LINEAR POSITION SENSORS

LINEAR POSITION SENSORS

Theory and Application

DAVID S. NYCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Nyce, David S.
Linear position sensors: theory and application / David S. Nyce.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-23326-9 (cloth)
1. Transducers. 2. Detectors. I. Title.
TK7872.T6N93 2003
681′.2—dc21
2003053455

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To Gwen, and our children Timothy, Christopher, and Megan,
whose love and support helped me complete this project

CONTENTS

PREFACE

Society and industry worldwide continue to increase their reliance on the availability of accurate and current measurement information. Timely access to this information is critical to effectively meet the indication and control requirements of industrial processes, manufacturing equipment, household appliances, onboard automotive systems, and consumer products. A variety of technologies are used to address the specific sensing parameters and configurations needed to meet these requirements.

Sensors are used in cars to measure many safety- and performance-related parameters, including throttle position, temperature, composition of the exhaust gas, suspension height, pedal position, transmission gear position, and vehicle acceleration. In clothes-washing machines, sensors measure water level and temperature, load size, and drum position variation. Industrial process machinery requires the measurement of position, velocity, and acceleration, in addition to chemical composition, process pressure, temperature, and so on. Position measurement comprises a large portion of the worldwide requirement for sensors. In this book we explain the theory and application of the technologies used in sensors and transducers for the measurement of linear position.

There is often some hesitation in selecting the proper word, *sensor* or *transducer*, since the meanings of the terms are somewhat overlapping in normal use. In Chapter 1 we present working definitions of these and other, sometimes confusing, terms used in the field of sensing technology. In Chapter 2 we explain how the performance of linear position transducers is specified. In the remaining chapters we present the theory supporting an understanding of the prominent technologies in use in linear position transducer products. Application guidance and examples are included.

The following are the owners of the trademarks as noted in the book:

CANbus	Robert Bosch GmbH, Stuttgart, Germany
HART	HART Communications Foundation, Austin, TX
Lincoder	Stegmann Corporation, Germany
NiSpan C	Huntington Alloys, Incorporated
Permalloy	B&D Industrial Mining Services, Inc.
Profibus	PROFIBUS International
Ryton	Phillips Petroleum Company
SSI	Stegmann Corportation, Germany
Temposonics	MTS Systems Corporation, Eden Prairie, MN
Terfenol D	Extrema Products, Inc., Ames, IA
Torlon	Amoco Performance Products, Inc.

CHAPTER 1

SENSOR DEFINITIONS AND CONVENTIONS

1.1 IS IT A SENSOR OR A TRANSDUCER?

A *transducer* is generally defined as a device that converts a signal from one physical form to a corresponding signal having a different physical form [29, p. 2]. Energy can be converted from one form into another for the purpose of transmitting power or information. Mechanical energy can be converted into electrical energy, or one form of mechanical energy can be converted into another form of mechanical energy. Examples of transducers include a loudspeaker, which converts an electrical input into an audio wave output; a microphone, which converts an audio wave input into an electrical output; and a stepper motor, which converts an electrical input into a rotary position change.

A *sensor* is generally defined as an input device that provides a usable output in response to a specific physical quantity input. The physical quantity input that is to be measured, called the *measurand*, affects the sensor in a way that causes a response represented in the output. The output of many modern sensors is an electrical signal, but alternatively, could be a motion, pressure, flow, or other usable type of output. Some examples of sensors include a thermocouple pair, which converts a temperature difference into an electrical output; a pressure sensing diaphragm, which converts a fluid pressure into a force or position change; and a linear variable differential transformer (LVDT), which converts a position into an electrical output.

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

Obviously, according to these definitions, a transducer can sometimes be a sensor, and vice versa. For example, a microphone fits the description of both a transducer and a sensor. This can be confusing, and many specialized terms are used in particular areas of measurement. (An audio engineer would seldom refer to a microphone as a sensor, preferring to call it a transducer.) Although the general term *transducer* refers to both input and output devices, in this book we are concerned only with sensing devices. Accordingly, we will use the term *transducer* to signify an input transducer (unless specified as an output transducer).

So, for the purpose of understanding sensors and transducers in this book, we will define these terms more specifically as they are used in developing sensors for industrial and manufacturing products, as follows:

An *input transducer* produces an electrical output, which is representative of the input measurand. Its output is conditioned and ready for use by the receiving electronics.

The receiving electronics can be an indicator, controller, computer, programmable logic controller, or other. The terms *input transducer* and *transducer* can be used interchangeably, as we do in this book.

A *sensor* is an input device that provides a usable output in response to the input measurand.

The sensing part of a transducer can also be called the *sensing element*, *primary transducer*, or *primary detector*. A sensor is often one of the components of a transducer.

Sometimes, common usage will have to override our theoretical definition in order to result in clear communication among engineers in a specific industry. The author has found, for instance, that automotive engineers refer to any measuring device providing information to the onboard controller, as a sensor. In the case of a position measurement, this includes the combination of sensing element, conditioning electronics, power supply, and so on. That is, the term *sensor* is used to name exactly what our definition strives to call a transducer. In automotive terminology, the word *sender* is also commonly used to name a sensor or transducer. In any case, we rely on the definition presented here, because it applies to most industrial uses.

An example of a sensor as part of a transducer may help the reader understand our definition. The metal diaphragm shown in Figure 1.1*a* is a sensor that changes pressure into a linear motion. The linear motion can be changed into an electrical signal by an LVDT, as in Figure 1.1*b*. The combination of the diaphragm, LVDT, and signal conditioning electronics would comprise a pressure transducer. A pressure transducer of this description, designed by the author, is shown in Figure 1.2.

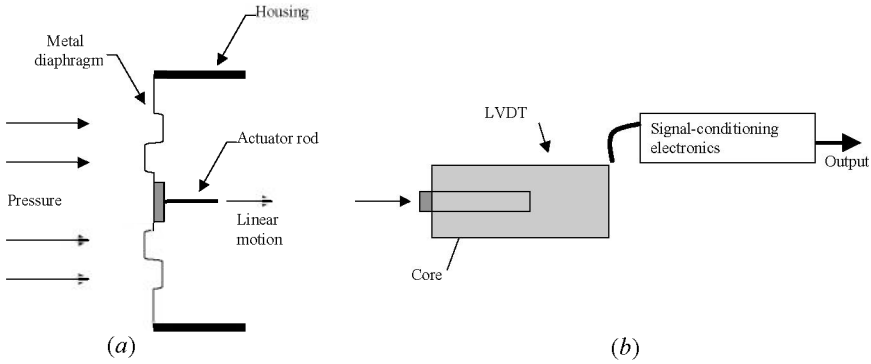

Figure 1.1 (*a*) The circular diaphragm (shown edgewise, cutaway) changes pressure into linear motion. (*b*) An LVDT changes linear motion to an electrical signal, comprising a transducer with the addition of signal-conditioning electronics.

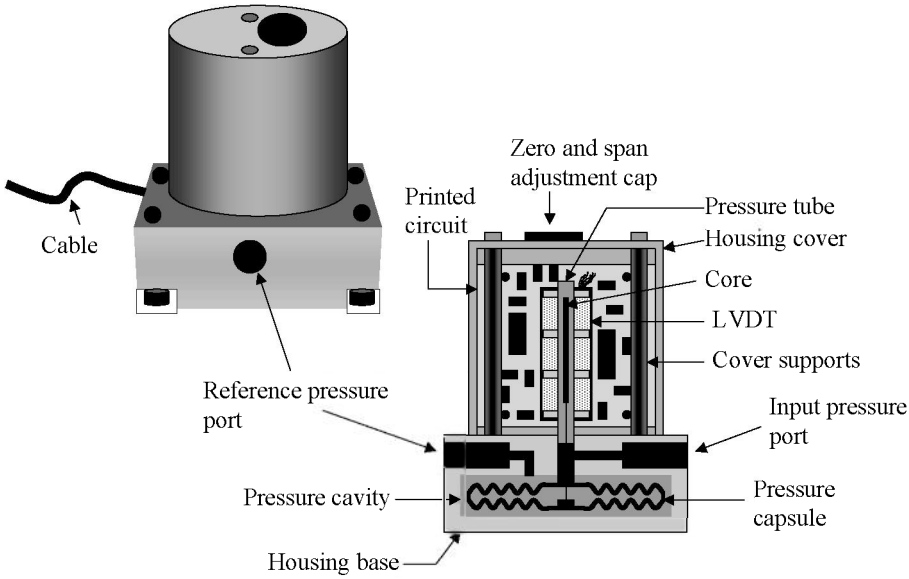

Figure 1.2 Commercially available pressure transducer according to Figure 1.1. Cutaway view with diaphragm in the lower cavity, and LVDT, core, and signal-conditioning electronics in the upper cavity.

1.2 POSITION VERSUS DISPLACEMENT

Since linear position sensors and transducers are presented in this work and many manufacturers confuse the terms *position* and *displacement*, the difference between position and displacement should be understood by the reader.

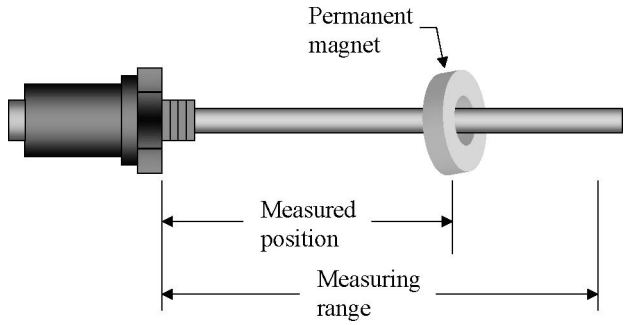

Figure 1.3 Magnetostrictive linear position transducer with position magnet. (Courtesy of MTS Systems Corporation.)

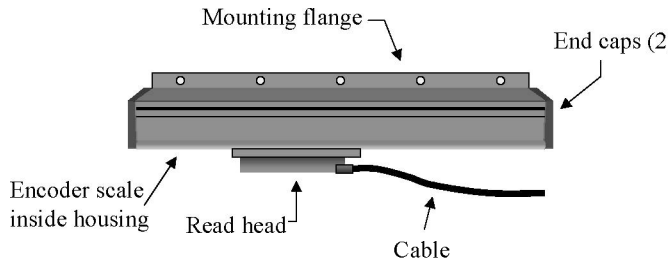

Figure 1.4 Incremental magnetic linear encoder.

A *position transducer* measures the distance between a reference point and the present location of the target. The word *target* is used in this case to mean that element of which the position or displacement is to be determined. The reference point can be one end, the face of a flange, or a mark on the body of the position transducer (such as a fixed reference datum in an absolute transducer), or it can be a programmable reference datum. As an example, consider Figure 1.3, which shows the components of the measuring range of a magnetostrictive absolute linear position transducer. This transducer measures the location of a permanent magnet with reference to a fixed point on the transducer. (More details on the magnetostrictive position transducer are presented in Chapter 9.)

Conversely, a *displacement transducer* measures the distance between the present position of the target and the position recorded previously. An example of this would be an incremental magnetic encoder (see Figure 1.4). Position transducers can be used as displacement transducers by adding circuitry to remember the previous position and subtract the new position, yielding the difference as the displacement. Alternatively, the data from a position transducer may be recorded into memory by a microcontroller, and differences calculated as needed to indicate displacement. Unfortunately, and con-

stituting another assault against clarity, it is common for many manufacturers of position transducers to call their products displacement transducers.

To summarize, *position* refers to a measurement with respect to a constant reference datum; *displacement* is a relative measurement.

1.3 ABSOLUTE OR INCREMENTAL READING

An absolute-reading position transducer indicates the measurand with respect to a constant datum. This reference datum is usually one end, the face of a flange, or a mark on the body of a position transducer. For example, an absolute linear position transducer may indicate the number of millimeters from one end of the sensor, or a datum mark, to the location of the target (the item to be measured by the transducer). If power is interrupted, or the position changes repeatedly, the indication when normal operation is restored will still be the number of millimeters from one end of the sensor, or a datum mark, to the location of the target. If the operation of the transducer is disturbed by an external influence, such as by an especially strong burst of electromagnetic interference (EMI), the correct reading will be restored once normal operating conditions return.

To the contrary, an incremental-reading transducer indicates only the changes in the measurand as they occur. An electronic circuit is used to keep track of the sum of these changes (the count) since the last time that a reading was recorded and the count was zeroed. If the count is lost due to a power interruption, or the sensing element is moved during power-down, the count when normal operating conditions are restored will not represent the present magnitude of the measurand. For example, if an incremental encoder is first zeroed, then moved upscale 25 counts, followed by moving downscale 5 counts, the resulting position would be represented by a count of 20. If there are 1000 counts per millimeter, the displacement is 0.02 mm. If power is lost and regained, the position would probably be reported as 0.00 mm. Also, if the count is corrupted by an especially strong burst of EMI, the incorrect count will remain when normal operation is restored.

1.4 CONTACT OR CONTACTLESS SENSING AND ACTUATION

One classification of a position transducer pertains to whether it utilizes a *contact* or *noncontact* (also called *contactless*) type of sensing element. With contactless sensing, another aspect is whether or not the transducer also uses contactless actuation. In a contact type of linear position sensor, the device making the conversion between the measurand and the sensor output incorporates a sliding electrical and/or mechanical contact. The primary example is the linear potentiometer, (see Figure 1.5). The actuator rod is connected internally to a wiper arm. The wiper arm incorporates one or more

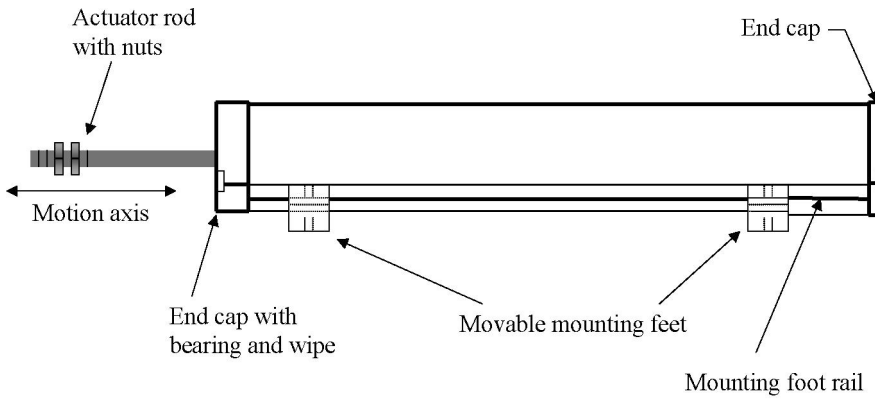

Figure 1.5 Linear potentiometer.

flexible contacts, which press against a resistive element. The potentiometer is powered by applying a voltage across the resistive element. Changing position along the motion axis causes the wiper(s) to rub against the resistive element, thus producing an output voltage as an indication of the measurand. A more complete description of the linear potentiometer is provided in Chapter 3.

It is because of the rubbing contact between the wiper and the resistive element that a linear potentiometer is called a contact sensor. The primary advantages are its simplicity and that it often does not require signal conditioning. It is also generally thought of as a low-cost sensing technique, although automation of manufacture of other types of sensors is closing the cost gap. The disadvantage of a contact sensor is that there is a finite lifetime associated with the rubbing elements. Further explanation of this in reference to potentiometric linear position transducers and the design trade-offs taken to optimize operating life are also presented in Chapter 3.

In a contactless linear position sensor, the device making the conversion between the measurand and the sensor output incorporates no physical connection between the moving parts and the stationary parts of the sensor. The "connection" between the moving parts and the stationary parts of the sensor is typically provided through the use of inductive, capacitive, magnetic, or optical coupling. Examples of contactless linear position sensing elements include the LVDT, Hall effect, magnetostrictive, and magnetoresistive sensors. These are explained further in their respective chapters later in the book, but as an example, we consider the LVDT here briefly.

An LVDT linear position transducer with core is shown in Figure 1.6. The core is attached to the movable member of the system being measured (the target). The LVDT housing is attached to the stationary member of the system. As the core moves within the bore of the LVDT, there is no physical contact between the core and the remainder of the LVDT. Inductive coupling between

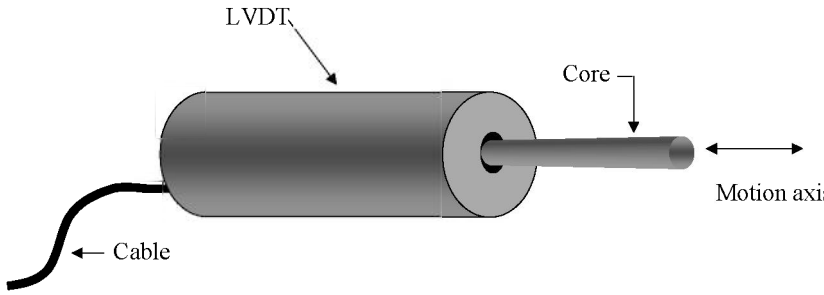

Figure 1.6 LVDT linear position transducer with core.

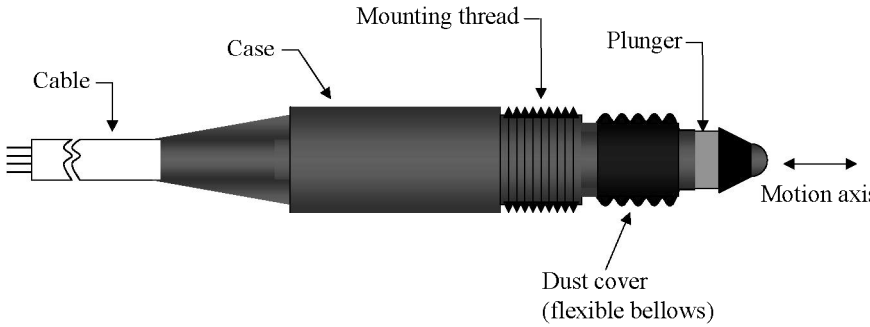

Figure 1.7 Contacting actuation in an LVDT gauge head.

the LVDT primary and its secondary windings, through the magnetically permeable core, afford the linkage. Contactless sensors are generally more complicated than linear potentiometers, and typically require signal conditioning electronics.

In addition to contactless operation within the sensor, a sensing system may utilize contactless *actuation* when there is no mechanical coupling between the sensing element and the movable physical element (the target) whose position is being measured. For an example of magnetic coupling, a permanent magnet can be mounted to a movable machine toolholder, and a magnetostrictive position transducer (as shown in Figure 1.3) can be mounted along the motion axis of the toolholder. The measurement of tool position is then made without any mechanical contact between the toolholder and the sensing element. Contactless actuation obviously does not utilize any rubbing parts, which can wear out and reduce the life or accuracy of the measurement. Conversely, contacting actuation is used with an inherently contactless sensor when the toolholder presses the spring-loaded plunger of an LVDT gauge head, for example (see Figure 1.7).

Even though the LVDT itself operates as a contactless sensor, the contact actuation of the plunger leaves the system somewhat open to reduced life and

varying accuracy, due to wear. In this example, repeated rubbing of the gauge head shaft against its bushings will eventually result in wear, possibly affecting performance through undesired lateral motion of the shaft, or increased operating force.

1.5 LINEAR AND ANGULAR CONFIGURATIONS

Linear position sensors and transducers operate by utilizing any of a large number of technologies, some of these being resistive, capacitive, inductive, Hall effect, magnetoresistive, magnetostrictive, and optical. Although this book presents the theory and application of linear position sensors, these same technologies are used to build angular sensors and transducers. For example, a resistive type of linear position sensor operates in much the same way as one constructed to measure an angular measurand. The angular (or rotary) sensor requires the addition of a rotating shaft to hold the wipers, and the resistive element is circular in shape. Other than that, the basic theory of operation is the same. If the reader is more interested in angular than in linear position sensing, the information in this book can still provide a good understanding of the technologies used. A detailed study of angular sensors, however, would include additional topics, such as angular momentum, rotational speed range, turn-counting techniques, torque requirements, end play, and bearing specification.

1.6 APPLICATION VERSUS SENSOR TECHNOLOGY

Linear position sensors can be designed that are based on one or more of a wide variety of technologies, as noted above and presented individually later in the book. When determining which sensor type to specify for use in a specific application, it may be important to match the technology of the sensor to the requirements of the application.

If the sensor will undergo continuous repetitive motion, as with constant vibration, contactless sensing and contactless actuation may be required to eliminate parts that could wear out. In this case, magnetic or optical coupling to the sensor can be used. If it is desired to use the same linear position sensor type for short strokes (tens of millimeters) as well as long strokes (several meters), a sensor technology with this operating range capability may be required. Magnetostrictive technology can be used in this case. Advantages and disadvantages for each technology are listed in the respective chapters, but Table 1.1 provides general information on application suitability.

The rated lifetime of a sensor element can be an important consideration in the application of a contact linear potentiometer in the presence of continuous vibration. A typical lifetime rating for a potentiometer is 20 million cycles. If the motion system has a constant dithering or vibration at 10 Hz, for

TABLE 1.1 Application Suitability of Various Sensors

Technology	Absolute	Noncontact	Lifetime	Resolution	Range	Stability
Resistive	Yes	No	Low	Medium	Medium	Medium
Capacitive	Yes	Some models	High	Low to high	Low	Low
Inductive	Yes	Yes	High	Medium	Medium	Low
LVDT	Yes	Yes	High	High	Medium	Medium
Hall effect	Yes	Yes	High	High	Low	Low
Magnetoresistive	Yes	Yes	High	High	Low	Low
Magnetostrictive	Yes	Yes	High	High	High	High
Encoder	Some models	Some models	Medium	Low to high	Medium	High

example, this number of cycles can be accumulated at a small spot on the element within two months. Many motion systems have two primary positions in which they operate over 90% of the time. The number of cycles of the example in each of these two positions is represented, per month, by the equation

$$\begin{aligned} &10\,\text{Hz} \times 2.59 \times 10^6\,\text{s/month} \times 50\%\text{/position} \times 90\%\,\text{duty} \\ &\quad = 11.6 \times 10^6\,\text{cycles/position/month} \end{aligned} \tag{1.1}$$

This assumes that the two primary positions are used about equally. Accordingly, a contact resistive sensor (potentiometer) exposed to 10 Hz dithering in two positions can wear out within months. See Chapter 3 for more details on resistive sensing.

CHAPTER 2

SPECIFICATIONS

2.1 ABOUT POSITION SENSOR SPECIFICATIONS

The list of parameters that are important to specify in characterizing a position sensor may be somewhat different from those that would be important to specify in, for example, a sensor for gas analysis. Compared to a gas sensor, the position sensor may have a similar need to list power supply requirements, operating temperature range, and nonlinearity but there will be differences related to the specific measuring technique. A position sensor specification should indicate whether it measures linear or angular motion, if the reading is absolute or incremental, and whether it uses contact or contactless sensing and actuation. Conversely, a gas sensor spec would indicate what kind of gas is detected, how well it ignores other interfering gases, if it measures gas by percent volume or partial pressure, and the shelf life (if it is an electrochemical type of gas sensor having a limited lifetime). So there exist a number of specifications that are important when describing the performance capability of a position transducer and its suitability for use in a given application. These specifications are presented here.

2.2 MEASURING RANGE

For it to provide an accurate reading, the measurand, or physical quantity being measured, must have a range that is within the capability of the trans-

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

ducer. A position transducer can have a measuring range specified from zero to full scale, or it can be specified as a ± full-scale range (FSR). It is common with an LVDT, for example, to specify bipolar ranges, such as ±100 mm FSR. In this case and with a ±10-V dc output specified, the output voltage would vary from −10 V direct current (dc) to +10 V dc for a measurand changing from −100 mm to +100 mm. In the center of travel, the output would be zero. Since the example transducer is specified over the range −100 to +100 mm, the full-scale range is 200 mm. If the corresponding output range were ±10 V dc, the full-range output (FRO) would span 20 V dc. These are the amounts used when other parameters are specified as a percent of FSR, or FRO. For example, with an LVDT and signal conditioner specified for a maximum zero shift of 1.0% per 100°C, an FSR of ±100 mm, and an FRO of ±10.0 V dc, a 100°C temperature change can produce an error of 2.0 mm or 0.20 V.

In a magnetostrictive position sensor, the sensing element measures a time period starting from one end, thus making an absolute, zero-based measurement. Even so, it is possible to produce a transducer having a bipolar range by adding an offset incorporated within the signal conditioning electronics; but the most common configuration is to have a zero to full scale range (unipolar), with zero being located near one end of the transducer. An example of a unipolar range is an output of 0.0 V dc to +10.0 V dc, corresponding to an input position of zero to 1.0 m.

2.3 ZERO AND SPAN

The terms *zero* and *span* are used to describe the measurand and/or the output of a transducer. On a unipolar scale, the zero is the lowest reading, and the span is the difference between the full-scale and zero readings. For example, a position transducer may have a measuring range of 0.0 to 1.0 m and produce an output of 4.0 to 20.0 mA. In this case the input measurand has a zero of 0.0 m and a full scale of 1.0 m. The span is also 1.0 m. The output has an offset, however. The output has a zero of 4.0 mA and a full scale of 20 mA. The span is therefore 16.0 mA. So 16 mA of output span represents, and is proportional to, 1 m of input measurand. The output sensitivity is thus 16.0 μA/mm. This output sensitivity means that from any starting point in the measuring range, the output will change by 16.0 μA for each millimeter of position change.

Understanding the distinction among zero, span, and full scale is important when troubleshooting errors, since knowing whether the error is a zero shift or a span shift can indicate the error source. If, for example, you are temperature-testing a position transducer with an output of 4 to 20 mA, corresponding to an input range of 0 to 100 mm, you would first set the position to zero. The output will be approximately 4 mA. As the temperature is varied in an environmental chamber, changes in the output are recorded as "zero" error.

Next, the position is set to 100.0 mm. The output will be approximately 20 mA. After again changing the temperature over the same range, record the output changes as FRO error. Subtract the zero error from the FRO error to find the span error. By analyzing these errors, the source(s) of any temperature sensitivity problems can be categorized. Things that cause zero error are offset-related errors. They can be mechanical, such as thermal expansion of a mounting feature or actuator rod, or electrical, such as input voltage drift of an amplifier or resistance change in a voltage-divider circuit.

Things that cause a span error are gain-related errors. They can also be mechanical, such as a changing spring rate; or electrical, such as change in a transistor gain, a resistance change in an amplifier feedback loop, or a capacitance change in a coupling capacitor for an alternating-current (ac) signal. Knowing this cause-and-effect link helps to guide one's efforts in the troubleshooting of transducer errors as well as when designing a sensor or transducer to meet the specifications required in the product development stage.

2.4 REPEATABILITY

When the transducer is exercised over a set of conditions, and then exactly the same conditions are met again, the difference between the consecutive readings is called *repeatability*. This is usually tested by maintaining fixed temperature, humidity, and other environmental conditions and then exercising the transducer by changing the measurand between fixed points. For example, the core of an LVDT can be exercised from zero, to full scale, to zero, then to half scale. A data point is taken at the last position. Then the movement of the core is continued to full scale, to zero, then to half scale again. The second data point is taken. This is done repeatedly to obtain a set of data. The standard deviation of this data set is the repeatability.

It is possible, theoretically, to have a repeatability that has a smaller value than the resolution, by adding noise to the system and making a statistical analysis of the resulting set of data; but this is not helpful to someone using the transducer. So the specified repeatability should not be smaller than the specified resolution. This assures that it is possible for the user to reproduce the specified level of performance. Repeatability can be the most important characteristic of a transducer if the receiving equipment is able to compensate for nonlinearity, temperature effects, calibration error, and so on. This is because repeatability is the only transducer characteristic that cannot be compensated. Also, in many control systems, repeatability is more important than transducer accuracy because the system can often be programmed to provide the output desired in response to a given input from the transducer, as long as the input received from the transducer is always the same for a given set of conditions.

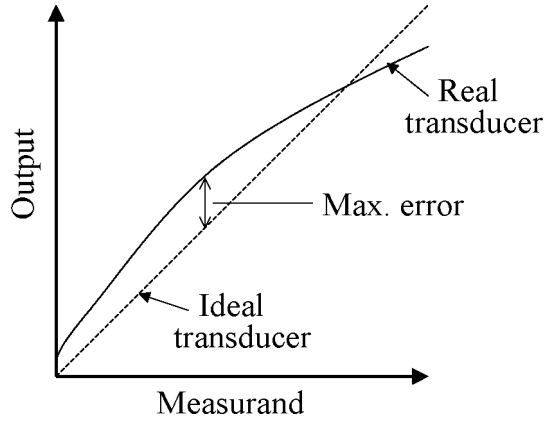

Figure 2.1 Nonlinearity, comparing an ideal transducer characteristic (a straight line) to the characteristic of a real transducer.

2.5 NONLINEARITY

The set of output data obtained from a theoretically perfect (ideal) linear position transducer, when exercising it throughout the specified operating range and recording the output data versus input stroke, should form a straight line from the zero reading to the full-scale reading. In a real transducer, the data do not form a perfectly straight line, and the endpoints are not exactly at the specified zero and full-scale points. This is shown in Figure 2.1, somewhat exaggerated for clarity. The maximum amount of difference between the transducer characteristic and the ideal characteristic is the maximum error. This could be reported as a percent of full range and called the percent accuracy, but instead, accuracy is normally reported as the individual components comprising it. This is appropriate, since there are other components that limit the accuracy of a transducer in a given application. The term *static error band* is properly used to indicate the sum of the effects of nonlinearity, repeatability, and hysteresis. Environmental effects are typically reported separately. Nonlinearity itself, however, can be interpreted in several ways, as presented next. Repeatability and hysteresis are presented in the following sections.

Typically, the most important characteristic of transducer accuracy is nonlinearity. A straight line is drawn that closely approximates the transducer characteristic. The difference between the straight line and an ideal line is calibration error. Calibration error can be broken down further into zero offset and gain (or span) error. The difference between the straight line and the transducer characteristic is the nonlinearity, reported as a percentage of full range. The *nonlinearity* error specification is often referred to improperly as the transducer "linearity." For example, if the maximum error (between the

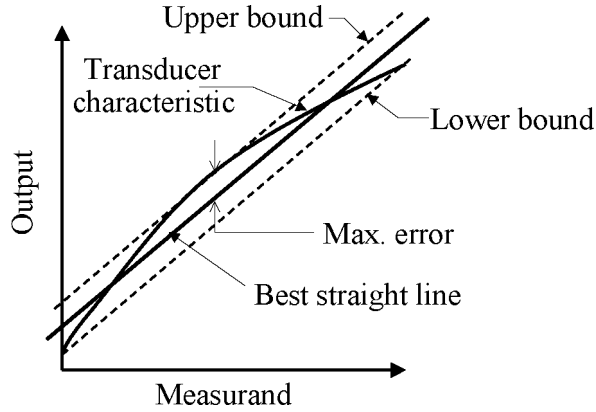

Figure 2.2 Finding the best straight line and maximum nonlinearity error.

transducer characteristic and a straight line) is 0.5 mm and the full-scale range is 100 mm, the nonlinearity is 0.5%. This sounds simple enough, but there are a number of ways to arrive at a "best" straight line, which closely approximates the transducer characteristic, and to which the transducer output data will be compared.

Best Straight Line Nonlinearity

The best straight line (BSL) can also be called the *best-fit straight line* or *independent BSL*. When BSL or best-fit straight line is all that is named as the nonlinearity reference in the specification, or an independent BSL is named, it is not required that any specific point on the BSL be drawn through any specific data point of the transducer output characteristic. The BSL does not have to go through zero or full-scale input, or either endpoint of the sensor data. The purpose is only to find a straight line that comes closest to matching all the output data points of the transducer. The stated nonlinearity is then the maximum deviation of any data point from this straight line. A good way to visualize this is shown in Figure 2.2.

Two lines are placed on the graph of the transducer characteristic, one above and one below the line representing the transducer data. These are called the *upper* and *lower bounds*. The two parallel lines should be brought as close together as possible while encompassing all the transducer data between them. They do not have to be parallel to the transducer data. A third straight line is then placed along the center between the two parallel lines. This third line is the best straight line. The maximum deviation (error) between this line and the transducer data, expressed as a percentage of full range, is the transducer BSL nonlinearity. This line can be defined in Y-intercept form as

$$Y = mX + B \tag{2.1}$$

where m is the slope of the line and B is the Y-intercept. This means that m is the scaling factor and B is the zero offset.

One can visualize that half of the distance between the two parallel lines drawn on the graph (measured vertically) is the BSL nonlinearity, being the absolute value of the amplitude of the maximum deviation of the output from a straight line. The method for calculating the BSL without using a graph, however, may not be evident at first glance. A practical way to find this line from the data is first to find the least-squares line through the data (see "Least-Squares Straight-Line Nonlinearity") and use this to derive a line equation in Y-intercept form [equation (2.1)]. Then use an iterative method with small changes in slope (m) and intercept (B) until a line equation is found that yields the minimum deviation from the transducer data.

Zero-Based Nonlinearity

When it is desired to ensure that the output indicates zero when the measurand is zero, a zero-based nonlinearity may be specified. This may be needed when the indication of a negative position would not make sense and the equipment receiving the transducer signal cannot make the correction. In this case, one end of a straight line is set equal to the zero measurand/zero output point (in a graph, the origin), and the other end of the line is moved up or down (changing the slope) until minimizing the maximum deviation of the sensor output data from the line (see Figure 2.3). There will usually be one or more points on the sensor characteristic that fall above the straight line, as

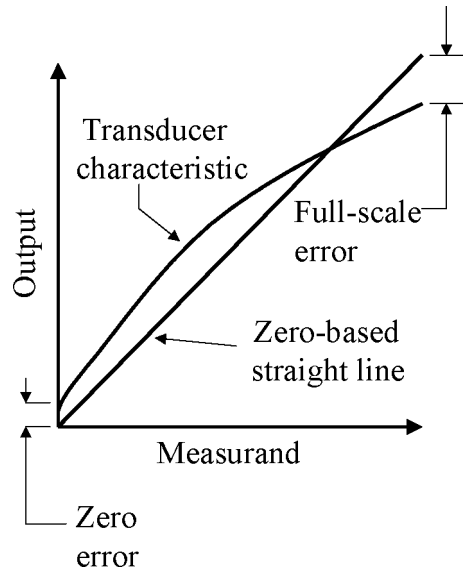

Figure 2.3 Zero-based nonlinearity.

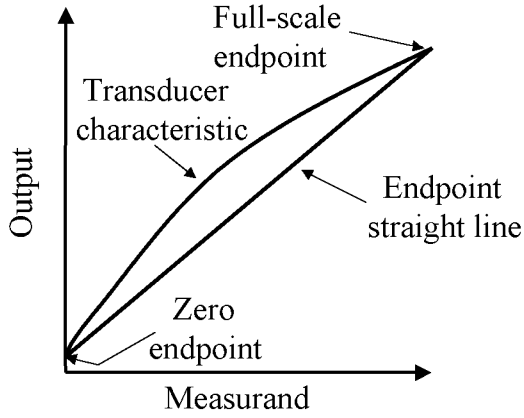

Figure 2.4 Endpoint nonlinearity.

well as one or more points that fall below it. In a uniformly curved characteristic as in Figure 2.3, there will be one maximum somewhere near the midpoint and another near full scale. These two error amounts should be approximately the same if the straight line is properly placed.

Endpoint Nonlinearity

A straight line can be drawn between the transducer outputs at zero measurand and at full scale (these two points are called the *endpoints*). The maximum deviation between this line and the transducer data is called the *endpoint nonlinearity* (see Figure 2.4). Transducer manufacturers prefer to specify nonlinearity according to one of the other methods, though, because the magnitude of the endpoint nonlinearity is in the range of two times the number obtained by one of the other methods. Endpoint nonlinearity may be of interest to a user whose equipment does not have a means for correcting gain errors of the transducer.

Least-Squares Straight-Line Nonlinearity

Nonlinearity based on a least-squares regression (LSR) of the input data versus the output data is the most popular type of specification because it can easily be calculated. The disadvantage is that it can be very close to the optimum line but is not necessarily the absolute best straight line, since it is a statistical estimation. The degree to which the LSR line actually represents the "best" straight line depends on the number of data points taken and the nonuniformity or erratic nature of the data. The result will be less representative when the data do not follow a continuous smooth curve and when the

number of data points is smaller. Still, it is the most popular way to find a BSL, since it is easy to implement mathematically.

If the LSR straight line is represented as $Y = mX + B$, the slope m, is found by

$$m = \frac{\sum_{d=1}^{n} X_d Y_d}{\sum_{d=1}^{n} X_d^2} \tag{2.2}$$

where X_d and Y_d are the data from the input measurand and transducer output, respectively, and n is the number of data points. Once the slope m is found, the Y-intercept that yields the lowest overall deviation must be found. Then the maximum deviation is reported as the least-squares nonlinearity. It is easy to implement on a set of data using a pocket calculator or spreadsheet program.

In a calculator, select the linear regression function. Enter the input measurand data consecutively as the set of values for the first variable of a two-variable array. Enter the corresponding sensor output data as the set of values for the second variable of the array. Select the calculate function.

In a spreadsheet program, select the linear regression analysis tool, this performs a linear regression using the least-squares method to fit a straight line through the data selected as input data columns in a spreadsheet. For example, in Excel, load the analysis tool pack. Then select the regression analysis tool. Make a spreadsheet with one column of input measurand data versus a second column with the corresponding sensor output data over the full range of transducer operation. Select the input measurand data (first column) as the input X range. Select the output data (second column) as the input Y range. Then calculate the slope of the LSR line using the SLOPE function. Find the Y-intercept using the INTERCEPT function. This will provide the slope and the Y-intercept of the least-squares regression line. Next, an "LSR line" (third column) is made, using the slope and Y-intercept applied to the X range (first column) according to the formula $Y = mX + b$. Then calculate the errors (fourth column) as the difference between the second and third columns. The maximum number in the error column (fourth column) is the least-squares nonlinearity. See Table 2.1 and also refer to Figure 2.5, which is a graph of the data of Table 2.1.

After the spreadsheet functions are used to find the slope and intercept of a least-squares line, the least-squares nonlinearity error is specified as the maximum difference between the transducer data and least-squares BSL data, divided by the FRO. The slope and intercept constants for a particular transducer can be entered or downloaded into the equipment using the transducer, thereby allowing the equipment to correct the transducer signal to improve overall system accuracy.

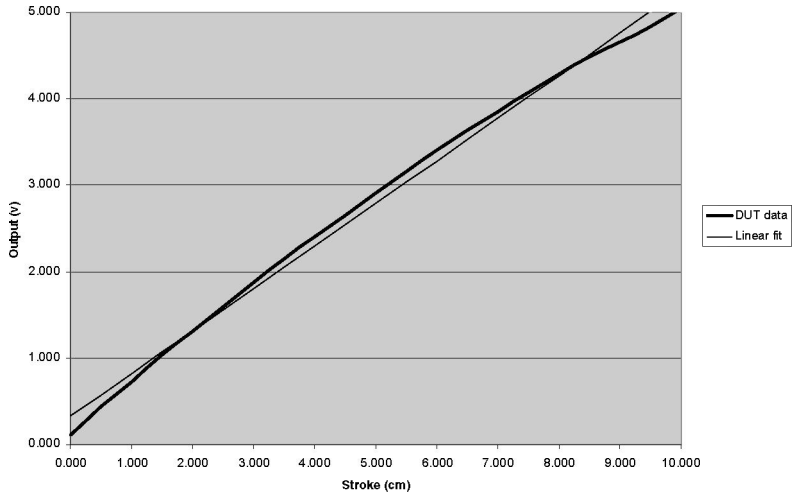

Figure 2.5 Graph of the transducer data and the least-squares straight line of Table 2.1.

TABLE 2.1 Example of an Excel Spreadsheet Showing Position, Output, and the Calculated Nonlinearity Error[a]

Ref. Position (cm)	Transducer Output (V)	Best Line Calc.	Best Line Data (V)	Error (V)	Error (cm)
0.000	0.112		0.327	–0.215	–0.436
0.500	0.441	Slope =	0.573	–0.132	–0.268
1.000	0.727	0.493	0.819	–0.092	–0.188
1.500	1.029		1.066	–0.037	–0.075
2.000	1.305		1.312	–0.007	–0.015
2.500	1.594	Intercept =	1.559	0.035	0.072
3.000	1.872	0.327	1.805	0.067	0.136
3.500	2.144		2.051	0.093	0.188
4.000	2.397		2.298	0.099	0.202
4.500	2.643		2.544	0.099	0.201
5.000	2.906		2.790	0.116	0.235
5.500	3.159		3.037	0.122	0.248
6.000	3.410		3.283	0.127	0.258
6.500	3.629		3.529	0.100	0.202
7.000	3.853		3.776	0.077	0.157
7.500	4.068		4.022	0.046	0.093
8.000	4.286		4.269	0.017	0.035
8.500	4.485		4.515	–0.030	–0.061
9.000	4.651		4.761	–0.110	–0.224
9.500	4.838		5.008	–0.170	–0.344
10.000	5.049		5.254	–0.205	–0.416

[a] Input position data as reference position and output data as transducer output. The least squares nonlinearity is 4.36% FRO.

2.6 HYSTERESIS

Regarding the output signal of a position transducer, hysteresis is the variation between upscale and downscale approaches to the same position. More specifically, when a sensor is steadily indicating an increasing output (moving upscale), crossing through position *a* of the measurand, then reversing direction and steadily indicating a decreasing reading (moving downscale), again passing through position *a*, there will be a slight difference in the reading recorded for the increasing and decreasing approaches to position *a*. This characteristic is shown, exaggerated, in Figure 2.6. Position *a* is shown as a point on the lower curve, corresponding to an increasing measurand. A different output is shown for the same point *a* on the upper curve, corresponding to a decreasing measurand. The difference between the two points is the maximum error due to hysteresis.

That which is typically called hysteresis may include mechanical backlash, the building of spring force before a wiper moves, magnetic remanence in a sensing element magnetic circuit, and plastic deformation of a sensing member, among others. These are typically reported only as an overall hysteresis error, and the individual elements are not reported separately.

In a position sensor based on the use of a magnetic field, for example, one cause of hysteresis is the magnetic remanence of the material, which is being affected by the magnetic field. An initially nonmagnetized material would first follow line (*a*) as shown in Figure 2.7 upon exposure to a magnetizing force. As the external field (magnetizing force) builds up, the magnetic material becomes magnetized. Then, when the magnetizing force is reduced, the remanence of that material causes some of the magnetic field to remain in the mag-

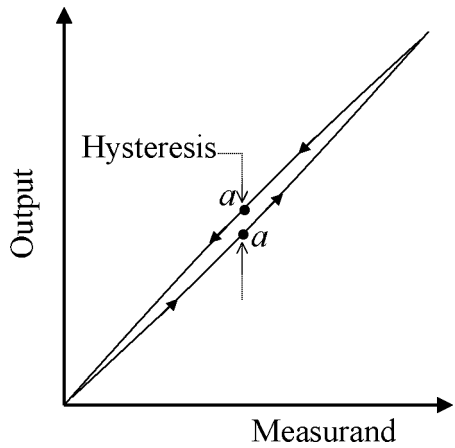

Figure 2.6 Hysteresis shown in a plot of measurand versus transducer output. The hysteresis error is the difference between upscale and downscale measurements of the same measurand, *a*.

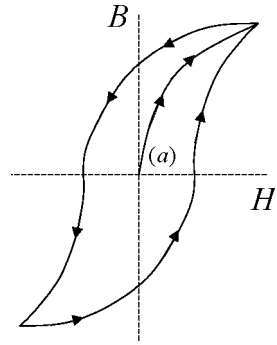

Figure 2.7 Remanent magnetic field in a magnetic material. Line (*a*) is an initial non-magnetized state.

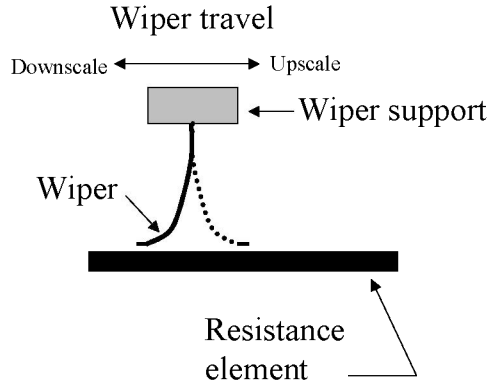

Figure 2.8 Wiper flexing causes different upscale and downscale readings. The upscale tracking position is shown as the solid wiper. The dashed-line wiper is the downscale tracking position.

netic material—it has become somewhat "magnetized." Thereafter, the field strength in the material would follow lines (*b*) and (*c*) when subjected to further reductions and increases in magnetizing force. The remanent field may result in an error in the output signal from the transducer. The measurement of the magnetic remanence of a material is the value of the flux density, *B*, retained with the magnetizing force, *H*, removed, after magnetizing the material to saturation [20, p. 333].

Hysteresis in a potentiometric type of position sensor comes from other sources. The wiper may flex slightly down as it is being moved up, and then start to flex slightly up as it is being moved down. The lagging of the output reading with respect to the input motion will cause a difference between the upscale and downscale readings (see Figure 2.8). The amount of flexing depends on the flexural strength of the wiper, the wiper force pressing it

against the resistive element, and the surface friction of the resistive element. There may also be backlash in the actuator that drives the wiper movement. Backlash is sometimes separated out in screw- or gear-driven potentiometers but has the same effect on performance in a given application.

The accepted way to measure hysteresis in the output of a position transducer is first to exercise the transducer throughout its full range in order to have a reproducible starting point. Then the position is varied smoothly to move the measurand starting from the zero reading, up to full scale, and then back to zero, while recording data at approximately uniformly spaced points along the range of the measurand. The upscale and downscale tracks are plotted. Then the maximum deviation between the two is noted. This deviation is reported as hysteresis and specified as a percent of full scale. Typically, the maximum error will be in the middle of the stroke. In an LVDT that travels in both a positive and negative direction, with respect to the null or zero position, the maximum hysteresis error is normally around the null point. Sometimes a bipolar range LVDT will have a unipolar hysteresis specification, as well as one for bipolar operation, or report a null hysteresis separately.

A related parameter of potentiometric sensors is friction error, due to the friction of the sliding wiper. This is usually included in the hysteresis spec as described above, but not always. In a potentiometer application that will be accompanied by constant vibration, the effect of wiper friction will be greatly reduced. To indicate this, a hysteresis error with friction-free measuring is sometimes stated in the specification of a contact sensor. During testing, a vibration is applied to overcome friction and allow the wiper to move to the friction-free point. This is also called *mechanical dithering*. (An interesting fact: An analogous electrical dithering can be used to average out quantizing error for increased resolution in some digital electronic circuits.)

2.7 CALIBRATED ACCURACY

A transducer exhibits a given performance, including nonlinearity, hysteresis, temperature sensitivity, and so on; however, the actual performance in the application is also affected by the accuracy to which the transducer output was calibrated to a known standard. For a position sensor, lengths for reference accuracy can be measured with a linear encoder (such as that manufactured by Stegmann and pictured in Figure 2.9), a laser interferometer (Figure 2.10), or another sensing technique capable of accuracy sufficiently higher than that expected from the sensor being measured. The normal requirement is that the reference standard should exhibit an error 10 times less than that of the device to be tested. In this case, the error in the reference device can be essentially ignored. Sometimes, though, this ratio of error is not practically available. When using a ratio of less than 10, an allowance should be made for this when evaluating the data.

Figure 2.9 Linear encoder (magnetic strip type). (Courtesy of Stegmann, Inc.)

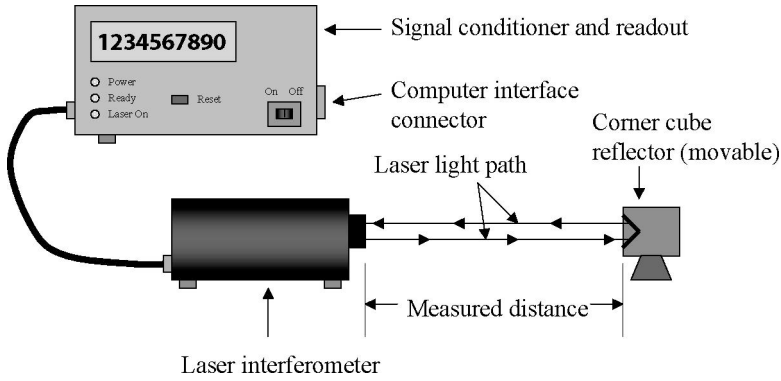

Figure 2.10 Laser interferometer.

Calibrated accuracy is the absolute accuracy of the individual transducer calibration and includes the accuracy of the standard used as well as the ability of the calibration technique to produce a setting that matches the standard. For example, if the setting is made by turning a potentiometer adjustment, the operator tries to obtain a setting that results in a particular output reading. The operator will be able to achieve this to within some level of tolerance. That tolerance will become part of the calibrated accuracy specification, in addition to any allowance made due to the accuracy of the reference standard that was used. Rather than specifying a calibrated accuracy of 99.9%, for example, it is more common to list a calibration error of 0.1%. When evaluating the total error budget of an application, the calibration error must be included as well as the nonlinearity, hysteresis, temperature error, and other factors.

2.8 DRIFT

Drift encompasses the changes in the transducer output that occur even though there are no changes in the measurand or environmental conditions. The only variable when measuring drift is the elapsed time. In a position transducer, this means that there is no position change (the sensor actuator is normally locked into a stationary position for this test). The test is run at constant temperature, constant humidity, constant power supply voltage, constant load impedance, and so on, while the transducer output is recorded. Drift is reported in two components: short- and long-term drift, and is expressed as a percentage of full-range output. On a typical position transducer, short-term drift is that which occurs in less than 24 hours. It is reported as error in percent of FRO per hour. Long-term drift is specified in the same way, but the time period is per month. On some reference-grade equipment, long-term drift time period may be per year.

Sources of short-term drift include such things as noise, instability in electronic circuits, mechanical instability, insufficient electrical or mechanical damping, and susceptibility to random low-level electrical noise in the environment, whereas long-term drift originates from changes in electrical component characteristics and mechanical wear. For example, electrolytic capacitors can change capacitance value or equivalent series resistance (ESR) as the electrolyte dries with age. Mechanical components can undergo wear or fatigue. Identifying the type of drift experienced (short or long term) can give clues to the possible sources of the drift.

2.9 WHAT DOES ALL THIS ABOUT ACCURACY MEAN TO ME?

An engineer tasked with implementing a position transducer into a control system must determine whether or not the control system will be capable of exhibiting the specified position accuracy when incorporating the feedback element (position transducer) that is planned to be used. Errors can be divided into the categories of either the static or dynamic type. Static errors in a transducer typically include nonlinearity, hysteresis, and repeatability. Dynamic errors include phase shift or amplitude variation due to the transducer frequency response, amplitude variation due to damping factor, and so on. Errors due to environmental conditions are normally reported separately and include errors from changes in temperature, humidity, moisture, pressure, salt spray, and so on.

The position transducer specification will probably not list an overall error that can be expected, which would include a combination of all the static and dynamic errors (i.e., performance over the temperature range, including nonlinearity, hysteresis, etc.). Rather, all the specifications will be listed individually, and it is up to the user to decide how to add up or otherwise choose to utilize the specified errors in determining the suitability of the transducer for

obtaining the desired system performance. Some sources of error will apply to the application being considered, and some will not. For example, the requirement may include a wide temperature range but only a slowly changing measurand. In this case, the temperature error will be important, but not the frequency response, phase shift, or damping factor. Also, the dynamic errors can sometimes be very pronounced and must each be considered in detail. The static error band includes several errors, however, all of which can be in the same range of magnitude. These errors must be added up in some way and evaluated for their accumulated effect on the performance of the transducer application. The sum of the static errors is called the *static error band*. At a Christmas party of a former employer, the author (who enjoys rock music) witnessed a performance of a rock musical group that called themselves "The Static Error Band." A good play on words, but their performance was dreadful.

If all the individual specifications were simply combined as an arithmetic sum, this could be used as the overall accuracy specification. Doing this, however, would not be realistic. It is not likely that all the errors would each be at its maximum simultaneously, and at the same time, for each to act in the worst-case direction so that their effects would add. Instead, some errors will be positive and some errors will be negative. Some errors will be near maximum, others will be around average, some are likely to be lower than average. One way to sum these error specifications statistically in the design of industrial products is to use a root-sum-of-squares (RSS) estimation. In the RSS method, each individual error percentage is squared, the results are added together, then the square root of this sum is calculated.

$$e_{\text{sum}} = \pm\sqrt{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2} \tag{2.3}$$

or

$$e_{\text{sum}} = \sqrt{\sum_{j=1}^{n} e_j^2} \tag{2.4}$$

where e is an individual source of error and n is the number of error sources. In using this method, it is assumed that each error acts independently and has an evenly symmetrical distribution. This simple RSS statistical sum solves for an interval of σ, accounting for approximately 63% of the specimens of the product, under the assumptions of the example. The standard deviation, σ, is also the square root of the dispersion.

For a production product, it is more reasonable to solve for a 3σ interval, assuring that nearly all specimens of the product (99.75%) will perform as calculated. Again assuming that the distribution is Gaussian, the maximum error for a 3σ interval is

$$e_{\text{sum}} = \sqrt{3\sum_{j=1}^{n} e_j^2} \tag{2.5}$$

For example, if the 1σ temperature error is 0.21%, the calibration error is 0.13%, and the nonlinearity error is 0.05%, the static error band calculation using each method would be:

simple sum: $0.21 + 0.13 + 0.05 = 0.39\%$

1σ RSS: $\sqrt{0.21^2 + 0.13^2 + 0.05^2} = 0.25\%$

3σ RSS: $\sqrt{3(0.21^2 + 0.13^2 + 0.05^2)} = 0.44\%$

If the error specification is based on a 3σ interval, the error over a σ interval can be listed as the "typical" error.

2.10 TEMPERATURE EFFECTS

Time and again, experience has shown that the greatest impediment to overall accuracy in a measuring and control system is usually due to the temperature sensitivity of the transducer and associated circuitry (unless it is actually a temperature sensor). Developing a temperature sensitivity characterization and implementing changes to reduce this are often a major part of the development program in the design of a new transducer. Commercial products that will likely be used indoors have a typical temperature range specification of 0 to 70°C. Outdoor and industrial sensors have the operating range –40 to 85°C. Automotive sensors have several ranges, depending on where they will be mounted in the car. Engine compartment devices and those near other heat sources, such as the exhaust system or shock absorber orifices or valves, range up to 150°C and sometimes higher. Sensors for use in the passenger compartment can have a more narrow operating temperature range.

In addition to operating temperature range, there may be a storage temperature range, and there will be a temperature sensitivity specification while in the operating temperature range. The storage temperature applies when the sensor is not required to operate and is a survivability specification (any effect on calibration may also be noted). The operating temperature sensitivity specification, however, is sometimes the most important system performance specification. On a linear position transducer, there should be a temperature sensitivity specification for zero and for span. Span is the difference between the zero reading and the full-scale reading (see Section 2.3).

Zero shift is due to thermal coefficient of expansion, the warping and shifting of mechanical components, as well as the changes in offset voltage of op amps, mismatching of temperature coefficients of resistor bridge circuits, and

so on. Span shift is due to changes in gain factor with temperature, which can be mechanical or electrical in origin. To measure span shift, zero shift is first measured. Then full-scale shift is measured. Span shift is obtained by finding the difference between zero shift and full-scale shift.

When developing a new product, it is important to separate temperature errors into those due to zero and those due to span errors. This enables the engineer to have a clue pointing to the source of the errors so the design can be optimized to reduce those errors. Zero shifts can be compensated by selecting thermal expansion coefficients of the materials of construction, for example, or by selecting an amplifier with a low input offset voltage drift with temperature (or chopper stabilized). Span shift might be compensated by using a Ni-Span C spring material, manganin resistance wire, or a resistor or capacitor with a controlled rate of sensitivity to temperature. A coil of nickel wire has been used by the author as a compensating resistor, because nickel has a nearly linear temperature coefficient of resistance of approximately +0.0067 Ω/Ω/°C. Manganin wire is sometimes used in winding the coils of an LVDT because it is made from an alloy that has a very low temperature coefficient of resistance, approximately +0.00002 Ω/Ω/°C [6, p. 75].

2.11 RESPONSE TIME

Response time, of course, is the amount of time elapsed between the application of a change in the measurand to the transducer input and the resulting indication of that change in the transducer output. This simple explanation begs for more detail, though, when trying to account for the actual differences between changes in the measurand and the sensor output signal. Thus, the total response time in a fully damped system may be further divided into a *lag time* before the start of response, a *time constant* based on natural frequency and damping, and a *stabilization* or *settling time* while the final reading is being approached (see Figure 2.11).

The lag time is the time that passes between the start of a change in the measurand and the start of a change in the sensor output ($t_1 - t_0$ in Figure 2.11). This can be due to propagation delay in electronic components, or the equivalent in mechanical, pneumatic, or other types of components. The time constant or main component of response time is usually based on the natural frequency of the sensing element, the maximum time between samples of a sampling type sensor, or the frequency of a filter somewhere in the signal path. Coupled with a damping factor, this results in most of the specified response time. It is usually specified, with a step input, as the time between the start of response until reaching 63% of the final response, $t_2 - t_1$ in Figure 2.11.

Final output is typically the level that would be indicated after waiting the lag time plus five time constants for stabilization of the output after a step change in the measurand ($t_3 - t_0$ in Figure 2.11). The amount of change in

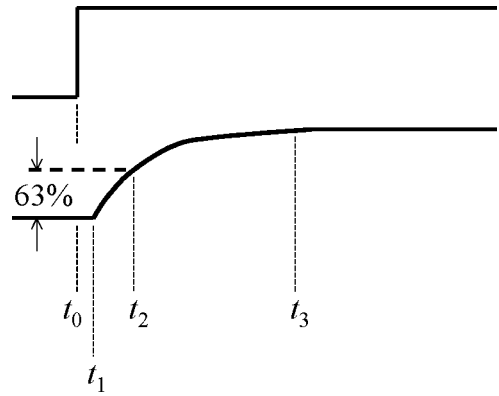

Figure 2.11 Input measurand (upper trace) and output signal (lower trace) versus time, showing lag time, time constant, and stabilization.

output after waiting one to five time constants is usually determined by a combination of mechanical damping and electronic filtering. Most position transducers include a low-pass filter in their output circuit. Higher-order filters offer a sharper cutoff rate and therefore can have a calculated operating frequency (f_o) that is closer to the natural frequency of the sensor system, thus having a shorter settling time than could be had when using a lower-order filter. This can appear as a faster response time or as a reduced error for a given frequency of variation in the measurand. A Butterworth filter solution is normally used for a maximally flat amplitude response, whereas a Bessel solution can be used for a constant phase shift over a range of frequency. A Tschebychev filter solution will offer a faster falloff rate with changing frequency, but at the expense of adding a signal amplitude ripple in the passband. This is normally not desired in a position transducer and can give excessively higher error when velocity or acceleration signals are derived from the position signal. An expedient means for deriving these various filter solutions is presented in reference 15.

Alternatively, sometimes response time is stated as the time between 10 and 90% of the final output response to a step input. This is less specific and may require testing to verify that the performance is suitable for your application if the response time is critical. The response time information given so far is based on the output amplitude as a percentage of the expected amplitude. In real-time feedback systems, it may also be important to look at the phase lag from the measurand input to signal output, in addition to the output amplitude variation. The combination could be specified, for example as –3 dB at 1 kHz with 10° phase lag. This would mean that with the measurand varying as a sine wave of frequency 1 kHz, the transducer output voltage would be 0.707 of the theoretical output and delayed by 10° as compared to the measurand.

2.12 OUTPUT TYPES

Transducer outputs are supplied in many variations of analog and digital formats. Popular *analog* outputs include 0 to 10 V dc, ±10 V dc, 0 to 5 V dc, 4 to 20 mA (occasionally, 1 to 5 mA), 10 to 90% ratiometric, 5 to 95% ratiometric, frequency, timed pulse, and pulse-width modulation (PWM).

Timed pulse and PWM are often called *digital outputs* because they are suitable to be interfaced directly to digital circuits but are, in fact, analog signals because they can be continuous with no quantization. Timed pulses and PWM do, however, usually provide their signal at the same voltage levels as digital signals (typically, where 0 V dc represents a logic low level and +5 V dc represents a logic high level), or alternatively, at voltages and impedances according to various differential signal standards.

Voltage output circuits, including 0 to 10 V dc, ±10 V dc, 0 to 5 V dc, 10 to 90% ratiometric, and 5 to 95% ratiometric are either operated into a high-impedance circuit or may have a load resistor within the customer's application circuit (see Figure 2.12). The transducer manufacturer provides a minimum load resistance specification. If the customer applies a load resistance lower than this, the output performance can be degraded, due to the limitation on current driving capability of the output amplifier. Current loop output circuits, including 4 to 20 mA and 1 to 5 mA, are operated in a low-resistance circuit or may have a precision load resistor within the customer's application to convert the current into a voltage (see Figure 2.13). A 4- to 20-mA transducer is commonly used with a precision 250-Ω load resistor to convert the output to 1 to 5 V dc. (This use still preserves the main advantage of using a current loop; voltage drops along the length of the cable are ignored.) The transducer manufacturer provides a maximum load resistance specification. If the customer applies a load resistance higher than this, the output performance can be degraded, due to the lack of a sufficiently high voltage to drive the output current.

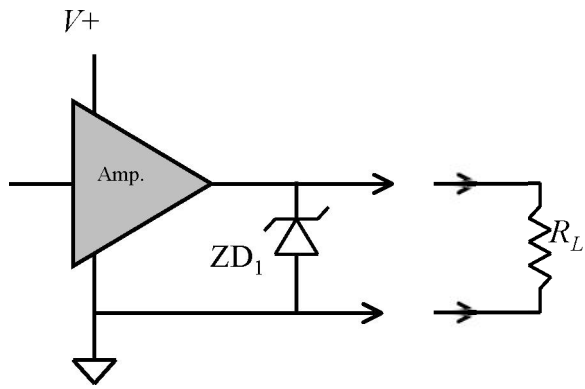

Figure 2.12 Voltage output transducer with load resistor, R_L. ZD_1 provides protection of the amplifier against electrostatic discharge (ESD).

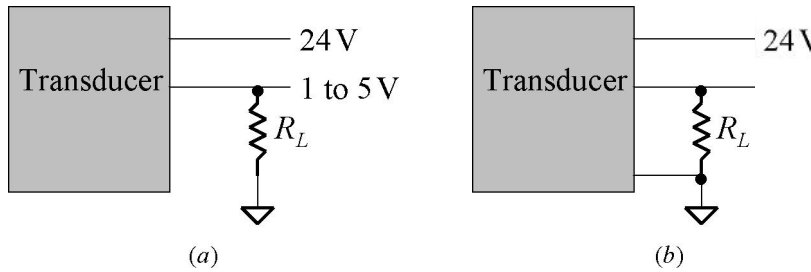

Figure 2.13 4 to 20 mA current output transducers: (*a*) two-wire; (*b*) three-wire.

Transducers with a current output can be operated with a separate power supply input (three or four wires), or the power and signal can be incorporated on one pair of wires (two wires). The single pair type is called a *loop-powered transmitter, two-wire transmitter*, or *transmitter*. The advantage of a loop-powered transmitter is that only two wires need to be run, saving installation cost. This is especially important with intrinsically safe systems, where each nongrounded conductor going into a hazardous area must be protected by a safety barrier device (see Section 2.16). With a loop-powered transmitter, the transmitter (a position transducer, for example) is powered over the same pair of wires as are used to indicate the transducer signal. The minimum signal level of 4 mA is sufficient to operate the transducer and indicates the minimum reading of the measurand. A larger measurand reading is produced by the transducer circuit drawing a greater amount of current.

Transducers with a ratiometric output have an output voltage that varies as a percentage of the power supply voltage to indicate the value in the measurand. For example, with a 10 to 90% ratiometric transducer having a 5-V power supply, a zero measurand input will produce an output voltage of 0.5 V dc, or 10% of the 5-V dc power supply. The output will vary from 0.5 to 4.5 V dc for a measurand change of zero to full scale. The advantage is that a voltage reference is not required in the transducer or the customer's receiving circuit. Voltage references are expensive and have their own error specification, which will add to the total system error. Since the transducer and the customer circuit refer to the same power supply voltage with a ratiometric indicating system, there is no additional error due to variations in voltage reference. However, when making laboratory measurements on a ratiometric transducer, a power supply voltage reading must be taken together with each output voltage reading. This is because the lab power supply may have small short-term variations, and the test meter will read this as transducer error, since the test meter normally is not ratiometric and also has error from its own internal voltage reference (which would not be a factor with a ratiometric reading).

The output types presented so far have been analog. In addition, there are many digital formats in wide commercial and industrial use, so a few common ones are presented here. Some popular digital protocols, suitable for use in

communicating with transducers, include SSI (serial synchronous interface), CANbus (controller area network), Profibus (application profile), and HART (highway addressable remote transducer). Each of these is described in their own fairly complex manual of hardware and software interface, so only short explanations are given here.

SSI was developed as a serial interface technique for use with absolute encoders in order to transfer data from the transducer to a controller. It was developed by Stegmann Corporation, an encoder manufacturing company in Germany. This interface option is available on many controllers and programmable logic controllers (PLCs). Absolute encoders generally produce a parallel output in Gray code (see Chapter 10). The SSI protocol allows serial communication of the parallel data available at the encoder using a very simple format. This technique is also available with other position transducer technologies, such as magnetostrictive linear position transducers. An SSI connection system comprises two power, two clock, and two data lines (wires). The data lines connect a shift register in the transducer to a shift register in the user's application circuit, through suitable voltage-level and impedance-matching circuitry. The user's application circuit sends clock pulses on the clock lines to shift the data out of the register located within the transducer and into the register located in the application circuit. The register length is usually 24 or 25 bits but can vary, depending on the type of transducer. The range of clock rates that can be used is specified for the transducer and varies with the cable length. A data transfer rate of up to 1.8 MHz is possible with a 15-m cable. Longer cables, up to 300 m, can be used if the clock rate is limited to 100 kHz. After all of the data bits are transferred, a synchronization period follows. The data line state remains "high" and no data are transferred during this time. The synchronization period is longer than the period of clock frequency. Synchronization is then possible by knowing that the next pulse on the clock lines after the synchronization period will be pulse 1. The first pulse (the first high-to-low transition) is the signal for the transducer to latch its data (i.e., to cease taking new measurements and freeze the data that are in the register). The next low-to-high transition is the start bit. Then the following sequence of high-to-low transitions shifts out the data. After the data are transferred, the clock line remains high for at least the minimum synchronization period. The cycle is repeated at a rate set according to the internal update rate of the transducer and the requirement of the application for new data. Differential driver/receiver circuits and termination resistors are used to limit the possibility of electrical interference. This is not a bus connection, but is a one-to-one connection between the transducer and controller. Data flows one way, from the transducer to the application (user) circuit.

CANbus was introduced at the SAE congress in Detroit in February 1986. Robert Bosch GmbH developed the CANbus communication bus system on behalf of BMW and Mercedes for use in automotive applications. As a high-speed serial data network, it was designed to replace wiring bundles and to provide connections among distributed controllers. It is now an international

standard (ISO 11898). Since it is a bus connection, several devices can be connected in parallel to the same set of four wires, two for power and two for signal. There are several modes of operation. In one mode, the user application circuit (we will call this the *controller*) sends out an address to all the transducers. The transducer that has that address acknowledges and either sends its data or waits for further instruction. If the controller sends a further instruction, the transducer responds with the corresponding information. In another mode, any device on the bus can send a request or data to any other device on the bus in a similar manner. In this mode, there could be a time when two or more devices try to use the bus at the same time. Called *bus contention*, this is solved based on a priority level assigned to each device. As each bit of a device address or instruction is placed on the bus consecutively, each device continues to listen until it reads a bit that does not match its own address or that does not match the instruction or data it was attempting to send, in which case it lets go of the bus. If a device is trying to send a message but is unable because the bus is in use, it will try again after a pause.

Profibus is an open digital communication system, which has been very popular in the fields of factory automation and process automation. It is a bus system based on *profiles* (also called *application profiles*)—hence the name Profibus. The profiles are specifications from manufacturers and users, which define performance features and other properties of transducers, devices, and systems. Profibus began in 1987 in Germany in cooperation with the German government, to establish a serial data communication fieldbus and standardize the interface to be used with field devices. The Profibus FMS (fieldbus message specification) protocol was defined first and specifies the most demanding communication requirements. Later, the Profibus DPs (decentralized peripherals) was defined for a simpler, faster configuration. By 1995, the system gained popularity.

The physical layer is commonly implemented using an RS485 type of hardware connection. This requires four wires: a twisted pair for data plus two wires for power. Other wired options include RS485-IS (intrinsically safe), MBP (Manchester coding, bus powered), MBP-IS, and MBP-LP (low power). The MBP connections are two-wire, having power and data transfer using only one pair of wires. The MBP data transmission rate is much slower than the RS485 rate.

Another option for Profibus data transfer is by use of a fiber optic cable instead of wires. This data transmission means may be selected when the system will operate in an environment of high electromagnetic field intensity (which could otherwise cause electrical interference with the data), high-voltage potentials (fiber optic cables provide electrical isolation), long transmission distances (over 10 km), or high data transfer rate. Since up to 32 devices can be installed per node, a station address is assigned to each device. This can be programmed as a *hard address* using switches within the device, or as a *soft address*, while assigning parameters in programmable memory during system startup.

HART (highway addressable remote transducer) protocol was originally developed by Rosemount in 1985. Ownership of the HART technology was transferred to the HART Communication Foundation, a nonprofit organization, in 1993. HART was developed as an open protocol for the purpose of improving the two-wire transmitter system, commonly in use for field devices at the time, by adding bidirectional digital communication capability. Utilizing a combination of analog and digital techniques, a HART transmitter can operate in a 4- to 20-mA analog loop just as described earlier. In addition, the transducer can be instructed or queried by using digital communication on the same two wires. The digital words are impressed on the analog lines by low-level sinusoidal oscillations that do not affect the indication of meters reading the 4- to 20-mA signal. The oscillations are frequency-shift keying (FSK) signals according to the Bell 202 telephone standard. A logical 0 is 2200 Hz, a logical 1 is 1200 Hz. Communication circuits within the transducer filter and detect the oscillations, converting them back into digital signals, which are usually read by a microcontroller. The main advantages of HART are that additional capability can be added to already installed two-wire systems, and the system is backward compatible with the installed base of two-wire 4- to 20-mA loop transmitters. Although the transducers can operate in standard current loops, as described, they can also be set to operate in a multidrop mode. In this case, the current is set to a constant low level, and up to 15 devices can be connected in parallel. Then only digital communications are used.

A HART data transaction comprises a master command and a slave response. Communication access is governed by token passing among the devices on a channel (a *channel* is a pair of wires). The transmitted message includes the information for the next passing of the token. If no communication takes place within the period of a preset timer, the token expires and control of the channel is again open. The protocol supports various device parameters, including process variable values, status and diagnostics, device identification, and calibration information. Cable lengths of up to 3000 m can be used, depending on the operating mode (standard or multidrop), individual product specifications, and cable specifications. Intrinsically safe transducers can be implemented with proper installation of the rated safety barriers.

This protocol supports a substantial installed base in process control applications and is currently the leading communication protocol for process instrumentation. It is not generally suitable for motion control and has not gained a large following throughout the industrial world, because it is relatively slow compared to other digital communication techniques.

2.13 SHOCK AND VIBRATION

Most position transducers will undergo some level of shock and vibration exposure during normal use. This specification is used to qualify the sensor for

two purposes: How will the normal shock and vibration levels affect accuracy, and how long will the transducer last under higher levels of shock and vibration before failure?

The installation components and technique can make a difference in the performance and survivability of the transducer in response to a given vibration input, thus increasing or decreasing the susceptibility to shock and vibration. Mounts made from a dampening type of elastomeric compound may improve the reliability of a position transducer, for example, by reducing the peak amplitude of the vibration, but may also reduce the transducer accuracy by allowing motion to occur between the sensor mounting flange and the rigid sensor mount (the elastomer being inserted between these two).

Sometime during the development of a position transducer, a sample should be mounted, in the intended way, into a vibration test fixture. The performance and survivability are tested over an extended time at several discrete frequencies and amplitudes. A frequency scan is also done to find any tendencies of the transducer to resonate at particular frequencies. If any resonant frequencies are found, the design can be modified to remove or dampen the undesired resonance. Alternatively, the transducer can be tested for an extended time at each found resonant frequency to make sure that the proper performance and reliability are maintained. A typical vibration machine setup is shown in Figure 2.14. Note that the vibration system comprises a signal source, power amplifier, forcer, transducer mounting fixture, and an accelerometer. Cooling is also normally required. The vibration level and frequency are

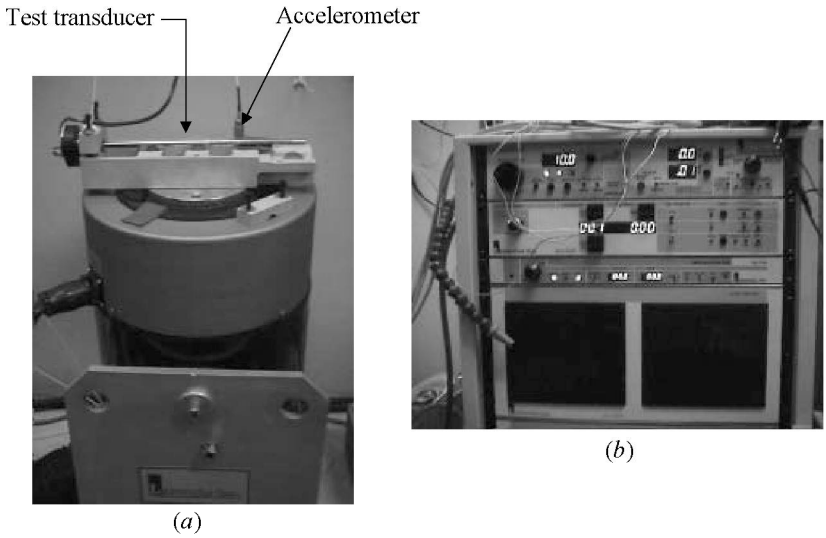

Figure 2.14 (*a*) Transducer mounted with a fixture to a vibration forcer; (*b*) source and amplifier. This setup utilizes forced air cooling. An accelerometer is mounted to the test fixture.

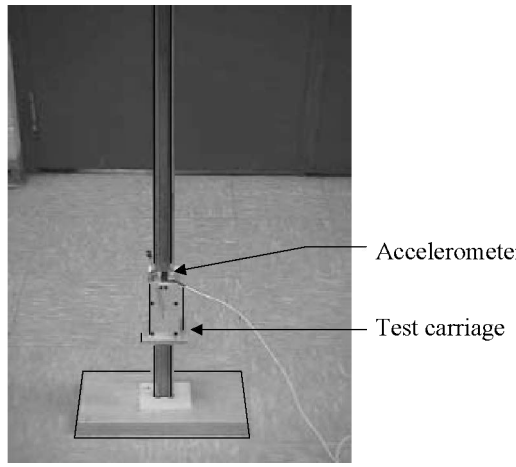

Figure 2.15 Drop-test shock tester, with accelerometer attached.

confirmed by mounting one or more accelerometers to the transducer or mounting fixture, wired to a driver/readout device.

The vibration testing conducted in the development lab is done to make sure that the design does not contain any major flaws. An additional vibration test is usually conducted at a certified testing lab to make sure that industry standards are met. At the same time, mechanical shock tests are performed. It is not common for a development lab to have shock testing equipment on site. Shock testing can be done by dropping the transducer and fixture a predetermined distance into a calibrated stopping area (see Figure 2.15), or a moving mass can be slammed into the transducer mounting fixture. The author has also visited a facility where larger equipment was submerged in water and explosive charges were used to produce the shock. A related test is the drop test. This can usually be performed in the development lab. The transducer is dropped from a predetermined height (usually, desktop height) onto a concrete floor. It is dropped at least six times, once on each normal face (top, bottom, four sides). In addition, if there is a particular angle in which the transducer could be expected to be more sensitive, dropping at this angle is also tested.

2.14 EMI/EMC

The *electromagnetic compatibility* (EMC) of a transducer is a rating of how well it will operate in, and not substantially contribute to, an environment of electromagnetic radiation. When electromagnetic radiation emitted from one or more devices affects the performance of another device, it is called *electromagnetic interference* (EMI). EMC is divided into two broad areas: emis-

sion and susceptibility. A well-designed transducer will not emit sufficient levels (emission) of electromagnetic energy to disturb the performance of other devices, nor will it be affected (susceptibility) by EMI emitted from other devices at reasonable levels.

Industrial product specifications also include ratings for resistance to electrostatic discharge (ESD), electrical fast transient (EFT) or "burst", and lightning surges. ESD is a high-voltage discharge similar to that experienced when one walks across a carpet and touches a doorknob. Voltages of more than 14 kV are possible. ESD is normally simulated by applying the specified voltage while using a small coupling capacitor that simulates human contact. The EFT test simulates the burst of high voltage/high range of frequencies that can occur when a set of relay contacts open. Arcing between the contacts as they open can produce high-voltage spikes over a range of frequencies, especially when the closed contacts were passing a substantial amount of current through an inductive load.

Lightning-induced surges do not include direct lightning strikes (as these can typically melt or vaporize portions of the equipment). Rather, a lightning surge test simulates the increased voltage levels that may be induced between conductors when a lightning strike occurs in the general area. Nearby strikes induce higher voltage peaks for shorter durations. Strikes farther away produce more moderate voltages for a longer duration. A lightning surge test can typically have a peak voltage in the range 600 to 5000 V and a duration of 1 ms to 10 μs, respectively. Durations of longer than 1 ms for moderate voltages can be simulated by multiple strikes. The waveform of the applied test voltage is a dual exponential one. There is an exponential rise time to the peak voltage, followed by an exponential decay back to zero over a longer time period. The surge immunity test is described in IEC 1000-4-5.

FCC (Federal Communications Commission) part 15 addresses EMI to some degree, but the standard to meet is generally considered to be the EMC portion of the requirements for the CE Mark (Communauté Européene). Since 1996, all products that are liable to cause, or be affected by, EMI must have the CE mark if they will be sold in any of the European Union countries. This includes all electrical and electronic products. For the purposes of compliance testing, electromagnetic energy may either be conducted or be radiated through space. When conducted, the energy comes into the device through the connection leads or through the case. The EMC standards for the CE mark come from the International Electrotechnical Commission (IEC). These are the IEC 1000 series of standards, formerly known as the IEC 801 series.

Circuits to provide the protection needed to meet EMC requirements include those contained within the integrated circuits, as well as additional components mounted to the printed circuit board (PCB), housing, and/or connectors. Integrated circuits that make external connections via the input and output leads must be able to withstand the EMI, ESD, and EFT, or components must be added to ensure compatibility. Back-to-back fast zener diodes

can be used to limit peak voltages. (Back-to-back connected diodes normally mean that two zener diodes are connected in series with their anodes connected together. One cathode goes to common or ground, and the other cathode goes to pin to be protected. Connecting the cathodes together, instead, would also work.) Sometimes additional impedance is needed in series with the zeners to limit peak current, thus avoiding damage to the zeners when a high-energy voltage spike is encountered. The impedance can be a resistor or may be an inductor to aid in reducing the passage of higher-frequency interference.

An inductor is often formed by adding a ferrite bead to a wire lead coming into the PCB. A parallel connection of a ceramic capacitor (typically in the range 0.1 to 10 nF) can be added and serves to shunt high-frequency energy to ground. These shunt capacitors can be connected between the transducer lead and the circuit common but are often connected between the transducer lead and the case. Transducers with high-frequency signals appearing in their sensing elements, microcontrollers, or communication circuits need special attention to avoid the "leaking" out of this energy and causing EMI. Switching power supply regulators also produce voltage spikes that can be radiated or conducted to the circuitry being supplied. The amplitude of the voltage spikes conducted can be reduced by ramping the switch points instead of switching sharply (some integrated circuits are available that provide this function). This also reduces the radiated EMI. Another way to reduce conducted EMI is to use a higher switching frequency (more than 1 MHz) so that smaller filter components can be used to attenuate the spikes. Radiated EMI can be spread across a wide frequency range (called *spread spectrum*) by constantly varying the switching frequency. This limits the energy in any given frequency, making it more likely to pass the CE tests.

Protection against a lightning-induced surge must include a larger energy-handling device. The voltage level may not be as high as with ESD, but the energy level may be higher with lightning surge than with ESD because of the longer time period during which the voltage may be present when due to lightning surge. In addition to the use of zener diodes, lightning surge protection often includes the use of spark gaps or metal-oxide varistors (MOVs) to handle the higher energy. Although MOVs are less expensive than fast, high-surge-current zeners, their performance can be degraded after handling multiple surges.

Testing to make sure that a transducer being developed will meet EMI requirements usually takes place in two steps: in-house laboratory testing and third-party lab testing. For most engineering companies, an anechoic chamber and a full set of EMC testing equipment to guarantee meeting the latest industry specifications is not cost-effective. Instead, most development labs have some basic equipment that can be used to get a good indication of whether or not the transducer will pass the "real" test. This lab equipment can also be used to troubleshoot a new product design for EMI-related problems.

After initial testing in the development lab, the transducer to be tested is sent to a third-party test lab [a third party (not the developer, not the customer) that is certified to perform the testing]. This lab supplies a test report that can be used to prove to the customer that the transducer was tested and meets the specified requirements for EMC.

2.15 POWER REQUIREMENTS

Older industrial equipment often required ±15 V dc power. This was needed to operate the analog amplifiers within the transducers, controllers, and so on. With the availability of lower-voltage and lower-power amplifiers beginning in the mid-1970s, however, the requirement for the negative supply voltage was reduced. By the 1990s, the negative power supply voltage was rarely used. It has since become the standard for basic industrial transducers to be connected into either +15- or +24-V dc power systems, the negative side of the power supply being called *common* (and often grounded at some point in the system). To enable a transducer to operate with either a +15- or a +24-V dc power supply, it should be designed for operation from 13.5 to 26.5 V dc. This is derived (approximately) as 15 V minus 10%, and 24 V + 10%. The additional 10% on both ends of the supply voltage range allows for variation in calibration from one power supply to the next, changes in the supply voltage with line and load variation, and it allows for long-term drift of the power supply voltage.

Mobile equipment, on the other hand, usually operates from a nominal 12- or 24-V battery system, and 42-V automotive systems are expected soon. Since a mobile environment has a charging system, a high starting load, and possible disconnection of the battery (which is otherwise normally providing a load to the charging system), a much wider range of power supply voltage must be accommodated than just the rated voltage of the battery. For example, in a 12-V automotive system, the battery voltage during cranking could be 8 V or lower, while normal operation of a transducer with this supply voltage is still expected. When the battery is being charged at normal cruise conditions, its voltage can be more than 14 V. Under conditions of *load dump* (which is when the charging system is operating normally, and the battery connection is broken abruptly), the system voltage can peak at several times the battery nominal voltage. So a complete specification of possible power supply voltages and durations is needed for the proper design of a transducer that will be used in a mobile application.

Common to all transducers, no matter which power supply type is specified, is the need for adequate circuit design to provide protection against reverse polarity of the power supply connections, overvoltage, and shorting of the output leads to power or ground. In addition, many mobile applications require protection against any combination of miswiring of the input and

output leads, including the shorting of any of them to power or ground potentials. Sometimes it is possible to limit the added expense of overvoltage protection in a transducer for an automotive application by specifying that it be connected to the 5-V regulated power available from the on-board controller or engine control unit (ECU).

2.16 INTRINSIC SAFETY, EXPLOSION PROOFING, AND PURGING

Transducers of various types must often be able to operate in areas that may experience a hazardous (flammable or explosive) atmosphere intermittently or continuously. *Intrinsic safety* is a method used to prevent combustion or explosion by removing the likelihood of the presence of ignition energy. *Explosion proofing* (also called *flameproofing*) is a method used to contain an explosion should it occur. *Purging* is a method that removes hazardous gas from the area surrounding the electrical equipment. These are the three main explosion prevention methods and are described here, although other methods are also used, such as encapsulation or sand immersion.

For the purposes of fire and explosion prevention, a hazardous atmosphere is one that contains flammable or explosive materials that can be ignited by a sufficient energy source when an oxidizer is present in a quantity capable of supporting combustion. The flammable or explosive material may be flammable gas or vapor (such as methane gas or alcohol vapor), flammable dust (such as coal dust or wheat flour), or flammable fibers or flyings (such as textile fibers). Liquids with low-flash-point temperatures (e.g., gasoline) generate flammable vapor at normal temperatures; so these vapors are therefore in the category of flammable gas or vapor. (Flash point is the lowest temperature at which a flammable liquid generates vapors at a rate high enough to support combustion in the presence of sufficient oxidizer and an ignition source.)

In North America, the classifications for hazardous locations, and specified in the United States by the National Electrical Code (NEC) are divided into three classes: class I (gas or vapor), class II (dust), and class III (fibers or flyings). In addition to the class, hazardous locations are separated into two divisions. A division 1 location is one where the hazardous condition may exist under normal operating conditions, such as in the vicinity of a process vessel having a hatch that can be opened during normal processing. A division 2 location is one where the hazardous condition can exist only under abnormal conditions, such as during repair, an accidental breach of a seal or failure of a purge system. Areas other than division 1 or 2 are labeled as nonhazardous. Divisions 1 and 2 are similar, but not identical, to European zones 1 and 2, respectively. In addition, there is a European zone 0 where the hazardous condition is expected to exist continuously or for long periods during normal operation.

The flammable or explosive materials of class I and class II locations are further broken down into groups, based on the minimum ignition energy of

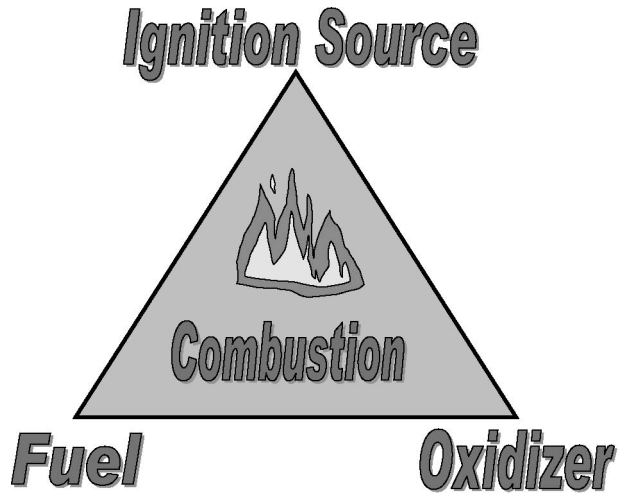

Figure 2.16 Combustion triangle. (From Ref. 18, p. 24.)

the most easily ignitable mixture of the material with air. For example, class I, groups A through D, include the following gases: A, acetylene; B, hydrogen; C, ethylene; and D, methane. As may be obvious, they are arranged from A, the most easily ignitable, through D, the least easily ignited. Similarly, examples of class 2 dust groups include E, aluminum dust, F, coal dust, and G, grain dust. Other gases and dust are included in these classes; the examples listed here are only representative.

The elements required to initiate combustion are shown in the combustion triangle, Figure 2.16. As depicted, a fuel, an oxidizer, and an ignition source are needed before combustion can occur. If any one of these elements is absent, fire or explosion is prevented. The starting point for evaluation is the fuel. A certain minimum amount of fuel must be available to support combustion. With a gas or vapor, this minimum concentration is called the *lower explosive limit* (LEL). If the fuel concentration is below the LEL for the type of fuel present, combustion cannot take place. In some special cases in closed containers, combustion can be prevented by keeping the fuel concentration above the *upper explosive limit* (UEL). Above the UEL, the fuel/air ratio is higher than that capable of supporting combustion (i.e., there is enough fuel but not enough oxidizer present).

If fuel is available in a concentration between the LEL and UEL, one needs to consider removing the oxidizer and/or ignition source. In practice, an ignition source could be present in the form of heat or an electric spark, and the oxidizer is atmospheric oxygen. Atmospheric air at zero relative humidity is approximately 20.94% oxygen (O_2) and 78.09% nitrogen (N_2), the remainder being argon (Ar), carbon dioxide (CO_2), and other gases. These percentages by volume are independent of the atmospheric pressure, but the partial pres-

sure due to each gas (expressed in absolute pressure) is proportional to the barometric pressure. Even with fuel present, combustion can be prevented by limiting the oxygen level to below the minimum concentration needed for combustion. Called *inerting*, this is normally accomplished by forcing nitrogen (a relatively inert gas) into the container to be protected, to displace enough air to reduce the oxygen level below that needed to support combustion. The author developed a system for this purpose in the early 1980s, which became a standard for fire prevention in the chemical, pharmaceutical, and paint mixing industries. This system measured the oxygen concentration within, and controlled the flow of nitrogen into, a process vessel by using a closed-loop measuring and control system. This system reduced the oxygen level to below the concentration needed to support combustion.

When fuel and oxygen are both present, making sure that sufficient energy to cause ignition is not availale can prevent combustion. This is the principle behind intrinsic safety.

Intrinsic Safety

In England in 1913, a gas explosion in a coal mine caused the loss of many lives [10, p. 5]. It was believed that the ignition was caused by a system for signaling the surface crew that a coal car was ready to be brought up. The signal was energized by shorting two contacts with a shovel. After a lot of research and work, it was determined that this problem could be avoided by limiting the energy available for a spark. This was the birth of the idea of intrinsic safety. When a device is rated as *intrinsically safe* (IS), its operating voltage, current, energy storage, and temperature are low enough that it is incapable of igniting the atmosphere when used according to the connection diagram recommended. An IS-rated device is designed so that it does not have any hotspots, has a low enough outside case temperature for the rated atmospheric gases, and does not store energy in excess of the prescribed level. The IS device is rated to operate in conjunction with one or more types or models of IS barriers. An IS system normally includes a series device called an *IS barrier* installed between the nonhazardous and hazardous areas (see Figure 2.17). This is mounted in the nonhazardous area and makes sure that any energy passed into the hazardous area is at such a low level that it cannot ignite the hazardous atmosphere.

IS barriers can be passive or active. The schematic of a single-channel passive IS barrier device (also called a *zener barrier*) is shown in Figure 2.18. This is the simplest type of IS barrier and is shown to give the reader a basic understanding of the theory. It includes a first zener diode, ZD_1, to limit the circuit voltage. A second zener diode, ZD_2, is included for redundancy in case the first one fails. A resistor, R_2, is connected between the zener diodes so that it is possible to test each diode. A resistor, R_1, is in series on the nonhazardous side to limit the current in an overvoltage condition, protecting the zener and limiting the temperature. A fuse, F_1, is also in series on the nonhazardous side

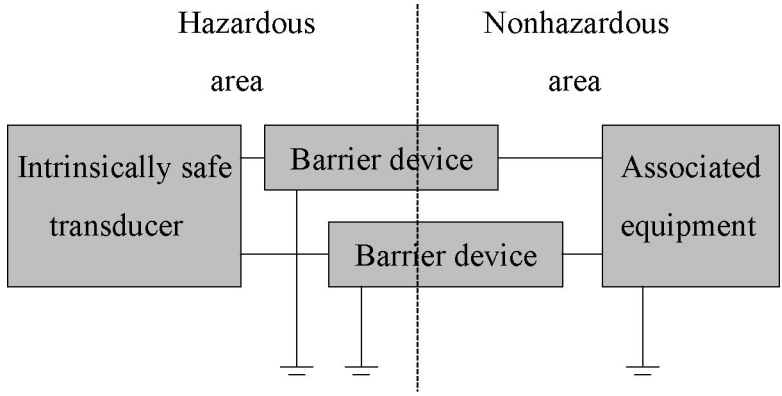

Figure 2.17 Intrinsically safe barrier devices installed to protect a hazardous area.

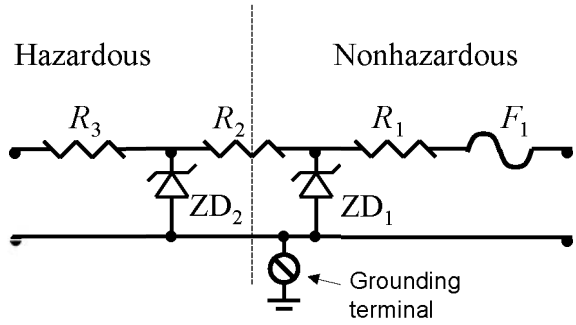

Figure 2.18 Schematic of a single-channel passive IS barrier.

to protect the zener, ZD_1, in case of a gross overvoltage from the equipment on the nonhazardous side, such as an inadvertent connection of line power (115 or 250 V ac) instead of the 24 V dc for which the usual system is designed. The fuse must be encapsulated to prevent it from becoming an ignition source, and in fact, the complete IS barrier device is typically encapsulated.

A resistor, R_3, is in series with the hazardous side to limit the current to a safe level, at the voltage rating of ZD_2, for the group rating of the expected flammable materials. Barrier devices are normally designed to be mounted on a metal plate or DIN rail (DIN is a European standard, Deutsches Institut für Normung). In a passive barrier, it is required that the ground connection be wired to the appropriate IS ground point of the system with a resistance of 1 Ω or less. This is sometimes a difficult task to accomplish reliably and is one reason why active barriers are popular.

An active barrier includes electronic circuitry to perform more complicated functions or to provide added convenience, in addition to providing intrinsic

safety. An active barrier can provide galvanic isolation between the hazardous and nonhazardous parts of the system by using oscillator and demodulator circuits with inductive, capacitive, or optical coupling. This removes the need for a zener barrier ground connection of less than 1 Ω. Active barriers are also available for interfacing with switch inputs, annunciator outputs, or conversion between voltage and current signals, and communication with various digital protocols. Each barrier device, whether passive or active, is labeled with the safety ratings, approval body, and limitations.

A simple device itself may not be required to carry an IS approval if it is used with an IS barrier approved for such use. For the purpose of IS requirements, a simple device is one that does not have energy storage capability or an elevated temperature. This includes switches, light-emitting diodes (LEDs), thermocouples, and photocells. Most other electrical devices must have approval from an accepted approval agency before they can be used in an approved IS system. In North America, the primary approval agencies are UL (Underwriters' Laboratories), FM (Factory Mutual), and CSA (Canadian Standards Association), but many smaller agencies are also registered. The IS standards among these agencies are similar and are UL 913, FM 3610, and CSA 22.2, respectively. There is some acceptance of these agencies worldwide, but typically, an approval for the country of use is required. Nineteen European countries are members of CENELEC (the European Committee for Electrotechnical Standardization). There are also many other affiliated countries worldwide. CENELEC members have, by regulation, agreed to accept each other's approval systems. Even so, it is common for a customer to require the approval of his or her own country's agency: PTB in Germany, for example.

Outside North America, the classifications are slightly different from those already described. The North American division 2 is somewhat similar to the European zone 2 (but not exactly identical), division 1 is similarly compared with zone 1. There is a zone 0 in Europe, which is assigned to areas where explosive mixtures are expected to be present continuously or present for long periods.

Explosion Proofing

In an explosion-proof housing, it is assumed that hazardous gases are not prevented from entry into an electronic equipment enclosure. It is also assumed that it may be possible for the device or wiring contained to cause an ignition. The housing is designed to be capable to withstand the resulting explosion or increase in pressure due to the ignition. The enclosure also has a pressure relief path, which vents the pressurized gas while cooling the gas to a temperature below the ignition temperature of the external atmosphere (see Figure 2.19).

The pressure relief and gas-cooling path is called a *flame path*. It is usually designed to include an area of parallel metal surfaces such that a gas-fueled flame passing along the path will be sufficiently cooled before escaping the housing. A sintered metal plug is sometimes used instead for this purpose.

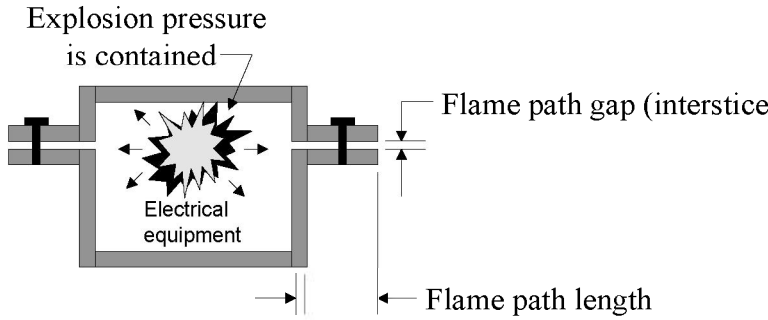

Figure 2.19 Explosion-proof housing.

After sufficient cooling, the gas will no longer be burning and will be cool enough that it cannot ignite the atmosphere external to the explosion-proof housing. The parallel metal surfaces, which naturally have a relatively high thermal conductivity, are designed to have a small gap between them, called the *interstice*. The flame path length is made long enough to provide for the needed cooling time. The flame path also provides the means to allow the high gas pressure (from the combustion) to escape. This is important so that the service person does not potentially risk injury due to removing the cover from a pressurized vessel for maintenance or repair of the equipment contained within. The flame path allows equalization of the pressure inside the enclosure with atmospheric pressure, so that after combustion or explosion takes place inside an explosion-proof enclosure, the pressure subsides quickly.

Purging

A third alternative for fire prevention is the use of a nonflammable gas to displace any flammable gas that may be present. This method is called purging. In Europe, the standard is European Norm (EN) 50016. In the United States the standard is National Fire Protection Agency (NFPA) 496. The NFPA specifies three types of purge for use with installations having various likelihoods of the presence of a hazardous gas: X, Y, and Z. A type Z purge system reduces the enclosure hazard rating from division 2 to nonhazardous. A type Y purge reduces the rating from division 1 to division 2, and a type X purge reduces the rating from division 1 to nonhazardous. Types Z and X are the most popular, because type Y reduces division 1 to division 2, and then division 2 standards must still be met, whereas with type Z or X, the classification is then nonhazardous. Schematics of typical type Z and X purge systems are shown in Figure 2.20.

The type Z purge is simpler and requires some attention by the operator. That's why it is only for use with division 2 installations, since there is a lower potential for danger than with division 1. In a type Z purge, an operator applies

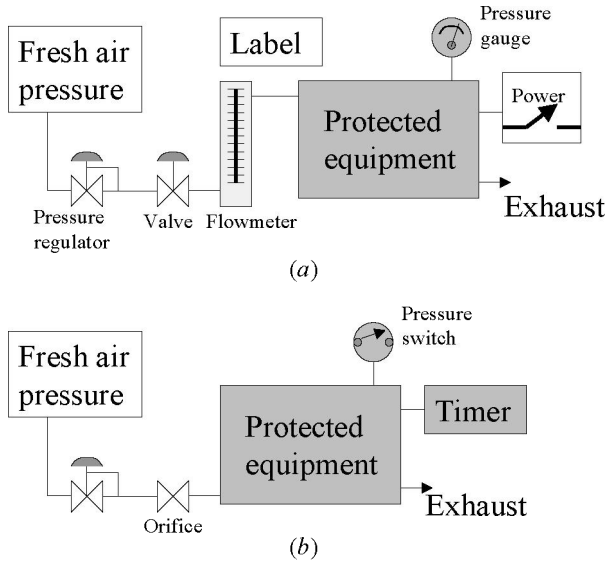

Figure 2.20 Purging systems: (*a*) type Z; (*b*) type X.

the purge gas at a specified flow rate, by operating a valve while reading a flow meter, for a predetermined amount of time before powering the system by operating the power switch manually. This assures that the potential hazardous gas has been displaced sufficiently by fresh air (or another nonflammable gas) before the presence of an ignition source is possible. If the purge is lost, the operator notices this on the flow meter and turns off the power. For a type Z purge, a label must be placed at the location of the protected equipment, listing a warning and the operating instructions.

In a type X purge, an automatic system is used to allow purge gas to flow for a predetermined time at a minimum pressure before the purge-protected equipment is energized. If the purge pressure in the enclosure is lost at any time, the power to the protected equipment is switched off. Sometimes a controlled leak is included to reduce the purging time. In any purge system, the electrical equipment is contained within a reasonably sealed enclosure. It does not have to be airtight but must be able to hold the required amount of purge pressure. During the initial purging time, the equipment is not yet protected, so that any electrical equipment operating at that time (e.g., timer, pressure switch) must have another means of protection. This protection is normally by means of encapsulation, explosion proof, or intrinsic safety.

Electrical equipment can be protected by any one of the methods described, but sometimes a system may be protected by two or more methods used together. An example would be using an explosion-proof or purged housing to contain intrinsic safety barriers within a hazardous area.

2.17 RELIABILITY

The probability that a product will perform and be free of failure is called *reliability*. Intrinsic reliability of a product is a function of the quality of design and indicates the reliability that is theoretically possible. Achieved reliability may be different, and is usually lower than the intrinsic reliability, due to factors such as unexpected conditions during use (environmentally or customer induced), a temporary lapse in quality of manufacture, or lack of maintenance. Achieved reliability is sometimes called *operational reliability*, and its measurement in the field is called *demonstrated reliability*.

Reliability is generally quantified by reporting *mean time between failures* (MTBF). This is a statistical representation of the reliability of a product, and relates to the estimated service life of a product until it requires repair or replacement. Although *mean time to failure* (MTTF) should be used for nonrepairable products, MTBF is typically used for both repairable and nonrepairable products. Contrary to intuition, the MTBF number is not very useful for predicting how long a particular example of a product will last before it fails. If a product has an MTBF of 250,000 hours, it does not mean that one can expect an individual example of that product to last that long. The MTBF is meant to be applied over a long period of time to a statistically large sample. Then the probability that a particular example of a product would last a given amount of time can be calculated. If a product has a failure rate that remains constant throughout its life, the MTBF is the inverse of the failure rate. This is normally assumed, although often it is not actually the case because a newly installed product may be subject to infant mortality, whereas an older product may contain parts that are wearing out.

To determine the probability that a device will perform without failure over a specified period, a statistical formula is used, but first, a base of information is needed. There are two ways by which to arrive at a number representing MTBF: calculated and demonstrated. When calculating reliability, the device is broken down into all of its component parts. Each component part, such as a solder joint, weld, connection pin, resistor, or transistor, is assigned an individual failure rate. Then all the failure rate numbers are added up to arrive at a total failure rate. The inverse of the total failure rate is the MTBF.

One method to calculate the individual failure rates, and the total, is described in MIL Standard 317. Another is contained in the Bellcore standard. The MIL standard was developed, of course, with military applications in mind. The Bellcore standard was developed for the telecommunications industry and is a better predictor of the actual measured performance of products in commercial use. Both standards are in common use for industrial products.

Calculating an MTBF from estimated failure rates may be the only way to predict reliability before a new product has been in the field. Conversely, if a large number of the product has been installed and operating in the field for a sufficiently long time, an MTBF can be determined from the field reliability

data. These data come from the internal reports of the manufacturer regarding products returned to the factory, for reasons related to reliability, by the end users. Since this is based on actual performance results, it is called *demonstrated MTBF*.

To determine the probability that a given device will perform without failure over a given time period, T, the following formula is used:

$$R(T) = e^{-T/MTBF} \tag{2.6}$$

where $R(T)$ is the reliability over period T and e is the natural log (approximately 2.71828). For example, if a transducer has an MTBF of 250,000 hours (about 28 years), a particular example of the transducer has an 84% chance of operating for a period T of 44,000 hours (about five years). This may sound pretty good. As mentioned above, however, a total failure rate can be expressed as the sum of individual rates. So if three of those transducers are installed on a single machine, there is only a 28% chance that all three will last for five years before at least one of them will need repair.

Another important aspect of reliability is the extent to which the transducer is fit for use at the time when it is actually needed. This is called *availability* and is expressed as

$$\frac{\text{uptime}}{\text{uptime} + \text{downtime}} \tag{2.7}$$

or alternatively, it can be expressed as

$$\frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}} \tag{2.8}$$

where MTTR is the mean time to repair. If a product never failed, the availability would be 100%.

CHAPTER 3

RESISTIVE SENSING

3.1 RESISTIVE POSITION TRANSDUCERS

Resistive linear position transducers are very popular, relatively inexpensive, and are also the most easily understood type of linear position transducer. They are normally three-wire devices, and are also called *potentiometric position transducers, linear potentiometers, potentiometers*, or *pots*. The linear potentiometer is a position *sensor*, as it measures a physical variable directly. But since the sensing element also provides a usable electrical signal with no additional signal conditioning, it is also properly called a position *transducer*.

The basic concept is the same as that used in volume and tone controls on many radios and TV sets, in that a voltage is applied across a resistive element and a conductive wiper slides along the resistive element, making electrical contact with it. This allows a voltage potential to be read from the wiper. As the wiper voltage varies, it indicates the position of the wiper along the resistive element.

When using a resistive position transducer, the voltage applied across the resistive element is normally dc. The wiper voltage is approximately determined by the distance of the wiper from one end, divided by the total element length, multiplied by the total voltage across the element. In electronics engineering terms, this kind of function is called a *voltage divider* and is shown in Figure 3.1. Figure 3.1*a* symbolizes a two-resistor voltage divider. The output voltage at the connection between the upper and lower resistors is defined by the formula:

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

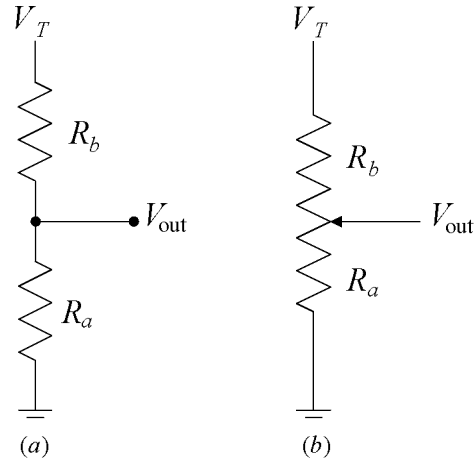

Figure 3.1 (*a*) Two-resistor voltage-divider circuit; (*b*) potentiometer circuit.

$$V_{out} = \frac{V_T R_a}{R_a + R_b} \tag{3.1}$$

Figure 3.1b symbolizes a potentiometer, where the arrow pointing to the middle of the resistor is called the *wiper*. The same formula for the output voltage applies when the resistance below the wiper connection is called R_a and that above the wiper is R_b.

3.2 RESISTANCE

Electrical resistance is the real part of the electrical impedance. This is the part responsible for the dissipation of energy [28, p. 8]. Electric current is often compared to the flow of water in a pipe. The volume of water flowing in a pipe per unit time, such as in liters per minute, is analogous to electrical current (1 ampere = 1 coulomb per second, where 1 coulomb is equal to the charge of 6.242×10^{18} electrons [11, p. 942]). The water pressure (e.g., in pascal) is like voltage, where a higher pressure can force a higher flow rate through the pipe if other parameters remain unchanged. The pipe diameter, roughness, and straightness control how easily water can flow. A smaller pipe diameter, greater degree of inner wall roughness, and curves or obstructions in the pipe create a resistance to water flow that is similar to electrical resistance. A pipe with a higher resistance to water flow requires a higher pressure difference (across the ends of the section of pipe in question) to achieve a given flow rate. Similarly, a higher voltage potential difference across a given resistance is required to obtain a higher current flow through the resistance. In electric circuits, the voltage is equal to the product of current and resistance:

$$E = IR \tag{3.2}$$

(This is a rearrangement of Ohm's law, normally written as $I = E/R$), where E is the voltage in volts (V), I the current in amperes (A), and R the resistance in ohms (Ω). For example, $10.0\,\text{V} = 0.10\,\text{A} \times 100\,\Omega$.

If several resistances are connected together in series, the total resistance, R_T, is equal to the sum of the individual resistances:

$$R_T = R_1 + R_2 + R_3 \tag{3.3}$$

When several resistances are connected in parallel, the total resistance is equal to the reciprocal of the sum of the reciprocals of the individual resistances:

$$R_T = \frac{1}{1/R_1 + 1/R_2 + 1/R_3} \tag{3.4}$$

3.3 HISTORY OF RESISTIVE LINEAR POSITION TRANSDUCERS

Ohm's law is named for Georg Simon Ohm, a German physicist who published this relationship in 1827. An electrical parameter called *resistance* was found to be the controlling factor in determining the current that would flow in a circuit with a given voltage. Later, as an electrical component, early resistors were formed of coils of wire, followed by versions using a carbon composition core having wire leads attached and coated with an insulating layer (*carbon comp resistors*). Later refinements in precision and manufacturing process included resistive elements in the form of carbon film, metal film, and Cermet types.

Wirewound variable resistors and resistive potentiometers were developed as electrical components, and rotary versions were used in early electrical and electronic circuits. They were also constructed with a variety of resistive element types, as described here in the resistive element section. It was a small step to start using a rotary potentiometer as a rotary position sensor. Linear measurements were made using a rotary potentiometer with a toothed gear (pinion gear) mounted on the shaft. The pinion gear would ride on a movable rack, and the rack position was therefore indicated by the potentiometer. When the potentiometer was refined for use as a linear position transducer, the linear resistive element was used with a contact wiper tracking along it. Further improvements were added to increase lifetime and performance (some of these are described in Section 3.4).

3.4 LINEAR POSITION TRANSDUCER DESIGN

Figure 3.2 shows the construction of a modern resistive linear position transducer. It comprises a resistive element, a parallel conductor, and a wiper

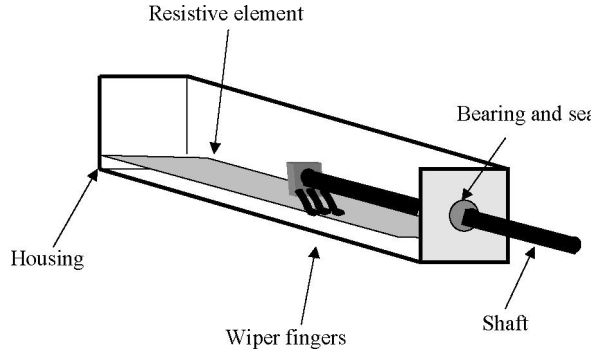

Figure 3.2 Rod-type resistive linear position sensor construction.

assembly as the major components. Also needed are a mechanical arrangement to keep the major components slidably positioned, and a housing, with bearings, seals, and wipes incorporated as needed to facilitate the performance and longevity of the major components.

In addition to the special considerations pertinent to designing each major component, as explained in the next few paragraphs, it is also important to execute the general design of the transducer in a way that can eliminate the sources of wear and inaccuracy. The main culprit in causing these problems is the admittance of dirt to the sliding parts and to the resistive element. Dirt is excluded in rod-actuated designs by adding one or more seals and wipes, which clean the rod as it is retracted into the bearing, and seal the rod-to-bearing sliding joint. A *wipe* is a tough fiber or elastomer annular element that rubs on the shaft to remove debris that may cling to the shaft as the shaft is retracted into the housing. A *seal* is a tight-fitting elastomer ring positioned around the shaft between the wipe and the bearing. The seal removes any smaller contaminants that may have gotten past the wipe and prevents the entry of liquids into the housing.

In the rodless type, it is a little more tricky to keep the dirt out. A flexible strip is positioned along a lengthwise side of the housing. That side of the housing would be open if the flexible strip was not in place. The housing is often made from an aluminum extrusion. The wiper assembly is attached to the underside of a car, which moves along tracks in the sides of the housing. As the car moves, the flexible strip is picked up ahead of the car and replaced against the housing behind the car (see Figure 3.3). Unfortunately, there always remains a small gap in the sealing area near the car through which it is possible for dirt to enter.

As the wiper slides along the track in the housing and makes electrical contact with the resistive element, it also contacts a conductor that runs parallel to the resistive element. The wiper provides electrical connection between the selected point on the resistive element and the parallel conductor. The conductor is brought out to form one of the three leads (or connector pins) to the

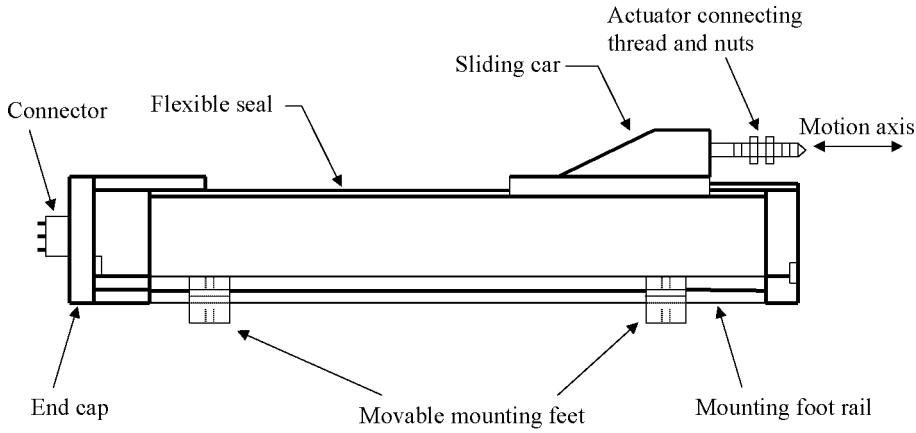

Figure 3.3 Rodless-type linear position transducer.

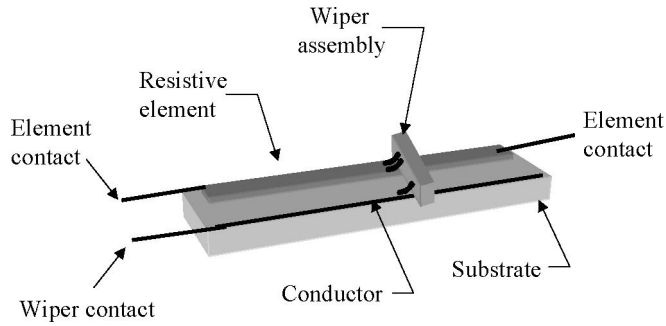

Figure 3.4 Connections to the wiper and resistive element.

position transducer. The conductor becomes the connection point for external circuits to connect to the various positions selected along the resistive element. The other two leads or connector pins (for a total of three) are connected to the ends of the resistive element (see Figure 3.4).

Besides resolution, nonlinearity, and hysteresis (see Chapter 2), there are a few additional electrical characteristics that must be controlled in a resistive position transducer. When the wiper is positioned all of the way at one end, there may be a part of the resistive element remaining before coming to the contact point for the connector pin. This is called the *end resistance*. There is an end resistance at each end of the resistive element. The end resistance limits the range of output voltage from the wiper so that the lowest output voltage is not quite as low as zero volts and the full-scale output is not quite as high as the supply voltage. When designing a resistive linear position transducer, the end resistance should be minimized, but if the wiper is allowed to go too

far, the opposite problem occurs: excess end travel. If there is a conductive area adjacent to the resistive element past where the lead connects, the wiper can move past the end of the resistive element or the lead connection point. Then the mechanical stroke is longer than the electrical stroke. The difference is called the *overtravel*. It is desirable in some applications, where the mechanical zero can be adjusted. Since there is no change in output in the overtravel areas, it could be a source of error when no mechanical adjustment is available.

Contact resistance is the resistance of the wiper contact to the resistive element and to the parallel conductor. If the output loading produces a measurable current in the wiper circuit, the voltage drop across the contact resistance will cause an offset in the output voltage. This offset can be adjusted out in some systems, but it is desirable to design the load circuit so that the offset is negligible for the application. Minimal loading is also important to retain the linear performance of the transducer. The load resistance should therefore be in the range of more than 50 times the resistance of the transducer resistive element.

3.5 RESISTIVE ELEMENT

In early resistive position transducers, the resistive element was either wirewound or constructed of a substrate onto which a carbon surface was formed. The wirewound element has a limitation of coarse resolution, because the minimum increment of resistance selection that is possible is one turn of the resistance wire. Some wirewound transducers suffer as well from a difficulty of maintaining good wiper contact with the resistance wire while the wiper remains stationary for long periods. Occasional movement of the wiper is sometimes required to clean the resistance wire surface and maintain a good contact with the wiper. Types of resistive element construction include wirewound, carbon film, conductive plastic, metal film, Cermet, and hybrid.

Wirewound elements are manufactured by winding a wire around a nonconductive mandrel. The wire has an electrically insulating coating, which is removed along one edge after winding. In the completed transducer, the wiper rides against the area of the winding where the insulation is removed. The mandrel can be made of ceramic, plastic, or metal that has an insulating coating. Ceramic is the most stable mandrel material, plastic is the cheapest, and metal conducts heat and is useful for higher power dissipation. The temperature coefficient of resistance of a wirewound element is relatively low, about 50 to 100 ppm/°C.

The disadvantages of a wirewound resistive element are the coarse resolution, noise due to breaking and making contact between adjacent windings, and the tendency to lose contact between the wiper and the resistive element when the wiper is not moved for long periods of time (days). The resolution is coarse because the minimum step from one setting of the wiper to the next is equal to the resistance of one turn of wire. If a wirewound element has 500

turns, the finest resolution possible is 0.2% (1/500). Noise is higher than with other element types because the wirewound element is inductive and the make-and-break action when the wiper moves from turn to turn causes voltage pulses on the output due to the rapid change in wiper current, d_i/d_t. When the wiper is not moved for long periods of time, an oxide layer forms on the surface of the wire resistance element. This increases contact resistance and can cause an open circuit between the wiper and the resistance element. Moving the wiper will clear the oxide and return the pot to the normal contact resistance.

Carbon film resistive elements are manufactured by mixing a paste of carbon and clay and screening it onto a nonconductive (usually ceramic) substrate material. The assembly is then fired to harden the carbon–clay mixture into a ceramic composite. This has the advantage of yielding a finer resolution than with a wirewound element because the wiper can follow a continuous path (not passing from one wire turn to the next, as with a wirewound element). Noise is also less than with a wirewound pot, because the surface is relatively smooth and the element is noninductive. One disadvantage is that they have a temperature coefficient of resistance (about 200 ppm/°C) that is larger than wirewound, metal film, or Cermet, but this is not so important when used in a voltage-divider circuit unless there is asymmetrical end resistance or external added resistance for trimming. Although electrical noise is much lower than with a wirewound element, there is still a substantial amount of noise generated as the wiper(s) scratches along the somewhat rough surface. The noise is a rapid change in resistance produced by large differences in contact resistance.

Conductive plastic elements are very smooth, allowing for very low noise from the wiper contact. They are made by adhesive mounting a conductive plastic film onto a rigid nonconductive substrate. The conductive plastic film is a composite mixture of carbon or another conductive powder with plastic. A wide range of resistance values is possible, but conductive plastic elements have the largest thermal coefficient of resistance (about 300 to 500 ppm/°C).

Metal film resistive elements are produced by depositing a film of metal alloy onto the surface of a ceramic substrate. This makes a very durable assembly. The resistance can be adjusted by cutting away some of the metal film. This also makes possible the correction of the change in resistance per millimeter, for a more linear performance of the position transducer. The performance is similar to that of a carbon film pot but is more rugged and has a lower temperature coefficient of resistance (about 50 to 100 ppm/°C).

Cermet is a mixture of conductive metal particles with clay. Varying amounts of silver, chromium, and lead oxide, for example, are mixed with the clay to provide the desired range of resistance. The mixture is screened onto a ceramic substrate and fired. This is nearly as rugged as metal film, somewhat adjustable, and has a relatively low temperature coefficient of resistance (about 50 to 100 ppm/°C).

Since a Cermet element is very rugged, with a low-temperature coefficient, and a plastic film has the smoothest surface, it is possible to adhere a conduc-

tive plastic film to the surface of a Cermet element to get the best properties of each. This is called a *hybrid* or *multilayer element*. One disadvantage of this construction is that the conductive plastic surface may not last as long as a cermet surface.

The resistivity of a material is the electrical resistance of the material per unit area or volume. Area resistance is used for measuring thin films and is called *surface resistivity*. Since it is common to measure the resistance of thin films, or surface resistivity, a simplified unit is used. It is indicated in ohms per square (Ω/sq). It does not matter if the linear unit is inches or millimeters, since both the length and width are affected (thus making the linear unit scale irrelevant). For example, when evaluating the resistance of a rectangular film of width w and length l, the number of squares is equal to l/w. The total resistance from end to end is then

$$R = \rho \frac{l}{w} \tag{3.5}$$

where R is the total resistance and ρ is the resistivity in Ω/sq.

Volume resistivity is used for three-dimensional materials such as wire, and is called *bulk resistivity*. Since it is a common form for which resistance is measured, there is also a simplified unit of resistivity for wire. Having a uniform cross-sectional area along its length, the resistivity of a given wire material and diameter can be indicated in units of ohm-meters (Ω·m), or ohm-centimeters. The resistance of a wire having length l is then

$$R = \rho \frac{l}{A} \tag{3.6}$$

where ρ is resistivity in Ω·m, for example, and A is the cross-sectional area.

3.6 WIPER

The wiper of a linear position transducer has two areas of electrical contact: one set of wipers contacts the resistive element, the other set contacts the parallel conductor. Multiple fingers are formed in each set of wipers to improve the contact reliability. If one wiper finger is not making contact with the surface of the resistive element at a particular instant, for example, at least one of the other fingers is likely to be making contact at that time. This way, nearly continuous contact can be ensured. The degree of difficulty in maintaining contact is partly due to the nonuniformity of the surface of the resistive element, the element having high and low spots, as well as areas of better and poorer surface conductivity. The other important part of maintaining contact is due to the ability of the wiper surface to maintain a clean and conductive surface in the contact area. This is accomplished by selecting the best

alloy for plating onto the contact area, which is electrochemically compatible with the resistive element material. The electronegativity of the wiper plating alloy must be very close to the electronegativity of the resistive element material. There is also a trade-off in the wiper design between higher contact pressure providing better electrical contact, versus lower contact pressure allowing for longer cycle life (due to less wear on the surface of the resistive element).

The base material of the wiper is a spring material of which the spring rate and physical dimensions are chosen to find a balance between contact pressure and other performance factors. Higher contact pressure improves (reduces) contact resistance but worsens (increases) hysteresis by causing flexing of the wiper assembly. Higher contact pressure also reduces the lifetime of the resistive element by rubbing off some of the resistive coating and forming particles that interfere with electrical contact. The chemical composition of the plating on the wiper must be chosen to maintain low surface resistivity in combination with the chosen composition of resistive element surface, especially when the wiper is not moved over long periods of time. Suitable plating materials include gold, palladium, or silver, and can be laid down as multilayers, or mixtures.

3.7 LINEAR MECHANICS

As mentioned earlier, it is typical for the potentiometer housing to be made of an aluminum alloy extrusion. This has several advantages in cost, producibility, and design flexibility. The tooling cost for an aluminum extrusion is very low compared with the tooling cost for a metal die casting or a plastic injection molding. A metal housing also has an EMC advantage over a plastic housing because an electrically conductive housing provides electromagnetic shielding. An extrusion can easily be designed to include a slot into which the resistive element can be mounted. Channels can also be included which provide holes at the ends of the extrusion suitable to accept screws to attach end plates for the housing. Metal end plates complete the shielding scheme.

In a rod-type position transducer, a bushing is needed at the rod entry end, and a guide is needed on the inside tip of the rod to keep the rod parallel with the housing. The tip guide rides against the inside wall of the housing to maintain alignment and can also accommodate the wiper assembly. To stop the undesired entry of particulate debris, and to protect the bushing from wear due to hard particles, one or more elastomeric or plastic wipes and seals are mounted outside the bushing.

3.8 SIGNAL CONDITIONING

A resistive sensor is normally used without signal conditioning circuitry. If one lead of the resistive element is connected to zero volts, and the other to a reg-

ulated supply voltage, the wiper output is a voltage between those two points and in proportion to the position. There is a need for a proper input circuit in the device that reads the pot output. The input circuit is a load on the pot output. If the load resistance is too low, it causes a lower signal level (if the pot is not going completely to full scale) and increases the nonlinearity of the position indication from the pot. Loss of signal level can be corrected with a gain adjustment, but the nonlinearity is more difficult to accommodate. To eliminate this problem, the load resistance has to be much larger than that of the resistive element. For example, a common value of resistance for a pot is 10 kΩ. In this case, the input impedance of the receiving circuit should be 1 MΩ or more. This is relatively easy to obtain by using a noninverting operational amplifier, or an instrumentation amplifier. The problem with a high-impedance input, however, is that it becomes easier to pick up electrical noise. Electrical noise is generated from motors, switches, and solenoids (among other things) in industrial equipment. Electronic filter circuits are typically added to the input signal conditioning circuit to limit the effect of the noise. Using a shielded cable or a twisted pair of wires to carry the signal from the pot into the input of an instrumentation amplifier is another way to limit admitting electrical noise into the circuit.

It is also possible to change the output of a potentiometer to a lower impedance, thus reducing the likelihood of picking up noise. This can be done by adding an op-amp circuit, within the housing of the pot, as a buffer. The conductors bringing the wiper signal to the amplifier are very short and are shielded by the potentiometer housing. So there is very little noise induced at this point. Then the output impedance of the op-amp circuit (on the order of less than 1 Ω) becomes the transducer output impedance. The op-amp circuit can be a voltage follower configuration, with the output tied to the inverting input and the power supply rails connected to the potentiometer resistive element. A circuit used by the author to condition the output of a potentiometer is shown in Figure 3.5.

The output voltage of the follower circuit is essentially the same as that at the noninverting input, V_{i+}. If the amplifier open-loop gain is 100,000, for example, the output voltage, V_o, is approximately [38, p. 364]

$$V_o = \frac{V_{i+}}{1 + 1/100{,}000} \tag{3.7}$$

since $V_{i+} = V_{i-}$.

One problem here would be that the pot then needs to be identified as to which lead of the resistive element is wired to positive, and which to negative or common. A possible solution to this problem is to power the op amp using a diode bridge to steer the power supply voltage. The slight disadvantage of the diode bridge is that the op-amp output cannot swing to the rails, being limited to about 0.6 V shy of reaching the rail voltages.

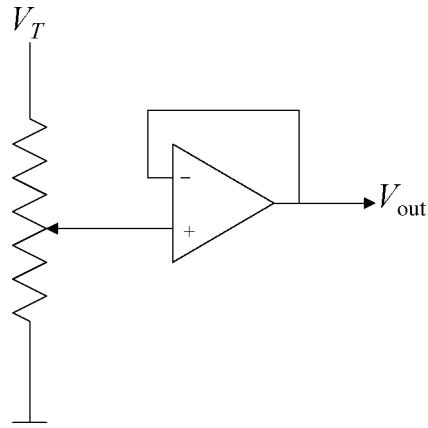

Figure 3.5 Linear potentiometer circuit with low output impedance.

3.9 ADVANTAGES AND DISADVANTAGES

Resistive linear position transducers are a popular choice because of their relatively low cost, simple wiring, and because they are easy to understand. Wirewound pots are still common, but transducers with Cermet elements are better suited for industrial use where finer resolution and high cycle lifetime are needed, especially in closed-loop control systems. It should be noted, however, that there is always the problem of long-term wear when using a contact type of transducer. So for high-reliability applications where periodic replacement of the transducer is not desired or is difficult, a noncontact transducer should be installed.

3.10 PERFORMANCE SPECIFICATIONS

Nonlinearity

The nonlinearity of a potentiometric linear position transducer is controlled by the uniformity to which the resistive element is manufactured. In a wirewound element, the wire must have uniform resistivity, but just as important is the accuracy in winding the wire onto the mandrel. The wire must be carefully layer wound, with each turn adjacent to the preceding turn, and with the same pitch. The turns must be tight to prevent movement when the wiper moves over them. In carbon-film or Cermet types, the resistive paste must be homogeneous and applied to the substrate in a uniform geometry. Uniformity of the conductive mixtures is determined primarily by the uniformity of the conductive particles and their dispersion throughout the paste. Uniformity of application is determined by the thickness as well as the width and straightness to which the paste is applied to the substrate.

It is a little more difficult to produce a highly linear plastic film element, and these are therefore used in lower-cost, lower-performance applications. Cermet elements, however, can be adjusted after firing. This is done by first measuring the linear response carefully and calculating the changes needed to reduce the nonlinearity. Then scratches are cut along the edge of the element, an area that is not used by the wiper, to increase the resistance slightly in those areas, as needed. A manual operation or a mechanized one can be used to apply the scratches mechanically. More accurate results can be obtained by making the scratches by laser etching in an automated process. To provide the lowest nonlinearity error, as required in higher-accuracy products, the nonlinearity can be adjusted by laser etching to reduce the nonlinearity error to the level of about ±0.05% of full range.

Overall resistance is not as important as uniformity, since the potentiometer is typically used in a voltage-divider circuit. The resistive load, however, can cause additional nonlinearity if it is too low, as mentioned earlier. An example is considered where a resistive element is connected across a power supply, with terminals at zero volts and V+, and a load resistor is connected from the wiper to zero volts. The load resistor has essentially no effect on the wiper voltage output when the wiper is at either extreme (ignoring contact resistance). However, when the wiper is at other positions, the load resistor is in parallel with the lower part of the element and changes the voltage-divider ratio. This variation in the effect from the load resistor is the source of the nonlinearity that can be induced by a load resistor. The maximum effect in the case cited is with a wiper position at 67% of full range. As also stated earlier, errors are reduced by making the resistance of the load resistor much higher than that of the resistive element.

Hysteresis

The hysteresis of a potentiometer comes from the mechanics of the wiper to element contact and from the wiper tracking mechanics. Normally, there are three or more fingers on a given wiper assembly (these fingers will collectively be called the wiper). The wiper rubs against the resistive element with a carefully controlled (by design) amount of contact force. The force is applied by spring tension of the flexed wiper. The wiper is made from a spring type of metal and is then plated with an alloy to enhance conductivity to the resistive element the surface, reduce oxidation, and reduce wear. The alloys typically used include palladium and gold.

Given the contact force, contact area, and surface friction, there is a resulting amount of force required to move the wiper across the element. This force acts against the flexural strength of the wiper to cause a movement of the wiper and its mount. As the measurand moves upscale, the wiper is flexed slightly in the downscale direction. Similarly, as the measurand changes in the downscale direction, the wiper flexes toward the upscale direction. This was shown in Figure 2.8. The difference between upscale and downscale flexing is the main

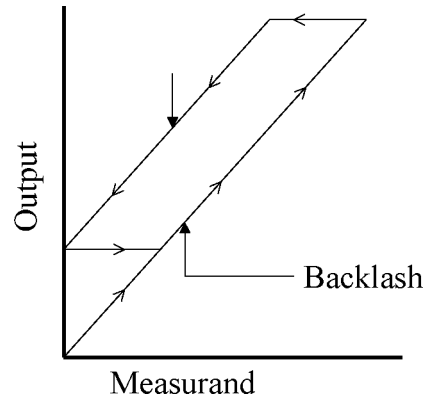

Figure 3.6 Shape of the transducer characteristic, including backlash (exaggerated for clarity), normally considered to be a part of hysteresis.

source of hysteresis in a linear potentiometer. The hysteresis just described is also called *backlash*, and has a different shape from the transducer hysteresis curve that was shown in Figure 2.6. The curve shape for backlash is shown in Figure 3.6.

Wear/Lifetime

Since the linear potentiometer is a contact device, a major consideration in its use is how long it will last before wearing out. When the sensor is new, the surface element is clean and smooth, the wiper is evenly coated with the alloy selected, and the tracking mechanics are tight. As the potentiometer wears over many cycles of use, all of these initial conditions change. The main problems to be encountered due to wear when a potentiometric linear position transducer has undergone a number of cycles nearing its rated lifetime are noise and dead spots. As debris accumulates on the surface of the resistive element, the wipers make and break contact with the surface as they ride over the debris particles. This increases the resistance variation and uncertainty of the output for a given position (which is called *noise*). If all the wipers that are on either the element or the parallel conductor rest on debris particles, an open circuit can result. When this problem occurs repeatedly at one location, or an open circuit results from the resistive coating being worn off at one location, it is called a *dead spot*.

Dead Zones

There may be electrical dead zones on each end of travel where a further change in the input measurand does not cause a further change in the output signal. This is possible when the design allows the wiper to ride up onto the

conductor lands, which are used for connecting the electrical element to the potentiometer end pins. Mechanical dead zones may also exist in which further travel of the wiper is mechanically prevented before the full expected range of the input measurand is reached. In this case, it is not mechanically possible to move the actuator rod to the extreme end of the stroke. This is usually due to misalignment of the transducer when mounting it.

3.11 TYPICAL PERFORMANCE SPECIFICATIONS AND APPLICATIONS

A typical set of specifications for a potentiometric linear position transducer (in this case, the Honeywell Longfellow II Series) is as follows:

Mechanical	
Total mechanical travel:	150 to 1200 mm
Starting forces:	0.45 kg
Total weight:	0.36 to 2 kg
Vibration:	20 g rms/0.75 mm 5–20 Hz
Shock:	50 g, 11 ms half sine wave
Backlash:	0.025 mm
Life:	1 billion dither operations
Electrical	
Theoretical electrical travel:	150 to 1200 mm
Independent nonlinearity:	0.1%
Total resistance:	5000 Ω
Resistance tolerance:	20%
Operating temperature:	–65 to 105 °C
Resolution:	Infinite
Insulation resistance:	1000 MΩ at 500 Vdc
Dielectric strength:	1000 V rms
Recommended wiper current:	<1 μA
Electrical connection:	Binder series 681
Maximum applied voltage:	30 Vdc

When matching a type of potentiometer to a real-world application, it is necessary to specify the required full-scale range, nonlinearity, hysteresis, and total resistance, but consideration must also be allowed for the type of resistive element, expected number of lifetime cycles, and mounting ability.

A wirewound element can provide good long-term reliability but should not be used in servo control loops or where fine resolution is needed. In a servo control loop, there can be a constant dithering between two adjacent turns on a wire wound element when the actual control point being sought lies between the two points of contact available. In this case, a higher-resolution pot with continuous output should be used.

Some advantages of the potentiometric linear position transducer over other types include the ease of understanding the transducer performance, ease of application to a specific use, the relatively low cost, no electronic circuit is needed to operate the transducer, and they have a high-level output signal voltage directly from the transducer. Some disadvantages include the limited frequency response, high operating force due to friction, high dynamic force due to the mass of the moving parts, and the limited lifetime due to wear.

CHAPTER 4

CAPACITIVE SENSING

4.1 CAPACITIVE POSITION TRANSDUCERS

Position transducers based on capacitive sensing are very popular in the industrial world because they provide a relatively simple technique to implement a noncontact measurement. The transducer may have an actuating member which contacts the item to be measured, but the noncontact sensing element provides the opportunity for designing a transducer that will impose only minimal mechanical loading forces and sustain minimal wear. A noncontact capacitive position transducer measures the location of a conductive target (see Figure 4.1).

Capacitive transducers require a set of driving and conditioning electronics, and therefore are inherently more complex than a contact resistive transducer. A typical capacitive linear position transducer comprises a variable capacitance sensing element, electronics, and suitable mechanical components to house them. The housing provides the means to maintain alignment of the movable elements while receiving the mechanical input of the measurand.

Contrary to the capability of a linear potentiometer, a capacitive sensing element does not produce a directly usable output. The basic concept is that the electronic circuit drives the sensing element with an alternating current, the sensing element changes capacitance due to changes in the measurand, and the resulting signal is demodulated by the electronic circuit. In addition, the circuit conditions the power supply and converts the demodulated signal

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

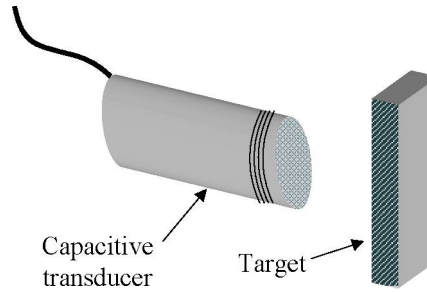

Figure 4.1 A capacitive position transducer installation includes a sensing element and a target.

into the desired output. A housing is usually designed to contain all the components, including the sensing element, electronics, cable strain relief or connector, actuator assembly, and mounting features.

4.2 CAPACITANCE

To understand and design capacitive sensors, it is helpful to be familiar with the nature of the electrical property of capacitance. The capacitance of a system is a measure of its capability to store electrostatic energy. An analogy can be drawn between an electrical circuit with capacitance and a water system with a storage tank. In this analogy, a tank with a large diameter (capacitance value) can hold a lot of water for a given height (voltage). Voltage acts like water pressure, and water pressure is often measured by its height (e.g., meters of water head, or pressure). A water tank extending up to a height of 20 m above the ground will develop a pressure of a little over 2 atm gauge pressure at ground level when the tank is full. If a pump (the electrical equivalent would be a battery or other power source) fills the tank at a certain flow rate in liters per second (electrical equivalent: charging a capacitor with a current in coulombs per second, or amperes), the water level (voltage) will rise. If a smaller-diameter pipe (larger-value resistor) is used to fill or drain the tank, the tank will fill (charge up) or drain (discharge) more slowly. The surface area of the tank in square meters, for example, is analogous to the capacitance value (in farads) of a capacitor. So a larger-diameter tank (greater capacitance) will fill or drain (change voltage) more slowly with a given flow rate (current). The product of the number of liters of water stored in the tank and the square of the height of the tank is proportional to the amount of energy stored [see equation (4.3)].

A simple capacitor (see Figure 4.2) is formed by two parallel conductive plates in close proximity, separated by a dielectric material (an electrical insulator) by distance d. The dielectric material may be vacuum or air, plastic, fiberglass (as in a printed circuit board), or almost any nonconductor of elec-

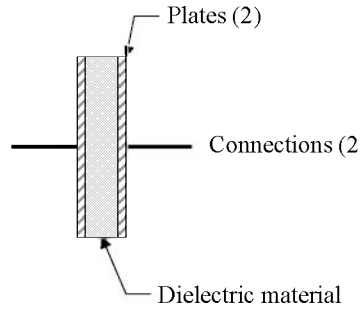

Figure 4.2 Construction of a simple parallel-plate capacitor.

tricity that is suitable for forming into the desired physical shape. The capacitance, C, would be given by

$$C = \varepsilon_0 \varepsilon_r \frac{A}{d} \tag{4.1}$$

where ε_0 is the permittivity (permittivity is explained further in Section 4.3) of free space in farads per meter (F/m), ε_r is the relative permittivity of the dielectric material (a ratio, without units, also called the *dielectric constant*), A is the effective area of the plates in square meters, and d is the distance between the plates in meters.

Once the capacitance is known, the amount of charge stored in the capacitor can be determined when it is charged to a particular voltage. The charge, Q, is given by

$$Q = CV \tag{4.2}$$

where Q is expressed in coulombs (the coulomb is the unit of charge), C the capacitance in farads, and V the voltage to which the capacitor is charged. A coulomb is the equivalent charge of 6.24×10^{18} electrons. One ampere of electrical current flow is the transfer of 1 coulomb of charge per second. The electrostatic energy, W_E, that is stored in a capacitor is equal to

$$W_E = \frac{1}{2} CV^2 \tag{4.3}$$

where W_E is energy in joules (or watt-seconds), C is in farads, and V is the applied voltage in volts. The stored energy has the letter W because energy is also known as the *ability to do work*. This is analogous to the potential energy stored in a compressed spring or that stored in the water in the elevated tank described earlier. When a given current is applied to charge a capacitor, the rate of change of voltage, dV/dt, varies in direct proportion to the capacitance:

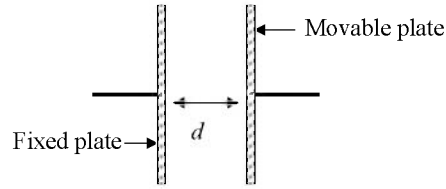

Figure 4.3 Rudimentary capacitive sensor. Capacitance changes with plate separation distance, *d*.

$$I = C\frac{dV}{dt} \tag{4.4}$$

As an example of a rudimentary capacitive position sensor, see Figure 4.3. According to equation (4.1), the capacitance will vary as the distance, *d*, between the parallel plates is varied. With dimensions in centimeters, air as the dielectric, *C* in picofarads, and neglecting fringe effects, equation (4.1) becomes [2, p. 38]

$$C = 0.08854\frac{A}{d} \tag{4.5}$$

When several capacitors are connected in parallel, the total capacitance, C_T, is the sum of the individual capacitances:

$$C_T = C_1 + C_2 + C_3 \tag{4.6}$$

When several capacitors are connected in series, the total capacitance is equal to the reciprocal of the sum of the reciprocals of the individual capacitances:

$$C_T = \frac{1}{1/C_1 + 1/C_2 + 1/C_3} \tag{4.7}$$

The reader may have noticed that the formula for adding parallel-connected capacitors is similar to that for adding series-connected resistors. Similarly, the formula for adding series-connected capacitors is similar to that for adding parallel-connected resistors.

4.3 DIELECTRIC CONSTANT

The permittivity of vacuum, ε_0, is 8.85×10^{-12} F/m. The relative permittivity of a material is equal to the ratio of the permittivity of the material to that of vacuum [19, p. 13/3]. Accordingly, relative permittivity is a ratio and has no units. The relative permittivity, ε_r, of vacuum is defined as 1. The ε_r value of dry

air is 1.0006, so it is almost the same as vacuum. The relative permittivity of a material is also called its *dielectric constant*, and this term is often used regarding the permittivity of dielectric materials of capacitors and capacitive sensors. The dielectric constant of a material is measured by using a standard capacitor: for example, a set of two parallel square plates, 1 cm on a side. The plates are made to sandwich a layer of the dielectric material to be measured and the capacitance recorded as the first reading. Then another measurement is made with the dielectric material replaced by air or vacuum (air and vacuum give almost identical results), with the same spacing between the capacitor plates as when the dielectric material in question was between them. This is the second reading. The first reading divided by the second reading is the *dielectric constant* (or *relative permittivity*).

The higher the dielectric constant or permittivity, the slower an electric field will travel, and the larger will be the capacitance of a given size and spacing of parallel plates. The velocity of electric field propagation is not a linear relationship with permittivity. The velocity of electric field propagation, v, in a material is inversely proportional to the square root of the permittivity and the permeability of the material:

$$v = \frac{1}{\sqrt{\varepsilon_a \mu_a}} \tag{4.8}$$

where ε_a is absolute permittivity, being the product of the permittivity of vacuum, ε_0, and the relative permittivity, ε_r, of the material:

$$\varepsilon_a = \varepsilon_0 \varepsilon_r \tag{4.9}$$

and where μ_a is absolute permeability, being the product of the permeability of vacuum, μ_0, and the relative permeability, μ_r, of the material:

$$\mu_a = \mu_0 \mu_r \tag{4.10}$$

4.4 HISTORY OF CAPACITIVE SENSORS

Capacitive linear position sensors are also called *variable-capacitance sensors*. They must be combined with electronic circuits in order to operate, and are therefore not as old in the art as are the potentiometric position sensors. Early capacitive linear position sensors were built for a particular application rather than being made as standard products for general use. Shorter stroke sensors used variable spacing of the sensing plates; with air, another gas, or vacuum as the dielectric. In the 1970s, the author used a pressure transducer that incorporated an early capacitive position sensor. The capacitive sensor was coupled to a metal diaphragm, which flexed in response to changes in pressure. The

capacitive sensor measured the diaphragm deflection and produced an analog voltage output representing the applied pressure. This technology was sold as a standard line of pressure transducers.

The relatively large effort required to design electronic circuits to drive and condition the signal from capacitive sensors was reduced substantially when integrated circuits (ICs) for this use became widely available in the 1990s. These ICs enabled a reduction in the size and complexity of the circuitry, and included both the driving and signal conditioning circuitry necessary for capacitive sensors. The lower cost also made the technology suitable for automotive use.

4.5 CAPACITIVE POSITION TRANSDUCER DESIGN

The capacitive sensor plates are typically constructed of one or more dielectric substrates onto which metallic layers are formed. In a two-plate sensor, the fixed plate can be a metal layer that is formed on a dielectric substrate by using standard electronic printed-circuit techniques. The second plate would be the movable plate, which is moved in response to the measurand. Other geometries are also possible, as explained later in this chapter. In addition to this basic technique, specialized circuitry, materials, and sensor configurations can be used to enhance resolution, signal/noise ratio, sensitivity, stability, and temperature performance.

According to equation (4.1), the capacitance will change with a change in distance d between the plates, area A of the aligned area of the plates, or relative permittivity ε_r of the dielectric material. So, in practical position transducers, the dielectric material, the plate area, and/or distance between the plates can be made to vary with the measurand. Figure 4.4 shows several representative diagrams for these configurations, providing a changing capacitance in response to a change in linear position.

There are trade-offs among cost, performance, and range of stroke length, which act to steer the designer toward choosing the best sensor configuration for a particular transducer design. Variable spacing, where the target being measured is the movable plate, is generally the least expensive approach. Longer-stroke sensors generally require variable area or variable dielectric designs. After choosing the sensing element configuration, a housing must be designed that will provide alignment among the several parts of the sensing element.

The type of capacitive position transducer that is most popular in industry works with a target supplied by the user. The transducer mounts with its sensing element near and parallel to the target. The target can be a metal plate, which moves in an axis perpendicular to the sensing element. A transducer of this configuration is shown in Figure 4.5. The parallel-plate arrangement of Figure 4.4*b* is a variable-area capacitive sensor. In this case, the capacitance measured between the plates will vary approximately with the percentage of

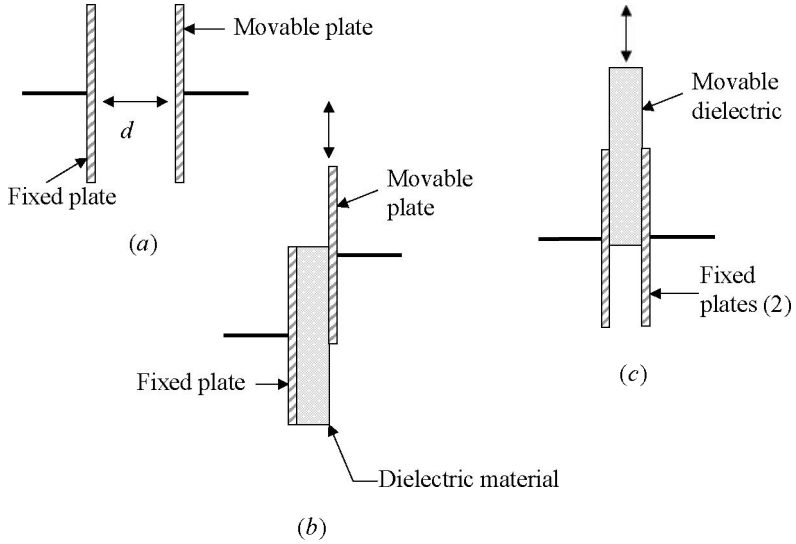

Figure 4.4 Plate configurations for variable capacitance in response to a variation in linear position: (*a*) variable spacing; (*b*) variable area; (*c*) variable dielectric constant.

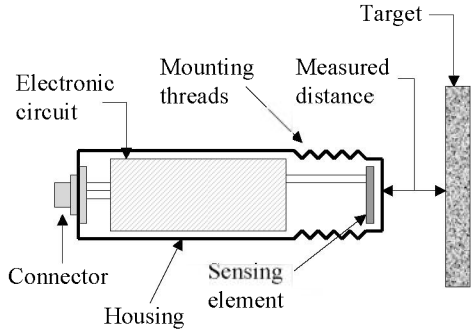

Figure 4.5 A capacitive linear position transducer, with user-supplied target, includes a sensing element, electronics, and a housing with mounting and connection means.

the area of the movable plate that is aligned directly over the fixed plate. This assumes that the distance between the plates is small compared to the dimensions along the sides of the plates. In Figure 4.5, the linear distance to be measured is in the left and right direction. This is called the *x-axis*, or *sensitive axis*. One error source, however, is that the capacitance will also vary with motion at 90° to the *x*-axis. We'll call this the *y-axis*, or *cross axis* (in the figure: into the paper).

An improvement to this basic arrangement, called *overlap*, is shown in Figure 4.6 as applied to a variable area sensor. Here the upper plate is not as

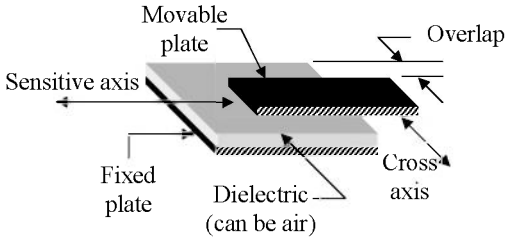

Figure 4.6 Improved plate design, with overlap, to reduce sensitivity to motion in the cross-axis.

wide as the lower plate. Thus, for a given position, there is little variation in capacitance with motion in the (y) cross-axis direction as long as the upper plate remains positioned over some part of the lower plate. A slight disadvantage is that the total capacitance is somewhat smaller due to the smaller width of the movable plate. There is a performance trade-off of accepting a lower signal level with the accompanying lower signal/noise (S/N) ratio to reduce the cross-axis sensitivity. This is an example of how the design of a sensor element can be improved to fulfill an application requirement during the initial testing of a proposed sensor configuration. For this reason, and because each element of a transducer interacts with each other element, transducer design often includes several stages of iteration.

According to equation (4.1), the capacitance will vary directly with the aligned area of the plates. Thus, in the arrangement of Figure 4.6, the capacitance will vary directly with the linear position of the upper electrode as it moves in the sensitive axis. Another consideration is that since the capacitance also varies inversely with the distance between the plates, d, the thickness of the dielectric material must remain constant over time, temperature changes, and motion in the sensitive axis. This can be accomplished in the case of an air dielectric by using a mechanical means to make sure that the two plates track in parallel, such as by fixing the bottom plate and having the upper plate ride in a track. An alternative way would be to use a plastic or ceramic dielectric plate sandwiched between the electrode plates instead of air. In addition to less cross-axis sensitivity, another advantage of this construction would be an increase in total capacitance. The factor by which capacitance would increase, according to equation (4.1), is approximately equal to the relative permittivity of the dielectric plate material.

An additional source of error in capacitive position sensors is the possible unwanted change in capacitance due to changes in the relative positions of nearby conductors. When a conductive body moves within the vicinity of the sensor, it can provide additional capacitive coupling between the two plates of the sensor. Similarly, nearby electric fields can also be detected and thereby affect the capacitance reading. A shield can be added to surround the sensing area to prevent this (see Figure 4.7).

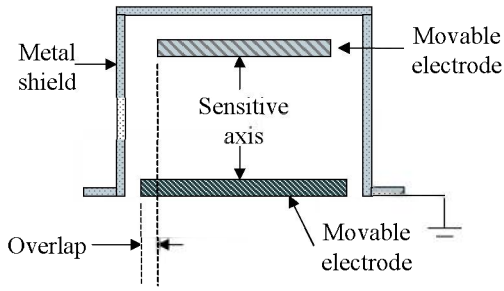

Figure 4.7 Adding a shield around a variable-spacing sensor to prevent unwanted error signals from nearby conductors or electric fields.

Overlap was explained relative to variable-area sensors with lateral motion to provide a variable area between the plates, but the variable-spacing sensors of Figures 4.5 and 4.7 also benefit from overlap. According to equation (4.1), the capacitance will vary inversely as the distance, d, between the parallel conductor plates. Like the variable-area design, though, there can be unwanted sensitivity in the cross-axes in a variable-spacing sensor. Sensitivity to cross-axis motion in the y or z axes (if the other cross-axis is called z, in this case) is reduced by making one plate of smaller dimensions than the other plate. Again, the capacitance level is less, but this disadvantage is sometimes not as important as the advantage of reducing sensitivity in the cross-axes.

4.6 ELECTRONIC CIRCUITS FOR CAPACITIVE TRANSDUCERS

The capacitive sensing element of a linear position transducer is driven with an ac or switching circuit. This allows the measurement of a variable frequency, phase shift, or amplitude change due to the change in capacitance in response to changes in the measurand. Typically, either an oscillator or a *time-period generator circuit* would be used. In a time-period generating circuit, also called a *monostable multivibrator*, or *one-shot*, the length of the timed period depends on the magnitude of the capacitance (see Figure 4.8). The trigger for the one-shot could come from a microcontroller circuit that sends the trigger pulse when a reading is desired, or the one-shot could be driven by a free-running pulse generator for continuous operation. A suitable free-running pulse generating circuit is shown in Figure 4.9.

An alternative is to drive the capacitive sensing circuit with a free-running oscillator in which the frequency is determined by the sensor capacitance, C_S, and therefore by the measurand (see Figure 4.10). The oscillator signal could then be fed to a circuit that produces the desired output signal from the frequency signal. This could be a frequency-to-voltage converter circuit, if a voltage output is desired, a block diagram of which is shown in Figure 4.11.

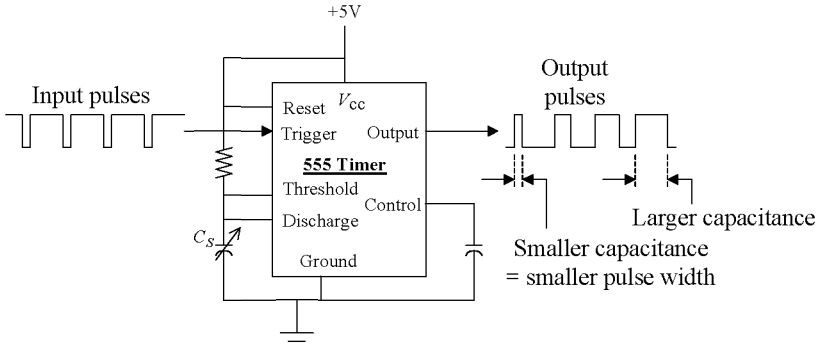

Figure 4.8 One-shot circuit in which the output pulse width depends on the variable value of sensor capacitance, C_S.

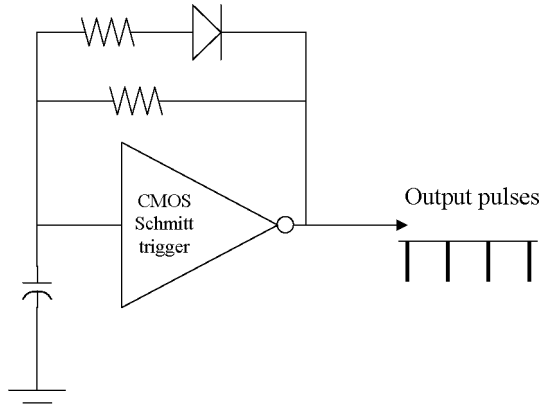

Figure 4.9 Free-running pulse generator to drive the one-shot circuit of Figure 4.8.

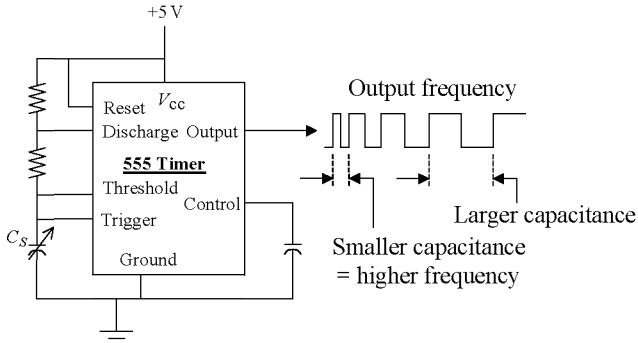

Figure 4.10 Free-running oscillator in which the sensing capacitance, C_S, determines the operating frequency.

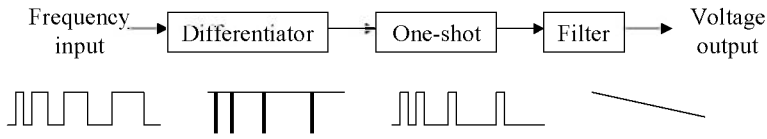

Figure 4.11 Frequency-to-voltage converter circuit block diagram.

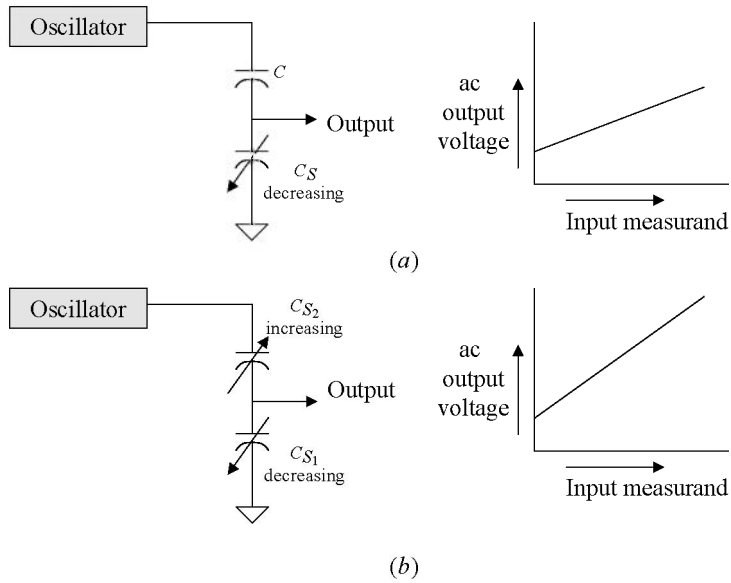

Figure 4.12 (*a*) Single-element circuit with variable voltage amplitude. The output characteristic versus measurand is shown to the right. (*b*) Dual-element circuit where one capacitance increases while the other decreases.

Another way to obtain a variable dc voltage signal from an oscillator and variable capacitance is to develop an ac voltage that has an amplitude that varies with the measurand, which can then be converted to a dc voltage. This is shown in Figure 4.12. In Figure 4.12*a*, C and C_S comprise a voltage-divider circuit. The ac voltage across C_S increases as the capacitance decreases, because its capacitive reactance (impedance) increases. The capacitive reactance, X_C, is given by the equation

$$X_C = \frac{1}{2\pi fC} \tag{4.11}$$

where X_C is the capacitive reactance in ohms, f the frequency of operation, and C the capacitance in farads (π has a value of approximately 3.1415926535).

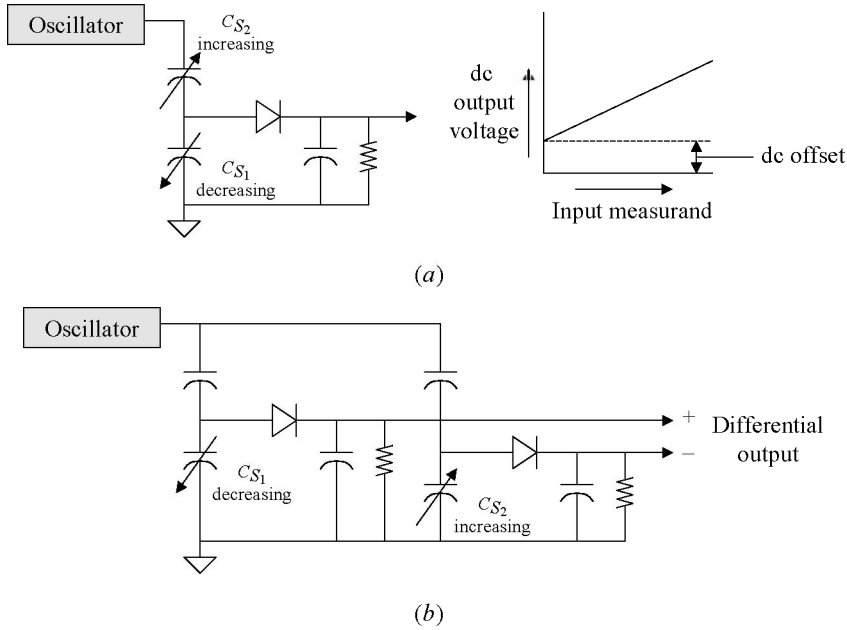

Figure 4.13 (*a*) Dual sensing element with a single-diode demodulator; (b) dual sensing element with a dual-diode demodulator.

With C_S having increased capacitive reactance, the voltage drop is greater across C_S and less across C. In the dual-element sensor, one capacitance increases with the measurand as the other capacitance decreases. This doubles the output signal full-scale range and thus increases the *S*/*N* ratio. Once the variable ac voltage amplitude signal is obtained, it is then necessary to convert it into a more usable form, such as a varying dc voltage. After the dc voltage is obtained, it can be changed to a digital signal by using an analog-to-digital (A/D) converter.

Two methods to convert an ac voltage amplitude to a dc voltage include the use of either a diode type of demodulator or a synchronous demodulator. A diode demodulator can be configured as in Figure 4.13. A single-diode demodulator provides a variable-voltage output with a dc offset, as shown in Figure 4.13*a*. The dual-diode demodulator provides a differential voltage output as shown in Figure 4.13*b*. The differential voltage can be connected to the inputs of a differential amplifier. If the (+) output of the demodulator is connected to the (+) input of a differential amplifier and the (–) to the (–), the amplifier output will center on zero volts. For the single-diode demodulator, a reference voltage or resistive voltage divider (not shown) can be set to equal the dc offset voltage of the demodulator and fed into the differential amplifier (–) input. The demodulator output would then be fed to the amplifier (+) input.

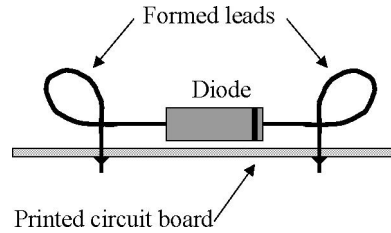

Figure 4.14 When using a demodulator diode provided with leads, thermally induced stress can be reduced by forming the leads into loops.

A diode demodulator is asynchronous, because there is no timing requirement between the signal phase and operation of the diode. The diode just conducts when the anode voltage is higher than the cathode voltage, thus charging up the filter capacitor. With silicon diodes, the anode must be approximately 0.60 V higher than the cathode for conduction of the diode to take place. This is called the *forward bias voltage*. One problem is that the magnitude of the forward bias voltage changes with diode conduction current, temperature, and mechanical stress on the diode. The temperature sensitivity is about −2.2 mV/°C. The temperature sensitivity is a major source of error in a diode demodulator. In a dual-diode demodulator, the temperature sensitivities of the two diodes approximately match each other, but temperature compensation may be needed to obtain high-accuracy signals. Using a matched diode pair molded into one case can help, but thermal gradient-induced stress can cause unpredictable shifts in forward bias voltage. One technique used to reduce the thermally induced stress on the die within a diode with leads (not surface mount) is to form a loop in the leads before soldering them into the printed circuit board, as shown in Figure 4.14.

In a synchronous demodulator, an electronic switch is used instead of a diode. The switch closes when the signal voltage amplitude is higher and opens when it is lower. Closing the switch allows the filter capacitor to charge, but without the errors of the forward-bias voltage of a diode demodulator. For proper operation, the switch(es) must be opened and closed at the appropriate times, synchronous to the timing of the oscillator. With a square-wave oscillator driving the capacitive sensor, the switch can be operated by the oscillator output as shown in Figure 4.15.

4.7 GUARD ELECTRODES

It is possible for a nongrounded movable plate of a capacitive sensor to pick up unwanted signals from nearby circuitry, as well as to be affected by capacitive coupling with nearby conductors. To minimize the errors caused by these sources, a guarding scheme is often used in conjunction with capacitive sensing elements and transducer circuits. The guard is a shield con-

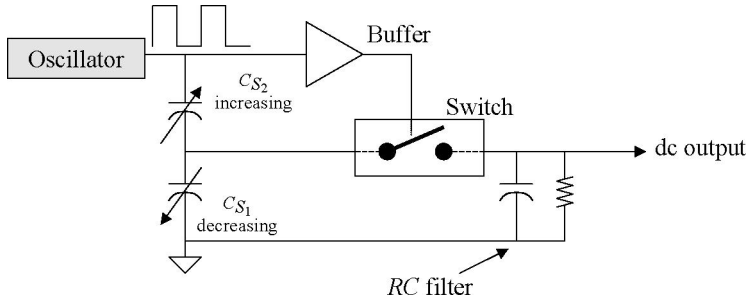

Figure 4.15 A synchronous demodulator derives a dc amplitude signal from a variable ac waveform by operating switches synchronously to the ac waveform.

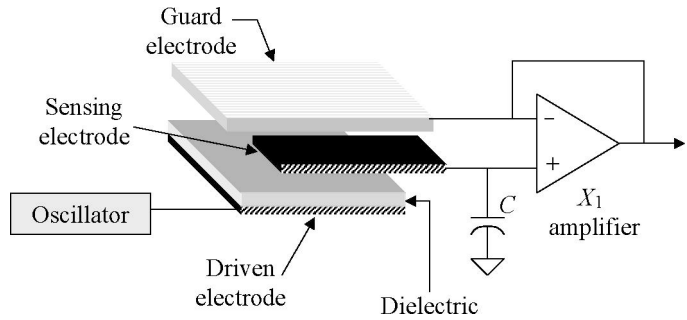

Figure 4.16 Driving the guard electrode with a voltage similar to that of the movable sensing electrode. The driven electrode and sensing electrode comprise the sensor capacitance. This is similar to Figure 4.12, with guard added, and C is swapped with C_S (because here, C is connected to common).

ductor that surrounds, or is adjacent to, the movable sensing plate. The guard is larger than, or moves with the plate, and is electrically driven with a voltage equal to that of the plate (see Figure 4.16). This is accomplished by driving the guard electrode from the low-impedance output of a buffer amplifier. Since the buffer output impedance is low, it can drive the guard electrode without being affected by the unwanted noise energy. There is no current flow between the sensing plate and the guard (other than the input bias current of the op amp) because there is no voltage potential between them.

4.8 EMI/RFI

An *RC* oscillator, with the capacitive sensor element comprising all or part of the frequency-determining capacitance, has a tendency to synchronize or beat

with interfering EMI sources that include energy at or near one of the harmonics of the frequency at which the capacitive sensor is operating. A harmonic of a particular frequency is that frequency or any integer multiple of it. The particular frequency is the fundamental, or first harmonic. Two times that frequency is the second harmonic, three times is the third harmonic, and so on.

As the capacitance of a capacitive sensor element is just starting to charge up, there is a larger voltage difference between the capacitor voltage and the switching point of the demodulator circuit connected to the capacitor. This larger voltage gap means that it is less likely that it will be surpassed by noise voltage peaks, which would trigger the switching circuit prematurely. As the capacitor voltage nears the switching voltage, however, the amount of noise voltage needed to trigger the switch prematurely becomes very small. So the result is a variation of the circuit sensitivity to a noise spike. This sensitivity changes over the period of the cycle (and the percent charge of the capacitor). The result is that the oscillator frequency tends to track the interference frequency at a rate proportional to the phase difference between the two frequencies. This is called *beating*. The difference between the fundamental frequency and the interfering harmonic is called the *beat frequency*. As the oscillator operating frequency beats with an interference signal near one of its harmonic frequencies, an amplitude modulation of the oscillator output occurs. [If you've ever heard two nonsynchronized outboard motors operating at the "same" speed, you can notice this. The sound gets louder and softer at a regular interval. That interval is the period of the beat frequency (the beat frequency is the difference in frequency between the two motors.] A motor synchronizer eliminates this by controlling the spark timing of one motor so that its rpm value will exactly match that of the other motor.) In a transducer, beating with an external energy source is a cause of error in the transducer output.

Even when there is not a sympathetic frequency relationship between the transducer electronics and a noise source, interference can still occur. Energy spikes in the noise level can induce error spikes in the transducer. Both of these effects can be reduced by shielding or guarding of the sensing element. Adequate bypassing of possible noise voltage on the power supply and other input–output lines can also help.

4.9 TYPICAL PERFORMANCE SPECIFICATIONS AND APPLICATIONS

The most popular type of capacitive linear position transducer is designed to detect the position of a metal target with respect to the transducer element, as was shown in Figure 4.5. A typical set of specifications for this configuration is as follows:

Housing dimensions:	25 mm in diameter by 75 mm long, with external mounting thread
Operating range:	0 to 8 mm from sensing face
Operating temperature:	–20 to 85°C
Output type:	analog voltage, 0 to 5 V dc
Power supply:	8 to 26 V dc
Repeatability:	less than 0.1% of FSR
Hysteresis:	less than 0.1% of FSR

Sensors of this type are often used in industrial processes where noncontact measurement of roller position or other moving parts is desired. No maintenance is required because there are no rubbing parts, although the full-scale range is limited when compared to magnetostrictive transducers or long-stroke LVDTs. The long-term accuracy is about the same as in an LVDT.

CHAPTER 5

INDUCTIVE SENSING

5.1 INDUCTIVE POSITION TRANSDUCERS

The first thoughts in designing an inexpensive position transducer often involve an inductive type of sensing element. This is because they can be simple in theory, as the basic sensing element is made from one or more coils of wire, together with a movable core (see Figure 5.1). In some applications, though, an acceptable trade-off among cost, performance, and measuring range can be difficult to achieve. Still, they have a wide range of practical applications where fairly accurate measurements are required over a relatively short measuring range (up to several tens of millimeters) or where medium accuracy over a medium range will suffice. They are generally not considered to be very practical over longer measuring ranges of more than 500 mm because of the difficulty of winding the coil. Longer ranges become more expensive than with other sensing techniques. It is also much more difficult to make each sensor perform the same as the next, due to the imperfect control possible in winding a long coil of magnet wire. Like a capacitive position-sensing element, an inductive one does not produce a directly usable output. An electronic circuit is needed to provide an ac drive to the sensor, then demodulate and condition the signal to produce the desired output.

The outward appearance of the transducer housing is often similar to that of an LVDT (see Chapter 6). Since the sensing element is a noncontact device, the complete transducer can operate with very little wear or mechanical loading force. Typical inductive position transducers do make contact to the

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

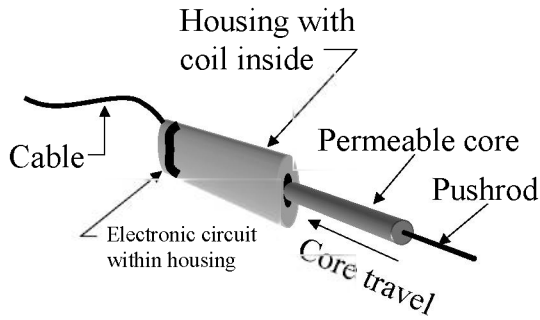

Figure 5.1 An inductive position transducer includes a sensing coil, a magnetically permeable core, electronic circuit, and housing. The electronic circuit can be included within the housing or mounted externally.

part being measured, however, through the use of a pushrod. One end of the pushrod rides against the measured part, while its other end moves a magnetically permeable core piece within the sensing element.

Inductive proximity sensors (*prox sensors*) are usually not called linear position transducers, so are not covered in this work, except for a short mention here. They are truly noncontact sensors, since they directly detect the magnetically permeable and/or conductive item to be measured. Since they operate at relatively high frequency, changes in the coil circuit due to the magnetic permeability of a ferromagnetic target or by eddy current dissipated in a conductive target can be detected. The signal from most prox sensors is a switched output that switches when the target comes within the calibrated range. Continuous functions are also possible but are typically nonlinear and affected by the shape and size of the target.

5.2 INDUCTANCE

Inductance is a property of all electrical conductors. This is because a current flowing in a conductor is always accompanied by a magnetic field. Electromagnetic energy is stored in the field. The inductance of a system is a measure of its capability to store electromagnetic energy. An analogy can be drawn between an electrical circuit with current flowing (in amperes, i.e., coulombs per second) through an inductance, and a water system with water flowing (e.g., in liters per second) in a pipe when the pipe wall has an elastic quality. For the purpose of this analogy, it is easier to envision a pipe made from rubber, although a metal pipe is also elastic; but the elasticity of a metal pipe is too small to observe visually. When water is forced to flow in the pipe, the forcing pressure tends to expand the elastic pipe as the flowing water progresses. During this time, when the pipe is still expanding, energy is being stored in the form of a volume of water at a pressure due to the pipe elastic-

ity. When the pipe has been fully expanded for the given forcing pressure, the water will be flowing out of the pipe at a steady rate as long as the forcing pressure at the input to the pipe is constant. Under this condition, there is a steady pressure drop across the length of the pipe (from one end to the other) that depends on the flow rate. If the force being applied to the water at one end of the pipe is increased or decreased, the total flow rate at the other end will tend to remain the same for a while because the elastic pipe will make up the difference. For example, with an abrupt increase in the input pressure, the pipe will expand further (building up more energy), thus delaying an increase in the output flow rate for awhile. For an abrupt decrease in the input pressure, the pipe will relax, thus supporting the flow rate for awhile.

Additionally, if a valve is placed in the output line and the opening is suddenly reduced, the inertia of the water will cause a pressure increase in the pipe. For a while, this increase in pressure will act to delay the output flow rate from changing by as much as the smaller opening size would otherwise dictate. The pressure drop from one end of the pipe to the other is analogous to the voltage drop across an inductor coil. A function of the inertia or mass of the moving water, together with the elasticity of the pipe, is analogous to the inductance of an electrical inductor. (The inertia and elasticity have effects similar to those of the core permeability and number of turns of an inductor.) The flow rate (in liters of water per second, for example) is analogous to the electrical current through the inductance (i.e., coulombs per second). An inductance in a circuit tends to impede a change in current flow, just as the elastic pipe and the inertia of the water work together to impede the change of water flow rate. Similarly, an inductance depends on both the core permeability and the number of turns of wire of the coil to form the inductance. A simple inductor is formed by a coil of wire, with the material in the center of the coil being called the *core*, as shown in Figure 5.2.

When several inductors are connected in series, the total inductance, L_T, is the sum of the individual inductances:

$$L_T = L_1 + L_2 + L_3 \tag{5.1}$$

When several inductors are connected in parallel, the total inductance is equal to the reciprocal of the sum of the reciprocals of the individual inductances:

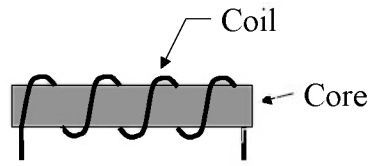

Figure 5.2 Construction of a simple inductor, with coil and core.

$$L_T = \frac{1}{1/L_1 + 1/L_2 + 1/L_3} \tag{5.2}$$

The formula for adding series-connected inductors is similar to that for adding series-connected resistors. The formula for adding parallel-connected inductors is similar to that for adding parallel-connected resistors.

The coil in Figure 5.2 is shown as a single layer, although real coils are usually made with many layers of wire. A multilayered coil that is wound onto a straight form is called a *solenoid-wound coil*. The core material may be air, but instead, it is often made from a material having a higher magnetic permeability than air, such as the alloys or mixtures of iron, nickel, or cobalt. This is the case shown in Figures 5.1 and 5.2 and normally used in the design of an inductive linear position transducer. With a core of higher permeability, a higher inductance is achieved for given coil dimensions and number of turns as compared to the same coil if wound with an air core. The formula for finding the inductance, L, of a coil varies with the shape of the coil and core, but for a solenoid-wound coil on a straight cylindrical core, where the core length is much greater than the core diameter, it is approximately

$$L = \frac{N^2 \mu_a A}{l} \tag{5.3}$$

where the unit for inductance is the henry, N the number of turns of wire wound around the core, A the cross-sectional area of the coil in square meters, μ_a the absolute permeability of the core in henries per meter, and l the core length in meters (absolute permeability, μ_a, is the product of the relative permeability of the material, μ_r, and the permeability of vacuum, μ_0; see Section 5.3).

The magnitude of electromagnetic energy, W, stored in an inductor is directly proportional to the inductance, L, and varies with the square of the current, I, flowing through the coil:

$$W = \frac{1}{2} L I^2 \tag{5.4}$$

where W is the energy expressed in joules, or watt-seconds. W remains constant when the current in the inductor is constant.

Inductance in an electrical circuit opposes a change in current by generating what is called a *back EMF*. *Electromotive force* (EMF), measured in volts, is the theoretical agent that tends to produce or maintain an electric current in a circuit. It is the electrical energy per unit charge:

$$E = \frac{W}{Q} \tag{5.5}$$

where E is EMF in volts, W is energy or work in joules, and Q is charge in coulombs. The direction of EMF in a portion of a circuit containing a source of EMF is the direction in which a positive charge would be forced in a circuit containing only this single source of EMF [11, p. 943]. The resulting EMF of a circuit is the sum of all the sources of EMF in the circuit, those acting in one direction being called positive and those in the opposite direction being called negative. Those acting in a direction opposite to the resulting EMF are called back EMF.

When the current through a coil with inductance, L, is changing, the EMF (E) across the coil is proportional to the rate of change of current, dI/dt:

$$E = L\frac{dI}{dt} \tag{5.6}$$

where the unit for inductance is the henry. The inductance of a conductor can be increased by forming it into a coil having cross-sectional area A, length l, and number of turns N. The resulting inductance of the coil is

$$L = \frac{N^2}{\Re} \tag{5.7}$$

where $\Re$ is the magnetic reluctance.

For a coil of uniform length, l, which is much larger than its cross-sectional dimension, and core material with relative permeability μ_r, the reluctance is

$$\Re = \frac{l}{\mu_0 \mu_r A} \tag{5.8}$$

where μ_0 is the magnetic permeability of vacuum ($4\pi \times 10^{-7}$ H/m) and μ_r is the relative permeability of the core material on which the coil is wound. Relative permeability of a material is the ratio of its permeability to that of vacuum (μ_r is near 1 for most materials, except that it is much higher for ferromagnetic materials such as iron, nickel, and cobalt). So then, for the case where the length, l, is much larger than the coil cross section, formulas (5.7) and (5.8) can be combined to find the coil inductance, which is approximately

$$L = \frac{N^2 \mu_0 \mu_r A}{l} \tag{5.9}$$

This is the derivation of formula (5.3), since μ_a is the product of μ_0 and μ_r. In addition to the inductance, a real coil has a parallel capacitance due to the adjacent turns of the wire. There is also a series resistance resulting from the resistance of the length of wire used to form the coil. A circuit approximation of a real inductor coil is shown in Figure 5.3.

While inductive reactance, X_L, increases with increasing frequency (equation 5.10), capacitive reactance, X_C, decreases (equation 5.11).

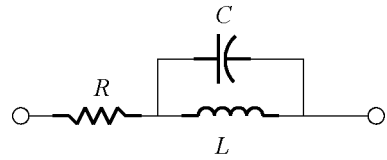

Figure 5.3 Circuit representation of the inductance, L, capacitance, C, and resistance, R, of a real inductor.

Because a real inductor has a parallel capacitance, there exists a frequency where the capacitive reactance is equal to inductive reactance. This frequency is called the *resonant frequency*. When the coil is energized at the resonant frequency, the impedance is resistive. (Impedance, in ohms, is the vector sum of capacitive reactance, inductive reactance, and resistance. By convention, the imaginary part is positive for capacitive and negative for inductive.) When below the resonant frequency, the impedance is inductive. Above it, the impedance is capacitive. The formula for inductive reactance is

$$X_L = 2\pi FL \tag{5.10}$$

where X_L is the inductive reactance in ohms, L the inductance in henries, and F the operating frequency. The formula for capacitive reactance is

$$X_C = \frac{1}{2\pi FC} \tag{5.11}$$

where X_C is the capacitive reactance in ohms and C is the capacitance in farads. So it follows that the condition for resonance is when

$$2\pi FL = \frac{1}{2\pi FC} \tag{5.12}$$

An inductive position transducer is usually designed to operate far away from its resonant frequency to avoid having a nonlinear response. However, a short stroke sensor can use resonance to increase sensitivity, or resonance can be used to correct a characteristic that is already nonlinear due to other constraints of the application.

5.3 PERMEABILITY

A magnetic field is often represented by lines of flux: The greater the magnetic flux density, B, the closer together the flux lines are drawn. The magnetic permeability, μ, of a material, is the ability of that material to support magnetic flux. In vacuum, the magnetic flux density is related to the magnetic field intensity by

$$B = \mu_0 H \tag{5.13}$$

where B is expressed in newtons per ampere-meter (N/A·m); H is the magnetic field intensity, or magnetizing force, in A/m; and μ_0 is the *permeability of vacuum*. The permeability of vacuum is $4\pi \times 10^{-7}$ H/m, or 1.257 μH/m. Since the relative permeability of a material is a ratio, equal to the ratio of the permeability of the material to that of vacuum, it has no unit of measurement. The relative permeability, μ_r, of vacuum is defined as 1. The relative permeability of dry air is 1.0006, so it is almost the same as vacuum.

All materials have magnetic properties to some extent. Those known as ferromagnetic materials have relatively high relative permeability, on the order of 50 to 5000, depending on the alloy used. These include iron, nickel, cobalt, and their alloys. An atom in a ferromagnetic material has a magnetic moment that tends to align itself with an applied magnetic field. Nonferromagnetic materials typically have a relative permeability very close to 1. Those with a relative permeability only slightly above that of vacuum, such as aluminum and oxygen, are called *paramagnetic*. Materials having a permeability less than that of vacuum, such as copper and nitrogen, are called *diamagnetic* [17, p. 396].

The higher the relative permeability, the slower a magnetic field will travel and the larger will be the inductance of a coil wound on this material and having a given size and turns count. The velocity of magnetic field propagation is not a linear relationship with permeability. As stated in Chapter 4, the velocity of magnetic field propagation, v, in a material is inversely proportional to the square root of the permittivity and the permeability of the material:

$$v = \frac{1}{\sqrt{\varepsilon_a \mu_a}} \tag{5.14}$$

where ε_a is the absolute permittivity of a material, being the product of the permittivity of vacuum, ε_0, and the relative permittivity, ε_r, of the material:

$$\varepsilon_a = \varepsilon_0 \varepsilon_r \tag{5.15}$$

and where μ_a is the absolute permeability of a material, being the product of the permeability of vacuum, μ_0, and the relative permeability, μ_r, of the material:

$$\mu_a = \mu_0 \mu_r \tag{5.16}$$

5.4 HISTORY OF INDUCTIVE SENSORS

The English scientist Michael Faraday discovered in 1831 that a changing magnetic field in one electric circuit induced a current into a nearby circuit. This is now called *electromagnetic induction*. It was already known at that time that

passing a current through a wire conductor caused the wire to be surrounded by a magnetic field, as indicated by its effect on a magnetic compass needle. Faraday showed that the reverse was also true (i.e., that in addition to a current causing a magnetic field, a magnetic field can cause a current). Faraday demonstrated this by moving a magnet within a coil and detecting the current induced into the coil. The relationship between a changing magnetic field and the current induced is now called *Faraday's law*. An American, Joseph Henry, made similar discoveries at about the same time. The unit of inductance, the henry, is named for him. In addition, when a magnetic field is produced by a current flow through a coil, electromagnetic energy is stored in that magnetic field. This is called *self-induction*.

Inductive linear position sensors, also known as *variable-inductance sensors*, have been widely used since the middle of the twentieth century. They were first used in the very beginning of the twentieth century. Being easy to fabricate but difficult with which to obtain high accuracy, variable inductance sensors have mostly been incorporated into measurement systems where simplicity was more important than performance. Their capability to be used in noncontact measuring configurations also lends their use to high-reliability applications or other areas where elimination of wearing components is important. The sensing element can be designed for use at high temperatures of over 150°C, although the electronics module may need to be mounted in a lower-temperature area unless specially designed for high-temperature use. High-temperature semiconductors have been built using silicon carbide (SiC), gallium arsenide (GaAs), silicon on insulator (SOS), and several dielectrically isolated CMOS processes. Since the basic sensor element, comprising a coil and a movable core is simple, most improvements in performance over the years have been accomplished by advances in the related circuitry. These improvements include higher-temperature operation, reduced temperature sensitivity, greater stability, smaller size, and lower power consumption.

5.5 INDUCTIVE POSITION TRANSDUCER DESIGN

A typical inductive sensor to be used as a component of a linear position transducer was shown in Figure 5.1. It comprises a movable core, made from a ferromagnetic material, and a coil of wire wound onto a bobbin. When the core is mostly outside the coil bobbin, the coil inductance is lower. As the core moves into the bobbin and increases the average coil permeability, the coil inductance increases according to equation (5.3). In a practical position transducer, a nonferromagnetic pushrod (e.g., aluminum or plastic) is attached to the core to break the magnetic circuit between the core and the movable element that will be measured. By way of the nonferromagnetic pushrod, the core is then caused to move with changes in the measurand.

Since the core can move in and out of the coil, the total sensor length with the core fully inserted is almost as short as the length of the coil. The total

Figure 5.4 Inductive linear position transducer with rod ends. (Courtesy of Positek.)

sensor length when the core is nearly pulled out of the coil approaches the sum of the coil and core lengths. This additional length is one of the drawbacks to using an inductive sensor, although the overall length is not as great as that in an LVDT for a given measurement range.

The transducer shown in Figure 5.4 has a housing, a movable rod, and eyes for attachment at each end. The eyes allow mounting the transducer between two movable pins or bolts, as when mounted in parallel to a shock absorber, so that each end is free to pivot around the mounting pins or bolts. Inside the transducer, the coil is rigidly mounted within the housing, while the core is attached to the rod that moves in and out. Between the rod and the housing are installed rod guides and wipes. The guides make sure that the rod moves freely without binding and help to extend the mechanical life of the transducer (life = number of motion cycles possible before wearing out the guides). The wipes prevent foreign material that may be stuck to the rod from entering the guides. Otherwise, the foreign material could damage the guides and reduce the life. Also included within the housing is the electronic circuit.

5.6 COIL

It is important to select carefully the proper material for the bobbin that supports the coil. Any warpage of the bobbin during its manufacture or later, due to temperature variations, can produce a change in the coil inductance

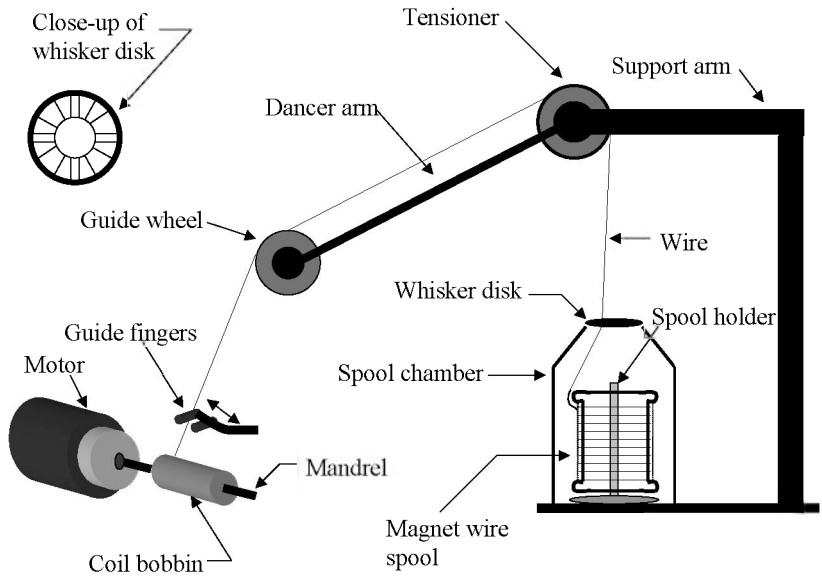

Figure 5.5 Basic elements of a single coil winder.

and result in an error in sensor output. A glass-filled high-temperature plastic is usually specified to provide this temperature stability. The coil-winding machine, coil supporting tools, and wire guide tool must be designed and calibrated to produce a repeatable winding profile, so that each sensor coil will be like the next. A spool of magnet wire is usually placed, with its axis in the vertical position, inside a spool chamber, which is a tapered clear plastic tube with the top open (see Figure 5.5). The top of the spool chamber is narrower than the spool; the bottom is wider.

A set of *whiskers* made from short, radially mounted strands of monofilament fishing line is mounted on top of the spool chamber. The whiskers add some resistance to the wire as it is stripped from the spool and help to eliminate tangles. Then the wire goes to a tensioner and dancer arm before winding onto the bobbin through the winding fingers. The tensioner comprises one or more disks over which the magnet wire passes. The disks have an adjustable brake to control the rotation force and provide the desired tension on the wire between the tensioner and the coil. The dancer arm is a long spring that takes up the nonuniformity in wire feed rate as the machine speeds up and slows down. The tensioner and dancer arm work together to prevent the magnet wire from breaking or stretching excessively during the winding operation. The winding machine itself has a mandrel to hold the bobbin, a traverse to hold and control the guide fingers, a motor to turn the mandrel, and a turns counter. There is usually a speed-up and slow-down cycle that engages when starting and stopping the winding process. The last few turns are wound at slow speed in order to count the correct number of turns accurately before stopping.

After the coil is wound, the ends must be attached to lead wires electrically and mechanically, since the coil wire is normally too small in diameter to be used as attachment leads. Before or after lead attachment, varnish and/or adhesive tape is often used to stabilize the coil assembly before placing it into the coil housing. Sometimes the wire ends are skeined before lead attach. This means that the last few centimeters of the wire are doubled over one or more times and twisted together to increase the strength of the wire where it attaches to the pin.

In addition to providing mounting means and protection for the coil, the housing provides magnetic shielding to prevent inductance changes due to nearby magnetic materials. To provide this shielding, the housing must be made from a ferromagnetic metal or alloy. A common selection is to use a (ferromagnetic) steel housing with nickel plating. The nickel plating is also ferromagnetic and adds corrosion protection and better appearance. The nickel plating is normally applied using an electroless process, where the part is immersed in a liquid that contains nickel ions that plate onto the surface. The nickel thickness is determined by the time and temperature of immersion. The finish is smooth and shiny, as opposed to an electroplating process, where polishing would be required to yield a shiny surface. After the coil is assembled into its case, a liquid potting material is often added, which then cures into a solid. This adds mechanical protection and seals against the entry of moisture.

The coil is usually wound with an enameled copper wire, called *magnet wire*. The insulating coating can be heat-strippable to aid in soldering to the leads. One disadvantage in using copper wire is that the temperature coefficient of copper is about 5400 parts per million per degree Celsius. If the sensor circuit is sensitive to resistance change, the change can be compensated by using either a separate compensating coil or a temperature-compensating resistor. The use of a compensating coil is explained in Section 5.8. A temperature-compensating resistor can be one having a negative temperature coefficient (NTC) in series with the copper wire coil or located on the electronic circuit board, or positive temperature coefficient (PTC) on the electronic circuit board. The normal NTC type of resistor is called a *thermistor* and has a nonlinear temperature coefficient where the resistance decreases as the temperature increases. The common type of PTC is based on the element nickel and has a relatively linear positive temperature coefficient. One can also fabricate a PTC resistor by winding nickel wire onto a form (as the author did for compensating a product designed in the early 1970s, before PTC resistors were commonly available). PTC or NTC resistors can also be used to compensate for temperature shifts that are due to the electronic circuit and not to the sensing element.

Another way to remove the problem of temperature-induced changes in the resistance of a coil is to wind it with a wire that has a near-zero temperature coefficient of resistance. This is called *manganin wire* [32, p. 90] and is an alloy of copper, manganese, and iron (Cu/Mn/Fe) with a temperature coefficient of resistance of $+0.00002\,\Omega/\Omega/^{\circ}C$ [6, p. 75]. It is available from magnet

wire manufacturers in the same sizes and with the same insulating coatings as those of other magnet wires.

5.7 CORE

The core material must be selected to provide the desired amount of inductance change as it moves into the coil. In addition to that, it is desirable that the core have some additional favorable properties. These properties include long-term stability, temperature stability, low power loss at the operating frequency (relating to magnetic hysteresis), and some corrosion resistance. Core materials are usually nickel–iron alloys, with some other materials added to tune the magnetic and mechanical properties. Some common ones include Ni-Span C, Hiperco 50B, and various forms of Permalloy. Ni-Span C has the advantage of very low permeability sensitivity to temperature change. Hiperco 50B has a higher absolute permeability. The manufacturer's data sheet must be consulted to define the proper annealing versus coldworking finished condition in order to obtain the desired magnetic properties. With most core materials, there is a final annealing process that relieves the mechanical stress from machining and makes the entire core magnetically uniform. A batch of cores is brought to an elevated temperature for a set period of time. The annealing schedule is available from the manufacturer of the alloy to be used. During the annealing process, oxygen is excluded to prevent oxidation, and a reducing atmosphere is generally used. The reducing gas can be hydrogen, which must be used with a method to prevent explosion (explained in Section 6.5). Alternatively, a forming gas can be used that has insufficient hydrogen concentration to support an explosion but sufficient hydrogen to chemically reduce the surface of the core material.

5.8 SIGNAL CONDITIONING

Since a change in the measurand results in a corresponding inductance change in a variable-inductance position sensor, an electronic circuit is needed to measure that inductance change. Accordingly, the electronic circuit includes the functions to drive the sensing coil, measure its inductance, and produce the desired output signal. The sensing coil is normally energized with an ac driving voltage from a sine-wave oscillator. The coil's inductance change can be indicated by a changing oscillator frequency or a changing amplitude. For compensating errors from temperature variations or other sources, two or more coils can be incorporated into the sensor and a comparison between them used to produce an output signal.

Figure 5.6 shows a simple circuit with a single variable inductor producing a change in amplitude corresponding to a change in inductance. In this case the voltage across the inductor is not zero when a zero output voltage is

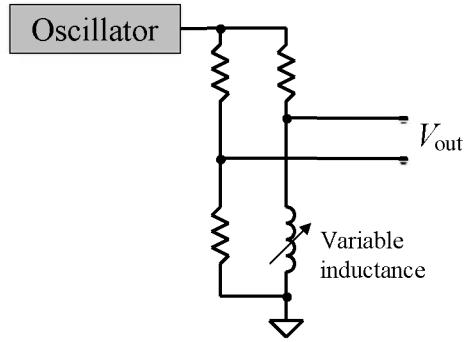

Figure 5.6 Change in the variable inductance, L, results in a change in the amplitude of the differential output voltage.

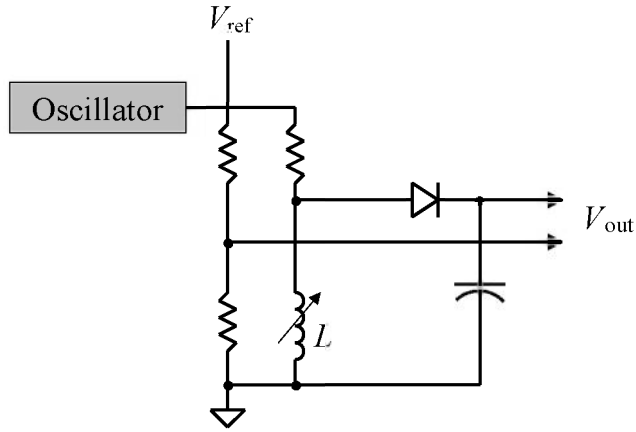

Figure 5.7 Variable-inductance sensor circuit with a dc output voltage.

desired. So a resistive divider is used to provide an offset equal to the same voltage as the inductor would have with a measurand of zero. A differential amplifier can be used to subtract these two voltages to obtain a zero-based output. The circuit of Figure 5.6 provides an ac output voltage of varying amplitude, but a dc output is often desired, with a positive voltage indicating displacement in one direction and a negative voltage indicates displacement in the opposite direction. This can be accomplished by using a circuit such as that shown in Figure 5.7.

The same considerations on demodulation techniques (diode demodulator versus a demodulator circuit using a semiconductor switch) would apply as those shown in Chapter 4 and will not be repeated here. A circuit using a variable inductance, a nonvariable inductance (for error compensation), and a diode demodulator is shown in Figure 5.8.

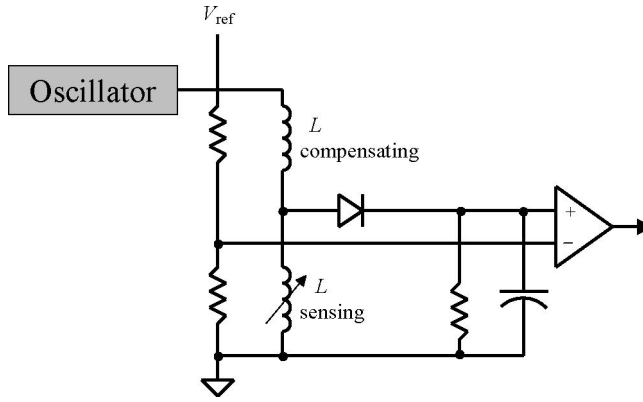

Figure 5.8 Sensing circuit with compensating coil and dc output voltage.

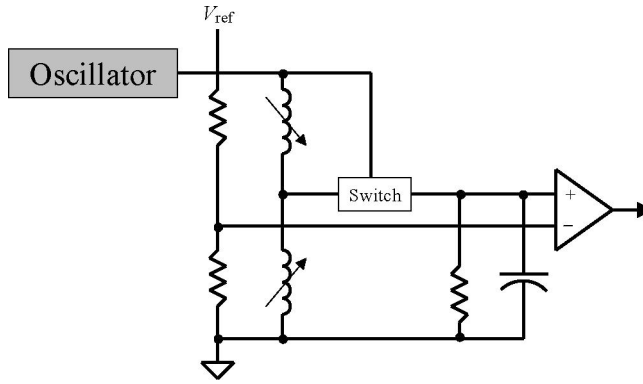

Figure 5.9 Dual-coil variable-inductance linear position sensor with semiconductor switch and dc output.

Greater sensitivity and measuring range can be obtained by using two sensing coils arranged as shown in Figure 5.9. When the core is centered between the coils, the ac voltage on each coil is the same. As the core moves more into one coil, the inductance of that coil increases as the inductance of the other coil decreases. The coil resistances vary the same as each other and do not need compensation. The voltage at the connection between the coils is lower than the voltage at the resistive voltage divider when the measurand is at minimum. The voltage at the connection between the coils is higher than the voltage at the resistive voltage divider when the measurand is at maximum. The differential amplifier produces an output voltage swing that can be centered on zero to provide a ±dc output, or it can be adjusted for zero to full-scale output voltage, as desired.

5.9 ADVANTAGES

A main advantage of variable-inductance position sensors is their simplicity. This advantage is somewhat eroded by the need for an electronic circuit with a driver and demodulator to form a complete transducer, but it is still a popular transducer type. Another advantage is that when there is no need for supporting bushings, bearings, and so on, the sensor can be implemented as a noncontact device, thus having a nearly infinite lifetime. In addition, typical inductive sensors with an analog output voltage have a nearly infinite resolution, determined only by the ability to read the signal over whatever noise may be present. Inductive sensors are also less affected by dust and humidity than are capacitive sensors. A disadvantage is the relative lack of precision and stability that is attainable with other technologies, such as magnetostrictive linear position transducers and some encoders.

5.10 TYPICAL PERFORMANCE SPECIFICATIONS AND APPLICATIONS

Whereas magnetostrictive linear position transducers can provide an accuracy in the range of 0.01% and an LVDT in the range 0.1 to 0.5%, inductive position transducers are generally limited to the approximate accuracy range 0.2 to 1.0%, depending on the measuring range. This is not a problem for many applications where the most important aspect is to have a monotonic response with high resolution, such as in many position control loops.

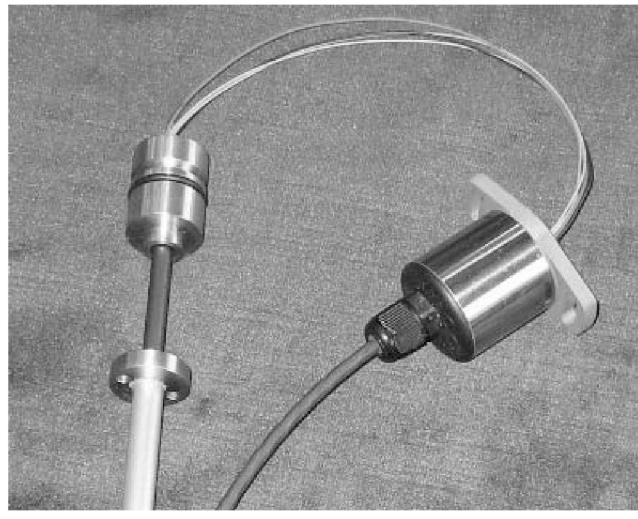

Figure 5.10 Inductive linear position transducer. (Courtesy of Positek.)

A typical inductive linear position sensor is shown in Figure 5.10, with external signal conditioning electronics. Typical specifications for an inductive linear position transducer are:

Full-scale range:	10 to 600 mm
Input voltage:	5 V dc
Supply current:	10 to 20 mA
Output:	0.5 to 4.5 V dc, ratiometric
Nonlinearity:	±0.25%
Resolution:	nearly infinite
Operating temperature range:	–40 to 125°C

One application of an inductive linear position sensor is the measurement of the extension of smaller sizes of hydraulic cylinder. The robustness of the inductive sensor lends itself well to this installation, but there is a limitation on accuracy and maximum length. Where a longer measuring range or higher accuracy is required, a magnetostrictive transducer is often used. The magnetic coupling of a magnetostrictive transducer also makes the design of the cylinder easier. Inductive transducers can be used successfully in many industrial and commercial applications. Valve positioners for process control on the factory floor are a good candidate, as well as measuring the drum position in a clothes-washing machine and other position-sensing requirements for consumer goods.

CHAPTER 6

THE LVDT

6.1 LVDT POSITION TRANSDUCERS

The linear variable differential transformer (LVDT) is a position transducer that is noncontact, absolute reading, and has essentially infinite resolution. It comprises three or more coils within which a magnetically permeable core moves to provide variable coupling between the primary coil and the secondary coils (usually two). Although the detection technique is noncontact, there is often a mechanical arrangement added to keep the core positioned in the coil throughout the stroke. One example of this is the configuration called an LVDT *gauge head*. Practical linear sensors can be designed with a nonlinearity of less than 0.2% and full-scale ranges (FSRs) from less than 1 to over 100 mm. Resolution is nearly infinite. Curved and rotary sensors are possible. Popular applications include industrial machinery, such as metal forming machines and in-process dimensional verification, as well as automotive and commercial products.

LVDTs require a set of driving and conditioning electronic circuits. A typical LVDT is sold as a sensing element with a core, the electronic circuit being supplied as a separate device. An LVDT that includes all the required electronics within its housing is often called a *dc LVDT* because it operates on a dc power supply and has a dc output, although the internal operation includes the normal ac driving, demodulation, and signal-conditioning circuitry.

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

6.2 HISTORY OF THE LVDT

Although variable differential transformers were constructed earlier, the linear variable differential transformer as a position transducer was first described by G. B. Hoadley, and U.S. patent 2,196,809 was issued to him in 1940 [13, p. 3–4]. Early uses were mostly military, because of the ruggedness possible with this sensor, and some quality assurance applications, because of the inherent high resolution of the measurement. Additional improvements in construction and performance were made throughout the 1950s and 1960s. By then, the LVDT was used widely for industrial applications. The improvements that followed in the 1970s and later were mostly in the electronics that drive the LVDT and condition the signal. This included more accurate drivers, phase adjustment, and demodulation circuitry. Integrated circuits (ICs) with the complete LVDT electronics function became available as a standard product in the 1980s. Dc LVDTs were available in the 1970s with "cordwood" construction; that is, axial-leaded parts were connected between two parallel printed circuit boards. These were very difficult to design and manufacture (as the author found out from personal experience.) They had minimal functionality and performance characteristics, due to the lack of available space to house a more complex circuit. With the ICs that were available in the 1980s, the dc LVDT became more practical to design and manufacture, with good performance resulting in the finished product.

6.3 LVDT POSITION TRANSDUCER DESIGN

A basic LVDT comprises one primary coil, two secondary coils, and a movable core. The coils are wound onto a tubular coil form, the interior of which is called the *bore* (see Figure 6.1). Communication between the moving and stationary parts of the transducer is achieved by means of inductive coupling. Therefore, any number of position changes can be made without incurring

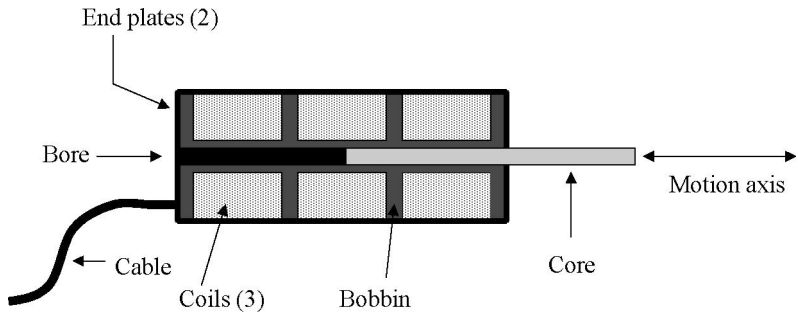

Figure 6.1 Cutaway view of an LVDT, with housing and core.

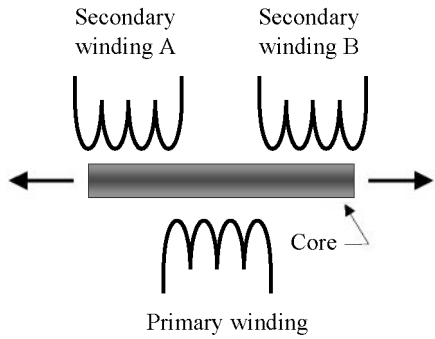

Figure 6.2 Configuration of the windings of an LVDT.

wear to the transducer parts. When using an LVDT to measure the dimensions of a formed metal part (such as a fender for an automobile), the LVDT core is linked to a stylus that contacts the surface of the formed metal part. Although there may be wearing of the stylus, there is no wear of the actual sensor element. In the example of sample testing the profile of an automotive fender, a form is made that matches the desired profile of the fender. Several tens of LVDTs are arranged in the form to measure all the critical areas of the fender shape. The data are sent to a computer for analysis. Damage and wear of the tool that is used to form the fender can be assessed. This information is used to make adjustments to the tool and to predict the maintenance that will be needed on the tool after a given number of parts are manufactured. This same method is used for sample testing of automotive windshields, except that approximately 100 LVDTs are normally used. The windshield profile must be very accurately controlled to provide good sealing and to prevent undue stress on the glass during its assembly to the car.

An LVDT core is normally a cylinder of permeable material and provides inductive coupling between the primary coil and the secondary coils. The core moves within the bore of the LVDT and does not touch the walls of the bore. As the core moves along within the bore, the distance from the center between secondaries to the center of the core is read electronically (see Figure 6.2). When the distance from the center of the core to the center between the secondaries is zero, the core is approximately in the position known as the *null*. The exact position of the null is where the sum of the outputs of the two secondaries is at its lowest value. This applies when the secondaries are connected in the series bucking configuration, explained further in Section 6.7.

To obtain a position signal from an LVDT, the primary must be driven by an ac source (an oscillator), and the voltages from the secondaries must be demodulated. This circuitry is called the signal conditioner. The basic elements of a simple signal-conditioning circuit are shown in Figure 6.3.

In transformer theory, two or more coils are coupled by mutual inductance. If a voltage is impressed onto one coil (the primary), the voltage produced at

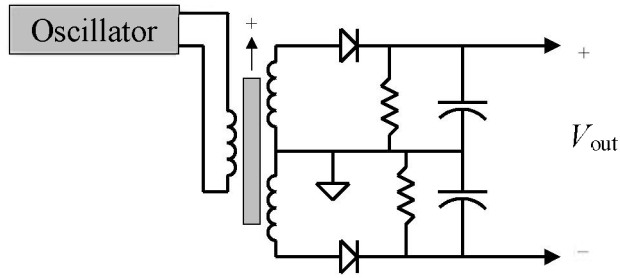

Figure 6.3 Simple LVDT signal conditioner with diode demodulator.

a second coil (a secondary) is related to the magnitude of the primary voltage by the ratio of the number of turns of the respective coils. The coil onto which the voltage is impressed is called the *primary*. The coils that produce the relative outputs are called the *secondaries*. For example, a secondary with twice the number of turns as the primary will produce an output of twice the voltage input amplitude on the primary, ignoring loading effects.

6.4 COILS

The three or more coils are wound onto a common bobbin. It is important to select the bobbin material for rigidity and stability. Any warpage of the bobbin during its manufacture or later, due to temperature variations, can produce an error in the sensor output. A glass-filled high-temperature plastic is usually specified (some common choices include glass filled-nylon, Ryton, and Torlon).

A winding machine is used to wind the primary and secondary coils onto the bobbin. The coil-winding machine can be hand operated, with a start switch to be pressed after the wire is manually started on the bobbin. The winding then stops when the set number of turns have been applied. Alternatively, an automated winding machine can be used when it is planned to make large quantities (usually, more than 50,000 parts per year). The coils are wound with light-gauge magnet wire, so it is important to control the wire tension closely as it is wound, to avoid stretching the wire. The average tension is controlled by an adjustable brake. The instantaneous tension is averaged out by using a spring-loaded guide called a *dancer arm* (see Figure 5.5).

After the coil is wound, the ends are soldered or welded to end pins or lead wires, which are heavier and more durable than the wire used to wind the coils. When the magnet wire is of a very small diameter, less than about American Wire Gauge (AWG) 38, stripping the varnish prior to soldering may be difficult to accomplish by an abrasive process without breaking the wire. In this case, it is advantageous to specify the use of a heat-strippable coating instead of varnish. This coating can be soldered directly, without stripping. Care must be taken to specify a heat-strippable coating that will operate satisfactorily at

the high end of the required operating temperature for the LVDT, while being able to be soldered at a temperature that will not excessively melt the plastic surrounding the end pins. On LVDTs rated for high-temperature use (above 100°C) this may not be possible. In that case, abrasive or chemical stripping must be used.

Varnish and/or adhesive tape are often applied over the coils to stabilize them. This ensures that the coil wires will not be damaged during the remaining process steps, provides more uniform heating and cooling, and improves the ruggedness of the completed LVDT in the field. The finished wound bobbin assembly is inserted into a housing and usually potted with epoxy to seal it and make it extremely rugged. The housing is usually made of a material that will shield the coils magnetically. This is accomplished by making the housing parts of a permeable material such as nickel-plated steel. In addition to the housing tube, end plates are added to complete the shield. This magnetic shielding prevents interference from nearby electromagnetic fields and the effects from magnetically permeable materials, which may pass in close proximity.

The temperature coefficient of the copper wire does not cause a noticeable error, since the secondary coils work together and the resistances track each other. But occasionally, it is desired to minimize the temperature sensitivity of the coil resistance. This can be accomplished by using manganin wire to wind the coils, because it has a near-zero temperature coefficient of resistivity.

Since the LVDT is a transformer, the output voltage is proportional to the ratio of the number of turns of wire in the secondary to that of the primary coils. This is true as long as the core is not saturated, and sufficient energy is induced to support the output power in the secondaries. To design a lower-power LVDT usually requires more turns in the primary coil. Then more turns are required in the secondaries to maintain the same output voltage level.

6.5 CORE

The core material is a cylindrical or tubular component made of the nickel–iron alloy Permalloy. It usually is threaded to enable attaching it to the element that is to be measured. After the core is the final shape, size, and is threaded, it is annealed. The annealing process removes mechanical stress and makes the permeability more uniform. This aids the completed LVDT to obtain a low-null, unit-to-unit repeatability, lower nonlinearity, and better temperature performance. During the annealing process, there is usually a reducing gas flowing to prevent oxidation. The annealing gas is usually hydrogen or a gas mixture that includes hydrogen. Therefore, steps need to be taken to prevent a fire or explosion (such as keeping the hydrogen percentage below that needed for combustion). If pure hydrogen gas is used, the processing operation must be performed within an explosion-protected room, and all the electric and electronic equipment and other energy-generating or energy-storing

equipment used must be either explosion-proof or intrinsically safe (see Section 2.16). The floor, walls, and ceiling of an explosion-protected room will usually be made of concrete, with a sloped ceiling and a vent at the peak, to allow and direct the escape of any hydrogen vented from the process. Since hydrogen is lighter than air, it will rise to the ceiling and be vented. The room may incorporate a weak wall or other means to limit pressures in the event of hydrogen ignition. The amount of hydrogen that could be vented from the process should be minimized by using tightly sealed process equipment, with the hydrogen flow rate indicated on a flow meter and the oxygen concentration in the process monitored by an oxygen meter (such as an electrochemical fuel cell or paramagnetic oxygen sensor). The oxygen concentration in the room should be monitored by an intrinsically safe oxygen monitor to assure that enough oxygen is available for the worker to breathe. A flammable gas monitor should also be installed in the room to warn if there is a process leak and the hydrogen concentration rises above the lower explosive limit (LEL; see Section 2.16). All intrinsically safe and explosion-proof equipment should be rated for hydrogen use (in the United States, this is group B). To avoid the use of an explosive gas while maintaining a reducing atmosphere, one can use *forming gas*. This is a mixture of inert gas and hydrogen in which the concentration of hydrogen gas is high enough to provide the needed reducing atmosphere (about 3 to 4%), but not high enough to support combustion when mixed with air in any ratio. Since forming gas has a percentage of hydrogen below 100%, the process system must have better sealing and a constant gas flow rate to keep the oxygen concentration in the low-ppm level (less than 20 ppm of oxygen).

Some common core materials include Ni-Span C, Hiperco 50B, and various forms of Permalloy. Ni-Span C has the advantage of very low sensitivity of the modulus of elasticity and magnetic properties to temperature changes in the normal range of temperatures for industrial use (–40 to 85°C). Hiperco 50B has a higher permeability than Ni-Span C. When using the core with an LVDT, the core should be connected mechanically to the member being measured by a nonferromagnetic pushrod. The pushrod, or actuator rod, is screwed into the internal thread of the core. This enables movement of the core without affecting the core permeability. The pushrod can be made from aluminum, plastic, or nonmagnetic stainless steel. However, a nonmagnetic stainless steel (such as the ASM 300 series) must be annealed after machining because the cold-working of the machining process can align the magnetic domains and cause the stainless steel to become magnetic.

When determining the dimensions of the core to be used with a new LVDT design, the most efficient path to take is to start with the core dimensions used with an LVDT currently on the market with somewhat similar characteristics to those desired. The core should extend partway into each secondary coil even when the core is positioned toward either end at full stroke. The primary criterion that determines core length is the need to exhibit the least nonlinearity error (see Section 2.5). First, take a set of readings of the output versus the

core position. Then plot or analyze the data to determine the curve shape. A straight line is desired. If the the curve bows up, so that the readings in the middle of the stroke are higher than the ideal line, the core is too short. If the curve bows down, so that the readings in the middle of the stroke are lower than the ideal line, the core is too long. Make a new set of five cores with length variations from the original in approximately 1% increments. Test these cores for nonlinearity error curve shape in the same way. When an S-curve is obtained, where the positive bow near one end of the stroke is approximately equal to the negative bow near the other end, the core length is optimum.

The LVDT housing can also have a great effect on the nonlinearity error, since it is generally made of ferromagnetic material in order to provide magnetic shielding. The author found that this is very important in developing a gauge head LVDT, as shown in Figure 6.10. To make the gauge head as small as possible, the coils are sometimes wound as free-standing coils without a bobbin. The coil is held together by the wire coating, which is softened with heat during winding, which bonds and forms a unitized coil when cooled. The coils (three) are inserted into the housing and held in place by epoxy. The author found that exact positioning of the coils within the housing had a great effect on nonlinearity error. So a fixture is used during manufacture, to epoxy them into the optimum positions.

6.6 CARRIER FREQUENCY

Although a few LVDTs are designed for operation at 60 Hz, excitation frequencies of 250 Hz to 10 kHz are more typical, with 1 kHz being the most common. Generally, a higher carrier frequency is desired in order to have a faster response of the LVDT to variation in the position. The response time is proportional to the filtering frequency for a given order (number of poles) of the filter circuit contained within the signal conditioning electronics. For a simple filter, the *carrier frequency* (also called the *excitation frequency*) must be 5 to 10 times higher than the corner frequency to which the filter is tuned, to obtain an acceptably low level of ripple in the dc output from the signal-conditioning electronics. In fact, the highest frequency of interest in the measured signal should be 0.1 times the excitation frequency [34, p. 278]. Generally, the current driving the primary at a given voltage is lower when the frequency is higher, due to the inductive reactance of the coil. The secondary impedance is also greater. A limiting factor, however, is that an excessively high carrier frequency (>10 kHz) leads to eddy current loss in the core and results in lower output signal level, more power dissipation, and greater temperature sensitivity, unless the core is very small, with a small wall thickness. So there is a trade-off between frequency response and other performance parameters when determining the carrier frequency that will be specified. The ratio between the carrier frequency and the effective filter frequency can be reduced, while maintaining a low ripple in the output voltage, by using a com-

bination of analog and digital filtering. The LVDT signal can first pass through an antialiasing analog filter that is tuned relatively close to the carrier frequency. Then the signal can be digitized by an A/D converter and fed to a microcontroller. The micro can execute a digital filtering program to reduce the signal ripple to the amount desired. If an analog voltage output is desired, the microcontroller sends the digital signal to a D/A converter. This is explained further in Section 6.8.

The digital filtering technique preferred by the author uses a set of five registers configured as a shift register. Each new datum of LVDT position information (a data cycle) coming from the A/D converter is fed into the input (new datum) end of the shift register by the microcontroller, pushing out the oldest datum from the other end and discarding it. The micro then reads the five data (four old plus one new), discards the highest and lowest readings, and takes the average of the remaining three; the average becomes the output datum. The result is that noise pulses are thrown away (likely to be the highest and/or lowest readings) and do not affect the reported measurement reading. A set of five registers, as described, is the smallest number that can be used effectively in this method; but more registers (>10) are often utilized. Using a larger number of registers yields a better (smoother) average but increases the response time to a changing measurand. A more common method uses five registers (for example), but after operating similarly on one set of five data, throws them all away and takes a new set of five data. The disadvantage of the second method is that with the example of five data, it yields a lag time of five data cycles; whereas the first method has a lag time of only one or two data cycles before some change is indicated.

It is also possible to adjust the carrier frequency in order to reduce the temperature sensitivity of the signal amplitude. There is a nominal phase-angle difference between the primary voltage and the secondary voltages for a given set of conditions. As the temperature changes, the phase-angle difference can change. By testing the LVDT output stability over temperature at several different carrier frequencies, an optimum frequency can be found that minimizes the temperature-induced span error. This frequency is called the *zero-phase frequency*.

6.7 DEMODULATION

The core is a magnetically permeable material, usually a nickel–iron alloy. When the core is centered on the primary and thus equally spaced relative to the two secondaries, the voltage induced into each of the secondaries is approximately equal. The secondaries are usually connected into a circuit such that the voltages are *series bucking*. This means that the secondary voltages subtract and will have a total series-connected output of nearly zero volts when the core is centered. A dot on the schematic indicates the polarity of the coil, and each lead that is marked with a dot has the same phase. A bucking

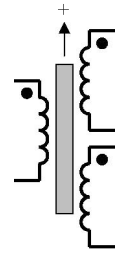

Figure 6.4 Schematic of the LVDT phasing connection for a series-bucking configuration, showing a dot to indicate same-phase leads.

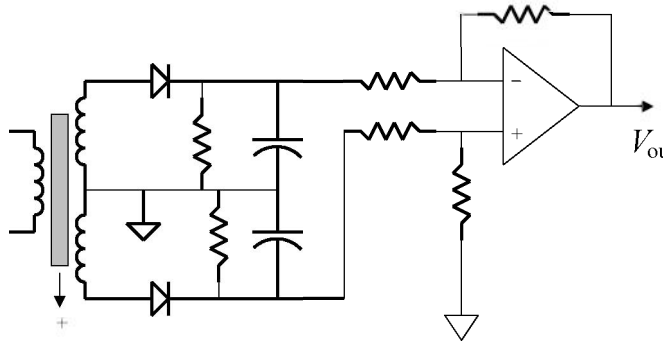

Figure 6.5 Diode demodulator with differential amplifier.

connection is as shown in Figure 6.4. When the core is centered, it is called the *null position*.

When the core is not at null, the secondary that is more coupled to the primary, due to the core being closer to more turns of its coils than to the turns of the other secondary coil, will have a greater voltage output. So the ac output voltage of the two secondaries connected in series bucking will be at the minimum when at null and increasing as the core moves away from null in either direction. To tell which direction the core has moved away from null, the secondary output must be demodulated to a dc voltage, so the polarity of the voltage will indicate the direction of core travel. The simplest way to obtain a dc output from an LVDT signal is to use the diode demodulator and differential amplifier circuit of Figure 6.5.

The resulting output across the diode cathodes is a varying dc differential voltage. That is, if we call the voltage at one cathode the reference, the voltage at the other cathode could vary from –100 to +100 mV for core positions from – full stroke to + full stroke, respectively. This differential voltage could then be amplified by the differential amplifier to provide a ±10-V dc transducer output, for example. Circuits similar to this are commonly used when the LVDT is a component part of a pressure transducer or gauge head, for

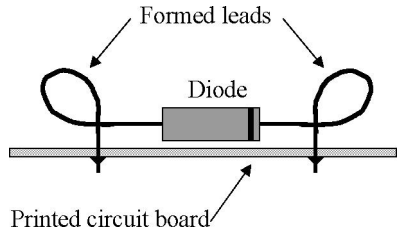

Figure 6.6 Forming the leads of a leaded discrete diode helps to reduce stress on the semiconductor die.

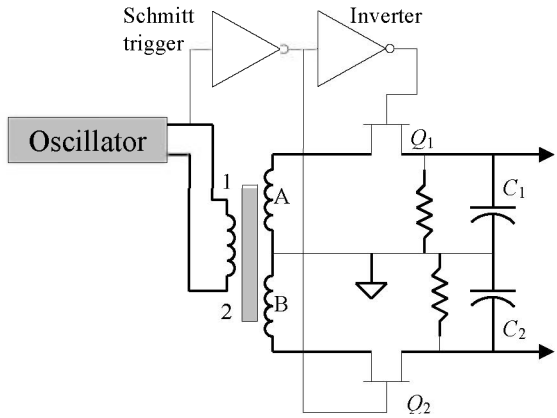

Figure 6.7 FET synchronous demodulator with square-wave excitation.

example, a disadvantage of this circuit is that the forward voltage drop of the (usually silicon) diodes has a strong temperature coefficient (about –2.2 mV/°C) and is also sensitive to mechanical stress. This is less of a problem when the demodulated dc signal from the LVDT is relatively high (≥±0.25 V). With lower signal levels, the problem is due primarily to changes in the voltage drop across one diode as compared to that across the other. There are variations in the temperature coefficient from one diode to the next, as well as variations in mechanical stress induced into the junctions by soldering and differences in thermal coefficient of expansion between the diode bodies and the circuit board that cause mechanical stress in the diode. Changing the mechanical stress on the diode changes the voltage drop. Mechanical stress can be reduced by using diodes with leads and forming them as in Figure 6.6.

A more precise method for demodulating a low-level LVDT signal is to use synchronous demodulation. A simple method developed by the author in the early 1970s, originally for use in a pressure transducer, is shown in Figure 6.7. Bipolar or Field-effect transistors replace the diodes that were shown in Figure 6.5. The voltage drop across the turned-on transistors is much smaller than the forward voltage drop across a diode (a few millivolts instead of 500 or

600 mV). The transistor voltage drop also experiences far less change over a given range of temperature (compared to that of a diode).

Although LVDTs have traditionally been driven by a sine wave, the author has found that a suitably designed LVDT can give good results when driven by an appropriate square wave and synchronously demodulated if the leads between the signal conditioner and the LVDT are relatively short (<10 cm). Short leads can be used when the LVDT is the sensing element of a transducer, such as in a pressure transducer where the LVDT measures the travel of a pressure-sensing diaphragm. This simple circuit does not include phase adjustment, as in the demodulator shown in Section 6.8, which is one reason why it is suitable for use only when the LVDT is located very close to the driver and demodulator circuits. The other reason that leads must be kept short is that a square wave can allow the generation of more complex waveforms, due to the reactance of a long cable, introducing signal variations if the cable is flexed or brought into close proximity with other dielectric or conductive materials.

In Figure 6.7, the primary is driven by a square wave (typically, 5 V dc). This square wave is used to switch on the demodulator transistors at appropriate times. For example, when the primary is going positive at pin 1, transistor Q_1 conducts to charge capacitor C_1 up to the voltage of secondary A (this explanation is for an enhancement-mode N-channel FET or an NPN bipolar transistor, but the opposite polarity may be used with the appropriate configuration). Similarly, when the primary is driven positive at pin 2, transistor Q_2 conducts to charge capacitor C_2 up to the voltage of secondary B. So the voltage across C_1 represents the voltage of coil A, and the voltage across C_2 represents the voltage of coil B. When the voltages are equal, the differential output is zero. When the voltage across C_1 is greater than that across C_2, the differential output is positive; when the voltage across C_1 is less than that across C_2, the differential output is negative.

6.8 SIGNAL CONDITIONING

The circuitry that provides excitation to the primary and demodulates and amplifies the signals from the secondaries is called a *signal conditioner*. When a signal conditioner is sold as a separate product, it normally has several adjustable features for use over a wider range of applications with a variety of LVDT models. A signal conditioner product will typically include a sine-wave generator in addition to a synchronous demodulator. Accurate demodulation is an important function of the signal conditioner, but many other features can also be critical to good performance. The LVDT excitation carrier must be stable with time and temperature and of relatively low distortion. The carrier is normally in the shape of a sine wave. The sine-wave generator circuit of Figure 6.8 was invented by the author in the 1970s as a way to produce a sine wave that is stable in amplitude and frequency. It also provides the signals

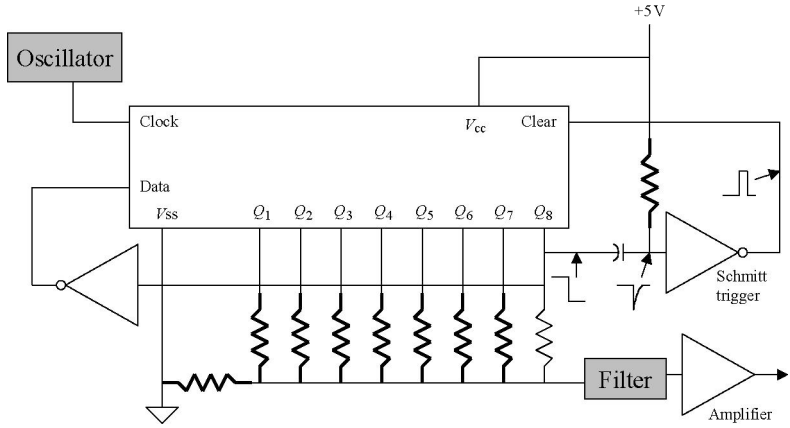

Figure 6.8 Digitally synthesized sine-wave generator.

to drive a wide-range phase adjustment circuit. Previously, phase adjustment was accomplished by analog circuitry, and the adjustment range was limited to ±45° (limited by the stability range of the analog amplifier). By using the circuit shown and adding only a ±22.5° analog phase adjustment circuit, a phase adjustment range of ±180° is easily obtained. The input of the analog phase adjustment circuit is simply switched among the Q outputs of the shift register of the sine-wave generator, as needed to come within 22.5° for the final adjustment.

The difficulty in designing a stable sine-wave oscillator is to produce a low-distortion sine-wave shape at various selectable frequencies at the same time that the frequency and amplitude are held stable with time and temperature. A Wien bridge circuit [21, p. 52] is a common way to attempt this but results in a difficult-to-compensate amplitude dependence on temperature, requiring temperature compensation of each individual circuit.

The circuit shown in Figure 6.8 synthesizes a sine wave digitally. This signal is derived from an easy-to-implement, stable, 5-V power supply. Since the CMOS Q outputs are switched between 0 and 5 V, the waveform amplitude will be as stable as the 5-V regulated supply. The input clock frequency can be adjusted while not affecting the accuracy of the sine wave shape. The values of the resistors are weighted to yield a voltage-divider output that varies according to the sine function. The output is then passed through a filter to smooth it and take away the switching spikes, leaving a clean sine wave (see Figure 6.9).

Because the sine wave is generated digitally through the shift register sequence, there is a square pulse available for every 22.5° of the full-wave period. A digital switch can select which 22.5° increment is utilized, the ±22.5° adjustment is then available from that starting point. In this way good resolution is obtained, with the adjustment covering only ±22.5°, but the

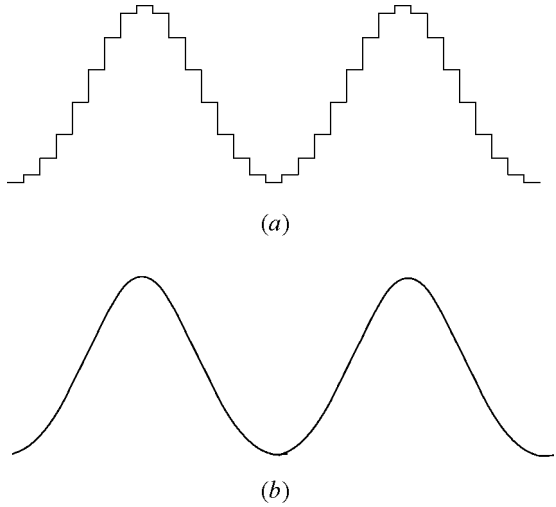

Figure 6.9 (*a*) The steps of the digitally generated staircase are weighted according to the sine function. (*b*) The filter circuit smoothes the waveform.

entire range of ±180° can be covered by using the digital switch and analog adjustment.

The typical way to drive an LVDT is to use a sine-wave driver with constant voltage. An alternative way is to drive the LVDT primary with a constant current in order to keep the current in the primary coil the same, as the coil resistance increases with increasing temperature and inductance changes with core position. Depending on the characteristic of the particular LVDT used, one or the other drive type (constant voltage or constant current) will provide the best performance over a temperature range.

Another method that can be used when a microprocessor is available is to utilize the function

$$\text{output voltage} = \frac{\text{coil 1 voltage} - \text{coil 2 voltage}}{\text{coil 1 voltage} + \text{coil 2 voltage}}$$

This tends to compensate for variation in the voltage of the sine-wave drive but assumes that the sum of the secondary voltages will remain the same as the core moves throughout its range. This assumption is true only if the core travel is somewhat limited to keep the core sufficiently within both coils so the sum A + B remains relatively constant.

6.9 ADVANTAGES

Since the LVDT core does not touch the inside of the coil bobbin, it is a noncontact sensing element. This means that many full-stroke cycles can be

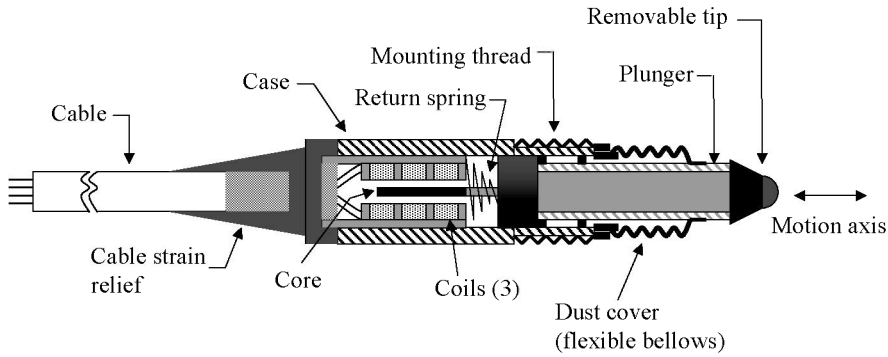

Figure 6.10 Gauge head LVDT.

endured without wear or degradation of the performance characteristics. In many applications, though, bearings are added to maintain the alignment between the core and the bore of the coil bobbin (as in the gauge head configuration of Figure 6.10). When bearing wear occurs, there is still no change to the LVDT accuracy, but there may be some additional force required from the measured element to drive the motion of the LVDT core.

An LVDT is an infinite-resolution sensor. The only limitations imposed on resolution are due to noise, characteristics of the signal-conditioning electronics, or limitations of the user's signal-receiving circuitry. Higher excitation voltages are used in noisy environments to maintain a high signal/noise ratio. Quantizing error in the receiving electronics may limit resolution due to the analog-to-digital converter that is often incorporated there by the user. Since the LVDT conditioned output is an analog signal, an A/D converter is needed to present the signal to a digital system such as a controller using a microprocessor. Use of a 16-bit A/D converter provides high resolution and reduces this limitation.

A hermetically sealed LVDT is a very rugged sensor. They can be used in high-humidity, high-vibration environments and over a wide temperature range. The metal end plates are typically TIG (tungsten–inert gas) welded to the metal housing. One end includes a hermetically sealed connector. For TIG welding, the end plates are held against a tube that is the housing. The housing is clamped into a collet to hold it firmly while it is rotated by a motor. A tungsten electrode is brought close to the edge where the end plate joins the housing tube. An inert gas, usually argon, flows around the tungsten electrode to eliminate oxidation of the molten metal during the welding process. An electric arc is struck between the tungsten electrode and the joint as the LVDT housing is rotated by a motor. The heat of the arc melts some metal from the end plate and the housing so that a weld bead is formed at the seam. The resulting part is totally sealed against the entry of water or gases. If a stainless steel material is used for the housing and end plates, a substantially corrosion-resistant sensor is produced.

6.10 TYPICAL PERFORMANCE SPECIFICATIONS AND APPLICATIONS

Several styles and housing diameters are standard in the industry. The basic style includes an LVDT housing and a separate core. The core has internal threads by which to attach a nonferromagnetic actuation rod. A second style is the gauge head, which uses a small-diameter coil with the core mounted internally and a plunger attached. The plunger is usually loaded by a spring but can also be supplied with air-actuated return instead of the spring. A third style is the dc LVDT, which has the driving and demodulation electronics contained within the LVDT housing. Here are some typical performance specifications for a basic style sensor:

Input voltage:	1 kHz ac, 5 V rms
Supply current:	5 to 10 mA
Output:	differential AC signal
Nonlinearity:	0.2 to 0.5%, depending on stroke length and type
Resolution:	nearly infinite
Operating temperature range:	–25 to 85°C

The LVDT is a popular choice for high-reliability military applications. They are often used as the position feedback sensor in the actuator for control surfaces in airplanes. LVDT gauge heads are used successfully throughout industry on the production floor as well as in quality assurance. An array of LVDT gauge heads can be assembled to a template of an automobile door, for example. Samples are taken from the production line and placed against the template. Deviations from the standard profile are measured and used to determine if the tool making the door panels needs adjustment or reworking to keep the finished parts within the allowed tolerances.

Incoming inspection tools use LVDTs to measure critical dimensions and read them into a computer that compares the data against the accepted tolerances. The gauge head of Figure 6.10 is a configuration often used in this way. There is a resistance-welding tool on the market that uses an LVDT to measure the submicron motion that occurs in the mating parts as the welding progresses (this measurement is called the weld *set-down*). LVDTs have been used as components in many other products, including pressure transducers and valve positioners for process control.

CHAPTER 7

THE HALL EFFECT

7.1 HALL EFFECT TRANSDUCERS

Position transducers based on the Hall effect are often used in automotive and industrial products because they can provide long life at a relatively low cost. Since the sensitivity of a Hall effect element is based on measuring the magnetic field at a specific point within the device package, a single element provides for a relatively short stroke linear position sensor (less than 25 mm stroke). Longer-stroke-length transducers can be made by using mechanical advantage or by incorporating an array of sensing elements, but the benefit of lower cost is then reduced.

Hall effect sensors measure the strength and polarity of a magnetic field. A Hall effect linear position transducer includes at least a Hall device, a position magnet, and associated electronic circuits. The position magnet is attached to the element to be measured. As the magnet approaches the Hall device, the strength of the magnetic field increases, and the output of the Hall device increases. Since the change in magnetic field strength is due to a change in the magnet position, the Hall device produces an electrical output that varies with changes in the position of the magnet. The polarity of the magnetic field is indicated by the polarity of the electrical output.

Hall effect sensing elements have a relatively low output voltage (tens of millivolts), so fabricating a position transducer requires the addition of an amplifier to increase the signal voltage level. A constant voltage or constant current is applied to drive the sensing element, and a differential output

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

voltage is then produced. The output voltage amplitude is proportional to the magnetic field strength and thus related to the distance between the sensing element and a position magnet. As mentioned earlier, the polarity of the output voltage is dependent on the polarity of the field produced by the position magnet (i.e., with lines of flux extending from the north pole or the south pole of the position magnet). So, depending on the connection configuration, as a magnetic *north* pole approaches a given Hall effect element from one side, the resulting output voltage may be positive whereas a negative output would be produced from a magnetic *south* pole approaching that same Hall effect element from the same side. A permanent magnet has two poles, north and south. The north pole is the one that seeks the magnetic north of the earth. Since opposite poles attract, this means that the magnetic north of the earth is analogous to what we call the south pole of a permanent magnet. This is important to know when checking the polarity of a permanent magnet by using it as a compass needle. The north pole of a magnet is also called the north-seeking pole.

7.2 THE HALL EFFECT

In a Hall effect sensing element, also called a *Hall device*, an electric current is passed through a conductive material while the material is subjected to a magnetic field. Electrical contacts to the conductive material provide the means for applying the current and for measuring the output voltage. The output from a Hall effect sensing element is called the *Hall voltage*, V_H. The direction of V_H is perpendicular to both the electric current and the magnetic field. The magnitude of V_H is proportional to the density of the magnetic flux, β, and to the amount of current, I, flowing through the sensing element (see Figure 7.1).

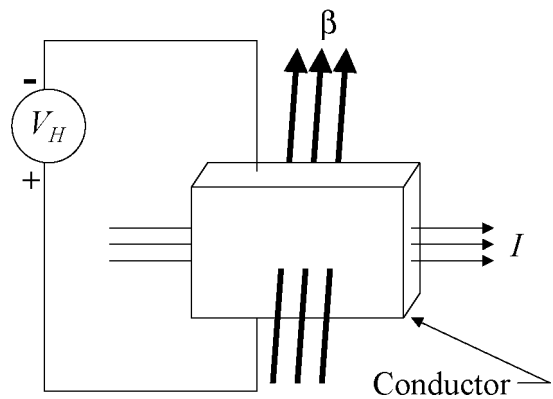

Figure 7.1 Hall effect: mutually perpendicular current, I, and magnetic flux, β, in a conductor result in generation of the Hall voltage, V_H.

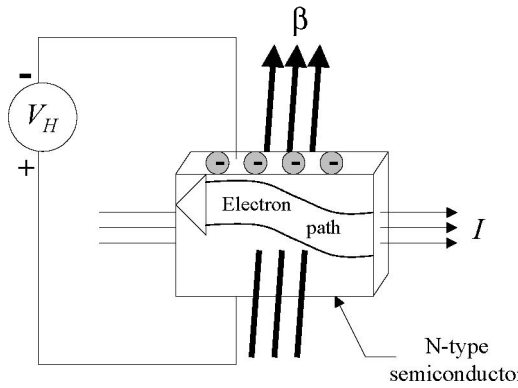

Figure 7.2 Concentration of carriers (–) is forced to one side of the Hall device by magnetic force and produces the voltage differential, V_H.

The magnetic field causes a gradient of carrier concentration to occur across the conductor. The carrier concentration is greater on one side than on the other. The larger number of carriers on one side of the conductor causes the voltage potential, V_H, to develop. In Figure 7.2, electrons being shown as the charge carriers in an n-type semiconductor, the concentration of electrons is greater toward the upper edge of the conductor. The electrons are supplied by the current flow, and they are directed toward the upper edge of the conductor by the force from the magnetic field. (Note that the direction of electron flow is opposite that of current flow, according to the standard convention.) A voltage potential difference is thus developed across the conductor of Figure 7.2, with the upper surface at a negative potential with respect to the lower surface.

The top surface of Figure 7.2 is negative because of the negative charge of the greater number of electrons there. This differential voltage, between the top and bottom of the conductor of Figure 7.2, is the Hall voltage, V_H. The amplitude of V_H varies with the amount of current flow and the magnetic field strength [7, p. 215] according to

$$V_H = \frac{K_H \beta I}{z} \tag{7.1}$$

where V_H is the Hall voltage, K_H the Hall constant, β the magnetic flux density, I the current flowing through the conductor, and z the thickness of the conductor. With commercially available Hall devices, the conductor dimension is already known by the manufacturer, and the sensitivity is specified for the particular model. In this case, formula (7.1) can be simplified to

$$V = K\beta I \tag{7.2}$$

where K is the sensitivity factor, usually in volts per gauss or volts per millitesla (mT), which is specified by the manufacturer.

7.3 HISTORY OF THE HALL EFFECT

In 1879 at Johns Hopkins University, physicist Edwin H. Hall discovered the effect that now bears his name [16, p. 473; 31, p. 1]. He found that a voltage potential appears across a conductor when a magnetic field is applied at right angles to the flow of an electric current in the conductor. This voltage potential is the Hall voltage (see Figure 7.1). The Hall effect has been used to develop sensing elements that measure the intensity and polarity of a magnetic field. These magnetic field sensors have been further developed into position transducers, where change in the position of a magnet attached to a target results in change in the magnetic flux density at the location of the Hall device. (A Hall effect–based sensing element is often called a Hall device.)

Although discovered in the late nineteenth century, the Hall effect had little commercial use until the 1950s when development of semiconductor compounds led to the first useful Hall effect laboratory instruments. The use of a Hall effect sensor allowed the measurement of a static magnetic field, which was not possible with the coil assemblies previously used for magnetic field measurement. Once the Hall effect sensor and some signal conditioning electronics were incorporated into a single integrated circuit in the late 1960s, many applications became practical. Most of the early applications were for switch-type sensors. These included keyboards, joysticks, and other industrial and commercial products.

The presence of an offset voltage and of temperature sensitivity noticeably limited the performance of single-element Hall effect sensors. To obtain a lower offset voltage and lower temperature sensitivity of the offset voltage, dual elements were used as a partial solution. This was followed by the development of quad elements in a bridge configuration. Most present-day Hall devices are of a bridge configuration. The resulting capability for operating over a wider temperature range made automotive sensors and control systems possible. Further integration of the circuitry and use of the bridge configuration improved the performance of switch-type devices and made linear transducers practical. Since the output voltage from a Hall device is very low, chopper-stabilized amplifiers were integrated to remove any offset voltage error that originated in the earlier dc amplifiers. A later method, also used to eliminate offset errors, involves constantly switching the polarity of the input current to the Hall device, thereby producing an ac signal and taking away the drift-prone dc bias by using an ac amplifier. Once it is amplified to a high level, the ac signal is converted back to dc.

After introduction of the bridge configuration Hall device and improved product uniformity achieved through high-volume manufacturing, few of the

performance increases developed thereafter were due to improvements in the Hall device itself. Instead, the integration and increasing complexity of the signal-conditioning circuitry brought most of the advancement in performance.

7.4 HALL EFFECT POSITION TRANSDUCER DESIGN

A simple position transducer can be devised by placing a permanent magnet and a Hall device in close proximity to each other, as shown in Figure 7.3. The output of the Hall device will vary as the distance between it and the position magnet varies, due to the change in magnetic field strength at the Hall device. As depicted in the output versus position curve, the relationship is nonlinear. As the position magnet is relatively far away from the Hall device, the sensitivity is low. As the magnet gets closer to the Hall device, the sensitivity increases steadily. This is due to the fact that the magnetic field strength at the Hall device varies approximately as the square of the separation distance between it and the position magnet. The response of the Hall device itself is essentially linear with respect to the magnetic field strength.

An alternative sensor configuration is shown in Figure 7.4. A longer sensing stroke is possible because of the orientation of the position magnet. As the position magnet moves along its stroke, the Hall device is affected by the north pole of the magnet at one end of the stroke and then by the south pole at the other end of the stroke. This produces a bipolar output voltage characteristic, as shown to the right in Figure 7.4.

An improved position sensor configuration is shown in Figure 7.5. Two magnets are held in place by a bracket. Like poles of the two magnets (south, in this case) face the Hall device, which is fixed in a location between the magnets and along their motion axis. The complete bracket and magnet assembly is moved with the measurand. With the Hall device positioned between

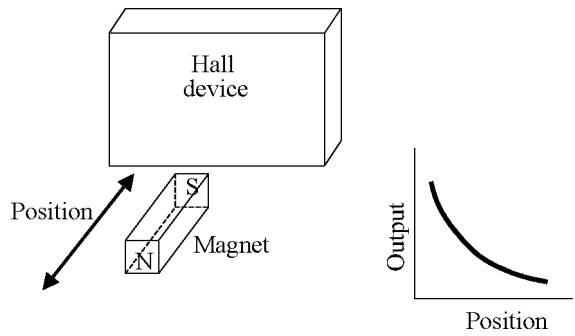

Figure 7.3 Simple position sensor based on a Hall device, together with a typical curve of the output voltage versus position.

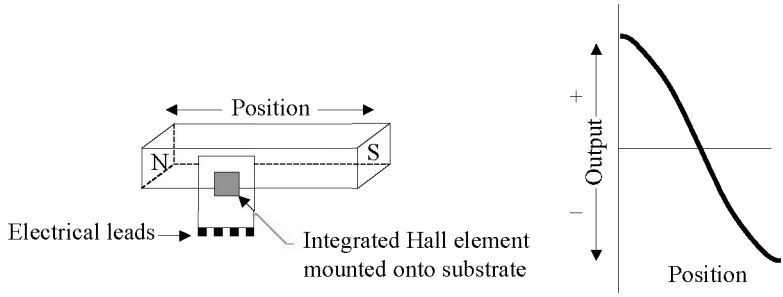

Figure 7.4 Alternative configuration of a Hall effect position sensor with longer stroke and bipolar output voltage.

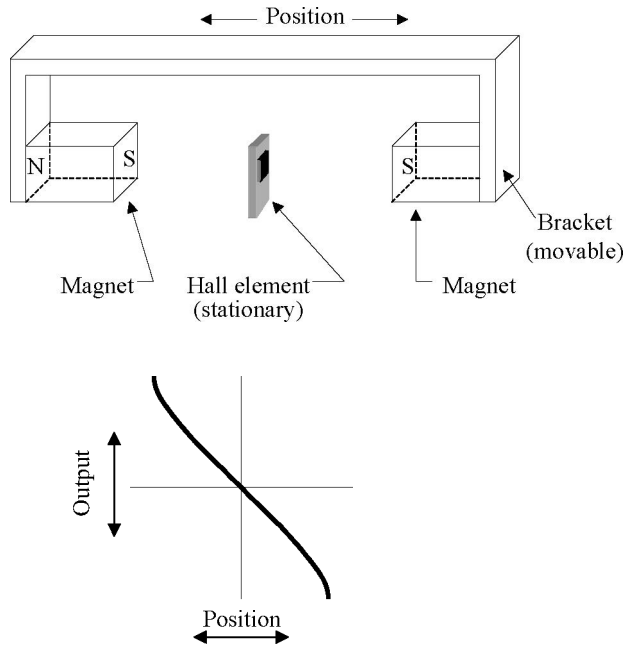

Figure 7.5 Hall effect position sensor with improved configuration for greater stability.

the two magnets, the field strength in the center is near zero even if the magnetic field strength changes with temperature (assuming that the thermal effect is the same on both magnets). As one of the magnets approaches the Hall device, the field strength and output increase.

As the same polarity of the other magnet approaches the Hall device from the opposite side, the direction of the output voltage is reversed. The bipolar output voltage characteristic shown in the curve is fairly linear if the total sep-

aration of the two magnets is not too great. This sensor configuration is suitable for short position transducers in the range ±1 cm or less because the field strength becomes too low at longer distances, increasing the nonlinearity. At one end of the measuring range (the full measuring range is also called the *stroke*), the magnetic field from the south pole of one position magnet penetrates the Hall device from that side. At the other end of the stroke, the magnetic field from the south pole of the other magnet penetrates the Hall device from the other side. These two directions of the magnetic field are what results in the bipolar output signal from the Hall device. That is, the output voltage goes from negative to positive, as shown in the figure. So, one way to obtain a bipolar output is to subject one side of a Hall device to the field from a magnetic *north pole* and then to a *south pole* (Figure 7.4). Another way to obtain a bipolar output is first to subject one side of a Hall device to the field from a magnetic south pole, and then to subject the other side of the Hall device to a south pole (Figure 7.5).

7.5 HALL EFFECT ELEMENT

The Hall device, or *Hall element*, is made from a thin sheet of conductive material. Input connections for the introduction of current flow are positioned perpendicular to the output connections, which are for detecting the Hall voltage. When in the presence of a magnetic field, the Hall voltage is proportional to the strength of the applied field. The polarity of the Hall voltage (+ or –) is determined by the polarity of the applied magnetic field (north or south).

The Hall constant, K_H, is larger in semiconductors than in metals [30, p. 79]. Because a larger Hall constant results in greater sensitivity, Hall devices normally utilize semiconductor materials in their construction. The semiconductor material may be either *p* or *n* type, depending on the polarity of the charge carriers (holes or electrons, respectively). A large Hall constant requires high carrier mobility. Semiconductor materials provide high carrier mobility after doping with an impurity selected to provide carriers of electrons or holes. A low resistance value will help to limit thermal noise and allows a higher signal/noise ratio. Since electrons tend to move faster than holes under a given set of conditions, greater sensitivity can be obtained by using *n-type* semiconductor material [31, p. 8].

When a Hall device is manufactured, it is important to accurately space the two contacts for the Hall voltage pickup. If the two contacts are not perfectly aligned, there will be a nonzero output when there is a zero magnetic field strength. This is due to the differential in the voltage drop resulting from the current flow in unequal resistances over the two sides of the element. Since perfect alignment is not likely, a small voltage at zero magnetic field strength will usually be found. This is the zero offset voltage. An adjustment can be included in the signal conditioning electronics to correct the offset voltage to zero.

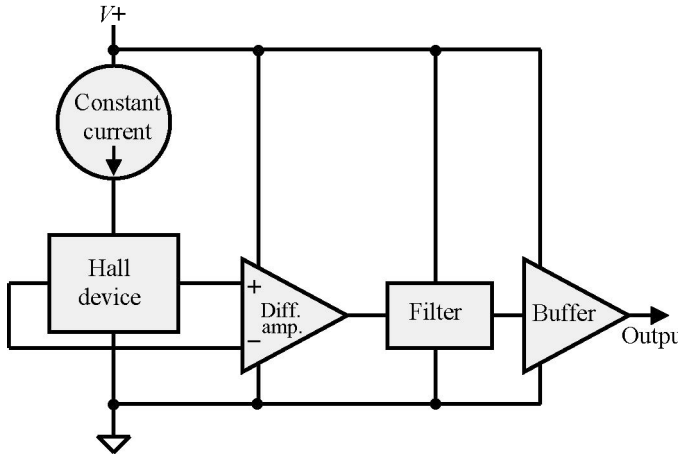

Figure 7.6 Block diagram of an analog circuit for a single Hall device or a bridge configuration of Hall devices.

7.6 ELECTRONICS

When using a single Hall device (containing either a single element or a bridge connection), a circuit similar to the block diagram shown in Figure 7.6 can be used. Since the Hall voltage is proportional to the current through the Hall device, a constant-current source is shown as supplying the drive current. Sometimes a constant-voltage source is used instead. Alternatively, if the Hall device is powered directly from the power source with no current or voltage regulator, the output will be ratiometric. Ratiometric operation, in which the output voltage span is proportional to the supply voltage, can allow lower cost by eliminating a voltage reference while improving performance by the amount of error that would have been added by a voltage reference (see Section 2.13).

The Hall voltage is a small differential voltage that develops across the Hall device and is then amplified by a differential amplifier. This amplifier also converts the differential voltage to a single-ended voltage, which means that it is referenced to zero volts. Next, a filter passes on the expected range of signal frequencies and rejects (does not pass) other frequencies. This function would normally require a low-pass filter, which passes signals of low frequency and attenuates those of higher frequencies [15, p. 6]. The rejected frequencies are not wanted because they are due to electrical or electromagnetic noise. If an analog output from the transducer is desired, a buffer amplifier conditions the signal to drive the output. If a digital output is needed, an A/D converter would be used instead of the buffer amplifier. A digital implementation is shown in Figure 7.7. Note that, even in a digital circuit with a microcontroller, some analog circuitry is still required. A filter is not shown in the digital version,

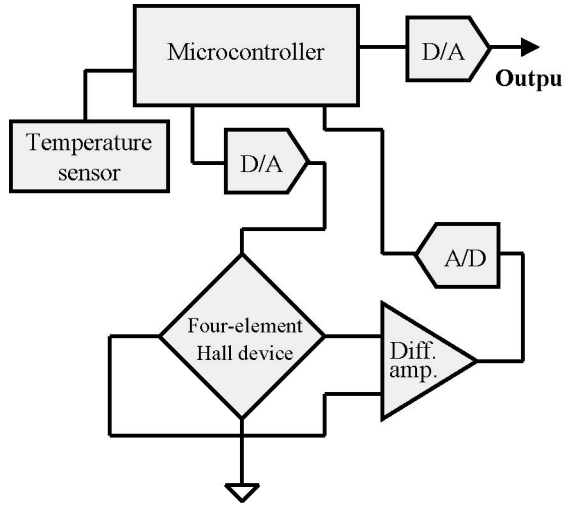

Figure 7.7 Block diagram of a Hall device digital circuit with a microcontroller.

which uses a microcontroller (μC), because a digital filter algorithm can be included in the μC code, although an antialiasing analog filter in hardware may still be needed in some cases.

In Figure 7.7, the microcontroller powers the Hall device through a digital-to-analog (D/A) converter. The D/A converter can be a voltage or a current output type. After the differential amplifier, an A/D converter is used to bring the Hall device signal into the μC. (These additional conversion circuits are typically required in order to use a μC with a purely analog sensing element, like a Hall device. A much easier interface to a μC can be had with a time-variant type of sensor such as a magnetostrictive linear position sensor.) After the μC code conditions the Hall device digitized signal properly, an analog output is produced by a second D/A converter. Alternatively, if a digital output is desired, the μC can provide a parallel or serial output without needing the second D/A converter. If temperature compensation is desired, the μC should be provided with a temperature input as shown in Figure 7.7. The temperature sensor can be implemented with a pulse-width-modulated output that can be measured by the micro with a timer. Also, temperature sensors are available that have a direct serial digital output. The temperature sensor should be mounted very close to the Hall device, or integrated into the Hall device chip, to provide an accurate reading of the Hall device temperature. If the temperature sensor is not close enough to the Hall element, thermal gradients across the transducer (due to power dissipation or to variation in ambient temperature) will cause error in the temperature indicated.

The output from the Hall device will have errors. Some of these include zero offset, gain error, offset voltage change with temperature, gain factor change with temperature, nonlinearity, and hysteresis. Potentiometers are

added to the amplifier circuit to provide the means to adjust the room-temperature zero and gain errors. These controls are called zero and span adjustments. Alternatively, this can be done in software if a microcontroller is used.

Compensation for temperature-induced errors is a little more difficult. The change in gain error with temperature is often similar for Hall devices made on the same production line. Therefore, one compensation can often be used for all transducers if highly accurate compensation is not needed. The temperature sensitivity of zero, however, can vary substantially from one transducer to the next. For the highest accuracy in laboratory equipment, each Hall device, or the complete transducer, must be characterized by running throughout the operating temperature range in an environmental chamber. The data are recorded and used to calculate the correction. The correction can be accomplished in analog circuitry by using temperature-compensating resistors, or in software if using a microcontroller.

Using four Hall elements together in a Wheatstone bridge configuration helps to compensate for thermal errors and increases the measurement sensitivity. The Wheatstone bridge configuration is usually incorporated into one substrate and used as if it were a single Hall device.

7.7 LINEAR ARRAYS

It was mentioned earlier that a single Hall device, or Wheatstone bridge connection of four devices, is only operable over a range of up to a maximum of 25 mm. Alternatively, multiple Hall devices can be aligned along a linear axis to fabricate a sensing element for a longer-stroke linear position transducer, as shown in Figure 7.8. In this configuration, an array of several or tens of Hall devices is positioned along a stationary measurement scale. A position magnet (permanent magnet) is moved along the array and represents the position to be measured. The one or two Hall devices in proximity to the magnet will indicate the magnetic field strength by the voltages of their outputs. Output connections from all the Hall devices are brought back to a signal processor, usually located at one end of the Hall device array. Multiplexing the Hall

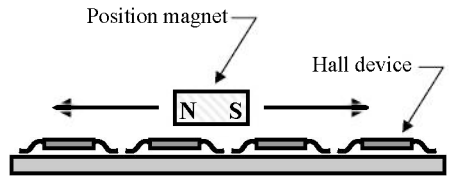

Figure 7.8 Expanded range linear position transducer. A number of Hall devices may be deployed in a linear array. The output signal is derived from a combination of the Hall device locations and their individual outputs.

device connection lines and scanning the signal amplitudes will indicate which Hall devices are close to the position magnet. The Hall device that is closest to the position magnet is first detected, for example, by way of determining which Hall device has the highest output. The known location of that particular device represents a *coarse position*. The actual reading from that device represents a *fine position*. The electronic circuit combines the coarse and fine position data to form a high-resolution output signal over a wider range. Position transducers with full-scale ranges of up to several meters can be made using this technique.

7.8 ADVANTAGES

Most Hall devices are made from semiconductor material. This means that they have the advantage to be constructed so that some or all of the signal-conditioning circuitry required can be incorporated on the same semiconductor substrate that comprises the sensor element. This can reduce size and cost. Hall effect position transducers have an advantage over contact position transducers such as a potentiometer. Since the Hall device is non-contact, there are no wearing parts to limit the life of the transducer. Hall effect–based position transducers also are normally absolute reading: There is no need for re-zeroing after a glitch or power cycling as would be required with an encoder.

Hall devices can operate at zero speed, this being an advantage over some encoders that have an uncertain output when approaching the edge of a transition between two adjacent readings when the approach speed is near zero. An encoder with this characteristic can have a constant dithering of the output signal. A Hall effect device, however, has an essentially infinite resolution and will indicate even small changes in position with good linearity and a monotonic characteristic. Besides zero speed, some Hall devices can operate at frequencies of over 100 kHz. Hall devices have a repeatable response and can operate over a broad temperature range (as wide as –40 to 150°C).

Nonlinearity and other accuracy limitations of Hall effect position transducers are about average within the range of available position transducers, being about the same as LVDTs but not as accurate as a magnetostrictive position transducer (e.g., Temposonics). Some position transducers, Such as an LVDT or some magnetostrictive position transducers, can be designed such that the basic sensor element can be separated from the electronics module by a length of cable. This allows the size of the sensor element to be minimized as well as to be located in a higher-temperature environment, where it would be difficult to use low-cost versions of electronics. This is not a possibility with Hall devices because the Hall device itself is made of semiconductor material, thus limiting the maximum temperature. Also, the signal-conditioning circuitry that is integrated onto the monolith including the Hall device (to maximize its performance and keep the cost within limits) has temperature limitations because the semiconductor material is formulated to enhance the Hall device

performance. High-temperature electronic circuits require different formulations of semiconductor materials and different fabrication techniques from those best suited for manufacturing a Hall device.

7.9 TYPICAL PERFORMANCE SPECIFICATIONS AND APPLICATIONS

A typical Hall effect transducer has a performance specification similar to this one from Diltronic/Midori:

Range:	20 mm
Power supply voltage:	5 V dc
Sensitivity:	200 mV/mm
Nonlinearity:	±2.0%
Repeatability:	±0.03 mm
Hysteresis:	±0.09 mm
Case size:	22.0 mm × 63.0 mm

A Hall device can be used as a component of a position transducer or as the position-sensing component of another type of transducer. Another type of transducer that uses a Hall effect–sensing element is a pressure transducer, where a metal diaphragm flexes in response to changes in pressure. A Hall device would measure the deflection of the diaphragm, and then after the signal is conditioned, the transducer would provide an electrical output which indicates the pressure (see Figure 7.9). The Hall device in this case, together with the position magnet and diaphragm, comprise the sensor or sensing element. A transducer is formed by the addition of the signal conditioning electronics, power supply, housing, and so on.

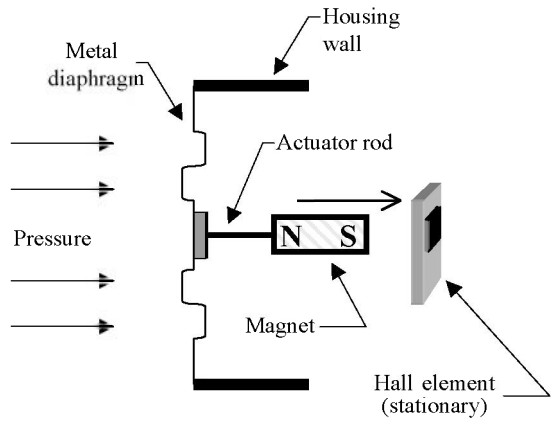

Figure 7.9 Pressure transducer configuration with a Hall device to measure the diaphragm position as a function of the applied pressure.

Some additional applications of Hall effect linear position sensors include: in the sending unit for measuring level in a fuel tank (used with a float containing the position magnet), measuring pedal position in a car, noncontact joystick, measuring liquid level over several meters using a linear array of Hall devices, and many others.

CHAPTER 8

MAGNETORESISTIVE SENSING

8.1 MAGNETORESISTIVE TRANSDUCERS

A magnetoresistive sensing element provides a changing electrical resistance when subjected to an external magnetic field. Generally, the resistance of nonmagnetic conductors increases when a field is applied but is nonlinear [14, p. 56]. This change is in response to the magnitude of the strength of the magnetic field (see Figure 8.1). As in Hall effect transducers, magnetoresistive linear position transducers are noncontact, absolute reading, and have essentially infinite resolution. The measuring range of an individual element is limited to a maximum of about 25 mm, but practical linear sensors have been designed by using multiple sensing elements in a linear array to obtain transducers with a full-scale range (FSR) of over 2 m. These transducers can have a nonlinearity of less than 0.05%. Popular applications include positioning of industrial machinery and measuring of small displacements in commercial products and industrial control systems.

The coupling between the moving and stationary parts of the transducer is achieved by means of a magnetic field. Therefore, any number of position changes can be made without incurring wear to the transducer parts. In a single-element magnetoresistive linear position transducer, a position magnet moves in relation to the sensor. This permanent magnet provides a magnetic field intensity that changes with the distance to the sensor. As the magnetic field strength increases perpendicular to the current flow, the resistivity

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

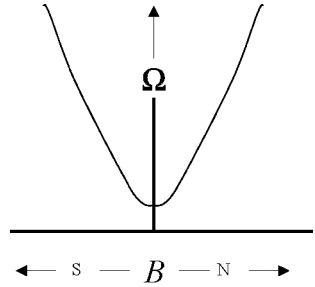

Figure 8.1 The resistance of a nonmagnetic conductor (in ohms), varies in response to the magnetic field strength, B, but not the polarity (north or south).

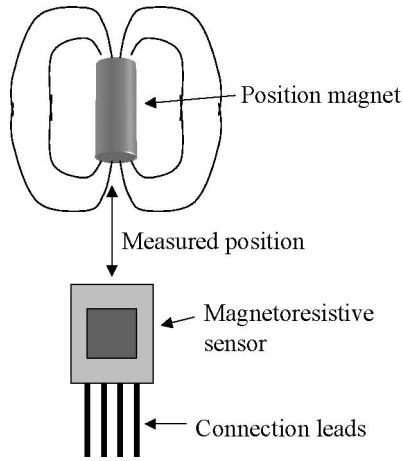

Figure 8.2 Magnetoresistive sensor with position magnet.

increases. A simple arrangement of a magnetoresistor with position magnet is shown in Figure 8.2. The resistance of the sensing element responds to the magnitude of the magnetic field, not to its polarity.

8.2 MAGNETORESISTANCE

Magnetoresistance is the change in resistance of a current-carrying conductor when a magnetic field is applied [9, p. 284]. A two-terminal device having this characteristic is called a *magnetoresistor*. Although the resistance of nonmagnetic conductors generally increases when a magnetic field is applied, in most magnetic materials, electrical resistance decreases with the increase of magnetic field strength when the magnetic field direction is perpendicular to the current flow through the magnetoresistor (see Figure 8.3).

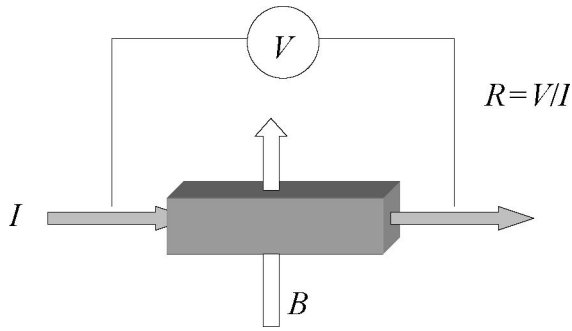

Figure 8.3 The magnetic field, B, is at a right angle to the current flow, I.

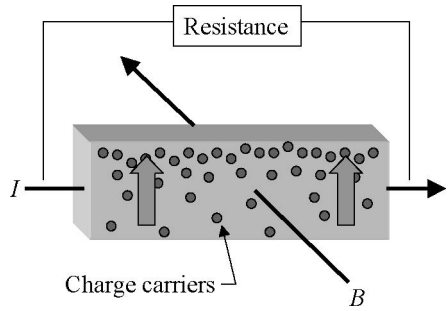

Figure 8.4 In a conductor, charge carriers are moved to one side by a magnetic field, B.

Almost all conductors exhibit some degree of magnetoresistance. The physical basis underlying the property of magnetoresistivity in a conductor is the Lorentz force. This force causes the charge carriers, electrons that are carrying the current, to move in curved paths [12, p. 449]. A Lorentz force is a force that acts on a moving charge in the presence of a magnetic field. The force is at right angles to the magnetic field vector as well as to the velocity vector of the moving charge. When a magnetic field is applied as shown in Figure 8.4, the charge carriers are aligned with the magnetic field.

More of the charge carriers are forced to one side (the top of the figure), leaving fewer carriers on the other side (the bottom). The magnetoresistor then undergoes an increase in resistance due to the availability of fewer charge carriers when the magnetic field is applied. This is called *galvanomagnetoresistance*. The change in resistance comes from the combination of two component parts. These are a reduction in forward carrier velocity as a result of the carriers being forced to move sideways as well as forward, as described above, and a reduction in the effective cross-sectional area of the conductor

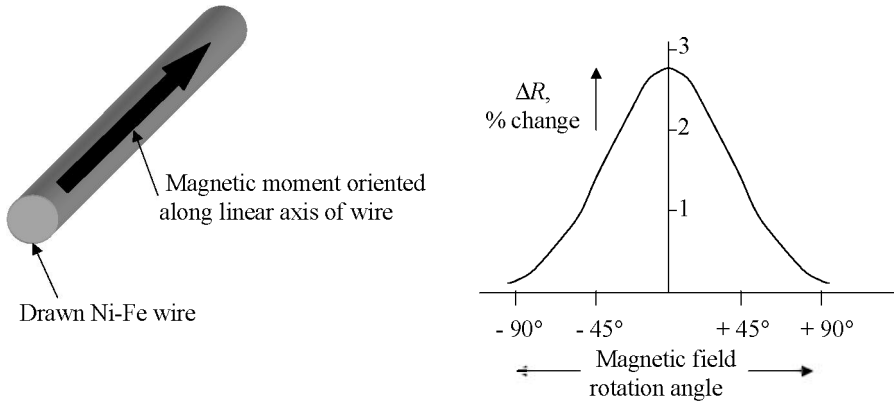

Figure 8.5 Magnetic domains tend to align in one direction when a ferromagnetic wire is drawn.

as a result of the carriers being crowded to one side [5, p. 293]. The resistance changes according to

$$\text{resistivity} = \frac{\text{voltage}}{\text{carrier density} \times \text{carrier velocity}} \tag{8.1}$$

The characteristic above applies to conductors, but magnetoresistance is especially large in ferromagnetic materials, particularly in nickel–iron alloys known as Permalloys, due to the magnetic properties of the material. Before film technologies were used, it was found that a drawn nickel–iron wire had the property of magnetoresistance. When the wire is drawn, the magnetic domains tend to align parallel to the linear axis of the wire, shown in Figure 8.5.

During deformation, the grains rotate as well as elongate, causing certain crystallographic directions and planes to become aligned. Consequently, preferred orientations, or textures, are developed which cause anisotropic behavior [1, p. 189]. This gives the wire an orientation of magnetization along this axis, called the *easy axis*. Changing the magnetic axis to be perpendicular to the current flow (the hard axis), by bringing a permanent magnet in close proximity causes an increase in the electrical resistance. This is called the *anisotropic magnetoresistive* (AMR) *effect* (see Figure 8.6).

A material is anisotropic if it has a predictable variation of a property (sensitivity to a magnetic field, in this case), depending on the direction in which it is measured. The opposite condition is called *isotropic*, which is *directional uniformity*. The charge carriers in a metal will be electrons that are free to move within the conductor. The resistance between the two ends is relative to the bulk resistance of the material per cross section times the length. This is assuming that nearly all the material is participating in the movement of electrons.

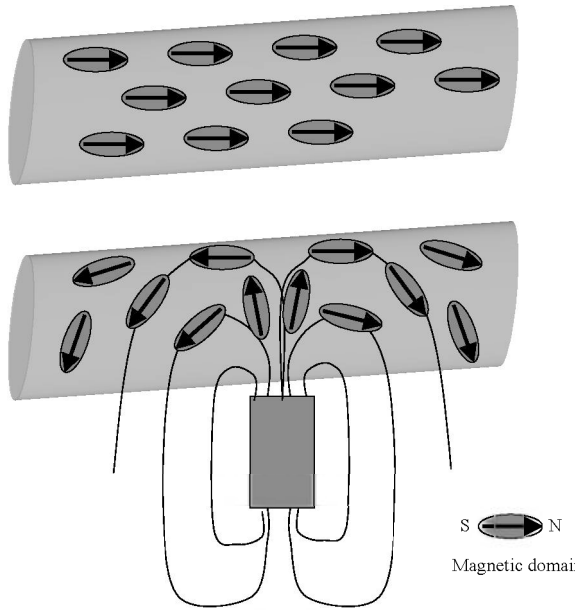

Figure 8.6 Anisotropic magnetoresistive effect. Magnetic domains align along the linear axis of a drawn Ni–Fe wire. When a magnetic field is applied, B, this alignment is changed, and the resistance increases.

The amount of change in the resistance in magnetoresistors varies approximately in proportion to the square of the magnetic field strength. The degree of magnetoresistance of a material for a given change in magnetic field strength is called the *magnetoresistive ratio* ($R_{\text{magnetoresistive}}$). It is the ratio of the change in resistance (due to the magnetic field) to the original resistance without the magnetic field:

$$R_{\text{magnetoresistive}} = \frac{R_{\max} - R_{\min}}{R_{\min}} \tag{8.2}$$

where $R_{\min}$ is the resistance before application of the magnetic field and $R_{\max}$ is the resistance with the full magnetic field strength applied.

If a material is able to maintain a given magnetoresistive sensitivity over a wider range of magnetic field strength, the magnetoresistive ratio that is possible will be higher. A magnetoresistive material of lesser capability may saturate at the higher magnetic field strengths and limit its operating range of magnetic field strength that can be used. Magnetoresistance occurs in pure ferromagnetic metals to a small degree, in the range of a 1% change in resistance. (The term *ferromagnetism* comes from the early association of the phenomenon with ferrous, or iron-containing, materials [33, p. 498]).

Nickel–iron alloys can have greater magnetoresistance than pure metals and can have a change in resistance in the range 2 to 5%.

A drawn nickel–iron wire was presented earlier, but a more practical magnetoresistor can be made by depositing a thin strip of permalloy onto a substrate while in the presence of a magnetic field. Most sensing elements are fabricated as a Permalloy thin film because the anisotropy can be made essentially uniaxial. Applying the magnetic field during deposition aligns the magnetic domains in the Permalloy along the desired axis. The resistance of the AMR material is at its lowest when the current and the magnetic moment are parallel (the current is parallel to the easy axis). The ratio of the change in resistance, ΔR, to the total resistance, R, is approximately proportional to two times the cosine of the angle, α, between them.

$$\frac{\Delta R}{R} \propto 2\cos \propto \tag{8.3}$$

The angle is changed by bringing a permanent magnet close to the AMR material. The resistance is then highest when the angle between the current and the magnetic moment is 90°.

Because of this relationship, the best linear region is in the area of 45° of rotation of the magnetic moment. To extend the linear range to lower levels of magnetic field strength, a special configuration of conductors has been used. Small strips of a nonmagnetic conductor such as aluminum have been placed at an angle of 45° to the long axis of an AMR material (see Figure 8.7). This is called a *barber pole configuration*. These strips have a lower resistance than the base AMR material and cause the current flow to approximate a 45° angle to the magnetization easy axis when no external magnetic field is applied. In addition to improving the linear response to low-intensity magnetic fields,

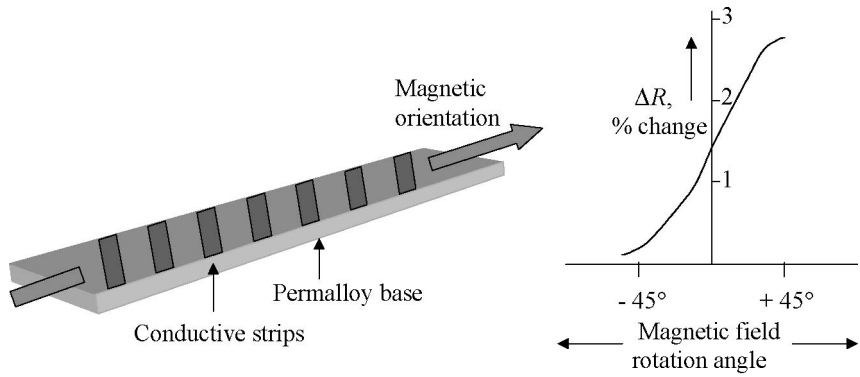

Figure 8.7 Placing conductive strips at 45° to improve linear response when sensing low-magnitude magnetic fields.

this also enables the determination of the polarity of the sensed field. The polarity is indicated by an increase or decrease in resistance, corresponding to an increase or decrease respectively in the 45° angle. One magnet pole causes an increase in the angle, the other causes a decrease. A consequence is that the conductors reduce the active part of the element by shielding it and reduce the magnitude of the sensor signal. As described earlier, the sensitivity of an AMR material is based on a preferred orientation of the magnetic moment. This orientation can be induced by mechanical means, such as drawing a wire, or by applying a magnetic field to orient the magnetic domains during fabrication. One drawback of magnetoresistive materials with oriented domains is that it is possible to change this orientation accidentally. If an AMR sensor is exposed to a strong magnetic field of an orientation different from the original, the MR material may be magnetized in the new direction, drastically changing the performance of the sensor. For this reason, it is important to prevent exposure to excessively strong magnetic fields.

Simple AMR construction has been used in the fabrication of many magnetically based sensors. The use of magnetoresistance sensing elements accelerated, however, with the development of special materials with a much greater percentage of resistance change for a given level of magnetic field strength. The property of these more sensitive materials is called *giant magnetoresistance* (GMR). Alternately layering ferromagnetic alloys with other conductors into one structure increases the sensitivity (see Figure 8.8).

Typical GMR structures have a magnetoresistance ratio of 5 to 20%, depending on the combination of layered materials and their construction. Further specialized materials designed for magnetoresistive properties at low temperatures can have a magnetoresistance ratio of 70% or more while still in a usable proportional range of resistance versus magnetic field strength. They are made by using a combination of metal oxides and semiconductor materials. This substantially greater sensitivity is called *colossal* magnetoresistance (CMR). The temperature sensitivity of such devices has, however, made them unsuitable for practical application so far.

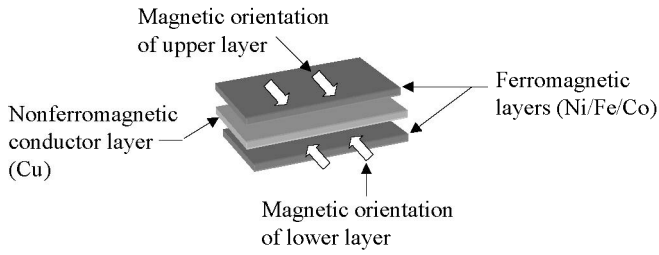

Figure 8.8 A GMR sensing element comprises alternate layers of ferromagnetic and nonferromagnetic conductors sandwiched together.

8.3 HISTORY OF MAGNETORESISTIVE SENSORS

The phenomenon of magnetoresistance was first observed by Lord Kelvin in the materials iron and nickel in 1856 [4, p. 875; 29, p. 109]. He observed the resistance change of a wire while passing a current through the wire and applying a magnetic field with a permanent magnet. Since the discovery of the property of giant magnetostriction most activity in magnetoresistive sensors and transducers have used GMR over the earlier simple anisotropic ferromagnetic materials. GMR was first observed in 1988 [22] in multilayered structures of thin sheets of magnetic and nonmagnetic conductors sandwiched together (as shown in Figure 8.8).

Early GMR sensors consisted of a section of this layered material with electrodes attached and packaging added. A position magnet was allowed to pass in proximity to the GMR element and to cause changes in the resistance of one or more legs of a Wheatstone bridge circuit a connection of four elements as shown in Figure 8.9. When the resistance of one leg, or two diagonal legs, increases (or decreases), the differential output voltage changes to a degree proportional to the resistance change. Two diagonal legs are allowed to respond to the sensed magnetic field. The other two legs are shielded from the magnetic field. This arrangement reduces temperature sensitivity because, although the individual magnetoresistors have a strong temperature dependence, they all respond similarly. The net result is that temperature changes have a much-reduced effect on both zero and span.

Later GMR sensors incorporated the sensing element into integrated circuits. This has the advantage of allowing for some conditioning as well as providing a means of addressing multiple elements. One such device with a linear array of sensing elements is shown in Figure 8.10. The basic magnetoresistive sensing elements, packaged and characterized, are available from several sources worldwide including NVE Corporation, Philips Semiconductor, Siemens, and Sony.

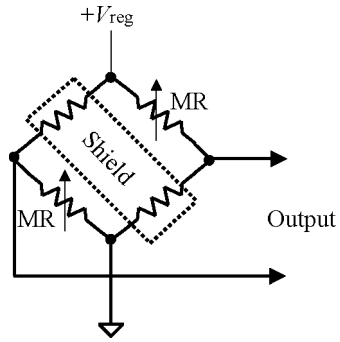

Figure 8.9 Set of magnetoresistors connected in a Wheatstone bridge configuration.

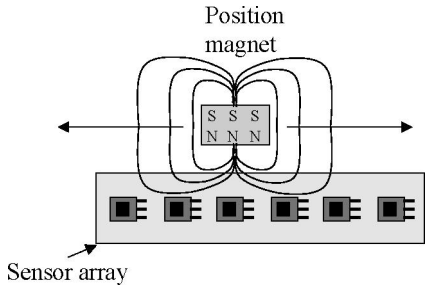

Figure 8.10 A longer-range linear position transducer can be configured by utilizing a linear array of magnetoresistors that sense the position magnet located location along the sensing axis.

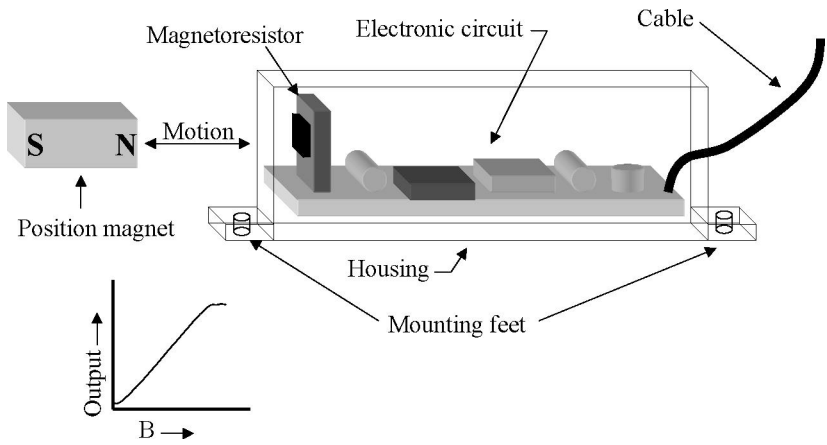

Figure 8.11 Magnetoresistive linear position transducer, with sensing element, housing, mounting feet, and cable.

8.4 MAGNETORESISTIVE POSITION TRANSDUCER DESIGN

A magnetoresistive linear position transducer comprises a sensing element, together with a housing, electronic circuit, and a separate permanent magnet. The permanent magnet is attached to the movable workpiece member of which the position is to be determined. The housing contains the sensing element and electronic circuits, as well as providing the means for mechanical mounting of the transducer to the stationary part of the workpiece (see Figure 8.11).

The electronic circuit provides the regulated power supply, resistance measuring circuit, and signal conditioning necessary to provide the desired output type. The housing is typically made from aluminum so that it does not affect the magnetic field experienced by the sensing element. The position magnet is sometimes enclosed within an aluminum housing itself to provide shielding

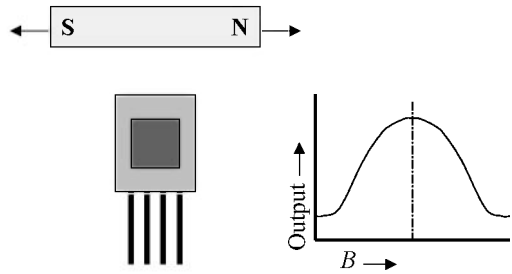

Figure 8.12 Configuration with a longer position magnet provides a longer measuring range.

as well as to provide a means for mounting it to the movable part of the workpiece. The position magnet can be configured to approach the sensing element along the axis of the poles, as in Figure 8.11, or can be a longer shape that operates with the sensing element positioned between the poles, as in Figure 8.12. Allowing the use of both of the magnet poles increases the total measuring range of the sensing element.

8.5 MAGNETORESISTIVE ELEMENT

A current-carrying conductor is normally surrounded by a magnetic field. As presented earlier, the change in resistance of a magnetoresistor is caused by the rotation of the magnetic field by the introduction of an external permanent magnet (the position magnet). In a position transducer the field is rotated by an angle of 90° because the position magnet is arranged to approach the element at a right angle. The change in resistance is approximately equal to the square of the magnetic field strength, until the ferromagnetic material in the sensor approaches magnetic saturation.

An MR or GMR sensing element is formed by depositing the MR materials or layers of material onto a semiconductor substrate. This is usually done in a meander pattern to increase the length of the element and thereby increase the sensitivity. Four elements are normally placed on one substrate and arranged in a Wheatstone bridge connection (see Figure 8.13). Two of the diagonal elements are used for sensing the magnetic field. The remaining two legs are shielded from the effects of a magnetic field as in the schematic of Figure 8.9. This configuration increases the sensitivity to a magnetic field and decreases the temperature sensitivity (compared to using a single MR element).

8.6 LINEAR ARRAYS

Signals from a linear array of individual magnetoresistors can be multiplexed into a microcontroller in order to expand the measuring range to the length

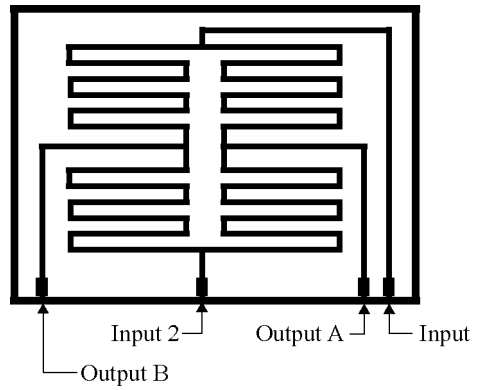

Figure 8.13 Wheatstone bridge connection pattern of four MR legs to form one sensing element.

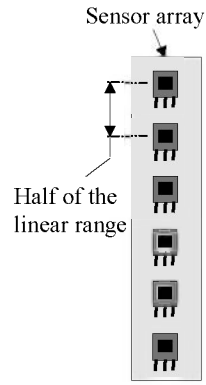

Figure 8.14 The sensors of an array should be spaced at one-half of their linear range.

of the sensor array. Each of the magnetoresistors is actually a Wheatstone bridge connection of four elements. The spacing between the magnetoresistive sensors should be approximately equal to half of their linear range (see Figure 8.14). A microcontroller analyzes the sensor signals to derive information that indicates which sensor is in the closest proximity to the position magnet, as well as to measure the magnetic field strength indicated by that particular element in the array. The combination of these two pieces of information forms the complete datum. A linear approximation between adjacent sensor readings is used to find the exact position of the magnet. The microcontroller strobes the multiplexer to sample the various sensing elements in the array, retrieve their data, and interprets the results into the desired transducer output.

8.7 ELECTRONICS

In a single-point position transducer (with a correspondingly limited measuring range) the Wheatstone bridge connection of four MR elements is wired to a regulated voltage source and a differential amplifier as shown in Figure 8.15. The amplifier output is fed to a filter to remove noise that is not within the expected frequency response characteristic of the transducer. After the filter, the position signal is calibrated and buffered in an output circuit that provides the desired type of output voltage, current, or digital signal to the end user.

In a longer position transducer that uses an array of MR sensors, it is required that a microcontroller be used to effectively derive the required mathematical interpretation of the signals from each of the sensors (see Figure 8.16). The signals from the sensor array are brought into a multiplexing circuit (mux) either directly or by using a bus structure. The switching of the mux is directed by the microcontroller. The signal selected is then amplified by the differential amplifier, filtered, and then converted to a digital representation by the analog-to-digital converter. With the digitized signals, the microcontroller can then interpret the correct reading and output a digital position

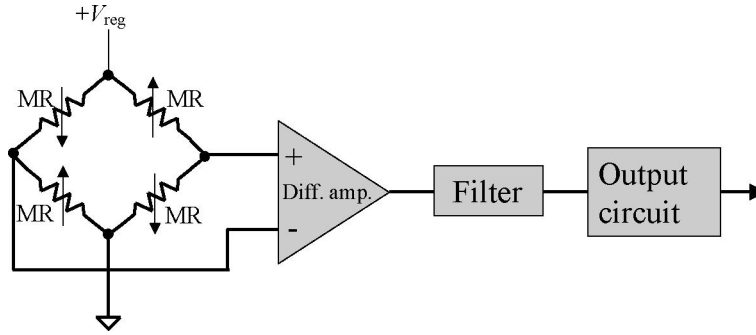

Figure 8.15 Single Wheatstone bridge sensor position transducer with analog output.

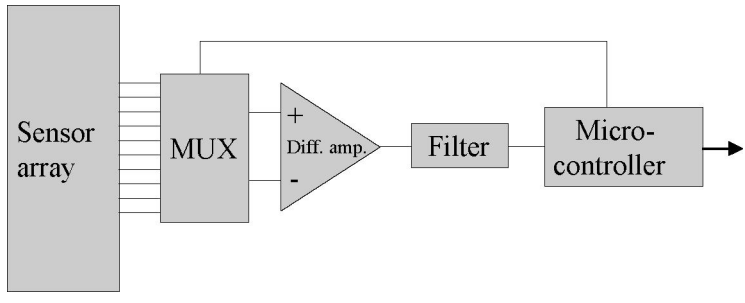

Figure 8.16 Linear array with microcontroller and digital or analog output.

directly, or can provide an analog output by means of a digital-to-analog converter.

8.8 ADVANTAGES

Magnetoresistors have an advantage over inductive sensors in that they can operate in a linear position transducer without a modulator. Compared to a Hall device, a magnetoresistor can generally operate at a higher frequency (or rate of change of position) before attenuation occurs. Magnetoresistors can operate over a wide temperature range (e.g., –55 to 200°C). A GMR magnetoresistor generally has greater sensitivity to a magnetic field than does a Hall device, so a less powerful magnet can be used, or a stronger magnet can be used at a greater distance.

The application of a magnetoresistive sensing element is generally similar to that of a Hall device, the primary difference being that a Hall device produces a change in voltage due to a change in magnetic field strength, whereas a magnetoresistor produces a change in resistance under similar conditions. Other differences include the sensitivity of a magnetoresistor to the magnitude of the magnetic field strength versus the Hall device being also sensitive to the magnetic field polarity. A magnetoresistor typically requires a lower amount of current to power it, since a Hall device needs a driving current to produce the Hall voltage.

8.9 TYPICAL PERFORMANCE SPECIFICATIONS AND APPLICATIONS

When using a magnetoresistive position transducer to monitor the motion of an industrial machine, for example, the position magnet is a permanent magnet that is fitted to the moving part of the machine. The distance to the position magnet is read by a magnetoresistive sensing element mounted along and parallel to the motion axis of the machine. The position magnet does not need to touch the sensing element, because the coupling is through the magnetic field. As the position magnet moves along parallel to the sensing element, the distance between the position magnet and the sensor electronics (head) end of the sensor is read electronically.

Although measuring ranges longer than 25 mm can be handled by using an array of MR sensors and a microcontroller, even longer ranges are possible by using a coded assembly of position magnets. Overall ranges of hundreds of meters can be outfitted with position magnet assemblies, each having its own code based on spacing of the individual magnets in the assembly. An MR sensor array is the movable part and is long enough to always read at least two of the coded magnet assemblies. The location of a particular coded magnet assembly being read represents a coarse reading, and the individual reading

Figure 8.17 C-Pomux transducer for long-range absolute measuring. (Courtesy of Stegmann, Inc.)

from that magnet assembly represents the fine reading. The combination of these two readings is the absolute position from the reference point. An example is the C-Pomux transducer of Figure 8.17.

The specification of this type of sensor depends on the measuring range, but here is a representative example:

Resolution:	100 μm
Repeatability:	±300 μm
Measurement accuracy:	±1000 μm
Moving mass:	3 kg
Operating temperature:	0 to 60°C
Maximum speed:	6.6 m/s
Sampling interval:	0.8 s
Supply voltage:	10 to 32 V dc
Current consumption:	0.3 A
Output:	24-bit SSI

Applications for a long-range transducer can include elevator shafts, where the elevator car position is monitored relative to the floor designation; inventory control systems, in which the location of the inventory retrieval robot is measured with respect to the inventory storage shelf; and directing the movement of mules, where materials are delivered to workstations on a production floor. Applications for short-range, single-sensor element transducers, and arrays up to 1 m in length include LVDT replacements, valve positioners, injection molding and other machine control, and pedal position for onboard automotive use.

CHAPTER 9

MAGNETOSTRICTIVE SENSING

9.1 MAGNETOSTRICTIVE TRANSDUCERS

Magnetostrictive position transducers are noncontact, absolute reading, and have essentially infinite resolution. Linear transducers are commercially available with a nonlinearity of less than .01% and full-scale ranges from less than 10 mm to over 20 m [24]. Curved and rotary sensors are possible. Popular applications include industrial machinery, such as injection molding machines and hydraulic cylinders, as well as automotive and commercial products.

The coupling between the moving and stationary parts of the transducer is achieved by means of a magnetic field. Therefore, any number of position changes can be made without incurring wear to the transducer parts. When using a magnetostrictive position sensor to monitor the motion of a machine tool, for example, a permanent magnet is fitted to the toolholder. The permanent magnet, called the *position magnet*, is read by a sensing element mounted along and parallel to the motion axis of the tool. The position magnet does not touch the sensing element. As the position magnet moves along parallel to the sensing element, the distance between the position magnet and the sensor electronics (head) end of the sensor is read electronically (see Figure 9.1).

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

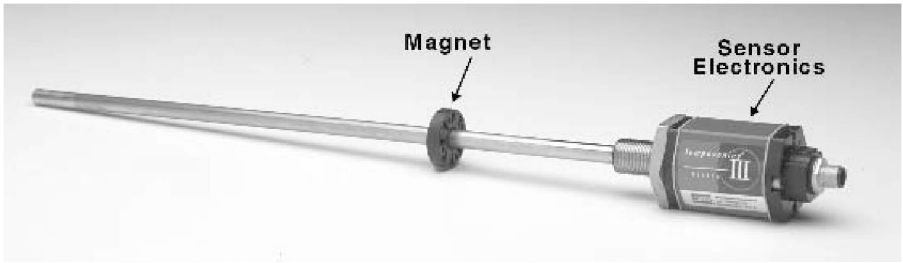

Figure 9.1 Magnetostrictive position transducer with position magnet. (Courtesy of MTS Systems Corporation.)

9.2 MAGNETOSTRICTION

Magnetostriction is a property of ferromagnetic materials, including nickel, iron, cobalt, and some of their alloys. When placed in a magnetic field, these materials change size and/or shape. James Prescott Joule discovered the magnetostriction of iron in 1842. He established the fact that when magnetized, iron increases in length in the direction of magnetization and contracts at right angles to this direction [4, p. 630; 35, p. 4].

The physical response of a ferromagnetic material is due to the presence of magnetic moments. Generally speaking, most materials have approximately as many electrons spinning in one direction as the other, resulting in a magnetically insensitive structure [36, p. 123]. However, in an element with unfilled subvalence shells, more electrons spin in one direction than in the other. Therefore, these elements have a net magnetic moment and can be understood by considering the material as a collection of tiny permanent magnets called *domains*. Each domain comprises many atoms. When a material is not magnetized, the domains are arranged randomly. However, when the material is magnetized, the domains are oriented with their axes approximately parallel to each other (see Figure 9.2). Interaction of an external magnetic field with the domains causes a magnetostrictive effect. Controlling the ordering of the domains through alloy selection, thermal annealing, coldworking, and magnetic field strength can optimize this effect [24].

The term *magnetostriction* generally refers to any change in dimensions due to magnetization. However, magnetostrictive position sensors utilize *Joule magnetostriction* which is the change in length of a ferromagnetic material due to magnetization. When a material has positive magnetostriction, it enlarges when placed in a magnetic field. Conversely, with negative magnetostriction, the material shrinks. The amount of magnetostriction in base elements and simple alloys is on the order of a few parts per million. Some exotic materials, when magnetized, change dimension by a factor of hundreds of times greater than this when under a very strong magnetizing force. One such material developed in the 1980s is Terfenol D, but it has a high cost and limited

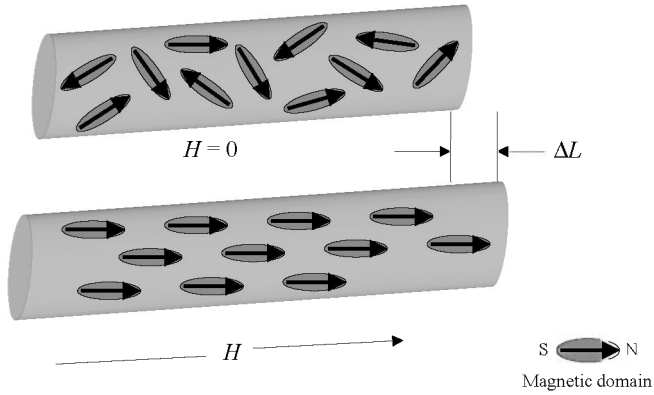

Figure 9.2 Positive magnetostriction. Magnetic domains align with magnetic field, H, and cause stress, inducing an increase in mechanical dimension, ΔL.

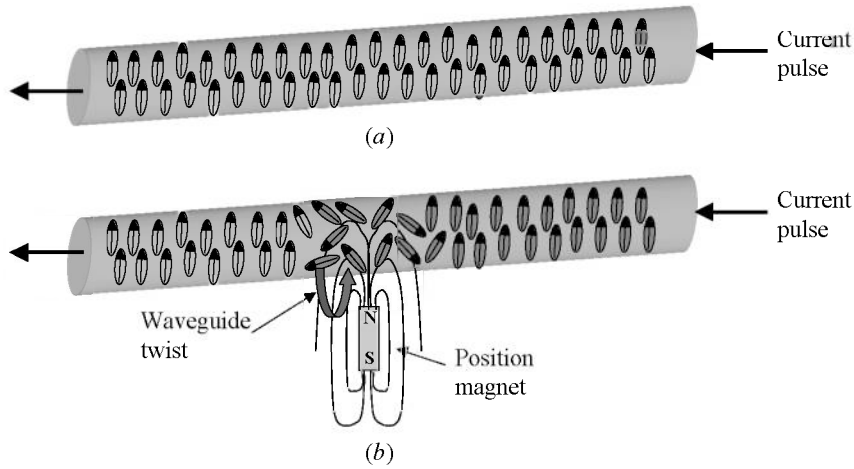

Figure 9.3 (*a*) Magnetic domains in a ferromagnetic wire are aligned by the current pulse; (*b*) Wiedemann effect—a torsional force occurs at the location of an axial magnetic field (position magnet) when the current is applied to the wire.

practical use in sensors because it is not ductile and therefore cannot be drawn into a wire or ribbon shape.

Just as magnetization stresses the material, causing dimensional changes, the reverse is also true: Applying stress to a magnetostrictive material changes its magnetic properties (e.g., magnetic permeability). This stress-induced effect on magnetic properties is called the *Villari effect*. An important characteristic of a wire made of a magnetostrictive material is the *Wiedemann effect*. It is the torsional force produced in a ferromagnetic wire at the location of an axial magnetic field when a current flows in the wire (see Figure 9.3). The torsional

force results from the vector addition of the magnetic field surrounding the waveguide, due to the current pulse, and the magnetic field supplied by the position magnet. The torsional force produces a slight mechanical twist in the waveguide. In a magnetostrictive position transducer, the current is applied as a short-duration pulse, about 1 to 2 μs. The minimum current density is along the center of the wire, and the maximum is at the wire circumference. This is due to the skin effect and means that the magnetic field intensity, due to the current, is also greatest at the wire surface. This aids in developing the waveguide twist.

9.3 HISTORY OF MAGNETOSTRICTIVE SENSORS

An important precursor of the magnetostrictive position sensor was the magnetostrictive delay line as used in an early memory device of the 1960s. A wire made from a nickel–iron alloy acted as a sonic waveguide, forming the memory core for storing serial data. A transducer at the memory input enabled excitation of sonic pulses onto the waveguide. Another transducer at the output end of the wire detected the presence or absence of the sonic pulses (see Figure 9.4). To store data, a series of pulses was impressed onto the waveguide at the input (transmitter) end. The presence of a sonic pulse on the waveguide represented a logical 1, and the absence of a pulse at the expected time represented a logical 0. The sonic waves traveled down the waveguide at the "speed of sound" in that material (approximately 3000 m/s). Additional data could be fed into the input end until just before the first pulse reached the output (amplifier) end. When the first pulse was just about to reach the output end, the memory was full and no further data could be entered. As a sonic

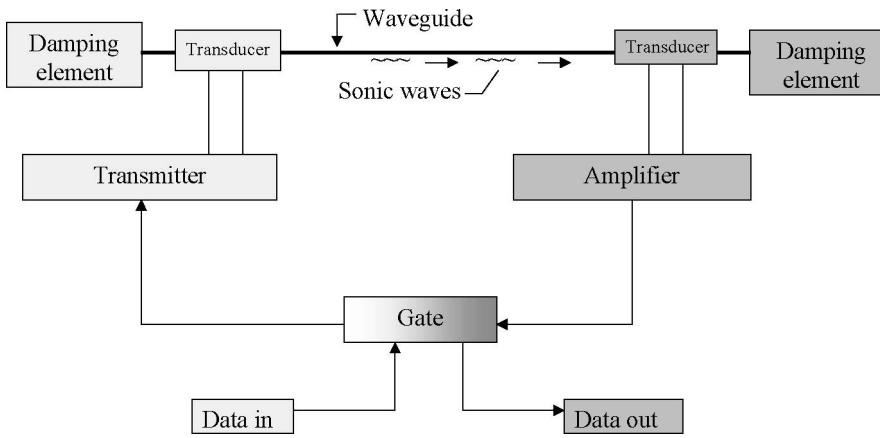

Figure 9.4 Serial memory device using a magnetostrictive wire (waveguide) delay line.

pulse reached the output end, it would be detected, amplified (to account for attenuation in the waveguide), and fed again into the input end. The data could be stored in the waveguide continuously. The data could be read serially after the synchronizing space between the last data pulse and the first pulse. Building on this delay line technology, Jacob Tellerman invented the magnetostrictive position transducer by adding a position magnet and introducing a current pulse into the waveguide. The first products based on this technology were trademarked Temposonics: *tempo* because the actual measurement is the time between the current pulse application and receipt of the position magnet pulse; *sonics* because a sonic wave enables the measurement.

9.4 MAGNETOSTRICTIVE POSITION TRANSDUCER DESIGN

To incorporate the ideas discussed above into a linear position sensor, the current is introduced as a short-duration pulse (called an *interrogation pulse*) of about 1 to 2 μs. This current pulse, together with the field from the position magnet, produces a mechanical torsional pulse at the location of the position magnet. The torsional pulse travels as a sonic wave in the waveguide at about 3000 m/s. The waveguide is typically a wire made from an iron–nickel alloy and is about 0.25 to 1 mm in diameter. The sonic pulse travels in both directions on the waveguide. The pulse traveling toward the pickup device is detected by the pickup. The pulse traveling in the other direction is eliminated by the damping device. This prevents interference due to sonic wave reflection from the waveguide end. In operation, a timer is started when the interrogation pulse is applied to the waveguide. An amount of time elapses as the sonic pulse travels from the position magnet location to the pickup. The timer is stopped when the pickup detects the sonic pulse. The elapsed time indicates the distance between the position magnet and the pickup.

A practical magnetostrictive position sensor comprises five basic components: waveguide, position magnet, pickup, damping element, and the electronic circuit. Design and optimization of each of these components for cost and performance depends partly on the type of application for which the sensor is intended. Industrial applications such as position feedback for control of hydraulic cylinders, require high resolution and fast update time. High-volume applications such as struts for automatic body control in cars require high reliability and low cost. The following paragraphs provide a closer look at some of these design considerations.

9.5 WAVEGUIDE

The waveguide material and manufacturing process are developed to control many performance characteristics. These include the coefficient of magnetostriction, sonic wave attenuation, sonic velocity temperature coefficient, vari-

ation of sonic velocity from unit to unit, variation of sonic velocity along the length of waveguide within one unit (this is the main source of nonlinearity), and hysteresis. Other less important properties include straightness, solderability, bendability, weldability, electrical resistance, and permeability. Short- or long-term drifts are not a property of the waveguide as long as it is operated at temperatures substantially lower than the Curie point. The technical properties of a magnetic material depend on both the intrinsic properties of the material and the microstructure [8, p. 377]. Microstructure is affected by coldworking and by thermal treatments.

The *coefficient of magnetostriction (c)* is the factor that relates the amount of strain (ε) produced in a magnetostrictive material to a given amount of magnetic field intensity (β):

$$\varepsilon = \beta c \tag{9.1}$$

One would think that it is important to have a high coefficient of magnetostriction in a magnetostrictive transducer. But to the contrary, it is more important to optimize parameters such as low hysteresis, low attenuation, and low sensitivity to temperature changes. For that reason, the coefficient of magnetostriction of the waveguide is normally only that of a simple nickel–iron alloy. A popular alloy to use for the waveguide is Ni-Span C, developed by the International Nickel Company. It was intended for use in springs to maintain a constant modulus of elasticity over a wide temperature range. Its magnetostrictive effect was also used in sonar transducers.

Attenuation is the gradual reduction in amplitude of the sonic wave as it travels over a longer distance in the waveguide. The amount varies with the processing of the waveguide material, and is logarithmic. This attenuation can be optimized (reduced) by proper selection of the waveguide material, the amount of coldwork, and annealing. When the waveguide wire is sized by passing it through progressively smaller dies, it becomes harder and is called coldwork hardening. Several times during the process, the wire must be heated to soften it, called *annealing*. The last stages of working and annealing result in a percent coldwork. Various manufacturers have their preferred targets, which are trade secrets.

The *temperature coefficient* of sonic velocity affects the thermal sensitivity of both the zero setting and the scale factor of the output. The effect on zero is because all readings are made relative to the distance of the position magnet from the pickup. The position magnet cannot come up against the pickup because it interferes with the pickup signal. So the zero position is a finite distance from the pickup. Since the popular alloy Ni-Span C has a low temperature coefficient of the modulus of elasticity, it also has a very low temperature coefficient of sonic velocity. With optimum processing, the temperature coefficient of Ni-Span C is less than 5 ppm/°C.

Sonic velocity varies somewhat from *unit to unit*. This can be minimized by tight control over the alloy composition, coldworking, and annealing process. The remaining variation is adjusted by individual calibration of the sensor,

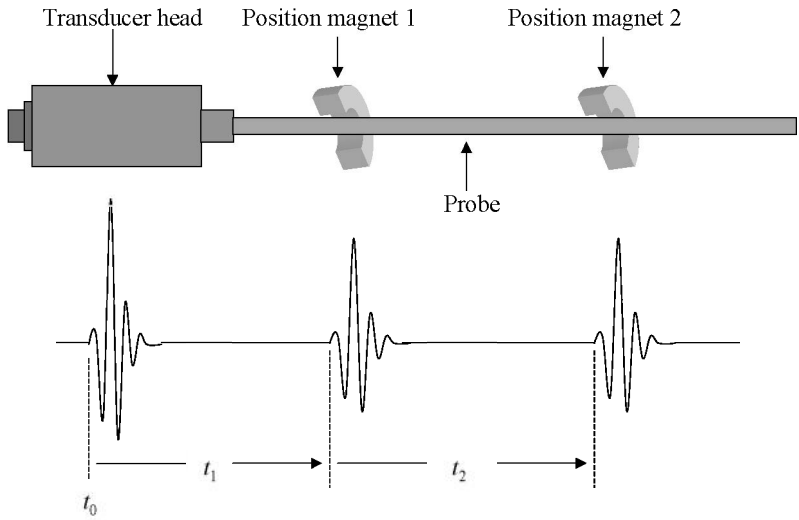

Figure 9.5 Oscilloscope trace of the signal on the pickup coil of a magnetostrictive position transducer.

loading a calibration factor into onboard memory that can be downloaded by the user, or by publishing a calibration factor for use with that sensor.

Nonlinearity that originates with the waveguide has two sources: processing/handling induced, and signal interference. Process/handling-induced errors can be reduced by the same methods as outlined above for sonic velocity. The signal interference induced error occurs mainly near the head end of the transducer. When the waveguide is interrogated, the pickup coil receives some of the magnetic field and rings for a short period of time, exhibiting a decaying sinusoidal wave of voltage across its terminals (see Figure 9.5). This ringing adds to the signal when the position magnet is close enough to the head that it is within that short time period.

Since the waveguide is a ferromagnetic material, it has a magnetic *hysteresis* when being magnetized or demagnetized. This manifests itself as a difference in sensor-indicated position when the position approaches a given point from an upscale position compared to approaching the same point from a downscale position. Once the waveguide material and processing have been selected, increasing the current level of the interrogation pulse can further reduce the hysteresis.

9.6 POSITION MAGNET

The position magnet can have any one of several shapes in order to fit the application requirement (see Figure 9.6). The most popular shape is the ring magnet. The ring may be a molded magnet material or can be a nonmagnetic

Figure 9.6 Position magnet shapes. (Courtesy of MTS Systems Corporation.)

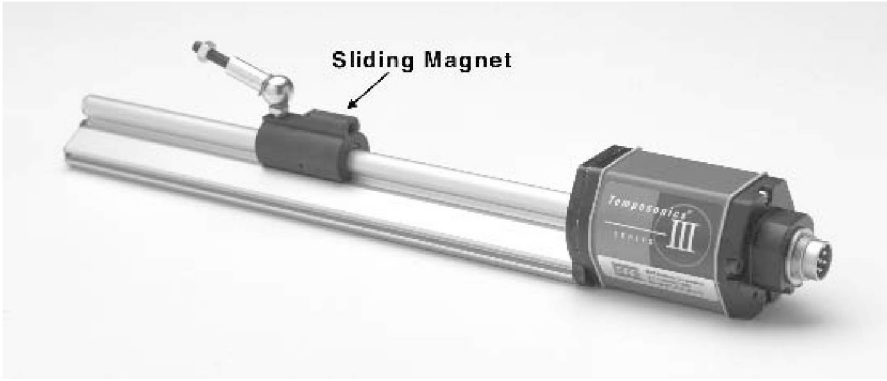

Figure 9.7 Sensor housing with a track to guide the sliding position magnet. (Courtesy of MTS Systems Corporation.)

ring with magnet rods inserted. With the waveguide assembly running through the center of the ring, the transducer output is less sensitive to the radial position of the waveguide within the ring (e.g., if the ring is moved 1 mm to the side, the output to the sensor will typically indicate less than 1/30 of that amount).

Sometimes the application requires something other than a ring for the position magnet. A cut into the ring (C-shaped magnet) allows the ring to be added or taken away from the waveguide without having access to the end of the waveguide. A bar magnet is the simplest configuration of position magnet but must track relatively parallel to the waveguide during its travel to preserve accuracy. This is sometimes accomplished in the sensor housing design as shown in Figure 9.7. The (bar-type) position magnet is mounted inside a sliding

member that follows a track, which is part of the sensor element housing. This maintains parallel alignment of the sensor element with the position magnet path without any effort by the user. This is popular in some factory machinery, such as plastic injection molding machines. An alternative that also requires good alignment is an outside ring or partial outside ring. In this case, the waveguide is located on the outside of the ring magnet so that the ring may be allowed to rotate while still energizing the waveguide.

9.7 PICKUP DEVICES

At least three types of *pickup* devices are in common use: radial tape, coaxial coil, and piezoelectric pickup element (see Figure 9.8). The radial ferromagnetic tape is the most difficult to manufacture but has the advantage that the tape material can be optimized as needed for a pickup device. This results in a much higher signal/noise ratio than in the other types, since the tape can be optimized for maximum output instead of for a low thermal coefficient of sonic velocity or low attenuation versus position. The tape is a ribbon of magnetostrictive material with a higher coefficient of magnetostriction than the waveguide and is welded to the waveguide. As the sonic wave reaches the intersection of the waveguide and tape, some of the sonic pulse energy travels down the tape. Here, the sonic pulse is a compression wave rather than a torsional wave. Due to the Villari effect (also called *reverse magnetostriction*), the compression wave causes an area of different permeability from the remainder of the tape material. The tape protrudes into the center of a coil and is magnetized by an adjacent *bias magnet*. As the sonic wave on the tape passes through the coil, the associated area of different permeability also passes through the coil. This causes a change in magnetic field intensity on the coil. Due to the Faraday effect, a voltage pulse is produced across the coil ends. This voltage pulse is used to detect the presence of a sonic pulse.

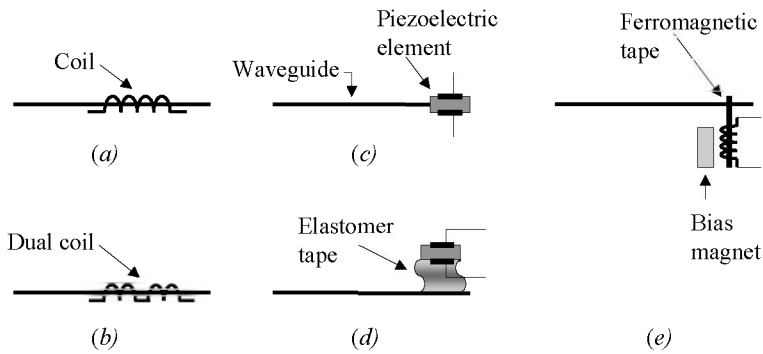

Figure 9.8 Types of pickup devices: (*a*) coaxial coil, (*b*) dual coaxial coil, (*c*) piezoelectric, (*d*) elastomer tape, and (*e*) ferromagnetic tape.

The coaxial coil pickup comprises a coil of wire through which the waveguide passes. This is simpler than the tape pickup because it eliminates the need for the tape and for the bias magnet. The current pulse used to generate the sonic wave also leaves a remanent magnetic field on the waveguide. This is enough to provide the magnetic field for the pickup coil. As the sonic wave on the waveguide passes through the coil, the associated area of different permeability also passes through the coil. This produces a voltage pulse across the coil leads in the same way as explained for the tape pickup. The disadvantage of this simpler design is that since the waveguide must be designed for low attenuation and temperature sensitivity, the coefficient of magnetostriction that results is relatively low. This results in a lower amplitude of signal and an associated lower signal/noise ratio.

A piezoelectric pickup can be used to detect the sonic wave from the waveguide without using a magnetic field. This provides the hope that it would be less sensitive to interference from external magnetic fields. In actual sensors, however, it has proven difficult to get as good an energy transfer as can be had with the ferromagnetic tape method. The associated low signal level results in a low signal/noise ratio, confounding the effort to reduce corruption from external interference. An additional area of concern with the piezoelectric pickups is that they are generally made from brittle ceramic materials. A lot of field problems have been caused by fracture of the piezoelectric element when the transducer was inadvertently dropped onto a hard surface.

A piezoelectric element has also been coupled to the waveguide by an elastomer tape in the hope of reducing vibration sensitivity. The result, however, has been reduced sensitivity and an increase in manufacturing time.

9.8 DAMP

A damping element (also called the *damp*) is incorporated at the tip end of the waveguide to remove sonic pulses that are moving away from the pickup. If they were not removed, they would reflect from the waveguide tip and travel back toward the pickup, causing interference with the desired signal. The damp has to have sufficient damping quality to remove the unwanted sonic pulse, but not be so effective at the front end that it causes a reflection directly from its own front end. Various combinations of shape and hardness of elastomeric materials are used to approximate the desired performance.

9.9 ELECTRONICS

The electronic circuit provides the interrogation pulse, amplifies, filters, and detects the signal pulses, and then provides the desired analog or digital output (see Figure 9.9). When designing a sensor that could be as short as a few millimeters or as long as several meters, the waveguide current must be controlled

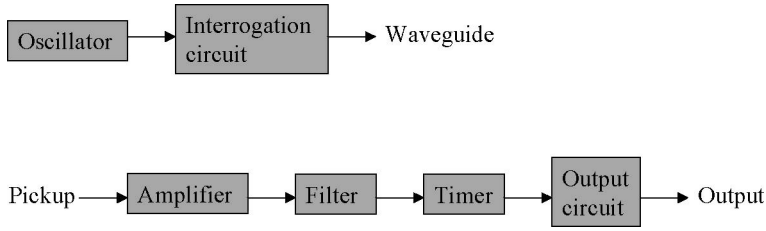

Figure 9.9 The electronic circuit applies the interrogation pulse, conditions the return pulse, measures the elapsed time, and converts it to the desired output signal.

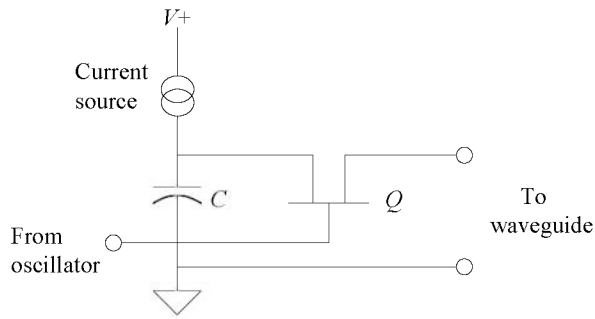

Figure 9.10 Interrogation circuit with hold capacitor, C, and field-effect transistor, Q.

to give sufficient current for good signal amplitude and low hysteresis, but not so large that ringing in the pickup limits the usefulness near the head. Sometimes a constant-current source circuit is used, and sometimes a microprocessor adjustable voltage or current source is used. Since the interrogation pulse occurs at a rate somewhere between 10 Hz and 5 kHz, but the pulse duration is only 1 to 2 μs, energy for the interrogation pulse is normally stored in a holding capacitor. This reduces the peak current demand on the sensor power supply leads (see Figure 9.10). Normally, a capacitor is charged up in one of various ways and then a field-effect transistor (FET) discharges it into the waveguide. The FET is used because it can have a low on-resistance at high current (2 to 5 A) but requires only microamperes at its control pin (gate).

After the waveguide is interrogated and the pickup supplies the signal pulse, the signal pulse is conditioned. The signal pulse is of a frequency in the range 100 to 300 kHz. A high-pass or bandpass filter can be used to reject noise from other mechanical and electrical sources. The signal amplitude can be several hundred millivolts when using a welded ferromagnetic tape type of pickup, but is only a few millivolts when using other pickups. So, with other pickups, amplification is needed. After the amplifier, the signal pulse is detected using a voltage comparator.

In operation, the interrogation pulse also sets a flip-flop and triggers a blanking timer, which waits for the interrogation ringing to die down. The flip-flop will be reset, after a minimum blanking time set by the timer, upon receiving the stop pulse. The output of the flip-flop is then a pulse-width-modulated signal, which can be used as the sensor output or can be conditioned further. Alternative outputs include analog voltage or current and various digital formats (SSI, CANbus, HART, Profibus, etc.). Although the current pulse to drive the waveguide usually has an amplitude of over 2 A, it is possible to make a two-wire magnetostrictive position transducer that operates on less than 4 mA, thus allowing the design of a two-wire 4- to 20-mA transmitter. This was first accomplished by a technique invented by the author and patented in 1991 [26].

9.10 ADVANTAGES

As mentioned earlier, the original magnetostrictive position sensing technology was trademarked Temposonics. Position transducers based on this technology are noncontact, absolute reading, and have essentially infinite resolution. Transducer lengths in production include a full-scale range (FSR) of less than 10 mm up to a FSR of over 20 m. Analog and digital outputs are available to indicate position, displacement, velocity, and acceleration. Competing technologies include LVDTs, inductive sensors, encoders, ultrasonic sensors, and potentiometers.

Due to the noncontact nature of magnetostrictive position transducers, there are no mechanical parts to wear out. This is an advantage over contact sensors, the most popular being the potentiometer. For example, even though some potentiometers list an operating life of over 10 million cycles, this can be exceeded in only a few months when the system experiences a continuous application-induced vibration or a control system-induced dithering. With a vibration or dithering frequency of only 10 Hz, you find

$$\begin{aligned} &10\,\text{cycles/s} \times 3600\,\text{s/h} \times 24\,\text{h/day} \times 30\,\text{days/month} \\ &\quad = 25\,\text{million cycles/month} \end{aligned} \tag{9.2}$$

Typically, a motion control system will have a few spots in the active stroke, which are used most often. So in a few months, these often-used spots can be traversed a number of times exceeding the rated lifetime. When a spot is worn on a potentiometer element, the result is an area of unstable or inaccurate readings.

Since magnetostrictive position transducers read the location of a position magnet, performance is not affected when nonmagnetic materials are placed between the position magnet and the waveguide. This leads to novel applications requiring the position magnet to be on the outside of a housing or other

TABLE 9.1 Magnetostrictive Position Transducers versus Other Technologies

Technology	Resolution	Nonlinearity	FSR[a] Available	Ruggedness
Magnetostriction	High	Low	10 mm–20 m	High
LVDT	High	Medium	2 mm–200 mm	High
Inductive	Medium	Medium	2 mm–500 mm	High
Encoder	Medium	Low	10 mm–2 m	Low
Ultrasonic	Low	High	100 mm–20 m	Medium
Potentiometer[b]	Medium	Medium	10 mm–500 mm	Medium

[a] FSR, full-scale range.
[b] The Potentiometer is a contact-type transducer; all others listed are noncontact.

functional member, and the waveguide on the inside (such as in a hydraulic cylinder). For example, the nonmagnetic materials can be aluminum, plastic, some stainless steels, and others.

The absolute measurement provided by magnetostriction-based position transducers offers an advantage over the incremental (or displacement) measurements of magnetic or optical linear scales, encoders, and so on. This is because an absolute position sensor measures the distance between a fixed datum and the point of interest, whereas a displacement sensor measures the distance between a previous measurement point and the current point of interest. An incremental sensor (such as an incremental encoder) counts the number of increments since the last reset of the count. The size of the increment determines the measurement resolution. If the previous position or count is forgotten, the displacement or incremental transducer must be reset to zero or a known position. Since the actual measurement of a magnetostrictive position transducer is presented as a variable time period, digital circuits can be used for all the electronic functions after detection of the return pulse. The "speed of sound" in the waveguide does not change with time or temperature; so the position output is extremely stable.

The time period representing the measured position is analog, with infinite resolution, although the timing of digital pulses indicates it. The resolution of the transducer signal output depends only on the resolution of the timing circuitry and its ability to recognize the signal over any noise that may be present. Temposonics industrial sensors have a resolution as fine as 1 μm. Table 9.1 compares some essential features of magnetostriction versus some competing technologies.

9.11 TYPICAL PERFORMANCE SPECIFICATIONS

Listed below are some of the specifications of the Temposonics III CANbus industrial sensor.

Measured variables:	position, displacement, velocity, set points
Resolution:	sensor: essentially infinite; CANbus output: 0.002 mm
Repeatability:	0.001%
Nonlinearity:	0.01% (relative to a least squares straight line)
Hysteresis:	0.004 mm
Update time:	≤1 ms
Output style:	CANbus (can be specified as various analog or digital types)

9.12 APPLICATION

Magnetostrictive position transducers are made in many configurations for use in a wide range of applications. Usually, the position magnet is separate from the sensor element and is attached to the member to be measured. This configuration is used in general machinery applications, primary and secondary woodworking equipment, hydraulic cylinders, shock absorbers, and many others. Figure 9.11 shows an example of a typical configuration.

An additional advantage of the magnetostrictive position transducer is that it can be used with multiple position magnets. On some standard models, up to 32 position magnets can be used simultaneously [25]. More magnets can be used in special applications. When an interrogation pulse is applied to a magnetostrictive waveguide having multiple position magnets, an ultrasonic pulse is launched at the location of each magnet. The pickup senses each return pulse in sequence. Each set of timing data, relating to the respective return pulse, is stored in an appropriate memory location. A microprocessor can then retrieve the individual position data, representing each magnet, as needed.

Recently, large-volume production of magnetostrictive transducers has begun for use in automotive applications. Automatic body control of passenger vehicles is one such use [27]. To accomplish this, a magnetostrictive posi-

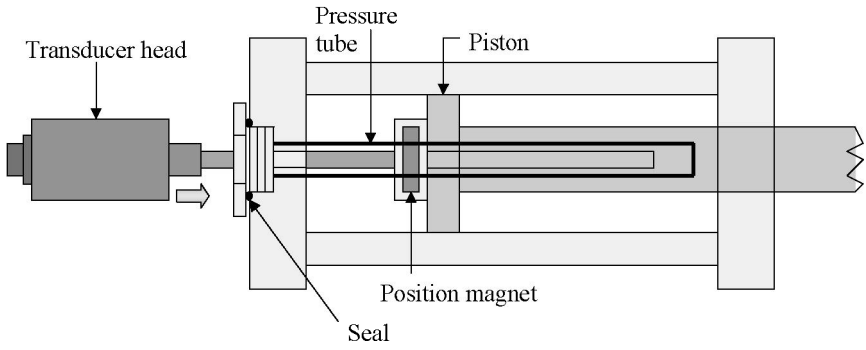

Figure 9.11 Magnetostrictive position transducer installed in a hydraulic cylinder.

TABLE 9.2 Applications of the Magnetostrictive Position Transducer

Industry	Application
Automotive	Production machinery and also onboard in suspension, transmission, and steering
Chip and wafer handling	Precision measurement and no wearing parts enable this application
Electric actuators	Both linear and rotary position can be measured with one sensor and two magnets (patent pending)
Hydraulic and pneumatic cylinders	The sensor is mounted within the rod and the magnet is fixed to the cylinder
Food and beverage	Milk tanks and can filling machines
Liquid level	Process control, leakage detection, and inventory control
Medical	Hospital bed positioning
Metalworking	The sensor is used for measurement and control in forges, presses, bending machines, and cutoff machines
Mobile equipment	Garbage trucks, agriculture, and grading and paving
Paper converting	Used to control slitters and flexographic presses
Plastics	Multimagnet transducers on plastic injection molding machines measure motion of the injector, ejector, and mold halves; also used in blow-molding
Primary metal	Walking beams and ladle control
Primary wood	Sawmills, portable sawmills, lathes, cutoff saws, positioning knees, and presses
Secondary wood	Saw positioning and tennoners
Testing equipment	Materials testing, automotive, military and aerospace testing, earthquake, and wavemaking machines
Textiles	Used in carpet tufters

tion transducer is mounted inside one sliding member of a shock tower or strut. The position magnet is mounted to the other sliding member. These position data are used with a controller and hydraulic system to improve ride and handling.

Because of the ruggedness, temperature stability, reliability, high performance, multiple magnet capability, and wide range of lengths, magnetostrictive position transducers are used in many different applications. Some of these are listed in Table 9.2.

CHAPTER 10

ENCODERS

10.1 LINEAR ENCODERS

Linear displacement can be measured and communicated by a device called an encoder without using any form of analog-to-digital conversion because the basic output signal is already in a digital format. Although the term *encoder* has also been applied to devices based on laser interferometers, the term *linear encoder* will be used here in reference to standard industrial transducers based on geometric patterns applied along a linear scale and detected by any one of several methods. Linear encoders are available as incremental or absolute reading and encompass various detection techniques, including brush, optical, magnetic, and capacitive types. Besides selecting whether the output will be absolute or incremental, encoder designers and users make trade-offs among important product features, including ruggedness, resolution, and physical size.

10.2 HISTORY OF ENCODERS

The earliest type of linear encoder was the *brush* type, shown in Figure 10.1, in which mechanical contact fingers rub along a metal pattern printed onto an insulating base. The path of the brush moves over conducting and insulating segments. When the brush is in contact with a conducting segment, a contact closure occurs [23, p. 106]. The pattern is formed onto the base in the same

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

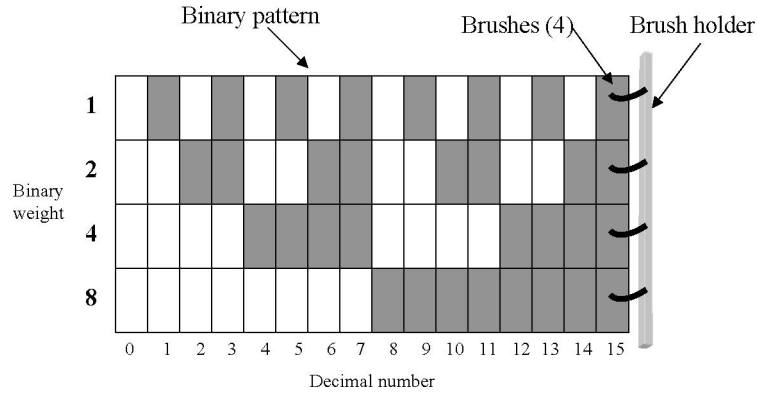

Figure 10.1 Brush type of linear position encoder with binary pattern. (See Section 10.8 to learn about other patterns.)

way as printed circuit boards are made for connecting electronic circuits. The major drawback of this contacting measurement technique was the wear of the metal pattern and fingers, sometimes resulting in errors, and eventually resulting in failure.

Noncontact encoders based on optical and magnetic techniques took over in the 1960s and 1970s because of their increased life and reliability. In the 1990s, the brush type largely became no longer acceptable for industrial measurement systems. For this reason, additional details of brush-type encoders are not included here, except where expedient to illustrate switching and coding theory.

10.3 CONSTRUCTION

A linear encoder can take many physical forms. Some have a housing similar to a potentiometric linear position transducer, comprising a housing, rod, bushing, wiper, scale, detector, and electronics. Others have a read head that rides in a track along the scale. Still others have a separate read head that must be mounted by the user so that it passes along parallel to the scale. An optical type is shown in Figure 10.2. The housing provides the base for combining the component parts, and includes means for mounting the encoder to the application.

In a *rod* type, the rod is usually made from hardened and polished steel and moves linearly into and out of the housing from one end. A bushing supports the rod, reducing wear and resisting the force due to side loading. One or more wipers clean particles away from the rod before it goes through the bushing. This reduces wear of the rod and bushing. The scale is an opaque member with slots (or reflective areas) for an optical transducer or a coded magnetic strip

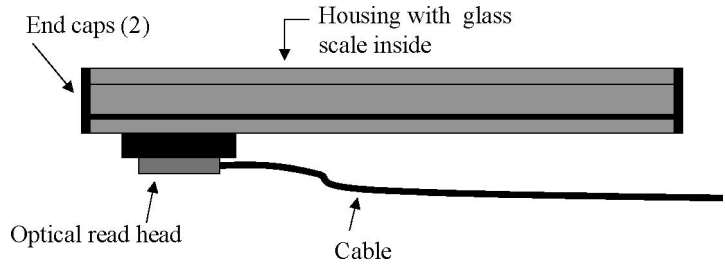

Figure 10.2 Optical type of linear position encoder.

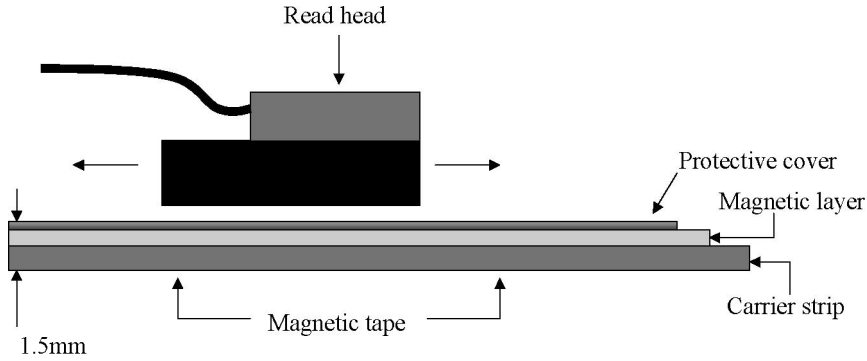

Figure 10.3 Incremental magnetic encoder with separate magnetic tape.

for a magnetic transducer. As the rod moves in or out, displacement or absolute position information is read from the scale via the detector.

An alternative construction is shown in Figure 10.3. In this magnetic encoder, the magnetic tape is supplied to the user as a separate component. The tape is attached to a stationary member, along the axis to be measured. The moving member is fitted with a magnetic pickup device. The combination of the coded tape and the pickup device comprises the linear encoder. The incremental pulse counting encoder circuit is zeroed at a starting position, often at one end of the tape. Then, as the pickup moves along the tape, the variations on the tape are counted and used to indicate the position as a distance from the starting position. As in all incremental encoders, this is actually a measurement of displacement rather than one of position.

10.4 ABSOLUTE VERSUS INCREMENTAL ENCODERS

It may be obvious that the information encoded onto the track of an incremental encoder can have a finer pitch than that of an absolute encoder. This

is because only one or two bits must be toggled while the position is changing with an incremental encoder. With an absolute encoder, sufficient information must be read at each position to represent the position totally at that point in the transducer stroke with respect to the reference datum. So if the position count is 2046 counts, an absolute encoder must encode the number 2046 (requiring 11 binary bits) at that exact location, whereas with an incremental encoder, there is only one bit of information located at that point (or two, counting the quadrature bit at 90° separation from the first, as presented later). In addition to the A quad B tracks, an incremental encoder also may have an index track. The index track has a mark at one point that indicates the index or starting point. Then the count can be rezeroed automatically each time the index point is passed. So an advantage of an incremental encoder is that of higher resolution for a given spacing of the detected feature (e.g., optical slots). An absolute encoder, however, has the advantage of instant startup after a corruption of power or data, without needing to rezero or reset the count.

10.5 OPTICAL ENCODERS

An optical encoder uses a transmitter/receiver pair of optoelectronic devices. The transmitter part of the pair is a light source, usually a light-emitting diode (LED). The receiver part is usually a phototransistor. A track having opaque and transparent sections arranged in series is passed through the optoelectronic pair as the position changes. (Alternatively, reflective and nonreflective sections can be used; see Figure 10.4.)

The light source has a means for focusing the light onto the detector. This may include a focusing lens, collimator, and/or a slit or pinhole. If the encoder is incremental, only two receivers are needed. The second receiver is spaced

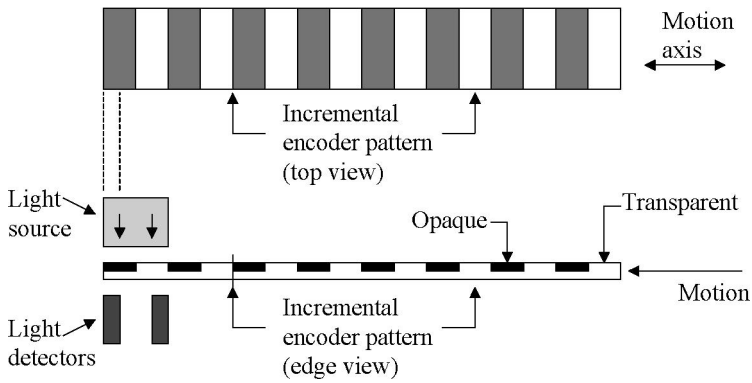

Figure 10.4 Incremental optical encoder with LED and phototransistor arranged for quadrature output.

from the first by a set distance to yield a 90° shift between the two outputs, resulting in a quadrature output (see Section 10.7). In this case, the alternating dark/light areas are counted to obtain the present position. A closer spacing between adjacent dark or light areas will yield a higher resolution. A third receiver may be used to detect an index mark.

In an absolute encoder, a sufficient number of light receivers are needed to represent the maximum number of bits to be indicated. The reciprocal of this number of bits is the resolution. For example, a 12-bit encoder represents 4096 counts and has a theoretical resolution of 0.024% or 1/4096. In this example there will be 4096 sets of position data along the full-range stroke of the sensor. Twelve parallel data bits will be read by 12 phototransistors. It is possible to use one long light bar to drive all of the phototransistors, or multiple LEDs can be used.

10.6 MAGNETIC ENCODERS

A magnetic encoder uses a magnetic tape onto which the position information is recorded, together with one or more magnetic sensors to read the data. The magnetic sensors in modern encoders are usually Hall effect or magnetoresistive types. Information on the Hall effect and on magnetoresistance is included here in Chapters 7 and 8, respectively. An incremental magnetic encoder is shown in Figure 10.5. Within the track is a magnetic tape having reversals of polarization along its length. These magnetic field variations are detected by the magnetoresistive pickups. If the encoder is incremental, only two pickups are needed. The second pair is spaced from the first pair by a set distance to yield a 90° shift between the two outputs, resulting in a quadrature output (see Section 10.7). In this case the pulses from the pickups are counted to obtain the present position. The highest resolution that is possible is limited by the smallest size of two adjacent field variations that can be

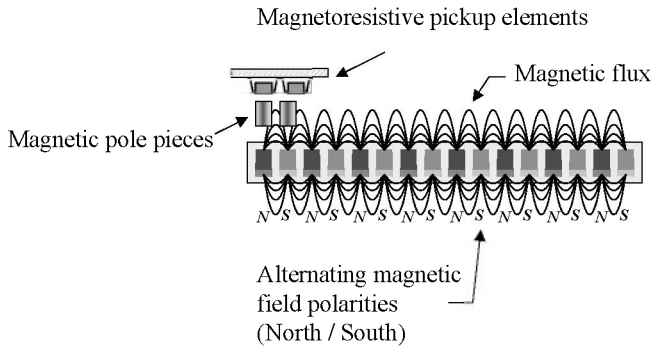

Figure 10.5 Incremental magnetic encoder with magnetoresistive pickup.

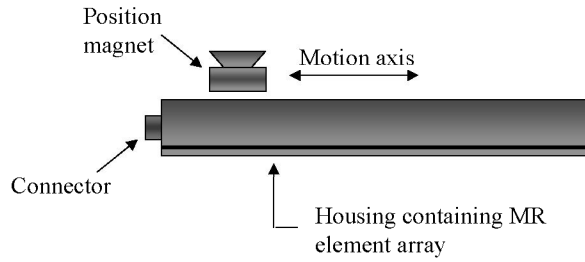

Figure 10.6 Absolute magnetic encoder.

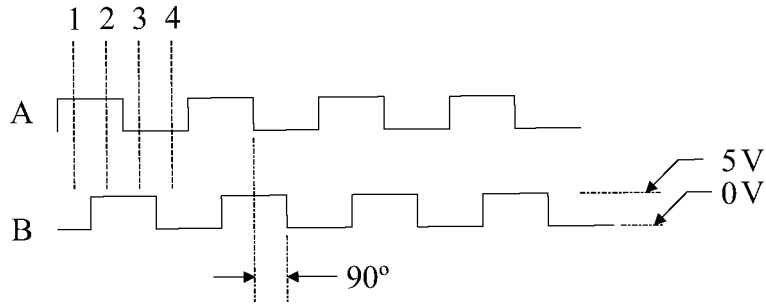

Figure 10.7 A quad B outputs are separated by 90°. States 1 through 4 occur during each count cycle.

differentiated by the pickup device. The magnetic encoder can be incremental or absolute (see Figure 10.6), following the same theory as presented for the optical type.

10.7 QUADRATURE

Incremental encoders have two outputs, called A and B. These are arranged in quadrature, which means that they are separated in phase by 90°, as shown in Figure 10.7. This arrangement is also called *A quad B*. It is called *quadrature* because "quad" means four, and a transition of one or the other data line (A or B) occurs four times per count cycle. A complete count cycle in most situations is generally considered to comprise 360° (the 360° is a count cycle, not the angle of rotation of a rotary sensor). So with four transitions per count cycle, one phase is delayed when compared to the other by 90°, or one-fourth of 360°. The four possible states are (1) A high, B low; (2) A high, B high; (3) A low, B high; and (4) A low, B low.

If the pulses from either A or B are counted, the change in position is indicated by the number of counts multiplied by the distance per count (e.g., 1000

counts with a resolution of 10 μm would be 1000 counts × 10 μm = 10 mm). The reason for having the other output (B or A) is to find the direction of motion, either incrementing or decrementing the count, so the current count represents the actual position. In Figure 10.7, for example, if A goes from low to high while B remains low, it is an increment. Conversely, if B is high at that time, it is a decrement. This basic explanation can be used to easily understand how to obtain a count and the count direction. It is also possible to obtain four counts per cycle by looking at all of the transitions as count inputs and looking at the relationship between the A and B inputs, at the times of the counted transitions, to determine the direction of the change.

Modern circuits that read the A quad B signals from a position transducer do not actually wait for a transition to occur and then count that transition. Instead, it is more common to monitor the states of A and B continuously at a higher sampling rate than the transitions are expected to occur. With this (state, rather than transition) information and, usually, a microcontroller, smoother operation can be obtained with less likelihood of error.

10.8 BINARY VERSUS GRAY CODE

In an absolute encoder, the output is typically either in binary or Gray code, but binary-coded decimal (BCD) is also available. BCD is similar to binary, except that the data bits are arranged into sets of 4 bits each. Four bits of BCD data equal one character and can have a value of zero through nine. For reference, hexadecimal is also shown because the reader may be familiar with this binary number representation. In hexadecimal, an 8-bit binary word is viewed as two 4-bit nibbles. The 4-bit nibble can have 16 possible values. These are represented as the decimal numbers 0 through nine, followed by letters A through F (for a total of 16 characters).

Natural binary or BCD is easy to interface directly to standard digital circuits but has the disadvantage that a change of only one increment involves the simultaneous change of more than one bit. This means that a large error can be indicated if all the bits do not switch at exactly the same time. For example, when the count changes from 7 to 8, it is a change from 0111 to 1000 in binary or BCD. But if the most significant bit (MSB) changes a few milliseconds before the rest of the bits change, there will be a reading of 15 (1111 in binary, F in hexadecimal, or an error in BCD) for those few milliseconds. This can represent a large error and cause stability problems in a servo control system. A system called the Gray code was developed to solve this problem.

The Gray code is arranged so that an increment of 1 bit always changes the output by only 1 bit [37, pp. 6–126]. The differences among BCD, hexadecimal, binary, and a Gray code are shown in Table 10.1. Figure 10.8 is a corresponding Gray code pattern. One can see that a change between any two adjacent numbers requires only the change of 1 bit in the Gray code. When the transducer operates using the Gray code, the controller or other device to

TABLE 10.1 Decimal Equivalents of Hexadecimal, Binary-Coded Decimal (BCD), Natural Binary, and Gray Code

Decimal	Hexadecimal	BCD		Natural Binary	Gray
0	0	0000	0000	0000	0000
1	1	0000	0001	0001	0001
2	2	0000	0010	0010	0011
3	3	0000	0011	0011	0010
4	4	0000	0100	0100	0110
5	5	0000	0101	0101	0111
6	6	0000	0110	0110	0101
7	7	0000	0111	0111	0100
8	8	0000	1000	1000	1100
9	9	0000	1001	1001	1101
10	A	0001	0000	1010	1111
11	B	0001	0001	1011	1110
12	C	0001	0010	1100	1010
13	D	0001	0011	1101	1011
14	E	0001	0100	1110	1001
15	F	0001	0101	1111	1000

Figure 10.8 Gray code pattern corresponding to Table 10.1.

which it is connected will usually incorporate a Gray-to-binary conversion. A Gray-to-binary converter can be built in hardware with gates as shown in Section 10.9, or it can be stored in a lookup table in memory. Then the controller can use the binary numbers for operation as usual after they are converted from the Gray code.

10.9 ELECTRONICS

The light source of an optical encoder is usually one or more LEDs driven by a simple reference voltage with current limiting resistor, or by a constant-

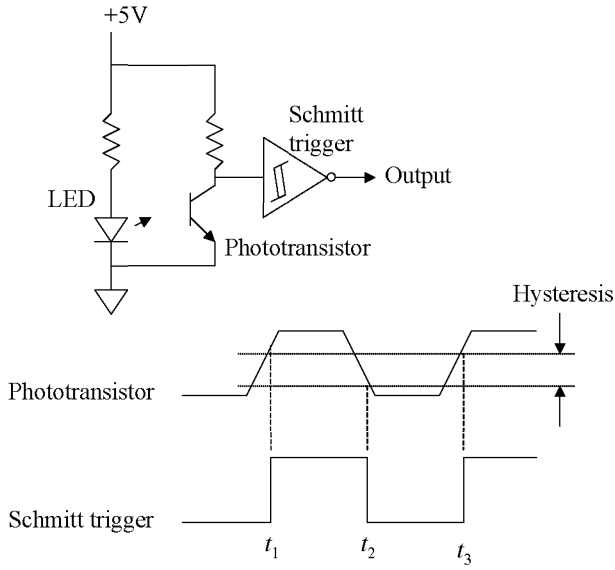

Figure 10.9 LED and phototransistor connection circuit, with Schmitt trigger.

current source. Temperature compensation capability may be added to maintain a constant light output with temperature variations. The LED and phototransistor are typically connected as shown in Figure 10.9. The signal directly from the phototransistor will not be a waveform with sharp transitions. The sharp transitions needed for the A quad B output are generated by a Schmitt trigger circuit. A Schmitt trigger transitions from one state to the other when the input crosses a voltage threshold. It has hysteresis built in so that the opposite transition will not take place until the input moves substantially past that same threshold. So there is a positive-going threshold and a negative-going threshold. In the figure, the positive-going threshold is approximately 4 V (with a 5-V power supply voltage). The negative-going threshold is approximately 1 V. This is a hysteresis of 3 V.

If an absolute encoder uses a Gray code pattern, the resulting output must be converted to natural binary before the data can be handled by a microcontroller. This conversion can be done by using a mathematical formula in the microcontroller, a lookup table, or in hardware. A schematic is shown in Figure 10.10 for converting the Gray code to natural binary.

10.10 ADVANTAGES

With resolution as fine as 0.01 μm (millionths of a meter), linear encoders are often the preferred method of position sensing in the precision environment

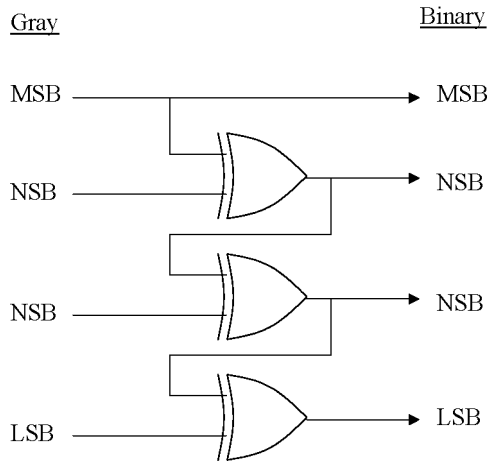

Figure 10.10 Schematic for converting Gray code to natural binary.

of machine tools. The highest-resolution transducers tend to be the incremental ones. A disadvantage to the incremental mode is that data can be corrupted due to electromagnetic noise or power fluctuations.

Magnetic linear encoders, with a magnetic tape and a sensor head, can have a long measuring range of up to 40 m (such as the Lincoder by Stegmann). The user makes the installation by attaching the magnetic tape to a fixed surface, along which the sensor head will travel. They retain the advantage of non-contact measuring, and so have a virtually unlimited life.

Shorter linear encoders can be magnetically or optically coupled and are self-contained with the magnetic tape or optical scale contained within a housing. Typical industrial models can have a resolution of better than 0.1 μm. The bushings and wipers, used with models having an actuator rod, have a finite lifetime. End of life, however, does not mean that the measurement accuracy is affected. It means that some mechanical drag is encountered as the rod rubs on worn bushings or bearings, sometimes accompanied by audible noise.

10.11 TYPICAL PERFORMANCE SPECIFICATION AND APPLICATIONS

Figure 10.2 showed an optical linear encoder. Representative specifications are as follows:

Full-scale range:	100 mm to 3 m
Resolution:	0.1 to 10 μm
Accuracy at 20°C:	0.1 to 10 μm
Hysteresis:	0.5 μm

Operating temperature:	0 to 50°C
Maximum speed:	1 m/s
Input power:	5 V dc at 180 mA maximum

Coding patterns can be photographically reproduced onto the measuring medium. Any nonuniformity of the pattern on the measuring medium is a source of error. In a linear encoder, these include the width and spacing of the optical, magnetic, or conductor tracks that represent the individual bits. Incremental encoders generally have finer resolution, in a given technology, than do absolute versions. Magnetic scales are somewhat more rugged than optical scales, because high-resolution optical scales are made from glass. Typical encoder applications include process machinery feedback and control, robotics, and measuring equipment. Equipment using an incremental encoder will often have a zeroing function when the machine is first turned on, and additional rezeroing cycles at opportune times during operation. This is to avoid, as much as possible, prolonged error of the position count if it becomes corrupted by erratic motions or externally induced noise.

REFERENCES

1. D. Askeland, *The Science and Engineering of Materials*. Boston: PWS-Kent, 1989.
2. L. K. Baxter, *Capacitive Sensors Design and Applications*. Piscataway, NJ: IEEE Press, 1997.
3. R. Boll, *Soft Magnetic Materials*. London: Heyden & Son, 1977.
4. R. M. Bozorth, *Ferromagnetism*. New York: D. Van Nostrand, 1951.
5. H. Burke, *Handbook of Magnetic Phenomena*. New York: Van Nostrand Reinhold, 1986.
6. J. J. Carr, *Sensors and Circuits*. Upper Saddle River, NJ: Prentice Hall, 1993.
7. J. R. Carstens, *Electrical Sensors and Transducers*. Upper Saddle River, NJ: Regents/ Prentice Hall, 1992.
8. D. Craik, *Magnetism Principles and Applications*. New York: Wiley, 1995.
9. B. D. Cullity, *Introduction to Magnetic Materials*. Reading, MA: Addison-Wesley, 1972.
10. Elcon Instruments, *Introduction to Intrinsic Safety*. Annapolis, MD: Elcon, 1989.
11. O. Esbach, *Handbook of Engineering Fundamentals*. New York: Wiley, 1975.
12. J. Fraden, *Handbook of Modern Sensors*. New York: Springer-Verlag, 1996.
13. E. Herceg, *Handbook of Measurement and Control*. Pennsauken, NJ: Schaevitz Engineering, 1976.
14. D. Jiles, *Introduction to Magnetism and Magnetic Materials*. London: Chapman & Hall, 1991.
15. D. E. Johnson and J. L. Hilburn, *Rapid Practical Designs of Active Filters*. New York: Wiley, 1975.

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9

16. R. Lerner and G. Trigg, *Encyclopedia of Physics*. New York: VCH Publishers, 1990.
17. P. Lorrain and D. Corson, *Electromagnetic Fields and Waves*. San Francisco: W.H. Freeman, 1962.
18. E. C. Magison, *Intrinsic Safety*. Research Triangle Park, NC: Instrument Society of America, 1984.
19. F. Mazda, *Electronics Engineer's Reference Book*, 6th ed. London: Butterworth, 1989.
20. P. Neelakanta, *Handbook of Electromagnetic Materials*. Boca Raton, FL: CRC Press, 1995.
21. J. C. Nelson, *Operational Amplifier Circuits*. Wobwin, MA: Butterworth-Heinemann, 1995.
22. NVSB series datasheet: Nonvolatile Electronics, Eden Prairie, MN. March 1996.
23. H. Norton, *Handbook of Transducers*. Upper Saddle River, NJ: Prentice Hall, 1989.
24. D. S. Nyce, Magnetostriction-based linear position sensors, *Sensors*, 11(4), 1994.
25. D. S. Nyce, Position sensors for hydraulic cylinders, *Hydraulics & Pneumatics*, November 2000.
26. D. S. Nyce, Low power magnetostrictive sensor, U.S. patent 5,070,485, 1991.
27. D. S. Nyce, Vehicle suspension strut having a continuous position sensor, U.S. patent 6,401,883, 2002.
28. H. Olson, *Dynamical Analogies*. New York: D. Van Nostrand, 1943.
29. R. Pallas-Areny and J. G. Webster, *Sensors and Signal Conditioning*, 2nd ed. New York: Wiley, 2001.
30. R. Philippe, *Electrical and Magnetic Properties of Materials*. Norwood, MA: Artech House, 1988.
31. E. Ramsden, *Hall Effect Sensors*. Cleveland, OH: Advanstar Communications, 2001.
32. R. Rose, L. Shepard, and J. Wulff, *The Structure and Properties of Materials*. New York: Wiley, 1966.
33. J. Shackelford, *Introduction to Materials Science for Engineers*. New York: Macmillan, 1985.
34. W. J. Tompkins, *Interfacing Sensors to the IBM PC*. Upper Saddle River, NJ: Prentice Hall, 1988.
35. E. D. Tremolet de Lacheisserie, *Magnetostriction: Theory and Applications of Magnetoelasticity*. Boca Raton, FL: CRC Press, 1993.
36. L. H. Van Vlack, *Elements of Materials Science*. Reading, MA: Addison-Wesley, 1964.
37. J. G. Webster, *The Measurement, Instrumentation, and Sensors Handbook*. Boca Raton, FL: CRC Press, 1999.
38. J. Williams, *Analog Circuit Design*. Wobwin, MA: Butterworth-Heinemann, 1991.

INDEX

Linear Position Sensors: Theory and Application, by David S. Nyce
ISBN 0-471-23326-9