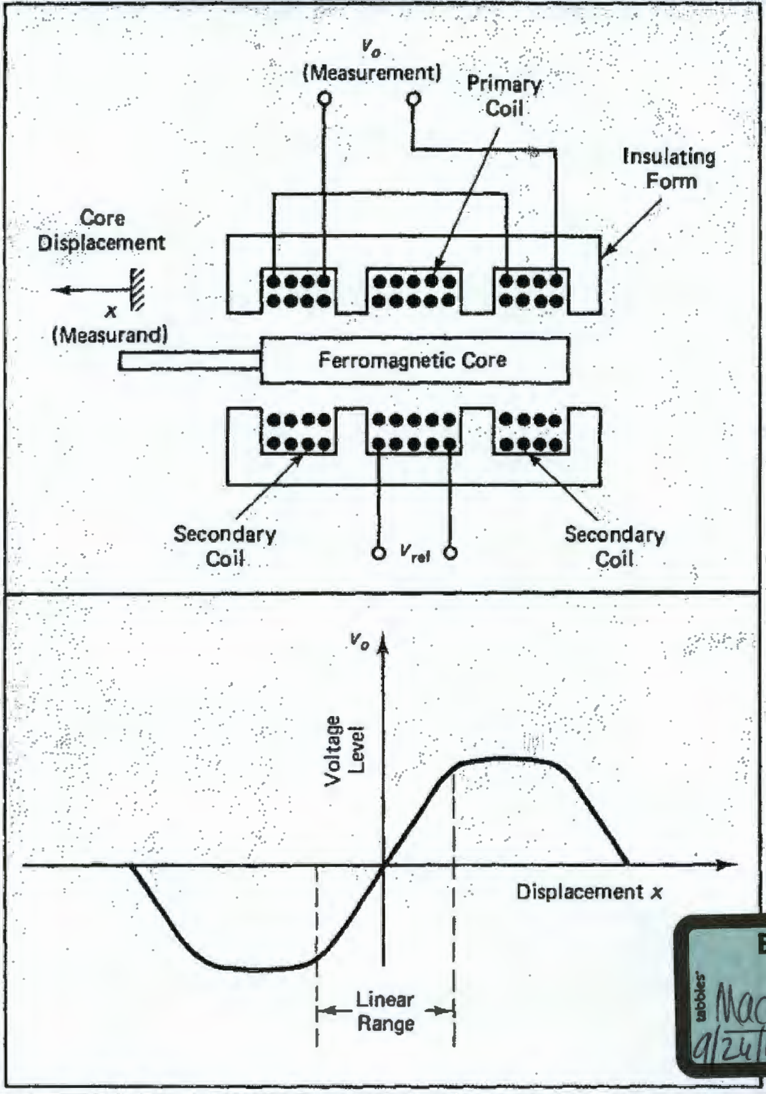
CONTROL SENSORS AND ACTUATORS



CLARENCE W. de SILVA

Control Sensors and Actuators

CLARENCE W. DE SILVA

PRENTICE HALL, Englewood Cliffs, New Jersey 07632

De Silva, Clarence W.
Control Sensors and actuators.

Includes bibliographies and index.
1. Automatic control. 2. Detectors. 3. Actuators.
I. Title.
TJ213.D38 1989 629.8 88-28998
ISBN 0-13-171745-6

To Charmaine, C.J., and Cheryl—as their senses *develop and as they become increasingly* active.

Editorial/production supervision and
interior design: David Ershun
Cover design: Ben Santora
Manufacturing buyer: Mary Ann Gloriande

The publisher offers discounts on this book when ordered
in bulk quantities. For more information, write:

Special Sales/College Marketing
Prentice Hall
College Technical and Reference Division
Englewood Cliffs, NJ 07632

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-171745-6

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Contents

1

Control, Instrumentation, and Design

1.1 INTRODUCTION

The demand for servnmechanisms in military applications during World War II provided much incentive and many resources for the growth of control technology. Early efforts were devoted to the development of analog controllers, which are electronic devices or circuits that generate proper drive signals for a plant (process). Parallel advances were necessary in actuating devices such as motors, solenoids, and valves that drive the plant. For feedback control, further developments in sensors and transducers became essential. With added snphistication in control systems, it was soon apparent that analog control techniques had serious limitations. In particular, linear assumptions were used to develop controllers even for highly nonlinear plants. Furthermore, complex and costly circuitry was often needed to generate even simple control signals. Consequently, mnst analog controllers were limited to on/off and proportional-integral-derivative (PID) actions, and lead and lag compensation networks were employed to compenstate for weaknesses in such simple contrnl actions.

The digital computer, first developed for large number-crunching jobs, was employed as a contrnller in complex control systems in the 1950s and 1960s. Originally, cost constraints restricted its use primarily to aerospace applications that required the manipulation of large amounts of data (complex models, several hundred signals, and thousands of system parameters) for control and that did not face serious cost restraints. Real-time contrnl requires fast computatinn, and this speed of computation is determined by the required control bandwidth (or the speed of control) and parameters (e.g., time constants, natural frequencies, and damping constants) of the process that is being controlled. For instance, prelaunch monitoring and control of a space vehicle would require digital data acquisition at very high sampling rates (e.g., 50,000 samples/second). As a result of a favorable decline of computation cost (both hardware and software) in subsequent years, widespread application of digital computers as control devices (i.e., digital control) has become feasible. Dramatic developments in large-scale integration (LSI) technology and microprocessors in the

1970s resulted in very significant drops in digital processing costs, which made digital control a very attractive alternative to analog cnntrol. Today, digital control has become an integral part of numerous systems and applications, including machine tools, robotic manipulators, automobiles, aircraft autopilots, nuclear power plants, traffic control systems, and chemical process plants.

Control engineers should be able to identify or select components for a control system, mndel and analyze individual components or overall systems, and choose parameter values so as to perform the intended functions of the particular system in accordance with some specifications. Component identification, analysis, selection, matching and interfacing, and system tuning (adjusting parameters to obtain the required response) are essential tasks in the instrumentatinn and design of a control system.

1.2 CONTROL SYSTEM ARCHITECTURE

Let us examine the generalized cnntrol system represented by the block diagram in figure 1.1. We have identified several discrete blocks, depending on various functions that take place in a typical contrnl system. Before proceeding, we must keep in mind that in a practical control system, this type of clear demarcation of components might be difficult; one piece of hardware might perform several functions, or more than one distinct unit of equipment might be associated with one function. Nevertheless, figure 1.1 is useful in understanding the architecture of a general control system. This is an analog control system because the associated signals depend on the continuous time variable; no signal sampling or data encoding is involved in the system.

Plant is the system or "process" that we are interested in controlling. By *control,* we mean making the system respond in a desired manner. To be able tn accomplish this, we must have access to the *drive system* or *actuator* of the plant. We apply certain *command signals,* or input, to the *controller* and expect the plant to behave in a desirable manner. This is the *open-loop control* situation. In this case, we do not use current information on *system response* to determine the control signals. In *feedback control systems,* the control loop has to be closed; *closed-loop control* means making *measurements* of system response and employing that information to generate control signals so as to correct any output errors. The output measurements are made primarily using analog devices, typically consisting of *sensor-transducer* units.

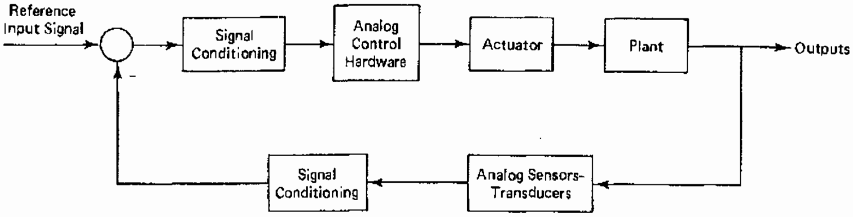


Figure 1.1. Components of a typical analog control system.

ligi-
has
hine
ants,

ntrol
oose
m in
tion,
e re-
ntrol

am in
func-
ep in
nents
more
erthe-
l sys-
on the
e sys-

y con-
ccom-
apply
behave
do not
n feed-
means
gener-
nts are
units.

Outputs

An important factor that we must consider in any practical control system is noise, including external disturbances. Noise may represent actual contamination of signals or the presence of other unknowns, uncertainties, and errors, such as parameter variations and modeling errors. Furthermore, weak signals will have to be amplified, and the form of a signal might have to be modified at various points of interaction. In these respects, *signal-conditioning methods* such as *filtering*, *amplification*, and *modulation* become important.

Identification of the hardware components (perhaps commercially available off-the-shelf items) corresponding to each functional block in figure 1.1 is one of the first steps of instrumentation. For example, in process control applications off-the-shelf analog proportional-integral-derivative (PID) controllers may be used. These controllers for process control applications have knobs or dials for control parameter settings—that is, proportional band or gain, reset rate (in repeats of the proportional action per unit time), and rate time constant. The control bandwidth (frequency range of operation) of these devices is specified. Various control modes—such as on/off, proportional, integral, and derivative, or combinations—are provided by the same control box.

Actuating devices (actuators) include DC motors, AC motors, stepper motors, solenoids, valves, and relays, which are also commercially available to various specifications. Potentiometers, differential transformers, resolvers, synchros, strain gauges, tachometers, piezoelectric devices, thermocouples, thermistors, and resistance temperature detectors (RTDs) are examples of sensors used to measure process response for monitoring performance and possible feedback. Charge amplifiers, lock-in amplifiers, power amplifiers, switching amplifiers, linear amplifiers, tracking filters, low-pass filters, high-pass filters, and notch filters are some of the signal-conditioning devices used in analog control systems. Additional components, such as power supplies and surge-protection units, are often needed in control, but they are not indicated in figure 1.1 because they are only indirectly related to control functions. Relays and other switching devices and modulators and demodulators may also be included.

1.3 DIGITAL CONTROL

Direct digital control (DDC) systems are quite similar to analog control systems. The main difference in a DDC system is that a digital computer takes the place of the analog controller in figure 1.1. Control computers have to be dedicated machines for real-time operation where processing has to be synchronized with plant operation and actuation requirements. This also requires a real-time clock. Apart from these requirements, control computers are basically no different from general-purpose digital computers. They consist of a processor to perform computations and to oversee data transfer, memory for program and data storage during processing, mass storage devices to store information that is not immediately needed, and input/output devices to read in and send out information. Digital control systems might utilize digital instruments and additional processors for actuating, signal-conditioning, or measuring functions, as well. For example, a stepper motor that responds with incremental motion steps when driven by pulse signals can be considered a digital actuator. Furthermore, it usually contains digital logic circuitry in its drive system. Similarly, a two-position solenoid is a digital (binary) actuator. Digital flow control may be accomplished using a digital control valve. A typical digital valve consists of a bank of orifices, each sized in proportion to a place value of a binary word $(2^i, i = 0, 1, 2, \ldots, n)$. Each orifice is actuated by a separate rapid-acting on/off solenoid. In this manner, many digital combinations of flow values can be obtained. Direct digital measurement of displacements and velocities can be made using shaft encoders. These are digital transducers that generate coded outputs (e.g., in binary or gray-scale representation) or pulse signals that can be coded using counting circuitry. Such outputs can be read in by the control computer with relative ease. Frequency counters also generate digital signals that can be fed directly into a digital controller. When measured signals are in the analog form, an analog front end is necessary to interface the transducer and the digital controller. Input/output interface boards that can take both analog and digital signals are available with digital controllers.

A block diagram of a direct digital control system is shown in figure 1.2. Note that the functions of this control system are quite similar to those shown in figure 1.1 for an analog control system. The primary difference is the digital controller (processor), which is used to generate the control signals. Therefore, analog measurements and reference signals have to be sampled and encoded prior to digital processing within the controller. Digital processing can be conveniently used for signal conditioning as well. Alternatively, digital signal processing (DSP) chips can function as digital controllers. However, analog signals are *preconditioned,* using analog circuitry prior to digitizing in order to eliminate or minimize problems due to *aliasing distortion* (high-frequency components above half the sampling frequency appearing as low-frequency components) and *leakage* (error due to signal truncation) as well as to improve the signal level and filter out extraneous noise. The drive sys-

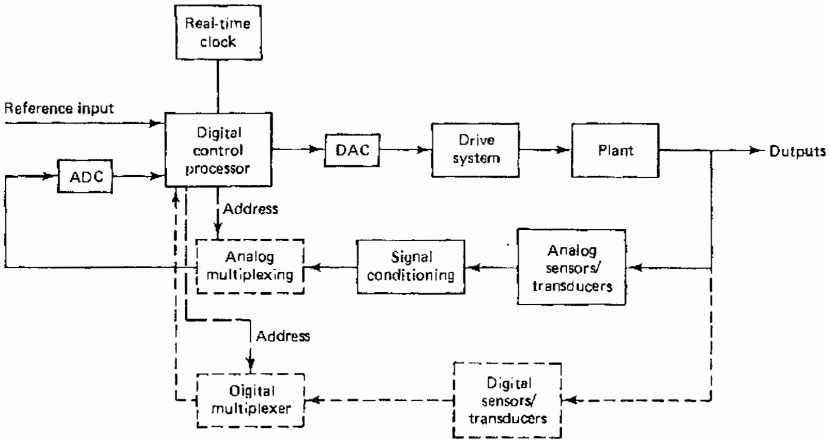


Figure 1.2. Block diagram of a direct digital control system.

tem of a plant typically takes in analog signals. Often, the digital output from controller has to be converted into analog form for this reason. Both *analog-to-digital conversion* (ADC) and *digital-to-analog conversion* (DAC) can be interpreted as signal-conditioning (modification) procedures. If more than one output signal is measured, each signal will have to be conditioned and processed separately. Ideally, this will require separate conditioning and processing hardware for each signal channel. A less expensive (but slower) alternative would be to time-share this expensive equipment by using a *multiplexer*. This device will pick one channel of data from a bank of data channels in a sequential manner and connect it to a common input device. Both analog and digital multiplexers are available. In a digital multiplexer, the input signals come from a bank of digital sensors, and the output signal itself, which would be in digital form, goes directly into the digital controller. High-speed multiplexers (e.g., over 50,000 switchings/second) use electronic switching.

For complex processes with a large number of input/output variables (e.g., a nuclear power plant) and with systems that have various operating requirements (e.g., the space shuttle), centralized direct digital control is quite difficult to implement. Some form of distributed control is appropriate in large systems such as manufacturing cells, factories, and multicomponent process plants. A favorite distributed control architecture is provided by heirarchical control. Here, distribution of control is available both geographically and functionally. An example for a three-level hierarchy is shown in figure 1.3. Management decisions, supervisory control, and coordination between plants are provided by the management (supervisory) computer, which is at the highest level (level 3) of the hierarchy. The next lower level computer generates control settings (or reference inputs) for each control region in the corresponding plant. Set points and reference signals are inputs to the direct digital control (DDC) computers that control each control region. The computers communicate using a suitable information network. Information transfer in both directions (up and down) should be possible for best performance and flexibility. In master–slave distributed control, only downloading of information is available.

1.4 SIGNAL CLASSIFICATION IN CONTROL SYSTEMS

A digital control system can be loosely interpreted as one that uses a digital computer as the controller. It is more appropriate, however, to understand the nature of the signals that are present in a control system when identifying it as a digital control system.

Analog signals are continuous in time. They are typically generated as outputs of a dynamic system. (Note that the dynamic system could be a signal generator or any other device, equipment, or physical system.) Analytically, analog signals are represented as functions of the continuous time variable t.

Sampled data are, in fact, pulse amplitude–modulated signals. In this case, information is carried by the amplitude of each pulse, with the width of the pulses kept constant. For constant sampling rate, the distance between adjacent pulses is also kept constant. In a physical situation, a pulse amplitude–modulated signal is generated through a *sample-and-hold operation*, in which the signal is sampled at

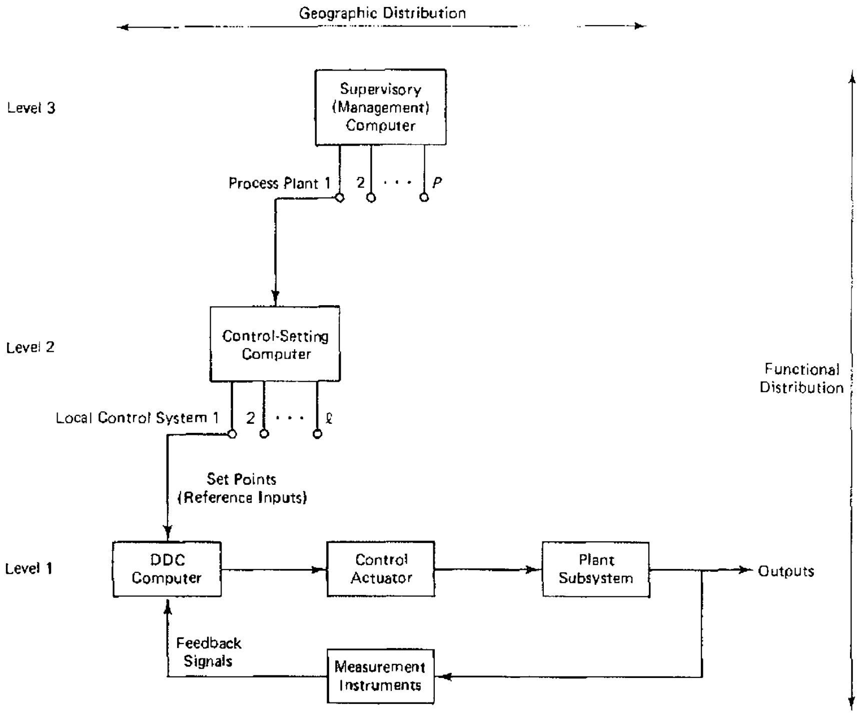


Figure 1.3. A three-level hierarchical control scheme.

the beginning of the *sampling period* and kept constant at that value, irrespective of the true value of the signal over that period. An important advantage of sampling is that expensive equipment can be shared among many signals. Furthermore, sampling is necessary in real-time digital processing to allow for the processing time. Analytically, sampled data consist of a *sequence* of numbers (or a function of integer variable).

Digital data are coded numerical data. For example, a binary code or ASCII (American Standard Code for Information Interchange) may be used to represent each value in a sequence of digital data. The code itself determines the actual value of a particular unit of digital information. Typically, digital data are generated by *digital processors, digital transducers, counters, encoders,* and other such digital devices.

Table 1.1 summarizes the identifying characteristics of these three types of data. *Analog systems* generate analog signals only. *Sampled-data systems* depend on analog data as well as sampled data. *Digital systems,* however, utilize all three types of signals, generally at different levels of interaction. Sampled-data systems and digital systems may be modeled using discrete-time models (see table 1.2).

TABLE 1.1 SIGNAL CATEGORIES FOR IDENTIFYING CONTROL SYSTEM TYPES

Signal (data) category	Description
Analog signals (data)	Continuous in time t; typically represents an output of a dynamic system
Sampled data	Pulse amplitude-modulated signals Information carried by pulse amplitude Typically generated by sample-and-hold process
Digital data	Coded numerical data; the particular code determines the numerical value Typically generated by digital processors, digital transducers, and counters

TABLE 1.2 REPRESENTATIVE ANALYTICAL CHARACTERISTICS OF CONTINUOUS-TIME AND DISCRETE-TIME SYSTEMS

	Analytical model	
System	Time domain	Transfer-function domain
Continuous-time systems	Differential equations	Laplace transfer functions or Fourier frequency response functions
Discrete-time systems	Difference equations	Z-transform transfer functions

Example 1.1

The sampling period ΔT for data acquisition is an important parameter in real-time digital control. Discuss the significance of the sampling period.

Solution On the one hand, ΔT has to be sufficiently large so that required processing and data transfer can be done during that time for each control step. This is crucial in real-time control. On the other hand, ΔT should be small enough to meet control bandwidth and process dynamics requirements. Shannon's sampling theorem states that in a sampled signal, the maximum meaningful frequency is the Nyquist frequency f_c, which is given by half the sampling rate:

$$f_c = \frac{1}{2\Delta T} \tag{1.1}$$

It follows that we should select ΔT such that the significant frequency content of the input/output signals of the particular process stays within the Nyquist frequency. Note that to be able to control the process effectively, all natural frequencies of interest in the plant should be smaller than f_c. Once f_c is chosen in this manner for digital control, it is also important to choose analog components, such as signal-conditioning devices in the control system, to have an operating bandwidth larger than f_c. For example, if the operating bandwidth of a robotic manipulator is specified to be 50 Hz, one must make sure that the associated analog sensors and transducers (resolvers, tachometers, etc.) and signal-conditioning devices (e.g., low-pass filters, charge amplifiers) have an operating bandwidth greater than 50 Hz—preferably about 200 Hz. Furthermore, the sampling period has to be smaller than 10 ms (from equation 1.1), preferably about 2 ms. It is then necessary to make sure that the control computer is capable of doing all the processing needed in each control increment within this time. Otherwise, distribution of control tasks might be needed. Parallel processing is another option. Another alternative is to employ a hardware implementation of the controller. Simplification of control algorithms should also be attempted, but without sacrificing the accuracy requirements. In general, distributed control is better than using a single control computer of larger capacity and faster speed.

1.5 ADVANTAGES OF DIGITAL CONTROL

The current trend toward using dedicated, microprocessor-based, and often decentralized (distributed) digital control systems in industrial applications can be rationalized in terms of the major advantages of digital control. The following are some of the important considerations.

1. Digital control is less susceptible to noise or parameter variation in instrumentation because data can be represented, generated, transmitted, and processed as binary words, with bits possessing two identifiable states.
2. Very high accuracy and speed are possible through digital processing. Hardware implementation is usually faster than software implementation.
3. Digital control can handle repetitive tasks extremely well, through programming.
4. Complex control laws and signal conditioning methods that might be impractical to implement using analog devices can be programmed.
5. High reliability in operation can be achieved by minimizing analog hardware components and through decentralization using dedicated microprocessors for various control tasks (see figure 1.3).
6. Large amounts of data can be stored using compact, high-density data storage methods.
7. Data can be stored or maintained for very long periods of time without drift and without being affected by adverse environmental conditions.
8. Fast data transmission is possible over long distances without introducing dynamic delays, as in analog systems.
9. Digital control has easy and fast data retrieval capabilities.
10. Digital processing uses low operational voltages (e.g., 0–12 V DC).
11. Digital control has low overall cost.

Some of these features should be obvious; the rest should become clear as we proceed through the book.

ust make
ers, etc.)
an oper-
the sam-
t 2 ms. It
ig all the
stribution
her alter-
n of con-
, require-
nputer of

n decen-
rational-
some of

strumen-
rocessed

g. Hard-

program-

impracti-

hardware
ssors for

a storage

nout drift

ucing dy-

s we pro-

1.6 FEEDFORWARD CONTROL

Many control systems have inputs that do nnt participate in feedback control. In other words, these inputs are not compared with feedback (measurement) signals to generate control signals. Some of these inputs might be important variables in the plant (process) itself. Others might be undesirable inputs, such as external disturbances that are unwanted yet unavoidable. Performance of a control system can generally be improved by measuring these (unknown) inputs and somehow using the information to generate control signals. Since the associated measurement and control (and compensation) take place in the forward path nf the control system, this method of control is known as *feedforward control.* Note that in feedback control, unknown "outputs" are measured and compared with known (desired) inputs to generate control signals. In feedforward control, unknown "inputs" are measured and that informatinn, alnng with desired inputs, is used to generate control signals that can reduce errors due tn these unknown inputs or variations in them.

A block diagram of a typical control loop that uses feedforward cnntrol is shown in figure 1.4. In this system, in addition to feedback cnntrnl, feedfnrward cnntrol is used to reduce the effects of a disturbance input that enters the plant. The disturbance input is measured and fed into the controller. The controller uses this informatinn tn modify the control action sn as to cnmpensate fnr the effect of the disturbance input.

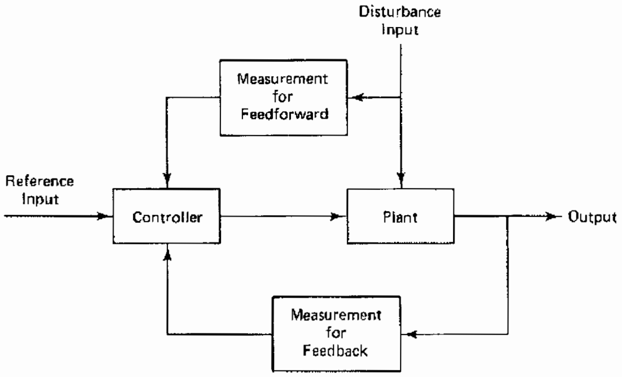


Figure 1.4. A typical feedback loop that uses feedforward control.

As a practical example, cnnsider the natural gas home heating system shown in figure 1.5a. A simplified block diagram of the system is shown in figure 1.5b. In conventional feedback control, the room temperature is measured and its deviation from the desired temperature (set point) is used to adjust the natural gas flow into the furnace. On/off control is used in most such applications. Even if propnrtinnal or three-mode (propnrtinnal-integral-derivative) control is employed, it is not easy to steadily maintain the room temperature at the desired value if there are large

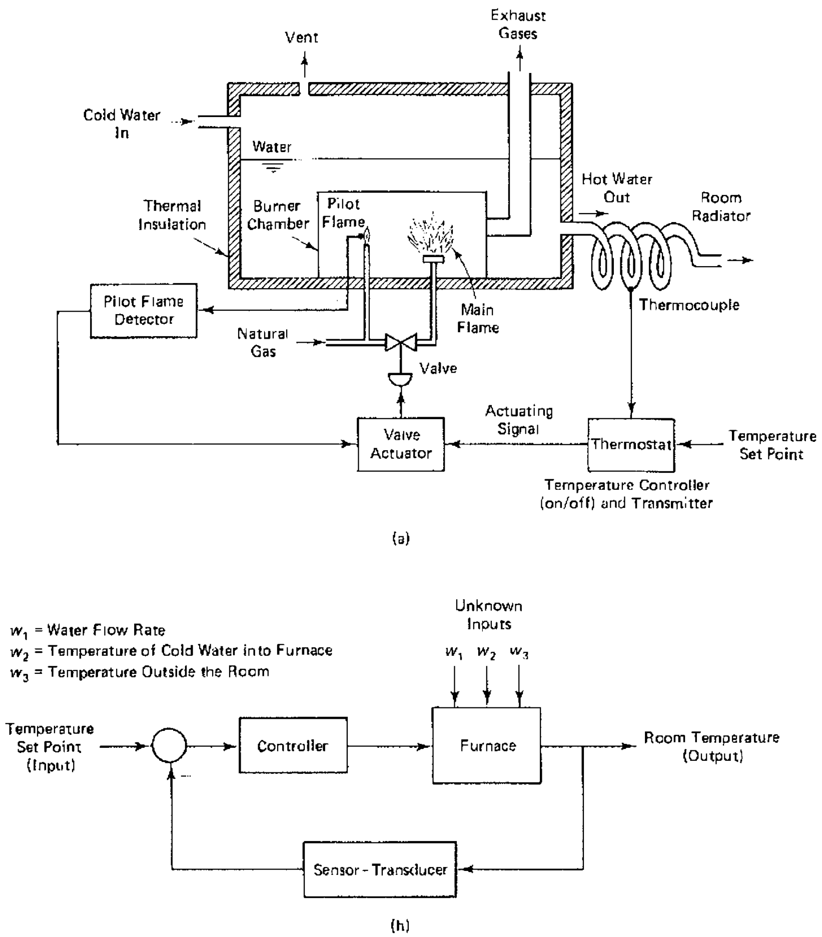


Figure 1.5. (a) A natural gas home heating system. (b) A block diagram representation of the system.

changes in other (unknown) inputs to the system, such as water flow rate through the furnace, temperature of water entering the furnace, and outdoor temperature. Better results can be obtained by measuring these disturbance inputs and using that information in generating the control action. This is feedforward control. Note that in the absence of feedforward control, any changes in the inputs w_1, w_2, and w_3 in figure 1.5 will be detected only through their effect on the feedback signal (room temperature). Hence, the subsequent corrective action can lag behind the cause (changes in w_i) considerably. This delay will lead to large errors and possible instability problems. With feedforward control, information on the disturbance inputs w_i will be available to the controller immediately, thereby speeding up the control action and also improving the response accuracy. Faster action and improved accuracy are two very desirable effects of feedforward control.

In some applications, control inputs are computed using accurate dynamic models for the plants, and the computed inputs are used for control purposes. This is a popular way for controlling robotic manipulators, for example. This method is also known as feedforward control. To avoid confusion, however, it is appropriate to denote this method as *computed-input control.*

1.7 INSTRUMENTATION AND DESIGN

In the previous discussion, we have identified several characteristic constituents of a control system. Specifically, we are interested in

- The *plant,* or the dynamic system to be controlled
- Signal *measurement* for system evaluation (monitoring) and for feedback and feedforward control
- The *drive system* that actuates the plant
- *Signal conditioning* by filtering and amplification and *signal modification* by modulation, demodulation, ADC, DAC, and so forth, into an appropriate form
- The *controller* that generates appropriate drive signals for the plant

Each function or operation within a control system can be associated with one or more physical devices, components, or pieces of equipment, and one hardware unit may accomplish several of the control system functions. By *instrumentation,* in the present context, we mean the identification of these various instruments or hardware components with respect to their functions, operation, and interaction with each other and the proper selection and interfacing of these components for a given application—in short, "instrumenting" a control system.

By *design,* we mean the process of selecting suitable equipment to accomplish various functions in the control system; developing the system architecture; matching and interfacing these devices; and selecting the parameter values, depending on the system characteristics, in order to achieve the desired objectives of the overall control system (i.e., to meet design specifications), preferably in an optimal manner and according to some performance criterion. In the present context, design is included as an instrumentation objective. In particular, there can be many designs that meet a given set of performance requirements.

Identification of key design parameters, modeling of various components, and analysis are often useful in the design process. This bonk provides fundamentals of sensing and actuation for electromechanical control systems. Emphasis is placed on control systems that perform motion- and force-related dynamic tasks. Sensors and transducers and actuators in this category will be discussed with respect to their performance specification, principles of operation, physical characteristics, modeling and analysis, selection, component interfacing, and determination of parameter values. Both analog and digital devices will be studied. Design examples and case studies drawn from applications such as automated manufacturing and robotics, transit vehicles, dynamic testing, and process control will be discussed throughout the book.

PROBLEMS

1.1. Giving appropriate examples, compare and contrast analog signals, sampled-data signals, and digital signals. What are the relative advantages and disadvantages of these three types of data?

1.2. What are differential equations and what are difference equations? Explain their significanace in the context of control system analysis, discussing the need for their solution as related to a digital control system.

1.3. **(a)** What is an open-loop control system and what is a feedback control system? Give one example of each case.

(b) A simple mass-spring-damper system (simple oscillator) is excited by an external force $f(t)$. Its displacement response y (see figure P1.3a) is given by the differential equation

$$m\ddot{y} + b\dot{y} + ky = f(t)$$

A block diagram representation of this sytem is shown in figure P1.3b. Is this a feedback control system? Explain and justify your answer.

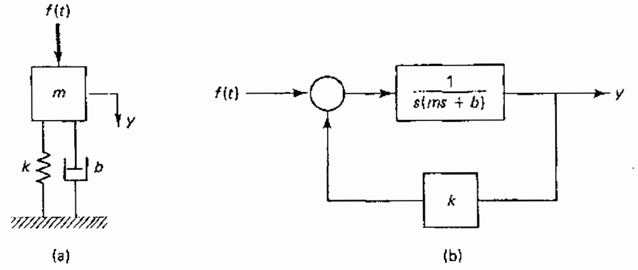


Figure P1.3. (a) A mechanical system representing a simple oscillator. (b) A block diagram representation of the simple oscillator.

1.4. You are asked to design a control system to turn on lights in an art gallery at night, provided that there are people inside the gallery. Explain a suitable control system, identifying the open-loop and feedback functions, if any, and describing the control system components.

1.5. Into what classification of control system components (actuators, signal modification devices, controllers, and measuring devices) would you put the following?

(a) Stepping motor
(b) Proportional-plus-intergration circuit
(c) Power amplifier
(d) ADC
(e) DAC
(f) Optical incremental encoder
(g) Process computer
(h) FFT analyzer

1.6. In feedforward control (computed-torque control) of robotic manipulators, joint torques (or forces) are computed from joint motion variables (displacements, velocities, and accelerations) using a suitable dynamic mndel fnr the manipulator. A 6-df (degree of freedom) model of a particular manipulatnr, formulated using nonrecursive Lagrangian dynamic equations, required aproximtely 10 s for one cycle of torque computation on a VAX 11/750 minicomputer. If a control bandwidth of at least 10 Hz is needed for typical applications of this robot, discuss whether real-time computed-torque control is possible in this case. Suggest improvements.

1.7. Discuss possible sources of error that can make open-loop control or feedforward control meaningless in some applications. How would you correct the situation?

1.8. A flexible manufacturing cell consists of a set nf machine tools, robots, parts transfer units (e.g., conveyors), and gaging stations (e.g., vision-based) managed by a single host computer. Each robot has a separate computer (e.g., Motorola M68000) for trajectory generation and fnr command setting. Each joint of a manipulator is controlled using a separate microprncessor (e.g., TMS-320), through direct digital control. Identifying control functions in each level, discuss the complete system as a hierarchical control system. A photograph of a flexible manufacturing cell is shown in figure P1.8.

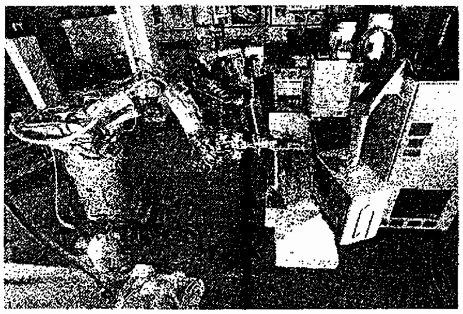


Figure P1.8. A flexible manufacturing cell (courtesy of the Robotics Institute, Carnegie Mellon University).

1.9. Compare analog control and direct digital control for motion control in high-speed applications of industrial manipulators. Give the advantages and disadvantages of each control method for this application.

1.10. Resolution of a device is defined as the smallest useful and detectable increment (change) in the output of the device. Dynamic range is the ratio of output range to resolution, expressed in decibels (dB). A digital control valve has four nrifices. It represents a four-digit binary number, and flow through each orifice is proportional to the corresponding position value. The smallest orifice allows a flow of f_o through it. Calculate the resolution and the dynamic range of the valve.

1.11. A soft-drink bottling plant uses an automated bottle-filling system. Desribe the operation of such a system, indicating various components in the control system and their functions. Typical components would include conveyor belt; a motor for the conveyor, with start/stop controls; a measuring cylinder, with inlet valve, exit valve, and level sensors; valve actuators; and an alignment sensor fnr the bottle and the measuring cylinder.

1.12. Consider the natural gas home heating system shown in figure 1.5. Describe the functions of various components in the system and classify them into controller, actuator, sensor, and signal modification function grnups. Explain the operation of the overall system and suggest pnssible improvements to obtain more stable and accurate temperature control.

1.13. In each of the following examples, indicate at least one (unknown) input that should be measured and used for feedforward control to improve the accuracy of the control system.

(a) A servo system for positioning a mechanical load. The servo motor is a field-controlled DC motor, with position feedback using a potentiometer and velocity feedback using a tachometer.

(b) An electric heating system for a pipeline carrying a liquid. The exit temperature of the liquid is measured using a thermocouple and is used to adjust the power of the heater.

(c) A room heating system. Room temperature is measured and compared with the set point. If it is low, a valve of a steam radiator is opened; if it is high, the valve is shut.

(d) An assembly robot that grips a delicate part to pick it up without damaging the part.

(e) A welding robot that tracks the seam of a part to be welded.

1.14. Consider the system shown by the block diagram in figure P1.14a. Note that

$G_p(s)$ = plant transfer function

$G_c(s)$ = controller transfer function

$H(s)$ = feedback transfer function

$G_f(s)$ = feedforward compensation transfer function

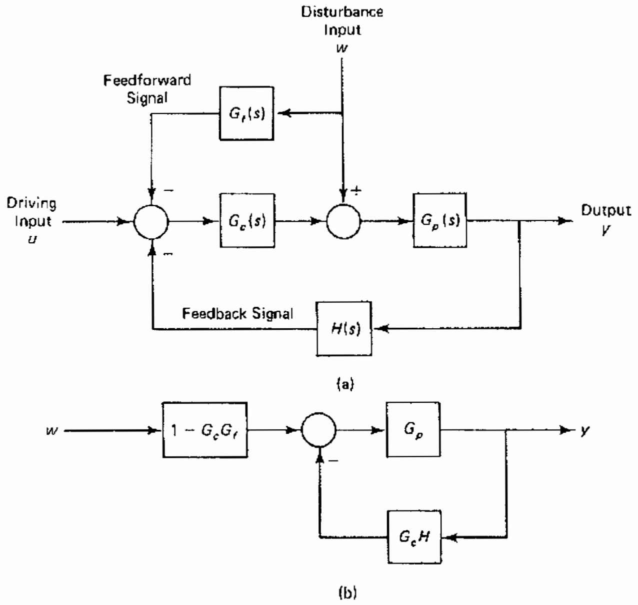


Figure P1.14. (a) A block diagram for a system with feedforward control. (b) Reduced form in the absence of the driving input.

ut that should be
; the control sys-

notor is a field-
ster and velocity

it temperature of
the power of the

ared with the set
iigh, the valve is

ut damaging the

ote that

Output
y

→ y

l control.

The disturbance input w is measured, compensated using G_f, and fed into the controller, along with the driving input u.

(a) Obtain the transfer function relationship between the output y and the driving input u in the absence of the disturbance input w.

(b) Show that in the absence of u, the block diagram can be drawn as in figure P1.14b. Obtain the transfer relationship between y and w in this case.

(c) From (a) and (b), write an expression for y in terms of u and w.

(d) Show that the effect of disturbance is fully compensated if the feedforward compensator is given by

$$G_f(s) = \frac{1}{G_c(s)}$$

1.15. A typical input variable is identified for each of the following examples of dynamic systems. Give at least one output variable for each system.

(a) Human body: neuroelectric pulses
(b) Company: infnrmation
(c) Power plant: fuel rate
(d) Automobile: steering wheel movement
(e) Robot: voltage to joint motor

1.16. Measuring devices (sensors-transducers) are useful in measuring outputs of a process for feedback control.

(a) Give other situations in which signal measurement would be important.
(b) List at least five different sensors used in an automobile engine.

1.17. Hierarchical control has been applied in many industries, including steel mills, oil refineries, glass works, and automated manufacturing. Most applications have been limited to two or three levels of hierarchy, however. The lower levels usually cnnsist of tight servo loops, with bandwidths on the order of 1 kHz. The upper levels typically control production planning and scheduling events measured in units of days or weeks.

(a) Estimate event duration at the lowest level and control bandwidth (in hertz) at the highest level for this type of application.

(b) A five-level hierarchy for a flexible manufacturing facility is as follows: The lowest level (level 1) handles servo control of robotic manipulator joints and machine tool degrees of freedom. The secnnd level performs activities such as coordinate transformation in machine tools that are required in generating control commands for various servo loops. The third level converts task commands into motion trajectories (of manipulator end effector, machine tool bit, etc.) expressed in world coordinates. The fnurth level converts complex and general task commands into simple task commands. The top level (level 5) performs supervisory control tasks for various machine tools and material-handling devices, including conrdination, scheduling, and definition of basic moves. Suppose that this facility is used as a flexible manufacturing cell for turbine blade production. Using diagrams to show tasks at various levels of the hierarchy, describe the operation of this manufacturing cell.

1.18. According to some observers in the process control industry, early brands of analog control hardware had a product life of about twenty years. New hardware controllers can become obsolete in a couple of years, even before their development costs are recovered. As a control instrumentation engineer responsible for developing an off-the-shelf process controller, what features would you incorporate in the controller to correct this problem to a great extent?

1.19. The programmable controller (PC) is a sequential control device that can sequentially and repeatedly activate a series of output devices (e.g., motors, valves, alarms, signal lights) on the basis of the states of a series of input devices (e.g., switches, two-state sensors). Show how a programmable controller and a vision system consisting of a solid-state camera and a simple image processor (say, with an edge-detection algorithm) could be used for sorting fruits on the basis of quality and size for packaging and pricing.

REFERENCES

BARNEY, G. C., *Intelligent Instrumentation*. Prentice-Hall, Englewood Cliffs, N.J., 1985.

BECKWITH, T. G., BUCK, N. L., and MARANGANI, R. D. *Mechanical Measurements,* 3d ed. Addison-Wesley, Reading, Mass., 1982.

DALLY, J. W., RILEY, W. F., and MCCONNELL, K. G. *Instrumentation for Engineering Measurements*. Wiley, New York, 1984.

DESILVA, C. W. *Dynamic Testing and Seismic Qualification Practice*. Lexington Books, Lexington, Mass., 1983.

———(Consulting Ed.). *Measurements and Control Journal,* all issues from February 1983 to December 1988.

FRANKLIN, G. F., and POWELL, J. D. *Digital Control of Dynamic Systems*. Addison-Wesley, Reading, Mass., 1980.

GIBSON, J. E., and TUTEUR, F. B. *Control System Components*. McGraw-Hill, New York, 1958.

HORDESKI, M. F. *The Design of Microprocessor, Sensor, and Control Systems*. Reston, Reston, Va., 1985.

JOHNSON, C. D. *Microprocessor-Based Process Control*. Prentice-Hall, Englewood Cliffs, N.J., 1984.

POTVIN, J. *Applied Process Control Instrumentation*. Reston, Reston, Va., 1985.

itches, two-state
. consisting of a
ge-detection al-
ze for packaging

N.J., 1985.
rements, 3d ed.
ıgineering Mea-
:on Books, Lex-
ı February 1983
\ddison-Wesley,
Iill, New York,
ns. Reston, Re-
glewood Cliffs,
985.

2

Performance Specification and Component Matching

2.1 INTRODUCTION

In feedback control systems, plant response is measured and compared with a reference input and the error is automatically employed in controlling the plant. It follows that a *measurement system* is an essential component in any feedback control system and forms a vital link between the plant and the controller. Measurements are needed in many engineering applications. The measurement process has to be automated, however, in control systems applications.

A typical measurement system consists of one or more *sensor-transducer* units and associated *signal-conditioning* (and modification) devices (see figure 2.1). Filtering to remove unwanted noise and amplification to strengthen a needed signal are considered signal conditioning. Analog-to-digital conversion (ADC), digital-to-analog conversion (DAC), modulation, and demodulation are signal modification methods. Note that signal conditioning can be considered under the general heading of signal modification. Even though data recording is an integral function in a typical data acquisition system, it is not a crucial function in a feedback cnntrol system. For this reason, we shall not go into details of data recording devices in this book. In a multiple measurement environment, a *multiplexer* could be employed prior to or following the signal-conditioning process, in order tn pick one measured signal at a time from a bank of data channels for subsequent processing. In this manner, one unit of expensive prncessing hardware can be time-shared between several signals. Sensor-transducer devices are predominantly analog components that generate analog signals, even though *direct digital transducers* are becoming increasingly

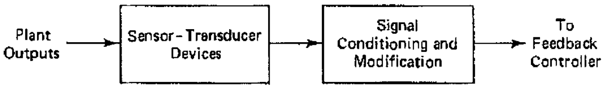

Plant Outputs
Sensor-Transducer Devices
Signal Conditioning and Modification
To Feedback Controller

Figure 2.1. Schematic representation of a measurement system.

popular in digital control applications. When analog transducers are employed, *analog-to-digital converters* (ADCs) have to be used to convert analog signals into digital data for digital control. This signal modification process requires sampling of analog signals at discrete time points. Once a value is sampled, it is encoded into a digital representation such as straight binary code, a gray code, binary-coded decimal (BCD) code, or American Standard Code for Information Interchange (ASCII). The changes in an analog signal due to its transient nature should not affect this process of ADC. To guarantee this, a *sample-and-hold operation* is required during each sampling period. For example, the value of an analog signal is detected (sampled) in the beginning of each sampling period and is assumed constant (held) throughout the entire sampling period. This is, in fact, the zero-order hold operatinn. The operations of multiplexing, sampling, and digitizing have to be properly synchronized under the control of an accurate timing device (a *clock*) for proper operation of the control system. This procedure is shown schematically in figure 2.2.

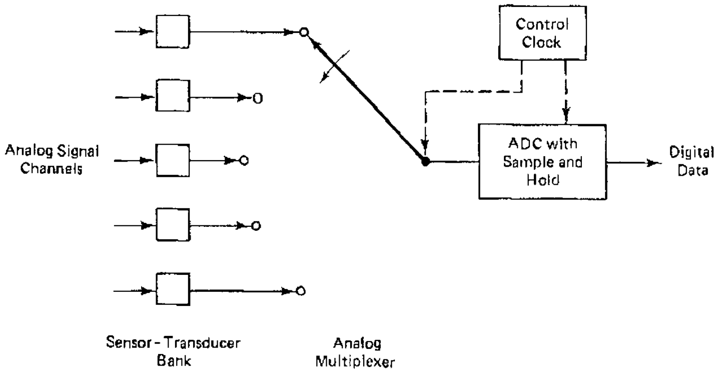


Figure 2.2. Measurement, multiplexing, and analog-to-digital conversion.

All devices that assist in the measurement procedure can be interpreted as components of the measurement system. Selection of available components for a particular application or design of new components should rely heavily on performance specifications for these components. A great majority of instrument ratings provided by manufacturers are in the form of static parameters. In control applications, however, dynamic performance specifications are also very important. In this chapter, we shall study instrument ratings and parameters for performance specification, pertaining to both static and dynamic characteristics of instruments.

When two or more components are interconnected, the behavior of individual components in the overall system can deviate significantly from their behavior when each component operates independently. Matching of components in a multicomponent system, particularly with respect to their impedance characteristics, should be done carefully in order to improve system performance and accuracy. In this chapter, we shall also study basic concepts of impedance and component matching. Al-

e employed,
; signals into
; sampling of
ıcoded into a
'-coded deci-
ıge (ASCII).
'fect this pro-
ıuired during
l is detected
ınstant (held)
: hold opera-
› be properly
or proper op-
ı figure 2.2.

Oigital
Oata

ɔn.

nterpreted as
ɔonents for a
ly on perfor-
ıment ratings
ıtrol applica-
ɪrtant. In this
performance
ıstruments.
of individual
ehavior when
multicompo-
:s, should be
In this chap-
ıatching. Al-

though the discussion is primarily limited to components in a measurement system, the ideas are applicable to many other types of components in a control system. Discussions and developments in this chapter are quite general; they do not address specific designs or hardware components. Specific instruments, their operating details, and physical hardware will be discussed in subsequent chapters.

2.2 SENSORS AND TRANSDUCERS

The *output variable* (or response) that is being measured is termed the *measurand.* Examples are acceleratinn and velocity of a vehicle, temperature and pressure of a process plant, and current through an electric circuit. A measuring device passes through two stages while measuring a signal. First, the measurand is *sensed.* Then, the measured signal is *transduced* (or converted) into a form that is particularly suitable for transmitting, signal conditioning, processing, or driving a controller or actuator. For this reason, output of the transducer stage is often an electrical signal. The measurand is usually an analog signal, because it represents the output of a dynamic system in feedback control applications. Transducer output is discrete in direct digital transducers. This facilitates the direct interface of a transducer with a digital processor. Since the majority of transducers used in control system applications are still analog devices, we shall consider such devices first (in chapters 3 and 4). Digital transducers will be discussed subsequently (in chapter 5).

The sensor and transducer stages of a typical measuring device are represented schematically in figure 2.3a. As an example, consider the operation of a piezoelectric accelerometer (figure 2.3b). In this case, acceleration is the measurand. It is first converted into an inertia force through a mass element and is exerted on a piezoelectric crystal within which a strain (stress) is generated. This is considered the sensing stage. The stress generates a charge inside the crystal, which appears as an electric signal at the output of the accelerometer. This stress-to-charge conversion or stress-to-voltage conversion can be interpreted as the transducer stage.

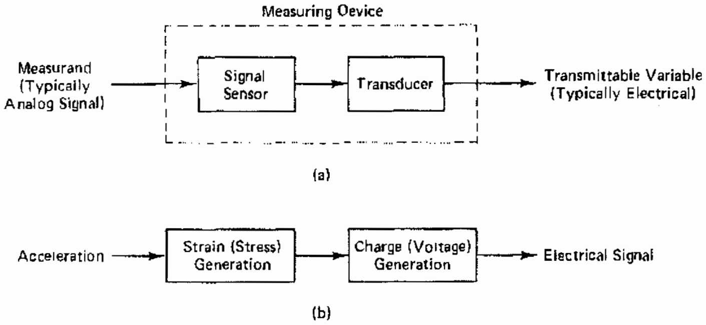


Figure 2.3. (a) Schematic representation of a measuring device. (b) Operation of a piezoelectric accelerometer.

A complex measuring device can have more than one sensing stage. More often, the measurand goes through several traosducer stages before it is available for control and actuating purposes. Sensor and transducer stages are functiooal stages, and sometimes it is not easy or even feasible to identify physical elements associated with them. Furthermore, this separation is not very important io using existing devices. Proper separation of sensor and transducer stages (physically as well as functionally) can be crucial, however, when designing new measuring instruments.

In some books, signal-conditioning devices such as electronic amplifiers are also classified as transducers. Since we are treating signal-conditioning and modification devices separately from measuring devices, this unified classification is avoided whenever possible in this book. Instead, the term *transducer* is used primarily in relation to measuring instruments. Following the common practice, however, the terms *sensor* and *transducer* will be used interchangeably to denote measuring instruments.

2.3 TRANSFER FUNCTION MODELS FOR TRANSDUCERS

Pure transducers depend on nondissipative coupling in the transduction stage. *Passive transducers* (sometimes called *self-generating transducers*) depend on their power transfer characteristics for operation. It follows that pure transducers are essentially passive devices. Some examples are *electromagnetic, thermoelectric, radioactive, piezoelectric,* and *photovoltaic* transducers. *Active transducers,* on the other hand, do not depend on power conversion characteristics for their operation. A good example is a *resistive* transducer, such as a potentiometer, that depends on its power dissipation to generate the output signal. Note that an active transducer requires a separate power source (power supply) for operation, whereas a passive transducer derives its power from a measured signal (measurand). In this classification of transducers, we are dealing with power in the immediate transducer stage associated with the measurand, not the power used in subsequent sigoal conditioning. For example, a piezoelectric charge generation is a passive process. But in order to condition the generated charge, a charge amplifier that uses an auxiliary power source would be needed.

There are advantages aod disadvantages in both types of transducers. In particular, since passive transducers derive their energy almost entirely from the measurand, they generally tend to distort (or load) the measured signal to a greater extent than an active transducer would. Precautions can be taken to reduce such loading effects, as will be discussed in a future section. On the other hand, passive transducers are generally simple in design, more reliable, and less costly.

A majority of sensors-transducers can be interpreted as *two-port elements* in which, under steady conditions, energy (or power) transfer into the device takes place at the *input port* and energy (or power) transfer out of the device takes place at the *output port*. Each port of a two-port transducer has a *through variable,* such as force or current, and an *across variable,* such as velocity or voltage, associated with it. Through variables are sometimes called *flux variables,* and across variables are called *potential variables*. Through variables are not always the same as *flow vari-*

ıge. More of-
; available for
:tional stages,
nts associated
.ısing existing
.ly as well as
instruments.
amplifiers are
ditioning and
lassification is
; used primar-
:ice, however,
ote measuring

on stage. *Pas-*
pend on their
;ducers are es-
noelectric, ra-
lucers, on the
ir operation. A
depends on its
transducer re-
a passive trans-
classification of
tage associated
ioning. For ex-
order to condi-
' power source

cers. In partic-
ɪm the measur-
ı greater extent
uch loading ef-
ive transducers

ort elements in
e device takes
e takes place at
:riable, such as
associated with
:s variables are
ıe as *flow vari-*

ables, which are used exclusively in *bond graph* models. Similarly, across variables are not the same as *effort variables,* which are used in bond graph terminology. For example, force is an effort variable, but it is also a through variable. Similarly, velocity is a flow variable and is also an across variable. The concept of effort and flow chanical impedance, but in analysis, mechanical impedance is not analogous to electrical impedance.

A two-port device can be modeled hy the transfer relation

$$\mathbf{G}\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} v_o \\ f_o \end{bmatrix} \tag{2.1}$$

where $\mathbf{G}$ is a 2×2 transfer function matrix, v_i and f_i denote the across and through variables at the input port, and v_o and f_o denote the corresponding variables at the output port (figure 2.4). This representation essentially assumes a *linear model* for transducer, so that the associated transfer functions (elements in matrix $\mathbf{G}$) are defined and valid. Such transducers are known as *ideal transducers.* Note that at a given port, if one variable is considered the input variable to the system, the other automatically becnmes the output variable of that system.

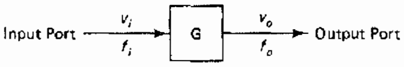


Figure 2.4. Two-port representation of a passive transducer.

Matrix transfer-function models are particularly suitable for transducers whose overall transduction process can be broken down into two or more simpler transducer stages. For example, consider a pressure transducer consisting of a bellows mechanism and a linear variable differential transformer (LVDT). In this device, the pressure signal is converted into a displacement by the pneumatic bellows mechanism. The displacement is converted, in turn, into a voltage signal by the LVDT. (The operation of an LVDT will be discussed in chapter 3.) If the transfer-function matrix for each transducer stage is known, the combined model is obtained by simply multiplying the two matrices in the proper order. To illustrate the method further, consider the *generalized series element* (*electrical impedance* or *mechanical mobility*) and the *generalized parallel element* (*electrical admittance* or *mechanical impedance*), which are denoted by Z and Y, respectively. The corresponding circuit representations are shown in figure 2.5. The model for the series-element transducer is

$$\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} 1 & Z \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_o \\ f_o \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & -Z \\ 0 & 1 \end{bmatrix}\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} v_o \\ f_o \end{bmatrix} \tag{2.2}$$

and the model for the parallel-element transducer is

$$\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ Y & 1 \end{bmatrix}\begin{bmatrix} v_o \\ f_o \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ -Y & 1 \end{bmatrix}\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} v_o \\ f_o \end{bmatrix} \tag{2.3}$$

These relations can be easily verified. The expressions for Z and Y for the three basic (ideal) electrical elements—resistance, inductance, and capacitance—are summarized in table 2.1. Note that s is the Laplace variable. In the frequency domain, s should be replaced by $j\omega$, where ω is the frequency (radians/second) and $j = \sqrt{-1}$.

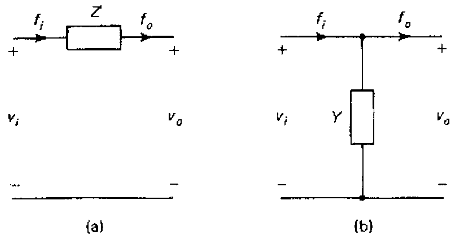


Figure 2.5. (a) Generalized series element (electrical impedance or mechanical mobility). (b) Generalized parallel element (electrical admittance or mechanical impedance).

TABLE 2.1 IMPEDENCE AND ADMITTANCE EXPRESSIONS FOR THE THREE IDEAL ELECTRIC ELEMENTS

Element	Impedance Z	Admittance Y
Resistance R	R	$\frac{1}{R}$
Inductance L	Ls	$\frac{1}{Ls}$
Capacitance C	$\frac{1}{Cs}$	Cs

A disadvantage of using through variables and across variables in the definition of impedance transfer functions is apparent when comparing electrical impedance with mechanical impedance. The definition of mechanical impedance is force/velocity in the frequency domain. This is a ratio of (through variable)/(across variable), whereas electrical impedance, defined as voltage/current in the frequency domain, is a ratio of (across variable)/(through variable). Since both force and voltage are "effort" variables and velocity and current are "flow" variables, it is convenient to use bond graph notation in defining impedance. Specifically,

$$\text{impedance (electrical or mechanical)} = \frac{\text{effort}}{\text{flow}}$$

In other words, impedance measures how much effort is needed to drive a system at unity flow.

Caution must be exercised in analyzing interconnected systems with mechanical impedance, because mechanical impedance cannot be manipulated using the rules for electrical impedance. For example, if two electric components are connected in series, the current (flow variable is the through variable) will be the same for both components, and voltage (effort variable is the across variable) will be additive. Accordingly, impedance of the series-connected electric system is just the sum of the impedances of the individual components. Now consider two mechanical components connected in series. Here the force (effort variable is the through variable) will be the same for both components, and velocity will be additive. Hence, it

ized series
ance or
Generalized
il admittance or

the definition
:al impedance
s force/veloc-
:oss variable),
ıcy domain, is
d voltage are
convenient to

ve a system at

with mechani-
ıted using the
ıents are con-
ll be the same
le) will be ad-
em is just the
wo mechanical
: through vari-
tive. Hence, it

is mobility, not impedance, that is additive in the case of series-connected mechanical components. It can be concluded that, analytically, mobility behaves like electrical impedance and mechanical impedance behaves like electrical admittance. Hence, in Figure 2.5a, the generalized series element Z could be electrical impedance or mechanical mobility, and in figure 2.5b, the generalized parallel element Y could be electrical admittance or mechanical impedance. Definitions of some mechanical transfer functions are given in table 2.2.

TABLE 2.2 DEFINITIONS OF SOME MECHANICAL TRANSFER FUNCTIONS

Transfer function	Definition (in the frequency domain)
Dynamic stiffness	Force/displacement
Dynamic flexibility, compliance, or receptance	Displacement/force
Impedance	Force/velocity
Mobility	Velocity/force
Dynamic inertia	Force/acceleration
Accelerance	Acceleration/force
Force transmissibility	Magnitude [output force/input force]
Velocity transmissibility	Magnitude [output velocity/input velocity]

Example 2.1

Consider a transducer modeled as in figure 2.6. Obtain a transfer-function relationship.

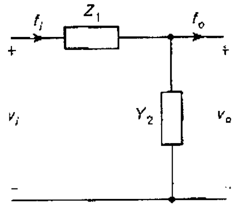


Figure 2.6. An example of a transfer model combination.

Solution This device has a transfer model given by

$$\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} 1 & Z_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ Y_2 & 1 \end{bmatrix} \begin{bmatrix} v_o \\ f_o \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ -Y_2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -Z_1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} v_o \\ f_o \end{bmatrix}$$

which results in the overall model:

$$\begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} 1+Z_1Y_2 & Z_1 \\ Y_2 & 1 \end{bmatrix} \begin{bmatrix} v_o \\ f_o \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & -Z_1 \\ -Y_2 & 1+Z_1Y_2 \end{bmatrix} \begin{bmatrix} v_i \\ f_i \end{bmatrix} = \begin{bmatrix} v_o \\ f_o \end{bmatrix}$$

Notice that when $Y_2 = 0$, equation 2.2 is obtained; and when $Z_1 = 0$, equation 2.3 results.

Example 2.2

The tachometer is a velocity-measuring device (passive) that uses the principle of electromagnetic generation. A DC tachometer is shown schematically in figure 2.7a. The field windings are powered by DC voltage v_f. The across variable at the input port is the measured angular speed ω_i. The corresponding torque T_i is the through variable at the input port. The output voltage v_o of the armature circuit is the across variable at the output port. The corresponding current i_o is the through variable at the output port. Obtain a transfer-function model for this device.

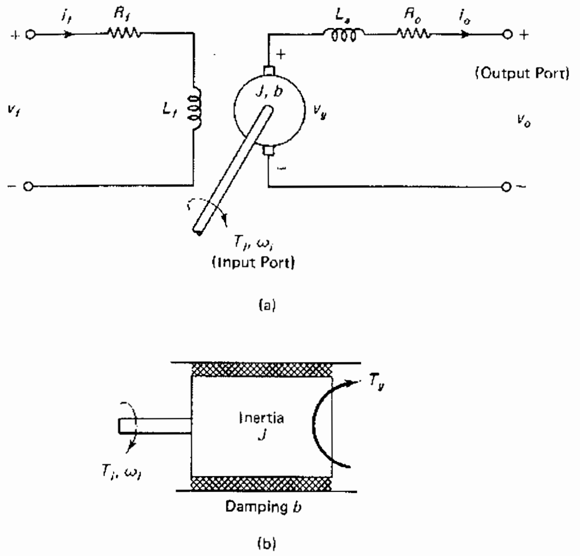


Figure 2.7. A DC tachometer example: (a) equivalent circuit; (b) armature free-body diagram.

Solution The generated voltage v_g at the armature (rotor) is proportional to the magnetic field strength of field windings (which, in turn, is proportional to the field current i_f) and the speed of the armature ω_i. Hence,

$$v_g = K' i_f \omega_i$$

Assuming constant field current, we have

$$v_g = K\omega_i \qquad \text{(i)}$$

The rotor magnetic torque T_g that resists the applied torque T_i is proportional to the magnetic field strengths of field windings and armature windings. Consequently,

$$T_g = K' i_f i_o$$

Since i_f is assumed constant, we get

$$T_g = K i_o \qquad \text{(ii)}$$

ciple of elec-
re 2.7a. The
input port is
h variable at
ariable at the
put port. Ob-

ture

l to the mag-
field current

(i)

tional to the
uently,

(ii)

Note that the same constant K is used in both equations (i) and (ii). This is valid when the same units are used to measure mechanical power and electrical power and no internal dissipation mechanisms are significant in the associated internal coupling. The equation for the armature circuit is

$$v_o = v_g - R_a i_o - L_a \frac{di_o}{dt} \tag{iii}$$

where R_a is the armature resistance and L_a is the *leakage inductance* in the armature circuit. With reference to Figure 2.7b, Newton's second law for a tachometer armature having inertia J and damping constant b is expressed as

$$J \frac{d\omega_i}{dt} = T_i - T_g - b\omega_i \tag{iv}$$

Now equation (i) is substituted into (iii) in order to eliminate v_g. Similarly, equation (ii) is substituted into (iv) in order to eliminate T_g. Next, the time derivatives are replaced by the Laplace variable s. This results in the two algebraic relations:

$$v_o = K\omega_i - (R_a + sL_a)i_o \tag{v}$$

$$(b + sJ)\omega_i = T_i - Ki_o \tag{vi}$$

Note that the variables v_o, i_o, ω_i, and T_i in equations (v) and (vi) are actually Laplace transforms (functions of s), not functions of t, as in equations (i) through (iv). Finally, i_o in equation (v) is eliminated using (vi). This gives the matrix transfer function relation

$$\begin{bmatrix} v_o \\ i_o \end{bmatrix} = \begin{bmatrix} K + (R_a + sL_a)(b + sJ)/K & -(R_a + sL_a)/K \\ -(b + sJ)/K & 1/K \end{bmatrix} \begin{bmatrix} \omega_i \\ T_i \end{bmatrix} \tag{vii}$$

The corresponding frequency domain relations are obtained by replacing s with $j\omega$, where ω represents the angular frequency (radians/second) in the frequency spectrum of a signal.

Even though transducers are more accurately modeled as two-port elements that have two variables associated with each port, it is useful and often essential, for practical reasons, to relate just one input variable and one output variable so that only one transfer function relating these two variables need be specified. This assumes some form of decoupling in the true model. If this assumption does not hold in the range of operation of the transducer, a measurement error would result. For instance, in the tachometer example, we like to express the output voltage v_o in terms of the measured speed ω_i. In this case, the off-diagonal term $-(R_a + sL_a)/K$ in equation (vii) of example 2.2 has to be neglected. This is valid when the tachometer gain parameter K is large and the armature resistance R_a is negligible, since the leakage inductance L_a is negligible in any case for most practical purposes. Note from equations (i) and (ii) that the tachometer gain K can be increased by increasing the field current i_f. This will not be feasible if the field windings are already saturated, however. Furthermore, K can be increased by increasing K'. Now K' depends on parameters such as number of turns and dimensions of the stator windings and magnetic properties of the stator core. Since there is a limitation on the physical size of the tachometer and the types of materials used in the construction, it follows that K cannot be increased arbitrarily. The instrument designer should take such factors into consideration in developing a design that is optimal in many respects. In practical transducers, the operating range is specified in order to minimize the effect of coupling terms, and the residual errors are accounted for by using correction curves. This approach is more convenient than using a coupled model, which introduces three more transfer functions (in general) into the model.

Another desirable feature for practical transducers is to have a static (nondynamic) input/output relationship so that the output instantly reaches the input value (or the measured variable). In this case, the transducer transfer function is a pure gain. This happens when the transducer time constants are small (i.e., the transducer bandwidth is high). Returning to example 2.2 again, it is clear from equation (vii) that static (frequency-independent) transfer-function relations are obtained when the electrical time constant

$$\tau_e = \frac{L_a}{R_a} \tag{2.4}$$

and the mechanical time constant

$$\tau_m = \frac{J}{b} \tag{2.5}$$

are both negligibly small. The electrical time constant is usually an order of magnitude smaller than the mechanical time constant. Hence, one must first concentrate on the mechanical time constant. Note from equation 2.5 that τ_m can be reduced by decreasing rotor inertia and increasing rotor damping. Unfortunately, rotor inertia depends on rotor dimensions, and this determines the gain parameter K, as we saw earlier. Hence, we face some constraint in reducing K. Next, turning to damping, it is intuitively clear that if we increase b, it will require a larger torque T_i to drive the tachometer, and this will load the system that generates the measurand ω_i, possibly affecting the measurand itself. Hence, increasing b also has to be done cautiously. Now, going back to equation (vii), we note that the dynamic terms in the transfer function between ω_i and v_o decrease as K is increased. So we see that increasing K has two benefits: reduction of coupling and reduction of dynamic effects (i.e., increasing the useful frequency range and bandwidth or speed of response).

2.4 PARAMETERS FOR PERFORMANCE SPECIFICATION

A *perfect measuring device* can be defined as one that possesses the following characteristics:

1. Output instantly reaches the measured value (fast response).
2. Transducer output is sufficiently large (high gain or low output impedance).
3. Output remains at the measured value (without drifting or being affected by environmental effects and other undesirable disturbances and noise) unless the measurand itself changes (stability).

take such fac-
ıy respects. In
nize the effect
ing correction
l, which intro-

have a static
ches the input
r function is a
nall (i.e., the
ear from equa-
is are obtained

(2.4)

(2.5)

rder of magni-
rst concentrate
be reduced by
/, rotor inertia
· K, as we saw
to damping, it
T_i to drive the
nd ω_i, possibly
one cautiously.
in the transfer
at increasing K
ffects (i.e., in-
nse).

'ollowing char-

mpedance).
ng affected by
oise) unless the

ing Chap. 2

4. The output signal level of the transducer varies in proportion to the signal level of the measurand (static linearity).
5. Connection of measuring device does not distort the measurand itself (loading effects are absent and impedances are matched).
6. Power consumption is small (high input impedance).

All of these properties are based on dynamic characteristics and therefore can be explained in terms of the dynamic behavior of the measuring device. In particular, items 1 through 4 can be specified in terms of the device (response), either in the *time domain* or in the *frequency domain*. Items 2, 5, and 6 can be specified using the *impedance* characteristics of device. In this section, we shall discuss response characteristics, leaving the discussion of impedance characteristics to section 2.5.

Time Domain Specifications

Figure 2.8 shows a typical *step response* in the dominant mode of a device. Note that the curve is normalized with respect to the steady-state value. We have identified several parameters that are useful for the time domain performance specification of the device. Definitions of these parameters are as follows:

Rise time. This is the time taken to pass the steady-state value of the response for the first time. In overdamped systems, the response is nonoscillatory; consequently, there is no overshoot. So that the definition is valid for all systems, rise time is often defined as the time taken to pass 90 percent of the steady-state value. Rise time is often measured from 10 percent of the steady-state value in order

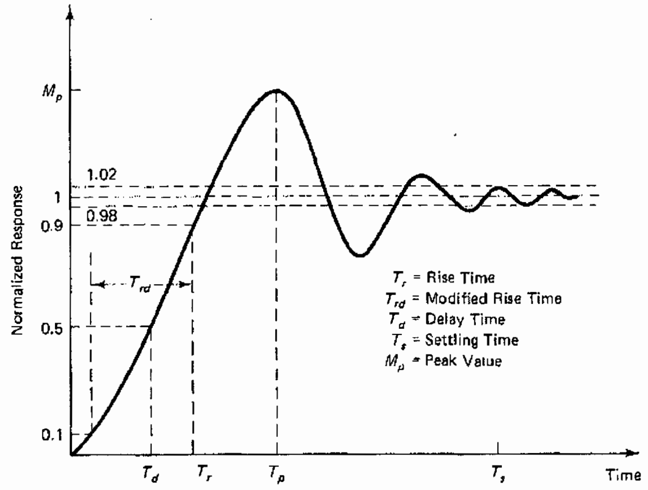


Figure 2.8. Response parameters for the time domain specification of performance.

to leave out start-up irregularities and time lags that might be present in a system. A modified rise time (T_{rd}) may be defined in this manner (see figure 2.8). An alternative definition of rise time, particularly suitable for nonoscillatory responses, is the reciprocal slope of the step response curve at 50 percent of the steady-state value, multiplied by the steady-state value. In process control terminology, this is in fact the *cycle time*. Note that no matter what definition is used, rise time represents the speed of response of a device: a small rise time indicates a fast response.

Delay time. This is usually defined as the time taken to reach 50 percent of the steady-state value for the first time. This parameter is also a measure of speed of response.

Peak time. This is the time at the first peak. This parameter also represents the speed of response of the device.

Settling time. This is the time taken for the device response to settle down within a certain percentage (e.g., ± 2 percent) of the steady-state value. This parameter is related to the degree of damping present in the device as well as degree of stability.

Percentage overshoot (P.O.). This is defined as

$$\text{P.O.} = 100\,(M_p - 1)\,\% \tag{2.6}$$

using the normalized-to-unity step response curve, where M_p is the peak value. Percentage overshoot is a measure of damping or relative stability in the device.

Steady-state error. This is the deviation of the actual steady-state value from the desired value. Steady-state error may be expressed as a percentage with respect to the (desired) steady-state value. In a measuring device, steady-state error manifests itself as an offset. This is a systematic (deterministic) error that normally can be corrected by recalibration. In servo-controlled devices, steady-state error can be reduced by increasing loop gain or by introducing lag compensation. Steady-state error can be completely eliminated using integral control (*reset*) action.

For the best performance of a measuring device, we wish to have the values of all the foregoing parameters as small as possible. In actual practice, however, it might be difficult to meet all specifications, particularly for conflicting requirements. For instance, T_r can be decreased by increasing the dominant natural frequency ω_n of the device. This, however, increases the P.O. and sometimes the T_s. On the other hand, the P.O. and T_s can be decreased by increasing device damping, but it has the undesirable effect of increasing T_r.

Frequency Domain Specifications

Figure 2.9 shows a representative *frequency transfer function* (often termed frequency response function) of a device. This constitutes *gain* and *phase angle* plots, using frequency as the independent variable. This pair of plots is commonly known

.t in a system. A
,8). An alterna-
·esponses, is the
:ady-state value,
y, this is in fact
.e represents the
onse.

ch 50 percent of
.sure of speed of

r also represents

.e to settle down
tate value. This
ıs well as degree

(2.6)

peak value. Per-
ıe device.

eady-state value
centage with re-
teady-state error
or that normally
y-state error can
ıon. Steady-state
ion.

ıve the values of
ice, however, it
ıg requirements.
frequency ω_n of
T_s. On the other
ıg, but it has the

'ten termed fre-
ıase *angle* plots,
ommonly known

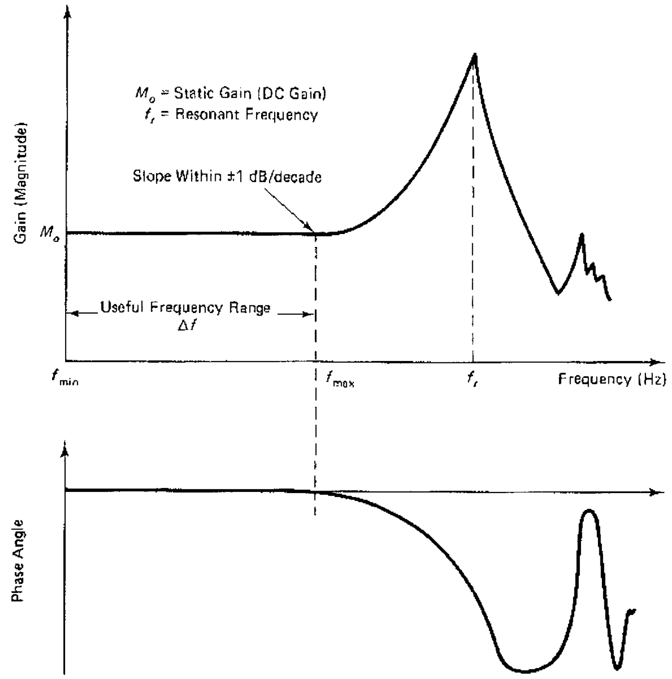


Figure 2.9. Response parameters for the frequency domain specification of performance.

as the *Bode diagram*, particularly when the magnitude axis is calibrated in *decibels* (dB) and the frequency axis in a log scale such as *octaves* or *decades*. Experimental determination of these curves can be accomplished either by applying a harmonic excitation and noting *amplitude gain* and *phase lead* in the response signal at steady state or by Fourier analysis of excitation and response signals for either transient or random excitations. Experimental determination of transfer functions is known as *system identification in the frequency domain*. Note that transfer functions provide complete information concerning system response to a sinusoidal excitation. Since any time signal can be decomposed into sinusoidal components through Fourier transform, it is clear that the response of a system to an arbitrary input excitation also can be determined using transfer-function information for that system. In this sense, transfer functions are frequency domain models that can completely describe linear systems. For this reason, one could argue that it is redundant to use both time domain specifications and frequency domain specifications, as they carry the same information. Often, however, both specifications are used simultaneously, because this can provide a better picture of the system performance. Frequency domain parameters are more suitable in representing some characteristics of a system under some types of excitation.

In the frequency domain, several system parameters have special significance for a measuring instrument:

Useful frequency range. This corresponds to the flat region (static region) in the gain curve and the zero-phase-lead region in the phase curve. It is determined by the dominant resonant frequency f_r of the instrument. The maximum frequency f_{max} in the useful frequency range is several times smaller than f_r for a typical measuring instrument (e.g., $f_{max} = 0.25 f_r$). Useful frequency range may also be determined by specifying the flatness of the static portion of the frequency response curve. For example, since a single pole or a single zero introduces a slope on the order of $\mp$ 20 dB/decade, a slope within 5 percent of this value (i.e., ± 1 dB/decade) may be considered flat for most practical purposes. Operation in the useful frequency range of a measuring device implies measurement of a signal whose significant frequency content is limited to this band. In that case, faithful measurement and fast response are guaranteed, because measuring device dynamics do not corrupt the measurement.

Instrument bandwidth. This is a measure of the useful frequency range of an instrument. Furthermore, the larger the bandwidth, the faster the speed of response of the device will be. Unfortunately, the larger the bandwidth, the more susceptible the instrument will be to high-frequency noise as well as stability problems. Filtering will be needed to eliminate unwanted noise. Stability can be improved by dynamic compensation. There are many definitions for bandwidth. Common definitions include the frequency range over which the transfer-function magnitude is flat, the resonant frequency, and the frequency at which the transfer-function magnitude drops to $1/\sqrt{2}$ (or 70.7 percent) of the zero-frequency (or static) level. The last definition corresponds to the *half-power bandwidth,* because a reduction of amplitude level by a factor of $\sqrt{2}$ corresponds to a power drop by a factor of 2.

Control bandwidth. This is used to specify speed of control. It is an important specification in both analog control and digital control. In digital control, the data sampling rate (in samples/second) has to be at least double the control bandwidth (in hertz) so that the control action can be generated at the full speed. This follows from *Shannon's sampling theorem.* Control bandwidth should be addressed from two points of view. For a system to respond faithfully to a control action (input), the control bandwidth has to be sufficiently small (i.e., input has to be slow enough) in comparison to the dominant (smallest) resonant frequency of the system. This is similar to the bandwidth requirement for measuring devices, mentioned previously. On the other hand, if a certain mode of response in a system is to be insensitive to control action, the control action has to be several times larger than the frequency of that mode. For example, if the bending natural frequency (in the fundamental mode) of a robotic manipulator is 10 Hz, control bandwidth has to be 30 Hz or more so that the robot actuators would not seriously excite that bending mode of the manipulator structure. In digital control, this will require a sampling rate of 60 samples/second or more. In other words, each control cycle in real-time control has to be limited to 1/60 s (approximately 17 ms) or less. Data acquisition

] significance

static region)
is determined
ım frequency
typical mea-
ılso be deter-
ncy response
ıpe on the or-
ɛ., ± 1 dB/
in the useful
signal whose
ıful measure-
amics do not

ency range of
speed of re-
he more sus-
ity problems.
improved by
h. Common
magnitude is
ctinn magni-
evel. The last
ction of am-
r of 2.

It is an im-
gital control,
control band-
l speed. This
be addressed
ontrol action
as to be slow
f the system.
entioned pre-
to be insen-
than the fre-
ency (in the
dth has to be
that bending
e a sampling
in real-time
a acquisition

and processing, including control computations, have to be done within this time. This calls for fast control processors, possibly hardware implementations, and efficient control algorithms.

Static gain. This is the gain (transfer function magnitude) of a measuring instrument within the useful range (or at low frequencies) of the instrument. It is also termed *DC gain*. A high value for static gain results in a high-sensitivity measuring device, which is a desirable characteristic.

Example 2.3

A mechanical device for measuring angular velocity is shown in figure 2.10. The main element of the tachometer is a rotary viscous damper (damping constant b) consisting of two cylinders. The outer cylinder carries a viscous fluid within which the inner cylinder rotates. The inner cylinder is connected to the shaft whose speed ω_i is to be measured. The outer cylinder is resisted by a linear torsional spring of stiffness k. The rotation θ_o of the outer cylinder is indicated by a pointer on a suitably calibrated scale. Neglecting the inertia of moving parts, perform a bandwidth analysis on this device.

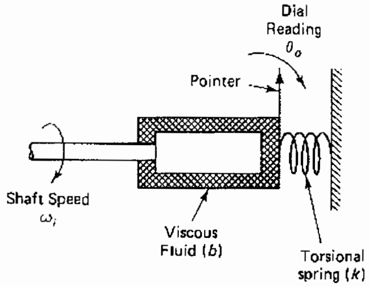


Figure 2.10. A mechanical tachometer.

Solution The damping torque is proportional to the relative velocity of the two cylinders and is resisted by the spring torque. The equation of motion is given by

$$b(\omega_i - \dot{\theta}_o) = k\theta_o$$

or (i)

$$b\dot{\theta}_o + k\theta_o = b\omega_i$$

The transfer function is determined by first replacing the time derivative by the Laplace operator s; thus,

$$\frac{\theta_o}{\omega_i} = \frac{b}{[bs + k]} = \frac{b/k}{[(b/k)s + 1]} = \frac{k_g}{[\tau s + 1]} \qquad \text{(ii)}$$

Note that the static gain or DC gain (transfer-function magnitude with $s = o$) is

$$k_g = \frac{b}{k} \qquad \text{(iii)}$$

and the time constant is

$$\tau = \frac{b}{k} \qquad \text{(iv)}$$

We face conflicting design requirements in this case. On the one hand, we like to have a large static gain so that a sufficiently large reading is available. On the other hand, the time constant must be small in order to obtain a quick reading that faithfully follows the measured speed. A compromise must be reached here, depending on the specific design requirements. Alternatively, a signal-conditioning device could be employed to amplify the sensor output.

Also, let us examine the half-power bandwidth nf the device. The frequency transfer functinn is

$$G(j\omega) = \frac{k_g}{[\tau j\omega + 1]} \tag{v}$$

By definition, the half-power bandwidth ω_b is given by

$$\frac{k_g}{|\tau j\omega_b + 1|} = \frac{k_g}{\sqrt{2}}$$

Hence

$$(\tau\omega_b)^2 + 1 = 2$$

Since bnth τ and ω_b are positive, we have

$$\tau\omega_b = 1$$

or

$$\omega_b = \frac{1}{\tau} \tag{vi}$$

Note that the bandwidth is inversely proportional to the time cnnstant. This confirms our earlier statement that bandwidth is a measure of the speed of respnnse.

Two other system parameters in the frequency dnmain that play crucial roles in interconnected devices are *input impedance* and *output impedance*. Impedance characteristics will be discussed in section 2.5.

Linearity

A device is considered linear if it can be mndeled by linear differrential equations, with time t as the independent variable. Nonlinear devices are often analyzed using linear techniques by considering small excursions about an nperating point. This linearization is accomplished by introducing incremental variables for inputs and outputs. If one increment can cover the entire operating range of a device with sufficient accuracy, it is an indication that the device is linear. If the input/output relations are nonlinear algebraic equations, it represents a *static nonlinearity*. Such a situation can be handled simply by using nonlinear calibration curves, which linearize the device without introducing nonlinearity errors. If, on the other hand, the input/output relations are nonlinear differential equations, analysis usually becomes quite complex. This situation represents a *dynamic nonlinearity*.

Transfer-function representation of an instrument implicitly assumes linearity. According to industrial terminology, a linear measuring instrument provides a mea-

: like to have
her hand, the
ıfully follows
n the specific
employed to

he frequency

(v)

(vi)

This confirms
se.

ucial roles in
edance char-

al equations,
ıalyzed using
int. This lin-
puts and out-
device with
put/output re-
arity. Such a
s, which lin-
her hand, the
ally becomes

nes linearity.
ovides a mea-

sured value that varies linearly with the value of the measurand. This is consistent with the definition of static linearity. All physical devices are *nonlinear* to some degree. This results from any deviation from the ideal behavior due to causes such as saturation, deviation from Hooke's law in elastic elements, coulomb friction, creep at joints, aerodynamic damping, backlash in gears and other loose components, and component wearout.

Nonlinearities in devices are often manifested as some peculiar characteristics. In particular, the following properties are important in detecting nonlinear behavior in dynamic systems:

Saturation. Nonlinear devices may exhibit saturation (see figure 2.11a). This may result from such causes as magnetic saturation, which is common in transformer devices such as differential transformers (see chapter 3), plasticity in mechanical components, or nonlinear deformation in springs.

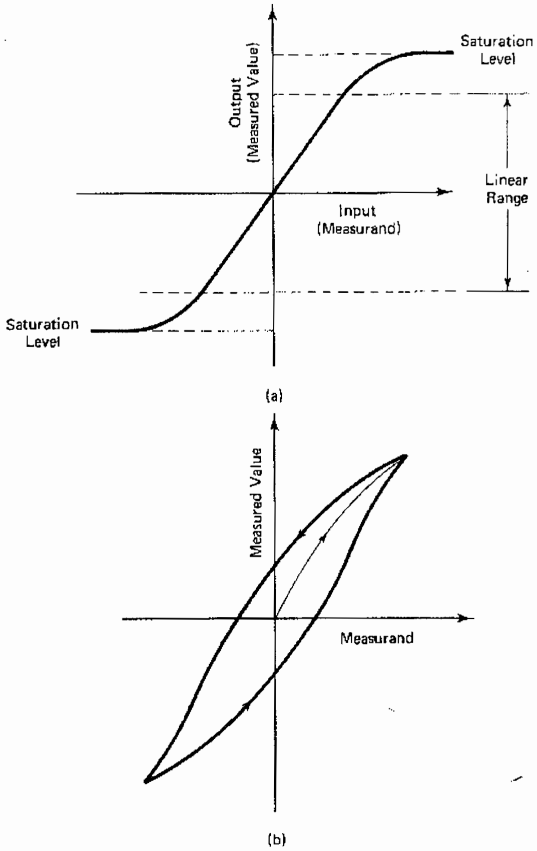
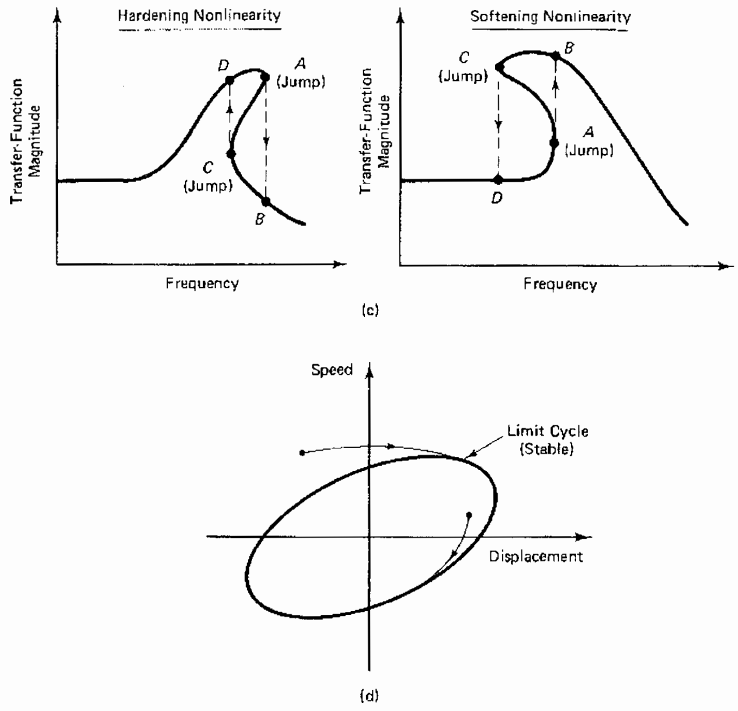


Figure 2.11. Common manifestations of nonlinearity in dynamic systems: (a) saturation; (b) hysteresis; (c) the jump phenomenon; (d) limit cycle response.



(c)

(d)

Figure 2.11. (*continued*)

Hysteresis. Nonlinear devices may produce hysteresis. In this case, the input/output curve changes, depending on the direction of motion (as indicated in figure 2.11b), resulting in a hysteresis loop. This is common in loose components such as gears, which have backlash; in components with nonlinear damping, such as coulomb friction; and in magnetic devices with ferromagnetic media and various dissipative mechanisms (e.g., eddy current dissipation). For example, consider a coil wrapped around a ferromagnetic core. If a DC current is passed through the coil, a magnetic field is generated. As the current is increased from zero, the field strength will also increase. Now, if the current is decreased back to zero, the field strength will not return to zero because of residual magnetism in the ferromagnetic core. A negative current has to be applied to demagnetize the core. It follows that the current-field strength curve looks somewhat like figure 2.11b. This is magnetic hysteresis. Note that linear viscous damping also exhibits a hysteresis loop in the force-displacement curve; this is a property of any mechanical component that dissipates energy. (Area within the hysteresis loop gives the energy dissipated in one cycle of motion.) In general, if force depends on both displacement (as in the case of a spring) and velocity (as in the case of a damping element), the value of force for a

ase, the in-
indicated in
components
ing, such as
and various
nsider a coil
h the coil, a
eld strength
eld strength
etic core. A
hat the cur-
agnetic hys-
in the force-
iat dissipates
one cycle of
ie case of a
f force for a

given value of displacement will change with velocity. In particular, the force when the component is moving in one direction (say, positive velocity) will be different from the force at the same location when the component is moving in the opposite direction (negative velocity), thereby giving a hysteresis loop in the force-displacement plane. If the relationship of displacement and velocity to force is linear (as in viscous damping), the hysteresis effect is linear. If the relationship is nonlinear (as in coulomb damping and aerodynamic damping), however, hysteresis is nonlinear.

The jump phenomenon. Some nonlinear devices exhibit an instability know as the jump phenomenon (or *fold catastrophe*) in the frequency response (transfer) function curve. This is shown in figure 2.11c for both hardening devices and softening devices. With increasing frequency, jump occurs from A to B; and with decreasing frequency, it occurs from C to D. Furthermore, the transfer function itself may change with the level of input excitation in the case of nonlinear devices.

Limit cycles. Nonlinear devices may produce limit cycles. An example is given in figure 2.11d on the phase plane of displacement and velocity. A limit cycle is a closed trajectory in the state space that corresponds to sustained oscillations without decay or growth. Amplitude of these oscillations is independent of the initial location from which the response started. In the case of a stable limit cycle, the response will move onto the limit cycle irrespective of the location in the neighborhood of the limit cycle from which the response was initiated (see figure 2.11d). In the case of an unstable limit cycle, the response will move away from it with the slightest disturbance.

Frequency creation. At steady state, nonlinear devices can create frequencies that are not present in the excitation signals. These frequencies might be harmonics (interger multiples of the excitation frequency), subharmonics (integer fractions of the excitation frequency), nr nonharmonics (usually rational fractions of the excitation frequency).

Example 2.4

Consider a nonlinear device modeled by the differential equation

$$\left\{\frac{dy}{dt}\right\}^{1/2} = u(t)$$

in which $u(t)$ is the input and y is the output. Show that this device creates frequency components that are different from the excitation frequencies.

Solution First, note that the steady-state response is given by

$$y = \int_0^t u^2(t)dt + y(0)$$

Now, for an input given by

$$u(t) = a_1 \sin \omega_1 t + a_2 \sin \omega_2 t$$

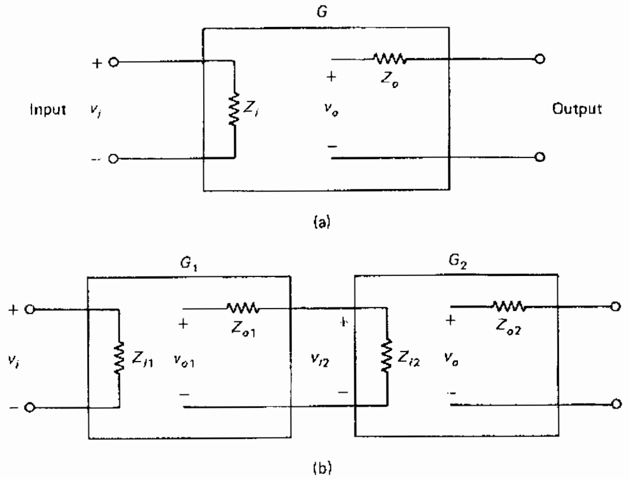


Figure 2.12. (a) Schematic representation of input impedance and output impedance. (b) The influence of cascade connection of devices on the overall impedance characteristics.

Now consider two devices connected in cascade, as shown in figure 2.12b. It can be easily verified that the following relations apply:

$$v_{o1} = G_1 v_i \tag{2.8}$$

$$v_{i2} = \frac{Z_{i2}}{Z_{o1} + Z_{i2}} v_{o1} \tag{2.9}$$

$$v_o = G_2 v_{i2} \tag{2.10}$$

These relations can be combined to give the overall input/output relation:

$$v_o = \frac{Z_{i2}}{Z_{o1} + Z_{i2}} G_2 G_1 v_i \tag{2.11}$$

We see from equation 2.11 that the overall frequency transfer function differs from the ideally expected product ($G_2 G_1$) by the factor

$$\frac{Z_{i2}}{Z_{o1} + Z_{i2}} = \frac{1}{Z_{o1}/Z_{i2} + 1} \tag{2.12}$$

Note that cascading has "distorted" the frequency response characeristics of the two devices. If $Z_{o1}/Z_{i2} << 1$, this deviation becomes insignificant. From this observation, it can be concluded that when frequency response characteristics (i.e., dynamic characteristics) are important in a cascaded device, cascading should be done such that the output impedance of the first device is much smaller than the input impedance of the second device.

utput

nd uutput
he overall

figure 2.12b. It

(2.8)

(2.9)

(2.10)

ɔut relation:

(2.11)

tion differs from

(2.12)

ristics of the two
om this observa-
cs (i.e., dynamic
ıld be done such
· than the input

Example 2.5

A lag network used as the compensatory element of a control system is shown in figure 2.13a. Show that its transfer function is given by

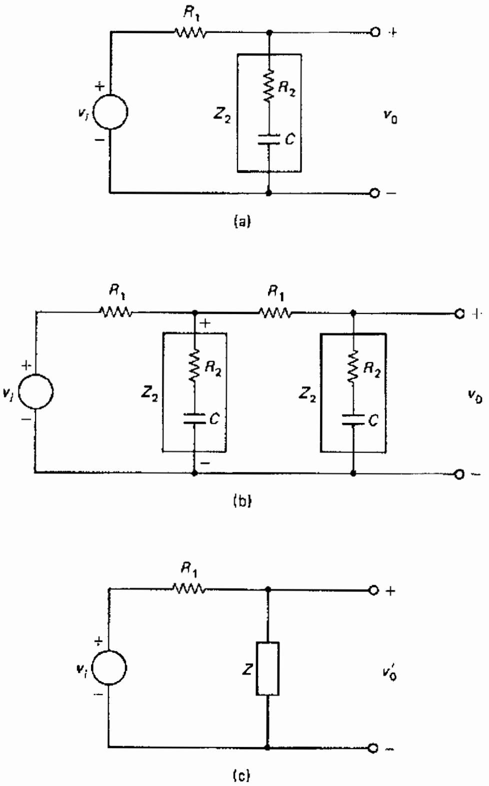
$$\frac{v_o}{v_i} = \frac{Z_2}{R_1 + Z_2}$$

where

$$Z_2 = R_2 + \frac{1}{Cs}$$

What is the input impedance and what is the output impedance for this circuit? Also, if two such lag circuits are cascaded as shown in figure 2.13b, what is the overall transfer function? How would you make this transfer function close to the ideal result:

$$\left\{\frac{Z_2}{R_1 + Z_2}\right\}^2$$



(a)

(b)

(c)

Figure 2.13. (a) Single circuit module. (b) Cascade connection of twu modules. (c) Equivalent circuit for (b).

Solution To solve this problem, first note that in figure 2.13a, voltage drop across the element $R_2 + 1/(Cs)$ is

$$v_o = \left(R_2 + \frac{1}{Cs}\right) \Big/ \left\{R_1 + R_2 + \frac{1}{Cs}\right\} v_i$$

$$= Z_2/(R_1 + Z_2)v_i$$

Hence,

$$\frac{v_o}{v_i} = \frac{Z_2}{R_1 + Z_2} \qquad \text{(i)}$$

Now, input impedance Z_i is derived by

$$\text{input current } i = \frac{v_i}{R_1 + Z_2}$$

$$Z_i = \frac{v_i}{i} = R_1 + Z_2$$

and output impedance Z_o is derived by

$$\text{short-circuit current } i_{sc} = \frac{v_i}{R_1}$$

$$Z_o = \frac{v_o}{i_{sc}} = \frac{Z_2/(R_1 + Z_2)v_i}{v_i/R_1} = \frac{R_1 Z_2}{R_1 + Z_2}$$

Next, consider the equivalent circuit shown in Figure 2.13c. Since Z is formed by connecting Z_2 and $(R_1 + Z_2)$ in parallel, we have

$$\frac{1}{Z} = \frac{1}{Z_2} + \frac{1}{R_1 + Z_2} \qquad \text{(ii)}$$

Voltage drop across Z is

$$v_o' = \frac{Z}{R_1 + Z} v_i \qquad \text{(iii)}$$

Now apply the single-circuit module result (i) to the second circuit stage in figure 2.13b; thus,

$$v_o = \frac{Z_2}{R_1 + Z_2} v_o'$$

Substituting equation (iii), we get

$$v_o = \frac{Z_2}{(R_1 + R_2)} \frac{Z}{(R_1 + Z)} v_i$$

The overall transfer function for the cascaded circuit is

$$G = \frac{v_o}{v_i} = \frac{Z_2}{(R_1 + Z_2)} \frac{Z}{(R_1 + Z)} = \frac{Z_2}{(R_1 + R_2)} \frac{1}{(R_1/Z + 1)}$$

ge drop across the

(i)

$\frac{Z_2}{+ Z_2}$

is formed by con-

(ii)

(iii)

iit stage in figure

$- 1)$

Now substituting equation (ii), we get

$$G = \left[\frac{Z_2}{R_1 + Z_2}\right]^2 \frac{1}{1 + R_1 Z_2/(R_1 + Z_2)^2}$$

We observe that the ideal transfer function is approached by making $R_1 Z_2/(R_1 + Z_2)^2$ small compared to unity.

Impedance-Matching Amplifiers

From the analysis given in the preceding section, it is clear that the signal-conditioning circuitry should have a considerably large input impedance in comparison to the output impedance of the sensor-transducer unit in order to reduce loading errors. The problem is quite serious in measuring devices such as piezoelectric sensors, which have very high output impedances. In such cases, the input impedance of the signal-conditioning unit might be inadequate to reduce loading effects; also, the output signal level of these high-impedance sensors is quite low for signal transmission, processing, and control. The solution for this problem is to introduce several stages of amplifier circuitry between the sensor output and the data acquisition unit input. The first stage is typically an *impedance-matching amplifier* that has very high input impedance, very low output impedance, and almost unity gain. The last stage is typically a stable high-gain amplifier stage to step up the signal level. Impedance-matching amplifiers are, in fact, *operational amplifiers* with feedback.

Operational amplifiers. Operational amplifiers (opamps) are voltage amplifiers with very high gain K (typically 10^5 to 10^9), high input impedance Z_i (typically greater than $1\text{M}\Omega$), and low output impedance Z_o (typically smaller than $100\ \Omega$). Thanks to the advances in integrated circuit technology, opamps—originally made with conventional transistors, diodes, and resistors—are now available as miniature units with integrated circuit elements. Because of their small size, the recent trend is to make signal-conditioning hardware an integral part of the sensor-transducer unit.

A schematic diagram for an opamp is showo in figure 2.14a. Supply voltage v_s is essential to power the opamp. This may be omitted, however, in schematic diagrams and equivalent circuits within the scope of present considerations. In the standard design of opamps, there are two input leads, denoted by 1 and 2 in figure 2.14a. If one of the two input leads is grnunded, it is a *single-ended amplifier*. If neither lead is grounded, it is a *differential amplifier* that requires two input signals. The latter arrangement rejects noise common to the two ioputs (e.g., line noise, thermal noise, magnetic noise) because signal 2 (at the − terminal of the opamp) is subtracted from signal 1 (at the + terminal) and amplified to give the output signal. Figure 2.14a is analogous to figure 2.13a, except that the amplifier gain K has replaced the transfer function $G(j\omega)$. Strictly speaking, K is a transfer function, and it depends on the frequency variable ω of the input signal. Typically, however, the bandwidth of an opamp is on the order of 10kHz; consequently, K may be assumed frequency-independent in that frequency range. This assumption is satisfactory for

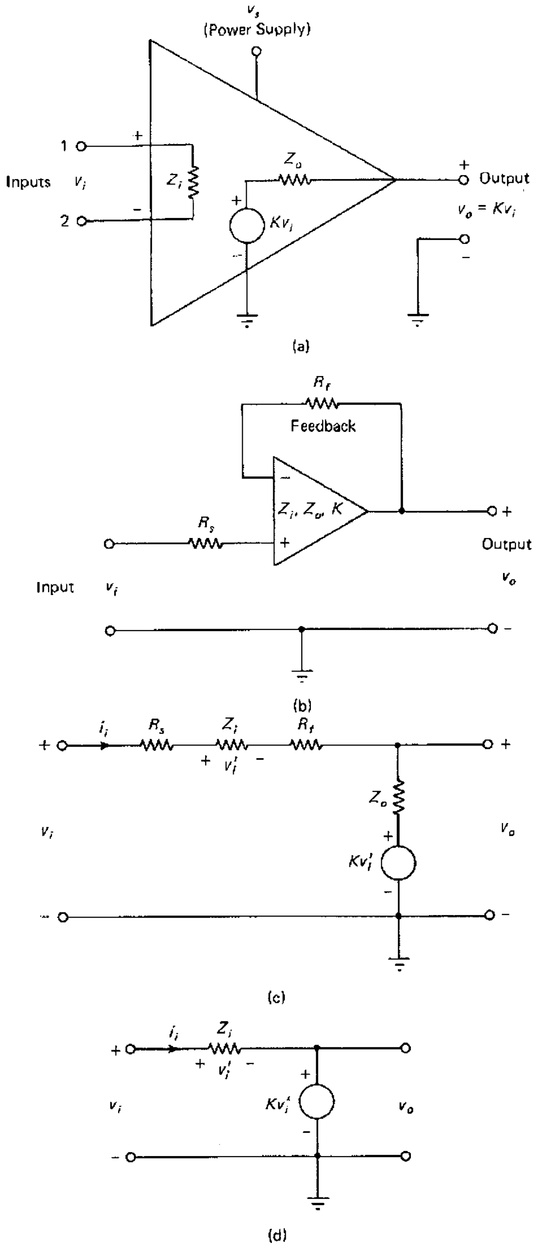
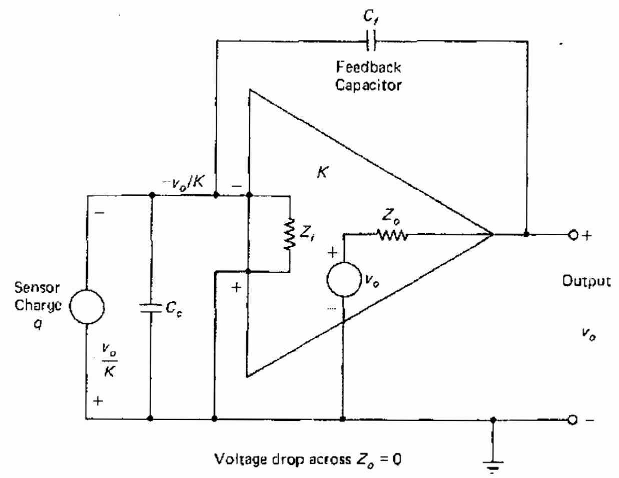


Figure 2.14. Impedance-matching amplifiers: (a) schematic representation of an operational amplifier; (b) schematic representation of a voltage follower; (c) equivalent circuit for the voltage follower; (d) Simplified equivalent circuit for the voltage follower; (e) charge amplifier.

ce-matching
representation
er; (b) schematic
ge follower; (c)
voltage
equivalent circuit
(e) charge



(e)

Figure 2.14. *(continued)*

most practical applications. Nevertheless, an operational amplifier in its basic form has poor stability characteristics; hence, the amplifier output can drift while the input is maintained steady. Furthermore, its gain is too high for direct voltage amplification of practical signals. For these reasons, additional passive elements, such as feedback resistors, are used in conjunction with opamps in practical applications.

Voltage followers. Voltage followers are impedance-matching amplifiers (or *impedance transformers*) with very high input impedance, very low output impedance, and almost-unity gain. For these reasons, they are suitable for use with high output impedance sensors such as piezoelectric devices. A schematic diagram for a voltage follower is shown in figure 2.14b. It consists of a standard (differential) opamp with a feedback resistor R_f connected between the output lead and the negative input lead. The sensor output, which is the amplifier input v_i, is connected to the positive input lead of the opamp with a series resistor R_s. The amplifier output is v_o, as shown. By combining figures 2.14a and 2.14b, the equivalent circuit for the voltage follower is drawn in figure 2.14c. Since the input impedance Z_i of the opamp is much larger than the other impedances (Z_o, R_s, and R_f) in the circuit, the simplified equivalent circuit shown in figure 2.14d is obtained. Note that v_i' is the voltage drop across Z_i.

To obtain an expression for gain $\tilde{K}$ of the voltage follower, examine figure 2.14d. It is clear that

$$v_i = v_i' + Kv_i' = (1 + K)v_i'$$

and

$$v_o = Kv_i'$$

The gain $\tilde{K}$ is given by

$$\frac{v_o}{v_i} = \frac{Kv_i'}{(1 + K)v_i'} = \frac{K}{1 + K}$$

or

$$\tilde{K} = \frac{K}{1 + K} \tag{2.13}$$

which is almost unity for large K.

To determine input impedance $\tilde{Z}_i$ of the voltage follower, first note that the input current i_i is given by

$$i_i = \frac{v_i'}{Z_i} = \frac{v_i}{(1 + K)Z_i}$$

It follows that the input impedance $\tilde{Z}_i$ is given by

$$\frac{v_i}{i_i} = (1 + K)Z_i$$

or

$$\tilde{Z}_i = (1 + K)Z_i \tag{2.14}$$

Since both Z_i and K are very large, it follows that a voltage follower clearly provides a high input impedance. Accordingly, it is able to reduce loading effects of sensors that have high uutput impedances.

To determine the output impedance $\tilde{Z}_o$ of a voltage follower, note that $v_o = 0$ when the output leads are shorted. Then, from figure 2.14c, the short-circuit output current is found to be (by current summation at the output node)

$$i_{sc} = \frac{v_i'}{Z_i} + \frac{Kv_i'}{Z_o}$$

Note that the value of v_i' under short-circuit conditiuns is different from that under open-circuit conditions. But v_i would not be affected by output shorting. When the output leads are shorted, it is clear from figure 2.14d that $v_i' \sim v_i$. Hence,

$$i_{sc} = \left(\frac{1}{Z_i} + \frac{K}{Z_o}\right)v_i$$

The output impedance is

$$\tilde{Z}_o = \frac{v_o}{i_{sc}} = \left(\frac{K}{1 + K}\right)v_i \Big/ \left[\left(\frac{1}{Z_i} + \frac{K}{Z_o}\right)v_i\right]$$

Now, since $Z_i >> Z_o/K$, we can neglect $1/Z_i$ in comparison to K/Z_o. Consequently,

$$\tilde{Z}_o = \frac{Z_o}{1 + K} \tag{2.15}$$

(2.13)

te that the in-

(2.14)

early provides
:cts of sensors

te that $v_o = 0$
t-circuit nutput

rom that under
ing. When the
ence,

Consequently,

(2.15)

Since Z_o is small to begin with and K is very large, it is clear that the output impedance of a vnltage follnwer is very small, as desired. Accordingly, the voltage follower has a unity gain, a very high input impedance, and a very low output impedance; it can be used as an impedance transformer. By connecting a voltage follower to a high-impedance measuring device (sensor-transducer), a low-impedance output signal is obtained. Signal amplification might be necessary befure this signal is transmitted or prncessed, hnwever.

In many data acquisition systems, output impedance of the output amplifier is made equal to the transmission line impedance. When maximum power amplification is desired, *conjugate matching* is recommended. In this case, input impedance and output impedance of the matching amplifier are made equal to the cnmplex conjugates of the source impedance and the load impedance, respectively.

Charge amplifiers. The principle of capacitance feedback is uitilized in charge amplifiers. They are commonly used for conditioning output signals from piezoelectric transducers. A schematic diagram for this device is shown in figure 2.14e. The feedback capacitance is denoted by C_f and the connecting cable capacitance by C_c. The charge amplifier views the sensnr as a charge source (q), even though there is an associated vnltage. Using the fact that charge = voltage × capacitance, a charge balance equation can be written:

$$q + \frac{v_o}{K}C_c + \left(v_o + \frac{v_o}{K}\right)C_f = 0$$

From this, we get

$$v_o = -\frac{K}{(K + 1)C_f + C_c}q \qquad (2.16)$$

If the feedback capacitance is large in comparison with the cable capacitance, the latter can be neglected. This is desirable in practice. In any event, for large values of gain K, we have the approximate relationship

$$v_o = -\frac{q}{C_f} \qquad (2.17)$$

Note that the output voltage is proportional to the charge generated at the sensor and depends only on the feedback parameter C_f. This parameter can be appropriately chosen in order to obtain the required output impedance characteristics. Practical charge amplifiers also have a feedback resistor R_f in parallel with the feedback capacitor C_f. Then the relationship corresponding to equation 2.16 becomes a first-order ordinary differential equation, which in turn determines the time constant of the charge amplifier. This time constant shnuld be high. If it is low, the charge generated by the piezoelectric sensor will leak out quickly, giving erroneous results at low frequencies (see problem 2.11).

Example 2.6

Suppose that the output signal from a sensor with output impedance Z_s is directly connected to an operational amplifier with gain K, input impedance Z_i, and output impedance Z_o. The resulting signal is read directly into a digital cnntroller using an analog-tn-digital conversinn (ADC) board with an equivalent load impedance Z_L. Pick parameters so as to reduce possible distortion of the digitized signal due to loading.

Solution A schematic representation of this arrangement of data acquisitinn is shnwn in figure 2.15. Straightforward analysis provides the following input/output relationship:

$$v_o = K\left(\frac{Z_i}{Z_s + Z_i}\right)\left(\frac{Z_L}{Z_o + Z_L}\right) v_i$$

If the input impedance Z_i of the opamp is very high in comparison with the sensor impedance Z_s, then $Z_i/(Z_s + Z_i)$ will approach unity. Furthermnre, if the load impedance Z_L is very high in comparison with the output impedance of the opamp, then $Z_L/(Z_o + Z_L)$ will also approach unity. In that case, the input/output relation reduces to

$$v_o = Kv_i$$

which corresponds to a simple amplification of measured voltage by the gain factor K. In practice, however, the parameter K may drift because of such reasons as bandwidth limitations and stability problems in the opamp. Hence, using an opamp is not the best way to achieve signal amplification.

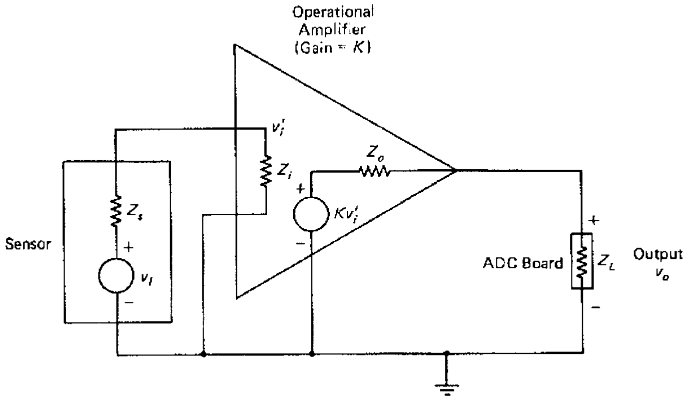


Figure 2.15. A data acquisition system example.

Measurement of Across Variables and Through Variables

Impedance concepts are very useful in selecting (and designing) instruments to measure across variables (voltage, velocity, pressure, temperature) and through variables (current, force, fluid flow rate, heat transfer rate). To develop some general concepts relating to the measurement of across variables and through variables, consider a

is directly con-
Z_i, and output
troller using an
edance Z_L. Pick
ıe to loading.

isition is shown
output relation-

with the sensor
'e, if the load
of the opamp,
tput relation re-

e gain factor K.
ns as bandwidth
p is not the best

Output
v_o

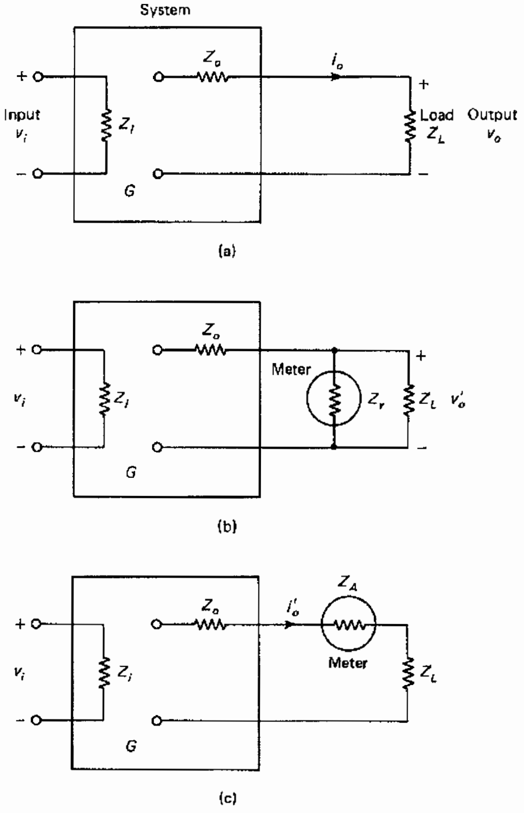
ıments to mea-
'ough variables
ıneral concepts
les, consider a

device (an electronic device such as an amplifier, filter, or control circuit or a mechanical, fluid, or thermal device) that has input impedance Z_i and output impedance Z_o. It is connected to a load of impedance Z_L. This device is shown schematically in figure 2.16a. Variable v_o across the load and variable i_o through the load are to be measured. It is seen that

$$v_o = Gv_i\frac{Z_L}{Z_o + Z_L}$$

or

$$v_o = \frac{G}{[Z_o/Z_L + 1]}v_i \tag{2.18}$$



(a)

(b)

(c)

Figure 2.16. Impedance of measuring instruments: (a) system representation; (b) measurement of an across variable; (c) measurement of a through variable.

and

$$i_o = \frac{G}{Z_o + Z_L} v_i \tag{2.19}$$

where G is the system (frequency) transfer function. In writing these relationships, we must remember that in the case of mechanical devices, the generalized impedance Z should be interpreted as mechanical mobility (velocity/force), not mechanical impedance (force/velocity). Otherwise, the combination rules used in getting these relationships (i.e., impedances additive in series and admittances additive in parallel) would not be valid.

Suppose that a meter of impedance Z_V is connected across the load to measure v_o, as shown in figure 2.16b. Since Z_V and Z_L are in parallel, their equivalent impedance Z is given by

$$\frac{1}{Z} = \frac{1}{Z_v} + \frac{1}{Z_L} \tag{2.20}$$

Again, for this relationship to be generally valid, impedance should be interpreted as (across variable)/(through variable), not as (effort variable)/(flow variable). This interpretation, however, contradicts the commonly used definition of mechanical impedance—force/velocity in the frequency domain. Alternatively, Z should be interpreted as "mobility" in mechanical systems. No such ambiguity exists in electrical, fluid, and thermal systems.

Due to loading effects from the meter, across variable v_o changes to v_o', and

$$v_o' = Gv_i \frac{Z}{Z_o + Z} = \frac{Gv_i}{Z_o/Z + 1}$$

In view of equation 2.20, we have

$$v_o' = \frac{Gv_i}{Z_o/Z_V + Z_o/Z_L + 1} \tag{2.21}$$

By comparing equation 2.21 with equation 2.18, we observe that for high accuracy of measurement (i.e., v_o' nearly equal to v_o), we must have either $Z_o/Z_V << 1$ or $Z_o/Z_V << Z_o/Z_L$. In other words, we must have $Z_V >> Z_o$ or $Z_V >> Z_L$. Hence, in general, a measuring instrument for an across variable must have a high impedance.

Now suppose that a meter of impedance Z_A is connected in series with the load to measure i_o, as shown in figure 2.16c. Because of instrument loading, the through variable i_o changes to i_o', and

$$i_o' = \frac{Gv_i}{Z_o + Z_L + Z_A} \tag{2.22}$$

By comparing equation 2.22 with equation 2.19, we note that for high accuracy (i.e., i_o' almost equal to i_o), we must have either $Z_A << Z_o$ or $Z_A << Z_L$. It follows that, in general, an instrument measuring a through variable has to be a low-impedance device.

(2.19)

; relationships, ıe generalized 'orce), not me- es used in get- tances additive

oad to measure heir equivalent

(2.20)

; interpreted as iable). This in- of mechanical Z should be in- xists in electri-

ges to v_o', and

(2.21)

r high accuracy $Z_o/Z_V << 1$ or $> Z_L$. Hence, in igh impedance. es with the load ng, the through

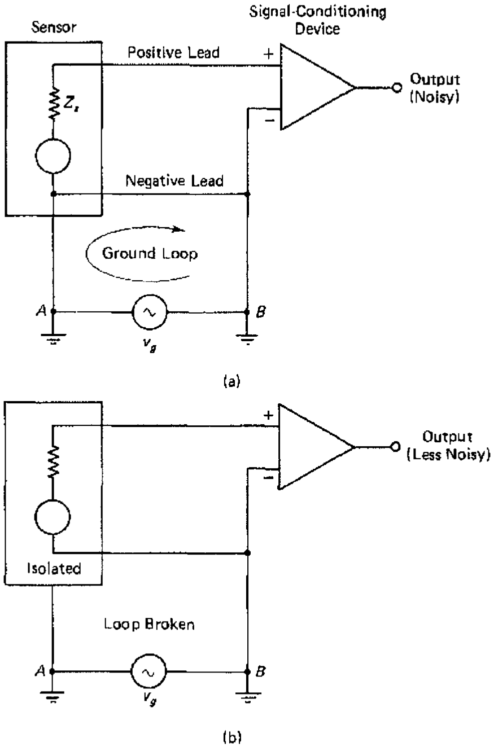
(2.22)

: high accuracy $< Z_L$. It follows s to be a low-

Ground Loop Noise

In devices that handle low-level signals (e.g., accelerometers and strain gage bridge circuitry), electrical noise can create excessive error. One form of noise is caused by fluctuating magnetic fields due to nearby AC lines. This can be avoided either by taking precautions not to have strong magnetic fields and fluctuating currents near delicate instruments or by using *fiber optic* (optically coupled) signal transmission (see chapter 3). Furthermore, if the two signal leads (positive and negative) are twisted or if shielded cables are used, the induced noise voltages become equal in the two leads, which cancel each other. Another cause of electrical noise is ground loops.

If two interconnected devices are grounded at two separate locations, ground loop noise can enter the signal leads because of the possible potential difference between the two ground points. The reason is that ground itself is not generally a uniform potential medium, and a nonzero (and finite) impedance may exist from point to point within the ground medium. This is, in fact, the case with typical ground media, such as instrument housings and common ground wire. An example is shown schematically in figure 2.17a. In this example, the two leads of a sensor are directly



(a)

(b)

Figure 2.17. (a) Illustration of a ground loop. (b) Device isolation to eliminate ground loops (an example of internal isolation).

connected to a signal-conditioning device such as an amplifier. Because of nonuniform ground potentials, the two ground points A and B are subjected to a potential difference v_g. This will create a ground loop with the common negative lead of the two interconnected devices. The solution to this problem is to isolate (i.e., provide an infinite impedance to) either one of the two devices. Figure 2.17b shows internal isolation of the sensor. External isolation, by insulating the casing, is also acceptable. Floating off the power supply ground will also help eliminate ground loops.

2.6 INSTRUMENT RATINGS

Instrument manufacturers do not usually provide complete dynamic information for their products. In most cases, it is unrealistic to expect complete dynamic models (in the time domain or the frequency domain) and associated parameter values for complex instruments. Performance characteristics provided by manufacturers and vendors are primarily static parameters. Known as instrument ratings, these are available as parameter values, tables, charts, calibration curves, and empirical equations. Dynamic characteristics such as transfer functions (e.g., transmissibility curves expressed with respect to excitation frequency) might also be provided for more sophisticated instruments, but the available dynamic information is never complete. Furthermore, definitions of rating parameters used by manufacturers and vendors of instruments are in some cases not the same as analytical definitions used in textbooks on dynamic systems and control. This is particularly true in relation to the term *linearity*. Nevertheless, instrument ratings provided by manufacturers and vendors are very useful in the selection, installation, operation, and maintenance of instruments. In this section, we shall examine some of these performance parameters.

Rating Parameters

Typical rating parameters supplied hy instrument manufacturers are

1. Sensitivity
2. Dynamic range
3. Resolution
4. Linearity
5. Zero drift and full-scale drift
6. Useful frequency range
7. Bandwidth
8. Input and output impedances

We have already discussed the meaning and significance of some of these terms with respect to dynamic behavior of instruments. In this section, we shall look at the conventional definitions given by instrument manufacturers and vendors.

se of nonuni-
to a potential
re lead of the
(i.e., provide
nows internal
s also accept-
und loops.

formation for
ic models (in
lues for com-
ers and ven-
ese are avail-
cal equations.
ty curves ex-
for more so-
ver complete.
nd vendors of
used in text-
elation to the
rers and ven-
enance of in-
e parameters.

se terms with
ok at the con-

Sensitivity of a transducer is measured by the magnitude (peak, rms value, etc.) of the output signal corresponding to a unit input of the measurand. This may be expressed as the ratio of (incremental output)/(incremental input) or, analytically, as the corresponding partial derivative. In the case of vectorial or tensorial signals (e.g., displacement, velocity, acceleration, strain, force), the direction of sensitivity should be specified. Cross-sensitivity is the sensitivity along directions that are orthogonal to the direction of sensitivity; it is expressed as a percentage of direct sensitivity. High sensitivity and low cross-sensitivity are desirable for measuring instruments. Sensitivity to parameter changes and noise has to be small in any device, however. On the other hand, in *adaptive control*, system sensitivity to control parameters has to be sufficiently high. Often, sensitivity and robustness are conflicting requirements.

Dynamic range of an instrument is determined by the allowed lower and upper limits of its input or output (response) so as to maintain a required level of measurement accuracy. This range is usually expressed as a ratio, in *decibels*. In many situations, the lower limit of dynamic range is equal to the resolution of the device. Hence, the dynamic range ratio is usually expressed as (range of operation)/(resolution).

Resolution is the smallest change in a signal that can be detected and accurately indicated by a transducer, a display unit, or any pertinent instrument. It is usually expressed as a percentage of the maximum range of the instrument or as the inverse of the dynamic range ratio. It follows that dynamic range and resolution are very closely related.

Example 2.7

The meaning of dynamic range (and resolution) can easily be extended to cover digital instruments. For example, consider an instrument that has a 12-bit analog-to-digital converter (ADC). Estimate the dynamic range of the instrument.

Solution In this example, dynamic range is determined (primarily) by the word size of the ADC. Each bit can take the binary value 0 or 1. Since the resolution is given by the smallest possible increment, a change by the least significant bit (LSB),

$$\text{digital resolution} = 1$$

The largest value represented by a 12-bit word corresponds to the case when all twelve bits are unity. This value is decimal $2^{12} - 1$. The smallest value (when all twelve bits are zero) is zero. Hence, using the definition

$$\text{dynamic range} = 20 \log_{10}\left[\frac{\text{range of operation}}{\text{resolution}}\right] \tag{2.23}$$

the dynamic range of the instrument is given by

$$20 \log_{10}\left[\frac{2^{12} - 1}{1}\right] = 72 \text{ dB}$$

Another (perhaps more correct) way of looking at this problem is to consider the resolution to be some value δy, rather than unity, depending on the particular application.

For example, δy may represent an output signal increment of 0.0025 V. Next, we note that a 12-bit word can represent a combination of 2^{12} values (i.e., 4,096 values), the smallest value being y_{min} and the largest value being

$$y_{max} = y_{min} + (2^{12} - 1)\,\delta y$$

Note that y_{min} can be zero, positive, or negative. The smallest increment between values is δy, which is, by definition, the resolution. There are 2^{12} values within y_{min} and y_{max}, the two end values inclusive. Then

$$\text{dynamic range} = \frac{y_{max} - y_{min}}{\delta y} = \frac{(2^{12} - 1)\delta y}{\delta y} = 12^{12} - 1 = 4{,}095 = 72 \text{ dB}$$

So we end up with the same result for dynamic range, but the interpretation of resolution is different.

Linearity is determined by the calibration curve of an instrument. The curve uf output amplitude (peak nr rms value) versus input amplitude under static conditions within the dynamic range of an instrument is known as the *static calibration curve.* Its closeness to a straight line measures the degree of linearity. Manufacturers provide this information either as the maximum deviation of the calibration curve from the least squares straight-line fit of the calibration curve or from some other reference straight line. If the least squares fit is used as the reference straight line, the maximum deviation is called *independent linearity* (more correctly, independent nonlinearity, because the larger the deviation, the greater the nonlinearity). Nonlinearity may be expressed as a percentage of either the actual reading at an operating point or the full-scale reading.

Zero drift is defined as the drift from the null reading of the instrument when the measurand is maintained steady for a long period. Note that in this case, the measurand is kept at zero or any other level that corresponds to null reading of the instrument. Similarly, *full-scale drift* is defined with respect to the full-scale reading (the measurand is maintained at the full-scale value). Usual causes of drift include instrument instability (e.g., instability in amplifiers), ambient changes (e.g., changes in temperature, pressure, humidity, and vibration level), changes in power supply (e.g., changes in reference DC voltage or AC line voltage), and parameter changes in an instrument (due to aging, wearout, nonlinearities, etc.). Drift due to parameter changes that are caused by instrument nonlinearities is known as *parametric drift*, *sensitivity drift*, or *scale-factor drift*. For example, a change in spring stiffness or electrical resistance due to changes in ambient temperature results in a parametric drift. Note that parametric drift depends on the measurand level. Zeru drift, however, is assumed to be the same at any measurand level if the other conditions are kept constant. For example, a change in reading caused by thermal expansion of the readout mechanism due to changes in ambient temperature is considered a zero drift. In electronic devices, drift can be reduced by using alternating current (AC) circuitry rather than direct current (DC) circuitry. For example, AC-coupled amplifiers have fewer drift problems than DC amplifiers. Intermittent checking for instrument response level with zero input is a popular way to calibrate for zero drift. In digital devices, for example, this can be done automatically from time to time be-

'ext, we note
values), the

between val-
thin y_{min} and

: 72 dB

on of resolu-

'he curve of
: conditions
tion curve.
nufacturers
ation curve
e other ref-
ht line, the
ndependent
y). Nonlin-
n operating

ment when
s case, the
ding of the
ale reading
rift include
ges (e.g.,
s in power
l parameter
Drift due to
as *paramet-*
spring stiff-
s in a para-
Zero drift,
conditions
xpansion of
ered a zero
rrent (AC)
oupled am-
king for in-
ero drift. In
to time be-

tween sample points, when the input signal can be bypassed without affecting the system operation.

Useful frequency range correspnnds to a flat gain curve and a zero phase curve in the frequency response characteristics of an instrument. The maximum frequency in this band is typically less than half (say, one-fifth) of the dominant resonant frequency of the instrument. This is a measure of instrument bandwidth.

Bandwidth of an instrument determines the maximum speed or frequency at which the instrument is capable of operating. High bandwidth implies faster speed of response. Bandwidth is determined by the dominant natural frequency ω_n or the dominant resonant frequency ω_r of the transducer. (Note: For low damping, ω_r is approximately equal to ω_n.) It is inversely proportional to rise time and the dominant time constant. Half-power bandwidth (defined earlier) is also a useful parameter. Instrument bandwidth has to be several times greater than the maximum frequency of interest in the measured signal. Bandwidth of a measuring device is impnrtant, particularly when measuring transient signals. Note that bandwidth is directly related to the useful frequency range.

Accuracy and Precision

The instrument ratings mentioned in the preceding section affect the overall *accuracy* of an instrument. Accuracy can be assigned either to a particular reading or to an instrument. Note that instrument accuracy depends not only nn the physical hardware of the instrument but also on the operating environment, including arbitrary factors such as the practices of a particular user. Usually, instrument accuracy is given with respect to a standard set of operating conditions (e.g., design conditions that are the normal steady operating conditions or extreme and transient conditions, such as emergency start-up and shutdown). *Meusurement accuracy* determines the closeness of the measured value to true value. *Instrument uccuracy* is related to the worst accuracy obtainable within the dynamic range of the instrument in a specific operating environment. *Measurement error* is defined as

$$\text{error} = (\text{measured value}) - (\text{true value})$$

Correction, which is the negative of error, is defined as

$$\text{correction} = (\text{true value}) - (\text{measured value})$$

Each of these can also be expressed as a percentage of the true value. Accuracy of an instrument may be determined by measuring a parameter whose true value is known, near the extremes of the dynamic range of instrument, under certain operating conditions. For this purpose, standard parameters or signals that can be generated at very high levels of accuracy would be needed. The National Bureau of Standards (NBS) is usually responsible for generation of these standards. Nevertheless, accuracy and error values cannot be determined to 100 percent exactness in typical applications, because the true value is not known to begin with. In a given situation, we can only make estimates for accuracy, by using ratings prnvided by the instrument manufacturer or by analyzing data from previous measurements and models.

Causes of error include instrument instability, external noise (disturbances), poor calibration, inaccurate information (e.g., poor analytical models, inaccurate control laws and digital control algorithms), parameter changes (e.g., due to environmental changes, aging, and wearout), unknown nonlinearities, and improper use of instrument.

Errors can be classified as *deterministic* (or *systematic*) and *random* (or *stochastic*). Deterministic errors are those caused by well-defined factors, including nonlinearities and offsets in readings. These usually can be accounted for by proper calibration and analysis practices. Error ratings and calibration charts are used to remove systematic errors from instrument readings. Random errors are caused by uncertain factors entering into instrument response. These include device noise, line noise, and effects of unknown random variations in the operating environment. A statistical analysis using sufficiently large amounts of data is necessary to estimate random errors. The results are usually expressed as a mean error, which is the systematic part of random error, and a standard deviation or confidence interval for instrument response. These concepts will be addressed in sectinn 2.7.

Precision is not synonymous with accuracy. Reproducibility (or repeatability) of an instrument reading determines the precision of an instrument. Two or more identical instruments that have the same high offset error might be able to generate responses at high precision, even though these readings are clearly inaccurate. For example, consider a timing device (clock) that very accurately indicates time increments (say, up to the nearest microsecond). If the reference time (starting time) is set incorrectly, the time readings will be in error, even though the clock has very high precision.

Instrument error may be represented by a random variable that has a mean value μ_e and a standard deviation σ_e. If the standard deviation is zero, the variable is deterministic. In that case, the error is said to be deterministic or repeatable. Otherwise, the error is said to be random. The precision of an instrument is determined by the standard deviation of error in the instrument response. Readings of an instrument may have a large mean value of error (e.g., large offset), but if the standard deviation is small, the instrument has high precision. Hence, a quantitative definition for precision would be

$$\text{precision} = (\text{measurement range})/\sigma_e \tag{2.24}$$

Lack of precision originates from random causes and poor construction practices. It cannot be compensated for by recalibration, just as precision of a clock cannot be improved by resetting the time. On the other hand, accuracy can be improved by recalibration. Repeatable (deterministic) accuracy is inversely proportional to the magnitude of the mean error μ_e.

In selecting instruments for a particular application, in addition to matching instrument ratings with specifications, several additional considerations should be looked into. These incude geometric limitations (size, shape, etc.), environmental conditions (e.g., chemical reactions including corrosion, extreme temperatures, light, dirt accumulation, electromagnetic fields, radioactive environments, and shock and vibration), power requirements, operational simplicity, availability, past record and reputation of the manufacturer and of the particular instrument, and cost

sturbances),
, inaccurate
lue to envi-
nproper use

random (or
s, including
or by proper
; used to re-
used by un-
: noise, line
ronment. A
to estimate
h is the sys-
erval for in-

epeatability)
wo or more
to generate
ccurate. For
time incre-
ing time) is
ck has very

has a mean
e variable is
ible. Other-
determined
of an instru-
he standard
quantitative

(2.24)

uction prac-
a clock can-
be improved
tional to the

natching in-
s should be
vironmental
mperatures,
ments, and
ability, past
nt, and cost

and related economic aspects (initial cost, maintenance cost, cost of supplementary components such as signal-conditioning and processing devices, design life and associated frequency of replacement, and cost of disposal and replacement). Often, these considerations becnme the ultimate deciding factors in the selection process.

2.7 ERROR ANALYSIS

Analysis of error is a very challenging task. Difficulties arise for many reasons, particularly the following:

1. True value is usually unknown.
2. The instrument reading may contain random error that cannot be determined exactly.
3. The error may be a complex (not simple) function of many variables (input variables and state variables or response variables).
4. The instrument may be made up of many components that have complex interrelations (dynamic coupling, multiple degree-of-freedom responses, nonlinearities, etc.), and each component may contribute to the overall error.

The first item is a philosophical issue that would lead to an argument similar to the chicken-and-egg controversy. For instance, if the true value is known, there is no need to measure it; and if the true value is unknown, it is impossible to determine exactly how inaccurate a particular reading is. In fact, this situation can be addressed to some extent by using statistical representations of error, which takes us to the second item listed. The third and fourth items may be addressed by error combination in multivariable systems and by error propagation in complex multicomponent systems. It is not feasible here to provide a full treatment of all these topics. Only an introduction to available analytical techniques will be given, using illustrative examples.

The concepts discussed in this section are useful not only in statistical error analysis but also in the field of *statistical process control* (SPC)—the use of statistical signals to improve performance of a process. Performing statistical analysis of a response signal and drawing its *control chart,* along with an *upper control line* and a *lower control line,* are key procedures in statistical process control.

Statistical Representation

We have noted that, in general, error is a random variable. It is defined as

$$\text{error} = (\text{instrument reading}) - (\text{true value}).$$

Randomness associated with a measurand can be interpreted in two ways. First, since the true value of the measurand is a fixed quantity, randomness can be interpreted as the randomness in error that is usually originating from the random factors in instrument response. Second, looking at the issue in a more practical manner, error analysis can be interpreted as an "estimation problem" in which the objective is to estimate the true value of a measurand from a known set of readings. In this latter point of view, "estimated" true value itself becomes a random variable. No matter what approach is used, however, the same statistical concepts may be used in representing error. First, let us review some important concepts in probability and statistics.

Cumulative probability distribution function. Consider a random variable X. The probability that the random variable takes a value equal to or less than a specific value x is a function of x. This function, denoted by $F(x)$, is termed *cumulative probability distribution function,* or simply *distribution functon*. Specifically,

$$F(x) = P[X \leq x] \tag{2.25}$$

Note that $F(\infty) = 1$ and $F(-\infty) = 0$, because the value of X is always less than infinity and can never be less than negative infinity. Furthermore, $F(x)$ has to be a monotonically increasing function, as shown in figure 2.18a, because negative probabilities are oot defined.

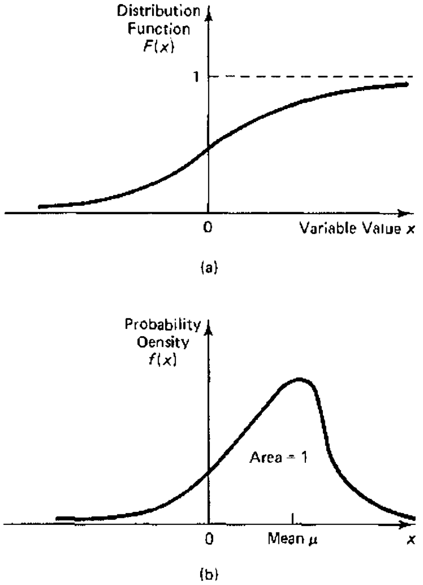


Figure 2.18. (a) A cumulative probability distribution function. (b) A probability density function.

Probability density function. Assuming that random variable X is a continuous variable and, hence, $F(x)$ is a continuous function of x, probability density function $f(x)$ is given by the slope of $F(x)$, as shown in figure 2.18b. Thus,

$$f(x) = \frac{dF(x)}{dx} \tag{2.26}$$

ich the objec-
f readings. In
n variable. No
nay be used in
robability and

random vari-
or less than a
termed *cumu-*
Specifically,

(2.25)

is always less
c, $F(x)$ has to
cause negative

ımulative
function. (b) A
:tion.

əle X is a con-
əability density
. Thus,

(2.26)

Hence,

$$F(x) = \int_{\infty}^{x} f(x)dx \tag{2.27}$$

Note that the area under the density curve is unity. Furthermore, the probability that the random variable falls within two values is given by the area under the density curve within these two limits. This can be easily shown using the definitions of $F(x)$ and $f(x)$:

$$P[a < X \leq b] = F(b) - F(a)$$

$$= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx \tag{2.28}$$

Mean value (expected value). If a random variable X is measured repeatedly a very large (infinite) number of times, the average of these measurements is the *mean value* μ or *expected value* $E(X)$. It should be easy to see that this may be expressed as the weighted sum nf all possible values of the random variable, each value being weighted by the associated probability of its occurrence. Since the probability that X takes the value x is given by $f(x)\ \delta x$, with δx apprnaching zero, we have

$$\mu = E(X) = \lim_{\delta x \to 0} \sum x f(x) \delta x$$

Since the right-hand-side summation becomes an integral in the limit, we get

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx \tag{2.29}$$

Root-mean-square (rms) value. The mean square value of a random variable X is given by

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx \tag{2.30}$$

The root-mean-square (rms) value is the square root of the mean square value.

Variance and standard deviation. Variance of a random variable is the mean square value nf the deviation from mean. This is denoted by Var(X) or σ^2 and is given by

$$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \tag{2.31}$$

By expanding equation 2.31, we can show that

$$\sigma^2 = E(X^2) - \mu^2 \tag{2.32}$$

Standard deviation σ is the square root of variance. Note that standard deviation is a measure of statistical "spread" of a random variable. A random variable with smaller σ is less random and its density curve exhibits a sharper peak, as shown in figure 2.19.

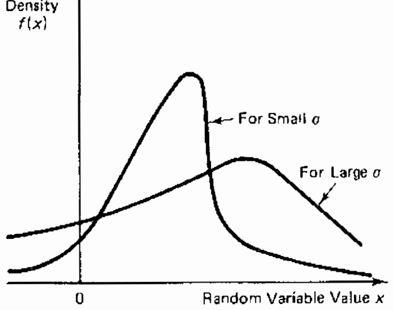


Figure 2.19. Effect of standard deviation on the shape of a probability density curve.

Some thinking should convince you that if the probability density function of random variable X is $f(x)$, then the probability density function of any (well-behaved) function of X is also $f(x)$. In particular, for constants a and b, the probability density function of $(aX + b)$ is also $f(x)$. Note, further, that the mean of $(aX + b)$ is $(a\mu + b)$. Hence, from equation 2.31, it follows that the variance of aX is

$$\mathrm{Var}(aX) = \int_{-\infty}^{\infty} (ax - a\mu)^2 f(x)dx$$

$$= a^2 \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

Hence,

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X) \tag{2.33}$$

Independent random variables. Two random variables, X_1 and X_2, are said to be independent if the event "X_1 assumes a certain value" is completely independent of the event "X_2 assumes a certain value." In other words, the processes that generate the responses X_1 and X_2 are completely independent. Furthermore, probability distributions of X_1 and X_2 will also be completely independent. Hence, it can be shown that for independent random variables X_1 and X_2, the mean value of the product is equal to the product of the mean values. Thus,

$$E(X_1 X_2) = E(X_1)E(X_2) \tag{2.34}$$

for independent random variables X_1 and X_2.

Now, using the definition of variance and equation 2.34, it can be shown that

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) \tag{2.35}$$

for independent X_1 and X_2.

Sample mean and sample variance. Consider N measurements $\{X_1, X_2, \ldots, X_N\}$ of random variable X. This set of data is termed a *data sample*. It generally is not possible to extract all information about the probability distribution of X from this data sample. We are able, however, to make some useful *estimates*. One would expect that the larger the data sample, the more accurate these statistical estimates would be.

An estimate for the mean value of X would be the *sample mean* $\bar{X}$, which is defined as

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.36}$$

An estimate for variance would be the *sample variance* S^2, given by

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N} (X_i - \bar{X})^2 \tag{2.37}$$

An estimate for standard deviation would be the *sample standard deviation*, S, which is the square root of the sample variance.

One might be puzzled by the denominator $N - 1$ on the right-hand side of equation 2.37. Since we are computing an "average" deviation, the denominator should have been N. But in that case, with just one reading ($N = 1$), we get a finite value for S, which is not correct because one cannot talk about a sample standard deviation when only one measurement is available. Since, according to equation 2.37, S is not defined (0/0) when $N = 1$, this definition of S^2 is more realistic. Another advantage of equation 2.37 is that this equation gives an *unbiased estimate* of variance. This concept will be discussed next. Note that if we use N instead of $N - 1$ in equation 2.37, the computed variance is called *population variance*. Its square root is *population standard deviation*. When $N > 30$, the difference between sample variance and population variance becomes negligible.

Unbiased estimates. Note that each term X_i in the sample data set $\{X_1, X_2, \ldots, X_N\}$ is itself a random variable just like X, because the measured value of X_i contains some randomness and is subjected to chance. In other words, if N measurements were taken at one time and then the same measurements were repeated, the values would be different from the first set, since X was random to begin with. It follows that $\bar{X}$ and S in equations 2.36 and 2.37 are also random variables. Note that the mean value of $\bar{X}$ is

$$E(\bar{X}) = E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{N\mu}{N}$$

Hence,

$$E(\bar{X}) = \mu \tag{2.38}$$

We know that $\bar{X}$ is an estimate for μ. Also, from equation 2.38, we observe that the mean value of $\bar{X}$ is μ. Hence, the sample mean $\bar{X}$ is an *unbiased estimate* of mean value μ. Similarly, from equation 2.37, we can show that the mean value of S^2 is

$$E(S^2) = \sigma^2 \tag{2.39}$$

assuming that X_i are independent measurements. Thus, the sample variance S^2 is an unbiased estimate of variance σ^2. In general, if the mean value of an estimate is equal to the exact value of the parameter that is being estimated, the estimate is said to be unbiased. Otherwise, it is a *biased estimate*.

Example 2.8

An instrument has a response X that is random, with standard deviation σ. A set of N independent measurements $\{X_1, X_2, \ldots, X_N\}$ is made and the sample mean $\bar{X}$ is computed. Show that the standard deviation of $\bar{X}$ is $\sigma/\sqrt{N}$.

Also, a measuring instrument produces a random error whose standard deviation is 1 percent. How many measurements should be averaged in order to reduce the standard deviation of error to less than 0.05 percent?

Solution To solve the first part of the problem, start with equation 2.36 and use the properties of variance given by equations 2.33 and 2.35:

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left[\frac{1}{N}(X_1 + X_2 + \cdots + X_N)\right]\\ &= \frac{1}{N^2}\text{Var}(X_1 + X_2 + \cdots + X_N)\\ &= \frac{1}{N^2}[\text{Var}(X_1 + \text{Var}X_2 + \cdots + \text{Var}X_N)]\\ &= \frac{N\sigma^2}{N^2}\end{aligned}$$

Here we used the fact that X_i are indpendent.

Hence,

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{N} \tag{2.40}$$

Accordingly,

$$\text{Std}(\bar{X}) = \frac{\sigma}{\sqrt{N}} \tag{2.41}$$

For the second part of the problem, $\sigma = 1$ percent and $\sigma/\sqrt{N} < 0.05$ percent. Then,

$$\frac{1}{\sqrt{N}} < 0.05$$

or

$$N > 400$$

Thus, we should average more than 400 measurements to obtain the specified accuracy.

Gaussian distribution. Gaussian distribution, or *normal distribution*, is probably the most extensively used probability distribution in engineering applications. Apart from its ease of use, another justification for its widespread use is provided by the *central limit theorem*. This theorem states that a random variable that is

nce S^2 is an
estimate is
mate is said

r. A set of N
an $\bar{X}$ is com-

ard deviation
luce the stan-

5 and use the

(2.40)

(2.41)

ercent. Then,

pecified accu-

tribution, is
ing applica-
l use is pro-
riable that is

formed by summing a very large number of independent random variables takes Gaussian distribution in the limit. Since many engineering phenomena are consequences of numerous independent random causes, the assumption of normal distribution is justified in many cases. The validity of Gaussian assumption can be checked by plotting data on *probability graph paper* or by using various tests such as the *chi-square test*.

The Gaussian probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2.42}$$

Note that only two parameters, mean μ and standard deviation σ, are necessary to determine a Gaussian distribution completely.

A closed algebraic expression cannot be given for the cumulative probability distribution function $F(x)$ of Gaussian distribution. It should be evaluated by numerical integration. Numerical values for the normal distribution curve are available io tabulated form, with the random variable X being normalized with respect to μ and σ according to

$$Z = \frac{X - \mu}{\sigma} \tag{2.43}$$

Note that the mean value of this normalized variable Z is

$$E(Z) = E[(X - \mu)/\sigma] = [E(X) - \mu]/\sigma$$
$$= (\mu - \mu)/\sigma$$

or

$$E(Z) = 0 \tag{2.44}$$

and the variance of Z is

$$\text{Var}(Z) = \text{Var}[(X - \mu)/\sigma] = \text{Var}(X - \mu)/\sigma^2$$
$$= \text{Var}(X)/\sigma^2 = \sigma^2/\sigma^2$$

or

$$\text{Var}(Z) = 1 \tag{2.45}$$

Furthermore, the probability density function of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \tag{2.46}$$

What is usually tabulated is the area under the density curve $f(z)$ of the normalized random variable Z for different values of z. A convenient form is presented in table 2.3, where the area under the $f(z)$ curve from 0 to z is tabulated up to four decimal places for different positive values of z up to two decimal places. Since the density curve is symmetric about the mean value (zero for the normalized case), values for negative z do not have to be tabulated. Furthermore, when $z \to \infty$, area A in table 2.3 approaches 0.5. The value for $z = 3.09$ is already 0.4990. Hence, for most practical purposes, area A may be taken as 0.5 for z values greater than 3.0. Since Z is norstraightforward integration using properties of trigonometric functions results in the following response:

$$y = (a_1^2 + a_2^2)\frac{t}{2} - \frac{a_1^2}{4\omega_1}\sin 2\omega_1 t - \frac{a_2^2}{4\omega_2}\sin 2\omega_2 t$$

$$+ \frac{a_1 a_2}{2(\omega_1 - \omega_2)}\sin(\omega_1 - \omega_2)t - \frac{a_1 a_2}{2(\omega_1 + \omega_2)}\sin(\omega_1 + \omega_2)t + y(0)$$

Note that the discrete frequency components $2\omega_1$, $2\omega_2$, $(\omega_1 - \omega_2)$, and $(\omega_1 + \omega_2)$ are created. Also, there is a continuous spectrum that is contributed by the linear function of t present in the response.

The fact that nonlinear systems create new frequency components is the basis of well-known *describing function analysis* of nonlinear control systems. In this case, the response of a nonlinear component to a sinusoidal (harmonic) input is represented by a Fourier series, with frequency cnmponents that are multiples of the input frequency. Details of the describing function approach can be found in textbooks on nonlinear control theory.

Several methods are available to reduce or eliminate nonlinear behavior in systems. They include calibration (in the static case), use of linearizing elements, such as resistors and amplifiers to neutralize the nonlinear effects, and the use of nonlinear feedback. It is alsn a good practice to take the following precautions:

1. Avoid operating the device over a wide range of signal levels.
2. Avoid operation over a wide frequency band.
3. Use devices that do not generate large mechanical motions.
4. Minimize coulomb friction.
5. Avoid loose joints and gear coupling (i.e., use *direct drive* mechanisms).

2.5 IMPEDANCE CHARACTERISTICS

When components such as measuring instruments, control boards, process (plant) hardware, and signal-conditioning equipment are interconnected, it is necessary to *match* impedances properly at each interface in order to realize their rated performance level. One adverse effect of improper impedance matching is the *loading effect*. For example, in a measuring system, the measuring instrument can distort the signal that is being measured. The resulting error can far exceed other types of measurement error. Loading errors result from connecting measuring devices with low input impedance to a signal source.

Impedance can be interpreted either in the traditional electrical sense or in the mechanical sense, depending on the signal being measured. For example, a heavy accelerometer can introduce an additional dynamic load that will modify the actual acceleration at the monitoring location. Similarly, a vnltmeter can modify the currents (and voltages) in a circuit, and a thermocouple junction can modify the temperature that is being measured. In mechanical and electrical systems, loading errors can appear as phase distortions as well. Digital hardware also can produce loading errors. For example, an analog-to-digital conversion (ADC) board can load the amplifier output from a strain gage bridge circuit, thereby significantly affecting digitized data (see chapter 4).

Another adverse effect of improper impedance consideration is inadequate output signal levels, which make signal processing and transmission very difficult. Many types of transducers (e.g., piezoelectric accelerometers, impedance heads, and microphones) have high output impedances on the order of a thousand megohms. These devices generate low output signals, and they would require conditioning to step up the signal level. *Impedance-matching amplifiers*, which have high input impedances and low output impedances (a few ohms), are used for this purpose (e.g., charge amplifiers are used in conjunction with piezoelectric sensors). A device with a high input impedance has the further advantage that it usually consumes less power (v^2/R is low) for a given input voltage. The fact that a low input impedance device extracts a high level of power from the preceding output device may be interpreted as the reason for loading error.

Cascade Connection of Devices

Consider a standard two-port electrical device. The *output impedance* Z_o of such a device is defined as the ratio of the open-circuit (i.e., no-load) voltage at the output port to the short-circuit current at the output port.

Open-circuit voltage at output is the output voltage present when there is no current flowing at the output port. This is the case if the output port is not connected to a load (impedance). As soon as a load is connected at the output of the device, a current will flow through it, and the output voltage will drop to a value less than that of the open-circuit voltage. To measure open-circuit voltage, the rated input voltage is applied at the input port and maintained constant, and the output voltage is measured using a voltmeter that has a very high (input) impedance. To measure short-circuit current, a very low-impedance ammeter is connected at the output port.

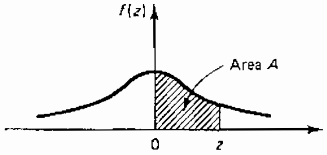
The *input impedance* Z_i is defined as the ratio of the rated input voltage to the corresponding current through the input terminals while the output terminals are maintained as an open circuit.

Note that these definitions are associated with electrical devices. A generalization is possible by interpreting voltage and velocity as *across variables*, and current and force as *through variables*. Then mechanical *mobility* should be used in place of electrical impedance.

Using these definitions, input impedance Z_i and output impedance Z_o can be represented schematically as in figure 2.12a. Note that v_o is the open-circuit output voltage. When a load is connected at the output port, the voltage across the load will be different from v_o. This is caused by the presence of a current through Z_o. In the frequency domain, v_i and v_o are represented by their respective *Fourier spectra*. The corresponding transfer relation can be expressed in terms of the complex frequency response (transfer) function $G(j\omega)$ under open-circuit (no-load) conditions:

$$v_o = Gv_i \qquad (2.7)$$

TABLE 2.3 A TABLE OF GAUSSIAN PROBABILITY DISTRIBUTION



$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$$

$$A = \int_0^z f(z)dz$$

Area A

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3233
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4758	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4799	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

malized with respect to σ, $z = 3$ actually corresponds to three times the standard deviation of the original random variable X. It follows that for a Gaussian random variable, most of the values will fall within $\pm 3\sigma$ about the mean value. It can be stated that approximately

- 68 percent of the values will fall within $\pm\ \sigma$ about μ
- 95 percent of the values will fall within $\pm 2\sigma$ about μ
- 99.7 percent of the values will fall within $\pm 3\sigma$ about μ

These can be easily verified using table 2.3.

8	.09
19	0.0359
14	0.0753
03	0.1141
80	0.1517
44	0.1879
90	0.2224
17	0.2549
23	0.2852
06	0.3233
65	0.3389
99	0.3621
10	0.3830
97	0.4015
62	0.4177
06	0.4319
29	0.4441
35	0.4545
25	0.4633
99	0.4706
61	0.4767
12	0.4817
54	0.4857
87	0.4890
13	0.4916
34	0.4936
51	0.4952
63	0.4964
73	0.4974
80	0.4981
86	0.4986
89	0.4990

ne standard
ian random
:. It can be

Statistical process control. In statistical process control (SPC), statistical analysis of process responses is used to generate control actions. This method of control is applicable in many situations of process control, including manufacturing quality control, control of chemical process plants, computerized office management systems, inventory control systems, and urban transit control systems. A major step in statistical process control is to compute control limits (or action lines) on the basis of measured data from the process.

Control Limits or Action Lines. Since a very high percentage of readings from an instrument should lie within ± 3σ about the mean value, according to the normal distribution, these boundaries (-3σ and $+3\sigma$) drawn about the mean value may be considered *control limits* or *action lines* in statistical process control. If any measurements fall outside the action lines, corrective measures such as recalibration, controller adjustment, or redesign should be carried out.

Steps of SPC. The main steps of statistical process control are as follows:

1. Collect measurements of appropriate response variables of the process.
2. Compute the mean value of the data, the upper control limit, and the lower control limit.
3. Plot the measured data and the two control limits on a control chart.
4. If measurements fall outside the control limits, take corrective action and repeat the control cycle (go to step 1).

If the measurements always fall within the control limits, the process is said to be in statistical control.

Example 2.9

Error in a satellite tracking system was monitored on-line for a period of one hour to determine whether recalibration or gain adjustment of the tracking controller would be necessary. Four measurements of the tracking deviation were taken in a period of five minutes, and twelve such data groups were acquired during the one-hour period. Sample means and sample variances of the twelve groups of data were computed. The results are tabulated as follows:

Period i	1	2	3	4	5	6	7	8	9	10	11	12
Sample mean $\bar{X}_i$	1.34	1.10	1.20	1.15	1.30	1.12	1.26	1.10	1.15	1.32	1.35	1.18
Sample variance S_i^2	0.11	0.02	0.08	0.10	0.09	0.02	0.06	0.05	0.08	0.12	0.03	0.07

Draw a control chart for the error process, with control limits (action lines) at $\bar{X} \pm 3\sigma$. Establish whether the tracking controller is in statistical control or needs adjustment.

Solution The overall mean tracking deviation,

$$\bar{X} = \frac{1}{12}\sum_{i=1}^{12}\bar{X}_i$$

is computed to be $\bar{X} = 1.214$. The average sample variance,

$$\bar{S}^2 = \frac{1}{12}\sum_{i=1}^{12}S_i^2$$

is computed to be $\bar{S}^2 = 0.069$. Since there are four readings within each period, the standard deviaton σ of group mean $\bar{X}_i$ can be estimated, using equation 2.41, as

$$S = \frac{\bar{S}}{\sqrt{4}} = \frac{\sqrt{0.069}}{\sqrt{4}} = 0.131$$

The upper control limit (action line) is at (approximately)

$$x = \bar{X} + 3S = 1.214 + 3 \times 0.131 = 1.607$$

The lower control limit (action line) is at

$$x = \bar{X} - 3S = 0.821$$

These two lines are shown on the control chart in figure 2.20. Since the sample means lie within the two action lines, the process is considered to be in statistical control, and controller adjustments would not be necessary. Note that if better resolution is required in making this decision, individual readings, rather than group means, should be plotted in figure 2.20.

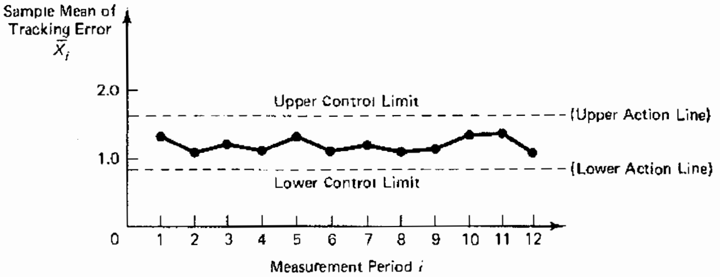


Figure 2.20. Control chart for the satellite tracking errur example.

Confidence intervals. The probability that the value of a random variable would fall within a specified interval is called a *confidence level.* As an example, consider a Gaussian random variable X that has mean μ and standard deviation σ. This is denoted by

$$X = N(\mu, \sigma) \tag{2.47}$$

Suppose that N measurements $\{X_1, X_2, \ldots, X_N\}$ are made. The sample mean $\bar{X}$ is an unbiased estimate for μ. We also know that the standard deviation of $\bar{X}$ is σ/N. Now consider the normalized random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \tag{2.48}$$

This is a Gaussian random variable with zero mean and unity standard deviation. The probability p that the values of Z fall within $\pm z_o$:

$$P(-z_o < Z \leq z_o) = p \tag{2.49}$$

can be determined from table 2.3 for a specified value of z_o. Now substituting equation 2.48 in 2.49, we get

$$P\left(-z_o < \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \leq z_o\right) = p$$

or

$$P\left(\bar{X} - \frac{z_o\sigma}{\sqrt{N}} \leq \mu < \bar{X} + \frac{z_o\sigma}{\sqrt{N}}\right) = p \tag{2.50}$$

Nnte that the lower limit has the $\leq$ sign and the upper limit has the $<$ sign within the parentheses. These have been used for mathematical precision, but for practical purposes, either $\leq$ or $<$ may be used in each limit. Now, from equation 2.50, it follows that the confidence level is p that the actual mean value μ would fall within $\pm z_o\,\sigma/\sqrt{N}$ of the estimated (sample) mean value $\bar{X}$.

Example 2.10

The angular resolution of a resolver (a rotary displacement sensnr—see chapter 3) was tested sixteen times, independently, and recorded in degrees as follows:

0.11, 0.12, 0.09, 0.10, 0.10, 0.14, 0.08, 0.08

0.13, 0.10, 0.10, 0.12, 0.08, 0.09, 0.11, 0.15

If the standard deviation of the angular resolution of this brand of resolvers is known tn be 0.01°, what are the odds that the mean resolution would fall within 5 percent of the sample mean?

Solution To solve this problem, we assume that resnlution is normally distributed. The sample mean is computed as

$$\bar{X} = \frac{1}{16}(0.11 + 0.12 + \cdots + 0.11 + 0.15) = 0.10625$$

In view of equation 2.50, we must have

$$\frac{z_o\sigma}{\sqrt{16}} = 5\% \text{ of } \bar{X}$$

Hence,

$$\frac{z_o \times 0.01}{\sqrt{16}} = \frac{5}{100} \times 0.10625$$

or

$$z_o = 2.125$$

Now, from table 2.3,

$$P(-2.125 < Z < 2.125) = 2 \times \frac{(0.4830 + 0.4834)}{2} = 0.9664$$

Sign test and binomial distribution. Sign test is useful in comparing accuracies of two similar instruments. First, measurements should be made on the same measurand (input signal to instrument) using the two devices. Next, the readings of one instrument are subtracted from the corresponding readings of the second instrument, and the results are tabulated. Finally, the probability of getting the number of negative signs (or positive signs) equal to what is present in the tabulated results is computed using *binomial distribution.*

Before discussing binomial distribution, let us introduce some new terminology. First, *factorial* r (denoted by $r!$) of an integer r is defined as the product

$$r! = r \times (r-1) \times (r-2) \times \cdots \times 2 \times 1 \tag{2.51}$$

Now, suppose that there are n distinct articles that are distinguishable from one another. The number of ways in which r articles could be picked from the batch of n, giving proper consideration to the order in which the r articles are picked (or arranged), is called the number of *permutations* of r from n. This is denoted by ^nP_r, which is given by

$$\begin{aligned} {}^nP_r &= n \times (n-1) \times (n-2) \times \cdots \times (n-r+2) \times (n-r+1) \\ &= \frac{n!}{(n-r)!} \end{aligned} \tag{2.52}$$

This can be easily verified, since the first article can be chosen in n ways and the second article can be chosen from the remaining $(n-1)$ articles in $(n-1)$ ways and kept next to the first article, and so on.

If we disregard the order in which the r articles are picked (and arranged), the number of possible choices of r articles is termed the number of *combinations* of r from n. This is denoted by ^nC_r. Now, since each combination can be arranged in $r!$ different ways (if the order of arrangement is considered), we have

$$^nC_r \times r! = {}^nP_r \tag{2.53}$$

Hence, using equation 2.52, we get

$$\begin{aligned} {}^nC_r &= \frac{n \times (n-1) \times (n-2) \times \cdots \times (n-r+2) \times (n-r+1)}{r!} \\ &= \frac{n!}{(n-r)!\,r!} \end{aligned} \tag{2.54}$$

With the foregoing notation, we can introduce binomial distribution in the context of sign test. Suppose that n pairs of readings are taken from the two instru-

nparing ac-
ade on the
t, the read-
the second
getting the
ie tabulated

v terminol-
oduct

(2.51)

om one an-
batch of n,
ked (or ar-
ited by ^nP_r,

+ 1)

(2.52)

iys and the
− 1) ways

anged), the
ations of r
anged in $r!$

(2.53)

(2.54)

:ion in the
two instru-

ments. If the probability that a difference in reading would be positive is p, then the probability that the difference wnuld be negative is $1 - p$. Note that if the systematic error in the two instruments is the same and if the random error is purely random, then $p = 0.5$.

The probability of getting exactly r positive signs among the n entries in the table is

$$p(r) = {}^nC_r p^r(1 - p)^{n-r} \tag{2.55}$$

To verify equation 2.55, note that this event is similar to picking exactly r items from n items and constraining each picked item to be positive (having probability p) and also constraining the remaining $(n - r)$ items to be negative (having probability $1 - p$). Note that r is a discrete variable that takes values $r = 1, 2, \ldots, n$. Furthermore, it can be easily verified that

$$\sum_{r=1}^{n} p(r) = \sum_{r=1}^{n} {}^nC_r p^r(1 - p)^{n-r} = (p + 1 - p)^n = 1 \tag{2.56}$$

Hence, $p(r)$, $r = 1, 2, \ldots, n$, is a discrete function that resembles a continuous probability density function $f(x)$. In fact, $p(r)$ given by equation 2.55 represents *binomial probability distribution*. Using equation 2.55, we can perform the sign test. The details of the test are conveniently explained by means of an example.

Example 2.11

To compare the accuracies of two brands of differential transformers (DTs, which are displacement sensors—see chapter 3), the same rotation (in degrees) of a robot arm joint was measured using both brands, DT1 and DT2. The following ten measurement pairs were taken:

DT1	10.3	5.6	20.1	15.2	2.0	7.6	12.1	18.9	22.1	25.2
DT2	9.8	5.8	20.0	16.0	1.9	7.8	12.2	18.7	22.0	25.0

Assuming that both devices are used simultaneously (so that backlash and other types of repeatability errors in manipulators do not enter into our problem), determine whether the two brands are equally accurate at the 70 percent level of significance.

Solution First, we form the sign table by taking the differences of corresponding measurements:

DT1 − DT2	0.5	−0.2	0.1	−0.8	0.1	−0.2	−0.1	0.2	0.1	0.2

Note that there are six positive signs and four negative signs. If we had tabulated DT2 − DT1, however, we would get four positive signs and six negative signs. Both these cases should be taken into account in the sign test. Furthermore, more than six positive signs or fewer than four positive signs would make the two devices less similar (in accuracy) than what is indicated by the data. Hence, the probability of getting six or more positive signs or four or fewer positive signs should be computed in this example in order to estimate the possible match (in accuracy) of the two devices.

If the error in both transducers is the same, we should have

$$P(\text{positive difference}) = p = 0.5$$

This is the hypothesis that we are going to test. Using equation 2.55, the probability of getting six or more positive signs or four or fewer negative signs is calculated as

$$1 - \text{probability of getting exactly 5 positive signs}$$

$$= 1 - {}^{10}C_5(0.5)^5 \times (0.5)^5 = 1 - \frac{10!}{5!5!} \times (0.5)^{10}$$

$$= 1 - 0.246 = 0.754$$

Note that the hypothesis of two brands being equally accurate is supported by the test data at a level of significance over 75 percent, which is better than the specified value of 70 percent.

Least squares fit. We have mentioned that instrument *linearity* may be measured by the largest deviation of the input/output data (or calibration curve) from the least squares straight-line fit of data. Since many algebraic expressions become linear when plotted to a logarithmic scale, linear (straight-line) fit is generally more accurate if log-log axes are used. Linear least squares fit can be thought of as an estimation method because it "estimates" the two parameters of an input/output model, the straight line, that fits a given set of data such that the squared error is a minimum. The estimated straight line is also known as the *linear regression line* or *mean calibration curve*.

Consider N pairs of data $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)\}$ in which X denotes the *independent variable* (input variable) and Y denotes the *dependent variable* (output variable).

Suppose that the estimated linear regression is given by

$$Y = mX + a \tag{2.57}$$

For the independent variable value X_i, the dependent variable value on the regression line is $(mX_i + a)$, but the actual (measured) value of the dependent variable is Y_i. Hence, the sum of squared error for all data points is

$$e = \sum_{i=1}^{N} (Y_i - mX_i - a)^2 \tag{2.58}$$

We have to minimize e with respect to the two parameters m and a. The required conditions are

$$\frac{\partial e}{\partial m} = 0 \quad \text{and} \quad \frac{\partial e}{\partial a} = 0$$

By carrying out these differentiations in equation 2.58, we get

$$\sum_{i=1}^{N} X_i(Y_i - mX_i - a) = 0$$

and

$$\sum_{i=1}^{N}(Y_i - mX_i - a) = 0$$

Dividing the two equations by N and using the definition of sample mean, we get

$$\frac{1}{N}\sum X_iY_i - \frac{m}{N}\sum X_i^2 - a\bar{X} = 0 \qquad \text{(i)}$$

$$\bar{Y} - m\bar{X} - a = 0 \qquad \text{(ii)}$$

Solving these two simultaneous equations for m, we obtain

$$m = \left(\frac{1}{N}\sum_{i=1}^{N} X_iY_i - \bar{X}\bar{Y}\right)\Bigg/\left(\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2\right) \qquad (2.59)$$

The parameter a does not have to be explicitly expressed, because from equations 2.57 and (ii), we can eliminate a and express the linear regression line as

$$Y - \bar{Y} = m(X - \bar{X}) \qquad (2.60)$$

Note from equation 2.57 that a is the Y-axis intercept (i.e., the value of Y when $X = 0$) and is given by

$$a = \bar{Y} - m\bar{X} \qquad (2.61)$$

Example 2.12

Consider the capacitor circuit shown in figure 2.21. First, the capacitor is charged to voltage v_o using a constant DC voltage source (switch in position 1); then it is discharged through a known resistance R (switch in position 2). Voltage decay during discharge is measured at known time increments. Three separate tests are carried out. The measured data are as follows:

Time t (sec)	0.1	0.2	0.3	0.4	0.5
Voltage v (volts)					
Test 1	7.3	2.8	1.0	0.4	0.1
Test 2	7.4	2.7	1.1	0.3	0.2
Test 3	7.3	2.6	1.0	0.4	0.1

If the resistance is accurately known to be 1,000 Ω, estimate the capacitance C in microfarads (μF) and the source voltage v_o in volts.

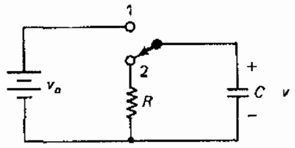


Figure 2.21. A circuit for the least squares estimation of capacitance.

probability of
ılated as

.10

'ted by the test
specified value

arity may be
n curve) from
sions become
enerally more
of as an esti-
output model,
ror is a mini-
ı line or *mean*

ıich X denotes
ıdent variable

(2.57)

the regression
variable is Y_i.

(2.58)

. The required

Solution To solve this problem, we assume the well-known expression for the free decay of voltage across a capacitor:

$$v(t) = v_o \exp[-t/(RC)] \tag{i}$$

Take the natural logarithm of equation (i):

$$\ln v = -\frac{t}{RC} + \ln v_o \tag{ii}$$

With $Y = \ln v$ and $X = t$, equation (ii) represents a straight line with slope

$$m = -\frac{1}{RC} \tag{iii}$$

and the Y-axis intercept

$$a = \ln v_o \tag{iv}$$

Using all the data, the overall sample means can be computed. Thus,

$$\bar{X} = 0.3 \quad \text{and} \quad \bar{Y} = -0.01335$$

$$\frac{1}{N}\sum X_i Y_i = -0.2067 \quad \text{and} \quad \frac{1}{N}\sum X_i^2 = 0.11$$

Now substitute these values in equations 2.59 and 2.61. We get

$$m = -10.13 \quad \text{and} \quad a = 3.02565$$

Next, from equation (iii), with $R = 1{,}000$, we have

$$C = \frac{1}{10.13 \times 1000}\,\text{F} = 98.72\ \mu\text{F}$$

From equation (iv),

$$v_o = 20.61 \text{ volts}$$

Note that in this problem, the estimation error would be tremendous if we did not use log scaling for the linear fit.

Least squares curve fitting is not limited to linear (i.e., straight-line) fit. The method can be extended to a polynomial fit of any order. For example, in *quadratic fit*, the data are fitted to a second-order (i.e., quadratic) polynomial. In that case, there are three unknown parameters, which would be determined by minimizing the quadratic error.

Error Combination

Error in a response variable of an instrument or in an estimated system parameter would depend on errors present in measured variables and parameter values that are used to determine the unknown variable or parameter. Knowing how component errors are propagated within a multicomponent system and how individual errors in system variables and parameters contribute toward the overall error in a particular response variable or parameter would be important in estimating error limits in comfor the free

(i)

(ii)

ɔe

(iii)

(iv)

ɔus if we did

ne) fit. The
n *quadratic*
ı that case,
imizing the

. parameter
ıes that are
ıponent er-
ıl errors in
. particular
its in com-

plex instruments. For example, if the output power in a gas turbine is computed by measuring torque and speed at the output shaft, error margins in the two measured "response variables" (torque and speed) would be directly combined into the error in the power computation. Similarly, if the natural frequency of a simple suspension system is determined by measuring mass and spring stiffness "parameters" of the suspension, the natural frequency estimate would be directly affected by possible errors in mass and stiffness measurements. Extending this idea further, the overall error in a control system depends on individual error levels in various components (sensors, actuators, controller hardware, filters, amplifiers, etc.) of the system and on the manner in which these components are physically interconnected and physically interrelated. For example, in a robotic manipulator, the accuracy of the actual trajectory of the end effector will depend on the accuracy of sensors and actuators at manipulator joints and on the accuracy of the robot controller. Note that we are dealing with a generalized idea of error propagation that considers errors in system variables (e.g., input and output signals, such as velocities, forces, voltages, currents, temperatures, heat transfer rates, pressures, and fluid flow rates), system parameters (e.g., mass, stiffness, damping, capacitance, inductance, and resistance), and system components (e.g., sensors, actuators, filters, amplifiers, and control circuits).

For the analytical development of a basic result in error combination, we will start with a functional relationship of the form

$$y = f(x_1, x_2, \ldots, x_r) \tag{2.62}$$

Here, x_i are the independent system variables or parameter values whose error is propagated into a dependent variable (or parameter value) y. Determination of this functional relationship is not always simple, and the relationship itself may be in error. Since our intention is to make a reasonable estimate for possible error in y due to the combined effect of errors from x_i, an approximate functional relationship would be adequate in most cases. Let us denote error in a variable by the differential of that variable. Taking the differential of equation 2.62, we get

$$\delta y = \frac{\partial f}{\partial x_1}\delta x_1 + \frac{\partial f}{\partial x_2}\delta x_2 + \cdots + \frac{\partial f}{\partial x_r}\delta x_r \tag{2.63}$$

For those who are not familiar with differential calculus, equation 2.63 should be interpreted as the first-order terms in a *Taylor series expansion* of equation 2.62. Now, rewriting equation 2.63 in the fractional form, we get

$$\frac{\delta y}{y} = \sum_{i=1}^{r}\left[\frac{x_i}{y}\frac{\partial f}{\partial x_i}\frac{\delta x_i}{x_i}\right] \tag{2.64}$$

Here, $\delta y/y$ represents the overall error and $\delta x_i/x_i$ represents the component error, expressed as fractions. We shall consider two types of estimates for overall error.

Absolute error. Since error δx_i could be either positive or negative, an upper bound for the overall error is obtained by summing the absolute value of each right-hand-side term in equation 2.64. This estimate e_{ABS}, which is termed *absolute error*, is given by

$$e_{\mathrm{ABS}} = \sum_{i=1}^{r} \left| \frac{x_i}{y} \frac{\partial f}{\partial x_i} \right| e_i \tag{2.65}$$

Note that component error e_i and absolute error e_{ABS} in equation 2.65 are always positive quantities; when specifying error, however, both positive and negative limits should be indicated or implied. (e.g., $\pm e_{\mathrm{ABS}}$, $\pm e_i$).

SRSS error. Equation 2.65 provides a conservative (upper bound) estimate for overall error. Since the estimate itself is not precise, it is often wasteful to introduce such a high conservatism. A nonconservative error estimate that is frequently used in practice is the *square root of sum of squares* (SRSS) error. As the name implies, this is given by

$$e_{\mathrm{SRSS}} = \left[\sum_{i=1}^{r} \left(\frac{x_i}{y} \frac{\partial f}{\partial x_i} e_i \right)^2 \right]^{1/2} \tag{2.66}$$

Note that this is not an upper bound estimate for error and that $e_{\mathrm{SRSS}} < e_{\mathrm{ABS}}$ when more than one nonzero error contribution is present. The SRSS error relation is particularly suitable when component error is represented by the standard deviation of the associated variable or parameter value and when the corresponding error sources are independent.

We shall conclude this chapter by giving several examples of error combination.

Example 2.13

Using the absolute value method for error combination, determine the fractional error in each item x_i so that the contribution from each item to the overall error e_{ABS} is the same.

Solution For equal contribution, we must have

$$\left| \frac{x_1}{y} \frac{\partial f}{\partial x_1} \right| e_1 = \left| \frac{x_2}{y} \frac{\partial f}{\partial x_2} \right| e_2 = \cdots = \left| \frac{x_r}{y} \frac{\partial f}{\partial x_r} \right| e_r$$

Hence,

$$r \left| \frac{x_i}{y} \frac{\partial f}{\partial x_i} \right| e_i = e_{\mathrm{ABS}}$$

Thus,

$$e_i = e_{\mathrm{ABS}} \Big/ \left(r \left| \frac{x_i}{y} \frac{\partial f}{\partial x_i} \right| \right) \tag{2.67}$$

Example 2.14

The result obtained in example 2.13 is useful in the design of multicomponent systems and in the cost-effective selection of instrumentation for a particular application. Using equation 2.67, arrange the items x_i in their order of significance.

ıed *absolute*

(2.65)

5 are always
ıegative lim-

nd) estimate
:ful to intro-
is frequently
he name im-

(2.66)

$< e_{\text{ABS}}$ when
lation is par-
deviation of
error sources

'or combina-

ractional error
ror e_{ABS} is the

(2.67)

ıonent systems
lication. Using

Solution Note that equation 2.67 may be written as

$$e_i = K \Big/ \left| x_i \frac{\partial f}{\partial x_i} \right| \tag{2.68}$$

where K is a quantity that does not vary with x_i. It follows that for equal error contribution from all items, error in x_i should be inversely proportional to $|x_i(\partial f/\partial x_i)|$. In particular, the item with the largest $|x(\partial f/\partial x)|$ should be made most accurate. In this manner, allowable relative accuracy for various components can be estimated. Since, in general, the most accurate device is also the most costly one, instrumentation cost can be optimized if components are selected according to the required overall accuracy, using a criterion such as that implied by equatinn 2.68.

Example 2.15

Tension T at point P in a cable can be computed with the knowledge of cable sag y, cable length s, cable weight w per unit length, and the minimum tension T_o at point O (see figure 2.22). The applicable relationship is

$$1 + \frac{w}{T_o} y = \sqrt{1 + \frac{w^2}{T^2} s^2}$$

For a particular arrangement, it is given that $T_o = 100$ lbf. The following parameter values were measured:

$$w = 1 \text{ lb/ft}, \qquad s = 10 \text{ ft}, \qquad y = 0.412 \text{ ft}$$

Calculate the tension T.

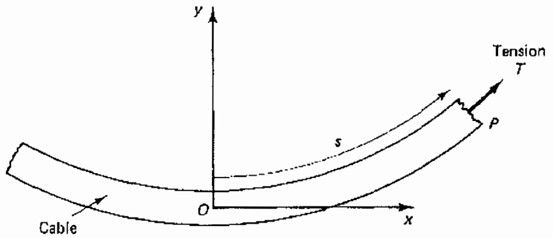


Figure 2.22. Cable tension example of error combination.

In addition, if the measurements y and s each have 1 percent error and the measurement w has 2 percent error in this example, estimate the percentage error in T.

Now suppose that equal contributions to error in T are made by y, s, and w. What are the corresponding percentage error values for y, s, and w so that the overall error in T is equal to the value computed in the previous part of the problem? Which of the three quantities y, s, and w should be measured most accurately, according to the equal contribution criterion?

Solution To make the analysis simpler, let us first square the given relationship:

$$\left(1 + \frac{w}{T_o} y\right)^2 = 1 + \frac{w^2 s^2}{T^2} \tag{i}$$

Substituting numerical values,

$$\left(1 + \frac{1 \times 0.412}{100}\right)^2 = 1 + \frac{1 \times 10^2}{T^2}$$

Hence,

$$T = 110 \text{ lbf}$$

Next, we differentiate equation (i) to get the differential relationship

$$2\left(1 + \frac{w}{T_o}y\right)\left[\frac{y}{T_o}\delta w + \frac{w}{T_o}\delta y\right] = \frac{2ws^2}{T^2}\delta w + \frac{2w^2 s}{T^2}\delta s - \frac{2w^2 s^2}{T^3}\delta T \qquad \text{(ii)}$$

Note that T_o is treated as a constant. The implication is that T_o is known with 100 percent accuracy. On rearranging the terms in equation (ii) and after straightforward algebraic manipulation, we get

$$\frac{\delta T}{T} = (1 - z)\frac{\delta w}{w} + \frac{\delta s}{s} - z\frac{\delta y}{y} \qquad \text{(iii)}$$

where

$$z = \frac{T^2 y}{s^2 w T_o}\left\{1 + \frac{wy}{T_o}\right\} \qquad \text{(iv)}$$

Using the absolute value method for error combination, we can express the error level in T as

$$e_{\text{ABS}} = |1 - z| e_w + e_s + z e_y \qquad \text{(v)}$$

Substituting the given numerical values,

$$z = \frac{110^2 \times 0.412}{10^2 \times 1 \times 100}\left(1 + \frac{1 \times 0.412}{100}\right) = 0.5$$

Hence,

$$e_{\text{ABS}} = 0.5\, e_w + e_s + 0.5\, e_y \qquad \text{(vi)}$$

Also, it is given that

$$e_w = 2\%, \qquad e_s = e_y = 1\%$$

Hence,

$$e_{\text{ABS}} = (1 - 0.5) \times 2 + 1 + 0.5 \times 1\% = 2.5\%$$

For equal contribution of error, in view of equation (vi), we have

$$0.5 e_w = e_s = 0.5 e_y = \frac{2.5}{3}\%$$

Hence,

$$e_w = 1.7\%, \qquad e_s = 0.8\%, \qquad e_y = 1.7\%$$

Note that the variable s should be measured most accurately according to the equal contribution criterion, because the tolerable level of error is the smallest for this variable.

(ii)

100 per-
ard alge-

(iii)

(iv)

:ror level

(v)

(vi)

qual con-
variable.

PROBLEMS

2.1. Discuss a type of device that could serve as a clock for a digital control system. Identify the main functions of such a clock in real-time control.

2.2. Explain the operation of two measuring devices of your choice, one to measure voltage and the other to measure temperature. Clearly identify the sensor stage and the transducer stage (or stages) in each of these devices.

2.3. Compare and contrast the following pairs of terms, giving suitable examples:
(a) Measurand and measured value
(b) Active transducers and passive transducers
(c) Through variables and across variables
(d) Effort variables and flow variables
(e) Impedance and admittance

2.4. Obtain transfer relations in the second-order vector-matrix form for the following two-port devices, clearly identifying the through and across variables at the input port and the output port.
(a) Gyroscope (or spinning top)
(b) Cam-follower mechanism
(c) Loudspeaker
(d) Viscous damper
(e) Thermocouple

2.5. What are pure transducers and what are ideal transducers? Discuss the five devices listed in problem 2.4 from this point of view.

2.6. The simple oscillator equation is given by

$$\ddot{y} + 2\zeta\omega_n\dot{y} + \omega_n^2 y = \omega_n^2 u(t)$$

Obtain the response y of this system for a unit step input $u(t)$ under zero initial conditions. In terms of the two parameters ζ and ω_n, obtain expressions for damping ratio, undamped natural frequency, damped natural frequency, rise time, percentage overshoot, and 2 percent settling time. Discuss the significance of each of these parameters with reference to the performance of a sensor-transducer device.

2.7. For the simple oscillator given in problem 2.6, what is the transfer function? Obtain an expression for its resonant frequency in terms of ζ and ω_n. If a sinusoidal excitation of frequency ω_n is applied to this system, what is the amplitude gain and phase lead in its response at steady state?

2.8. What do you consider a perfect measuring device? Suppose that you are asked to develop an analog device for measuring angular position in an application related to control of a kinematic linkage system (a robotic manipulator, for example). What instrument ratings (or specifications) would you consider crucial in this application? Discuss their significance.

2.9. Define electrical impedance and mechanical impedance. Identify a defect in these definitions in relation to the force-current analogy. What improvements would you suggest? What roles do input impedance and output impedance play in relation to the accuracy of a measuring device?

2.10. Discuss and contrast the following terms:
(a) Measurement accuracy
(b) Instrument accuracy

(c) Measurement error
(d) Precision
Also, for an analog sensor-transducer unit of your choice, identify and discuss various sources of error and ways to minimize or account for their influence.

2.11. A schematic diagram for a charge amplifier (with resistive feedback) is shown in figure P2.11. Obtain the differential equation governing the response of the charge amplifier. Identify the time constant of the device and discuss its significance. Would you prefer a charge amplifier to a voltage follower for conditioning signals from a piezoelectric accelerometer? Explain.

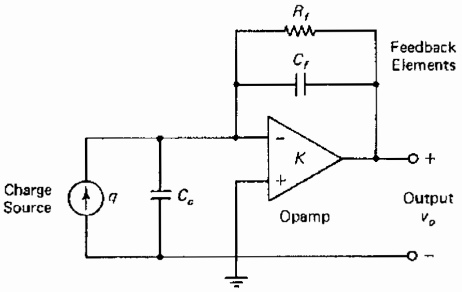


Figure P2.11. Schematic diagram for a charge amplifier.

2.12. List several response characteristics of nonlinear dynamic systems that are not exhibited by linear systems in general. Also, determine the response y of the nonlinear system

$$\left[\frac{dy}{dt}\right]^{1/3} = u(t)$$

when excited by the input $u(t) = a_1 \sin \omega_1 t + a_2 \sin \omega_2 t$. What characteristic of nonlinear systems does this result show?

2.13. What is meant by "loading error" in a signal measurement? Also, suppose that a piezoelectric sensor of output impedance Z_s is connected to a voltage-follower amplifier of input impedance Z_i. The sensor signal is v_i volts and the amplifier output is v_o volts. The amplifier output is connected to a device with very high input impedance. Plot to scale the signal ratio v_o/v_i against the impedance ratio Z_i/Z_s for values of the impedance ratio in the range 0.1 to 10.

2.14. Discuss how the accuracy of a digital controller may be affected by
(a) Stability and bandwidth of amplifier circuitry
(b) Load impedance of the analog-to-digital conversion circuitry.
Also, what methods do you suggest to minimize problems associated with these parameters?

2.15. From the point of view of loading, discuss why an active transducer is generally superior to a passive transducer. Also, discuss why impedance matching amplifiers are generally active devices.

2.16. Suppose that v_i and f_i are the across variable and the through variable at the input port of a two-port device and that v_o and f_o are the corresponding variables at the output port. A valid form for representing transfer characteristics of the device is

$$\begin{bmatrix} v_o \\ f_i \end{bmatrix} = G\begin{bmatrix} v_i \\ f_o \end{bmatrix}$$

arious

figure
plifier.
refer a
ric ac-

m for a

exhib-
ear sys-

of non-

that a
mplifier
v_o volts.
Plot to
of the

th these

lly supe-
are gen-

nput port
ne output

What are the advantages and disadvantages of this representation compared to that given by equation 2.1? Also, express transfer relations for a DC tachometer (example 2.2) in the foregoing form.

2.17. Thevenin's theorem states that with respect to the characteristics at an output port, an unknown subsystem consisting of linear passive elements and ideal source elements may be represented by a single across-variable (voltage) source v_{eq} connected in series with a single impedance Z_{eq}. This is illustrated in figures P2.17a and P2.17b. Note that v_{eq} is equal to the open-circuit across variable v_{oc} at the output port, because the current through Z_{eq} is zero. Consider the network shown in figure P2.17c. Determine the equivalent voltage source v_{eq} and the equivalent series impedance Z_{eq}, in the frequency domain, for this circuit.

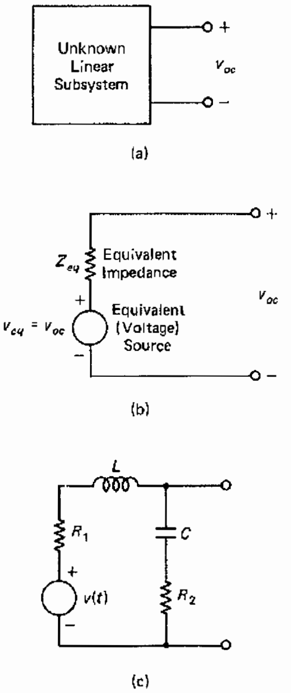


Figure P2.17. Illustration of Thevenin's theorem: (a) unknown linear subsystem; (b) equivalent representation; (c) example.

2.18. Using suitable impedance circuits, explain why a voltmeter should have a high resistance and an ammeter should have a very low resistance. What are some of the design implications of these general requirements for the two types of measuring instruments, particularly with respect to instrument sensitivity, speed of response, and robustness? Use a classical moving-coil meter as the model for your discussion.

2.19. A two-port nonlinear device is shown schematically in figure P2.19. The transfer relations under static equilibrium conditions are given by

$$v_o = F_1(f_o, f_i)$$

$$v_i = F_2(f_o, f_i)$$

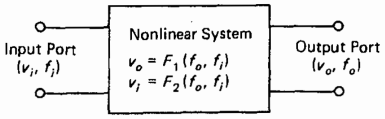


Figure P2.19. Impedance characteristics of a nonlinear system.

where v denotes an across variable, f denotes a through variable, and the subscripts o and i represent the output port and the input port, respectively. Obtain expressions for input impedance and output impedance of the system in the neighborhood of an operating point, under static conditions, in terms of partial derivatives of the functions F_1 and F_2. Explain how these impedances could be determined experimentally.

2.20. The damping constant b of the mounting structure of a machine is determined experimentally. First, the spring stiffness k is determined by applying a static load and measuring the resulting displacement. Next, mass m of the structure is directly measured. Finally, damping ratio ζ is determined using the logarithmic decrement method, by conducting an impact test and measuring the free response of the structure. A model for the structure is shown in figure P2.20. Show that the damping constant is given by

$$b = 2\zeta\sqrt{km}$$

If the allowable levels of error in the measurements of k, m, and ζ are ±2 percent, ±1 percent, and ±6 percent, respectively, estimate a percentage absolute error limit for b.

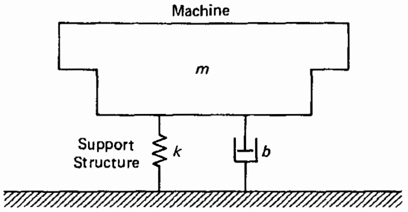


Figure P2.20. A model for the mounting structure of a machine.

2.21. In example 2.13 in the text, suppose that the square root of sum of squares (SRSS) method is used for error combination. What are the corresponding component error limits?

2.22. In example 2.15 in the text, suppose that the percentage error values specified are in fact standard deviations in the measurements of y, s, and w. Estimate the standard deviation in the estimated value of tension T.

2.23. The quality control system in a steel rolling mill uses a proximity sensor to measure the thickness of rolled steel (steel gage) at every two feet along the sheet, and the mill controller adjustments are made on the basis of the last twenty measurements. Specifically, the controller is adjusted unless the probability that the mean thickness lies within ±1 percent of the sample mean exceeds 0.99.

A typical set of twenty measurements (in millimeters) is as follows:

5.10, 5.05, 4.94, 4.98, 5.10, 5.12, 5.07, 4.96, 4.99, 4.95,
4.99, 4.97, 5.00, 5.08, 5.10, 5.11, 4.99, 4.96, 4.90, 4.10,

Check whether adjustments would be made in the gage controller on the basis of these measurements.

em.

cripts o
ons for
operat-
: F_1 and

experi-
id mea-
asured.
iod, by
model
iven by

ent, ±1
it for *b*.

e
e.

: (SRSS)
ent error

ed are in
dard de-

measure
the mill
rements.
thickness

95,
10,

s of these

2.24. Consider a mechanical component whose response x is governed by the relationship

$$f = f(x, \dot{x})$$

where f denotes applied (input) force and $\dot{x}$ denotes velocity. Three special cases are

(a) Linear spring:

$$f = kx$$

(b) Linear spring with a viscous (linear) damper:

$$f = kx + b\dot{x}$$

(c) Linear spring with coulomb frictinn:

$$f = kx + f_c\,\mathrm{sgn}(\dot{x})$$

Suppose that a harmonic excitation of the form

$$f = f_a \sin \omega t$$

is applied in each case. Sketch the force-displacement curves for the three cases at steady state. Which components exhibit hysteresis? Which components are nonlinear? Discuss your answers.

2.25. A tactile (distributed touch) sensor of a robotic manipulator gripper consists of a matrix of piezoelectric sensor elements placed 2 mm apart. Each element generates an electric charge when it is strained by an external load. Sensor elements are multiplexed at very high speed in order to avoid charge leakage and to read all data channels using a single high-performance charge amplifier. Load distribution on the surface of the tactile sensor is determined from the charge amplifier readings, since the multiplexing sequence is known. Each sensor element can read a maximum load of 50 N and can detect load changes on the order of 0.01 N.

(a) What is the spatial resolution of the tactile sensor?

(b) What is the load resolution (in N/m^2) of the tactile sensor?

(c) What is the dynamic range?

2.26. Four sets of measurements were taken on the same response variahle of a process using four different sensors. The true value of the response was known to be constant. Suppose that the four sets of data are as shown in figure P2.26a–d. Classify these data sets, and hence the corresponding sensors, with respect to precision and deterministic (repeatable) accuracy.

2.27. Dynamics and control of inherently unstable systems, such as rockets, can be studied experimentally using simple scaled-dnwn physical models of the prototype systems. One such study is the classic inverted pendulum problem. An experimental setup for the inverted pendulum is shown in figure P2.27. The inverted pendulum is supported on a trolley that is driven on a tabletop in a straight line, using a chain-and-sprocket transmission operated by a DC motor. The motor is turned by commands from a microprocessor that is interfaced with the control circuitry of the motor. The angular position of the pendulum rod is measured using a resolver and is transmitted to the microprocessor. A strategy of statistical process control is used to balance the pendulum rod. Specifically, control limits are established from an initial set of measurement samples of the pendulum angle. Subsequently, if the angle exceeds one control limit, the trolley is accelerated in the opposite direction, using an automatic command to the motor. The control limits are also updated regularly. Suppose that the fnllowing

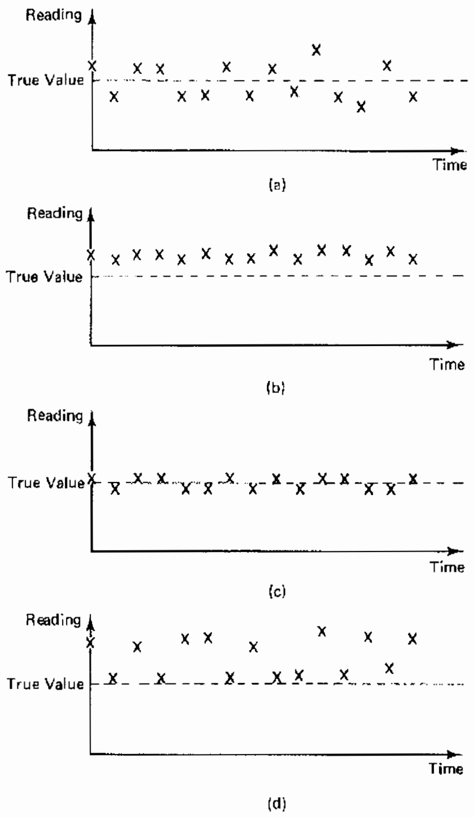


Figure P2.26. Four sets of measurements on the same response variable using different sensors.

twenty readings of the pendulum angle were measured (in degrees) after the system had operated for a few minutes:

0.5,	−0.5,	0.4,	−0.3,	0.3,	0.1,	−0.3,	0.3,	4.0,	0.0,
0.4,	−0.4,	0.5,	−0.5,	−5.0,	0.4,	−0.4,	0.3,	−0.3,	−0.1

Establish whether the system was in statistical control during the period in which the readings were taken. Comment on this method of control.

2.28. **(a)** What is a two-pnrt element? Discuss how dynamic coupling in a two-port device affects its accuracy when the device is used as a measuring device.

(b) What is the significance of bandwidth in a measuring device? Discuss methods to improve bandwidth.

(c) Using a two-port model for a DC tachometer, discuss methods of decreasing dynamic coupling and improving the useful frequency range.

nse

system

.0,

.1

hich the

t device

thods to

ising dy-

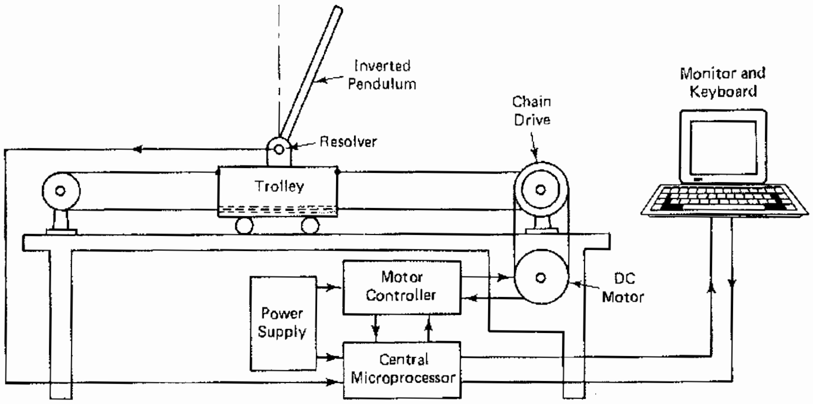


Figure P2.27. A microprocessor-controlled inverted pendulum—an application of statistical process control.

2.29. (a) Explain why mechanical loading error due to tachometer inertia can be significantly higher when measuring transient speeds than when measuring constant speeds.

(b) A DC tachometer has an equivalent resistance $R_a = 20\ \Omega$ in its rotor windings. In a position plus velocity servo system, the tachometer signal is connected to a feedback control circuit with equivalent resistance 2 kΩ. Estimate the percentage error due to electrical loading of the tachometer at steady state.

(c) If the conditinns were not steady, how would the electrical loading be affected in this application?

2.30. A single-degree-of-freedom model of a mechanical manipulator is shown in figure P2.30a. The joint motor has rotor inertia J_m. It drives an inertial load that has moment of inertia J_ℓ through a speed reducer of gear ratio $1 : r$ (Note: $r < 1$). The cnntrol scheme used in this system is the so-called feedforward control (strictly, *computed-torque control*) methnd. Specifically, the motor torque T_m that is required to accelerate or decelerate the load is computed using a suitable dynamic model and a desired motion trajectory for the manipulator, and the mntor windings are excited so as to generate that torque. A typical trajectory would consist of a constant angular acceleration segment followed by a constant angular velocity segment, and finally a constant deceleration segment, as shown in figure P2.30b.

(a) Neglecting friction (particularly bearing friction) and inertia of the speed reducer, show that a dynamic model for torque computation during accelerating and decelerating segments of the motion trajectory would be

$$T_m = (J_m + r^2 J_\ell)\ddot{\theta}_\ell / r$$

where $\ddot{\theta}_\ell$ is the angular acceleration of the load, hereafter denoted by α_ℓ. Show that the overall system can be modeled as a single inertia rotating at the motor speed. Using this result, discuss the effect of gearing on a mechanical drive.

(b) Given that $r = 0.1$, $J_m = 0.1$ kg m^2, $J_\ell = 1.0$ kg m^2, and $\alpha_\ell = 5.0$ rad/s^2, estimate the allowable error for these four quantities so that the combined error in the

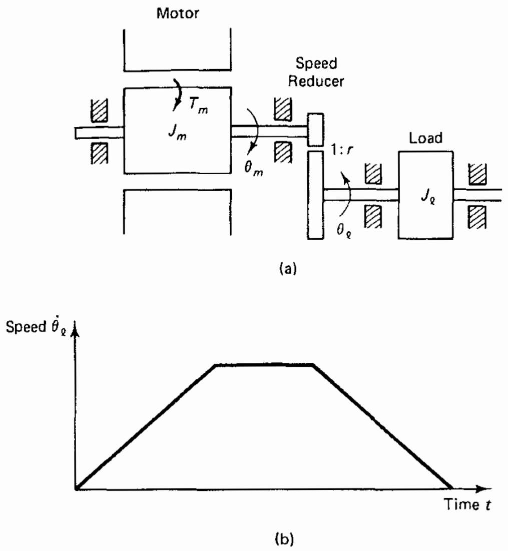


Figure P2.30. (a) A single-degree-of-freedom model of a mechanical manipulator. (b) A typical reference (desired) speed trajectory for computed-torque control.

computed torque is limited to ±1 percent and so that each of the four quantities contributes equally toward this error in computed T_m. Use the absolute value method for error combination.

(c) Arrange the four quantities r, J_m, J_ℓ, and α_ℓ in the descending order of required accuracy for the numerical values given in the problem.

(d) Suppose that $J_m = r^2 J_\ell$. Discuss the effect of error in r on the error in T_m.

2.31. A useful rating parameter for a mechanical tool is *dexterity*. Though not complete, an appropriate analytical definition for dexterity of a device is

$$\text{dexterity} = \frac{\text{number of degrees of freedom}}{\text{motion resolution}}$$

where the number of degrees of freedom is equal to the number of independent variables that is required to completely define an arbitrary position increment of the tool (i.e., for an arbitrary change in its kinematic configuration).

(a) Explain the physical significance of dexterity and give an example of a device for which the specification of dexterity would be very important.

(b) The power rating of a tool may be defined as the product of maximum force that can be applied by it in a controlled manner and the corresponding maximum speed. Discuss why the power rating of a manipulating device is usually related to the dexterity of the device. Sketch a typical curve of power versus dexterity.

2.32. Resolution of a feedback sensor (or resolution of a response measurement used in feedback) has a direct effect on the accuracy that is achievable in a control system. This is true because the controller cannot correct a deviation of the response from the desired value (set point) unless the response sensor can detect that change. It follows that the resolution of a feedback sensor will govern the minimum (best) possible deviation band of system response under feedback control. An angular position servo uses a resolver as its feedback sensor. If peak-to-peak oscillations of the servo load under steady-state conditions have to be limited to no more than two degrees, what is the worst tolerable resolution of the resolver? Note that, in practice, the feedback sensor should have a resolution better (smaller) than this worst value.

2.33. An actuator (e.g., electric motor, hydraulic piston-cylinder) is used to drive a terminal device (e.g., gripper, hand, wrist with active remote center compliance) of a robotic manipulator. The terminal device functions as a force generator. A schematic diagram for the system is shown in figure P2.33. Show that the displacement error e_x is related to the force error e_f through

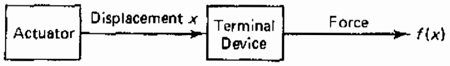


Figure P2.33. Block diagram for a terminal device of a robotic manipulator.

$$e_f = \frac{x}{f}\frac{df}{dx}e_x$$

The actuator is known to be 100 percent accurate for practical purposes, but there is an initial position error δx_o (at $x = x_o$). Obtain a suitable transfer relation $f(x)$ for the terminal device so that the force error e_f remains constant throughout the dynamic range of the device.

2.34. Consider, again, the mechanical tachometer shown in figure 2.10 (example 2.3). Write expressions for sensitivy and bandwidth for the device. Using the example, show that the two performance ratings, sensitivity and bandwidth, generally conflict. Discuss ways to improve the sensitivity of this mechanical tachometer.

REFERENCES

Brignell, J. E., and Rhodes, G. M. *Laboratory On-Line Computing*. Wiley, New York, 1975.

Brdch, J. T. *Mechanical Vibration and Shock Measurements*. Brüel and Kjaer, Naerum, Denmark, 1980.

Crandall, S. H., Karnopp, D. C., Kurtz, E. F., Jr., and Pridmore-Brown, D. C. *Dynamics of Mechanical and Electromechanical Systems*. McGraw-Hill, New York, 1968.

Dally, J. W., Riley, W. F., and McConnell, K. G. *Instrumentation for Engineering Measurements*. Wiley, New York, 1984.

deSilva, C. W. *Dynamic Testing and Siesmic Qualification Practice*. Lexington Books, Lexington, Mass., 1983.

———. "Motion Sensors in Industrial Robots." *Mechanical Engineering* 107(6): 40–51, June 1985.

Doebelin, E. O. *Measurement Systems*, 3d ed. McGraw-Hill, New York, 1983.

Harris, C. M., and Crede, C. E. *Shock and Vibration Handbook*, 2d ed. McGraw-Hill, New York, 1976.

Herceg, E. E. *Handbook of Measurement and Control*. Schaevitz Engineering, Pennsauken, N.J., 1972.

Rosenberg, R. C., and Karnopp, D. C. *Introduction to Physical Systems Dynamics*. McGraw-Hill, New York, 1983.

5

Digital Transducers

5.1 INTRODUCTION

Any transducer that presents information as discrete samples and that does not introduce a *quantization error* when the reading is represented in the digital form may be classified as a digital transducer. A digital processor plays the role of controller in a digital control system. This facilitates complex processing of measured signals and other known quantities in order to obtain control signals for the actuators that drive the plant of the control system. If the measured signals are in analog form, an analog-to-digital conversion (ADC) stage is necessary prior to digital processing. There are several other shortcomings of analog signals in comparison to digital signals, as outlined in chapter 1. These considerations help build a case in favor of direct digital measuring devices for digital control systems.

Digital measuring devices (or digital transducers, as they are commonly known) generate discrete output signals such as pulse trains or encoded data that can be directly read by a control processor. Nevertheless, the sensor stage of digital measuring devices is usually quite similar to that of their analog counterparts. There are digital measuring devices that incorporate microprocessors to perform numerical manipulations and conditioning locally and provide output signals in either digital or analog form. These measuring systems are particularly useful when the required variable is not directly measurable but could be computed using one or more measured outputs (e.g., power = force × speed). Although a microprocessor is an integral part of the measuring device in this case, it performs not a measuring task but, rather, a conditioning task. For our purposes, we shall consider the two tasks separately.

The objective of this chapter is to study the operation and utilization of several types of direct digital transducers. Our discussion will be limited to motion transducers. Note, however, that by using a suitable auxiliary front-end sensor, other measurands—such as force, torque, and pressure—may be converted into a motion and subsequently measured using a motion transducer. For example, altitude (or pres-

; not intro-
rm may be
troller in a
signals and
; that drive
; form, an
processing.
digital sig-
avor of di-

commonly
ta that can
: of digital
arts. There
numerical
r digital or
e required
more mea-
: is an inte-
g task but,
tasks sepa-

of several
n transduc-
other mea-
motion and
le (or pres-

sure) measurements in aircraft and aerospace applications are made using a pressure-sensing front end, such as a bellows or diaphragm device, in conjunction with an optical encoder to measure the resulting displacement. Motion, as manifested in physical systems, is typically continuous in time. Therefore, we cannot speak of digital motion sensors in general. Actually, it is the transducer stage that generates the discrete output signal in a digital motion measuring device. Commercially available *direct digital transducers* are not as numerous as analog sensors, but what is available has found extensive application.

When the output of a digital transducer is a pulse signal, a counter is used either to count the pulses or to count clock cycles over one pulse duration. The count is first represented as a digital word according to some code; then it is read by a data acquisition and control computer. If, on the other hand, the output of digital transducer is automatically available in a coded form (e.g., binary, binary-coded decimal, ASCII), it can be directly read by a computer. In the latter case, the coded signal is normally generated by a parallel set of pulse signals; the word depends on the pattern of the generated pulses.

5.2 SHAFT ENCODERS

Any transducer that generates a coded reading of a measurement can be termed an encoder. Shaft encoders are digital transducers that are used for measuring angular displacements and angular velocities. Applications of these devices include motion measurement in performance monitoring and control of robotic manipulators, machine tools, digital tape-transport mechanisms, servo plotters and printers, satellite mirror positioning systems, and rotating machinery such as motors, pumps, compressors, turbines, and generators. High resolution (depending on the word size of the encoder output and the number of pulses per revolution of the encoder), high accuracy (particularly due to noise immunity of digital signals and superior construction), and relative ease of adaption in digital control systems (because transducer output is digital), with associated reduction in system cost and improvement of system reliability, are some of the relative advantages of digital transducers over their analog counterparts.

Encoder Types

Shaft encnders can be classified into two categories, depending on the nature and the method of interpretation of the transducer output: (1) incremental encoders and (2) absolute encoders. The output of an incremental encoder is a pulse signal that is generated when the transducer disk rotates as a result of the motion that is being measured. By counting the pulses or by timing the pulse width using a clock signal, both angular displacement and angular velocity can be determined. Displacement, however, is obtained with respect to some reference point on the disk, as indicated by a reference pulse (index pulse) generated at that location on the disk. The index pulse count determines the number of full revolutions.

An absolute encoder (or whole-word encoder) has many pulse tracks on its transducer disk. When the disk of an absolute encoder rotates, several pulse trains—equal in number to the tracks on the disk—are generated simultaneously. At a given instant, the magnitude of each pulse signal will have one of two signal levels (i.e., a binary state), as determined by a level detector. This signal level corresponds to a binary digit (0 or 1). Hence, the set of pulse trains gives an encoded binary number at any instant. The pulse windows on the tracks can be organized into some pattern (code) so that each of these binary numbers corresponds to the angular position of the encoder disk at the time when the particular binary number is detected. Furthermore, pulse voltage can be made compatible with some form of digital logic (e.g., transistor-to-transistor logic, or TTL). Consequently, the direct digital readout of an angular position is possible, thereby expediting digital data acquisition and processing. Absolute encoders are commonly used to measure fractions of a revolution. However, complete revolutions can be measured using an additional track that generates an index pulse, as in the case of incremental encoder.

The same signal generation (and pick-off) mechanism may be used in both types of transducers. Four techniques of transducer signal generation can be identified:

1. Optical (photosensor) method
2. Sliding contact (electrical conducting) method
3. Magnetic saturation (reluctance) method
4. Proximity sensor method

For a given type of encoder (incremental or absolute), the method of signal interpretation is identical for all four types of signal generation. Thus, we shall describe the principle of signal generation for all four mechanisms, but we will consider only the optical encoder in the context of signal interpretation and processing.

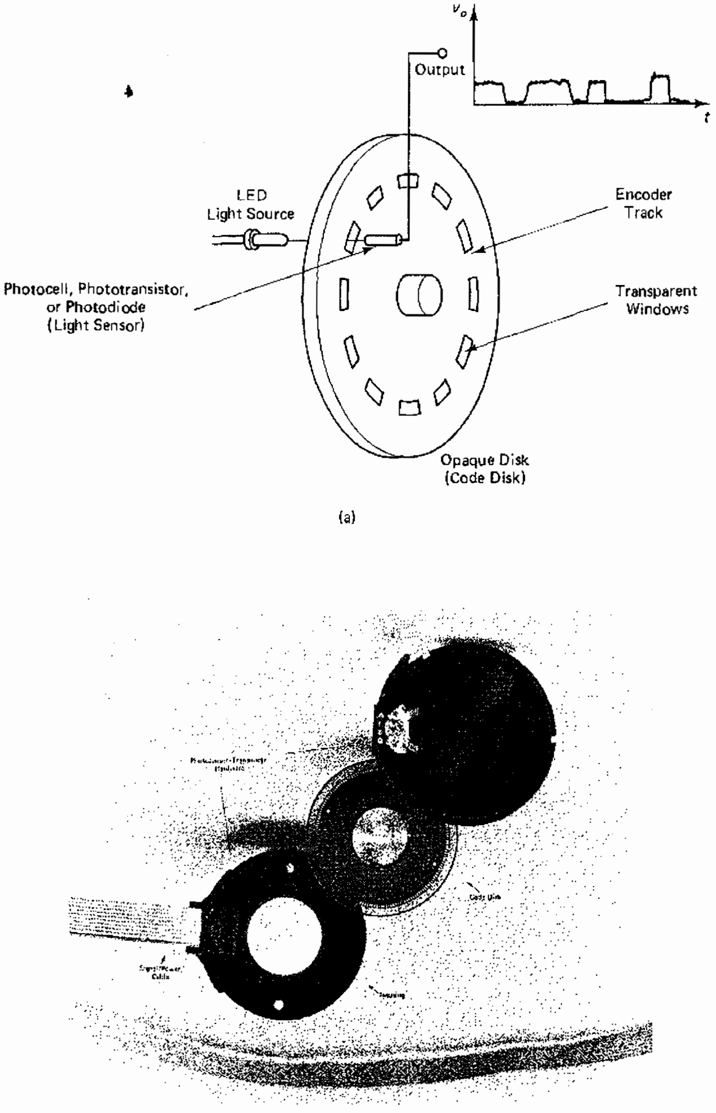
The optical encoder uses an opaque disk (code disk) that has one or more circular tracks, with some arrangement of identical transparent windows (slits) in each track. A parallel beam of light (e.g., from a set of light-emitting diodes or a tungsten lamp) is projected to all tracks from one side of the disk. The transmitted light is picked off using a bank of photosensors on the other side of the disk that typically has one sensor for each track. This arrangement is shown in figure 5.1a, which indicates just one track and one pick-off sensor. The light sensor could be a silicon photodiode, a phototransistor, or a photovoltaic cell. Since the light from the source is interrupted by the opaque areas of the track, the output signal from the probe is a series of voltage pulses. This signal can be interpreted to obtain the angular position and angular velocity of the disk. Note that in the standard terminology, the sensor element of such a measuring device is the encoder disk that is coupled to the rotating object (directly or through a gear mechanism). The transducer stage is the conversion of disk motion into the pulse signals. The opaque background of transparent windows (the window pattern) on an encoder disk is produced by contact printing index pulse

acks on its
lse trains—
. At a given
vels (i.e., a
sponds to a
ary number
ome pattern
position of
d. Further-
logic (e.g.,
adout of an
nd process-
revolution.
k that gen-

sed in both
on can be

al interpre-
describe the
der only the

more circu-
its) in each
r a tungsten
tted light is
at typically
which indi-
e a silicon
the source
e probe is a
lar position
, the sensor
the rotating
the conver-
transparent
act printing



(a)

(b)

Figure 5.1. (a) Schematic representation of an optical encoder. (b) Components of an incremental optical encoder (courtesy of Hewlett-Packard Conmpany).

techniques. The precision of this production procedure is a major factor that determines the accuracy of optical encoders. Note that a transparent disk with opaque spots will work equally well as the encoder disk of an optical encoder. The code disk, housing, and signal/power cable of a commercially available incremental optical encoder are shown in figure 5.1b.

In a sliding contact encoder, the transducer disk is made of an electrically insulating material (see figure 5.2). Circular tracks on the disk are formed by implanting a pattern of conducting areas. These conducting regions correspond to the transparent windows on an optical encoder disk. All conducting areas are connected to a common slip ring on the encoder shaft. A constant voltage v_{ref} is applied to the slip ring using a brush mechanism. A sliding contact such as a brush touches each track, and as the disk rotates, a voltage pulse signal is picked off by it (see figure 5.2). The pulse pattern depends on the conducting–nonconducting pattern on each track as well as the nature of rotation of the disk. The signal interpretation is done as it is for optical encoders. The advantages of sliding contact encoders include high sensitivity (depending on the supply voltage) and simplicity of construction (low cost). The disadvantages include the familiar drawbacks of contacting and commutating devices (e.g., friction, wear, brush bounce due to vibration, and signal glitches and metal oxidation due to electrical arcing). A transducer's accuracy is very much dependent upon the precision of the conducting patterns of the encoder disk. One method of generating the conducting pattern on the disk is electroplating.

Magnetic encoders have high-strength magnetic areas imprinted on the encoder disk using techniques such as etching, stamping, or recording (similar to tape recording). These magnetic areas correspond to the transparent windows on an opti-

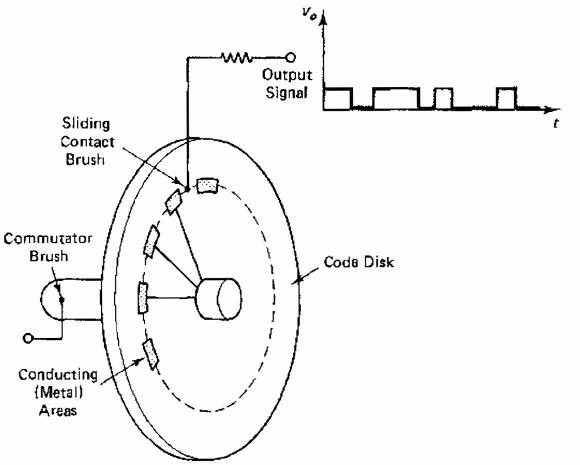


Figure 5.2. Schematic representation of a sliding contact encoder.

cal encoder disk. The signal pick-off device is a microtransformer that has primary and secondary windings on a circular ferromagnetic core. This pick-off sensor resembles a core storage element in older mainframe computers. The encoder arrangement is illustrated schematically in figure 5.3. A high-frequency (typically 100 kHz) primary voltage induces a vnltage in the secondary winding of the sensing element at the same frequency, operating as a transformer. A magnetic field of sufficient strength can saturate the core, however, thereby significantly increasing the reluctance and dropping the induced voltage. By demodulating the induced voltage, a pulse signal is obtained. This signal can be interpreted in the usual manner. Note that a pulse peak correspnnds to a nonmagnetic area and a pulse valley corresponds to a magnetic area on each track. Magnetic encoders have noncontacting pick-off sensors, which is an advantage. They are more costly than the contacting devices, however, primarily because of the cost of transformer elements and demodulating circuitry for generating the output signal.

Proximity sensor encoders use a proximity sensor as the signal pick-off element. Any type of proximity sensor may be used—for example, a magnetic induction probe or an eddy current probe, as discussed in chapter 3. In the magnetic induction probe, for example, the disk is made of ferromagnetic material. The encoder tracks have raised spots of the same material (see figure 5.4), serving a purpose analogous to that of the windows on an optical encoder disk. As a raised spot approaches the probe, the flux linkage increases as a result of the associated decrease in reluctance, thereby raising the induced voltage level. The output voltage is a pulse-modulated signal at the frequency of the supply (primary) voltage of the proximity sensor. This is then demodulated, and the resulting pulse signal is interpreted. In principle, this device operates like a conventional digital tachometer. If an eddy current probe is used, pulse areas in the track are plated with a conducting material. A flat plate may be used in this case, because the nonconducting areas on the disk do not generate eddy currents.

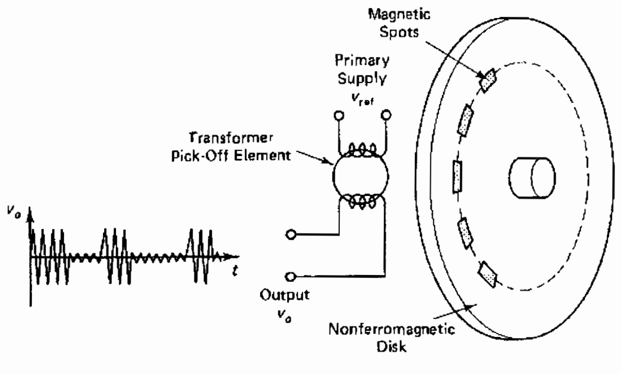


Figure 5.3. Schematic representation of a magnetic encoder.

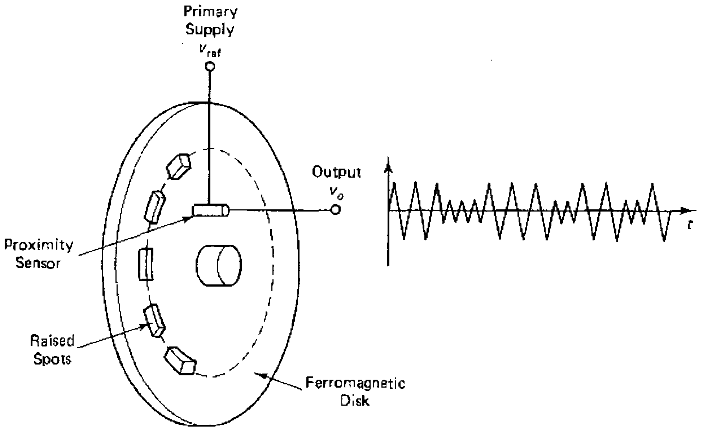


Figure 5.4. Schematic representation of a proximity probe encoder.

Note that an incremental encoder disk requires only one primary track that has equally spaced and identical window (pick-off) areas. The window area is equal to the area of the interwindow gap. Usually, a reference track that has just one window is also present in order to generate a pulse (known as the index pulse) to initiate pulse counting for angular position measurement and to detect complete revolutions. In contrast, absolute encoder disks have several rows of tracks, equal in number to the bit size of the output data word. Furthermore, the track windows are not equally spaced but are arranged in a specific pattern on each track so as to obtain a binary code (or a gray code) for the output data from the transducer. It follows that absolute encoders need at least as many signal pick-off sensors as there are tracks, whereas incremental encoders need one pick-off sensor to detect the magnitude of rotation and an additional sensor at a quarter-pitch separation (pitch = center-to-center distance between adjacent windows) to identify the direction of rotation. Some designs of incremental encoders have two identical tracks, one a quarter-pitch offset from the other, and the two pick-off sensors are placed radially without offset. A pick-off sensor for receiving a reference pulse is also used in some designs of incremental encoders (three-track incremental encoders).

In many control applications, encoders are built into the plant itself, rather than being externally fitted onto a rotating shaft. For instance, in a robot arm, the encoder might be an integral part of the joint motor and may be located within its housing. This reduces coupling errors (e.g., errors due to backlash, shaft flexibility, and resonances added by the transducer and fixtures), installation errors (e.g., eccentricity), and overall cost.

Since the signal interpretation techniques are quite similar for the various types of encoder signal generation techniques, we shall limit further discussion to optical encoders. These are the predominantly employed types of shaft encoders in practical applications. Signal interpretation depends on whether the particular optical encoder is an incremental device or an absolute device.

5.3 INCREMENTAL OPTICAL ENCODERS

There are two possible configurations for an incremental encoder disk: (1) the offset sensor configuration and (2) the offset track configuration. The first configuration is shown schematically in figure 5.5. The disk has a single circular track with identical and equally spaced transparent windows. The area of the opaque region between adjacent windows is equal to the window area. Two photodiode sensors (pick-offs 1 and 2 in figure 5.5) are positioned facing the track a quarter-pitch (half the window length) apart. The ideal forms of their output signals (v_1 and v_2) after passing them through pulse-shaping circuitry are shown in figure 5.6a and 5.6b for the two directions of rotation.

In the second configuration of incremental encoders, two identical tracks are used, one offset from the other by a quarter-pitch. In this case, one pick-off sensor is positioned facing each track—on a radial line, without any circumferential offset—unlike the previous configuration. The output signals from the two sensors are the same as before, however (figure 5.6).

In both configurations, an additional track with a lone window and associated sensor is also usually available. This track generates a reference pulse (index pulse) per revolution of the disk (see figure 5.6c). This pulse is used to initiate the counting

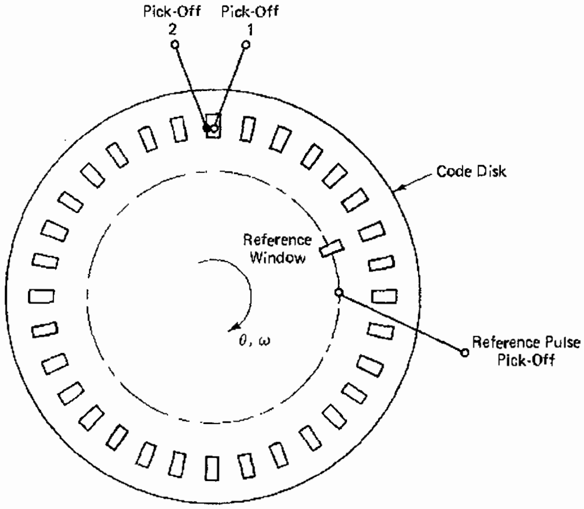


Figure 5.5. An incremental encoder disk (offset sensor configuration).

has
d to
dow
iate
ons.
r to
ally
iary
lute
reas
tion
dis-
igns
om
-off
en-

her
the
. its
ity,
ec-

pes
ical
ical

). 5

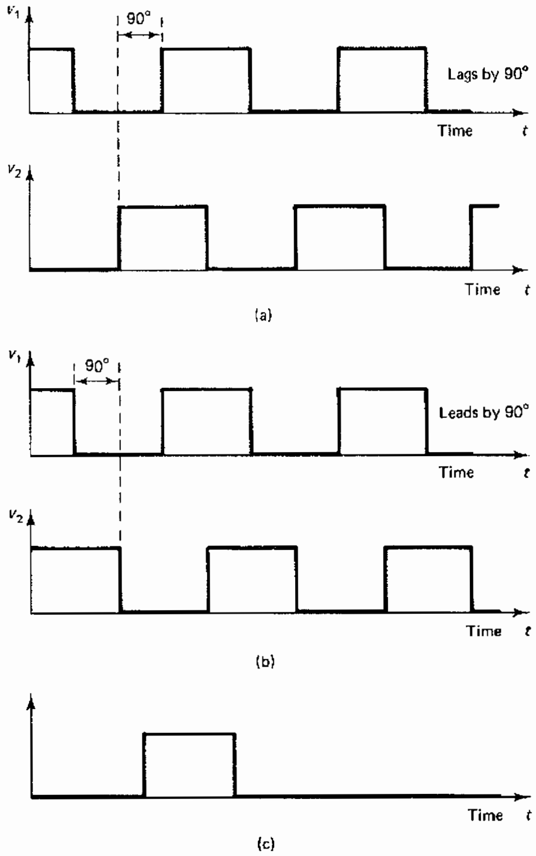


Figure 5.6. Shaped pulse signals from an incremental encoder: (a) for clockwise rotation; (b) for counterclockwise rotation; (c) reference pulse signal.

operation. Furthermore, the index pulse count gives the number of complete revolutions, which is required in absolute angular rotation measurements. Note that the pulse width and pulse-to-pulse period (encoder cycle) are constant in each sensor output when the disk rotates at constant angular velocity. When the disk accelerates, the pulse width decreases continuously; when the disk decelerates, the pulse width increases.

Direction of Rotation

The quarter-pitch offset in sensor location or track position is used to determine the direction of rotation of the disk. For example, figure 5.6a shows idealized sensor outputs (v_1 and v_2) when the disk rotates in the clockwise direction; and figure 5.6b shows the sensor outputs when the disk rotates in the counterclockwise direction. It is clear from these two figures that in clockwise rotation, v_1 lags v_2 by a quarter of a cycle (i.e., a phase lag of 90°); and in counterclockwise rotation, v_1 leads v_2 by a quarter of a cycle. Hence, the direction of rotation is obtained by determining the phase difference of the two output signals, using phase-detection circuitry.

One method for determining the phase difference is to time the pulses using a high-frequency clock signal. For example, if the counting (timing) operation is initiated when the v_1 signal begins to rise, and if n_1 = number of clock cycles (time) until v_2 begins to rise and n_2 = number of clock cycles until v_1 begins to rise again, then $n_1 > n_2 - n_1$ corresponds to clockwise rotation and $n_1 < n_2 - n_1$ corresponds to counterclockwise rotation. This should be clear from figures 5.6a and 5.6b.

Construction Features

The actual internal hardware of commercial encoders is not as simple as what is suggested by figure 5.5. (see figure 5.1b). A more detailed schematic diagram of the signal generation mechanism of an optical incremental encoder is shown in figure 5.7. The light generated by the light-emitting diode (LED) is collimated (forming parallel rays) using a lens. This pencil of parallel light passes through a window of the rotating code disk. The grating (masking) disk is stationary and has a track of windows identical to that in the code disk. A significant amount of light passes through the grating window only if it is aligned with a window of the code disk. Because of the presence of the grating disk, more than one window of the code disk may be illuminated by the same LED, thereby improving the intensity of light received by the photosensor but not introducing any error caused by the diameter of the pencil of light being larger than the window length. When the windows of the code disk face the opaque areas of the grating disk, virtually no light is received by the photosensor. Hence, as the code disk moves, alternating light and dark spots (a moiré pattern) are seen by the photosensor. Note that the grating disk helps increase

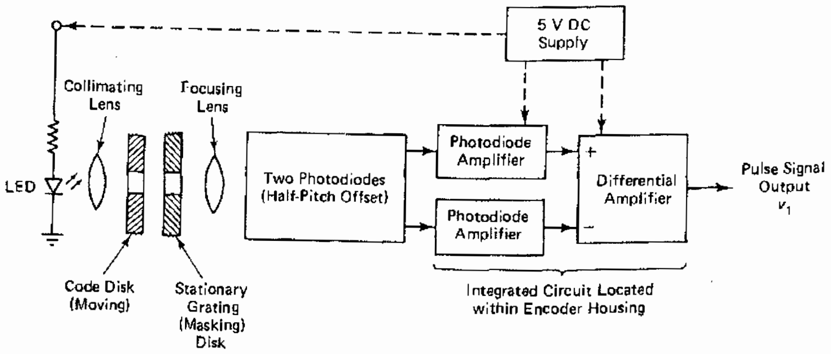


Figure 5.7. Internal hardware of an optical incremental encoder (for a single output pulse signal).

the output signal level significantly. But the supply voltage fluctuations also directly influence the light level received by the photosensor. If the sensitivity of the photosensor is not high enough, a low light level might be interpreted as no light, which would result in measurement error. Such errors due to instabilities and changes in the supply voltage can be eliminated by using two photosensors, one placed half a pitch away from the other along the window track. This arrangement should not be confused with the quarter-of-a-pitch offset arrangement that is required for direction detection. This arrangement is for contrast detection. The sensor facing the opaque region of the masking disk will always read a low signal. The other sensor will read either a high signal or a low signal, depending on whether it faces a window or an opaque region of the code disk. The two signals from these two sensors are amplified separately and fed into a differential amplifier. If the output is high, we have a pulse. In this manner, a stable and accurate output pulse signal can be obtained even under unstable voltage supply conditions. The signal amplifiers are integrated circuit devices and are housed within the encoder itself. Additional pulse-shaping circuitry may also be present. The power supply has to be provided separately as an external component. The voltage level and pulse width of the output pulse signal are logic-compatible (e.g., transistor-to-transistor logic, or TTL) so that they may be read directly using a digital board. The schematic diagram in figure 5.7 shows the generation of only one (v_1) of the two quadrature pulse signals. The other pulse signal (v_2) is generated using identical hardware but at a quarter of a pitch offset. The index pulse (reference pulse) signal is also generated in a similar manner. The cable of the encoder (usually a ribbon cable) has a multipin connector (see figure 5.1h). Three of the pins provide the three output pulse signals. Another pin carries the DC supply voltage (typically 5 V) from the power supply into the encoder. Note that the only moving part in the system shown in figure 5.7 is the code disk.

Displacement and Velocity Computation

A digital processor computes angular displacements and velocities using the digital data read into it from encoders, along with other pertinent parameters. To compute the angular position θ, suppose that the maximum count possible is M pulses and the range of the encoder is $\pm\theta_{max}$. Then the angle corresponding to a count of n pulses is

$$\theta = \frac{n}{M}\theta_{max} \tag{5.1}$$

Note that if the data size is r bits, allowing for a sign bit,

$$M = 2^{r-1} \tag{5.2}$$

where zero count is also included. Strictly speaking,

$$M = 2^{r-1} - 1$$

if zero count is not included. Note that if θ_{max} is 2π and $\theta_{min} = 0$, for example, then θ_{max} and θ_{min} will correspund to the same positiun of the code disk. To avoid this ambiguity, we use

$$\theta_{\min} = \frac{\theta_{\max}}{2^{r-1}}$$

Then equation 5.2 leads to the conventional definition for digital resolution:

$$\frac{(\theta_{\max} - \theta_{\min})}{(2^{r-1} - 1)}$$

Two methods are available for determining velocities using an incremental encoder: (1) the pulse-counting method and (2) the pulse-timing methnd. In the first method, the pulse count over the sampling period of the digital processor is measured and is used to calculate the angular velocity. For a given sampling period, there is a lower speed limit below which this method is not very accurate. In the second method, the time for one encoder cycle is measured using a high-frequency clock signal. This method is particularly suitable for measuring low speeds accurately.

To compute the angular velocity ω using the first method, suppose that the count during a sampling period T is n pulses. Hence, the average time for one pulse is T/n. If there are N windows on the disk, the average time for one revolution is NT/n. Hence,

$$\omega = \frac{2\pi n}{NT} \tag{5.3}$$

For the second method of velncity computation, suppose that the clock frequency is f Hz. If m cycles nf the clock signal are counted during an encoder period (interval between two adjacent windows), the time for that encoder cycle (i.e., the time to rntate through one encoder pitch) is given by m/f. With a total of N windows on the track, the average time for one revolution of the disk is Nm/f. Hence,

$$\omega = \frac{2\pi f}{Nm} \tag{5.4}$$

Note that a single incremental encoder can serve as both position sensor and speed sensor. Hence, a position loop and a speed loop in a control system can be closed using a single encoder, without having to use a conventional (analog) speed sensor such as a tachometer. The speed resolution of the encoder (depending on the methnd of speed computation—pulse counting or pulse timing) can be chosen to meet the accuracy requirements for the speed control loop. A further advantage of using an encoder rather than a conventional (analog) motion sensor is that an analog-to-digital converter (ADC) would be unnecessary. For example, the pulses generated by the encoder could be used as ***interrupts*** for the control computer. These interrupts are then directly counted (by an up/down counter or indexer) and timed (by a clock) within the control computer, thereby providing position and speed readings. Another way of interfacing an encoder to the control computer (without the need fnr an ADC) will be explained next.

Data Acquisition Hardware

A method for interfacing an incremental encoder to a digital processor (digital controller) is shown schematically in figure 5.8. The pulse signals are fed into an up/down counter that has circuitry to detect pulses (for example, by rising-edge detection or by level detection) and logic circuitry to determine the direction and to code the count. A pulse in one direction (say, clockwise will increment the count by one (an upcount), and a pulse in the opposite direction will decrement the count by one (a downcount). The coded count may be directly read by the processor through its input/output (I/O) board without the need for an ADC. The count is transferred to a latch buffer so that the measurement is read from the buffer rather than from the counter itself. This arrangement provides an efficient means of data acquisition because the counting process can continue without interruption while the count is being read by the processor from the latch buffer. The processor identifies various components in the measurement system using addresses, and this information is communicated to the individual components through the address bus. The start, end, and nature of an action (e.g., data read, clear the counter, clear the buffer) are communicated to various devices by the processor through the control bus. The processor can command an action to a component in one direction of the bus, and the component can respond with a message (e.g., job completed) in the opposite direction. The data (e.g., the count) are transmitted through the data bus. While the processor reads (samples) data from the buffer, the control signals guarantee that no data are transferred to that buffer from the counter. It is clear that the data acquisition consists of handshake operations between the processor and the auxiliary components. More than one encoder may be addressed, controlled, and read by the

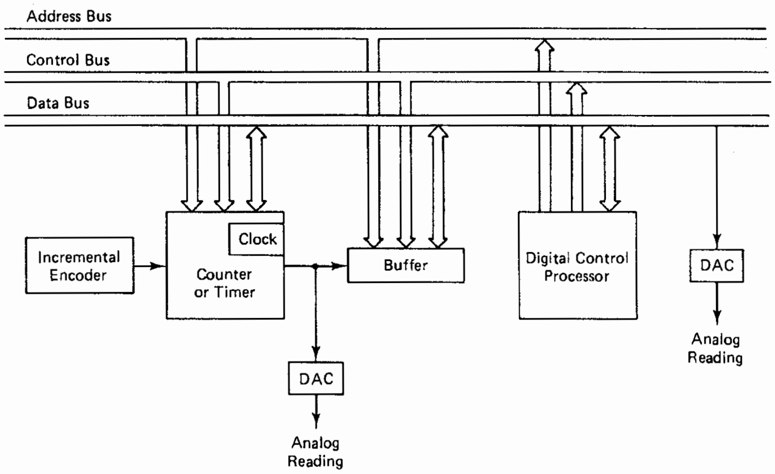
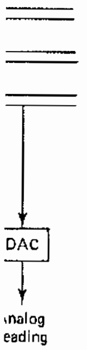


Figure 5.8. A data acquisition system for an incremental encoder.

gital con-
to an up/
ge detec-
d to code
nt by one
nt by one
rough its
erred to a
from the
sition be-
unt is be-
s various
mation is
he start,
uffer) are
The pro-
, and the
ite direc-
e the pro-
e that no
a acquisi-
ary com-
d by the



same three buses. The buses are typically multicore cables carrying signals in parallel logic. Slower communication in serial logic is also common.

In measuring position using an incremental encoder, the counter may be continuously monitored through a digital-to-analog converter (DAC in figure 5.8). However, the count is read by the processor only at every sampling instant. Since a cumulative count is required in displacement measurement, the buffer is not cleared once the count is read in by the processor.

For velocity measurement by the pulse-counting method, the buffer is read at intervals of T, which is also the counting-cycle time. The counter is cleared every time a count is transferred to the buffer, so that a new cnunt can begin. With this method, a new reading is available at every sampling period.

In the pulse-timing method of velocity computation, the counter is actually a timer. The encoder cycle is timed using a clock (internal or external), and the count is passed on tn the huffer. The counter is then cleared and the next timing cycle is started. The buffer is read by the processor periodically. With this method, a new reading is available at every encoder cycle. Note that under transient velocities, the encoder-cycle time is variable and is not directly related to the sampling period. Nevertheless, it is desirable to make the sampling period smaller than the encoder-cycle time, in general, so that no count is missed by the processor.

More efficient use of the digital processor may be achieved by using an interrupt routine. With this method, the counter (or buffer) sends an interrupt request to the processor when a new count is ready. The processor then temporarily suspends the current operation and reads in the new data. Note that the processor does not continuously wait for a reading in this case.

Displacement Resolution

The resolution of an encoder represents the smallest change in measurement that can be measured realistically. Since an encoder can be used to measure both displacement and velocity, we can identify a resolution for each case. Displacement resolution is governed by the number of windows N in the code disk and the digital size (number of bits) r of the buffer (counter output). The physical resolution is determined hy N. If only one pulse signal is used (i.e., no direction sensing), and if the rising edges of the pulses are detected (i.e., full cycles of the encoder are counted), the physical resolution is given by $(360/N)°$. But if both pulse signals (quadrature signals) are available and the capability to detect rising and falling edges of a pulse is also present, four counts can be made per encoder cycle, thereby improving the resolution by a factor of four. Hence, the physical resolution is given by

$$\Delta\theta_p = \frac{360°}{4N} \tag{5.5}$$

To understand this, note in figure 5.6a (or figure 5.6b) that when the two signals v_1 and v_2 are added, the resulting signal has a transition at every quarter of the encoder cycle. This is illustrated in figure 5.9. By detecting each transition (through edge detection or level detection), four pulses can be counted within every main cy-

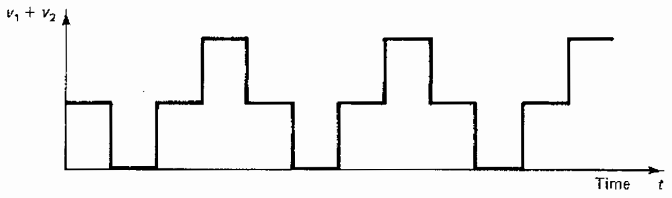


Figure 5.9. Quadrature signal addition to improve physical resolution.

cle. It should be mentioned that each signal (v_1 or v_2) separately has a resolution of half a pitch, provided that transitions (rising edges and falling edges) are detected and counted instead of pulses being counted. Accordingly a disk with 10,000 windows has a resolution of 0.018° if only one pulse signal is used (and both transitions, rise and fall, are detected). When both signals (with a phase shift of a quarter of a cycle) are used, the resolution improves to 0.009°. This resolution is achieved directly from the mechanics of the transducer; no interpolation is involved. It assumes, however, that the pulses are nearly ideal and, in particular, that the transitions are perfect. In practice, this cannot be achieved if the pulse signals are noisy; pulse shaping might be necessary.

Assuming that the maximum angle measured is 360° (or ±180°), the digital resolution is given by (see equations 5.1 and 5.2)

$$\Delta\theta_d = \frac{180°}{2^{r-1}} = \frac{360°}{2^r} \tag{5.6}$$

This should be clear, because a digital word containing r bits can represent 2^r different values (unsigned). As mentinned earlier, we have used $360°/2^r$ instead of $360°/(2^r - 1)$ in equation 5.6 for digital resolution, and this is further supported by the fact that 0° and 360° represent the same pnsition of the code disk. An ambiguity does not arise if we take the minimum value of θ to be $360°/2^r$, not zero. Then, by definition, the digital resolution is given by

$$\frac{(360° - 360°/2^r)}{(2^r - 1)}$$

This result is exactly the same as what is given by equation 5.6.

The larger of the two resolutions in equations 5.5. and 5.6 governs the displacement resolution of the encoder.

Example 5.1

For an ideal design of an incremental encoder, obtain an equation relating the parameters d, w, and r, where

d = diameter of encoder disk
w = number of windows per unit diameter of disk
r = word size (bits) of angle measurements

Assume that quadrature signals are available. If $r = 12$ and $w = 1{,}000$/in., determine a suitable disk diameter.

Solution In this problem, we are required to assign a word size for the resolution available from the number of windows. The position resolution by physical constraint (assuming that quadrature signals are available) is

$$\Delta\theta_p = \frac{1}{4}\left(\frac{360}{wd}\right)^\circ$$

The resolution available from the digital word size of the buffer is

$$\Delta\theta_d = \left(\frac{360}{2^r}\right)^\circ$$

Note that $\Delta\theta_p = \Delta\theta_d$ provides an ideal design. Hence,

$$\frac{1}{4}\frac{360}{\mathrm{wd}} = \frac{360}{2^r}$$

Simplifying, we have

$$wd = 2^{r-2}$$

with $r = 12$ and $w = 1{,}000$/in.,

$$d = \left(\frac{2^{12-2}}{1{,}000}\right) \text{ in.} = 1.024 \text{ in.}$$

The physical resolution of an encoder can be improved by using step-up gearing so that one rotation of the moving object that is being monitored corresponds to several rotations of the code disk of the encoder. This improvement is directly proportional to the gear ratio. Backlash in the gearing mechanism introduces a new error, however. For best results, this backlash error should be several times smaller than the resolution with no backlash. Note that the digital resolution is not improved by gearing, because the maximum angle of rotation of the moving object (say, 360°) still corresponds to the buffer size, and the change in the least significant bit (LSB) of the buffer corresponds to the same change in the angle of rotation of the moving object. In fact, the overall displacement resolution can be harmed in this case if excessive backlash is present.

Example 5.2

By using high-precision techniques to imprint window tracks on the code disk, it is possible to attain the window density of 1,000 windows/in. of diameter. Consider a 3,000-window disk. Suppose that step-up gearing is used to improve resolution and the gear ratio is 10. If the word size of the output buffer is 16 bits, examine the displacement resolution of this device.

Solution First consider the case in which gearing is not present. With quadrature signals, the physical resolution is

$$\Delta\theta_p = \frac{360^\circ}{4 \times 3{,}000} = 0.03^\circ$$

Now, for a range of measurement given by ±180°, a 16-bit output provides a digital resolution of

$$\Delta\theta_d = \frac{180^\circ}{2^{15}} = 0.005^\circ$$

Hence, in the absence of gearing, the overall displacement resolution is 0.03°. On the other hand, with a gear ratio of 10, and neglecting gear backlash, the physical resolution improves to 0.003°, but the digital resolution remains unchanged at best. Hence, the overall displacement resolution has improved to 0.005° as a result of gearing.

In summary, the displacement resolution of an incremental encoder depends on the following factors:

1. Number of windows on the code track (or disk diameter)
2. Gear ratio
3. Word size of the measurement buffer

The angular resolution of an encoder can be further improved, through interpolation, by adding equally spaced pulses in between every pair of pulses generated by the encoder circuit. These auxiliary pulses are not true measurements, and they can be interpreted as a linear interpolation scheme between true pulses. One method of accomplishing this interpolation is by using the two pick-off signals that are generated by the encoder (quadrature signals). These signals are nearly sinusoidal prior to shaping (say, by level detection). They can be filtered to obtain two sine signals that are 90° out of phase (i.e., a sine signal and a cosine signal). By weighted combination of these two signals, a series of sine signals can be generated such that each signal lags the preceding signal by any integer fraction of 360°. By level detection or edge detection (rising and falling edges), these sine signals can be converted into square wave signals. Then, by logical combination of the square waves, an integer number of pulses can be generated within each encoder cycle. These are the interpolation pulses that are added to improve the encoder resolution. In practice, about twenty interpolation pulses can be added between adjacent main pulses by this method.

Velocity Resolution

An incremental encoder is also a velocity-measuring device. The velocity resolution of an incremental encoder depends on the method that is employed to determine velocity. Since the pulse-counting method and the pulse-timing method are both based on counting, the resolution corresponds to the change in angular velocity that results from changing (incrementing or decrementing) the count by one. If the pulse-counting method is employed, it is clear from equation 5.3 that a unity change in the count n corresponds to a speed change of

$$\Delta\omega_c = \frac{2\pi}{NT} \qquad (5.7)$$

a digital

. On the
ıl resolu-
. Hence,
ing.

depends

ɡh inter-
enerated
and they
: method
are gen-
dal prior
e signals
d combi-
that each
ection or
rted into
n integer
: interpo-
ɔe, about
; by this

esolution
'mine ve-
ɔth based
ıat results
se-count-
ge in the

(5.7)

where N is the number of windows in the code track and T is the sampling period. Equation 5.7 gives the velocity resolution by this method. Note that this resolution is independent of the angular velocity itself. The resolution improves, however, with the number of windows and the sampling period. But under transient conditions, the accuracy of a velocity reading decreases with increasing T (the sampling frequency has to be at least double the highest frequency of interest in the velocity signal). Hence, the sampling period should not be increased indiscriminately.

If the pulse-timing method is employed, the velocity resolution is given by (see equation 5.4)

$$\Delta\omega_t = \frac{2\pi f}{Nm} - \frac{2\pi f}{N(m+1)} = \frac{2\pi f}{Nm(m+1)} \tag{5.8}$$

where f is the clock frequency. For large m, $(m + 1)$ can be approximated by m. Then, by substituting equation 5.4 in 5.8, we get

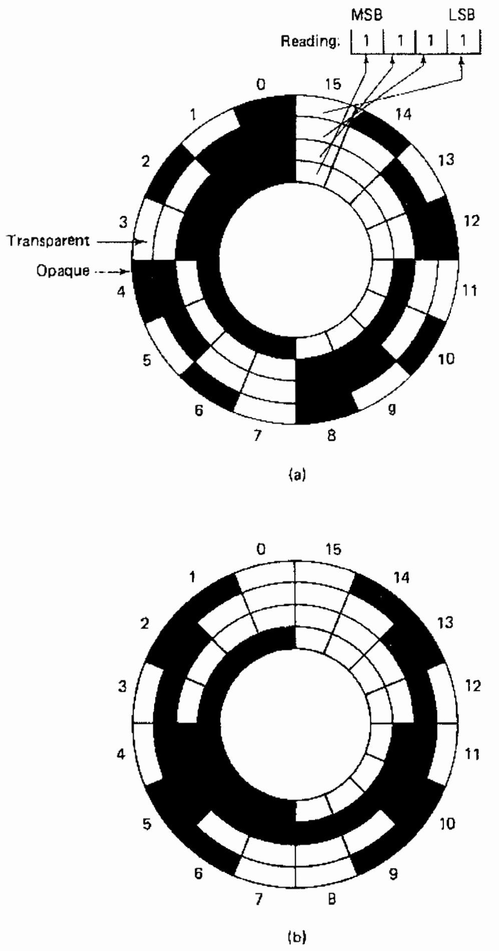
$$\Delta\omega_t = \frac{N\omega^2}{2\pi f} \tag{5.9}$$

Note that in this case, the resolution degrades quadratically with speed. This observation confirms the previnus suggestion that the pulse-timing method is appropriate for low speeds. For a given speed, the resolution degrades with increasing N. The resolution can be improved, however, by increasing the clock frequency. Gearing up has a detrimental effect on the speed resolution in the pulse-timing method, but it has a favorable effect in the pulse-cnunting method (see problem 5.6). In summary, the speed resolution of an incremental encoder depends on the following factors:

1. Number of windows N
2. Sampling period T
3. Clock frequency f
4. Speed ω
5. Gear ratio

5.4 ABSOLUTE OPTICAL ENCODERS

Absolute encoders directly generate coded data to represent angular positions using a series of pulse signals. No pulse counting is involved in this case. A simplified code pattern on an absolute encoder disk that utilizes the direct binary code is shown in figure 5.10a. The number of tracks (n) in this case is 4, but in practice n is on the order of 14. The disk is divided into 2^n sectors. Each partitioned area of the matrix thus formed corresponds to a hit of data. For example, a transparent area may correspond to binary 1 and an opaque area to binary 0. Each track has a pick-off sensor similar to those used in incremental encoders. The set of n pick-off sensors is arranged on a radial line and facing the tracks on one side of the disk. A light source (e.g., light-emitting diode or LED) illuminates the other side of the disk. As the disk rotates, the bank of pick-off sensors generates a set of pulse signals. At a given in-



(a)

(b)

Figure 5.10. Schematic diagram of an absolute encoder disk pattern: (a) binary code; (b) gray code.

stant, the combination of signal levels will provide a coded data word that uniquely determines the position of the disk at that time, with a resolution given by the sector angle. In figure 5.10a, the word size of the data is 4 bits. This can represent decimal numbers from 0 to 15, as given by the sixteen sectors of the disk. In each sector, the outermost element is the least significant bit (LSB) and the innermost element is the most significant bit (MSB). The direct binary representation of the disk sectors (position) is given in table 5.1. The angular resolution for this simplified example is $(360/2^4)°$, or 22.5°. If $n = 14$, the angular resolution improves to $(360/2^{14})°$, or 0.022°.

.agram of an
n: (a) binary

at uniquely
· the sector
nt decimal
sector, the
nent is the
ectors (po-
example is
$0/2^{14})°$, or

TABLE 5.1 SECTOR CODING FOR THE ABSOLUTE ENCODER EXAMPLE

Sector number	Straight binary code (MSB → LSB)	Gray code
0	0 0 0 0	0 1 1 1
1	0 0 0 1	0 1 1 0
2	0 0 1 0	0 1 0 0
3	0 0 1 1	0 1 0 1
4	0 1 0 0	0 0 0 1
5	0 1 0 1	0 0 0 0
6	0 1 1 0	0 0 1 0
7	0 1 1 1	0 0 1 1
8	1 0 0 0	1 0 1 1
9	1 0 0 1	1 0 1 0
10	1 0 1 0	1 0 0 0
11	1 0 1 1	1 0 0 1
12	1 1 0 0	1 1 0 1
13	1 1 0 1	1 1 0 0
14	1 1 1 0	1 1 1 0
15	1 1 1 1	1 1 1 1

Gray Coding

There is a data interpretation problem associated with the straight binary code in absolute encoders. Notice in table 5.1 that in straight binary, the transition from one sector to the adjacent sector may need more than one switching of bits of the binary data. For example, the transition from 0011 to 0100 or from 1011 to 1100 requires three bit switchings, and the transition from 0111 to 1000 or from 1111 to 0000 requires four bit switchings. If the pick-off sensors are not properly aligned along a radius of the encoder disk, or if excessive manufacturing error tolerances were allowed in imprinting the code pattern on the disk, or if environmental effects have resulted in large irregularities in the sector matrix, then bit switching will not take place simultaneously. This results in ambiguous readings during the transition period. For example, in changing from 0011 to 0100, if the LSB switches first, the reading becomes 0010. In decimal form, this incorrectly indicates that the rotation was from angle 3 to angle 2, whereas, it was actually a rotation from angle 3 to angle 4. Such ambiguities can be avoided by using a gray code, as shown in figure 5.10b for this example. The coded representation of the sectors is given in table 5.1. Note that in this case, each adjacent transition involves only one bit switching. A disadvantage of utilizing a gray code is that it requires additional logic to convert the gray-coded number to the corresponding binary number.

As with incremental encoders, the resolution of an absolute encoder can be improved by interpolation using auxiliary pulses. This requires an interpolation track and two pick-off sensors placed a quarter-pitch apart. This is equivalent to having an incremental encoder and an absolute encoder in the samc unit. The resolutiun is limited by the word size of the output data. Step-up gear mechanisms can also be employed to improve encoder resolution.

Absolute encoders can be used for angular velocity measurement as well. For this, either the pulse-timing method or the angle-measurement method may be used. With the first method, the interval between two consecutive readings is strobed (or timed) using a high-frequency strobe (clock) signal, as in the case of an incremental encoder. Typical strobing frequency is 1 MHz. The start and stop of strobing are triggered by the coded data from the encoder. The clock cycles are counted by a counter, as in the case of an incremental encoder, and the count is reset (cleared) after each counting cycle. The angular speed can be computed using these data, as shown earlier for an incremental encoder. With the second method, the change in angle is measured from one sample to the next, and the angular speed is computed as the ratio (angle change)/(sampling period).

Because the code matrix on the disk is more complex in an absolute encoder, and because more light sensors are required, an absolute encoder can be nearly twice as expensive as an incremental encoder. An absolute encoder does not require digital counters and buffers, however, unless data interpolation is done using an auxiliary track or pulse-timing is used for velocity calculation. Also, an absolute encoder has the advantage that if a reading is missed, it will not affect the next reading, whereas, a missed pulse in an incremental encoder would carry an error to the subsequent readings until the counter is cleared. Furthermore, incremental encoders have to be powered throughout operation of the system. Thus, a power failure can introduce an error unless the reading is reinitialized (calibrated). An absolute encoder must be powered *and* monitored only whcn a reading is taken.

5.5 ENCODER ERROR

Errors in shaft encoder readings can come from several factors. The primary sources of these errors are as follows:

1. Quantization error (due to digital word size limitations)
2. Assembly error (eccentricity, etc.)
3. Coupling error (gear backlash, belt slippage, loose fit, etc.)
4. Structural limitations (disk deformation and shaft deformation due to loading)
5. Manufacturing tolerances (errors from inaccurately imprinted code patterns, inexact positioning of the pick-off sensors, limitations and irregularities in signal generation and sensing components, etc.)
6. Ambient effects (vibration, temperature, light noise, humidity, dirt, smoke, etc.)

These factors can result in erroneous displacement and velocity readings and inexact direction detection.

One form of error in an encoder reading is the hysteresis. For a given position of the moving object, if the encoder reading depends on the direction of motion, the measurement has a hysteresis error. In that case, if the object rotates from position A to position B and back to position A, for example, the initial and the final readings of the encoder will not match. The causes of hysteresis include backlash in gear couplings, loose fits, mechanical deformation in the code disk and shaft, delays in electronic circuitry (electrical time constants), and noisy pulse signals that make the detection of pulses (say, by level detection or edge detection) less accurate.

The raw pulse signal from an optical encoder is somewhat irregular, primarily because of noise in the signal generation circuitry, including the noise created by imperfect light sources and photosensors. Noisy pulses have imperfect edges. As a result, pulse detection through edge detection can result in errors such as multiple triggering for the same edge of a pulse. This can be avoided by including a Schmitt trigger (a logic circuit with electronic hysteresis) in the edge-detection circuit, so that slight irregularities in the pulse edges will not cause erroneous triggering, provided that the noise level is within the hysteresis band of the trigger. A disadvantage of this method, however, is that hysteresis will be present even when the encoder itself is perfect. Virtually noise-free pulses can be generated if two photosensors are used to detect adjacent transparent and opaque areas on a track simultaneously and a separate circuit (a comparator) is used to create a pulse that depends on the sign of the voltage difference of the two sensor signals. (We described this method earlier. A schematic diagram of this arrangement is given in figure 5.7.)

Eccentricity Error

Eccentricity (denoted by e) of an encoder is defined as the distance between the center of rotation C of the code disk and the geometric center G of the circular code track. Nonzero eccentricity causes a measurement error known as the *eccentricity error*. The primary contributions to eccentricity are

1. Shaft eccentricity (e_s)
2. Assembly eccentricity (e_a)
3. Track eccentricity (e_t)
4. Radial play (e_p)

Shaft eccentricity results if the rotating shaft on which the code disk is mouoted is imperfect, so that its axis of rotation does not coincide with its geometric axis. Assembly eccentricity is caused if the code disk is improperly mounted on the shaft, so that the center of the code disk dnes not fall on the shaft axis. Track eccentricity comes from irregularities in the code track imprinting process, so that the center of the track circle does not coincide with the nominal geometric center of the disk. Radial play is caused by any looseness in the assembly in the radial direction. All four of these parameters are random variables that have mean values μ_s, μ_a, μ_t, and μ_p, respectively, and standard deviations σ_s, σ_a, σ_t, and σ_p, respectively. A very conservative upper bound for the mean value of the overall eccentricity is given by the sum of the individual mean values, each value being considered positive. A more reasonable estimate is provided by the *root-mean-square (rms)* value, as given by

$$\mu = \sqrt{\mu_s^2 + \mu_a^2 + \mu_t^2 + \mu_p^2} \tag{5.10}$$

Furthermore, assuming that the individual eccentricities are independent random variables, the standard deviation of the overall eccentricity is given by

$$\sigma = \sqrt{\sigma_s^2 + \sigma_a^2 + \sigma_t^2 + \sigma_p^2} \tag{5.11}$$

Knowing the mean value μ and the standard deviation σ of the overall eccentricity, it is possible to obtain a reasonable estimate for the maximum eccentricity that can occur. It is reasonable to assume that the eccentricity has a Gaussian (or normal) distribution, as shown in figure 5.11. The probability that the eccentricity lies between two given values is obtained by the area under the probability density curve within these two values (points) on the x-axis (also see chapter 2). In particular, for the normal distribution, the probability that the eccentricity lies within $\mu - 2\sigma$ and $\mu + 2\sigma$ is 95.5 percent, and the probability that the eccentricity falls within $\mu - 3\sigma$ and $\mu + 3\sigma$ is 99.7 percent. We can say, for example, that at a confidence level of 99.7 percent, the net eccentricity will not exceed $\mu + 3\sigma$.

Example 5.3

The mean values and the standard deviations of the four primary contributions to eccentricity in a shaft encoder are as follows (in millimeters):

Shaft eccentricity = (0.1, 0.01)
Assembly eccentricity = (0.2, 0.05)
Track eccentricity = (0.05, 0.001)
Radial play = (0.1, 0.02)

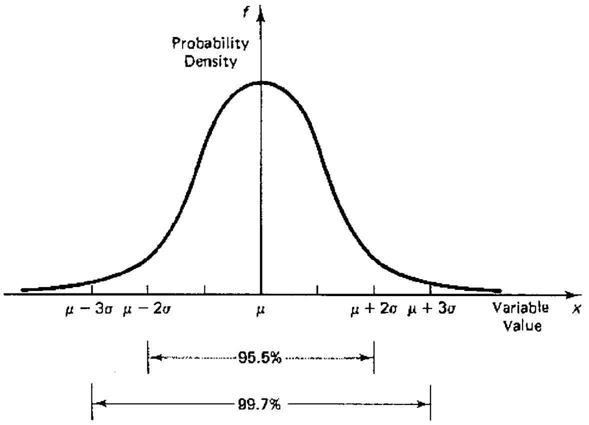


Figure 5.11. Gaussian (normal) probability density function.

Estimate the overall eccentricity at a confidence level of 96 percent.

Solution The mean value of the overall eccentricity may be estimated as the rms value of the individual means; thus, frnm equation 5.10,

$$\mu = \sqrt{0.1^2 + 0.2^2 + 0.05^2 + 0.1^2} = 0.25 \text{ mm}$$

Using equation 5.11, the standard deviation of the overall eccentricity is estimated as

$$\sigma = \sqrt{0.01^2 + 0.05^2 + 0.001^2 + 0.02^2} = 0.055 \text{ mm}$$

Now, assuming Gaussian distribution, an estimate for the overall eccentricity at a confidence level of 96 percent is given by

$$\hat{e} = 0.25 + 2 \times 0.055 = 0.36 \text{ mm}$$

Once the overall eccentricity is estimated in the foregoing manner, the corresponding measurement error can be determined. Suppose that the true angle of rotation is θ and the corresponding measurement is θ_m. The eccentricity error is given by

$$\Delta\theta = \theta_m - \theta \tag{5.12}$$

The maximum error can be shown to exist when the line of eccentricity (CG) is symmetrically located within the angle of rotation, as shown in figure 5.12. For this configuration, the sine rule fnr triangles gives

$$\frac{\sin(\Delta\theta/2)}{e} = \frac{\sin(\theta/2)}{r}$$

where r denotes the code track radius, which can be taken as the disk radius for most practical purposes. Hence, the eccentricity error is given by

$$\Delta\theta = 2\sin^{-1}\left(\frac{e}{r}\sin\frac{\theta}{2}\right) \tag{5.13}$$

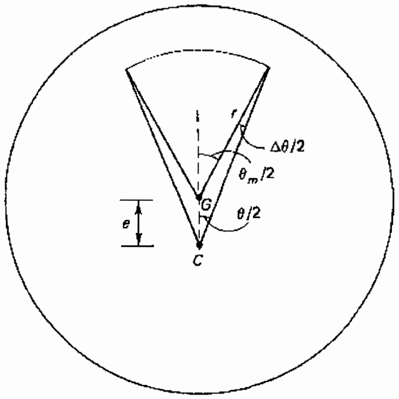


Figure 5.12. Nomenclature for eccentricity error (C = center of rotation, G = geometric center of the code track).

Example 5.4

Show analytically that the eccentricity error of an encoder disk does not enter measurements of complete revolutions of the disk.

Solution It is intuitively clear that the eccentricity error should not enter measurements of complete revolutions, and this can be shown analytically by using equation 5.13. In this case, $\theta = 2\pi$. Accordingly, $\Delta\theta = 0$. For multiple revolutions, the eccentricity error is periodic with period 2π.

For small angles, the sine of an angle is approximately equal to the angle itself, in radians. Hence, for small $\Delta\theta$, the eccentricity error may be expressed as

$$\Delta\theta = \frac{2e}{r}\sin\frac{\theta}{2} \tag{5.14}$$

Furthermore, for small angles of rotation, the fractional eccentricity error is given by

$$\frac{\Delta\theta}{\theta} = \frac{e}{r} \tag{5.15}$$

which is, in fact, the worst fractional error. As the angle of rotation increases, the fractional error decreases (as shown in figure 5.13), reaching the zero value for a full revolution. From the point of view of gross error, the worst value occurs when $\theta = \pi$, which corresponds to half a revolution. From equation 5.13, it is clear that the maximum gross error due to eccentricity is given by

$$\Delta\theta_{\max} = 2\sin^{-1}\frac{e}{r} \tag{5.16}$$

If this value is less than half the resolution of the encoder, the eccentricity error becomes inconsequential. For all practical purposes, since e is much less than r, we may use the following expression for the maximum eccentricity error:

$$\Delta\theta_{\max} = \frac{2e}{r} \tag{5.17}$$

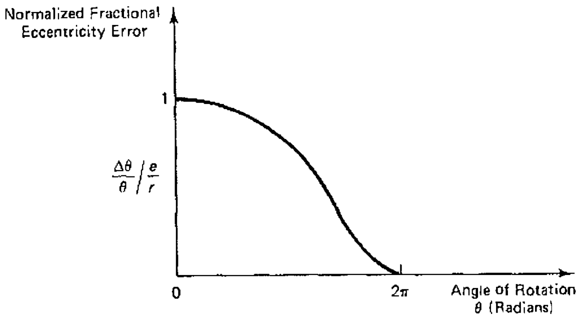


Figure 5.13. Fractional eccentricity error variation with the angle of rotation.

Example 5.5

Suppose that in example 5.3, the radius of the code disk is 5 cm. Estimate the maximum error due to eccentricity. If each track has 1,000 windows, determine whether the eccentricity error is significant.

Solution With the given level of confidence, we have calculated the overall eccentricity to be 0.36 mm. Now, from equation 5.16 or 5.17, the maximum angular error is

$$\Delta\theta_{\max} = \frac{2 \times 0.36}{50} = 0.014 \text{ rad} = 0.83^\circ$$

Assuming that quadrature signals are used to improve the encoder resolution, we have

$$\text{resolution} = \frac{360^\circ}{4 \times 1{,}000} = 0.09^\circ$$

Note that the maximum error due to eccentricity is more than ten times the encoder resolution. Hence, eccentricity will significantly affect the accuracy of the encoder.

Eccentricity also affects the phase angle between the quadrature signals of an incremental encoder if a single track and two pick-off sensors (with circumferential offset) are used. This error can be reduced using the two-track arrangement, with the two sensors positioned along a radial line, so that eccentricity affects the two outputs equally.

5.6 DIGITAL RESOLVERS

Digital resolvers, or mutual induction enenders, operate somewhat like analog resolvers, using the principle of mutual induction. They are known commercially as Inductosyns (Ferrand Contrnls, Valhalla, N.Y.). A digital resolver has two disks facing each other (but not touching), one (the stator) stationary and the other (the rotor) coupled to the rotating object. The rotor has a fine electric conductor foil imprinted on it, as shown schematically in figure 5.14. The printed pattern is a closely spaced set of radial pulses, all of which are connected to a high-frequency AC supply of voltage v_{ref}. The stator disk has two separate printed patterns that are identical to the rntor pattern, but one pattern on the stator is shifted by a quarter-pitch from the other pattern. The primary voltage in the rotor circuit induces voltages in the two secondary (stator) foils at the same frequency. As the rotnr turns, the level of the induced voltage changes, depending on the relative position of the foil patterns on the two disks. When the foil pulse patterns coincide, the induced voltage is maximum (positive or negative), and when the rntor foil pattern has a half-pitch offset from the stator foil pattern, the induced voltage in adjacent parts cancel each other, producing a zero output. If the speed of rotation is constant, the output voltages v_1 and v_2 in the two foils of the stator become signals that have a carrier frequency (supply frequency) component modulated by periodic and nearly sinusoidal signals with a

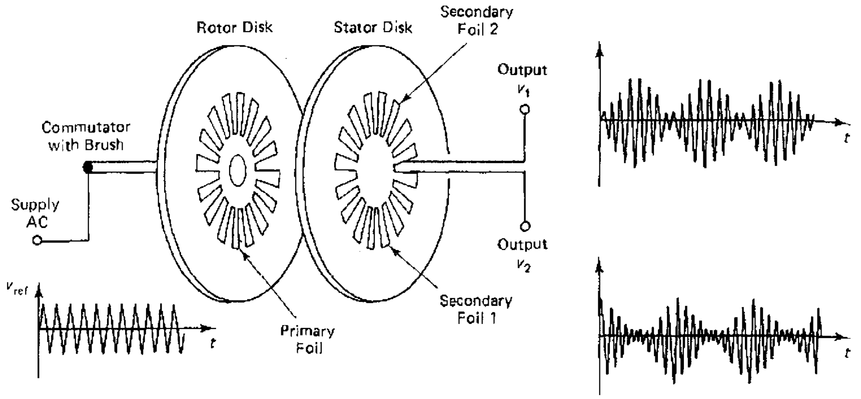


Figure 5.14. Schematic diagram of a digital resolver.

phase shift of 90°. The modulation signals can be extracted by demodulation and converted into pulse signals. When the speed is not constant, pulse width will vary with time. As in the case of an incremental encoder, angular displacement and angular velocity are determined by counting or timing the pulses. The direction of rotation is determined by the phase difference in the two modulating output signals. (In one direction, the phase shift is 90°; in the other direction, it is −90°.) Very fine resolutions are obtainable from a digital resolver; it is usually not necessary to use step-up gear systems or other techniques to improve resolution. Resolutions up to 0.0005° can be obtained from these transducers, but they are usually more expensive than optical encoders.

5.7 DIGITAL TACHOMETERS

Since shaft encoders are also used for measuring angular velocities, they can be considered tachometers. In classic terminology, a digital tachometer is a device that employs a toothed wheel to measure angular velocities. A schematic diagram of one such device is shown in figure 5.15. This is a magnetic induction tachometer of the variable-reluctance type. The teeth on the wheel are made of ferromagnetic material. The two magnetic induction (and variable-reluctance) proximity probes are placed facing the teeth radially, a quarter-pitch apart. When the toothed wheel rotates, the two probes generate output signals that are 90° out of phase. One signal leads the other in one direction of rotation and lags the other in the opposite direction of rotation. In this manner, directional readings are obtained. The speed is computed either by counting pulses over a sampling period or by timing the pulse width, as in the case of an incremental encoder.

Alternative types of digital tachometers use eddy current proximity probes or capacitive proximity probes (see chapter 3). In the case of an eddy current tachome-

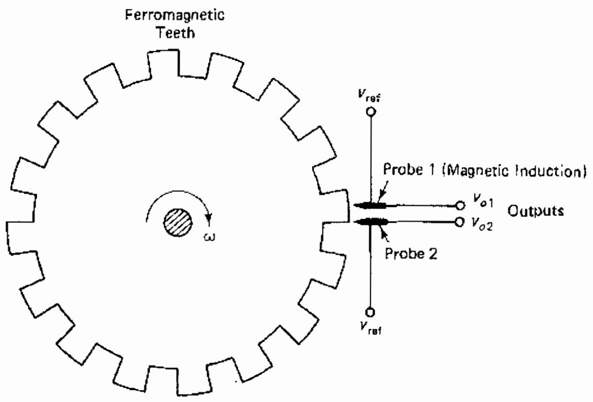


Figure 5.15. Schematic diagram of a pulse tachometer.

ter, the teeth of the pulsing wheel are made of or plated with electricity-conducting material, and the probe emits a radio-frequency magnetic field. In the case of a capacitive tachometer, the toothed wheel forms one plate of the capacitor; the other plate is the probe and is kept stationary. As the wheel turns, the capacitor gap width fluctuates. If the capacitor is excited by an AC voltage of high frequency (typically 1 MHz), a near-pulse-modulated signal at that carrier frequency is obtained. This can be detected by a suitable capacitance bridge circuit. By demodulating the output signal, the modulating-pulse signal can be extracted. This pulse signal is used in the angular velocity computation.

The advantages of these digital (pulse) tachometers over optical encoders include simplicity, robustness, and low cost. The disadvantages include poor resolution (determined by the number of teeth, the speed of rotation, and the word size used for data transmission), and mechanical errors due to loading, hysteresis, and manufacturing irregularities.

5.8 HALL EFFECT SENSORS

Consider a semiconductor element subject to a DC voltage v_{ref}. If a magnetic field is applied perpendicular to the direction of this voltage, a voltage v_o will be generated in the third orthogonal direction within the semiconductor element. This is known as the Hall effect (observed by E. H. Hall in 1879). A schematic representation of a Hall effect sensor is shown in figure 5.16.

A Hall effect sensor may be used for motion sensing in many ways—for example, as an analog proximity sensor, a digital limit switch, or a digital shaft encoder. Since the output voltage v_o increases as the distance from the magnetic source to the semiconductor element decreases, the output signal v_o can be used as a measure of proximity. Alternatively, a certain threshold level of output voltage v_o can be used

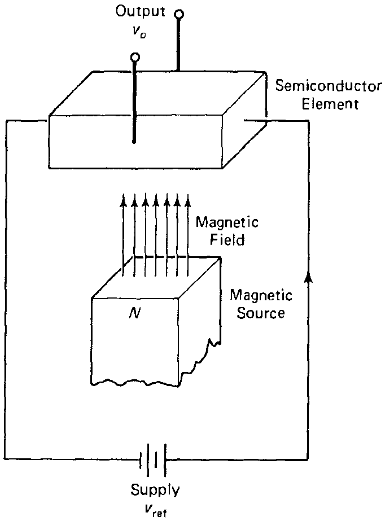


Figure 5.16. Schematic representation of a Hall effect sensor.

to activate a digital switch or to create a digital output, hence forming a digital limit switch.

A more practical arrangement would be to have the semiconductor element and the magnetic source fixed relative to one another in a single package. By moving a ferromagnetic member into the air gap between the magnetic source and the semiconductor element, the flux linkage can be altered. This changes v_o. This arrangement is suitable both as an analog proximity sensor and as a limit switch. Furthermore, if a toothed ferromagnetic wheel is used to change v_o, we have a shaft encoder or a digital tachometer (see figure 5.17).

The longitudinal arrangement of a proximity sensor, in which the moving element approaches head-on toward the sensor, is not suitable when there is a danger of overshooting the target, since it will damage the sensor. A more desirable configuration is the lateral arrangement, in which the moving member slides by the

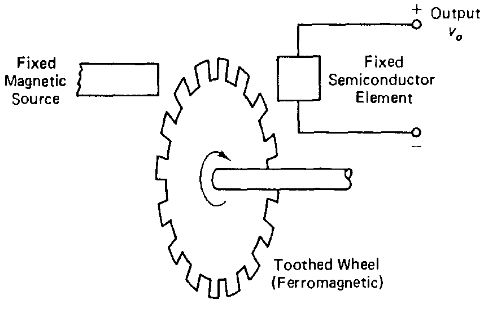


Figure 5.17. Schematic diagram of a Hall effect shaft encoder or digital tachometer.

sensing face of the sensor. The sensitivity will be lower, however, with this lateral arrangement.

The relationship between the output voltage v_o and the distance x of a Hall effect sensor measured from the moving member is nonlinear. Linear Hall effect sensors use calibration to linearize their output.

5.9 MEASUREMENT OF TRANSLATORY MOTIONS

Digital rectilinear transducers are useful in many applications. Typical applications include *x-y* positioning tables, machine tools, valve actuators, read-write heads in disk drive systems, and robotic manipulators (e.g., at prismatic joints) and robot hands. The principles used in angular motion transducers described so far in this book can be used in measuring rectilinear motions as well. In rectilinear encoders, for example, rectangular flat plates moving rectilinearly, instead of rotating disks, are used with the same types of signal generation and interpretation mechanisms.

Cable Extension Sensors

In many applications, rectilinear motion is produced from a rotary motion (say, of a motor) through a suitable transmission device, such as rack and pinion or lead screw and nut. In these cases, rectilinear motion can be determined by measuring the associated rotary motion, assuming that errors due to backlash, flexibility, and so forth, in the transmission device can be neglected. Another way to measure rectilinear motions using a rotary sensor is to use a modified sensor that has the capability to convert a rectilinear motion into a rotary motion within the sensor itself. An example would be the ***cable extension method*** of sensing rectilinear motions. This method is particularly suitable for measuring motions that have large excursions. The cable extension method uses an angular motion sensor with a spool rigidly coupled to the rotating part of the sensor (e.g., the encoder disk) and a cable that wraps around the spool. The other end of the cable is attached to the object whose rectilinear motion is to be sensed. The housing of the rotary sensor is firmly mounted on a stationary platform, so that the cable can extend in the direction of motion. When the object moves, the cable extends, causing the spool to rotate. This angular motion is measured by the rotary sensor. With proper calibration, this device can give rectilinear measurements directly. As the object moves toward the sensor, the cable has to retract without slack. This is guaranteed by using a device such as a spring motor to wind the cable back. The disadvantages of the cable extension method include mechanical loading of the moving object, time delay in measurements, and errors caused by the cable, including irregularities, slack, and tensile deformation.

Moiré Fringe Displacement Sensors

Another device for measuring rectilinear motions employs the moiré fringe technique. The operation of this motion sensor is somewhat analogous to that of a digital resolver, and the signal generation method is similar to that in an optical encoder.

Thus, this device can be viewed as an optical encoder with improved sensitivity and resolution. A transparent plate with a series of opaque lines arranged in parallel in the transverse direction forms the stationary plate (grating plate) of the transducer. This is called the mask plate. A second transparent plate, with an identical set of ruled lines, forms the moving plate. The lines on both plates are evenly spaced, and the line width is equal to the spacing between adjacent lines. A light source is placed on the moving plate side, and light transmitted through the common area of the two plates is detected on the other side using one or more photosensors. When the lines on the two plates coincide, the maximum amount of light will pass through the common area of the two plates. When the lines on one plate fall on the transparent spaces of the other plate, virtually no light will pass through the plates. Accordingly, as one plate moves relative to the other, a pulse train is generated by the photosensor, and it can be used to determine rectilinear displacement and velocity, as in the case of an incremental encoder. Moiré fringes are the shadow patterns formed in this manner. They can also be detected and observed by photographic means. With this technique, very small resolutions (e.g., 0.0002 in.) can be realized. Note that the method provides improved sensitivity over a basic optical encoder because light passing through many gratings is received by the same photosensor. Also, finer line spacing (in conjunction with wider light sensors) can be used in this method, thereby providing increased resolution.

The moiré device is used to measure rigid-body movements of one plate of the sensor with respect to the other, and it can be used to detect deformations (e.g., elastic deformations) of one plate with respect to the other in the direction orthogonal to the grating lines. In this case, depending on the nature of the plate deformation, some transparent lines of one plate will be completely covered by the opaque lines of the other plate, and some other transparent lines of the first plate will have coinciding transparent lines on the second plate. Thus, the observed image will have dark lines (moiré fringes) corresponding to the regions with clear/opaque overlaps of the two plates and bright lines corresponding to the regions with clear/clear overlaps of the two plates. Hence, the moiré fringe pattern will provide the deformation pattern of one plate with respect to the other.

Example 5.6

Suppose that each plate of a moiré fringe deformation sensor has a line pitch of 0.01 mm. A tensile load is applied to one plate in the direction perpendicular to the lines. Five moiré fringes are observed in 10 cm of the moiré image under tension. What is the tensile strain in the plate?

Solution There is one moiré fringe in every 10/5 = 2 cm of the plate. Hence, extension of a 2 cm portion of the plate = 0.01 mm, and

$$\text{tensile strain} = \frac{0.01\ \text{mm}}{2 \times 10\ \text{mm}} = 0.0005\epsilon = 500\mu\epsilon$$

In the foregoing example, we have assumed that the strain distribution (or deformation) of the plate is uniform. Under nonuniform strain distributions, the observed moiré fringe pattern generally will not be parallel straight lines.

5.1

5.10 LIMIT SWITCHES

Limit switches are sensors used in detecting limits of mechanical motions. A limit of a movement can be detected by using a simple contact mechanism to close a circuit or trigger a pulse. Hence, the information provided by a limit switch takes only two states (on/off, present/absent, go/no-go, etc.); it can be represented by one bit. In this sense, a limit switch is considered a digital transducer. Additional logic is needed if the direction of contact is also needed. Limit switches are available for both rectilinear and angular motions.

A microswitch is a solid-state switch that can be used as a limit switch. Microswitches are commonly used in counting operations—for example, to keep a count of completed products in a factory warehouse.

Although a purely mechanical device consisting of linkages, gears, ratchet wheels and pawls, and so forth, can serve as a limit switch, electrical and solid-state switches are usually preferred for such reasons as accuracy, durability, a low activating force (practically zero) requirement, low cost, and small size. Any proximity sensor could serve as the sensing element of a limit switch. The proximity sensor signal is then used in the required manner—for example, to activate a counter, a mechanical switch, or a relay circuit, or simply as an input to a control computer.

PROBLEMS

5.1. Identify active transducers among the following types of shaft encoders, and justify your claims. Also, discuss the relative merits and drawbacks of the four types of encoders.
(a) Optical encoders
(b) Sliding contact encoders
(c) Magnetic encoders
(d) Proximity sensor encoders

5.2. Explain why the speed resolution of a shaft encoder depends on the speed itself. What are some of the other factors that affect speed resolution? The speed of a DC motor was increased from 50 rpm tn 500 rpm. How would the speed resolution change if the speed was measured using an incremental encoder
(a) By the pulse-counting method?
(b) By the pulse-timing method?

5.3. Discuss construction features and operation of an optical encoder for measuring *rectilinear* displacements and velocities.

5.4. What is hysteresis in an optical encoder? List several causes of hysteresis and discuss ways to minimize hysteresis.

5.5. Describe methods of improving resolution in an encoder. An incremental encoder disk has 5,000 windows. The word size of the output data is 12 bits. What is the angular resolution of the device? Assume that quadrature signals are available but that no interpolation is used.

5.6. A shaft encoder that has N window per track is connected to a shaft through a gear system with gear ratio p. Derive formulas for calculating angular velocity of the shaft by
(a) The pulse-counting method
(b) The pulse-timing method

What is the speed resolution in each case? What effect does step-up gearing have on the speed resolution?

5.7. An optical encoder has n windows/inch diameter (in each track). What is the eccentricity tolerance e below which readings are not affected by eccentricity error?

5.8. Show that in the single-track, two-sensor design of an incremental encoder, the phase angle error (in quadrature signals) due to eccentricity is inversely proportional to the second power of the radius of the code disk for a given window density. Suggest a way to reduce this error.

5.9. Encoders that can provide 50,000 counts/turn with ±1 count accuracy are commercially available. What is the resolution of such an encoder? Describe the physical construction of an encoder that has this resolution.

5.10. A particular type of multiplexer can handle ninety-six sensors. Each sensor generates a pulse signal with variable pulse width. The multiplexer scans the incoming pulse sequences, one at a time, and passes the information onto a control computer.
(a) What is the main objective of using a multiplexer?
(b) What type of sensors could be used with this multiplexer?

5.11. Suppose that a feedback control sytem is expected to provide an accuracy within $\pm\Delta y$ in a response variable y. Explain why the sensor that measures y should have a resolution of $\pm(\Delta y/2)$ or better for this accuracy to be possible. An x-y table has a travel of 2 m. The feedback control system is expected to provide an accuracy of ±1 mm. An optical encoder is used to measure position for feedback in each direction (x and y). What is the minimum bit size that is required for each encoder output buffer? If the motion sensor used is an absolute encoder, how many tracks and how many sectors should be present on the encoder disk?

5.12. Discuss the advantages of solid-state limit switches over mechanical limit switches. Solid-state limit switches are used in many applications, particularly in the aircraft and aerospace industries. One such application is in landing gear control, to detect up, down, and locked conditions of the landing gear. High reliability is of utmost importance in such applications. Mean time between failure (MTBF) of over 100,000 hours is possible with solid-state limit switches. Using your engineering judgment, give an MTBF value for a mechanical limit switch.

5.13. Explain how resolution of a shaft encuder could be improved by pulse interpolation. Suppose that a pulse generated from an incremental encoder can be approximated by

$$v = v_o\left(1 + \sin\frac{2\pi\theta}{\Delta\theta}\right)$$

where θ denotes the angular position of the encoder window with respect to the photosensor position. Let us consider rotations of a half-pitch or smaller (i.e., $0 \le \theta \le \Delta\theta/2$, where $\Delta\theta$ is the window pitch angle). By using this sinusoidal approximation for a pulse, show that we can improve the resolution of an encoder indefinitely simply by measuring the shape of each pulse at clock cycle intervals using a high-frequency clock signal.

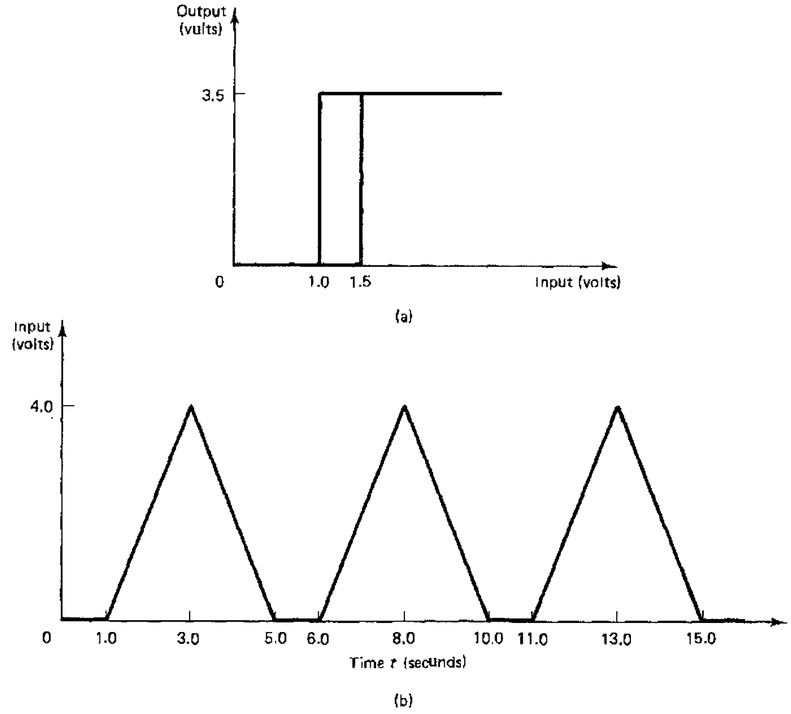
5.14. What is a Hall effect tachometer? Discuss the advantages and disadvantages of a Hall effect motion sensor in comparison to an optical motion sensor (e.g., an optical encoder).

5.15. The pulses generated by the coding disk of an incremental optical encoder are approximately triangular in shape. Explain the reason for this. Describe a method for converting these triangular pulses into sharp rectangular pulses.

n the
:cen-
)hase
> the
ı way
mer-
con-
ates a
:e se-
±Δy
:solu-
l of 2
n op-
What
iotion
ıld be
tches.
ft and
ct up,
mpor-
hours
ive an
lation.
:d by
photo-
(i.e.,
pprox-
initely
;h-fre-
a Hall
al cn-
proxi-
nvert-
ıap. 5

5.16. A brand of autofocusing camera uses a microprocessor-based feedback control system consisting of a charge-coupled device (CCD) imaging system, a microprocessor, a drive motor, and an optical encoder. The purpose of the control system is to focus the camera automatically, based on the image of the subject as sensed by a matrix of CCDs (a set of metal oxide semiconductor field-effect transistors, or MOSFETs). The light rays from the subject that pass through the lens will fall onto the CCD matrix. This will generate a matrix of charge signals, which are shifted one at a time, row by row, into an output buffer and passed on to the microprocessor after conditioning the resulting video signal. The CCD image obtained by sampling the video signal is analyzed by the microprocessor to determine whether the camera is focused properly. If not, the lens is moved by the motor so as to achieve focusing. Draw a schematic diagram for the autofocusing control system and explain the function of each component in the control system, including the encoder.

5.17. A Schmitt trigger is a semiconductor device that can function as a level detector or a switching element with hysteresis. The presence of hysteresis can be used, for example, to eliminate chattering during switching caused by noise in the switching signal. The input/output characteristic of a Schmitt trigger is shown in figure P5.17a. If the input signal is as shown in figure P5.17b, determine the output signal.



(a)

(b)

Figure P5.17. (a) The input/output characteristic of a Schmitt trigger. (b) A triangular input signal.

5.18. Displacement sensing and speed sensing are essential for a position servo. If a digital controller is employed to generate the servo signal, one option would be to use an analog displacement sensor and an analog speed sensor, along with analog-to-digital converters (ADCs) to produce the necessary digital feedback signals. Alternatively, an incremental encoder could be used to provide both displacement and speed feedbacks. In this case, ADCs are not needed. Encoder pulses will provide interrupts to the digital controller. Displacement is obtained by counting the interrupts. The speed is obtained by timing the interrupts. In some applications, analog speed signals are needed. Explain how an incremental encoder and a frequency-to-voltage converter (FVC) may be used to generate an analog speed signal.

5.19. Consider the two quadrature pulse signals (say, *A* and *B*) from an incremental encoder. Using sketches of these signals, show that in one direction of rotation, signal *B* is at a high level during the up-transition of signal *A*, and in the opposite direction of rotation, signal *B* is at a low level during the up-transition of signal A. Note that the direction of motion can be determined in this manner by using level detection of one signal during the up-transition of the other signal.

REFERENCES

Barney, G. C. *Intelligent Instrumentation.* Prentice-Hall, Englewood Cliffs, N.J., 1985.

Cerni, R. H., and Foster, L. E. *Instrumentation for Engineering Measurement.* Wiley, New York, 1962.

deSilva, C. W. "Motion Sensors in Industrial Robots." *Mechanical Engineering* 107(6): 40–51, June 1985.

———. "Counters/Frequency Tachometers." *Measurements and Control Journal* 116: 201, April 1986.

Franklin, G. F., and Powell, J. D. *Digital Control of Dynamic Systems.* Addison-Wesley, Reading, Mass., 1980.

Woolvet, G. A. *Transducers in Digital Systems.* Peter Peregrinus, London, 1979.

CLARENCE W. de SILVA

CONTROL SENSORS AND ACTUATORS

Suitable for both an undergraduate course in control system instrumentation and a graduate course in sensors and actuators for feedback control, this text is also a useful reference tool for practicing control system engineers. The text's seven chapters comprise a thorough foundation in control system instrumentation. Background information is provided in chapter 1. Component modeling, rating, matching, and error analysis aspects are discussed early, in chapter 2, so that the relevance and significance of these considerations can be explored in subsequent chapters. Chapters 3, 4, and 5 are devoted to sensors and transducers (e.g., digital and analog motion sensors, torque, force, and tactile sensors), and chapters 6 and 7 consider stepper motors (permanent-magnet, variable reluctance, and hybrid types) and continuous-drive actuators such as DC motors, induction motors, synchronous AC motors, and hydraulic actuators.

The author treats the basic types of control sensors and actuators in separate chapters without losing sight of the fact that various components in a control system function interdependently in accomplishing the specific control objective. In-depth discussions of some types of sensors and actuators include operating principles, modeling, design considerations, ratings, specifications, and applications. While other components are covered in less detail, the student will have gained the background needed to extend concepts and approaches to components that are functionally or physically similar.

Component integration and design considerations are addressed primarily through numerous worked end-of-chapter examples and problems. These are drawn from application systems such as robotic manipulators, machine tools, ground transit vehicles, aircraft, thermal and fluid process plants, and digital-computer components. Beyond their traditional role, the problems serve as a valuable source of information in addition to the main text. Answers to the numerical problems only are given at the end of the book to encourage independent thinking.

PRENTICE HALL, Englewood Cliffs, N.J. 07632

ISBN 0-13-171745-6