass and ceramic parts of the magnetron and also prevent its shifting during operation. Special attention must be paid to the mechanical connection for the extraction of the magnetron energy with the flange of the wave guide channel. The connection must be reliable and ensure precise matching of the wave guide openings. Misalignment between the wave guide openings causes a sharp decrease of the channel TWR which leads to melting of the dielectric in the coupling window and to failure of the magnetron.

When the conductor carrying the filament voltage is connected to the heater leads, one should make certain that the wire running from the terminal of the filament transformer connected with the modulator output is connected to the lead marked with a dot or with the letter "K" (Figure 8.17,a). In case of incorrect connection (Figure 8.17,b) the entire anode current passes through the heater which may cause failure of the magnetron.

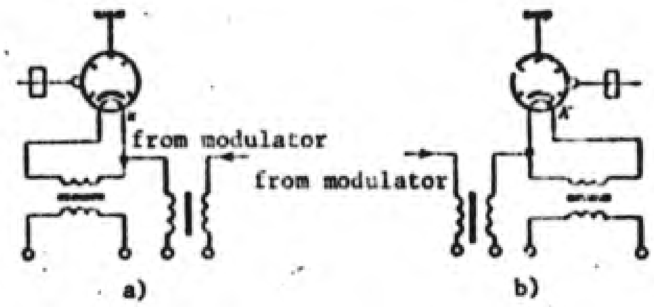


Figure 8.17. The arrangement for connecting the filament circuit to a magnetron:

a - the correct connection; b - an incorrect connection.

If the magnetron has been stored for a long time and during storage it did not undergo conditioning, then before it is switched on to the rated conditions it should pass through a preliminary processing. The necessity for the preliminary processing is due to the fact that during the storage, as a result of liberation of gas by the internal parts of the magnetron, the vacuum deteriorates. When anode voltage is applied to it, a gas discharge begins as a result of ionization. This gas discharge causes arcing in the magnetron and may in the final result lead to its failure. The arcing shows up in the form of unsteady readings of the "magnetron current" instrument.

The essence of the preliminary magnetron processing is that, when it is warmed up, a reverse process takes place — the process of absorption of gas by the magnetron internal parts. The absorption will be more intense when an electric field is present between the magnetron cathode and anode. The strength of the electric field must not cause a gas discharge. The preliminary processing consists of increasing the anode voltage not at once but gradually, passing through a series of discreet values. The discreet values of the anode voltage which are selected should correspond to the anode current value at which unstable operation of the magnetron arises. For each value of the anode voltage, the selected condition is maintained until the magnetron operation becomes stable. After the magnetron operates stably for 10 - 15 minutes at the rated anode voltage, the preliminary processing may be terminated.

8.2. A High-Frequency Amplifier Using a Traveling Wave Tube

H-f amplifiers using traveling wave tubes (TWT) have found wide use in receiver devices of modern radars.

The main advantages of amplifiers using TWTs are:

- a low noise factor (3 - 10 dB);

- large power amplification of a high-frequency signal (up to 15 - 35 dB) which practically eliminates the effect of noise from the mixer, klystron and i-f amplifier on the receiver sensitivity;

- the broad range of frequencies is uniformly amplified (the frequency range of low-noise TWTs in which an amplification of more than 20 dB is maintained is hundreds and thousands of megacycles);

- the ability to attenuate high-power signals due to the distinctive features of the amplitude characteristic of a TWT. As a consequence of this, reliable protection is ensured for semiconductors against the action of the main pulses and pulses of high-frequency oscillations from radars located nearby.

The amplification of high-frequency oscillations in a TWT (Figure 8.18) is based on the interaction of an electromagnetic wave with an electron flux moving in the same direction as the wave, but with a velocity somewhat larger than the wave phase velocity.

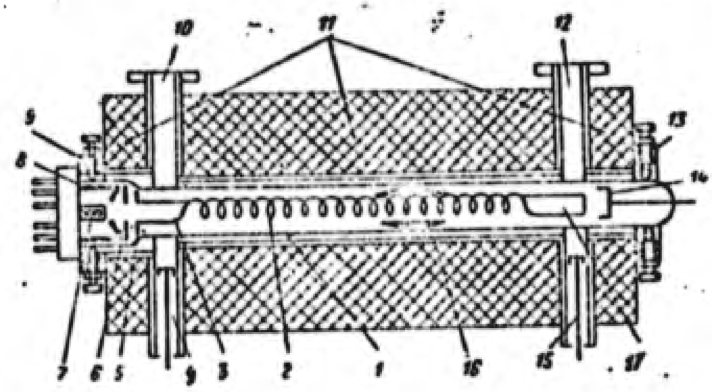


Figure 8.18. The schematic diagram of a TWT device.

Fundamentally a TWT is a glass cylinder with a length l of around 30 cm and a diameter of 1.5 - 2 cm. A pressure of 10^{-7} - 10^{-9} mm of Hg is created inside the cylinder. A decelerating system 2 made in the form of a spiral is positioned in the central part of the cylinder. This decelerating system is used for reducing the phase velocity of the electromagnetic wave. Miniature rod antennas 3 are attached to both ends of the spiral. The miniature antennas fulfill the role of elements for coupling the input and output of the decelerating system with the high-frequency transmission line. An electron gun is mounted at one end of the cylinder which forms the base. This electron gun consists of a cathode 7, a control electrode 8, a first anode 6 and a second anode 5. The electron gun is used for creating the electron flux of the necessary density, velocity and ahape. There is a collector 14 at the other end of the cylinder. The purpose of the collector is to catch the electrons after they emerge from the spiral.

The tube is placed inside a reinforcing framework which structurally is a cylindrical pipe 16. The pipe together with the tube spiral form a coaxial line. In order to connect this line with the high-frequency transmission line, an input wave guide section 10 and an output wave guide section 12 with tuning elements 4 are connected to the pipe at the places where the coupling elements are located. A focusing system is installed around the pipe. The focusing system is either a solenoid 11 in tubes with electromagnetic focusing or a series of magnetic rings in tubes with periodic magnetic focusing. In order that the axis of the electron flux coincides with the spiral axis, which is necessary in order to prevent electrons from hitting the spiral, a centering device 13 is provided for in the reinforcing framework.

The Main Parameters of a TWT and Their Dependence on the Power Supply Conditions

A traveling wave tube is characterized by a large number of parameters. Many of them are of interest only for particular cases of TWT use. Therefore, we will limit ourselves to an investigation of only those parameters which are the controlling parameters when a TWT operates as a high-frequency amplifier.

The following is a list of these parameters:

- the amplification factor;
- the noise factor;
- the operating range of frequencies.

The amplification factor K_a is the quantity characterizing the capacity of a TWT to transform energy and it numerically equals the ratio of the power at the tube output P_{out} to the power which is fed to its input P_{in}:

$$K_a = \frac{P_{out}}{P_{in}}.$$

The amplification factor usually is expressed in decibels:

$$K_y = 10 \lg \frac{P_{вых}}{P_{вх}} \text{ dB.}$$

The amplification factor of a TWT changes within broad limits. This is explained by its dependence on many factors, the main ones of which are: the amount of the input power, the tube electric parameters and the signal frequency.

Let us investigate the degree of influence of the listed factors on this most important parameter of the tube, and let us determine the possibility of influencing them during operation.

The dependence of the output power and consequently also of the amplification factor on the power of the input signal is graphically illustrated by the amplitude characteristic of a TWT. As is seen from Figure 8.19 three characteristic regions corresponding to various operating conditions may be distinguished on the amplitude characteristic.

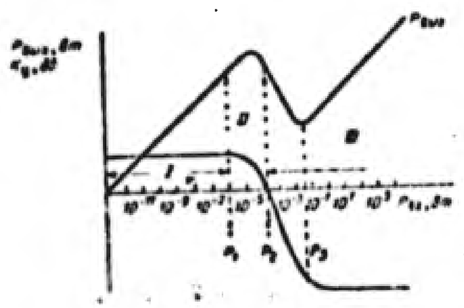


Figure 8.19. The amplitude characteristic of a TWT:

I - region of linear conditions; II - region of non-linear conditions; III - region of suppression.

Operation in the first region is characterized by a linear dependence between the power at the output and input of the TWT. The amplification factor in the linear region is constant and does not depend on the input signal power, which ensures amplification of the signal without distortion. When the input signal reaches a power level P_1, the amplifier operates in a nonlinearity region. The amplification factor in this nonlinear region is smaller than in the linear region and it is reduced to a greater extent, the larger the input power. Consequently, there is distortion of the signal being amplified. The upper limit of this region corresponds to that level of the input power P_2 at which the amplification factor equals one. A further increase of the input signal power causes the TWT to operate in a suppression mode for which a negative value of the amplification factor is characteristic. The tube attenuates an input signal, and the attenuation is more intense the greater the input signal power. Starting with the value of the input signal power P_3, the tube enters a mode in which it acts as a passive transmission line with an attenuation of 30 - 60 dB.

One should point out that the capacity of a TWT to attenuate signals with a power of more than 10^{-4} watts has a favorable effect on the operation of radars. In receiver devices without TWTs, when high-power input signals act on it, the noises and conversion losses of the semiconductor mixer increase significantly. This leads to a reduction of the receiver sensitivity.

The amplification factor of a TWT depends to a considerable degree on the tube electric parameters. As follows from the principle of operation of a TWT, in order to obtain a given amplification factor it is necessary to select electric conditions such that:

- the electron flux entering the decelerating system input completely falls on the collector, not branching off in the spiral. Also it must interact with the braking superhigh frequency field along its entire path, since only in this case will it deliver maximum energy to the electromagnetic wave;

- the losses are minimal when a signal is introduced at the TWT input and the output signal is removed.

The first condition is fulfilled with the precise centering of the tube and the correct focusing of the electron flux. In a TWT with electromagnetic focusing, the degree of a focusing of the beam is determined by the value of the solenoid current. The dependence of the output power and consequently also of the tube amplification factor on the solenoid current is shown in Figure 8.20. It attests to the fact that there is an optimum value of the solenoid current at which the amplification factor has a maximum value even if there is no critical value.

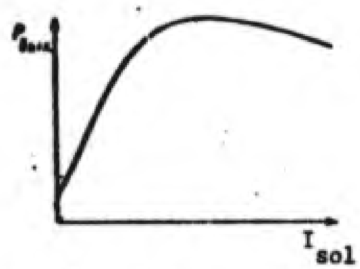


Figure 8.20. A TWT output power as a function of the solenoid current.

A reduction of the solenoid current by 15 - 20% of the rated vlaue leads to substantial defocusing of the beam. When this occurs, a considerable part of the electrons hit the spiral and the interaction of the electron flux with the electromagnetic wave deteriorates. As a consequence of this, the output power and also the tube amplification factor drop.

An increase of the solenoid current above the rated value also reduces the TWT amplification factor. This is explained by the fact that, with excessive compression of the electron flux, it moves away from the spiral surface and interacts with the wave in a region where the high-frequency electromagnetic field is weaker.

The focusing quality also depends on the voltage on the first anode and control electrode of the TWT. Moreover the voltage on the electrodes determines the magnitude of the beam current. As was shown in [21], the change of the amplification factor (in decibels) is proportional to the cube root of the beam current. Consequently, a deviation of the voltage on the first anode and control electrode from the established values will be especially pronounced for tubes with a high amplification factor. Let us investigate this with an example.

Let us assume that the beam current increases by a factor of 3 because of a voltage change on the gun electrodes. When this occurs, the amplification factor increases by a factor of 1.44 and for tubes with an initial amplification factor of 15 dB and 35 dB, respectively, it will equal 21.6 and 50 dB. For the first tube the change is 6.6 dB or 4.5 times, and for the second tube it is 15 dB, that is 31.5 times.

Since low-noise TWTs which have a rather high amplification factor are used in radars as the high-frequency amplifier, the setting of the optimum electric tube parameters must be performed with great care.

The values of the voltages on the TWT electrodes which correspond to the optimum conditions are indicated in the certificate affixed to each tube. The effectiveness of the interaction of the electron flux with the electromagnetic wave depends mainly on the voltage applied between the decelerating system (spiral) and the cathode. As is seen from the graph of Figure 8.21 there is a specific value of the voltage on the decelerating system, which is called the optimum interaction voltage, at which the TWT amplification is a maximum. The amplification factor strongly depends on the voltage on the spiral. A deviation of the voltage from the optimum value by only 5% reduces the amplification by 8 dB, that is by a factor of 6.3.

One should note that the optimum voltage of the decelerating system (spiral) depends on the value of the input power and frequency A larger input power and a lower frequency corresponds to a higher optimum voltage. The difference of the optimum voltage from the spiral voltage value indicated on the certificate may be explained by this dependence. As will be shown below, the voltage setting on the decelerating system is made on the basis of maximum amplification when tuning is performed.

The amplification factor as a function of the input signal frequency is shown graphically by the TWT frequency characteristic (Figure 8.22). The tube amplification factor in the operating range of frequencies does not remain constant. This is explained by the

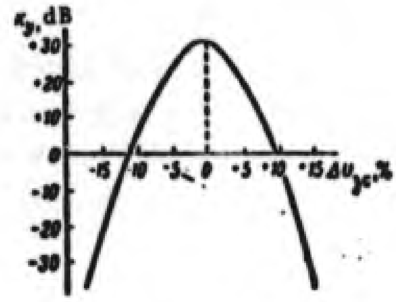


Figure 8.21. A TWT amplification factor as a function of the change of the voltage applied to the decelerating system.

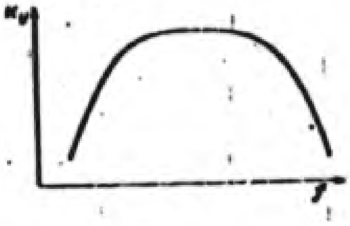


Figure 8.22. The frequency characteristic of a TWT.

fact that with a change of the frequency, the phase velocity of the wave changes, even if only slowly. As a result of this, while the voltage of the decelerating system (spiral) remains constant, the optimum relationship between the velocity of the electron flux and the phase velocity of the wave is disturbed. Moreover, it is impossible to achieve complete matching of the TWT input and output with the station high-frequency channel over a broad range of frequencies. This also has an influence on the change of the TWT amplification factor.

The change of the amplification in the operating range of frequencies is evaluated on the basis of the nonuniformity of the amplification, by which we mean the ratio of the maximum amplification to the minimum amplification. For low-noise TWTs, the nonuniformity of the amplification is 6 - 8 dB.

As a consequence of the nonuniformity of the TWT amplification in radars having a tuneable magnetron, the frequency tuning of the high-frequency amplifier must be done at the mean frequency of the retuning range.

<u>The noise factor F</u> is the controlling parameter of low-noise TWTs.

By the noise factor of a TWT, we understand the number which indicates how many times the signal-to-noise ratio of the power at the tube output is smaller than at its input. The noise factor of modern low-noise TWTs is 3 - 10 dB. The lower the noise factor, the better the TWT operates as a h-f amplifier.

Let us investigate which factors affect the noise factor and what measures may be taken in order to reduce it during radar operation. The main source producing noise is the chaotic change of the electron flux which is caused by the nonuniform output of electrons from the cathode, by their rearrangement in the beam and by the capture of part of the electrons by the tube electrodes. Consequently, a reduction of the noise level may be achieved by reducing to a minimum the causes giving rise to a change of the electron flux parameters (density, velocity, shape).

A reduction of the noise caused by fluctuations of the electron flux density and velocity in the cathode-decelerating system (spiral) space is provided by the construction of the electron gun and by the selection of the voltages on its electrodes. The setting of the power supply voltages corresponding to the minimum noise factor is performed experimentally at the manufacturing plant for each individual TWT. The optimum values of the power supply voltages are recorded in the tube certificate. During operation these voltages must be monitored since a deviation from the conditions indicated in the certificate leads to an increase of the TWT noise factor (Figure 8.23).

As is seen from Figure 8.23,e, with a simultaneous change of the voltages on all the electrodes of less than 5% from the optimum value the noise factor increases by more than a factor of two. The noise factor depends most strongly on the voltage on the decelerating system (spiral). It was shown above that this voltage also has a substantial effect on the amplification factor. One should note that agreement between the decelerating system voltage values corresponding to the minimum noise factor and the maximum amplification factor is

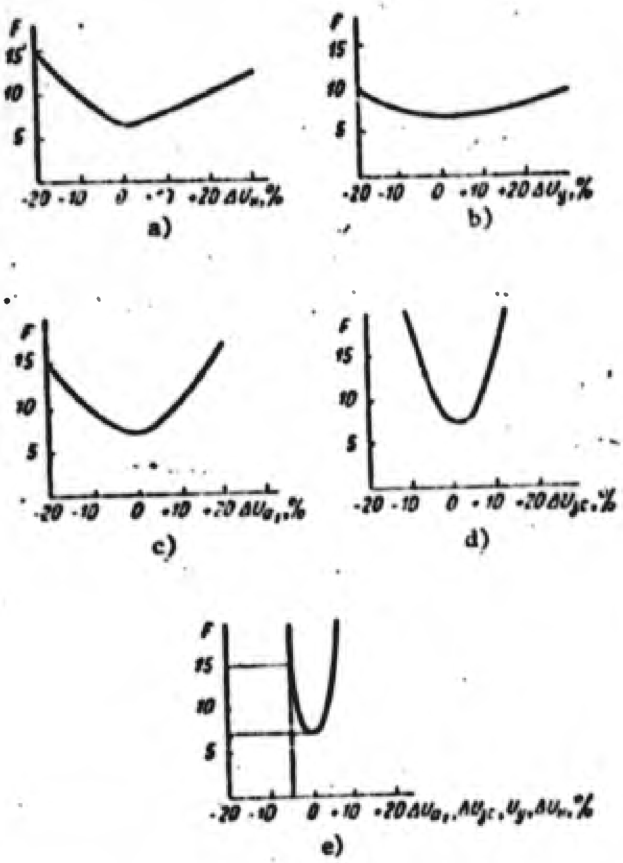


Figure 8.23. The dependence of TWT noise factor in connection with deviations of the voltages on its electrodes from the optimum values:

a - filament voltage; b - control electrode voltage; c - first anode voltage; d - decelerating system (spiral) voltage; e - simultaneous change of all the voltages.

characteristic for low-noise TWTs (Figure 8.24). This simplifies the tuning of a TWT.

An increase of the noise level also is connected with defocusing of the electron flux, since in this case forces arise which cause thermal motion of the electrons and in addition part of them begin to be captured by the decelerating system spiral. Consequently, a decrease of the solenoid current has an unfavorable affect on the noise factor. This solenoid current decrease may be caused by an increase of the resistance of the windings as a consequence of their heating. Therefore in a majority of cases, forced cooling of the TWT is used.

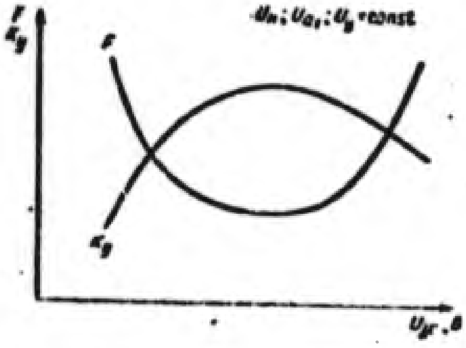


Figure 8.24. The amplification factor and noise factor of a TWT as a function of the voltage on the decelerating system.

The noise factor indirectly depends on the matching of the TWT input with the high-frequency transmission line, since reflections of the useful signal because of a nonuniformity in the line lead to a decrease of the signal/noise ratio at the TWT input. This is equivalent to an increase of the noise factor.

<u>The operating range of frequencies of a TWT</u>. A characteristic feature of a TWT is the broad range of uniformly amplified frequencies. In order to evaluate this property of a TWT, the concept of an "operating range of frequencies of a TWT" is introduced. The operating range of frequencies is the region of frequencies within the limits of which a specified amplification factor is maintained. Its value is established at the manufacturing plant and depends on the type of tube. Since the concept of "operating range of frequencies of a TWT" is equivalent to the concept of a "pass band" for ordinary amplifiers, the operating range of a TWT very frequently is not determined on the basis of a specified amplification factor, but at a certain level — 3 dB (for the power).

A given change of the amplification factor associated with a change of the input signal frequency is connected with the fact that with a given voltage of the decelerating system (spiral), the most effective exchange of energy between the electromagnetic wave and the electron flux occurs for a definite phase velocity of the wave. This depends on the frequency. Since the phase velocity only slightly changes with a change in the frequency, a TWT with a fixed voltage on the spiral has a uniform amplification over a broad range of frequencies. This quality of a TWT is especially valuable for radars in which shifting of the transmitter operating frequency is possible. With a TWT it is not necessary to retune the high-frequency amplifier when shifting from one operating frequency to another.

Checking the Electric Conditions of a TWT

When investigating the main parameters of a TWT, it was shown that the effective use of a tube is determined to a considerable degree by the electric conditions which are characterized by:

- the voltages on the tube electrodes;
- the current distribution between the tube electrodes;
- the solenoid magnetic field intensity and its distribution along the tube axis.

Both a DC voltage and an AC voltage are required for the TWT power supply:

- an adjustable stabilized DC voltage for feeding the control electrode;
- an adjustable stabilized DC voltage for the first anode power supply;
- an adjustable stabilized DC voltage for the second anode (spiral) power supply;
- an unregulated stabilized DC voltage for the collector power supply;
- an adjustable AC voltage for the filament power supply;
- an unstabilized DC voltage for the solenoid power supply.

The necessity for the adjustment of certain of the power supply voltages is caused by the fact that each individual TWT has electric parameters which are peculiar only to it and distinct from other types of TWT.

A typical power supply circuit for a low-noise TWT is shown in Figure 8.25. Using this circuit as an example, let us investigate how the TWT parameters are established and how one checks that their setting is correct. The power supply of the tube itself is a rectifier with electronic stabilization of the voltage. The solenoid power supply is provided from a separate, nonstabilized source. In order to obtain several different voltage values from a single rectifier, a voltage divider made up of the resistors $R_1 - R_6$ is used.

The voltage from the divider is fed to the collector. Usually the voltage on the collector is not regulated, since its value is a constant for various individual tubes of the same type. Adjustment of the voltages on the spiral, first anode, and control electrode is done by means of potentiometers R_2, R_4 and R_6, respectively. The voltage adjustment limits are selected on the basis of the scatter of the parameters of a given type of TWT. In the circuit being investigated, the stability of the power supply voltages depends on the relationship of the currents of the divider and load. The divider current must exceed by no less than a factor of 10 the total current of all the electrodes of the tube. In this case, a change of one voltage will have practically no effect on the others. The filament power supply is provided from a separate winding of the transformer and is regulated by potentiometer R_7. The voltage drop across the potentiometer is maintained constant by means of a current regulator (battery) L_1. This eliminates the effect of a voltage change of the radar power supply source on the value of the filament voltage and consequently also on the tube noise factor.

Monitoring jacks $G_1 - G_6$ are provided for measuring the voltages on the tube electrodes with the setting and monitoring of the TWT conditions. The measurements should be made with a DC voltmeter of precision class no lower than 1.0 with an internal resistance of not less than 0.5 Megohm. In some stations there is a built-in instrument for monitoring the power supply voltages. This instrument is connected to the various TWT circuits by means of a switch.

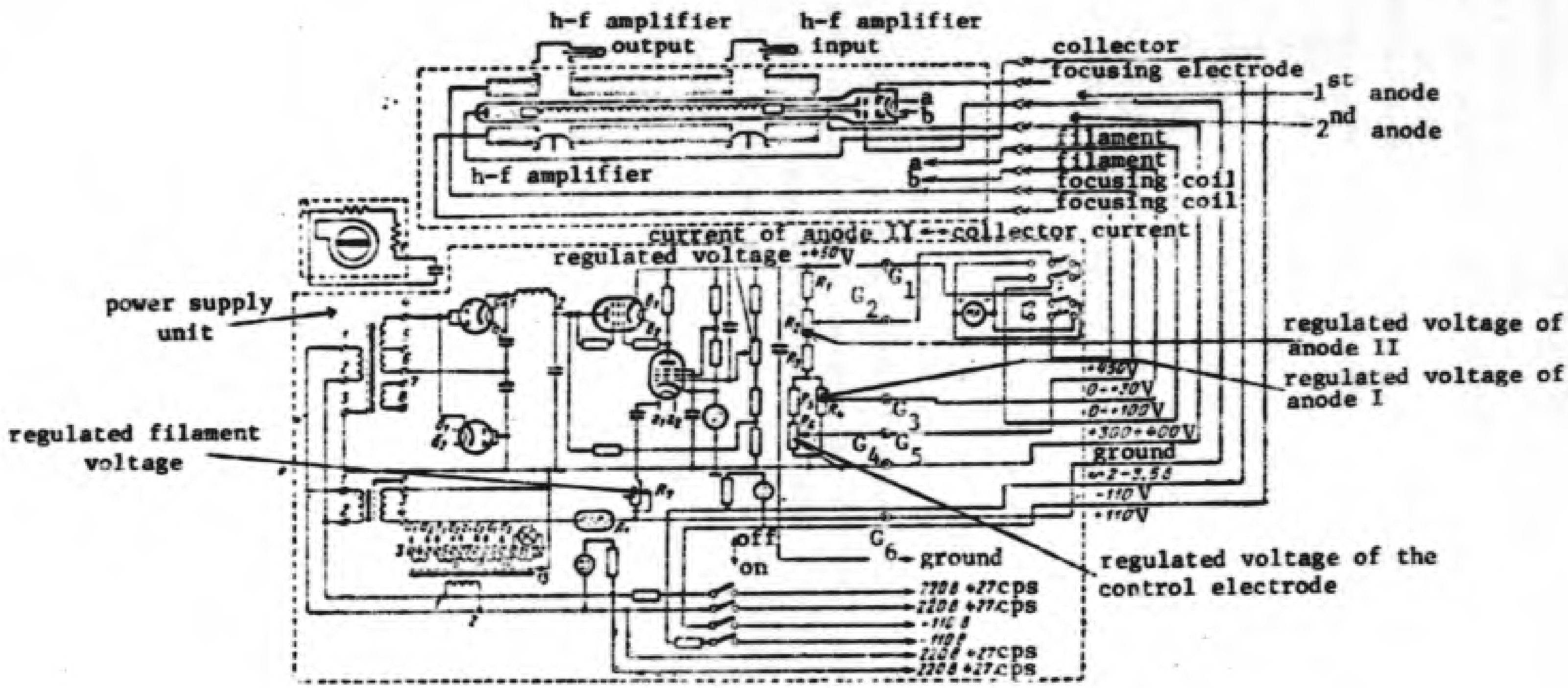


Figure 8.25. A typical power supply circuit for a low-noise TWT.

Earlier it was shown that the electron beam current (collector current) has a considerable influence on the amplification factor and noise factor of a TWT. The size of this beam current depends on the voltages on the tube electrodes (Figure 8.26). A deviation of the beam current from the rated value attests to a disturbance of the TWT electric parameters. Consequently one may, with sufficient efficiency and speed, monitor a TWT operating efficiency by measuring the collector current. It is also very important to measure the second anode (spiral) current.

An increase of the second anode current arises with poor centering of the tube or a disturbance of its focusing which may be caused by a change of the power supply voltage and a reduction of the intensity of the longitudinal magnetic field. The magnetic field intensity and its distribution along the axis of the magnetic system are selected when the solenoid is designed and are not adjusted during operation.

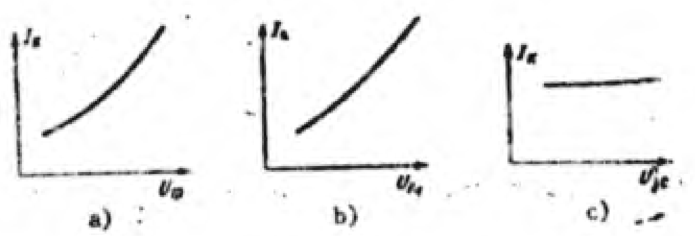


Figure 8.26. The TWT collector current as a function of the voltage of the focusing electrode (a), of the first anode (b) and of the decelerating system (c).

For low-noise TWT's, the magnetic flux along the tube must be uniform. The deviation of the magnetic field intensity at the places where the input wave guide and output wave guide adjoin the reinforcing framework must not be larger than 20% of the intensity at in the middle part of the magnetic system. During operation it may be necessary to dismantle and repair the magnetic system, as a result of which disturbances of the field structure are possible. These disturbances may be detected only by magnetic measurements. A typical form of the magnetic field distribution along the solenoid axis is shown in Figure 8.27.

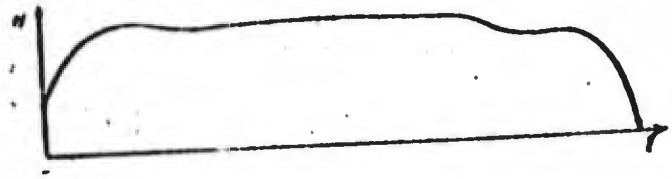


Figure 8.27. The distribution of the magnetic field along the solenoid axis.

8.3. Semiconductor Mixers

The receiver devices of the majority of radars are based on superheterodyne circuits, the standard element of which is a frequency mixer. The overall design and selection of the mixer nonlinear element are determined by the range in which the radar operates. Vacuum triodes and diodes in the decimeter range and semiconductor diodes in the centimeter wavelength range have found wide application as mixers in receiver devices.

The number of diodes used for converting the frequency serves as the basis for classifying mixers. Mixers are distinguished as half-wave mixers and balanced mixers.

The construction of a half-wave coaxial mixer operating in the range 2,500 - 3,500 Mc is shown in Figure 8.28. The signal from the cavity resonator of the ATR tube is sent through a coupling loop 1 to the coaxial line 7. The load for the line is a semiconductor diode 2. The diode input impedance must be matched with the line wave impedance since only in this case will there be no losses of the reflected signal. This is achieved by selecting the diameter of the coaxial line inner conductor and the distance between the diode and the coupling loop. A coaxial T-joint 5 is connected to the main line 7 at a distance of $\lambda/4$ from the coupling loop. The heterodyne signal

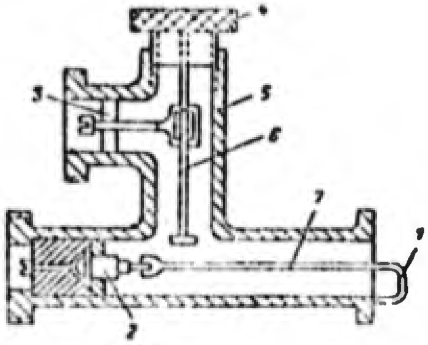


Figure 8.28. The construction of a coaxial diode mixer.

is fed to the diode through this T-joint. The coupling of the heterodyne with the mixer is capacitive and may be adjusted by displacing pin 6 by means of screw 4. In order to reduce the effect of a change in the coupling value on the operation of the heterodyne, a disk resistor 3 is included in the heterodyne line. The resistance of this disk resistor approximately equals the wave impedence of the heterodyne line.

Figure 8.29 shows the construction of a half-wave wave guide mixer. The coupling of the mixer with the discharger is achieved through a coupling window 1, the dimensions of which are selected on the basis of matching conditions. The heterodyne energy is fed to the wave guide section by means of probe 2 of klystron 3. The amount of power fed to diode 4 from the heterodyne may be adjusted by changing the immersion depth of the probe into the wave guide. The distance from the probe to the coupling window must equal an odd multiple of λ/8 in order to prevent the passage of the heterodyne signal to the antenna. The diode input impedance is matched with the wave guide by means of a plunger 5.

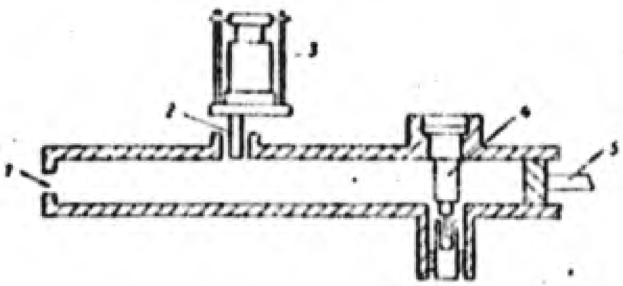


Figure 8.29. Construction of a half-wave wave guide mixer.

The circuit of a half-wave mixer in which a directional coupler is used as the coupling element between the klystron and the mixer is a rather simple circuit (Figure 8.30). In order to reduce the reflected signal losses in the heterodyne circuit, the coupling attenuation of the coupler must be no less than 13 dB. Adjustment of the amount of the power P_G sent to the mixer is made by means of an attenuator A.

Half-wave diode mixers have very serious shortcomings, among which one should first of all list:

- the transmission of heterodyne noise to the input of the i-f amplifier which increases the overall noise level and reduces the amplifier sensitivity;

- the shunting of part of the reflected signal energy to the heterodyne circuit;

- the penetration of part of the heterodyne energy to the antenna, as a result of which the heterodyne becomes a source of interference for other radars operating in this same range;

- the absence of decoupling between the klystron and the discharger which may be the cause of unstable operation of the klystron.

A balanced mixer circuit is free of these defects to a considerable extent. A double wave guide T-joint (Figure 8.11) is the basis of a balanced mixer in the centimeter wave range. The energy of the received signal and the energy of the heterodyne are sent through branches 1 and 2, respectively. Such a connection eliminates the coupling between the discharger and the klystron.

Use of a balanced mixer increases the receiver sensitivity as a result of suppression of the heterodyne noise. This is explained by the fact that, in a balanced mixer circuit, the currents of the intermediate-frequency signal which flow through the diodes 3 are found

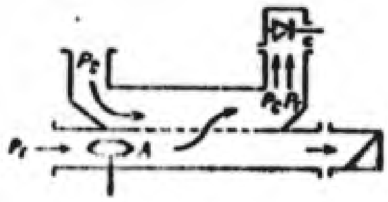


Figure 8.30. Circuit of a diode mixer with a directional coupler.

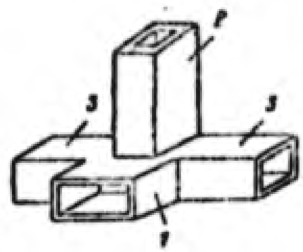


Figure 8.31. Balanced mixer with a double wave guide T-joint.

to be in opposite phase, but the corresponding noise components are in phase. The mixer output is connected to the balanced input of the i-f amplifier, as a consequence of which the signals sent from both diodes are summed, but the noises of the heterodyne components cancel.

In order to adjust the power being fed from the heterodyne to the mixer, a variable attenuator is mounted in the wave guide branch of the heterodyne.

Balanced mixers which are based on a slot bridge (Figure 8.32) have found wide use. Because of the properties of a slot bridge to divide the signal power into equal parts in the branches which are opposite the input branch and to change the phase of the oscillations when passing through the slot by 90°, a mixer of such construction ensures suppression of the klystron noise. Since adjacent branches of a slot bridge are not coupled (the amount of the decoupling is around 20 dB), such a mixer construction is free from the other defects which are inherent in half-wave mixers. There is practically no coupling between the klystron and the receiver channel in such a mixer.

Besides the indicated advantages, the use of a balanced circuit improves the protection of the semiconductor diodes, since the energy leaking through the discharger is reduced by a factor of two for each diode in comparison with a half-wave circuit. One should also mention that, when one of the diodes of the balanced circuit fails, the mixer operating efficiency is maintained with only an increase of the receiver noise factor by 5 - 6 dB.

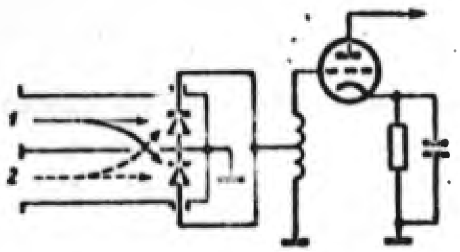


Figure 8.32. A balanced mixer with a slot bridge.

As was shown in Chapter 2, the quality of the mixer operation has a strong influence on the receiver device sensitivity, and it should be evaluated first of all on the basis of the transmission factor of the normal power k_p and on the basis of the internal noise level. For the majority of modern diode mixers, $k_p = 0.1 - 0.3$. The quantity which is the reciprocal of k_p is called the conversion loss. The signal energy losses in the mixer are the sum of losses connected with the signal reflection, with the shunting of the signal energy through the heterodyne channel, and with losses in the structure of the crystal holder and during conversion in the mixer. In order to reduce the losses caused by reflection, one may increase the mixer TWR from the signal input direction by the proper selection of the diodes . In order to reduce the losses in the heterodyne circuits, one should, when tuning the klystron, select the generating zone corresponding to the greatest power, since in this case the coupling of the mixer with the heterodyne will be weaker.

The noise level of the diode mixer is higher than the thermal fluctuations of its equivalent resistance. The noise properties of a mixer are evaluated by the relative noise temperature t_c, which is the factor which indicates how many times it is necessary to increase the temperature of the equivalent resistance in comparison with room temperature (15° C) in order that the rated power of its thermal noise be equal to the rated output noise power of the mixer. For diode mixers, the value of t_c lies in the limits of 1.2 - 3.5.

If k_p and t_c are known, one may determine the mixer noise factor:

$$F_c = \frac{t_c}{k_p}.$$

The relative noise temperature, the power transmission factor, and consequently also the noise factor of a diode mixer depend on the amount of power delivered from the heterodyne to the mixer. The indicated relationships are presented in Figure 8.33. From the figure it is seen that there is a certain value of the delivered power at which the mixer noise factor is a minimum.

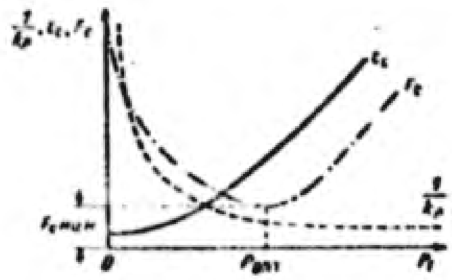


Figure 8.33. The noise temperature, conversion loss and noise factor of a mixer as a function of the heterodyne power.

The conditions of the diode mixer at which its noise is a minimum are called the optimum conditions. These conditions prevail when the receiver is tuned by selecting the value of the coupling between the heterodyne and the mixer. The conditions are monitored by means of a microammeter connected to the constant component circuit of the mixer current. The value of the mixer current constant component corresponding to the minimum value of F_c usually is indicated on its certificate, and for the majority of diodes it equals 0.4 - 0.5 mA, as indicated above.

Semiconductor diodes have a relatively low electric strength. Under the action of signals whose power exceeds acceptable values, partial or total "burnout" of the diode occurs, as a result of which the diode conversion properties deteriorate or are completely lost.

The main sources of such overly powerful signals are:

- transmitter pulses leaking through the ATR tube;

- pulses of high-frequency oscillations from the antennas of radars located nearby;

- high-power parasitic oscillations arising during unstable operation of the magnetron and leaking to the receiver input;

- static charges and induced emf arising in the low-frequency circuits when they are poorly shielded.

One may approximately evaluate the conversion properties of a diode mixer by comparing its forward and reverse resistance. The ratio of the forward resistance to the reverse resistance characterizes the slope of the diode volt-ampere characteristic, and consequently also the transmission factor (Figure 8.34). The action of signals of considerable power on the detector produces a change of the volt-ampere characteristic, which shows up as a sharp increase of the diode reverse current.

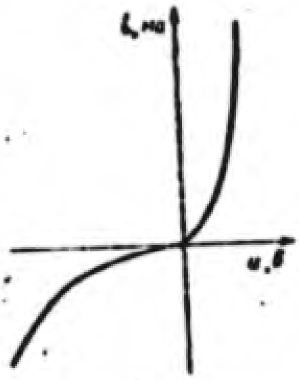


Figure 8.34. The volt-ampere characteristic of a mixer semiconductor diode.

Consequently, a reduction of the ratio of the mixer diode reverse resistance to the forward resistance indicates deterioration of its conversion properties. For diodes in good working order, this ratio equals 6 - 10, and the value of the forward resistance should lie in the limits of 400 - 800 ohm. However such a check still does not guarantee that the diode parameters are satisfactory. A complete check of diodes is carried out only by means of special instruments.

Since the mixer is a very sensitive element of the receiver and the state of its parameters determines to a considerable degree the quality of the radar operation, when working with it it is necessary to take precautionary measures and to make sure that the diode electric conditions do not exceed the maximum acceptable values. Operational experience with semiconductor diodes shows that, with an ATR tube in good working order and stable operation of the magnetron, burnout of diodes practically never occurs. A rather frequent cause of diode burnout is improper handling. Let us briefly investigate what precautionary measures should be taken when working with a diode mixer.

When installing a diode, it may be damaged by the charge of static electricity which has been accumulated on the operator body. Therefore, before inserting the diode in the holder, the operator must take hold of the mixer compartment frame with his free hand. In this way, the induced charge is drained from his body. This same rule should also be adhered to when removing a diode. In order to protect the diode from the action of high-frequency fields, it is stored in a special wrapper. After removing the diode from this wrapper, it should immediately be put in a place where there are no operating devices with high-power radiation. Diode circuits must be well shielded.

8.4. Heterodyne

A heterodyne is an integral element of a superheterodyne receiver. The heterodyne, together with the nonlinear element which is used for frequency conversion, forms a frequency converter or mixer. Let us investigate what requirements must be imposed on this element of the receiver.

Frequency stability is the most important requirement. In radars of the centimeter range, even a very small relative change of the heterodyne frequency (by several hundredths of a percent) leads to a considerable deviation of the intermediate frequency from the rated value. As a consequence of this, there will be a sharp reduction of the receiver sensitivity and a deterioration of the reproduction quality of the signal being received.

When investigating a mixer operation, it was shown that optimum signal conversion occurs at a fully determined value of the power being fed from the heterodyne to the mixer. Consequently, the second requirement is sufficiently high stability of the power of the generated oscillations.

In order to reduce the reflected signal losses in the heterodyne circuits and also to reduce the dependence of the generated oscillation frequency and power on the parameters of the mixer (which is the load for the heterodyne), the coupling between the mixer and the heterodyne must be very weak. From this it follows that the power of the oscillations being generated by the heterodyne must exceed the power required for optimum conversion of the signal by a factor of forty to sixty.

The heterodyne frequency retuning range must span the radar operating frequency range. It is desirable that the heterodyne have electronic frequency retuning. Finally, the heterodyne internal noise must be a minimum. A reflex klystron most completely satisfy these requirements in the decimeter and centimeter wavelength range. Thus the reflex klystron continues to remain the main heterodyne device for radars of these ranges. Below, main attention will be paid to questions connected with the operation of a reflex klystron.

One should mention that improvement of superhigh frequency instruments led to the creation of devices having an electronic frequency retuning range significantly broader than that for a reflex klystron. In recent years, these instruments have begun to be used as heterodynes of superhigh frequency receivers. Among them one should first of all mention traveling wave tubes and mitrons.

Reflex Klystron

A reflex klystron is an electric vacuum device in which the principle of electron velocity modulation is used for converting direct current energy into the energy of superhigh frequency oscillations (Figure 8.35). It is assumed that the reader is familiar with the theory of a reflex klystron operation. Therefore, we will limit ourselves to an investigation of the general character of the processes in an operating klystron, which is necessary in order to explain the principles of tuning it.

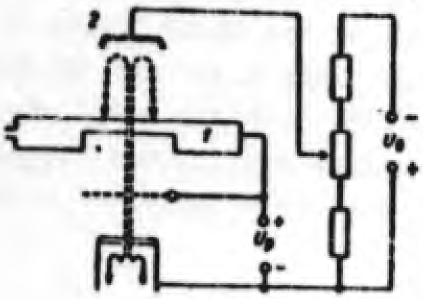


Figure 8.35. The arrangement of a klystron device.

The electron flux under the action of the voltage U_p moves with a uniform acceleration toward the resonator 1. By interacting with the high-frequency field of the resonator, the electrons acquire either a positive or a negative acceleration depending on the phase of the voltage between the resonator walls U_g. As a consequence of this, the electrons in the resonator-repeller 2 space are combined into a cluster around the electrons passing through the resonator when $U_g = 0$. These electrons return to the resonator under the action of the electric field created by the source U_0. Obviously generation in the klystron will be sustained when the electron bundles returning to the resonator deliver their energy to the field by interacting with the high-frequency field. This energy will be sufficient to cover the losses. Consequently, by changing the time of flight of the electron bundle one may either bring the klystron into a generating mode or cut off generation. In normal circuits, the change of the electron time-of-flight is achieved by changing the voltage on the repeller electrode. Those regions of the voltage U_0 at which generation is possible usually are called the generation zones of the klystron. Let us investigate the main parameters and characteristics of reflex klystrons.

Among the main parameters characterizing the operational properties of a klystron when it is used as the heterodyne of a radar receiver, we have:

- the range of mechanical frequency retuning, $f_{p\ min} - f_{p\ max}$;
- the range of electronic frequency retuning, Δf_0;
- the output power, P;
- the slope of electronic tuning, s.

The range of mechanical frequency retuning of the oscillations generated by the klystron is bounded by the maximum and minimum frequencies corresponding to the extreme value of the power in the generation operating zone.

The range of electronic tuning is determined as the frequency difference between the points of the generation zone at which the power drops to half of the maximum value.

By the rated output power of a klystron we understand the power delivered to a matched high-frequency load when the values of the resonator voltage and filament voltage are in accordance with specification certificate and there is an optimum voltage on the repeller.

The slope of electronic tuning characterizes the dependence of the klystron frequency on the voltage on the repeller. Numerically it equals the ratio of the frequency change Δf_G to the corresponding increment of the voltage on the repeller ΔU_0. This dependence is nonlinear. For the various types of klystrons, the slope of electronic tuning varies from tenths of a megacycle to several megacycles per volt.

The main data for several types of reflex klystrons are listed in Table 8.1.

The operational properties of a reflex klystron may be most fully evaluated by means of its operating, load and noise characteristics.

The operating characteristic of a reflex klystron (Figure 8.36) reflects the dependence of the generated oscillation output power and frequency on the change of the voltage on the repeller electrode with fixed resonator tuning and constant values of the voltage on the other electrodes, while the klystron is operating with a matched load.

TABLE 8.1. MAIN PARAMETERS OF SEVERAL REFLEX KLYSTRONS

Type of Klystron	Range of Mechanical Retuning, Mc	Range of Electronic Tuning, Mc	Output Power, mW	Slope of Electronic Tuning, Mc/V	Resonator Voltage, V	Repeller Voltage, V
2K28	1200 - 3750	± 10	90	0.7	300	-300
2K29	3400 - 3960	± 20	110	1.3	300	-250
2K48	2000 - 4000	± 10	110	-	1000	-250
2K25	8500 - 9660	± 30	25	4	300	-150

From an investigation of the operating characteristic, it follows that for voltage values on the repeller of U_{01}, U_{02}, U_{03} and U_{04} at which the condition of optimum interaction of the electron bundle with the resonator high-frequency field is fulfilled, the power of the generated oscillations is a maximum and the frequency is the same and equals the resonator natural frequency. By changing the resonator cavity by means of mechanical retuning, one is able to change the frequency of the oscillations being generated by the klystron within sufficiently broad limits. This change of the frequencies is in the limits ± (5 - 30%) of the mean value.

With a deviation of the voltage on the repeller from the indicated values, the power of the oscillations drops and the frequency changes within small limits (± 0.3 - 0.5% of the mean value). With an increase of the negative voltage on the repeller in respect to the values U_{01}, U_{02}, U_{03} and U_{04}, the frequency of the generated oscillations increases. With a decrease of this voltage, the frequency decreases. This is explained by the fact that, in the first case, the electron bundles return to the resonator before, and in the second case after, the resonator retarding field reaches the maximum value. Consequently, with mechanical frequency retuning of the klystron in the direction of increasing or decreasing frequencies, it is

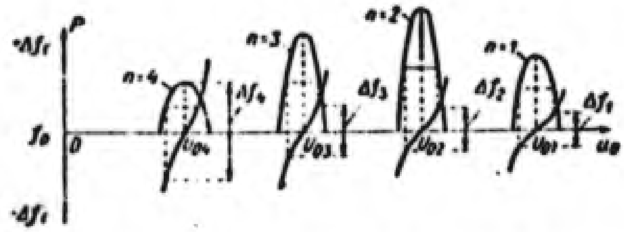


Figure 8.36. The operating characteristic of a reflex klystron with fixed dimensions of the resonator and voltage on the resonator.

necessary to increase or decrease the voltage on the repeller (Figure 8.37).

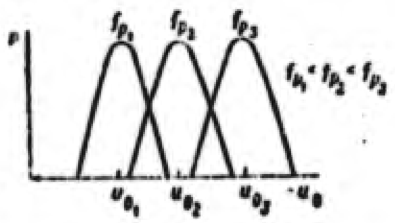


Figure 8.37. The relative position of klystron generation zones for various values of the dimensions of the resonator (f_{p_1}, f_{p_2}, f_{p_3}).

The operating characteristic of a klystron enables one to determine which of the generation zones should be selected as the operating zone. The generation zones of a klystron are not equivalent. From the operating characteristics it is seen that, with an increase of the zone number n, the range and slope of electronic frequency tuning increase, but there is a decrease of the maximum of the generated power. In order to increase the operating accuracy of the AFC circuit, preference should be given to zones where the range and slope of electronic tuning is larger. However, decreasing the reflected signal losses in the heterodyne circuits and weakening the influence of the mixer on the klystron output parameters forces one to return to the zone with the optimum power value.

The maximum power in the generation zone for a klystron without a load has the greatest value in the zone with the number n = 1, and it decreases with an increase of n, since in the first zone the transmission of energy by the electron bundles to the resonator field takes place during each period of the high-frequency oscillations, whereas in a zone with the number n - m, it takes place only once during m periods. However, when the klystron operates with a load, the voltage between the resonator walls decreases, since the resonator quality factor decreases. This causes a decrease of the percent of velocity modulation of the electron flux, as a result of which the time-of-flight of the electrons corresponding to the first generation zone turns out to be insufficient for the purpose of providing good quality clustering of them and the power in this zone decreases.

As follows from the operating characteristic, when operating on a matched load, the largest power occurs under this condition not in the first zone, but in the zones n = 2 or n = 3. The slope of electronic tuning also is acceptable in these same zones. One of the indicated zones, more precisely the one which corresponds to the optimum power value, should be selected as the operating zone when tuning is performed.

From what has been said above, it follows that the output parameters of a klystron depend to a considerable degree on the magnitude and character of the high-frequency load. The effect of the load on the value of the power and frequency of the oscillations being generated by the klystron is illustrated by its load characteristic (Figure 8.38), which is a family of curves of constant values of the output power (continuous lines) and frequency (dashed lines) plotted in the plane of complex impedance.

Since the standing-wave ratio (SWR) and phase of the reflected wave at each point of the line uniquely determines the ratio of the complex impedance at this point to the line wave impedance, the standing-wave ratio and phase may be used as the input data for the load characteristic. Therefore circles of SWR values are frequently plotted on the circular diagram of complex impedances (in Figure 8.38

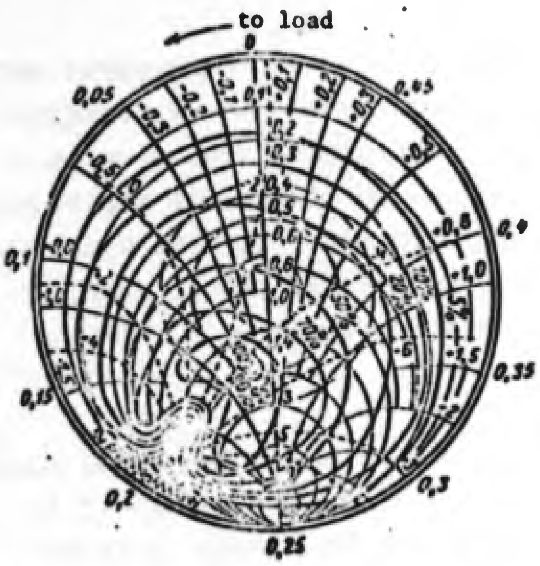


Figure 8.38. The load characteristic of a reflex klystron.

the dash-dot lines). These SWR circles are arranged around a common center of the diagram where the value is SWR = 1.

It follows from the load characteristics that the region of the most stable klystron operation corresponds to SWR values of the load of no more than 1.4. Larger reflections of high-frequency power from the load with an unfavorable phase of the reflection lead to unstable operation and even to stopping the oscillations (the crosshatched section of the complex impedance plane). In order to weaken the power of the reflected wave, a decoupling attenuator is frequently inserted between the load and the klystron.

Since, when using a reflex klystron as a heterodyne, very rigid requirements are imposed on the power stability and especially on the frequency stability of the generated oscillations, the problem of matching the klystron with the load when tuning a radar receiver is of very great importance. In radars, diode mixers which unfortunately have a considerable scatter of the parameters serve as the load for the klystron. Therefore, it turns out that for one pair of diodes, the mixer SWR satisfies the requirements for stable operation of the klystron and for other pairs it does not.

Evaluation of a mixer SWR may be done rather simply on the basis of the shape of the klystron operating generation zone. This may be observed on an oscillograph screen. One possible monitoring circuit is shown in Figure 8.39. An AC voltage is fed through isolation capacitors C_1 and C_2 with a capacitance of 0.1 - 0.2 microfarad and resistor R_2 = 200 - 300 kohm to the klystron repeller. By continuously changing the position of the autotransformer slide, a value of the AC voltage on the repeller is selected for which 100% amplitude modulation of the voltage on the repeller occurs. In this case, the signal fed to the oscillograph input from the load of the mixer diode (R_1 = 200 - 300 ohm) will reproduce the envelope of the klystron high-frequency oscillations.

The shape of the klystron generation zone and consequently also the quality of the padding of the klystron with the elements of the radar high-frequency section are considered satisfactory if there are no breaks or distortions in the image of the generation zone on the oscillograph at a level of more than 0.5 of the maximum value. Satisfactory shapes of the generation zone are shown in Figure 8.40,a. Figure 8.40,b shows generation zones which have unacceptable distortions. One is able to roughly evaluate the shape of the generation zone on the basis of the value of the mixer current. If, with a change of the voltage on the repeller, the readings of the instrument increase from zero up to some maximum value and then decrease but do not drop to zero and again continue to increase, this attests to unacceptable distortions in the given generation zone. One should keep in mind that distortions in the shape of the klystron generation zone, which are similar to those which are caused by a mismatched load, may be caused by insufficient cathode emission.

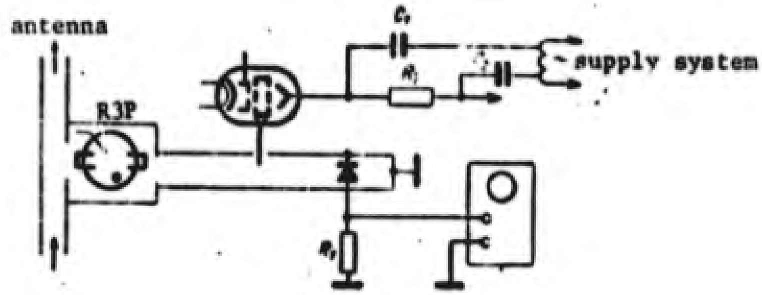


Figure 8.39. The circuit for observing the shape of a klystron generation zone.

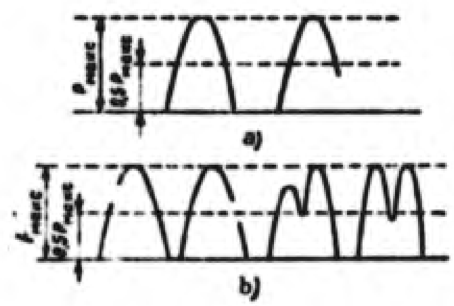


Figure 8.40. Shapes of a klystron generation zones.

Let us proceed to an investigation of the noise characteristics of a klystron. In receiver devices of centimeter-wave range radars, the noise at the converter output is determined by the noise of the reflex klystron. The noise characteristic of the klystron reflects the dependence of the klystron noise factor on the voltage of the repeller electrode.

Within the limits of the generation zone, the absolute level of the klystron noise varies insignificantly, increasing slightly in the range of higher frequencies of the electronic tuning range [46]. A significant increase of the noise level is observed with unstable operation of the klystron and in nonoperational generation regions. It is not the absolute noise factor of the klystron which is of practical interest, but the effective noise factor. The effective noise factor is determined both by the internal noise and by the value of the coupling between the klystron and mixer.

Above it was shown that for optimum signal conversion, the power fed from the klystron to the mixer must be a completely determined value. The indicated power level may be provided with a smaller coupling of the klystron with the mixer and a larger power of the oscillations being generated by the klystron. Therefore the noise characteristic which expresses the dependence of the klystron noise factor on the voltage of the repeller has the form shown in Figure 8.41.

From the investigation of the klystron noise characteristic, it follows that, in order to provide the minimum value of the effective noise factor, it is necessary to tune the klystron in such a way that the frequency of the oscillations being generated by it has the rated value when the voltage on the repeller corresponds to the maximum power of the generated oscillations.

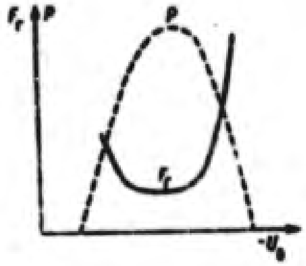


Figure 8.41. Noise characteristic of a reflex klystron.

8.5. Antenna Switch

The antenna switch couples three independent radar devices: the antenna, transmitter and receiver. The quality of its operation determines to a considerable extent the influence of these devices on the operation of one another. Therefore, in addition to the main purpose — the alternate connection of the transmitter and receiver to the antenna — the antenna switch must fulfill the following conditions:

- at the moment of radiation of the main pulse, it must not allow power capable of causing breakdown of the mixer diode to leak to the receiver input;
- it must ensure minimum signal losses both in the receiving mode and also in the transmitting mode;
- it must eliminate the effect of the load (the antenna or antenna equivalent) on transmitter operation.

For automatic switching of the antenna to the transmitter and receiver, the antenna switch must contain a nonlinear element. In existing types of radar antenna switches, the nonlinearity of the characteristic of such an element may be either a function of the power (switches using gas-discharge arresters in conjunction with resonance lines) or a function of the direction of the electromagnetic energy propagation (switches based on the use of ferrite properties).

The most important component of antenna switches of the first type is the resonance discharger (soft rhumbatron) which is a combination of a cavity resonator and a gas-filled discharger.

Antenna switches which are constructed for using the properties of gas-discharge arresters and resonance lines usually contain two dischargers: the transmitter discharger and the receiver protection discharger (ATR tube). The transmitter discharger prevents the leakage of the reflected signal energy to the superhigh frequency oscillator circuit. The receiver discharger protects the receiver from the action of the high-power main pulses of the transmitter.

Two types of ATR tubes exist: a narrow band type which allows retuning in a relatively small frequency range (10-15%) and a broad-band type with fixed tuning. Based on their structural execution, dischargers can be distinguished as those having an external resonator or an internal resonator. In the first case, the cavity resonator is a structural part of the antenna switch. The electrodes of such a discharger are placed in a glass cylinder and have disk outlets for connecting with the cavity resonator. The range in which dischargers with an external resonator are used is limited. In practice, dischargers of this type have not found application at frequencies above 6000 Mc. This is explained by the considerable reduction of the resonator quality factor at the indicated frequencies as a consequence of the increase of losses in the glass of the cylinder.

During the operation of the magnetron, the discharger has a pulsed high-frequency voltage applied to it. Before the onset of the discharge, the discharger resistance is very large. The voltage on its electrodes increases up to the value at which firing of the discharger occurs U_{fir}. Then the resistance of the discharger decreases in proportion to the growth of the gas discharge, and at the beginning of the steady discharge the voltage on its electrodes is reduced to the burning voltage U_{burn}. The envelope of the high-frequency voltage on the discharger electrodes during transmission of the main pulse has the form shown in Figure 8.42. As is seen from the figure, the greatest danger to the mixer diode consists of leakage of the peak energy to the receiver input during the time interval between the instant when the magnetron begins generation and the instant when the steady discharge begins. In order to avoid deterioration of the diode parameters or their complete destruction, the peak energy must

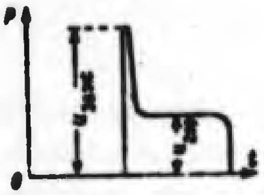


Figure 8.42. The envelope of high-frequency voltage on the electrodes of a discharger.

be less than 0.1 ergs and the leakage power must not exceed 100 mW.

In dischargers with a high quality factor, the leakage power remains constant with a change of the transmitter power of more than a factor of 10^3 (Figure 8.43) [3]. The magnitude of the peak power depends on the properties of the gas filling the discharger, the initial concentration of electrons in the discharger gap and the width of the spark gap. Consequently, a decrease of the peak energy may be achieved by increasing the concentration of electrons in the discharger gap.

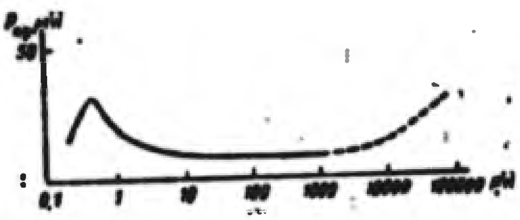


Figure 8.43. The power leaking through the resonance discharger for the receiver protection (ATR tube) as a function of the transmitter power.

A glow discharge caused by secondary ignition is the source of such electrons. The igniter electrode for this is the cathode. In order to reduce the voltage of the power supply source for the ignition circuit, the ignition electrode is activated. The glow discharge must, on the one hand, ensure reduction of the peak energy to a safe value and, on the other hand, not cause a noticeable reduction in the power of the reflected signals. Moreover, one should consider that with an increase of the ignition current, the discharger lifetime is shortened. In a majority of types of dischargers, an acceptable value of the peak energy may be achieved when the secondary emission current is about 30 microamp. However, with this value of the current, the secondary discharge is not sufficiently stable and there exists the danger of relaxation oscillations arising in the ignition circuit. The oscillations arising in this case are similar to the oscillations of a relaxation oscillator with a neon lamp. This is seen in Figure 8.44, in which the equivalent circuit of the secondary electrode circuit is shown.

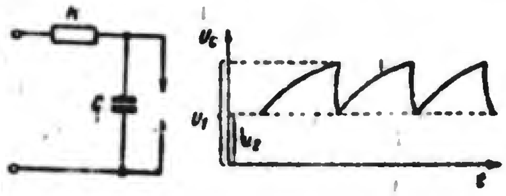


Figure 8.44. The equivalent circuit of the secondary ignition circuit of a resonance discharger.

In this circuit, R is the limiting resistance and C is the capacitance of the electrode and its power supply conductor. If the value of the secondary ignition current is insufficient for maintaining a continuous discharge, then the circuit will operate in the following way. The voltage on the capacitor increases at a rate determined by the time constant CR to the value U_1 at which breakdown of the spark gap occurs. After this, the capacitor rapidly discharges through the low resistance of the spark gap until the voltage reaches the quenching potential U_2.

Relaxation oscillations with reference to the i-f amplifier are a source of impact excitation, as a consequence of which they may cause self-excitation of the receiver. Besides this, with a small ignition current, the resistance of the gas discharge for a DC current is commensurate with the leakage resistance along the discharger cylinder. This may be the reason for the destruction of the discharge. In practice the discharger ignition current is selected as approximately equal to 100 microamp. With this value of the ignition current, the power losses of the reflected signal do not exceed 1%.

After the transmitter pulse ends, the gas in the spark gaps must be deionized. However, deionization may not be performed instantaneously. The time during which attenuation of the input signal introduced by the resonance discharger is reduced to three dB is called restoration time. An increase of the restoration time leads to an increase of the minimum effective range of the radar.

The quality of the discharger deteriorates when a discharger has been stored for a long time and during operation. In operating radars, deterioration of the discharger parameters may be detected on the basis of a decrease of the discharger current, an increase of the radar minimum effective range and by partial breakdown of the mixer diodes as a consequence of the increase of the leakage power.

In the majority of cases, deterioration of the discharger quality is connected with a change of the composition of the gas filling the discharger cylinder. The rate at which the gas composition changes depends mainly on ignition current and increases with it.

During operation, monitoring of the resonance discharger electric conditions is reduced to determining the discharger ignition current and to checking that there are no relaxation oscillations in the secondary ignition circuit. In order to monitor the ignition current, a built-in microammeter is usually provided in the radar. The presence of relaxation oscillations in the ignition circuit may be detected by means of an oscillograph. The arrangement for connecting the oscillograph is shown in Figure 8.45.

The use of broad-band superhigh frequency oscillators in radar transmitting devices and the subsequent increase in the power of transmitting devices leads to an increase of the overall size and complexity of the construction of antenna switches [56]. This required new solutions. One of the variations of such solutions was the creation of antenna switches using ferrites.

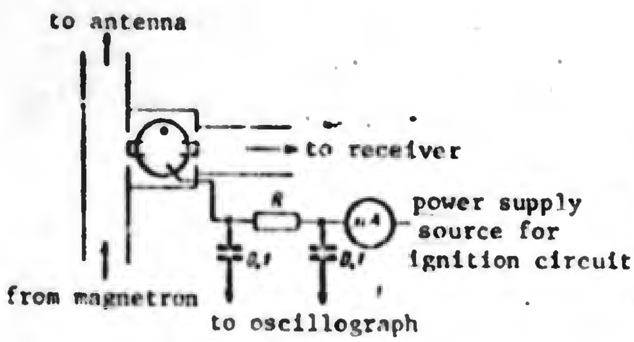


Figure 8.45. The arrangement for monitoring the stability of the secondary ignition circuit toward relaxation oscillations.

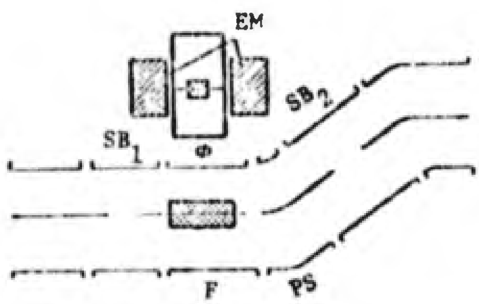


Figure 8.46. The arrangement of an antenna switch constructed to use the uncorrelated phase shifts in transversely magnetized ferrites.

Figure 8.46 shows the arrangement of a switch which uses uncorrelated phase shifts in transversely magnetized ferrites. The switch consists of two slot bridges (SB_1 and SB_2), between which there are a ferrite section F and a phase-shifting section PS.

The ferrite section consists of two waveguides having a common narrow wall. A ferrite plate is fastened in each wave guide symmetrically with respect to the common walls. The plates are magnetized by the transverse magnetic field created by an electromagnet EM.

When an electromagnetic wave is propagated through the ferrite section, the wave acquires a certain phase shift as a result of interaction of its magnetic field with the intrinsic magnetic moments of the electrons in the ferrite atoms. The amount of the phase shift depends on the strength of the constant magnetic field, the type of ferrite and its dimensions. The sign of the phase is determined by the direction of the electromagnetic energy propagation, by the polarity of the magnet and by the position of the plate with respect to the waveguide axis.

By adjusting the current in the winding of the electromagnet, that value of the magnetic field intensity is selected at which each

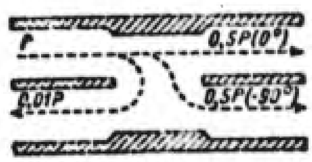


Figure 8.47. A slot bridge.

of the ferrites creates a phase shift equal to 45°. Since the plates are arranged symmetrically with respect to the common wall of the wave guides, the phase shifts created by them will have opposite signs. Consequently, the waves will have a resultant phase shift equal to 90° at the output of the ferrite section.

The slot bridge (Figure 8.47) has the following properties:

- it divides the high-frequency power entering one of the branches into equal parts between the opposite branches;
- no more than 0.01 of the entering power is shunted to the adjacent branch;
- with its passage through the slot, the phase of the wave changes by 90° in the direction of a lag.

The phase-shifting section consists of two wave guides of different lengths which are bent along their narrow wall. The difference in lengths of the wave guides is selected in such a way that the phase advance in the longer wave guide is 90° larger than the phase advance in the shorter wave guide.

The operation of a switch in the <u>transmission</u> mode for the case when the transmitter load is the antenna, and in the <u>reception</u> mode is illustrated in Figures 8.48 and 8.49. Obviously, in a switch of the given type, in order to switch the transmitter output from the antenna to the dummy antenna, it is sufficient to change the polarity of the electromagnet. Figure 8.50 illustrates the operation of the switch in the <u>transmission</u> mode for the case when the transmitter load is the dummy antenna. In addition to the main function, an antenna switch of the given type fulfills the role of a

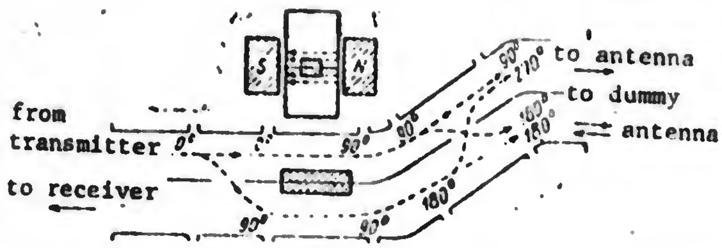


Figure 8.48. The arrangement illustrating the operation of the antenna switch in the transmission mode (the transmitter load is the antenna).

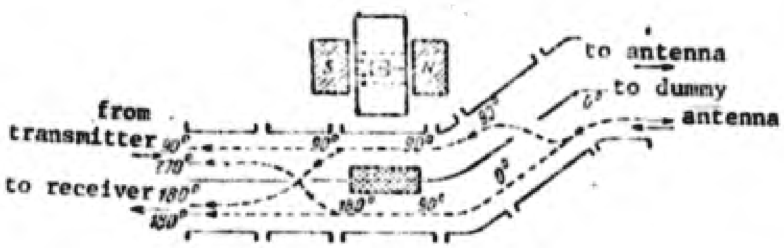


Figure 8.49. Arrangement illustrating the operation of the antenna switch in the reception mode.

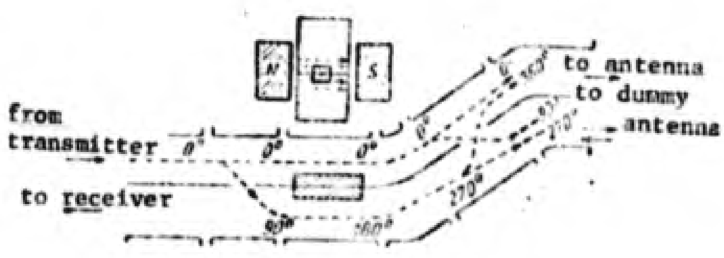


Figure 8.50. Arrangement illustrating the operation of the antenna switch in the transmission mode (the transmitter load is the dummy antenna).

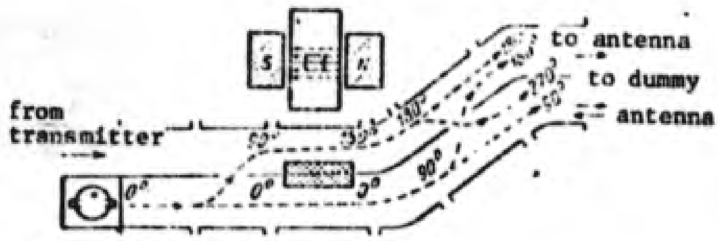


Figure 8.51. Arrangement illustrating the protection of the magnetron against the effect of a mismatched load by means of a ferrite rectifier.

rectifier, thereby eliminating the effect of a mismatched load on magnetron operation. It is not difficult to establish this by tracing the path of the reflected wave.

The energy of the reflected wave from the antenna is fed to the receiver input in the same way as the signal from the target. Since during magnetron generation, the discharger short circuits the receiver input, the energy of the reflected wave does not reach the receiver. The energy is reflected (Figure 8.51) and sent to the branch of the switch to which the dummy antenna is connected. The energy of the reflected wave is absorbed by the dummy antenna and in this way its multiple oscillations between the oscillator and an unmatched load are eliminated.

During operation, the parameters of a ferrite switch remain practically unchanged. The reason for breakdown of a switch may be damage of the ferrite plates. Electrical breakdown occurs in the wave guide at the place where the damaged plate is located when the ferrite plates are damaged. This electrical breakdown may be easily detected by its characteristic sound.

8.6. Methods of Tuning a Radar to Maximum Effective Range

In operational practice, the following methods of operational tuning of a radar to maximum effective range are used:

- on the basis of an isolated distant object;
- by means of a monitoring resonator (echo chamber);
- by means of a combined test instrument.

Tuning a radar to maximum effective range on the basis of an isolated object consists of the following procedure. Some characteristic radar object is selected in the area where the radar is located. One may use a corner reflector as the object. The station antenna is directed toward the selected object. Then maximum brightness (amplitude) of the marker signal from the isolated object is achieved on the indicator screen by tuning the elements of the station high-frequency section.

One should keep in mind that the effective range obtained by using this tuning method is not always the maximum effective range of the radar. This is explained in as follows. With this method, the magnitude of the signal from the distant object is the criterion for optimum tuning. Since the signal value is characterized not only by the station power potential but also by the radar observation conditions, this method does not allow one to detect a reduction of the station power potential due to deterioration in the operation of the elements in the high-frequency section. Thus, for example, when tuning a radar having a mixer with deteriorated parameters, the maximum range will be achieved, but it will be less than that range which might be obtained if the diode were replaced.

Among the defects of the tuning method based on a distant object one must also list the radiation of energy into space. Moreover, tuning a station on a distant object is not always possible. For example, it is impossible to use this method for tuning a radar mounted on a ship, when the ship is on the open seas. An advantage of this method is that when the radar is tuned, it is operating under actual conditions.

Tuning a radar by means of a monitoring resonator has become very common. The monitoring resonator is tuned to the station transmitter frequency beforehand, and then the maximum duration of illumination of the monitoring resonator signal is achieved on the screen of the radar indicator by tuning the elements of the high-frequency section.

An advantage of this method is that tuning is accomplished without radiation of the transmitter energy into space, since in the majority of radars, the monitoring resonator may be used when the radar operates on the dummy antenna. Under this condition, the dummy antenna must present to the magnetron precisely the same load as the antenna. If this is not the case, the frequency and power of the oscillations being generated by the magnetron change when switching from the equivalent load to the actual load. This leads to a reduction of the radar effective range. This is a defect of this method.

The method of tuning by means of a test instrument is based on the replacement of the actual reflected signal by an imitation signal which is shaped by the test instrument. The tuning process is similar to tuning on a distant object. The advantage of the method is that, after tuning, one is able to evaluate its effectiveness by measuring the radar main technical parameters which determine its effective range. However, this method is not free of defects, the main one of which is that the imitated signal is not absolutely the same as the real signal.

8.7. Tuning the High-Frequency Section of a Radar

Tuning the high-frequency section is one of the most important and most frequently performed operational tunings of a radar.

The importance of tuning is explained by the fact that the quality with which it is performed substantially affects the station effective range.

The necessity for frequently performing the indicated tuning is due to the very strong dependence of the parameters of the superhigh frequency instruments on the electric conditions and the environmental conditions. The concerted operation of the superhigh frequency instruments combined in the high-frequency section is characterized by the complicated influence of one instrument on the other. Therefore, a deviation of the output parameters of any of the instruments from the rated values causes a disturbance of the normal operation of the other instruments, and in certain cases also of the high-frequency section as a whole. Thus, for example, deterioration of the parameters of the ATR tube leads to deterioration of the converter properties of the diode mixer or to burnout. The problem of tuning the high-frequency section of a station consists of providing those conditions for the concerted operation of the instruments and elements included in the high-frequency section at which the maximum possible value of the power potential and consequently also of the radar effective range occurs.

Tuning a high-frequency section consists of several operations. The sequence of performing these operations, as was shown in Chapter 7, is important not only from the point of view of expenditure of time, but also from consideration of the quality of the tuning.

The technique discussed below for tuning a high-frequency section provides the most rational sequence of operations when tuning.

Tuning a Magnetron

If a tunable magnetron is used in the radar, then after it is mounted, it must be tuned to the station operating frequency. This tuning may be performed most simply by means of a monitoring resonator (echo chamber). The essence of the tuning consists of the following. A reading corresponding to the station operating frequency is set on the scale of the monitoring resonator. The radar is switched on, and those operating conditions of the magnetron are established at which the value of its average current equals the rated value. After this, by changing the setting of the magnetron frequency retuning unit, a maximum value of the monitoring resonator detector current is achieved. Obviously, the maximum value of the detector current corresponds to the case when the natural frequency of the monitoring resonator equals the frequency of the oscillations being generated by the magnetron. When a magnetron operating at a fixed frequency is used in the radar, it does not require tuning. However, since the stability of the magnetron operation turns out to have a considerable influence on the operation of the other elements of the high-frequency section of the station, one should check magnetron operation before beginning to tune the station. The shape of the high-frequency pulse spectrum contains considerable information on the quality of magnetron operation. As was shown in Section 8.1, the shape of the high-frequency pulse spectrum enables one to judge as to the constancy of the frequency during the duration of the pulse and also as to the frequency distribution of energy in the pulse. Monitoring the spectrum shape may be accomplished by means of a spectrum analyzer or by a monitoring resonator using the technique discussed in Chapter 2.

If the monitoring results show that the spectrum shape has distortions and a noticeable broadening of the frequency band in which the main part of the pulse energy is concentrated, this attests either to a disturbance of the matching of the magnetron with the load or to unacceptable changes of the amplitude of the modulating voltage pulse during magnetron operation. In this case, one should first of all measure the TWR of the antenna wave guide channel. If this parameter is found to be at the norm, one should monitor the shape of the modulating voltage pulse. When investigating the magnetron load characteristics (see Figure 8.6), it was shown that with the same value of the TWR there are regions of more stable magnetron operation and less stable magnetron operation depending on the phase of the reflection coefficient. Therefore, in certain radars, where there are no ferrite rectifiers, a phase-shifting section is mounted in the wave guide channel.

The phase-shifting section, which changes the wavelength in the wave guide, enables one to establish that phase of the reflected wave at which the operating point on the load characteristic is located in a region of more stable magnetron operation. In the case of unsatisfactory shape of the spectrum and when there is a phase shifter in the radar being tuned, one should adjust the phase of the reflected wave.

Preliminary Tuning of a TWT

The tuning of a TWT is a very important operation in the tuning of a radar high-frequency section. It consists of providing those operating conditions of the tube at which the tube has the highest amplification factor with a minimum noise factor. The preliminary tuning is limited to setting the power supply voltages on the tube electrodes in accordance with the values indicated in the ceritificate. Since different individual tubes of the same type have electrical conditions corresponding to most effective operation and which differ from one another, one should make certain that the number indicated on the certificate corresponds to the number marked on the tube before beginning to tune the TWT. The setting of the power supply voltages on the tube electrodes must be performed in a strict sequence. Any departure from this sequence leads to a considerable shortening of the tube lifetime, and in certain cases it may be the reason for the immediate breakdown of a tube. The technique for tuning a TWT recommended below shows this sequence.

Before switching on the TWT power supply source, one should set the adjustment units of the power supply voltages in a position which ensures that the minimum voltages are fed to the electrodes.

The switching on of the TWT begins with feeding the power supply to the tube forced cooling circuit. A disturbance of the thermal conditions has an especially unfavorable effect on the quality of the beam focusing, since heating of the solenoid winding leads to an increase of its resistance. Then the solenoid power supply is switched on, and the rated value of the winding current is established. One should note that, in the case when there is no solenoid current, further switching on of the tube is intolerable, since under this condition the entire electron flux will strike the spiral which may lead to immediate breakdown of the tube. Therefore, automatic removal of the power supply from the tube is usually provided in the radar when there is a break in the solenoid power supply circuit.

After this, the TWT power supply unit is switched on, and the rated value of the filament voltage is established. During the interval of time required for heating the cathode (for the majority of low-noise tubes, this time is 2 - 3 minutes) one must begin to adjust the voltages on the other electrodes of the tube. In order to prevent an unacceptable current increase in the decelerating system (spiral) which may lead to a breakdown of the tube, it is recommended to initially establish, first on the anode and then on the spiral, voltage values equal to approximately 75% of the values indicated on the certificate on the control electrode.

With these conditions, by monitoring the collector current and spiral current, the tube is adjusted by means of the centering device to a position for which the collector current has a maximum value and the spiral current a minimum value.

Then the voltages on the tube electrodes are increased to the rated value and once again the tube is centered, but this time more precisely. After this, the position of the centering units is fixed. One must judge as to the tuning results on the basis of the spiral current and collector current. Correspondence of the spiral current and collector current to the certificate value attests to the fact that the tube conditions have been correctly established.

Tuning a Heterodyne

Optimum amplification of intermediate-frequency signals will occur when the heterodyne frequency f_G differs from the superhigh frequency oscillator frequency f_C by precisely the amount of the intermediate frequency f_{if}:

$$f_r = f_c \pm f_{up}.$$

Since a radar receiver passband amounts to several megacycles, even an insignificant deviation of the frequency difference $f_p = f_G - f_C$ from the intermediate frequency f_{if} caused by detuning of the heterodyne leads to a sharp decrease of the receiver sensitivity. In that case, when the difference frequency signal is found to be beyond the limits of the i-f amplifier's passband (Figure 8.52), reception of signals is impossible. The tuning of the heterodyne as a consequence of this reduces to fulfilling the condition at which $f_p = f_{if}$.

When tuning the heterodyne of a centimeter range radar, the following technique is best. Before starting the tuning, all the receiver automatic adjustments (AFC, CAVC, AVC, etc.) are switched off.

The potentiometer axis of the manual frequency adjustment (MFA) (Figure 8.53) used to change the voltage on the repeller within the limits of the width of the klystron generation zone is set in the beyond position. This enables an excursion of the klystron frequency within the limits of the electronic frequency retuning range to be compensated during operation irregardless of any sign of detuning. By changing the voltage on the repeller by means of the <u>receiver tuning</u> potentiometer, the klystron operational generation zone is selected. Based on the considerations discussed in Section 8.4, this must be the zone which corresponds to maximum power generated by the klystron. The center of the generation zone is determined by the klystron. The center of the generation zone is determined on the basis of the receiver maximum mixer current. However, if the generation zone has a distorted shape, the mixer current maximum does not correspond to the center of the generation zone. Operation in such a zone is unacceptable. One may evaluate the approximate shape of the generation zone on the basis of readings from the instrument monitoring the mixer current as was already indicated in this chapter.

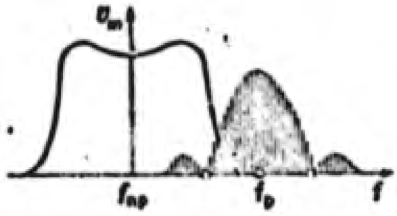


Figure 8.52. Displacement of the frequency spectrum of the intermediate-frequency radio pulse in respect to the frequency characteristic of the i-f amplifier.

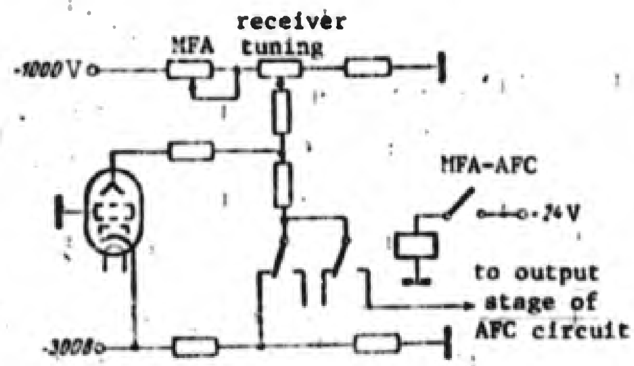


Figure 8.53. A klystron power supply circuit.

A more precise evaluation of the shape of the generation zone may be made by scanning it on an oscillograph screen according to the technique discussed in Section 8.4.

When there is a satisfactory shape of the generation zone, one may begin to tune the klystron to the frequency at which the condition $f_p = f_{if}$ is fulfilled. This tuning is performed by changing the natural frequency of the klystron resonator (either by means of the plungers in klystrons with an external resonator or by means of the tuning screw in klystrons with an internal resonator). One should bear in mind that a fully screwed-down tuning screw (plunger) corresponds to the maximum frequency of the klystron operating range, and a fully screwed-out tuning screw corresponds to the minimum frequency. Depending on which tuning method is used, the condition $f_p = f_{if}$ is fixed either on the basis of the maximum amplitude (brightness) of the signal from an isolated target (test instrument) or on the basis of the maximum duration of the monitoring resonator signal. When tuning a klystron, it is necessary to take into account the fact that the indicated equality is fulfilled for two frequencies of the heterodyne, i.e., when:

$$f_{г} = f_{c} + f_{пр}$$

and when

$$f_{г} = f_{c} - f_{пр}$$

In both cases, reception of useful signals is possible. However, the selection of the heterodyne frequency influences the correct operation of the AFC circuit.

When selecting the heterodyne frequency, it is necessary to consider the following [27]. If the frequency characteristic of the discriminator of the AFC circuit has the form shown in Figure 8.54, a then the correct tuning corresponds to the case $f_G > f_C$. With the other characteristic of the discriminator (Figure 8.54, b), the correct tuning corresponds to the case $f_G < f_C$. When these conditions are not met, there is a false response of the AFC circuit. In this case, the heterodynes will be noticeably detuned, and the reception of signals is practically impossible.

Establishing the Optimum Coupling Between the Mixer and the Heterodyne

The optimum coupling should correspond to that level of the power being fed from the heterodyne to the mixer at which the mixer noise is a minimum. For the majority of types of diodes, the optimum power level corresponds to a value of the detector current constant component of 0.4 - 0.5 mA.

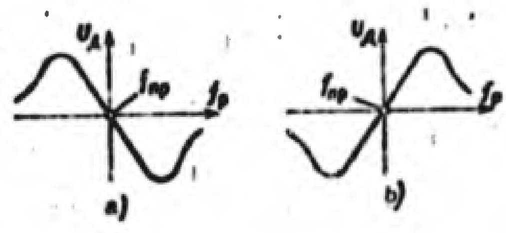


Figure 8.54. Frequency characteristics of the discriminator of the AFC circuit.
a - left characteristic; b - right characteristic.

Final Tuning of a TWT

The final tuning of a TWT reduces to setting the optimum voltage on the second anode (spiral) and matching the input and output of the tube decelerating system with the radar wave guide channel. One can judge as to the correspondence of the voltage value on the second anode and the optimum value either on the basis of the maximum amplitude (brightness) of the signal from an isolated target (test instrument), or on the basis of maximum duration of a monitoring resonator signal. In Section 8.2 it was shown that for low-noise tubes, the optimum voltage on the second anode corresponding to the minimum noise factor coincides with the voltage value at which the tube has a maximum amplification factor. Therefore, the maximum current value of the reciever second detector also may serve as the criterion for setting the optimum voltage on the second anode. Matching of the decelerating system input and output with the wave guide channel is achieved by means of fine tuning elements (see Figure 8.18) and by shifting the tube along the re-enforcing framework axis. The latter operation is explained by the fact that the position of the coupling antennas attached to the ends of the spiral with respect to the windows of the input and output wave guides determines the degree of coupling between the wave guide and the spiral. Initially, the matching is done by displacing the plungers of the fine tuning elements. Then by shifting the tube along the re-enforcing framework axis, its optimum position is found with respect to the windows of the input and output wave guides. Optimum tuning corresponds to the maximum amplitude (brightness) of the signal from an isolated target (monitoring instrument). One should keep in mind that the indicated tuning is a rather fine adjustment and requires definite skills.

Establishing the Optimum Coupling Between the Mixer of the AFC and the Heterodyne

The heterodyne power level and the power level of the main pulses at the mixer input must have a definite relationship. If this is not the case, currents of combination frequencies arise at the mixer output. This is a cause of unstable operation of the AFC circuit. If the ratio of the indicated powers equals 3, then the current of the strongest combination frequency will be attenuated by a factor of 10 in comparison with the current of the useful signal. This ensures stable operation of the AFC circuit.

Consequently, the power of the main pulses delivered to the AFC mixer must exceed by approximately a factor of three the power supplied from the heterodyne.

The amount of the power delivered from the heterodyne to the mixer is monitored by means of a "crystal current" instrument which measures the constant component of the mixer current.

Monitoring the amount of the main pulse power sent to the mixer cannot be done in a similar way, since the constant component of the current of these pulses is very small.

In practice, the indicated tuning is performed in the following way. By means of an attenuator included in the mixer branch through which the main pulse power is delivered, the coupling between the AFC mixer and magnetron is reduced until stable operation of the AFC circuit no longer occurs. Unstable operation is detected on the basis of fluctuations in the pointer of the "crystal current" instrument. The position of the attenuator corresponding to this state is recorded. Then the coupling is increased until the operation of the AFC circuit again becomes unstable. After this, the attenuator is set approximately in the middle position with respect to the indicated boundary values.

REFERENCES

1. Алпатов Н. И., Ферриты в электронных схемах. Воениздат, 1962.
2. Астафьев А. Ф. Окружающая среда и надежность радиотехнической аппаратуры. Изд-во «Энергия», 1965.
3. «Антенные переключатели». Пер. с англ., под ред. Р. Н. Перец. Изд-во «Советское радио», 1950.
4. Балашов В. П. и др. Автоматизация радиоизмерений. Изд-во «Советское радио», 1966.
5. Белов Ф. И., Соловейчик Ф. С. Вопросы надежности радиоэлектронной аппаратуры. Госэнергоиздат, 1961.
6. Берж К. Теория графов и ее применения. Изд-во иностранной литературы, 1962.
7. Берман Я. И., Гольдин Б. М. Настройка радиолокационной аппаратуры. Судпромгиз, 1957.
8. Бобнева М. И. Техническая психология. Изд-во «Наука», 1966.
9. Брюле Д. Д., Джонсон Р. А., Клетский Е. Д. Отыскание неисправностей в технических устройствах. «Зарубежная радиоэлектроника», 1961, № 7.
10. Бунин Г. Г., Васенькин В. А. Отражательные клистроны. Изд-во «Советское радио», 1966.
11. Важенина З. П. Фантастронные генераторы. Изд-во «Советское радио», 1965.
12. Валитов Р. А. Радиотехнические измерения. Изд-во «Советское радио», 1963.
13. Васильев Б. В. Надежность и эффективность радиоэлектронных устройств. Изд-во «Советское радио», 1964.
14. Вентцель Е. С. Теория вероятностей. Изд-во «Наука», 1964.
15. Верзаков Г. Ф., Киништ Н. В., Рабинович В. И., Тимонен Л. С. Введение в техническую диагностику. Изд-во «Энергия», 1968.
16. Григорьянц В. Г. Технические показатели радиолокационных станций. Воениздат, 1963.
17. Гусев В. П. Расчет электронных допусков радиоэлектронной аппаратуры. Изд-во «Советское радио», 1963.
18. «Два метода автоматического исследования импедансов». «Экспресс-информация», Радиотехническая промышленность, вып. 44, 1958.
19. Евтянов С. И. Переходные процессы в приемно-усилительных схемах. Связьиздат, 1948.
20. Евтянов С. И. Радиопередающие устройства. Связьиздат, 1950.

21. Жуков Б. С., Перегонов С. А. Лампы бегущей волны. Изд-во «Советское радио», 1967.

22. «Судовая радиолокационная станция «Донец-2». Изд-во «Морской транспорт», 1964.

23. «Измерения на сверхвысоких частотах». (Пер. с англ.) Изд-во «Советское радио», 1952.

24. Капланов М. Р., Левин В. А. Автоматическая подстройка частоты. Госэнергоиздат, 1956.

25. Карибский В. В., Пархоменко П. П., Согомонян Е. С. Техническая диагностика объектов контроля. Изд-во «Энергия», 1967.

26. Кофман А., Фор Р. Займемся исследованием операций. Изд-во «Мир», 1966.

27. Крохин В. В. Элементы радиоприемных устройств СВЧ. Изд-во «Советское радио», 1964.

28. Латинский С. М. Девиация судовых радиолокационных станций. Изд-во «Судостроение», 1966.

29. Лернер А. Начало кибернетики. Изд-во «Наука», 1967.

30. Луцкий В. А. Расчет надежности и эффективности радиоэлектронной аппаратуры. Изд-во АН УССР, 1963.

31. Мозгалевский А. В., Гаскаров Д. В., Глазунов Л. П., Ерастов В. Д. Автоматический поиск неисправностей. Изд-во «Машиностроение», 1967.

32. Месяцев П. П. Надежность производства электронно-вычислительных машин. Машгиз, 1963.

33. Нейман М. С. Курс радиопередающих устройств. Изд-во «Советское радио», 1965.

34. Орэ О. Графы и их применение. Изд-во «Мир», 1965.

35. «Общая теория систем». Изд-во «Мир», 1966.

36. «Основы использования магнетронов». Под ред. Ю. Н. Хлопова. Изд-во «Советское радио», 1967.

37. Павличенко А. Д. и др. Надежность радиоэлектронной аппаратуры. Изд-во «Советское радио», 1963.

38. Панин Н. С., Панченко Л. В. Импульсные магнетроны. Изд-во «Советское радио», 1966.

39. Раков В. И. Индикаторные устройства радиолокационных станций. Судпромгиз, 1962.

40. Райкин А. Л. Элементы теории надежности для проектирования технических систем. Изд-во «Советское радио», 1967.

41. У. Росс Эшби. Введение в техническую кибернетику. Изд-во «Иностранная литература», 1959.

42. «Радиолокационная техника», т. 1 и 2. (Пер. с англ.) Изд-во «Советское радио», 1949.

43. «Ремонтопригодность радиоэлектронной аппаратуры». (Пер. с англ.). Изд-во «Советское радио», 1952.

44. Сиверс А. П. Радиолокационные приемники. Изд-во «Советское радио», 1953.

45. Ситник Н. В. Автоматизация испытаний радиоэлектронного оборудования. Изд-во «Советское радио», 1959.

46. Сретенский В. Н. Основы применения электронных приборов сверхвысоких частот. Изд-во «Советское радио», 1963.

47. Кейнз С. Н. Поиск неисправностей в радиоэлектронных

системах методом функциональных проб. Изд-во «Советское радио», 1965.

48. Тимонен Л. С. О построении оптимальных программ диагностики состояния сложных технических систем. «Техническая кибернетика», 1966, № 4.

49. Тимонен Л. С. О построении оптимальных программ контроля работоспособности. «Автометрия», 1966, № 1.

50. Трошанов Н. А. Радиоаппаратура на лампах бегущей волны. Судпромгиз, 1961.

51. Уолфф. Метод самокалибровки радиолокационных станций. «Вопросы радиолокационной техники», 1955, № 2 (26).

52. Фано Р. Передача информации. Изд-во «Мир», 1965.

53. Хикс Г. Основные принципы планирования эксперимента. Изд-во «Мир», 1967.

54. Хаффмэн Д. А. Метод построения кодов с минимальной избыточностью. Кибернетический сборник. Изд-во «Иностранной литературы», 1961.

55. Харди Г. Г., Литтльвуд Д. Е., Полиа Г. Н. Неравенства. Изд-во «Иностранная литература», 1948.

56. Чедич. Современные разработки узлов сверхвысоких частот для радиолокационных станций. «Вопросы радиолокационной техники», 1957, № 6 (42)

57. Чудновский И. Я. Устранение неисправностей в усилителях. Изд-во «Искусство», 1965.

58. Шарапов В. И. Функционально-эквивалентные схемы для обучения специалистов по ремонту радиоэлектронной аппаратуры. «Вопросы психологии», 1966, № 1.

59. Шарапов В. И. Применение теории графов для решения задачи контроля работоспособности радиоэлектронных систем. Материалы республиканской конференции. Минск, 1967.

60. Мозгалевский А. В., Шарапов В. И. Техническая диагностика. ЛДНТП, Ленинград, 1968.

61. Широков А. М. Основы надежности и эксплуатации радиоэлектронной аппаратуры. Изд-во «Наука и техника», Минск, 1965.

62. Шишонок Н. А. и др. Основы теории надежности и эксплуатации радиоэлектронной техники. Изд-во «Советское радио», 1964.

63. Шеннон К. Работы по теории информации. Изд-во «Мир», 1966.

64. Glass B. An optimum policy for detecting a fault in a complex system. Operat. Res., 1959, 7, № 4.

65. Ramamoorthy C. V. Analysis of graphs by connectivity consideration, «J. Assoc. Comput. Mach.», 1966, 13, № 2.

66. Land A. H., Doig A. An Automatic Method of Solving Discrete Programming Problems. Econometrica, 1960, 28, p. 497—520.

67. Lawler E. L., Wood D. E. Branch—and—Bound methods: a survey. Operat. Res 1966, № 4 (Русский перевод: «Экспресс-информация». Техническая кибернетика, 1966, № 47.)

68. ГОСТ 13377-67. Надежность в технике. Термины. Изд-во стандартов, 1968.

69. Ксенз С. П., Капунников Ю. Ф., Малаксиянов М. Н., Никольский В. Н. Устранение неисправностей судовых РЛС. Изд-во «Морской транспорт», 1962.

SYMBOL LIST

Russian	Typed	Meaning
с	s	signal
и	p	pulse power
пр	rec	receiving
ц	t	effective reflecting area
макс	max	maximum
з	p	power
дб	dB	decibel
сог	sog	power generated by equivalent noise source
п	f	forward wave
о	r	reflected wave
мин	min	minimum
в	w	wave guide
вх	in	input
ф	phase	phase
ср	av	average
пр	lim	limiting
э	e	error
вс	ax	auxiliary
и	i	instrument
н	r	rise time
с	d	decay time
п	i	incident
гш	ng	noise generator
атт	att	attenuator
вч	hf	high frequency
пч	if	intermediate frequency
с	m	mixer
и	power	power
зв	rin	ringing

Russian	Typed	Meaning
ПР	f	forward
ПМ	sm	submodulator
П	g	generator
ПР	rec	receiver
КГ	qo	quartz-crystal oscillator
ИЗМ	m	measured
ИСТ	t	true
Р	r	radar
Г	g	geodesic
ПВ	CA	not defined
КП	MC	not defined
КВ	MA	not defined
К	M	monitoring
ЗАД	del	delay
Б	b	sighting device
[illegible]	p	preference
ПК	pm	monitoring period
П	s	search
КС	ds	not defined
ПК ЭФФ	pm eff	efficient monitoring periodicity
ВН	su	sudden
П	ga	gradual
ОТК	fail	failure
ПР	for	forecast
ВЫХ	out	output
ПР	pro	production
Т	temp	temperature
СТ	age	aging
ВХ Р	in p	input pulse
П	t	trimmer
КУ	md	monitoring device
СД	sd	not defined
СН	sn	not defined
Ц	ad	advisability

Russian	Typed	Meaning
З	e	electric
И	p	pulse
З	e	electron
З	o	overshoot
ЗаЖ	fir	firing
ГОР	burn	burning
Г	G	heterodyne
ПР	if	intermediate frequency
Д	d	discriminator
П	p	not defined
И	i	not defined
О	b	breaks
З	s	short
ПО	g	gradual
ВО	s	sudden