Displacement Measurement, Linear and Angular

Keith Antonelli
Kinetic Sciences Inc.

Viktor P. Astakhov
Concordia University

Amit Bandyopadhyay
Rutgers University

Vikram Bhatia
Virginia Tech

Richard O. Claus
Virginia Tech

David Dayton
ILC Data Device Corp.

Halit Eren
Curtin University of Technology

Robert M. Hyatt, Jr.
Howell Electric Motors

Victor F. Janas
Rutgers University

Nils Karlsson
National Defense Research Establishment

Andrei Kholkine
Rutgers University

James Ko
Kinetic Sciences, Inc.

Wei Ling Kong

Shyan Ku
Kinetic Sciences Inc.

J.R. René Mayer
Ecole Polytechnique de Montreal

David S. Nyce
MTS Systems Corp.

Teklic Ole Pedersen
Linkopings Universitet

Ahmad Safari
Rutgers University

Anbo Wang
Virginia Tech

Grover C. Wetsel
University of Texas at Dallas

Bernhard Günther Zagar
Technical University Graz

6.1 Resistive Displacement Sensors

Keith Antonelli, James Ko, and Shyan Ku

Resistive displacement sensors are commonly termed potentiometers or "pots." A pot is an electromechanical device containing an electrically conductive *wiper* that slides against a fixed *resistive element* according to the position or angle of an external shaft. See Figure 6.1. Electrically, the resistive element is "divided" at the point of wiper contact. To measure displacement, a pot is typically wired in a "voltage divider" configuration, as shown in Figure 6.2. The circuit's output, a function of the wiper's position, is an analog voltage available for direct use or digitization. Calibration maps the output voltage to units of displacement.

Table 6.1 lists some attributes inherent to pots. This chapter describes the different types of pots available, their electrical and mechanical characteristics, and practical approaches to using them for precision measurement. Sources and typical prices are also discussed. Versatile, inexpensive, and easy-to-use, pots are a popular choice for precision measurement.

Precision Potentiometers

Pots are available in great variety, with specific kinds optimized for specific applications. Position measurement requires a high-quality pot designed for extended operation. Avoid pots classified as trimmers, rheostats, attenuators, volume controls, panel controls, etc. Instead, look for *precision potentiometers.*

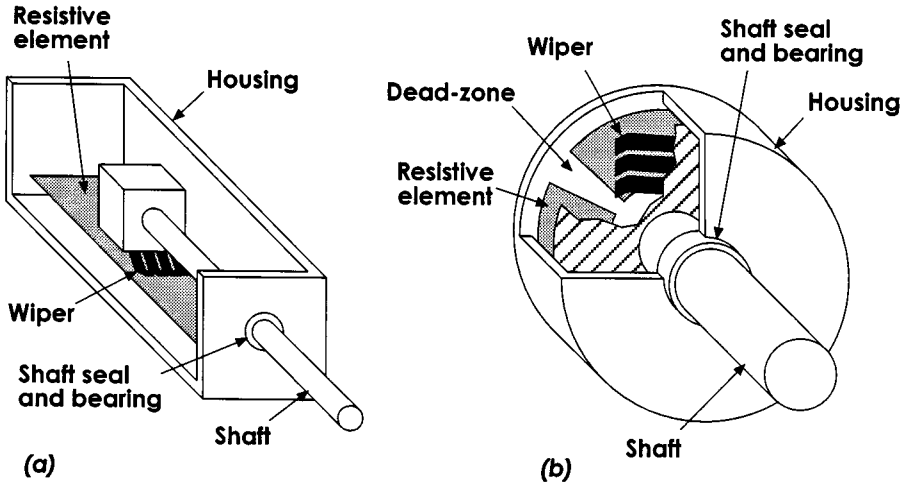


FIGURE 6.1 Representative cutaways of linear-motion (*a*) and rotary (*b*) potentiometers.

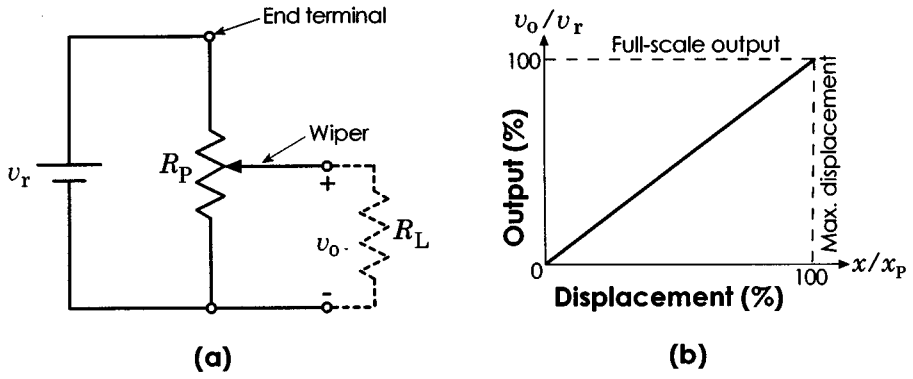


FIGURE 6.2 (*a*) Schematic diagrams depict a potentiometer as a resistor with an arrow representing the wiper. This schematic shows a pot used as a variable voltage divider — the preferred configuration for precision measurement. R_P is the total resistance of the pot, R_L is the load resistance, v_r is the reference or supply voltage, and v_o is the output voltage. (*b*) shows an ideal linear output function where x represents the wiper position, and x_P is its maximum position.

TABLE 6.1 Fundamental Potentiometer Characteristics

Advantages	Disadvantages
Easy to use	Limited bandwidth
Low cost	Frictional loading
Nonelectronic	Inertial loading
High-amplitude output signal	Wear
Proven technology	

Types of Precision Potentiometers

Precision pots are available in *rotary*, *linear-motion*, and *string pot* forms. String pots — also called *cable pots*, *yo-yo pots*, *cable extension transducers*, and *draw wire transducers* — measure the extended length of a spring-loaded cable. Rotary pots are available with single- or multiturn abilities: commonly 3, 5, or 10 turns. Linear-motion pots are available with maximum strokes ranging from roughly 5 mm to over 4 m [1, 2]. String pots are available with maximum extensions exceeding 50 m [3]. Pot manufacturers usually specify a pot's type, dimensions, resistive element composition, electrical and mechanical parameters, and mounting method.

Resistive Element

Broadly, a pot's resistive element can be classified as either *wirewound*, or *nonwirewound*. Wirewound elements contain tight coils of resistive wire that quantize measurement in step-like increments. In contrast, nonwirewound elements present a continuous sheet of resistive material capable of essentially unlimited measurement resolution.

Wirewound elements offer excellent temperature stability and high power dissipation abilities. The coils quantize measurement according to wire size and spacing. Providing the resolution limits are acceptable, wirewound elements can be a satisfactory choice for precision measurement; however, conductive plastic or hybrid elements will usually perform better and for considerably more cycles. These and other popular nonwirewound elements are described in more detail below.

Conductive plastic elements feature a smooth film with unlimited resolution, low friction, low noise, and long operational life. They are sensitive to temperature and other environmental factors and their power dissipation abilities are low; however, they are an excellent choice for most precision measurement applications.

TABLE 6.2 Characteristics of Conductive Plastic, Wirewound, and Hybrid Resistive Elements

	Conductive plastic	Wirewound	Hybrid
Resolution	Infinitesimal	Quantized	Infinitesimal
Power rating	Low	High	Low
Temperature stability	Poor	Excellent	Very good
Noise	Very low	Low, but degrades with time	Low
Life	10^6–10^8 cycles	10^5–10^6 cycles	10^6–10^7 cycles

TABLE 6.3 Potentiometer Terminal Markings

Terminal	Possible color codings			Rotary pot	Linear-motion pot
1	Yellow	Red	Black	CCW limit	Fully retracted limit
2	Red	Green	White	Wiper	Wiper
3	Green	Black	Red	CW limit	Fully extended limit

Hybrid elements feature a wirewound core with a conductive plastic coating, combining wirewound and conductive plastic technologies to realize some of the more desirable attributes of both. The plastic limits power dissipation abilities in exchange for low noise, long life, and unlimited resolution. Like wirewounds, hybrids offer excellent temperature stability. They make an excellent choice for precision measurement.

Cermet elements, made from a ceramic-metal alloy, offer unlimited resolution and reasonable noise levels. Their advantages include high power dissipation abilities and excellent stability in adverse conditions. Cermet elements are rarely applied to precision measurement because conductive plastic elements offer lower noise, lower friction, and longer life.

Carbon composition elements, molded under pressure from a carbon–plastic mixture, are inexpensive and very popular for general use, but not for precision measurement. They offer unlimited resolution and low noise, but are sensitive to environmental stresses (e.g., temperature, humidity) and are subject to wear.

Table 6.2 summarizes the distinguishing characteristics of the preferred resistive elements for precision measurement.

Electrical Characteristics

Before selecting a pot and integrating it into a measurement system, the following electrical characteristics should be considered.

Terminals and Taps

Table 6.3 shows the conventional markings found on the pot housing [4, 5]; CW and CCW indicate clockwise and counter-clockwise rotation as seen from the front end. Soldering studs and eyelets, integral connectors, and flying leads are common means for electrical connection. In addition to the wiper and end terminals, a pot may possess one or more terminals for *taps*. A tap enables an electrical connection to be made with a particular point along the resistive element. Sometimes, a *shunt resistor* is connected to a tap in order to modify the output function. End terminations and taps can exhibit different electrical characteristics depending on how they are manufactured. See [2] for more details.

Taper

Pots are available in a variety of different tapers that determine the shape of the output function. With a linear-taper pot, the output varies linearly with wiper motion, as shown in Figure 6.2. (Note that a pot with a linear taper should not be confused with a linear-motion pot, which is sometimes called a "linear pot.") Linear-taper pots are the most commonly available, and are widely used in sensing and control

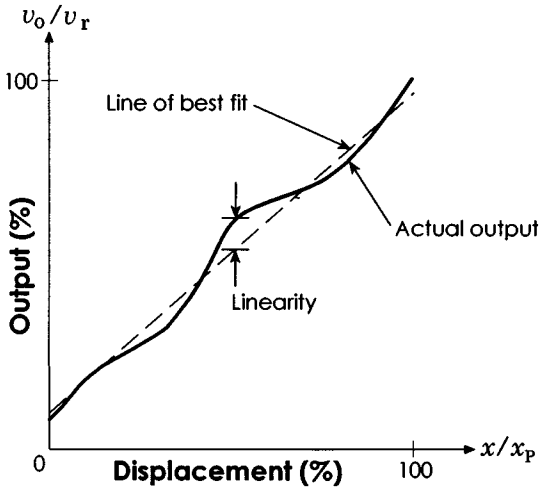


FIGURE 6.3 Independent linearity is the maximum amount by which the actual output function deviates from a line of best fit.

applications. Pots with nonlinear tapers (e.g., logarithmic, sine, cosine, tangent, square, cube) can also be useful, especially where computer control is not involved. Nonstandard tapers can be custom-manufactured or alternatively, certain types of output functions can be produced using shunt resistors, by combining outputs from ganged pots or by other means. (Refer to [6, 7] for more details.) Of course, if a computer is involved, the output function can always be altered through a software lookup table or mapping function.

Electrical Travel

Figure 6.2 shows how the ideal output of a pot changes with wiper position. In practice, there is a small region at both ends where output remains constant until the wiper hits a mechanical stop. *Mechanical travel* is the total motion range of the wiper, and *electrical travel* is the slightly smaller motion range over which the electrical output is "valid." Thus, when using a pot as a sensor, it is important to ensure that the wiper motion falls within the electrical travel limits.

Linearity

Linearity is the maximum deviation of the output function from an ideal straight line. *Independent linearity* is commonly specified, where the straight line is defined as the line that minimizes the linearity error over a series of sampled points, not necessarily measured over the full range of the pot. See Figure 6.3. Other linearity metrics, such as *terminal-based linearity, absolute linearity,* and *zero-based linearity,* are also sometimes used. Refer to [8] for more details. Pots are commonly available with independent linearities ranging from under 0.1% to 1%. When dealing with nonlinear output functions, *conformity* is specified since it is the more general term used to describe deviation from any ideal function. Conformity and linearity are usually expressed as a percentage of full-scale output (FSO).

Electrical Loading

Loading can significantly affect the linearity of measurements, regardless of a pot's quality and construction. Consider an ideal linear pot connected to an infinite load impedance (i.e., as in Figure 6.2). Since no current flows through the load, the output changes perfectly linearly as the wiper travels along the length of the pot. However, if the load impedance is finite, the load draws some current, thereby affecting the output as illustrated in Figure 6.4. Circuit analysis shows that:

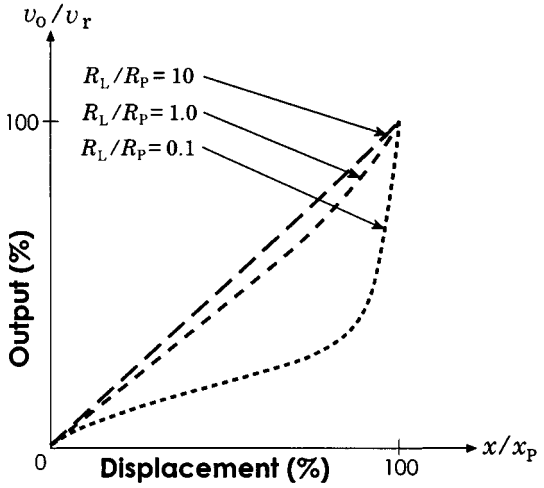


FIGURE 6.4 Linearity can be greatly influenced by the ratio of load resistance, R_L, to potentiometer resistance, R_P.

$$\frac{v_o}{v_r} = \frac{\left(x/x_P\right)\left(R_L/R_P\right)}{\left(R_L/R_P\right)+\left(x/x_P\right)-\left(x/x_P\right)^2} \tag{6.1}$$

Therefore, R_L/R_P should be maximized to reduce loading effects (this also involves other trade-offs, to be discussed). A minimum R_L/R_P value of 10 is sometimes used as a guideline since loading error is then limited to under 1% of full-scale output. Also, some manufacturers recommend a minimum load impedance or maximum wiper current in order to minimize loading effects and prevent damage to the wiper contacts. The following are some additional strategies that can be taken:

- Use a regulated voltage source whose output is stable with load variations
- Use high input-impedance signal conditioning or data acquisition circuitry
- Use only a portion of the pot's full travel

Resolution

Resolution defines the smallest possible change in output that can be produced and detected. In wirewound pots, the motion of the wiper over the coil generates a quantized response. Therefore, the best attainable resolution is $r = (1/N) \times 100\%$, where N is the number of turns in the coil. Nonwirewound pots produce a smooth response with essentially unlimited resolution. Hybrid pots also fall into this category. In practice, resolution is always limited by factors such as:

- Electrical noise, usually specified as *noise* for wirewound pots and *smoothness* for nonwirewound pots, both expressed as a percentage of full-scale output [10]
- Stability of the voltage supply, which can introduce additional noise into the measurement signal
- Analog-to-digital converter (ADC) resolution, usually expressed in "bits" (e.g., 10 mm travel digitized using a 12-bit ADC results in 10 mm/4096 = 0.0024 mm resolution at best)
- Mechanical effects such as stiction

Power Rating

The power dissipated by a pot is $P = v_r^2/R_P$. Therefore, power rating determines the maximum voltage that can be applied to the pot at a given temperature. With greater voltage supplied to the pot, greater

output (and noise) is produced but more power is dissipated, leading to greater thermal effects. In general, wirewound and cermet pots are better able to dissipate heat, and thus have the highest power ratings.

Temperature Coefficient

As temperature increases, pot resistance also increases. However, a pot connected as shown in Figure 6.2 will divide the voltage equally well, regardless of its total resistance. Thus, temperature effects are not usually a major concern as long as the changes in resistance are uniform and the pot operates within its ratings. However, an increase in pot resistance also increases loading nonlinearities. Therefore, temperature coefficients can become an important consideration. The temperature coefficient, typically specified in ppm $^{\circ}C^{-1}$, can be expressed as $\alpha = (\Delta R_P/R_P)/\Delta t$, where Δt is the change in temperature and ΔR_P is the corresponding change in total resistance. In general, wirewound pots possess the lowest temperature coefficients. Temperature-compensating signal-conditioning circuitry can also be used.

Resistance

Since a pot divides voltage equally well regardless of its total resistance, resistance tolerance is not usually a major concern. However, total resistance can have a great impact on loading effects. If resistance is large, less current flows through the pot, thus reducing temperature effects, but also increasing loading.

AC Excitation

Pots can operate using either a dc or an ac voltage source. However, wirewound pots are susceptible to capacitive and inductive effects that can be substantial at moderate to high frequencies.

Mechanical Characteristics

The following mechanical characteristics influence measurement quality and system reliability, and thus should be considered when selecting a pot.

Mechanical Loading

A pot adds inertia and friction to the moving parts of the system that it is measuring. As a result, it increases the force required to move these parts. This effect is referred to as *mechanical loading*. To quantify mechanical loading, rotary pot manufacturers commonly list three values: the equivalent *mass moment of inertia* of the pot's rotating parts, the *dynamic* (or *running*) *torque* required to maintain rotation in a pot shaft, and the *starting torque* required to initiate shaft rotation. For linear-motion pots, the three analogous loading terms are *mass*, *starting force*, and *dynamic* (or *running*) *force*.

In extreme cases, mechanical loading can adversely affect the operating characteristics of a system. When including a pot in a design, ensure that the inertia added to the system is insignificant or that the inertia is considered when analyzing the data from the pot. The starting and running force or torque values might also be considered, although they are generally small due to the use of bearings and low-friction resistive elements.

Mechanical Travel

Distinguished from electrical travel, *mechanical travel* is the wiper's total motion range. A mechanical stop delimits mechanical travel at each end of the wiper's range of motion. Stops can withstand small loads only and therefore should not be used as mechanical limits for the system. Manufacturers list maximum loads as the *static stopping strength* (for static loads) and the *dynamic stopping strength* (for moving loads).

Rotary pots are also available without mechanical stops. The shaft of such an "unlimited travel" pot can be rotated continuously in either direction; however, electrical travel is always less than 360° due to the discontinuity or "dead-zone" where the resistive element begins and ends. (See Figure 6.1.) Multiple revolutions can be measured with an unlimited travel pot in conjunction with a counter: the counter maintains the number of full revolutions while the pot measures subrevolution angular displacement.

Operating Temperature

When operated within its specified temperature range, a pot maintains good electrical linearity and mechanical integrity. Depending on construction, pots can operate at temperatures from as low as –65°C

to as high as 150°C. Operating outside specified limits can cause material failure, either directly from temperature or from thermally induced misalignment.

Vibration, Shock, and Acceleration

Vibration, shock, and acceleration are all potential sources of contact discontinuities between the wiper and the resistive element. In general, a contact failure is considered to be a discontinuity equal to or greater than 0.1 ms [2]. The values quoted in specification sheets are in *g*s and depend greatly on the particular laboratory test. Some characterization tests use sinusoidal vibration, random vibration, sinusoidal shock, sawtooth shock, or acceleration to excite the pot. Manufacturers use mechanical design strategies to eliminate weaknesses in a pot's dynamic response. For example, one technique minimizes vibration-induced contact discontinuities using multiple wipers of differing resonant frequencies.

Speed

Exceeding a pot's specified maximum speed can cause premature wear or discontinuous values through effects such as wiper bounce. As a general rule, the slower the shaft motion, the longer the unit will last (in total number of cycles). Speed limitations depend on the materials involved. For rotary pots, wirewound models have preferred maximum speeds on the order of 100 rpm, while conductive plastic models have allowable speeds as high as 2000 rpm. Linear-motion pots have preferred maximum velocities up to 10 m s^{-1}.

Life

Despite constant mechanical wear, a pot's expected lifetime is on the order of a million cycles when used under proper conditions. A quality film pot can last into the hundreds of millions of cycles. Of wirewound, hybrid, and conductive plastic pots, the uneven surface of a wirewound resistive element inherently experiences the most wear and thus has the shortest expected operating life. Hybrids improve on this by using a wirewound construction in combination with a smooth conductive film coating. Conductive plastic pots generally have the longest life expectancy due to the smooth surface of their resistive element.

Contamination and Seals

Foreign material contaminating pots can promote wear and increase friction between the wiper and the resistive element. Consequences range from increased mechanical loading to outright failure (e.g., seizing, contact discontinuity). Fortunately, sealed pots are available from most manufacturers for industrial applications where dirt and liquids are often unavoidable. To aid selection, specifications often include the type of *case sealing* (i.e., mechanisms and materials) and the *seal resistance* to cleaning solvents and other commonly encountered fluids.

Misalignment

Shaft misalignment in a pot can prematurely wear its bearing surfaces and increase its mechanical loading effects. A good design minimizes misalignment. (See *Implementation*, below.) Manufacturers list a number of alignment tolerances. In linear-motion pots, *shaft misalignment* is the maximum amount a shaft can deviate from its axis. The degree to which a shaft can rotate around its axis is listed under *shaft rotation*. In rotary pots, *shaft end play* and *shaft radial play* both describe the amount of shaft deflection due to a radial load. *Shaft runout* denotes the shaft diameter eccentricity when a shaft is rotated under a radial load.

Mechanical Mounting Methods

Hardware features on a pot's housing determine the mounting method. Options vary with manufacturer, and among rotary, linear-motion, and string pots. Offerings include custom bases, holes, tabs, flanges, and brackets — all of which secure with machine screws — and threaded studs, which secure with nuts. Linear-motion pots are available with rod or slider actuation, some with internal or external return springs. Mounting is typically accomplished by movable clamps, often supplied by the pot manufacturer. Other linear-motion pots mount via a threaded housing. For rotary pots, the two most popular mounting methods are the *bushing mount* and the *servo mount*. See Figure 6.5.

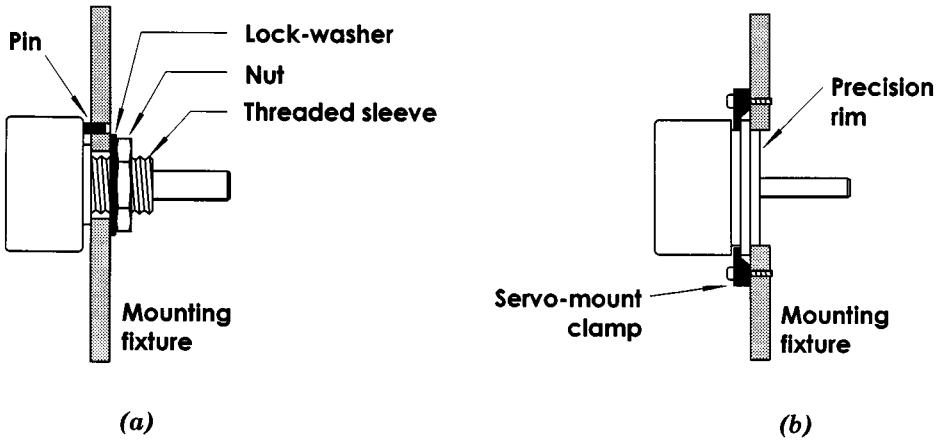


FIGURE 6.5 The two most common rotary pot mounts are the bushing mount (*a*), and the servo mount (*b*).

TABLE 6.4 Sources of Small Mechanical Components

PIC Design
86 Benson Road, P.O. Box 1004
Middlebury, CT 06762-1004
Tel: (800) 243-6125, (203) 758-8272; Fax: (203) 758-8271
www.penton.com/md/mfg/pic/

Stock Drive Products/Sterling Instrument
2101 Jericho Turnpike, Box 5416
New Hyde Park, NY 11042-5416
Tel: (516) 328-3300; Fax: (800) 737-7436, (516) 326-8827
www.sdp-si.com

W.M. Berg, Inc.
499 Ocean Ave.
East Rockaway, NY 11518
Tel: (800) 232-2374, (516) 599-5010; Fax: (800) 455-2374, (516) 599-3274
www.wmberg.com

Bushing mount

The pot provides a shaft-concentric, threaded sleeve that invades a hole in a mounting fixture and secures with a nut and lock-washer. An off-axis tab or pin prevents housing rotation. Implementing a bushing mount requires little more than drilling a hole; however, limited rotational freedom and considerable play before tightening complicate precise setup.

Servo mount

The pot provides a flanged, shaft-concentric, precision-machined rim that slips into a precision-bored hole in a mounting fixture. The flange secures with symmetrically arranged, quick-releasing *servo mount clamps*, available from Timber-Top, Inc. [9] and also from the sources listed in Table 6.4. (These clamps are also called *synchro mount clamps* and *motor mount cleats*, since servo-mounting synchros and stepper motors are also available.) Servo mounts are precise and easy to adjust, but entail the expense of precision machining.

Measurement Techniques

To measure displacement, a pot must attach to mechanical fixtures and components. The housing typically mounts to a stationary reference frame, while the shaft couples to a moving element. The *input motion* (i.e., the motion of interest) can couple directly or indirectly to the pot's shaft. A direct connection, although straightforward, carries certain limitations:

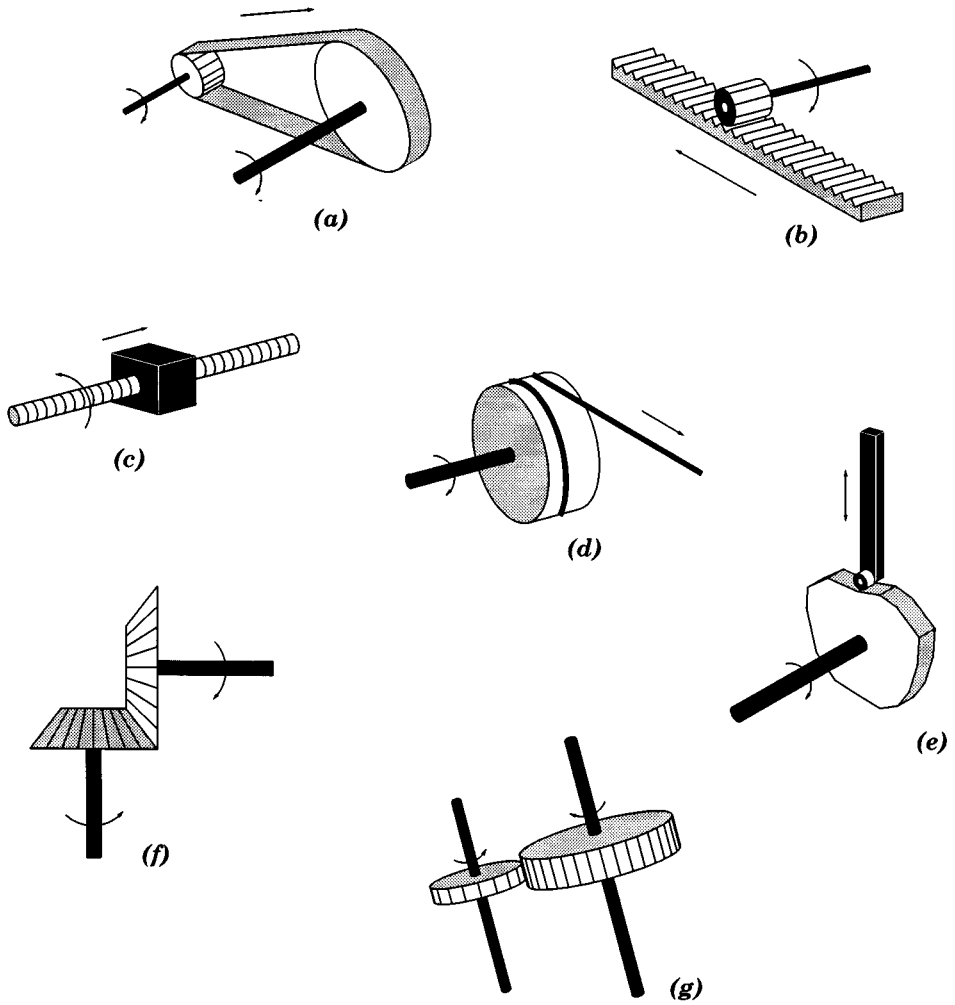


FIGURE 6.6 Mechanisms that extend a precision potentiometer's capabilities include belts and pulleys (*a*), rack-and-pinions (*b*), lead-screws (*c*), cabled drums (*d*), cams (*e*), bevel gears (*f*), and spur gears (*g*).

- The input motion maps 1:1 to the shaft motion
- The input motion cannot exceed the pot's mechanical travel limits
- Angle measurement requires a rotary pot; position measurement requires a linear-motion pot
- The pot must mount close to the motion source
- The input motion must be near-perfectly collinear or coaxial with the shaft axis

Figure 6.6 shows ways to overcome these limitations. Mechanisms with a mechanical advantage scale motion and adjust travel limits. Mechanisms that convert between linear and rotary motion enable any type of pot to measure any kind of motion. Transmission mechanisms distance a pot from the measured motion. Compliant mechanisms compensate for misalignment. Examples and more details follow. Most of the described mechanisms can be realized with components available from the sources in Table 6.4.

Gears scale the mapping between input and pot shaft motions according to gear ratio. They also displace rotation axes to a parallel or perpendicular plane according to type of gear (e.g., spur vs. bevel). Gears introduce backlash. Friction rollers are a variation on the gear theme, immune to backlash but prone to slippage. The ratio of roller diameters scales the mapping between input and pot shaft motions.

Rack-and-pinion mechanisms convert between linear and rotary motion. Mapping is determined by the rack's *linear pitch* (i.e., tooth-to-tooth spacing) compared to the number of teeth on the pinion. Backlash is inevitable.

Lead-screws convert rotary motion to linear motion via the screw principle. Certain low-friction types (e.g., ball-screws) are also capable of the reverse transformation (i.e., linear to rotary). Either way, mapping is controlled by the screw's *lead* — the distance the nut travels in one revolution. Lead-screws are subject to backlash.

Cabled drums convert between linear and rotary motion according to the drum circumference, since one turn of diameter D wraps or unwraps a length πD of cable. An external force (e.g., supplied by a spring or a weight) might be necessary to maintain cable tension.

Pulleys can direct a string pot's cable over a complex path to a motion source. Mapping is 1:1 unless the routing provides a mechanical advantage.

Pulleys and *belts* transmit rotary motion scaled according to relative pulley diameters. The belt converts between linear and rotary motion. (See Figure 6.6(a)) The empty area between pulleys provides a convenient passageway for other components. Sprocket wheels and chain have similar characteristics. Matched pulley-belt systems are available that operate with negligible slip, backlash, and stretch.

Cams map rotary motion into linear motion according to the function "programmed" into the cam profile. See [10] for more information.

Linkages can be designed to convert, scale, and transmit motion. Design and analysis can be quite complex and mapping characteristics tend to be highly nonlinear. See [11] for details.

Flexible shafts transmit rotary motion between two non-parallel axes with a 1:1 mapping, subject to torsional windup and hysteresis if the motion reverses.

Conduit, like a bicycle brake cable or *Bowden cable*, can route a cable over an arbitrary path to connect a pot to a remote motion source. The conduit should be incompressible and fixed at both ends. Mapping is 1:1 with some mechanical slop. Lubrication helps mitigate friction.

A mechanism's mapping characteristics impact measurement resolution and accuracy. Consider a stepper motor turning a lead-screw to translate a nut. A linear-motion pot could measure the nut's position directly to some resolution. Alternatively, a rotary pot linked to the lead-screw could measure the position with increased resolution if the mechanism mapped the same amount of nut travel to considerably more wiper travel. Weighing the resolution increase against the uncertainty due to backlash would determine which approach was more accurate.

Implementation

Integrating a pot into a measurement system requires consideration of various design issues, including the impact of the pot's physical characteristics, error sources, space restrictions, and wire-routing. The pot's shaft type and bearings must be taken into consideration and protected against excessive loading. A good design will:

- Give the pot mount the ability to accommodate minor misalignment
- Protect the shaft from thrust, side, and bending loads (i.e., not use the pot as a bearing)
- Provide hard limit stops within the pot's travel range (i.e., not use the pot's limit stops)
- Protect the pot from contaminants
- Strain-relieve the pot's electrical connections

A thorough treatment of precision design issues appears in [12].

Coupling to the Pot

Successful implementation also requires practical techniques for mechanical attachment. A string pot's cable terminator usually fastens to other components with a screw. For other types of pots, coupling technique is partly influenced by the nature of the shaft. Rotary shafts come with various endings, including plain, single-flatted, double-flatted, slotted, and knurled. Linear-motion shafts usually terminate in threads, but are also available with roller ends (to follow surfaces) or with spherical bearing ends (to accommodate misalignment). With care, a shaft can be cut, drilled, filed, threaded, etc.

In a typical measurement application, the pot shaft couples to a mechanical component (e.g., a gear, a pulley), or to another shaft of the same or different diameter. Successful couplings provide a positive link to the shaft without stressing the pot's mechanics. Satisfying these objectives with rotary and linear-motion pots requires a balance between careful alignment and compliant couplings. Alignment is not as critical with a string pot. Useful coupling methods include the following.

Compliant couplings. It is generally wise to put a compliant coupling between a pot's shaft and any other shafting. A compliant coupling joins two misaligned shafts of the same or different diameter. Offerings from the companies in Table 6.4 include bellows couplings, flex couplings, spring couplings, spider couplings, Oldham couplings, wafer spring couplings, flexible shafts, and universal joints. Each type has idiosyncrasies that impact measurement error; manufacturer catalogs provide details.

Sleeve couplings. Less expensive than a compliant coupling, a rigid sleeve coupling joins two shafts of the same or different diameter with the requirement that the shafts be perfectly aligned. Perfect alignment is difficult to achieve initially, and impossible to maintain as the system ages. Imperfect alignment accelerates wear and risks damaging the pot. Sleeve couplings are available from the companies listed in Table 6.4.

Press fits. A press fit is particularly convenient when the bore of a small plastic part is nominally the same as the shaft diameter. Carefully force the part onto the shaft. Friction holds the part in place, but repeated reassembly will compromise the fit.

Shrink fits. Components with a bore slightly under the shaft diameter can be heated to expand sufficiently to slip over the shaft. A firm grip results as the part cools and the bore contracts.

Pinning. Small hubbed components can be pinned to a shaft. The pin should extend through the hub partway into the shaft, and the component should fit on the shaft without play. Use roll pins or spiral pins combined with a thread-locking compound (e.g., Loctite 242).

Set-screws. Small components are available with hubs that secure with set-screws. The component should fit on the shaft without play. For best results, use two set-screws against a shaft with perpendicular flats. Dimple a plain shaft using the component's screw hole(s) as a drill guide. Apply a thread-locking compound (e.g., Loctite 242) to prevent the set-screws from working loose.

Clamping. Small components are also available with split hubs that grip a shaft when squeezed by a matching hub clamp. Clamping results in a secure fit without marring the shaft.

Adhesives. Retaining compounds (e.g., Loctite 609) can secure small components to a shaft. Follow manufacturer's instructions for best results.

Spring-loaded contact. A spring-loaded shaft will maintain positive contact against a surface that moves at reasonable speeds and without sudden acceleration.

Costs and Sources

Precision pots are inexpensive compared to other displacement measurement technologies. Table 6.5 lists approximate costs for off-the-shelf units in single quantity. Higher quality generally commands a higher price; however, excellent pots are often available at bargain prices due to volume production or surplus conditions. Electronic supply houses offer low-cost pots (i.e., under $20) that can suffice for short-term projects. Regardless of price, always check the manufacturer's specifications to confirm a pot's suitability for a given application.

Table 6.6 lists several sources of precision pots. Most manufacturers publish catalogs, and many have Web sites. In addition to a standard product line, most manufacturers will custom-build pots for high-volume applications.

TABLE 6.5 Typical Single-quantity Prices ($US) for Commercially Available Pots

Potentiometer type	Approximate price range
Rotary	$10–$350
Linear-motion	$20–$2000
String	$250–$1000

TABLE 6.6 Sources of Precision Pots

Company	Potentiometer types
Betatronix, Inc. 110 Nicon Court Hauppauge, NY 11788 Tel: (516) 582-6740; Fax (516) 582-6038 www.betatronix.com	Exotic linear-motion, rotary
BI Technologies Corp. 4200 Bonita Place Fullerton, CA 92635 Tel: (714) 447-2345; Fax: (714) 447-2500	Rotary
Bourns, Inc. Sensors & Controls Division 2533 N. 1500 W. Ogden, UT 84404 Tel: (801) 786-6200; Fax: (801) 786-6203 www.bourns.com	Mostly rotary, some linear-motion
Celesco Transducer Products, Inc. 7800 Deering Avenue Canoga Park, CA 91309 Tel: (800) 423-5483, (818) 884-6860 Fax: (818) 340-1175 www.celesco.com	String
Data Instruments 100 Discovery Way Acton, MA 01720-3648 Tel: (800) 333-3282, (978) 264-9550 Fax: (978) 263-0630 www.datainstruments.com	Linear-motion, rotary
Duncan Electronics Division BEI Sensors & Systems Company 15771 Red Hill Avenue Tustin, CA 92680 Tel: (714) 258-7500; Fax: (714) 258-8120 www.beisensors.com	Linear-motion, rotary
Dynamation Transducers Corp. 348 Marshall Street Holliston, MA 01746-1441 Tel: (508) 429-8440; Fax: (508) 429-1317	Linear-motion, rotary
JDK Controls, Inc. 424 Crown Pt. Circle Grass Valley, CA 95945 Tel: (530) 273-4608; Fax: (530) 273-0769	Rotary, “do-it-yourself” rotor/wiper assemblies
Midori America Corp. 2555 E. Chapman Ave, Suite 400 Fullerton, CA 92631 Tel: (714) 449-0997; Fax: (714) 449-0139 www.midori.com	Linear-motion, rotary, string; also magneto-resistive
New England Instrument 245 Railroad Street Woonsocket, RI 02895-1129 Tel: (401) 769-0703; Fax: (401) 769-0037	Linear-motion, rotary, resistive elements
Novotechnik U.S., Inc. 237 Cedar Hill Street Marlborough, MA 01752 Tel: (800) 667-7492, (508) 485-2244 Fax: (508) 485-2430 www.novitechnik.com	Linear-motion, rotary

TABLE 6.6 (continued) Sources of Precision Pots

Company	Potentiometer types
Servo Systems, Co. 115 Main Road, PO Box 97 Montville, NJ 07045-0097 Tel: (800) 922-1103, (973) 335-1007 Fax: (973) 335-1661 www.servosystems.com	Linear-motion, rotary (surplus)
SpaceAge Control, Inc. 38850 20th Street East Palmdale, CA 93550 Tel: (805) 273-3000; Fax: (805) 273-4240 www.spaceagecontrol.com	String
Spectrol Electronics Corp. 4051 Greystone Drive Ontario, CA 91761 Tel: (909) 923-3313; Fax: (909) 923-6765	Rotary
UniMeasure, Inc. 501 S.W. 2nd Street Corvallis, OR 97333 Tel: (541) 757-3158 Fax: (541) 757-0858 www.unimeasure.com	String
Axsys Technologies, Inc. Vernitron Sensor Systems Division Precision Potentiometer Division 2800 Anvil Street North St. Petersburg, FL 33710 Tel: (813) 347-2181; Fax: (813) 347-7520 www.axsys.com	Linear-motion, rotary

Evaluation

Precision pots are a mature technology, effectively static except for occasional developments in materials, packaging, manufacturing, etc. Recent potentiometric innovations — including momentary-contact membrane pots [13] and solid-state digital pots [14] — are unavailing to precision measurement.

The variable voltage divider is the traditional configuration for precision measurement. The circuit's output, a high-amplitude dc voltage, is independent of variations in the pot's total resistance, and is highly compatible with other circuits and systems. Other forms of output are possible with a precision pot configured as a variable resistor. For example, paired with a capacitor, a pot could supply a position-dependent *RC* time constant to modulate an oscillator's output frequency or duty cycle. In this setup, the pot's stability and ac characteristics would be important.

An alternative resistive displacement sensing technology is the *magneto-resistive potentiometer*, available in rotary and linear-motion forms. Magneto-resistive pots incorporate a noncontacting, permanent magnet "wiper" that rides above a pair of magneto-resistive elements. The elements, configured as a voltage divider, change their resistances according to the strength of the applied magnetic field, and thus divide the voltage as a function of the magnet's position. The output is approximately linear over a limited range of motion (e.g., 90° in rotary models). Magneto-resistive pots offer unlimited resolution and exceptionally long life, but may require temperature compensation circuitry. See [15] for more information.

References

1. Bourns, Inc., *Electronic Components RC4 Solutions Guide*, 1995, 304.
2. Vernitron Motion Control Group, (New York, NY) *Precision Potentiometers*, Catalog #752, 1993.

3. UniMeasure, Inc., *Position & Velocity Transducers*, UniMeasure document No. 400113-27A, Corvallis, OR.
4. Instrument Society of America (Research Triangle Park, NC), *ISA-S37.12-1977 (R1982) Specifications and Tests for Potentiometric Displacement Transducers*, 1982.
5. E. C. Jordan (Ed.), *Reference Data for Engineers: Radio, Electronics, Computer, and Communications,* 7th ed., Indianapolis: H.W. Sams, 1985, 5–16.
6. D. C. Greenwood, *Manual of Electromechanical Devices*, New York: McGraw-Hill, 1965, 297–299.
7. E. S. Charkey, *Electromechanical System Components*, New York: Wiley-Interscience, 1972, 302–303.
8. Variable Resistive Components Institute (Vista, CA), *VRCI-P-100A Standard for Wirewound and Nonwirewound Precision Potentiometers*, 1988.
9. Timber-Top, Inc., P.O. Box 517, Watertown, CT 06795, Tel: (860)-274-6706; Fax (860)-274-8041.
10. J. Angeles and C. S. López-Cajún, *Optimization of Cam Mechanisms*, Dordrecht, The Netherlands: Kluwer Academic, 1991.
11. P. W. Jensen, *Classical and Modern Mechanisms for Engineers and Inventors*, New York: Marcel Dekker, 1991.
12. A. H. Slocum, *Precision Machine Design*, Englewood Cliffs, NJ: Prentice Hall, 1992.
13. Spectra Symbol Inc., data sheet: *SoftPot® (Membrane Potentiometer),* Salt Lake City, UT, 1996.
14. Dallas Semiconductor Corp., *Digital Potentiometer Overview*, web page: www.dalsemi.com/Prod_info/Dig_Pots/, December 1997.
15. Midori America Corp., (Fullerton, CA) *Midori Position Sensors 1995 Catalog.*

6.2 Inductive Displacement Sensors

Halit Eren

Inductive sensors are widely used in industry in many diverse applications. They are robust and compact, and are less affected by environmental factors (e.g., humidity, dust) in comparison to their capacitive counterparts.

Inductive sensors are primarily based on the principles of magnetic circuits. They can be classified as self-generating or passive. The self-generating types utilize an electrical generator principle; that is, when there is a relative motion between a conductor and a magnetic field, a voltage is induced in the conductor. Or, a varying magnetic field linking a stationary conductor produces voltage in the conductor. In instrumentation applications, the magnetic field may be varying with some frequency and the conductor may also be moving at the same time. In inductive sensors, the relative motion between field and conductor is supplied by changes in the measurand, usually by means of some mechanical motion. On the other hand, the passive transducer requires an external source of power. In this case, the action of the transducer is simply the modulation of the excitation signal.

For the explanation of the basic principles of inductive sensors, a simple magnetic circuit is shown in Figure 6.7. The magnetic circuit consists of a core, made from a ferromagnetic materia,l with a coil of n number of turns wound on it. The coil acts as a source of magnetomotive force (mmf) which drives the flux Φ through the magnetic circuit. If one assumes that the air gap is zero, the equation for the magnetic circuit can be expressed as:

$$\text{mmf} = \text{Flux} \times \text{Reluctance} = \Phi \times \Re \quad \text{A-turns} \tag{6.2}$$

such that the reluctance $\Re$ limits the flux in a magnetic circuit just as resistance limits the current in an electric circuit. By writing the mmf in terms of current, the magnetic flux may be expressed as:

$$\Phi = ni/\Re \quad \text{weber} \tag{6.3}$$

In Figure 6.7, the flux linking a single turn is by Equation 6.3; but the total flux linking by the entire n number of the turns of the coil is

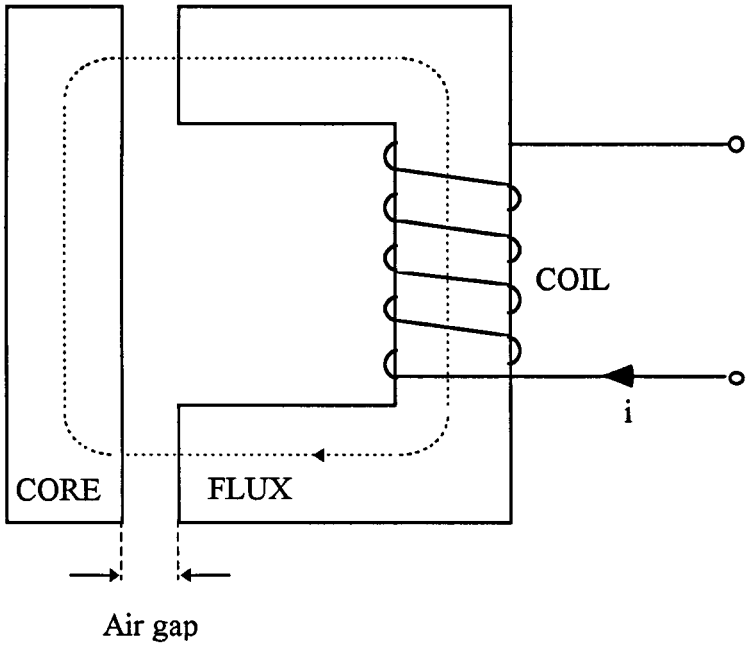


FIGURE 6.7 A basic inductive sensor consists of a magnetic circuit made from a ferromagnetic core with a coil wound on it. The coil acts as a source of magnetomotive force (mmf) that drives the flux through the magnetic circuit and the air gap. The presence of the air gap causes a large increase in circuit reluctance and a corresponding decrease in the flux. Hence, a small variation in the air gap results in a measurable change in inductance.

$$\Psi = n\Phi = n^2 i/\Re \qquad \text{weber} \tag{6.4}$$

Equation 6.4 leads to self inductance L of the coil, which is described as the total flux (Ψ weber) per unit current for that particular coil; that is

$$L = \Psi/I = n^2/\Re \tag{6.5}$$

This indicates that the self inductance of an inductive element can be calculated by magnetic circuit properties. Expressing $\Re$ in terms of dimensions as:

$$\Re = l/\mu\mu_0 A \tag{6.6}$$

where l = the total length of the flux path
μ = the relative permeability of the magnetic circuit material
μ_0 = the permeability of free space ($= 4\pi \times 10^{-7}$ H/m)
A = the cross-sectional area of the flux path

The arrangement illustrated in Figure 6.7 becomes a basic inductive sensor if the air gap is allowed to vary. In this case, the ferromagnetic core is separated into two parts by the air gap. The total reluctance of the circuit now is the addition of the reluctance of core and the reluctance of air gap. The relative permeability of air is close to unity, and the relative permeability of the ferromagnetic material is of the order of a few thousand, indicating that the presence of the air gap causes a large increase in circuit reluctance and a corresponding decrease in the flux. Hence, a small variation in the air gap causes a measurable change in inductance. Most of the inductive transducers are based on these principles and are discussed below in greater detail.

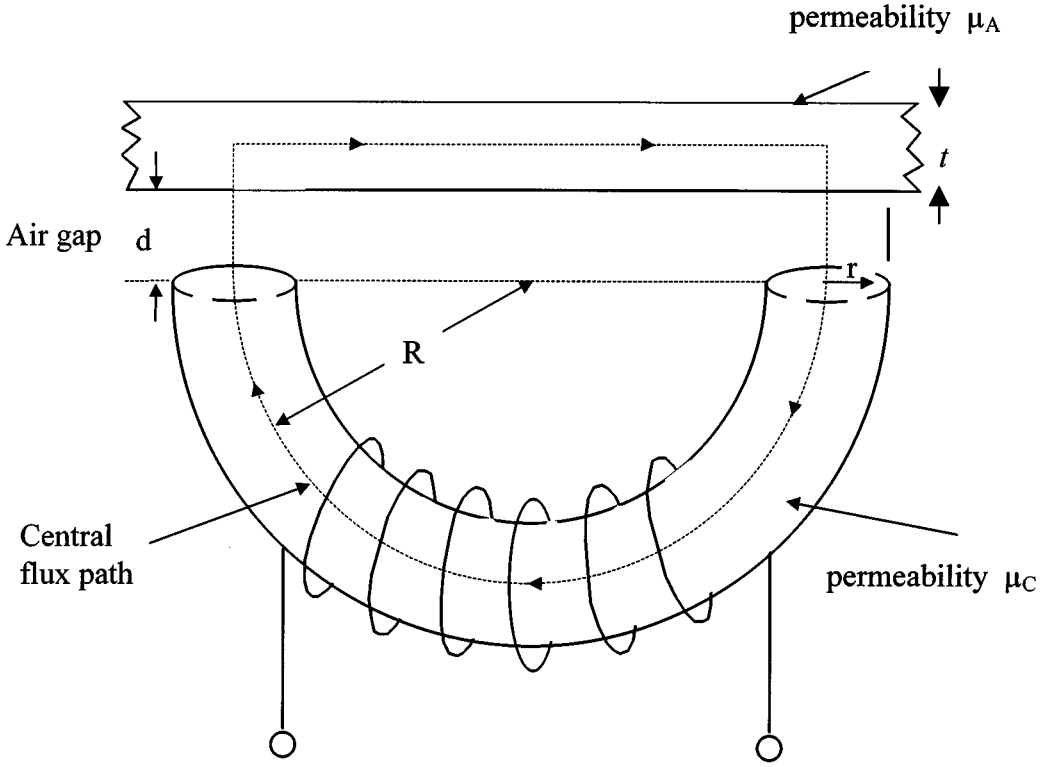


FIGURE 6.8 A typical single-coil, variable-reluctance displacement sensor. The sensor consists of three elements: a ferromagnetic core in the shape of a semicircular ring, a variable air gap, and a ferromagnetic plate. The reluctance of the coil is dependent on the single variable. The reluctance increases nonlinearly with increasing gap.

Linear and Rotary Variable-Reluctance Transducer

The variable-reluctance transducers are based on change in the reluctance of a magnetic flux path. This type of transducer finds application particularly in acceleration measurements. However, they can be constructed to be suitable for sensing displacements as well as velocities. They come in many different forms, as described below.

The Single-Coil Linear Variable-Reluctance Sensor

A typical single-coil variable-reluctance displacement sensor is illustrated in Figure 6.8. The sensor consists of three elements: a ferromagnetic core in the shape of a semicircular ring, a variable air gap, and a ferromagnetic plate. The total reluctance of the magnetic circuit is the sum of the individual reluctances:

$$\Re_T = \Re_C + \Re_G + \Re_A \tag{6.7}$$

where $\Re_C$, $\Re_G$, and $\Re_A$ are the reluctances of the core, air gap, and armature, respectively.

Each one of these reluctances can be determined by using the properties of materials involved, as in Equation 6.6. In this particular case, the reluctance $\Re_T$ can be approximated as:

$$\Re_T = R/\mu_C \mu_0 r^2 + 2d/\mu_0 \pi r^2 + R/\mu_A \mu_0 rt \tag{6.8}$$

In obtaining Equation 6.8, the length of flux path in the core is taken as πR. The cross-sectional area is assumed to be uniform, with a value of πr^2. The total length of the flux path in air is $2d$, and it is assumed that there is no fringing or bending of the flux through the air gap, such that the cross-sectional area of the flux path in air will be close to that of the cross section of the core. The length of an average central flux path in the armature is $2R$. The calculation of the appropriate cross section area of the armature is difficult, but it may be approximated to $2rt$, where t is the thickness of the armature.

In Equation 6.8 all of the parameters are fixed except for the one independent variable — the air gap. Hence, it can be simplified as:

$$\Re_T = \Re_0 + kd \tag{6.9}$$

where $\Re_0 = R/\mu_0\, r\,[1/\mu_C r + 1/\mu_A t]$, and
$k = 2/\mu_0\, \pi r^2$

Using Equations 6.5 and 6.9, the inductance can be written as:

$$L = n^2 \Big/ \left(\Re_0 + kd\right) = L_0 \Big/ \left(1 + \alpha d\right) \tag{6.10}$$

where L_0 = the inductance at zero air gap
$\alpha = k/\Re_0$

The values of L_0 and α can be determined mathematically: they depend on the core geometry, permeability, etc., as explained above. It can be seen from Equation 6.10 that the relationship between L and α is nonlinear. Despite this nonlinearity, these types of single coil sensors find applications in some areas, such as force measurements and telemetry. In force measurements, the resultant change in inductance can be made to be a measure of the magnitude of the applied force. The coil usually forms one of the components of an LC oscillator, for which the output frequency varies with the applied force. Hence, the coil modulates the frequency of the local oscillator.

The Variable-Differential Reluctance Sensor

The problem of the nonlinearity can be overcome by modifying the single coil system into a variable-differential reluctance sensor (also known as push-pull sensor), as shown in Figure 6.9. This sensor consists of an armature moving between two identical cores, and separated by a fixed distance of $2d$. Now, Equation 6.10 can be written for both coils as:

$$\begin{aligned} L_1 &= L_{01} \Big/ \left[1 + \alpha\left(d - x\right)\right], \\ L_2 &= L_{02} \Big/ \left[1 + \alpha\left(d + x\right)\right] \end{aligned} \tag{6.11}$$

Although the relationship between L_1 and L_2 is still nonlinear, the sensor can be incorporated into an ac deflection bridge to give a linear output for small movements. The hysteresis errors of these transducers are almost entirely limited to the mechanical components. These sensors respond to both static and dynamic measurements. They have continuous resolution and high outputs, but they may give erratic performance in response to external magnetic fields. A typical sensor of this type has an input span of 1 cm, a coil inductance of 25 mH, and a coil resistance of 75 Ω. The resistance of the coil must be carefully considered when designing oscillator circuits. The maximum nonlinearity is 0.5%.

A typical commercially available variable differential sensor is shown in Figure 6.10. The iron core is located halfway between the two E-shaped frames. The flux generated by primary coils depends on the reluctance of the magnetic path, the main reluctance being the air gap. Any motion of the core increases the air gap on one side and decreases it on the other side, thus causing reluctance to change, in accordance

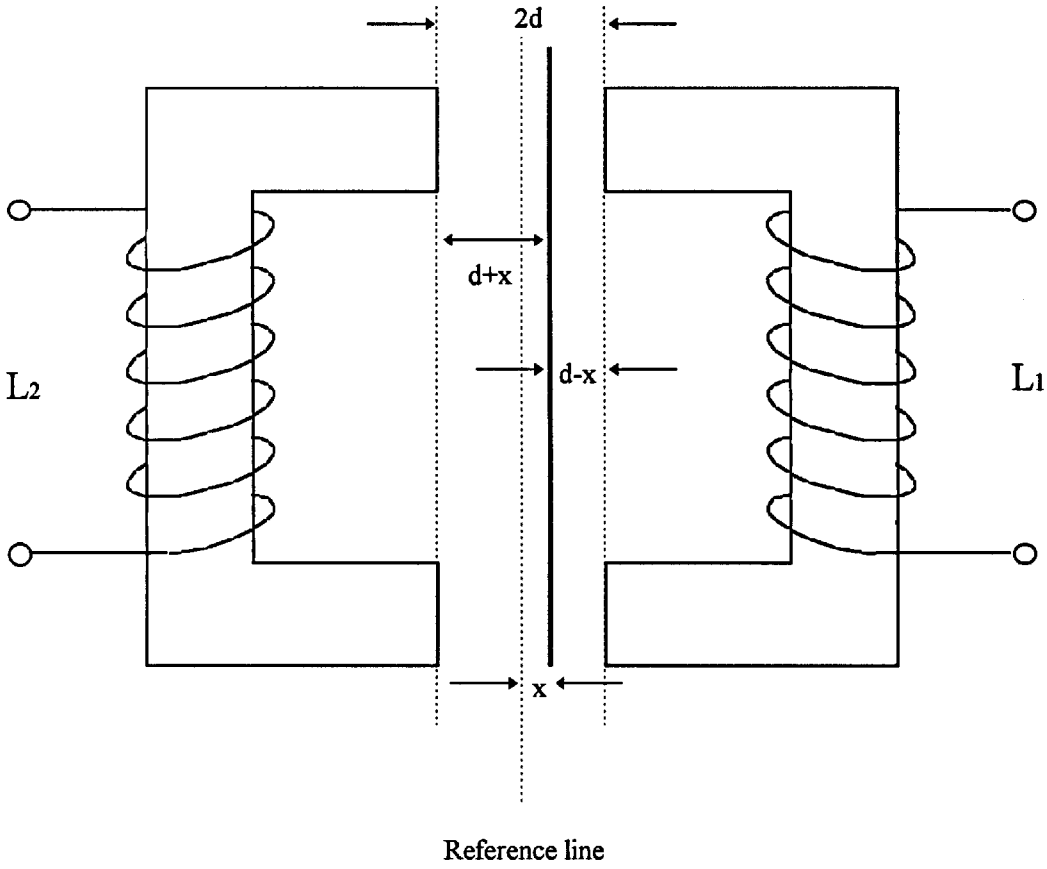


FIGURE 6.9 A variable-differential reluctance sensor consists of an armature moving between two identical cores separated by a fixed distance. The armature moves in the air gap in response to a mechanical input. This movement alters the reluctance of coils 1 and 2, thus altering their inductive properties. This arrangement overcomes the problem of nonlinearity inherent in single coil sensors.

with the principles explained above, and thereby inducing more voltage on one of the coils than on the other. Motion in the other direction reverses the action with a 180° phase shift occurring at null. The output voltage can be modified, depending on the requirements in signal processing, by means of rectification, demodulation, or filtering. In these instruments, full-scale motion may be extremely small — on the order of few thousandths of a centimeter.

In general, variable reluctance transducers have small ranges and are used in specialized applications such as pressure transducers. Magnetic forces imposed on the armature are quite large and this severely limits their application. However, the armature can be constructed as a diaphragm; hence, suitable for pressure measurements.

Variable-Reluctance Tachogenerators

Another example of a variable reluctance sensor is shown in Figure 6.11. These sensors are based on Faraday's law of electromagnetic induction; therefore, they may also be referred to as electromagnetic sensors. Basically, the induced emf in the sensor depends on the linear or angular velocity of the motion.

The variable-reluctance tachogenerator consists of a ferromagnetic, toothed wheel attached to a rotating shaft, and a coil wound onto a permanent magnet, extended by a soft iron pole piece. The wheel moves in close proximity to the pole piece, causing the flux linked by the coil to change, thus inducing an emf in the coil. The reluctance of the circuit depends on the width of the air gap between the rotating wheel and the pole piece. When the tooth is close to the pole piece, the reluctance is minimum and it

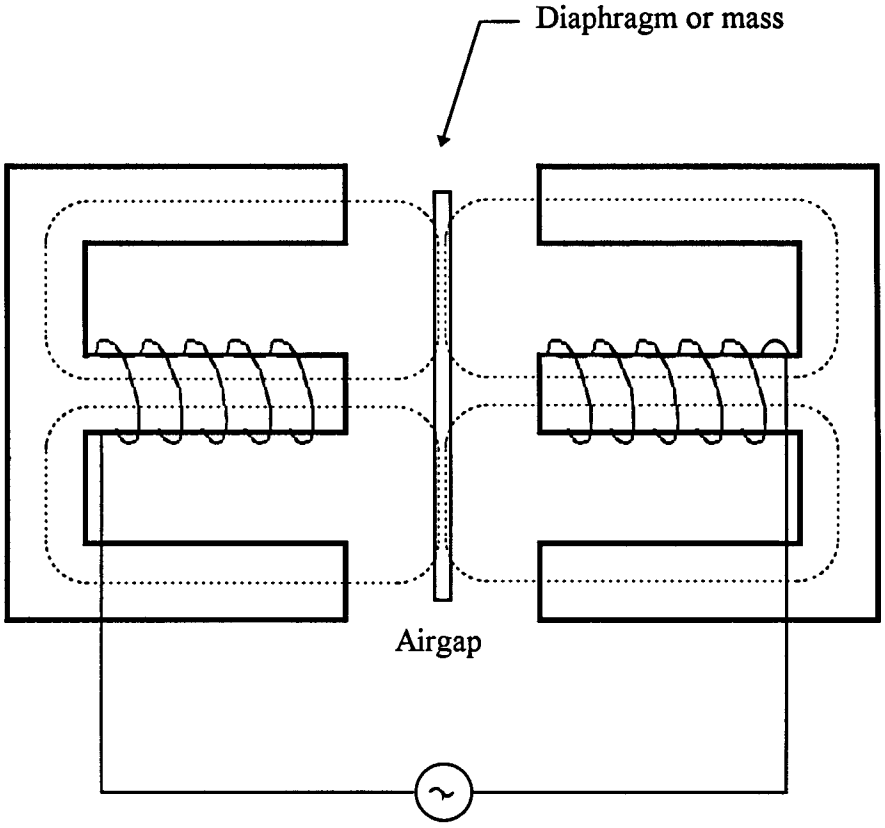


FIGURE 6.10 A typical commercial variable differential sensor. The iron core is located half-way between the two E frames. Motion of the core increases the air gap for one of the E frames while decreasing the other side. This causes reluctances to change, thus inducing more voltage on one side than the other. Motion in the other direction reverses the action, with a 180° phase shift occurring at null. The output voltage can be processed, depending on the requirements, by means of rectification, demodulation, or filtering. The full-scale motion may be extremely small, on the order of few thousandths of a centimeter.

increases as the tooth moves away from the pole. When the wheel rotates with a velocity ω, the flux may mathematically be expressed as:

$$\Psi(\theta) = A + B\cos m\theta \tag{6.12}$$

where A = the mean flux
B = the amplitude of the flux variation
m = the number of teeth

The induced emf is given by:

$$E = -d\Psi(\theta)/dt = -\left(d\Psi(\theta)/d\theta\right) \times \left(d\theta/dt\right) \tag{6.13}$$

or

$$E = bm\omega \sin m\omega t \tag{6.14}$$

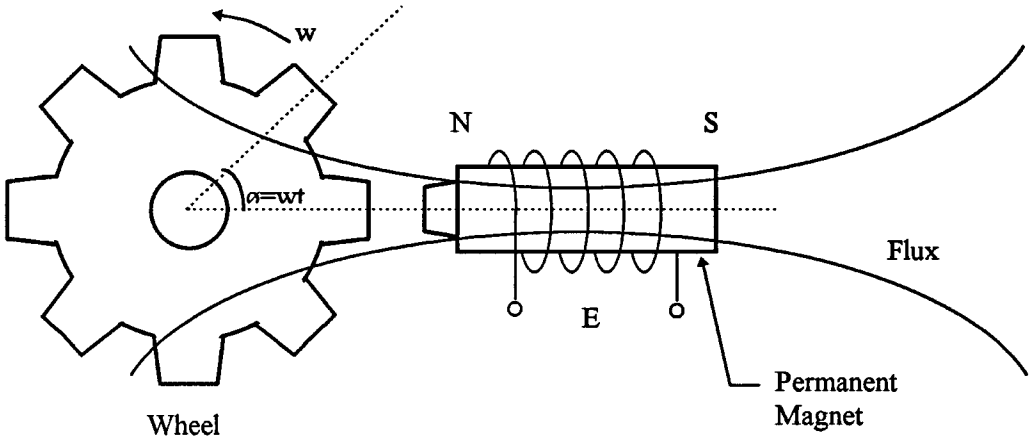


FIGURE 6.11 A variable-reluctance tachogenerator is a sensor which is based on Faraday's law of electromagnetic induction. It consists of a ferromagnetic toothed wheel attached to the rotating shaft and a coil wound onto a permanent magnet extended by a soft iron pole piece. The wheel rotates in close proximity to the pole piece, thus causing the flux linked by the coil to change. The change in flux causes an output in the coil similar to a square waveform whose frequency depends on the speed of the rotation of the wheel and the number of teeth.

Both amplitude and frequency of the generated voltage at the coil are proportional to the angular velocity of the wheel. In principle, the angular velocity ω can be found from either the amplitude or the frequency of the signal. In practice, the amplitude measured may be influenced by loading effects and electrical interference. In signal processing, the frequency is the preferred option because it can be converted into digital signals easily.

The variable-reluctance tachogenerators are most suitable for measuring angular velocities. They are also used in the volume flow rate measurements and the total volume flow determination of fluids.

Microsyn

Another commonly used example of variable-reluctance transducer is the Microsyn, as illustrated in Figure 6.12. In this arrangement, the coils are connected in such a manner that at the null position of the rotary element, the voltages induced in coils 1 and 3 are balanced by voltages induced in coils 2 and 4. The motion of the rotor in the clockwise direction increases the reluctance of coils 1 and 3 while decreasing the reluctance of coils 2 and 4, thus giving a net output voltage e_o. The movement in the counterclockwise direction causes a similar effect in coils 2 and 4 with a 180° phase shift. A direction-sensitive output can be obtained by using phase-sensitive demodulators, as explained in LVDT section of this chapter.

Microsyn transducers are used extensively in applications involving gyroscopes. By the use of microsyns, very small motions can be detected, giving output signals as low as 0.01° of changes in angles. The sensitivity of the device can be made as high as 5 V per degree of rotation. The nonlinearity may vary from 0.5% to 1.0% full scale. The main advantage of these transducers is that the rotor does not have windings and slip-rings. The magnetic reaction torque is also negligible.

Synchros

The term *synchro* is associated with a family of electromechanical devices that can be discussed under different headings. They are used primarily in angle measurements and are commonly applied in control engineering as parts of servomechanisms, machine tools, antennas, etc.

The construction of synchros is similar to that of wound-rotor induction motors, as shown in Figure 6.13. The rotation of the motor changes the mutual inductance between the rotor coil and the three stator coils. The three voltage signals from these coils define the angular position of the rotor.

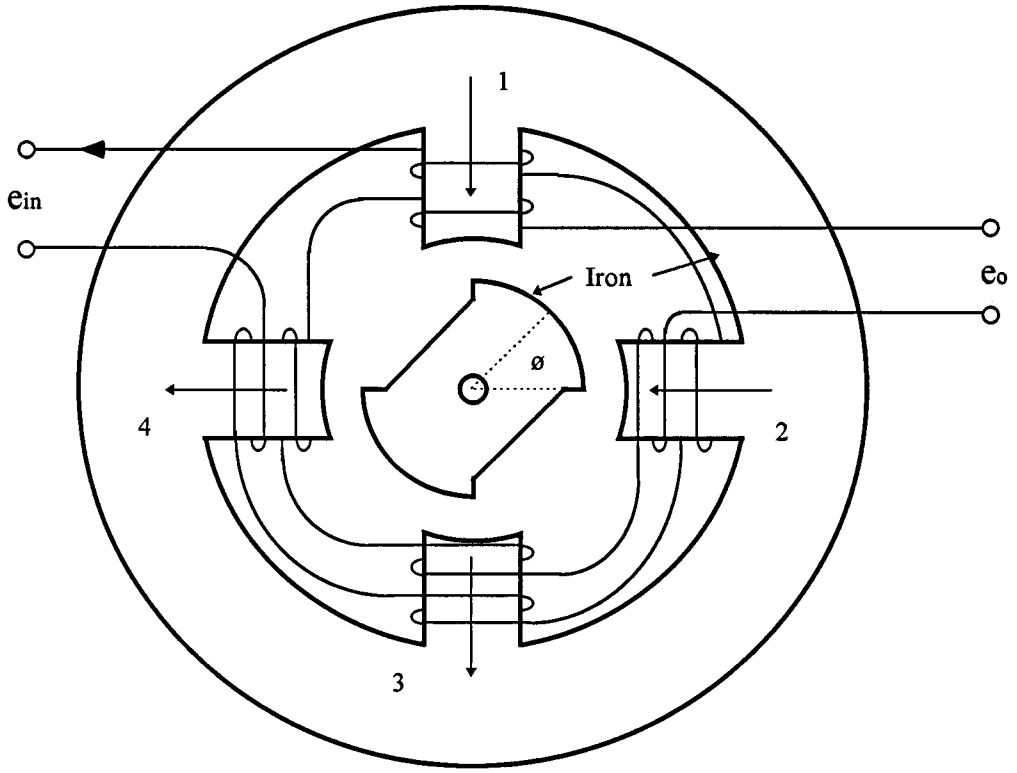


FIGURE 6.12 A microsyn is a variable reluctance transducer that consists of a ferromagnetic rotor and a stator carrying four coils. The stator coils are connected such that at the null position, the voltages induced in coils 1 and 3 are balanced by voltages induced in coils 2 and 4. The motion of the rotor in one direction increases the reluctance of two opposite coils while decreasing the reluctance in others, resulting in a net output voltage e_o. The movement in the opposite direction reverses this effect with a 180° phase shift.

Synchros are used in connection with variety of devices, including: control transformers, Scott T transformers, resolvers, phase-sensitive demodulators, analog to digital converters, etc.

In some cases, a control transformer is attached to the outputs of the stator coils such that the output of the transformer produces a resultant mmf aligned in the same direction as that of the rotor of the synchro. In other words, the synchro rotor acts as a search coil in detecting the direction of the stator field of the control transformer. When the axis of this coil is aligned with the field, the maximum voltage is supplied to the transformer.

In other cases, ac signals from the synchros are first applied to a Scott T transformer, which produces ac voltages with amplitudes proportional to the sine and cosine of the synchro shaft angle. It is also possible to use phase-sensitive demodulations to convert the output signals to make them suitable for digital signal processing.

Linear-Variable Inductor

There is a little distinction between variable-reluctance and variable-inductance transducers. Mathematically, the principles of linear-variable inductors are very similar to the variable-reluctance type of transducer. The distinction is mainly in the pickups rather than principles of operations. A typical linear variable inductor consists of a movable iron core that provides the mechanical input and two coils forming two legs of a bridge network. A typical example of such a transducer is the variable coupling transducer, which is discussed next.

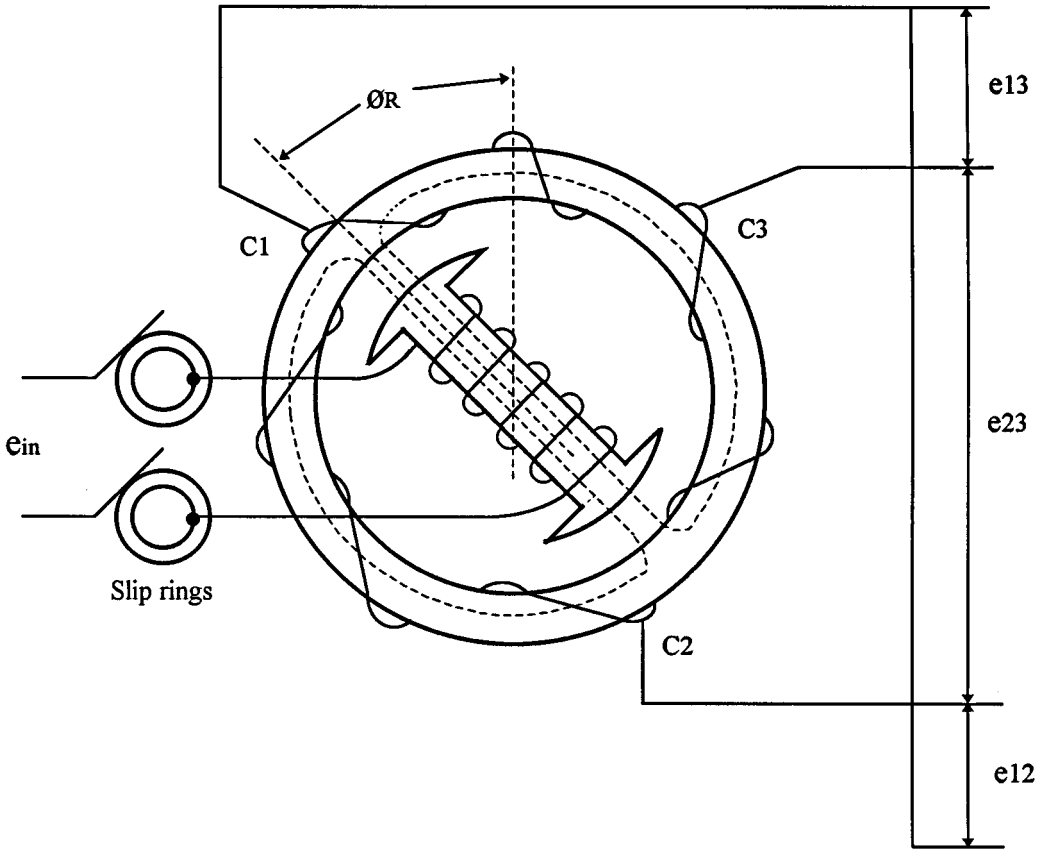


FIGURE 6.13 A synchro is similar to a wound-rotor induction motor. The rotation of the rotor changes the mutual inductance between the rotor coil and the stator coils. The voltages from these coils define the angular position of the rotor. They are primarily used in angle measurements and are commonly applied in control engineering as parts of servomechanisms, machine tools, antennas, etc.

Variable-Coupling Transducers

These transducers consist of a former holding a center tapped coil and a ferromagnetic plunger, as shown in Figure 6.14.

The plunger and the two coils have the same length *l*. As the plunger moves, the inductances of the coils change. The two inductances are usually placed to form two arms of a bridge circuit with two equal balancing resistors, as shown in Figure 6.15. The bridge is excited with ac of 5 V to 25 V with a frequency of 50 Hz to 5 kHz. At the selected excitation frequency, the total transducer impedance at null conditions is set in the 100 Ω to 1000 Ω range. The resistors are set to have about the same value as transducer impedances. The load for the bridge output must be at least 10 times the resistance, *R*, value. When the plunger is in the reference position, each coil will have equal inductances of value *L*. As the plunger moves by δL, changes in inductances $+\delta L$ and $-\delta L$ creates a voltage output from the bridge. By constructing the bridge carefully, the output voltage can be made as a linear function displacement of the moving plunger within a rated range.

In some transducers, in order to reduce power losses due to heating of resistors, center-tapped transformers can be used as a part of the bridge network, as shown in Figure 6.15(*b*). In this case, the circuit becomes more inductive and extra care must be taken to avoid the mutual coupling between the transformer and the transducer.

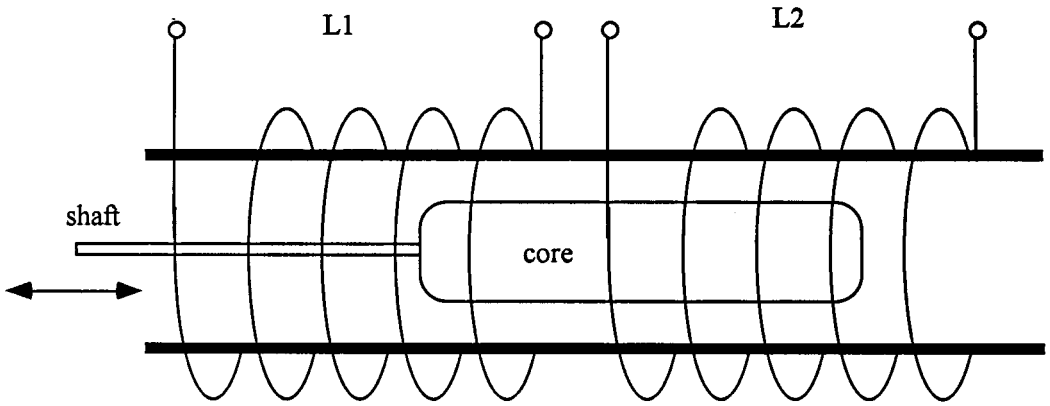


FIGURE 6.14 A typical linear-variable inductor consists of a movable iron core inside a former holding a center-tapped coil. The core and both coils have the same length *l*. When the core is in the reference position, each coil will have equal inductances of value *L*. As the core moves by δl, changes in inductances $+\delta L$ and $-\delta L$ create voltage outputs from the coils.

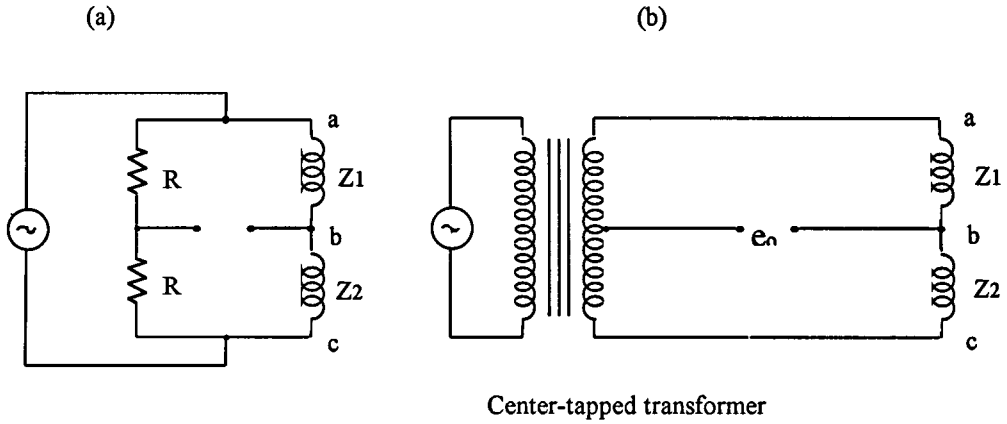


FIGURE 6.15 The two coils of a linear-variable inductor are usually placed to form two arms of a bridge circuit, also having two equal balancing resistors as in circuit (*a*). The bridge is excited with ac of 5 V to 25 V with a frequency of 50 Hz to 5 kHz. At a selected excitation frequency, the total transducer impedance at null conditions is set in the 100 Ω to 1000 Ω range. By careful construction of the bridge, the output voltage can be made a linear function displacement of the core within a limited range. In some cases, in order to reduce power losses due to heating of resistors, center-tapped transformers may be used as a part of the bridge network (*b*).

It is particularly easy to construct transducers of this type, by simply winding a center-tapped coil on a suitable former. The variable-inductance transducers are commercially available in strokes from about 2 mm to 500 cm. The sensitivity ranges between 1% full scale to 0.02% in long stroke special constructions. These devices are also known as linear displacement transducers or LDTs, and they are available in various shape and sizes.

Apart from linear-variable inductors, there are rotary types available too. Their cores are specially shaped for rotational applications. Their nonlinearity can vary between 0.5% to 1% full scale over a range of 90° rotation. Their sensitivity can be up to 100 mV per degree of rotation.

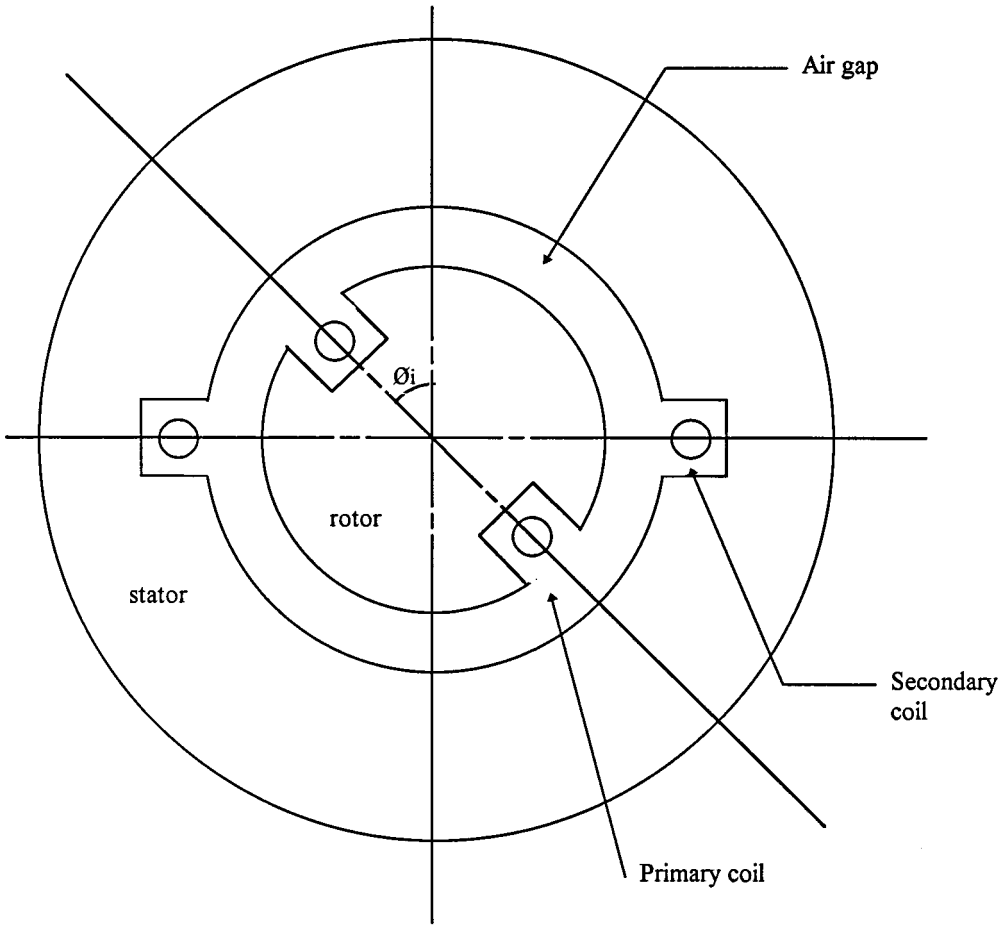


FIGURE 6.16 An induction potentiometer is a linear-variable inductor with two concentrated windings wound on the stator and on the rotor. The rotor winding is excited with ac, inducing voltage in the stator windings. The amplitude of the output voltage is dependent on the relative positions of the coils, as determined by the angle of rotation. For concentrated coils, the variation of the amplitude is sinusoidal, but linearity is restricted in the region of the null position. Different types of induction potentiometers are available with distributed coils that give linear voltages over an angle of 180° of rotation.

Induction Potentiometer

One version of a rotary type linear inductor is the induction potentiometer shown in Figure 6.16. Two concentrated windings are wound on the stator and rotor. The rotor winding is excited with an ac, thus inducing voltage in the stator windings. The amplitude of the output voltage is dependent on the mutual inductance between the two coils, where mutual inductance itself is dependent on the angle of rotation. For concentrated coil type induction potentiometers, the variation of the amplitude is sinusoidal, but linearity is restricted in the region of the null position. A linear distribution over an angle of 180° can be obtained by carefully designed distributed coils.

Standard commercial induction pots operate in a 50 to 400 Hz frequency range. They are small in size, from 1 cm to 6 cm, and their sensitivity can be on the order of 1 V/deg rotation. Although the ranges of induction pots are limited to less than 60° of rotation, it is possible to measure displacements in angles from 0° to full rotation by suitable arrangement of a number of induction pots. As in the case of most

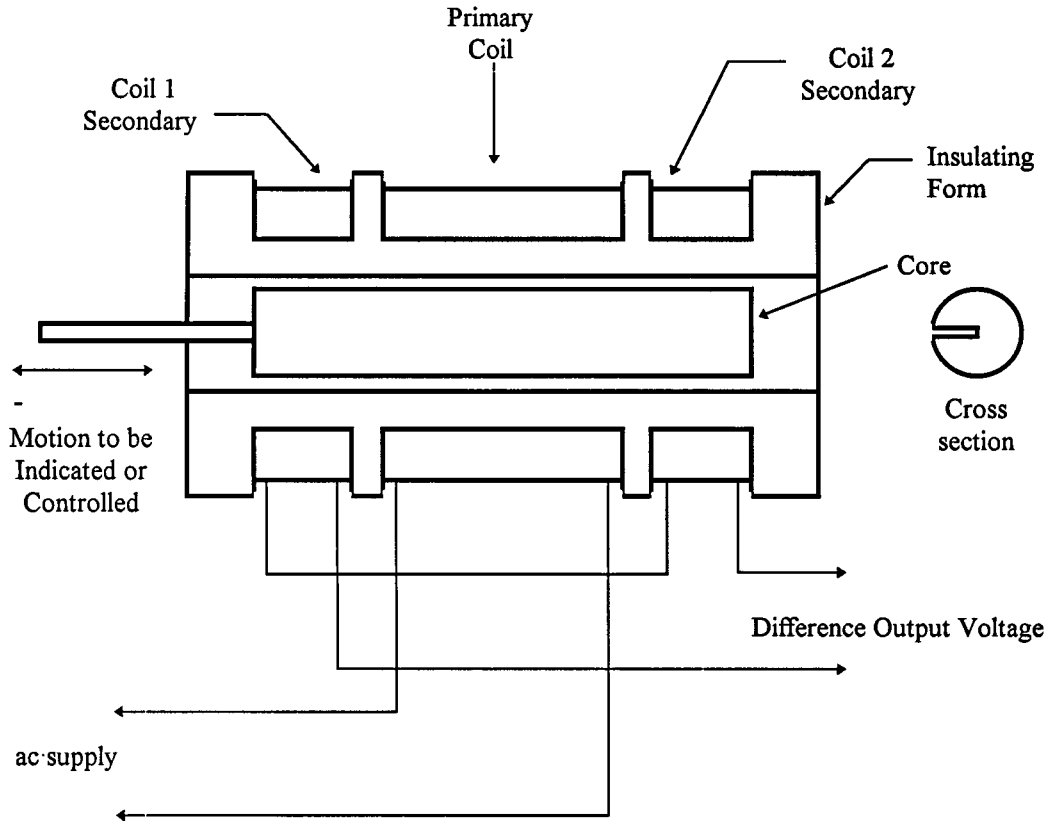


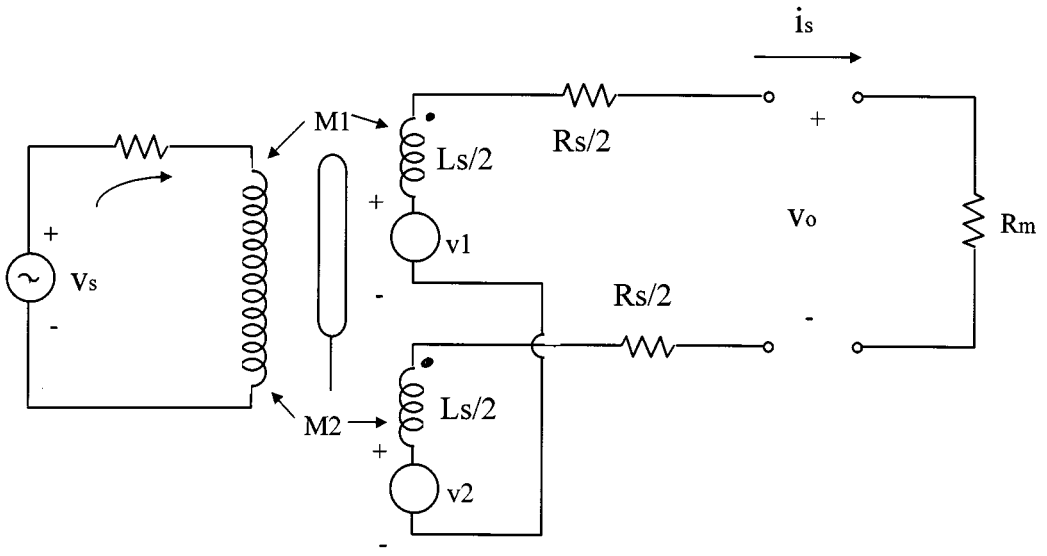
FIGURE 6.17 A linear-variable-differential-transformer LVDT is a passive inductive transducer consisting of a single primary winding positioned between two identical secondary windings wound on a tubular ferromagnetic former. As the core inside the former moves, the magnetic paths between primary and secondaries change, thus giving secondary outputs proportional to the movement. The two secondaries are made as identical as possible by having equal sizes, shapes, and number of turns.

inductive sensors, the output of an induction pot may need phase-sensitive demodulators and suitable filters. In many cases, additional dummy coils are used to improve linearity and accuracy.

Linear Variable-Differential Transformer (LVDT)

The linear variable-differential transformer, LVDT, is a passive inductive transducer finding many applications. It consists of a single primary winding positioned between two identical secondary windings wound on a tubular ferromagnetic former, as shown in Figure 6.17. The primary winding is energized by a high-frequency 50 Hz to 20 kHz ac voltage. The two secondary windings are made identical by having an equal number of turns and similar geometry. They are connected in series opposition so that the induced output voltages oppose each other.

In many applications, the outputs are connected in opposing form, as shown in Figure 6.18(*a*). The output voltages of individual secondaries v_1 and v_2 at null position are illustrated in Figure 6.18(*b*). However, in opposing connection, any displacement in the core position x from the null point causes amplitude of the voltage output v_o and the phase difference α to change. The output waveform v_o in relation to core position is shown in Figure 6.18(*c*). When the core is positioned in the middle, there is



(a)

FIGURE 6.18 The voltages induced in the secondaries of a linear-variable differential-transformer (*a*) may be processed in a number of ways. The output voltages of individual secondaries v_1 and v_2 at null position are illustrated in (*b*). In this case, the voltages of individual coils are equal and in phase with each other. Sometimes, the outputs are connected opposing each other, and the output waveform v_o becomes a function of core position x and phase angle α, as in (*c*). Note the phase shift of 180° as the core position changes above and below the null position.

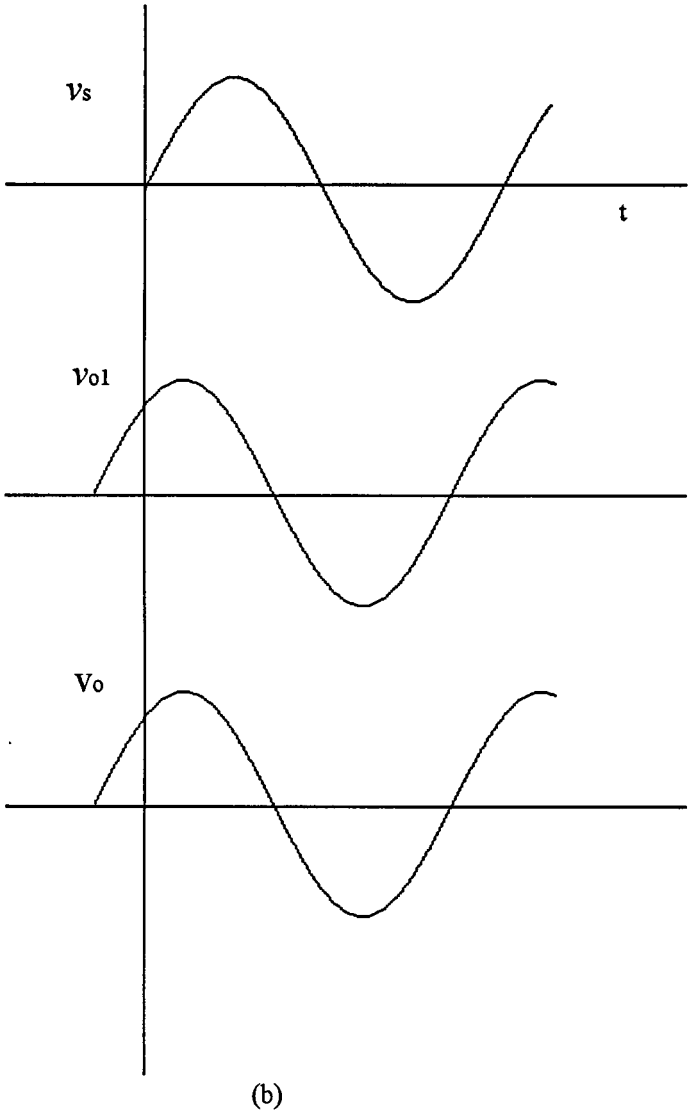
an equal coupling between the primary and secondary windings, thus giving a null point or reference point of the sensor. As long as the core remains near the center of the coil arrangement, output is very linear. The linear ranges of commercial differential transformers are clearly specified, and the devices are seldom used outside this linear range.

The ferromagnetic core or plunger moves freely inside the former, thus altering the mutual inductance between the primary and secondaries. With the core in the center, or at the reference position, the induced emfs in the secondaries are equal; and since they oppose each other, the output voltage is zero. When the core moves, say to the left, from the center, more magnetic flux links with the left-hand coil than with the right-hand coil. The voltage induced in the left-hand coil is therefore larger than the induced voltage on the right-hand coil. The magnitude of the output voltage is then larger than at the null position and is equal to the difference between the two secondary voltages. The net output voltage is in phase with the voltage of the left-hand coil. The output of the device is then an indication of the displacement of the core. Similarly, movement in the opposite direction to the right from the center reverses this effect, and the output voltage is now in phase with the emf of the right-hand coil.

For mathematical analysis of the operation of LVDTs, Figure 6.18(*a*) can be used. The voltages induced in the secondary coils are dependent on the mutual inductance between the primary and individual secondary coils. Assuming that there is no cross-coupling between the secondaries, the induced voltages may be written as:

$$\nu_1 = M_1 s i_p \quad \text{and} \quad \nu_2 = M_2 s i_p \tag{6.15}$$

where M_1 and M_2 are the mutual inductances between primary and secondary coils for a fixed core position; s is the Laplace operator; and i_p is the primary current



(b)

FIGURE 6.18 (continued)

In the case of opposing connection, no load output voltage v_o without any secondary current may be written as:

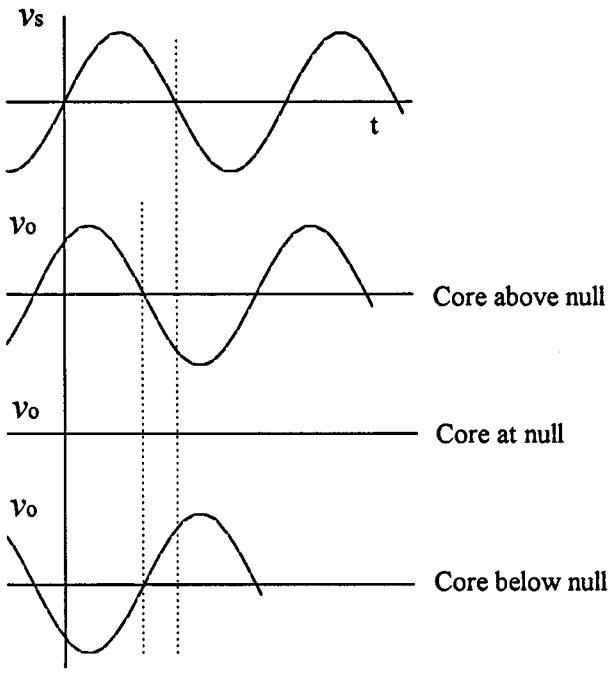
$$\nu_o = \nu_1 - \nu_2 = \left(M_1 - M_2\right) s i_p \tag{6.16}$$

writing

$$\nu_s = i_p \left(R + sL_p\right) \tag{6.17}$$

Substituting i_p in Equation 6.15 gives the transfer function of the transducer as:

$$\nu_o / \nu_s = \left(M_1 - M_2\right) s / \left(R + sL_p\right) \tag{6.18}$$



(c)

FIGURE 6.18 (continued)

However, if there is a current due to output signal processing, then describing equations may be modified as:

$$\nu_o = R_m\, i_s \tag{6.19}$$

where $i_s = (M_1 - M_2)\, si_p/(R_s + R_m + sL_s)$
and

$$\nu_s = i_p\left(R + sL_p\right) - \left(M_1 - M_2\right) si_s \tag{6.20}$$

Eliminating i_p and i_s from Equations 4.19 and 4.20 results in a transfer function as:

$$\nu_o/\nu_s = R_m\left(M_1 - M_2\right)s \Big/ \left\{\left[\left(M_1 - M_2\right)^2 + L_sL_p\right]s^2 + \left[L_p\left(R + R_m\right) + RL_s\right]s + \left(R_s + R_m\right) + R\right\} \tag{6.21}$$

This is a second-order system, which indicates that due to the effect of the numerator of Eq. 6.21, the phase angle of the system changes from +90° at low frequencies to –90° at high frequencies. In practical applications, the supply frequency is selected such that at the null position of the core, the phase angle of the system is 0°.

The amplitudes of the output voltages of secondary coils are dependent on the position of the core. These outputs may directly be processed from each individual secondary coil for slow movements of the core, and when the direction of the movement of the core does not bear any importance. However, for fast movements of the core, the signals can be converted to dc and the direction of the movement from the null position can be detected. There are many options to do this; however, a *phase-sensitive demodulator*

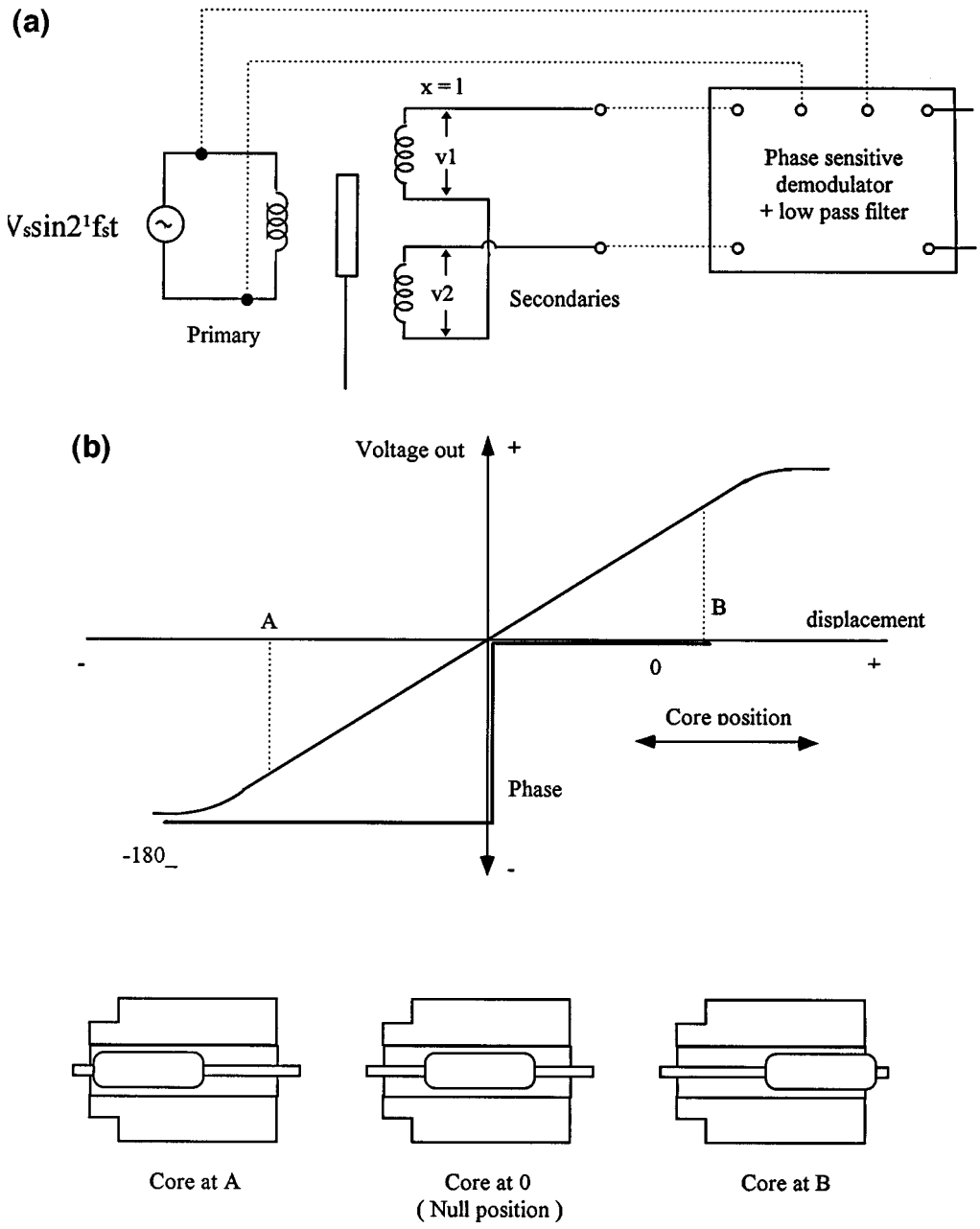


FIGURE 6.19 Phase-sensitive demodulator and (*a*) are commonly used to obtain displacement proportional signals from LVDTs and other differential type inductive sensors. They convert the ac outputs from the sensors into dc values and also indicate the direction of movement of the core from the null position. A typical output of the phase-sensitive demodulator is shown in (*b*). The relationship between output voltage v_o and phase angle α. is also shown against core position *x*.

and filter arrangement is commonly used, as shown in Figure 6.19(*a*). A typical output of the phase-sensitive demodulator is illustrated in Figure 6.19(*b*), in relation to output voltage v_o, displacement *x*, and phase angle α.

The phase-sensitive demodulators are used extensively in differential type inductive sensors. They basically convert the ac outputs to dc values and also indicate the direction of movement of the core

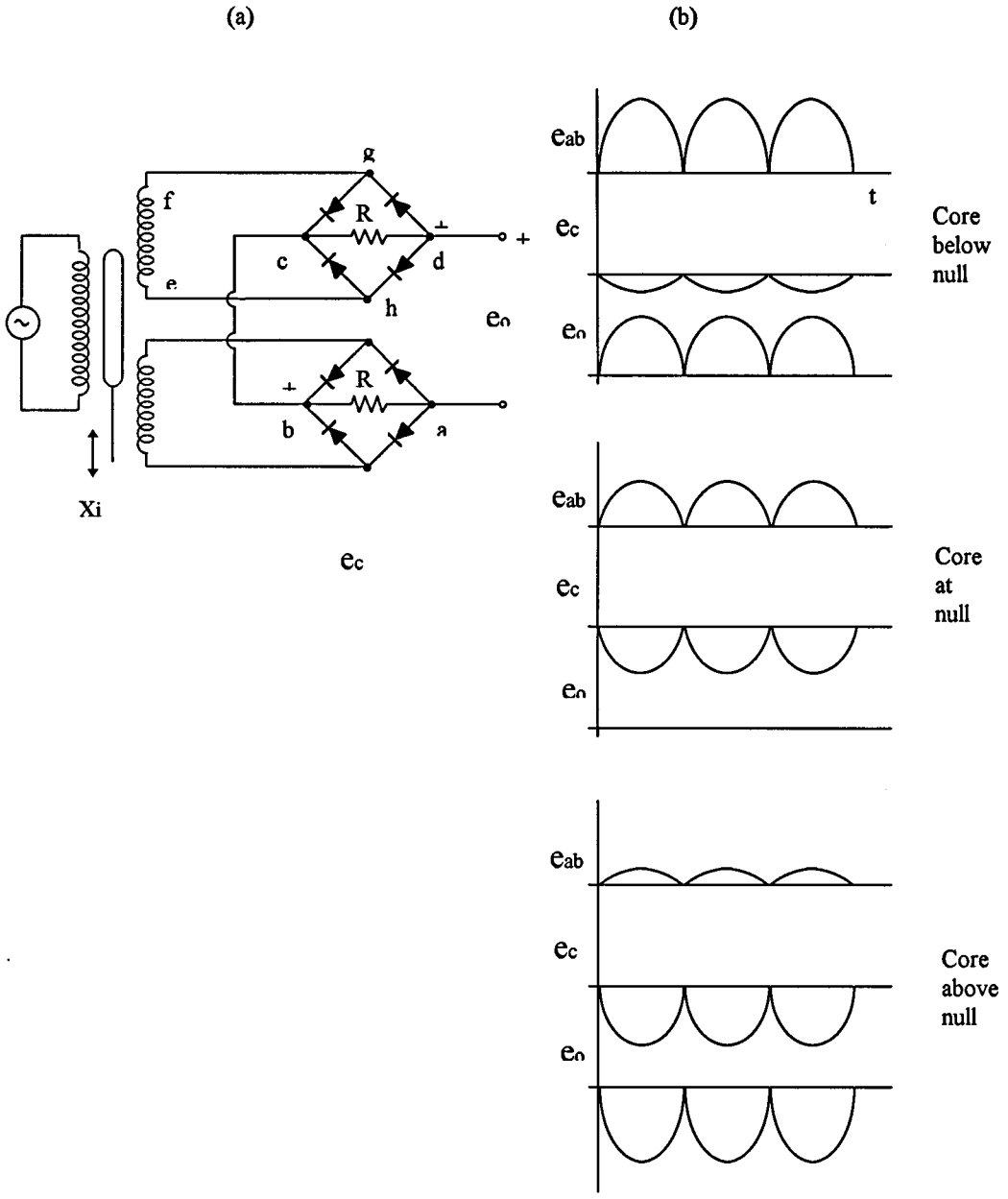


FIGURE 6.20 A typical phase-sensitive demodulation circuit based on diode bridges as in (*a*). Bridge 1 acts as a rectification circuit for secondary 1, and bridge 2 acts as a rectifier for secondary 2 where the net output voltage is the difference between the two bridges, as in (*b*). The position of the core can be determined from the amplitude of the dc output, and the direction of the movement of the core can be determined from the polarity of the voltage. For rapid movements of the core, the output of the diode bridges must be filtered, for this, a suitably designed simple *RC* filter may be sufficient.

from the null position. A typical phase-sensitive demodulation circuit may be constructed, based on diodes shown in Figure 6.20(*a*). This arrangement is useful for very slow displacements, usually less than 1 or 2 Hz. In this figure, bridge 1 acts as a rectification circuit for secondary 1, and bridge 2 acts as a rectifier for secondary 2. The net output voltage is the difference between the outputs of two bridges, as

in Figure 6.20(*b*). The position of the core can be determined from the amplitude of the dc output, and the direction of the movement of the core can be determined from the polarity of the dc voltage. For rapid movements of the core, the outputs of the diode bridges need to be filtered, wherein only the frequencies of the movement of the core pass through and all the other frequencies produced by the modulation process are filtered. For this purpose, a suitably designed simple RC filter may be sufficient.

There are phase-sensitive demodulator chips available in the marketplace, such as AD598 offered by Analog Devices Inc. These chips are highly versatile and flexible to use to suit particular application requirements. These chips offer many advantages over conventional phase-sensitive demodulation devices; for example, frequency off excitation may be adjusted to any value between 20 Hz and 20 kHz by connecting an external capacitor between two pins. The amplitude of the excitation voltage can be set up to 24 V. The internal filters may be set to required values by external capacitors. Connections to analog-to-digital converters are easily made by converting the bipolar output to unipolar scale.

The frequency response of LVDTs is primarily limited by the inertia characteristics of the device. In general, the frequency of the applied voltage should be 10 times the desired frequency response. Commercial LVDTs are available in a broad range of sizes and they are widely used for displacement measurements in a variety of applications. These displacement sensors are available to cover ranges from ±0.25 mm to ±7.5 cm. They are sensitive enough to be used to respond to displacements well below 0.0005 mm. They have operational temperature ranges from –265°C to 600°C. They are also available in radiation-resistant designs for operation in nuclear reactors. For a typical sensor of ±25 mm range, the recommended supply voltage is 4 V to 6 V, with a nominal frequency of 5 kHz, and a maximum nonlinearity of 1% full scale. Several commercial models are available that can produce a voltage output of 300 mV for 1 mm displacement of the core.

One important advantage of the LVDT is that there is no physical contact between the core and the coil form, and hence no friction or wear. Nevertheless, there are radial and longitudinal magnetic forces on the core at all times. These magnetic forces may be regarded as magnetic springs that try to displace the core to its null position. This may be a critical factor in some applications.

One problem with LVDTs is that it may not be easy to make the two halves of the secondary identical; their inductance, resistance, and capacitance may be different, causing a large unwanted quadrature output in the balance position. Precision coil winding equipment may be required to reduce this problem to an acceptable value.

Another problem is associated with null position adjustments. The harmonics in the supply voltage and stray capacitances result in small null voltages. The null voltage may be reduced by proper grounding, which reduces the capacitive effects and center-tapped voltage source arrangements. In center-tapped supplies, a potentiometer may be used to obtain a minimum null reading.

The LVDTs have a variety of applications, including control for jet engines in close proximity to exhaust gases and measuring roll positions in the thickness of materials in hot-slab steel mills. Force and pressure measurements may also be made by LVDTs after some mechanical modifications.

Rotary Variable-Differential Transformer

A variation from the linear-variable differential transformer is the rotary core differential transformer shown in Figures 6.21(*a*) and 6.21(*b*). Here, the primary winding is wound on the center leg of an E core, and the secondary windings are wound on the outer legs of the core. The armature is rotated by an externally applied force about a pivot point above the center leg of the core. When the armature is displaced from its reference or balance position, the reluctance of the magnetic circuit through one secondary coil decreases, simultaneously increasing the reluctance through the other coil. The induced emfs in the secondary windings, which are equal in the reference position of the armature, are now different in magnitude and phase as a result of the applied displacement. The induced emfs in the secondary coils are made to oppose each other and the transformer operates in the same manner as an LVDT. The rotating variable transformers may be sensitive to vibrations. If a dc output is required, a demodulator network can be used, as in the case of LVDTs.

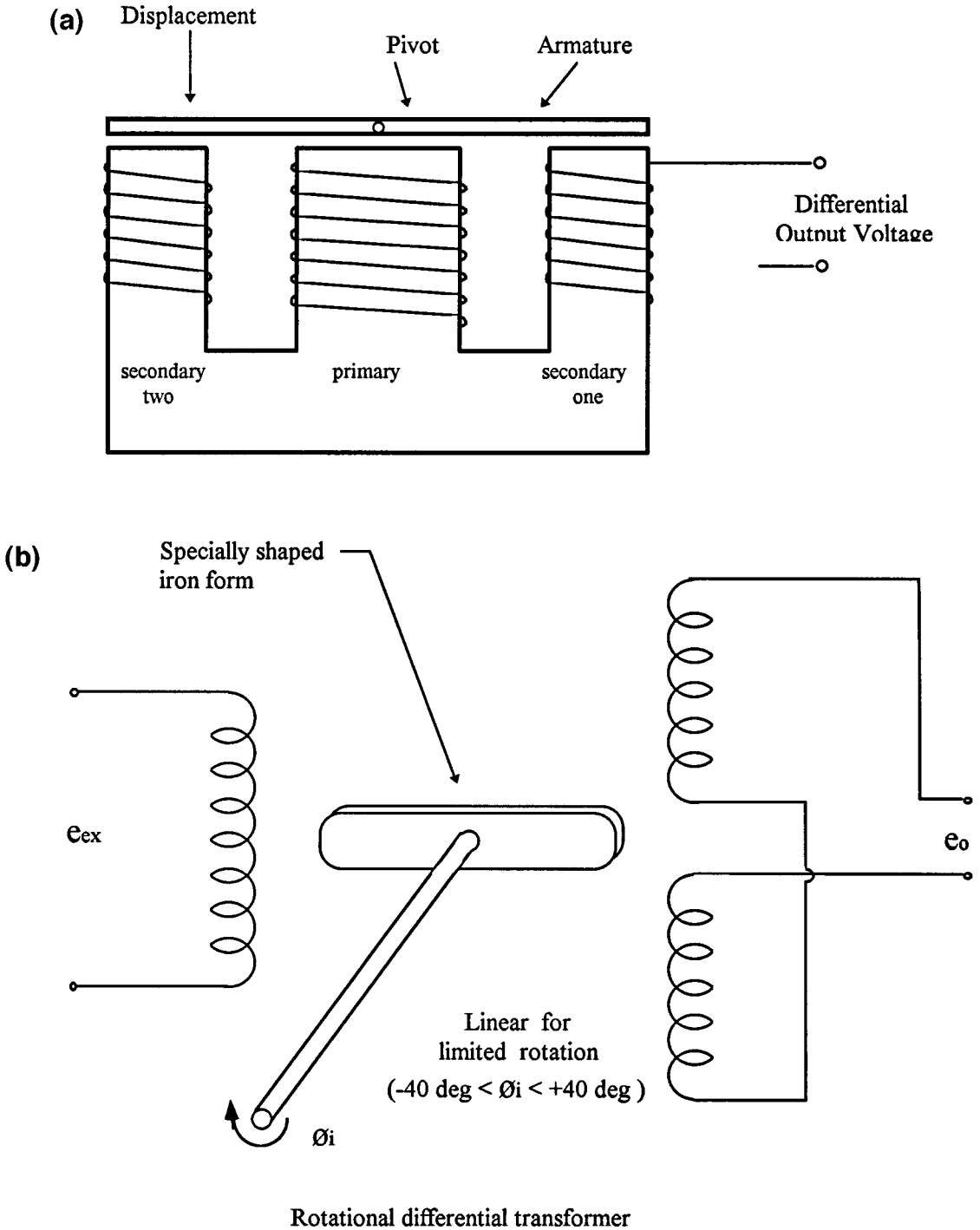


FIGURE 6.21 A rotary core differential transformer has an E-shaped core, carrying the primary winding on the center leg and the two secondaries on the outer legs, as in (*a*). The armature is rotated by an externally applied force about a pivot point above the center leg of the core (*b*). When the armature is displaced from its reference or balance position, the reluctance of the magnetic circuit through one secondary coil is decreased, increasing the reluctance through the other coil. The induced emfs in the secondary windings are different in magnitude and phase as a result of the applied displacement.

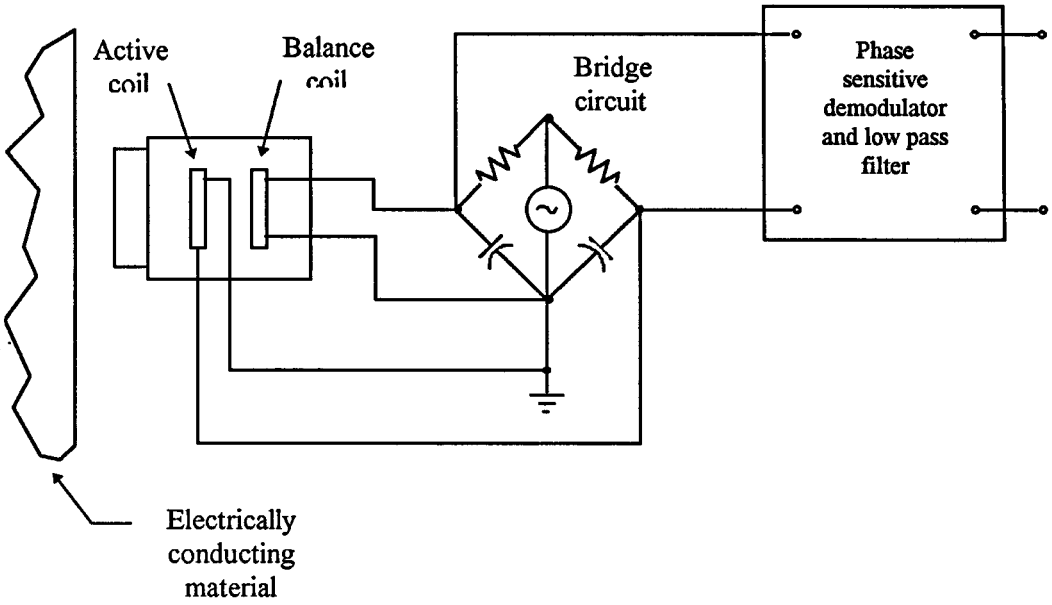


FIGURE 6.22 Eddy current transducers are inductive transducers using probes. The probes contain one active and one balance coil. The active coil responds to the presence of a conducting target, while the balance coil completes a bridge circuit and provides temperature compensation. When the probe is brought close the target, the flux from the probe links with the target, producing eddy currents within the target that alter the inductance of the active coil. This change in inductance is detected by a bridge circuit.

In most rotary linear-variable differential transformers, the rotor mass is very small, usually less than 5 g. The nonlinearity in the output ranges between ±1% and ±3%, depending on the angle of rotation. The motion in the radial direction produces a small output signal that can affect the overall sensitivity. However, this transverse sensitivity is usually kept below 1% of the longitudinal sensitivity.

Eddy Current

Inductive transducers based on eddy currents are mainly probe types, containing two coils as shown in Figure 6.22. One of the coils, known as the active coil, is influenced by the presence of the conducting target. The second coil, known as the balance coil, serves to complete the bridge circuit and provides temperature compensation. The magnetic flux from the active coil passes into the conductive target by means of a probe. When the probe is brought close to the target, the flux from the probe links with the target, producing eddy currents within the target.

The eddy current density is greatest at the target surface and become negligibly small, about three skin depths below the surface. The skin depth depends on the type of material used and the excitation frequency. While thinner targets can be used, a minimum of three skin depths is often necessary to minimize the temperature effects. As the target comes closer to the probe, the eddy currents become stronger, causing the impedance of the active coil to change and altering the balance of the bridge in relation to the target position. This unbalance voltage of the bridge may be demodulated, filtered, and linearized to produce a dc output proportional to target displacement. The bridge oscillation may be as high as 1 MHz. High frequencies allow the use of thin targets and provide good system frequency response.

Probes are commercially available with full-scale diameter ranging from 0.25 to 30 mm with a nonlinearity of 0.5% and a maximum resolution of 0.0001 mm. Targets are usually supplied by the clients, involving noncontact measurements of machine parts. For nonconductive targets, conductive materials of sufficient thickness must be attached to the surface by means of commercially available adhesives. Since the target material, shape, etc. influence the output, it is necessary to calibrate the system statistically

for a specific target. The recommended measuring range of a given probe begins at a standoff distance equal to about 20% of the stated range of the probe. In some cases, a standoff distance of 10% is recommended for which the system is calibrated as standard. A distance greater than 10% of the measuring range can be used as long as the calibrated measuring range is reduced by the same amount.

Flat targets must be of the same diameter as the probe or larger. If the target diameter is smaller than the probe diameter, the output drops considerably, thus becoming unreliable. Curved-surface targets can behave similar to flat surfaces if the diameter exceeds about three or four times the diameter of the probe. In this case, the target essentially becomes an infinite plane. This also allows some cross-axis movement without affecting the system output. Target diameters comparable to the sensor could result in detrimental effects in measurements due to cross-axis movements.

For curved or irregular shaped targets, the system must be calibrated using the exact target that is seen in the operation. This tends to eliminate any errors caused by the curved surfaces during application. However, special multiprobe systems are available for orbital motions of rotating shafts. If the curved (shaft) target is about 10 times larger than the sensor diameter, it acts as an infinite plane and does not need special calibrations. Care must be exercised to deal with electrical runout due to factors such as inhomegeneities in hardness, etc., particularly valid for ferrous targets. However, nonferrous targets are free from electrical runout concerns.

Shielding and Sensitivity of Inductive Sensors to Electromagnetic Interference

Magnetic fields are produced by currents in wires and more strongly by the coils. The fields due to coils are important due to magnetic coupling, particularly when there are two or more coils in the circuit. The magnetic coupling between coils may be controlled by large spacing between coils, orientation of coils, the shape of the coils, and by shielding.

Inductive sensors come in different shape and sizes. While some sensors have closed cores such as toroidal shapes, others have open cores and air gaps between cores and coils. Closed cores can have practically zero external fields, except for small leakage fluxes. Even if the sensors do not have closed cores, most variable inductor sensors have rather limited external fields, due to two neighboring sets of coils connected in opposite directions that minimize the external fields.

Inductive sensors are made from closed conductors. This implies that if the conductor moves in a magnetic field, a current will flow. Alternatively, a magnetic change produces current in a stationary closed conductor. Unless adequate measures are taken, there may be external magnetic fields linking (interference) with the sensor coils, thus producing currents and unwanted responses.

Due to inherent operations, inductive sensors are designed to have high sensitivity to magnetic flux changes. External electromagnetic interference and external fields can seriously affect the performance of the sensors. It is known that moderate magnetic fields are found near power transformers, electrical motors, and power lines. These small fields produce current in the inductive sensors elements. One way of eliminating external effects is accomplished by magnetic shielding of the sensors and by grounding appropriately. In magnetic shielding, one or more shells of high-permeability magnetic materials surround the part to be shielded. Multiple shells may be used to obtain very complete shielding. The ends of each individual shell are separated by insulation so that the shell does not act as a single shorted turn, thus accommodating high current flows. Similarly, in the case of multiple shielding, shells are isolated from each other by proper insulation.

Alternating magnetic fields are also screened by interposing highly conductive metal sheets such as copper or aluminum on the path of the magnetic flux. The eddy currents induced in the shield give a counter mmf that tends to cancel the interfering magnetic field. This type of shielding is particularly effective at high frequencies. Nevertheless, appropriate grounding must still be observed.

In many inductive sensors, stray capacitances can be a problem, especially at the null position of the moving core. If the capacitive effect is greater than a certain value, say 1% of the full-scale output, this effect may be reduced by the use of center-tapped supply and appropriate grounding.

References

1. J. P. Bentley, *Principles of Measurement Systems,* 2nd ed., United Kingdom: Longman Scientific and Technical, 1988.
2. E. O. Doebelin, *Measurement Systems: Application and Design,* 4th ed., New York: McGraw-Hill, 1990.
3. J. P. Holman, *Experimental Methods for Engineers,* 5th ed., New York: McGraw-Hill, 1989.
4. W. J. Tompkins and J. G. Webster, *Interfacing Sensors to the IBM PC,* Englewood Cliffs, NJ: Prentice-Hall, 1988.

Appendix to Section 6.2

LIST OF MANUFACTURERS

Adsen Tech. Inc.
18310 Bedford Circle
La Puente, CA 91744
Fax: (818) 854-2776

Dynalco Controls
3690 N.W. 53rd Street
Ft. Lauderdale, FL 33309
Tel: (954) 739-4300 & (800) 368-6666
Fax: (954) 484-3376

Electro Corporation
1845 57th Street
Sarasato, FL 34243
Tel: (813) 355-8411 & (800) 446-5762
Fax: (813) 355-3120

Honeywell
Dept 722
11 W. Spring Street
Freeport, IL 61032
Tel: (800) 537-6945
Fax: (815) 235-5988

Kaman Inst. Co.
1500 Garden of the Gods Rd.
Colorado Springs, CO 80907
Tel: (719) 599-1132 & (800) 552-6267
Fax: (719) 599-1823

Kavlico Corporation
14501 Los Angeles Avenue
Moorpark, CA 93021
Tel: (805) 523-2000
Fax: (805) 523-7125

Lucas
1000 Lucas Way
Hampton, VA 23666
Tel: (800) 745-8008
Fax: (800) 745-8004

Motion Sensors Inc.
786 Pitts Chapel Rd.
Alizabeth City, NC 27909
Tel: (919) 331-2080
Fax: (919) 331-1666

Rechner Electronics Ind. Inc.
8651 Buffalo Ave.
Niagara Falls, NY 14304
Tel: (800) 544-4106
Fax: (716) 283-2127

Reed Switch Developments Co. Inc.
P.O. Drawer 085297
Racine, WI 53408
Tel: (414) 637-8848
Fax: (414) 637-8861

Smith Research & Technology Inc.
205 Sutton Lane, Dept. TR-95
Colorado Springs, CO 80907
Tel: (719) 634-2259
Fax: (719) 634-2601

Smith Systems Inc.
6 Mill Creek Dr.
Box 667
Brevard, NC 28712
Tel: (704) 884-3490
Fax: (704) 877-3100

Standex Electronics
4538 Camberwell Rd.
Dept. 301L
Cincinnati, OH 45209
Tel: (513) 871-3777
Fax: (513) 871-3779

Turck Inc.
3000 Campus Drive
Minneapolis, MN 55441
Tel: (612) 553-7300 & (800) 544-7769
Fax: (612) 553-0708

Xolox Sensor Products
6932 Gettysburg Pike
Ft. Wayne, IN 46804
Tel: (800) 348-0744
Fax: (219) 432-0828

6.3 Capacitive Sensors—Displacement

Halit Eren and Wei Ling Kong

Capacitive sensors are extensively used in industrial and scientific applications. They are based on changes in capacitance in response to physical variations. These sensors find many diverse applications — from humidity and moisture measurements to displacement sensing. In some cases, the basic operational and sensing principles are common in dissimilar applications; and in other cases, different principles can be used for the same applications. For example, capacitive microphones are based on variations of spacing between plates in response to acoustical pressure, thus turning audio signals to variations in capacitance. On the other hand, a capacitive level indicator makes use of the changes in the relative permittivity between the plates. However, capacitive sensors are best known to be associated with displacement measurements for rotational or translational motions, as will be described next. Other applications of capacitance sensors such as humidity and moisture will be discussed.

Capacitive Displacement Sensors

The measurement of distances or displacements is an important aspect of many industrial, scientific, and engineering systems. The displacement is basically the vector representing a change in position of a body or point with respect to a reference point. Capacitive displacement sensors satisfy the requirements of applications where high linearity and wide ranges (from a few centimeters to a couple of nanometers) are needed.

The basic sensing element of a typical displacement sensor consists of two simple electrodes with capacitance C. The capacitance is a function of the distance d (cm) between the electrodes of a structure, the surface area A (cm^2) of the electrodes, and the permittivity ε (8.85×10^{-12} F m^{-1} for air) of the dielectric between the electrodes; therefore:

$$C = f\left(d, A, \varepsilon\right) \tag{6.22}$$

There are three basic methods for realizing a capacitive displacement sensor: by varying d, A, or ε, as discussed below.

Variable Distance Displacement Sensors

A capacitor displacement sensor, made from two flat coplanar plates with a variable distance x apart, is illustrated in Figure 6.23. Ignoring fringe effects, the capacitance of this arrangement can be expressed by:

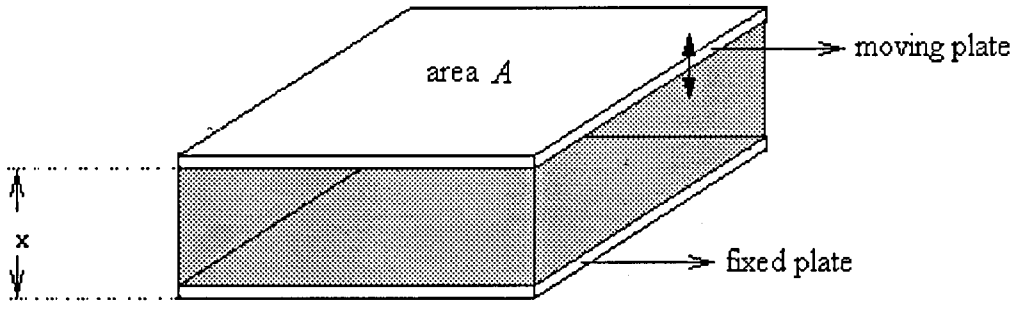


FIGURE 6.23 A variable distance capacitive displacement sensor. One of the plates of the capacitor moves to vary the distance between plates in response to changes in a physical variable. The outputs of these transducers are nonlinear with respect to distance x having a hyperbolic transfer function characteristic. Appropriate signal processing must be employed for linearization.

$$C(x) = \varepsilon A/x = \varepsilon_r \varepsilon_0 A/x \tag{6.23}$$

where ε = the dielectric constant or permittivity
ε_r = the relative dielectric constant (in air and vacuum $\varepsilon_r \approx 1$)
$\varepsilon_0 = 8.854188 \times 10^{-12}$ F/m^{-1}, the dielectric constant of vacuum
x = the distance of the plates in m
A = the effective area of the plates in m^2

The capacitance of this transducer is nonlinear with respect to distance x, having a hyperbolic transfer function characteristic. The sensitivity of capacitance to changes in plate separation is

$$dC/dx = -\varepsilon_r \varepsilon_0 A/x^2 \tag{6.24}$$

Equation 6.24 indicates that the sensitivity increases as x decreases. Nevertheless, from Equations 6.23 and 6.24, it follows that the percent change in C is proportional to the percent change in x. This can be expressed as:

$$dC/C = -dx/x \tag{6.25}$$

This type of sensor is often used for measuring small incremental displacements without making contact with the object.

Variable Area Displacement Sensors

Alternatively, the displacements may be sensed by varying the surface area of the electrodes of a flat plate capacitor, as illustrated in Figure 6.24. In this case, the capacitance would be:

$$C = \varepsilon_r \varepsilon_0 (A - wx)/d \tag{6.26}$$

where w = the width
wx = the reduction in the area due to movement of the plate

Then, the transducer output is linear with displacement x. This type of sensor is normally implemented as a rotating capacitor for measuring angular displacement. The rotating capacitor structures are also used as an output transducer for measuring electric voltages as capacitive voltmeters.

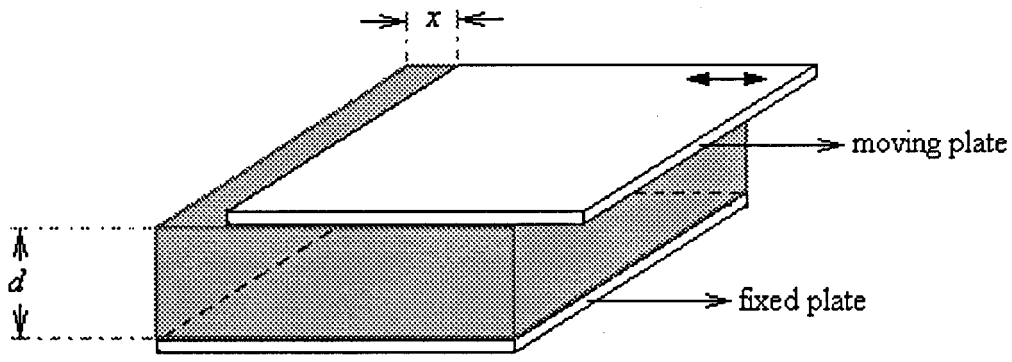


FIGURE 6.24 A variable area capacitive displacement sensor. The sensor operates on the variation in the effective area between plates of a flat-plate capacitor. The transducer output is linear with respect to displacement x. This type of sensor is normally implemented as a rotating capacitor for measuring angular displacement.

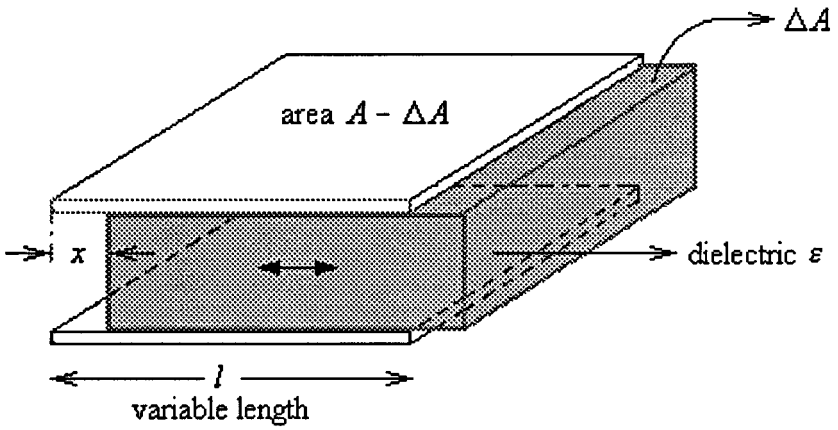


FIGURE 6.25 A variable dielectric capacitive displacement sensor. The dielectric material between the two parallel plate capacitors moves, varying the effective dielectric constant. The output of the sensor is linear.

Variable Dielectric Displacement Sensors

In some cases, the displacement may be sensed by the relative movement of the dielectric material between the plates, as shown in Figure 6.25. The corresponding equations would be:

$$C = \varepsilon_0 \, w\left[\varepsilon_2 l - \left(\varepsilon_2 - \varepsilon_1\right) x\right] \tag{6.27}$$

where ε_1 = the relative permittivity of the dielectric material
ε_2 = the permittivity of the displacing material (e.g., liquid)

In this case, the output of the transducer is also linear. This type of transducer is predominantly used in the form of two concentric cylinders for measuring the level of fluids in tanks. A nonconducting fluid forms the dielectric material. Further discussion will be included in the level measurements section.

Differential Capacitive Sensors

Some of the nonlinearity in capacitive sensors can be eliminated using differential capacitive arrangements. These sensors are basically three-terminal capacitors, as shown in Figure 6.26. Slight variations in the construction of these sensors find many different applications, including differential pressure measurements. In some versions, the central plate moves in response to physical variables with respect to the fixed plates. In others, the central plate is fixed and outer plates are allowed to move. The output from the center plate is zero at the central position and increases as it moves left or right. The range is equal to twice the separation *d*. For a displacement *d*, one obtains:

$$2\delta C = C_1 - C_2 = \varepsilon_r \varepsilon_0 lw \Big/ \left(d - \delta d\right) - \varepsilon_r \varepsilon_0 lw \Big/ \left(d + \delta d\right) = 2\varepsilon_r \varepsilon_0 lw\delta d \Big/ \left(d^2 + \delta d^2\right) \tag{6.28}$$

and

$$C_1 + C_2 = 2C = \varepsilon_r \varepsilon_0 lw \Big/ \left(d - \delta d\right) + \varepsilon_r \varepsilon_0 lw \Big/ \left(d + \delta d\right) = 2\varepsilon_r \varepsilon_0 lwd \Big/ \left(d^2 + \delta d^2\right) \tag{6.29}$$

Giving approximately:

$$\delta C / C = \delta d / d \tag{6.30}$$

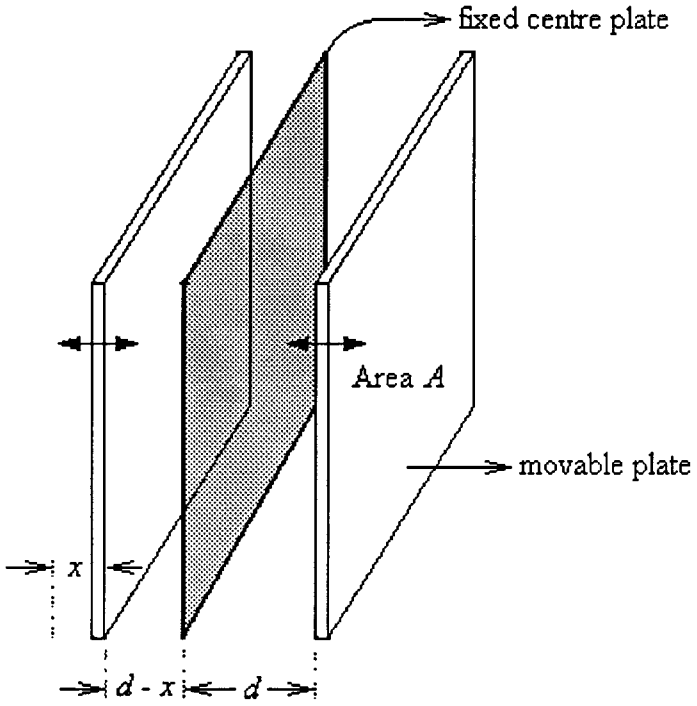


FIGURE 6.26 A differential capacitive sensor. They are essentially three terminal capacitors with one fixed center plate and two outer plates. The response to physical variables is linear. In some versions, the central plate moves in response to physical variable with respect to two outer plates, and in the others, the central plate is fixed and outer plates are allowed to move.

This indicates that the response of the device is more linear than the response of the two plate types. However, in practice some nonlinearity is still observed due to defects in the structure. Therefore, the outputs of these type of sensors still need to be processed carefully, as explained in the signal processing section.

In some differential capacitive sensors, the two spherical depressions are ground into glass disks; then, these are gold-plated to form the fixed plates of a differential capacitor. A thin, stainless-steel diaphragm is clamped between the disks and serves as a movable plate. With equal pressure applied to both ports, the diaphragm is then in neutral position and the output is balanced at a corresponding bridge. If one pressure is greater than the other, the diaphragm deflects proportionally, giving an output due to the differential pressure. For the opposite pressure difference, there is a phase change of 180°. A direction-sensitive dc output can be obtained by conventional phase-sensitive demodulation and appropriate filtering. Details of signal processing are given at the end of this chapter. In general, the differential capacitors exhibit better linearity than single-capacitor types.

Integrated Circuit Smart Capacitive Position Sensors

Figure 6.27 shows a typical microstructure capacitive displacement sensor. The sensor consists of two electrodes with capacitance, C_x. Since the system is simple, the determination of the capacitance between the two electrodes is straightforward. The smaller electrode is surrounded by a guard electrode to make C_x independent of lateral and rotational movements of the system parallel to the electrode surface. However, the use of a guard electrode introduces relative deviations in the capacitance C_x between the two electrodes. This is partially true if the size of the guard electrode is smaller than:

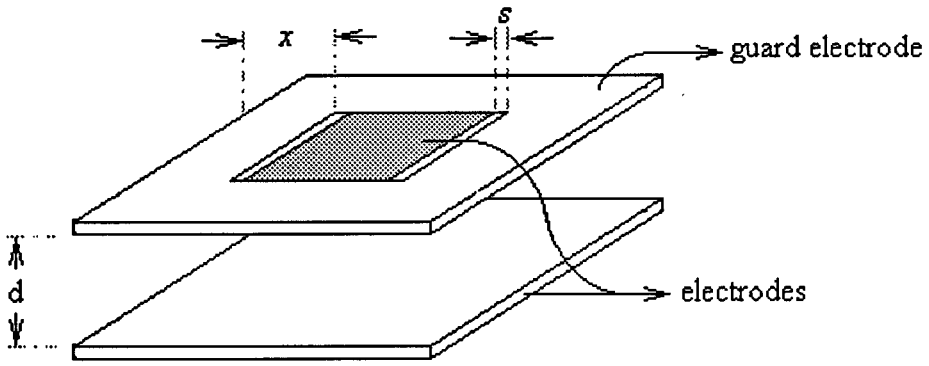


FIGURE 6.27 A typical smart capacitive position sensor. This type of microstructure position sensor contains three electrodes, two of which are fixed and the third electrode moves infinitesimally relative to the others. Although the response is highly nonlinear the integrated chip contains linearization circuits. They feature a 0 mm to 1 mm measuring range with 1 μm accuracy.

$$\delta < \exp\left(-\pi x / d\right) \tag{6.31}$$

where x is the width of the guard and d is the distance between the electrodes. Since this deviation introduces nonlinearity, δ is required to be less than 100 ppm. Another form of deviation also exists between the small electrode and the surrounding guard, particularly for gaps

$$\delta < \exp\left(-\pi d / s\right) \tag{6.32}$$

where s is the width of the gap. When the gap width, s, is less than 1/3 of the distance between electrodes, this deviation is negligible.

For signal processing, the system uses the three-signal concept. The capacitor C_x is connected to an inverting operational amplifier and oscillator. If the external movements are linear, by taking into account the parasitic capacitors and offsetting effects, the following equation can be written:

$$M_x = mC_x + M_{off} \tag{6.33}$$

where m is the unknown gain and M_{off} is the unknown offset. By performing the measurement of a reference C_{ref}, by measuring the offset, M_{off}, and by making $m = 0$, the parameters m and M_{off} can be eliminated. The final measurement result for the position, P_{os}, can be defined as:

$$P_{os} = \frac{M_{ref} - M_{off}}{M_x - M_{off}} \tag{6.34}$$

In this case, the sensor capacitance C_x can be simplified to:

$$C_x = \frac{\varepsilon A_x}{d_0 + \Delta d} \tag{6.35}$$

where A_x is the area of the electrode, d_0 is the initial distance between them, ε is the dielectric constant, and Δd is the displacement to be measured. For the reference electrodes, the reference capacitance may be found by:

$$C_{\text{ref}} = \frac{\varepsilon A_{\text{ref}}}{d_{\text{ref}}} \tag{6.36}$$

with A_{ref} the area and d_{ref} the distance. Substitution of Equations 6.35 and 6.36 into Equations 6.33 and 6.34 yields:

$$P_{\text{os}} = \frac{A_{\text{ref}}\left(d_0 + \Delta d\right)}{A_x d_{\text{ref}}} = a_1 \frac{\Delta d}{d_{\text{ref}}} + a_0 \tag{6.37}$$

P_{os} is a value representing the position if the stable constants a_1 and a_0 are unknown. The constant $a_1 = A_{\text{ref}}/A_x$ becomes a stable constant so long as there is good mechanical matching between the electrode areas. The constant $a_0 = (A_{\text{ref}}\ d_0)/(A_x\ d_{\text{ref}})$ is also a stable constant for fixed d_0 and d_{ref}. These constants are usually determined by calibration repeated over a certain time span. In many applications, these calibrations are omitted if the displacement sensor is part of a larger system where an overall calibration is necessary. This overall calibration usually eliminates the requirement for a separate determination of a_1 and a_0.

The accuracy of this type of system could be as small as 1 μm over a 1 mm range. The total measuring time is better than 0.1 s. The capacitance range is from 1 pF to 50 fF. Interested readers should refer to [4] at the end of this chapter.

Capacitive Pressure Sensors

A commonly used two-plate capacitive pressure sensor is made from one fixed metal plate and one flexible diaphragm, as shown in Figure 6.28. The flat circular diaphragm is clamped around its circumference and bent into a curve by an applied pressure *P*. The vertical displacement *y* of this system at any radius *r* is given by:

$$y = 3\left(1 - v^2\right)\left(a^2 - r^2\right)\ P/16\ Et^3 \tag{6.38}$$

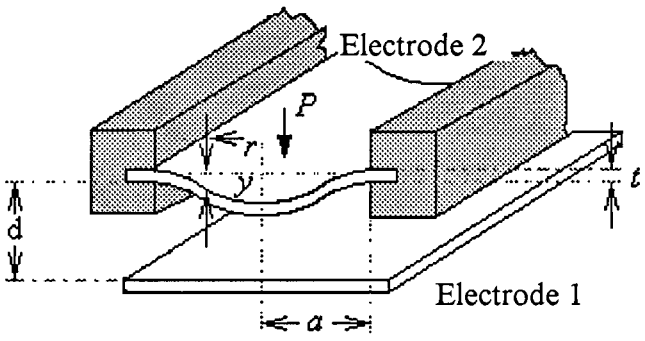


FIGURE 6.28 A capacitive pressure sensor. These pressure sensors are made from a fixed metal plate and a flexible diaphragm. The flat flexible diaphragm is clamped around its circumference. The bending of the flexible plate is proportional to the applied pressure *P*. The deformation of the diaphragm results in changes in capacitance.

where a = the radius of diaphragm
t = the thickness of diaphragm
E = Young's modulus
v = Poisson's ratio

Deformation of the diaphragm means that the average separation of the plates is reduced. Hence, the resulting increase in the capacitance ΔC can be calculated by:

$$\Delta C/C = \left(1 - v^2\right) a^4 P/16\, Et^3 \tag{6.39}$$

where d is the initial separation of the plates and C is the capacitance at zero pressure.

Another type of sensor is the differential capacitance pressure sensor shown in Figure 6.29. The capacitances C_1 and C_2 of the sensor change with respect to the fixed central plate in response to the applied pressures P_1 and P_2. Hence, the output of the sensor is proportional to $(P_1 - P_2)$. The signals are processed using one of the techniques described in the "Signal Processing" section of this chapter.

Capacitive Accelerometers and Force Transducers

In recent years, capacitive-type micromachined accelerometers, as illustrated in Figure 6.30, are gaining popularity. These accelerometers use the proof mass as one plate of the capacitor and use the other plate as the base. When the sensor is accelerated, the proof mass tends to move; thus, the voltage across the capacitor changes. This change in voltage corresponds to the applied acceleration.

In Figure 6.30, let $F(x)$ be the positive force in the direction in which x increases. Neglecting all losses (due to friction, resistance, etc.), the energy balance of the system can be written for an infinitesimally small displacement dx, electrical energy dE_e, and field energy dE_f of the electrical field between the electrodes as:

$$\mathrm{d}E_m + \mathrm{d}E_e = \mathrm{d}E_f \tag{6.40}$$

in which:

$$\mathrm{d}E_m = F\left(x\right) dx \tag{6.41}$$

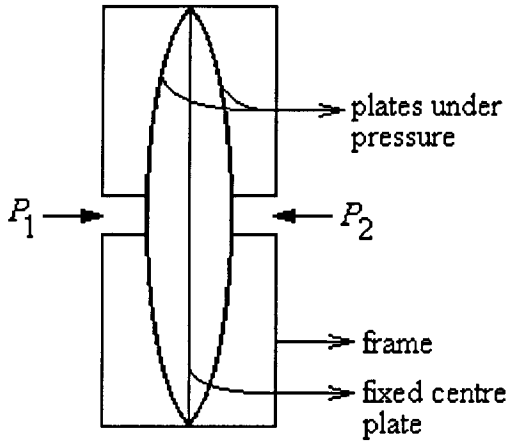


FIGURE 6.29 A differential capacitive pressure sensor. The capacitances C_1 and C_2 of the sensor changes due to deformation in the outer plates, with respect to the fixed central plate in response to the applied pressures P_1 and P_2.

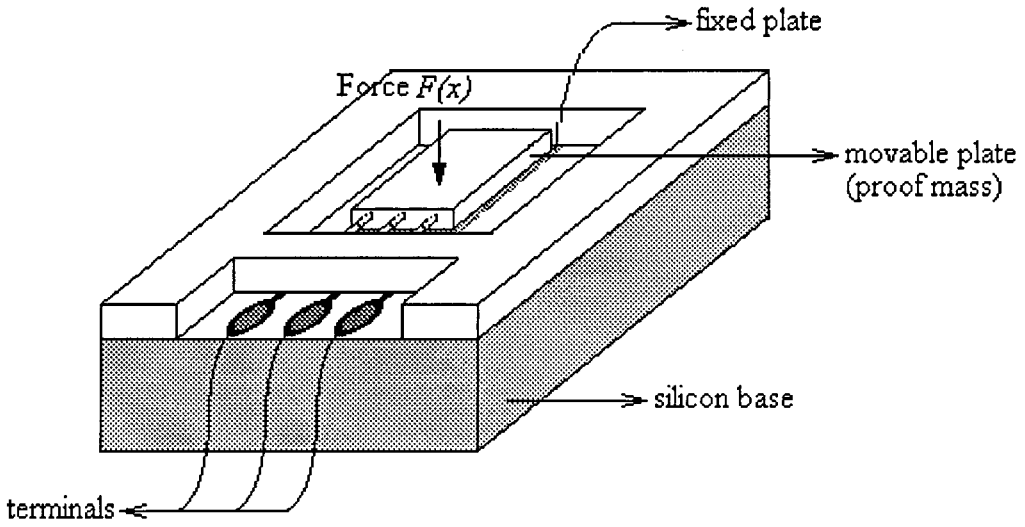


FIGURE 6.30 A capacitive force transducer. A typical capacitive micromachined accelerometer has one of the plates as the proof mass. The other plate is fixed, thus forming the base. When the sensor is accelerated, the proof mass tends to move, thus varying the distance between the plates and altering the voltage across the capacitor. This change in voltage is made to be directly proportional to the applied acceleration.

Also,

$$dE_m = d(QV) = Q\,dV + V\,dQ \tag{6.42}$$

If the supply voltage V across the capacitor is kept constant, it follows that $dV = 0$. Since $Q = VC(x)$, the Coulomb force is given by:

$$F(x) = -V^2 \frac{dC(x)}{dx} \tag{6.43}$$

Thus, if the movable electrode has complete freedom of motion, it will have assumed a position in which the capacitance is maximal; also, if C is a linear function of x, the force $F(x)$ becomes independent of x.

Capacitive silicon accelerometers are available in a wide range of specifications. A typical lightweight sensor will have a frequency range of 0 to 1000 Hz, and a dynamic range of acceleration of ± 2 g to ± 500 g.

Capacitive Liquid Level Measurement

The level of a nonconducting liquid can be determined by capacitive techniques. The method is generally based on the difference between the dielectric constant of the liquid and that of the gas or air above it. Two concentric metal cylinders are used for capacitance, as shown in Figure 6.31. The height of the liquid, h, is measured relative to the total height, l. Appropriate provision is made to ensure that the space between the cylindrical electrodes is filled by the liquid to the same height as the rest of the container. The usual operational conditions dictate that the spacing between the electrodes, $s = r_2 - r_1$, should be much less than the radius of the inner electrode, r_1. Furthermore, the tank height should be much greater than r_2. When these conditions apply, the capacitance is approximated by:

$$C = \frac{\varepsilon_1(l) + \varepsilon_g(h-1)}{4.6\log\left[1-(s/r)\right]} \tag{6.44}$$

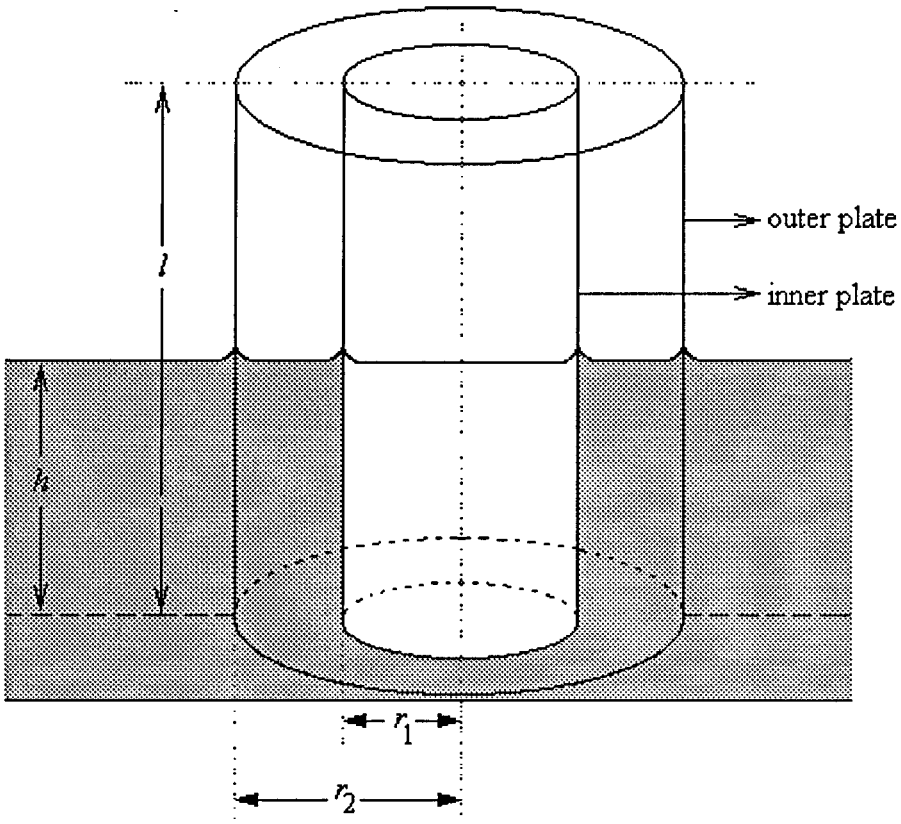


FIGURE 6.31 A capacitive liquid level sensor. Two concentric metal cylinders are used as electrodes of a capacitor. The value of the capacitance depends on the permittivity of the liquid and that of the gas or air above it. The total permittivity changes depending on the liquid level. These devices are usually applied in nonconducting liquid applications.

where ε_l and ε_g are the dielectric constants of the liquid and gas (or air), respectively. The denominator of the above equation contains only terms that relate to the fixed system. Therefore, they become a single constant. A typical application is the measurement of the amount of gasoline in a tank in airplanes. The dielectric constant for most compounds commonly found in gasoline is approximately equal to 2, while that of air is approximately unity. A linear change in capacitance with gasoline level is expected for this situation. Quite high accuracy can be achieved if the denominator is kept quite small, thus accentuating the level differences. These sensors often incorporate an ac deflection bridge.

Capacitive Humidity and Moisture Sensors

The permittivities of atmospheric air, of some gases, and of many solid materials are functions of moisture content and temperature. Capacitive humidity devices are based on the changes in the permittivity of the dielectric material between plates of capacitors. The main disadvantage of this type sensor is that a relatively small change in humidity results in a capacitance large enough for a sensitive detection.

Capacitive humidity sensors enjoy wide dynamic ranges, from 0.1 ppm to saturation points. They can function in saturated environments for long periods of time, a characteristic that would adversely affect many other humidity sensors. Their ability to function accurately and reliably extends over a wide range of temperatures and pressures. Capacitive humidity sensors also exhibit low hysteresis and high stability with minimal maintenance requirements. These features make capacitive humidity sensors viable for many specific operating conditions and ideally suitable for a system where uncertainty of unaccounted conditions exists during operations.

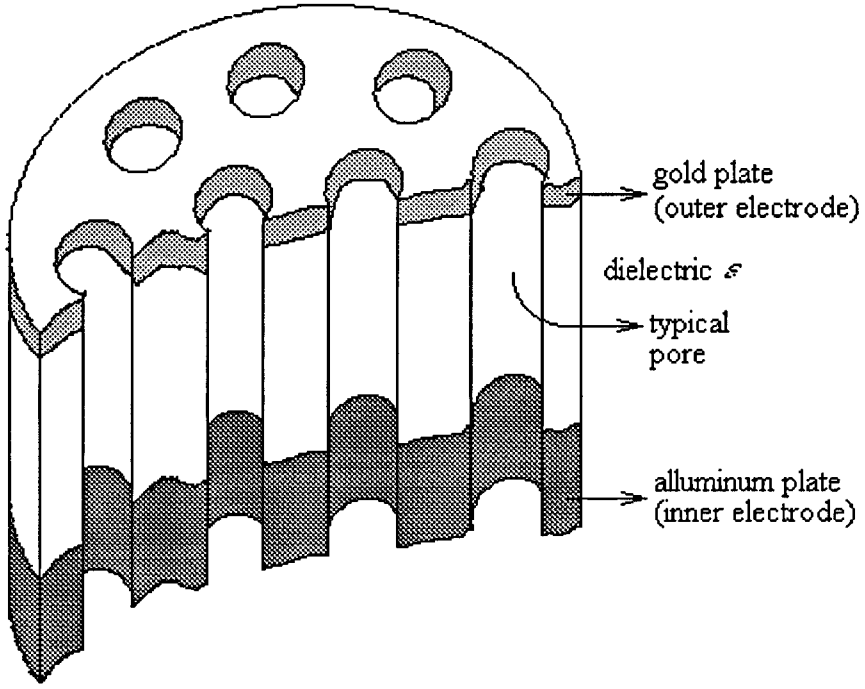


FIGURE 6.32 A typical capacitive humidity sensor. The sensors have pore or cracked mosaic structure for the moisture in air or gas to reach the dielectric material. The characteristics of the dielectric material change with the amount of water absorbed, thus reducing the resistance and increasing the capacitance. The quantity measured can be either resistance, capacitance, or impedance.

There are many types of capacitive humidity sensors. Aluminum, tantalum, silicon, and polymer types are introduced here.

Aluminum Type Capacitive Humidity Sensors

The majority of capacitive humidity sensors are aluminum oxide type sensors. In these type of sensors, high-purity aluminum is chemically oxidized to produce a prefilled insulating layer of partially hydrated aluminum oxide, which acts as the dielectric. A water-permeable but conductive gold film is deposited onto the oxide layer, usually by vacuum deposition, which forms the second electrode of the capacitor.

In another type, the aluminum-aluminum oxide sensor has a pore structure as illustrated in Figure 6.32. The oxide, with its pore structure, forms the active sensing material. Moisture in the air reaching the pores reduces the resistance and increases the capacitance. The decreased resistance can be thought of as being due to an increase in the conduction through the oxide. An increase in capacitance can be viewed as due to an increase in the dielectric constant. The quantity measured can be either resistance, capacitance, or impedance. High humidities are best measured by capacitance because resistance changes are vanishingly small in this region.

In addition to the kind of transducer design illustrated here, there are many others available with a number of substantial modifications for particular properties, such as increased sensitivity or faster response. Although most of these modifications result in a change in physical dimensions or appearance, the sensing material of the transducer — the aluminum oxide — remains the same.

In some versions, the oxide layer is formed by parallel tubular pores that are hexagonally packed and perpendicular to the plane of the base layer. These pores stop just before the aluminum layer, forming a very thin pore base. Water absorbed in these tubules is directly related to the moisture content of the gas in contact with it. The porous nature of the oxide layer produces a large area for the absorption of water vapor. At low humidities, the capacitance is due entirely to the mixed dielectric formed between the

oxide, water vapor, and air. However, at higher humidities, parallel conductance paths through the absorbed water are formed down the pore surfaces. Near saturation, this pore surface resistance becomes negligible, implying that the measured capacitance is virtually that between the very thin pore base and the aluminum core.

Tantalum Type Capacitive Humidity Sensors

In some versions of capacitive humidity sensors, one of the capacitor plates consists of a layer of tantalum deposited on a glass substrate. A layer of polymer dielectric is then added, followed by a second plate made from a thin layer of chromium. The chromium layer is under high tensile stress such that it cracks into a fine mosaic structure that allows water molecules to pass into the dielectric. The stress in the chromium also causes the polymer to crack into a mosaic structure. A sensor of this type has an input range of 0% to 100% relative humidity, RH. The capacitance is 375 pF at 0% RH and a linear sensitivity of 1.7 pF per % RH. The error is usually less than 2% due to nonlinearity and 1% due to hysteresis.

Silicon Type Capacitive Humidity Sensors

In other capacitive humidity sensors, silicon is used as the dielectric. The structure and operation of silicon humidity sensors are very similar to the aluminum oxide types. Some silicon-type humidity sensors also use the aluminum base and a thin-film gold layer as the two electrodes. The silicon dielectric has a very large surface area, which means that the sensitivity is still relatively large even if the sensing area is very small. This is an important feature with the increasing trend of miniaturization. Both sensor types are now typically found as extremely small wafer-shaped elements, placed on a mechanical mount with connecting lead wires. The formation of porous silicon is a very simple anodization process and, since no elaborate equipment is needed, devices can be made at relatively low cost. Also, by controlling the formation conditions, the structure of the porous silicon can easily be modified so devices can be tailored to suit particular applications.

In both the silicon and the aluminum oxide capacitive humidity sensors, the radii of the pores in the dielectric are such that they are specifically suited for water molecules. Most possible contaminants are too large in size to pollute the dielectric. However, contaminants can block the flow of water vapor into the sensor material, thus affecting the accuracy of the instrument. For example, in dust-contaminated streams, it may be possible to provide a simple physical barrier such as a sintered metal or plastic hoods for the sensor heads. Many sensors come with some form of casing to provide protection.

Polymer Type Capacitive Humidity Sensors

In some sensors, the dielectric consists of a polymer material that has the ability to absorb water molecules. The absorption of water vapor of the material results in changes in the dielectric constant of the capacitor. By careful design, the capacitance can be made directly proportional to percentage relative humidity of the surrounding gas or atmosphere.

In general, an important key feature of capacitive humidity sensors is the chemical stability. Often, humidity sensing is required in an air sample that contains vapor contaminants (e.g., carbon monoxide) or the measurements are performed on a gas sample other than air (e.g., vaporized benzene). The performance of these sensors, and in particular the silicon types, is not affected by many of these gases. Hydrocarbons, carbon dioxide, carbon monoxide, and CFCs do not cause interference. However, the ionic nature of the aluminum oxide dielectric makes it susceptible to certain highly polar, corrosive gases such as ammonia, sulphur trioxide, and chlorine. Silicon is inert; its stable nature means that these polar gases affect the sensor element to a far lesser degree.

Capacitive Moisture Sensors

Capacitive moisture measurements are based on the changes in the permittivity of granular or powder type dielectric materials such as wheat and other grains containing water. Usually, the sensor consists of a large cylindrical chamber, (e.g., 150 mm deep and 100 mm in diameter), as shown in Figure 6.33. The chamber is filled with samples under test. The variations in capacitance with respect to water content are processed. The capacitor is incorporated into an oscillatory circuit operating at a suitable frequency.

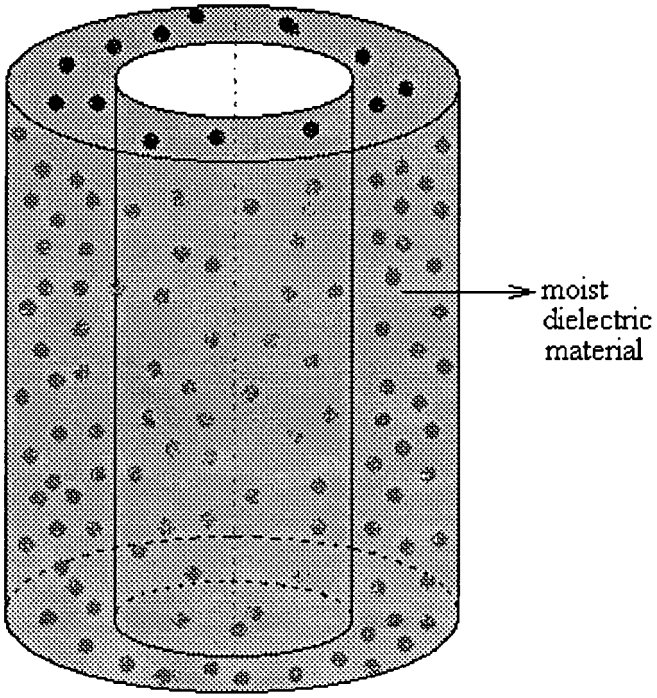


FIGURE 6.33 A capacitive moisture sensor. The permittivity of material between two cylindrical or parallel plates with fixed dimensions changes, depending on the moisture level of the materials in the chamber. The variations in capacitance values with respect to water content is processed. The capacitor is incorporated as a part of an oscillatory circuit operating at a suitable frequency, usually at radio frequencies.

Capacitive moisture sensors must be calibrated for samples made from different materials, as the materials themselves demonstrate different permittivities. Accurate temperature is necessary as the dielectric constant may be highly dependent on temperature. Most of these devices are built to operate at temperature ranges of 0°C to 50°C, supported by tight temperature compensation circuits. Once calibrated for a specific application, they are suitable for measuring moisture in the range of 0% to 40%.

Signal Processing

Generally, capacitive type pickups require relatively complex circuitry in comparison to many other sensor types, but they have the advantage of mechanical simplicity. They are also sensitive, having minimum mechanical loading effects. For signal processing, these sensors are usually incorporated either in ac deflection bridge circuits or oscillator circuits. In practice, capacitive sensors are not pure capacitances but have associated resistances representing losses in the dielectric. This can have an important influence in the design of circuits, particularly in oscillator circuits. Some of the signal processing circuits are discussed below.

Operational Amplifiers and Charge Amplifiers

One method of eliminating the nonlinearity of the relationship between the physical variable, (e.g., two-plate displacement sensors) and capacitance C is through the use of operational amplifiers, as illustrated in Figure 6.34. In this circuit, if the input impedance of the operational amplifier is high, the output is not saturated, and the input voltage is small, it is possible to write:

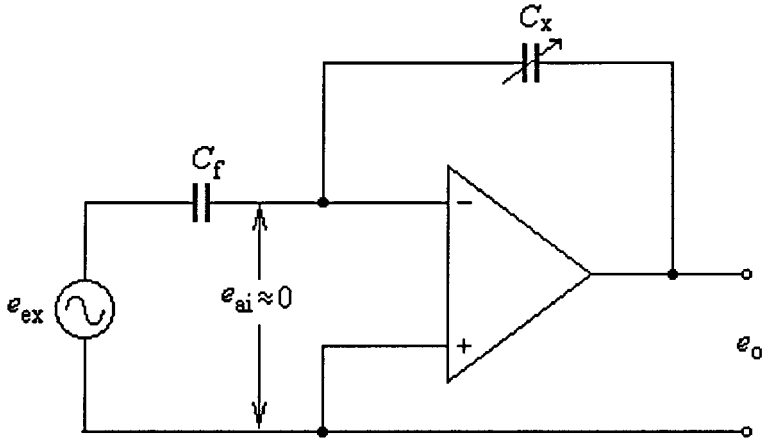


FIGURE 6.34 An operational amplifier signal processor. This method is useful to eliminate the nonlinearity in the signals generated by capacitive sensors. By this type of arrangement, the output voltage can be made directly proportional to variations in the signal representing the nonlinear operation of the device.

$$1/C_f = \int i_f \, dt = e_{ex} - e_{ai} = e_{ex} \tag{6.45}$$

$$1/C_x = \int i_x \, dt = e_0 - e_{ai} = e_0 \tag{6.46}$$

$$i_f + i_x - i_{ai} = 0 = i_f + i_x \tag{6.47}$$

Manipulation of these equations yields:

$$e_0 = -C_f \, e_{ex}/C_x \tag{6.48}$$

Substituting the value of C_x yields:

$$e_0 = -C_f x \, e_{ex}/\varepsilon A \tag{6.49}$$

Equation 6.49 shows that the output voltage is directly proportional to the plate separation x, thus giving linearity for all variations in motion.

However, a practical circuit requires a resistance across C_f to limit output drift. The value of this resistance must be greater than the impedance of C_f at the lowest frequency of interest. Also, because the transducer impedance is assumed to be purely capacitive, the effective gain is independent of frequency.

A practical charge amplifier circuit is depicted in Figure 6.35. In this case, the effective feedback resistance R_{ef} is given by:

$$R_{ef} = R_3\left(R_1 + R_2\right)/R_2 \tag{6.50}$$

It is possible to reduce the output drift substantially by selecting the resistors suitably. The accuracy of this circuit can be improved further by cascading two or more amplifiers. In this way, a substantial improvement in the signal-to-noise ratio can also be achieved. In the inverting input, the use of resistor R_4 is necessary because of bias currents.

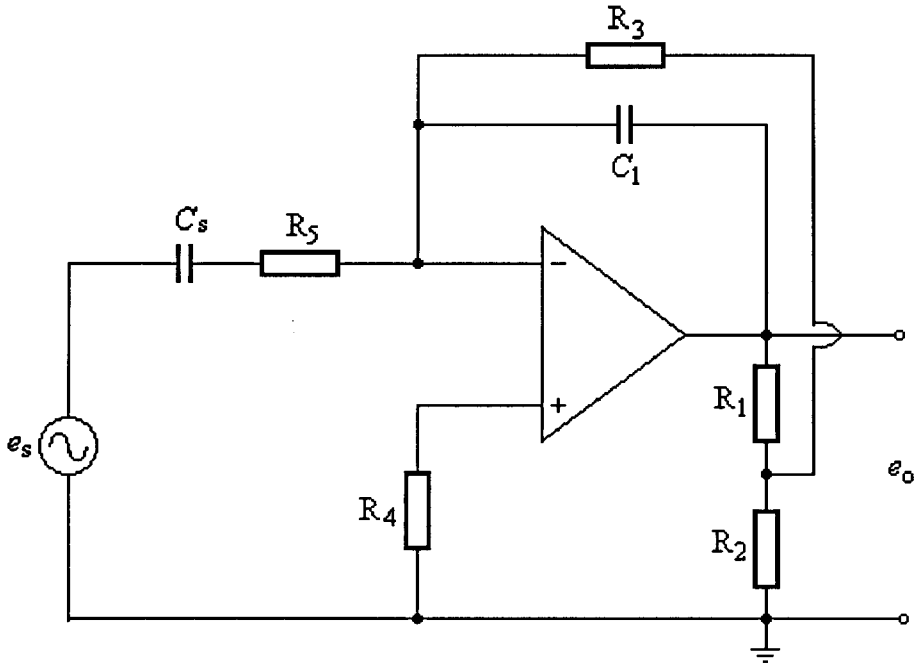


FIGURE 6.35 A practical charge amplifier. The effective feedback resistance is a function of other resistances. It is possible to reduce the output drift substantially by selecting the resistors suitably. The accuracy of this circuit can be improved further by cascading two or more amplifiers, thereby substantially improving the signal-to-noise ratio.

Pulse Width Modulation

As in the case of some capacitive vibrational displacement sensors, the output of the sensor may be an amplitude-modulated wave as shown in Figure 6.36. When rectified, the average value of this wave gives the mean separation of the plates. The vibration amplitude around this mean position may be extracted by a demodulator and a low-pass filter circuit. The output of the low-pass filter is a direct indication of vibrations, and the waveform can be viewed on an oscilloscope.

Square Wave Linearization

Another linearization technique applied in capacitive pressure transducers and accelerometers is pulse width modulation. The transducer consists of two differential capacitors as shown in Figure 6.37. The voltages of these capacitors, e_1 and e_2, switch back and forth with a high excitation frequency (e.g., 400 kHz) between excitation voltage and ground. The system is arranged in such a way that the output voltage is the average voltage difference between e_1 and e_2. At null position, $e_1 = e_2$, the output is a symmetrical square wave with zero average value. As the relative positions of the plates change, due to vibration, the average value of the output voltage shifts from the zero average value and becomes positive or negative depending on the direction of the displacement. Hence, the output voltage can be expressed by:

$$e_o = e_{ex}\left(C_1 - C_2\right)\Big/\left(C_1 + C_2\right) \tag{6.51}$$

Substituting:

$$C_1 = C_0\, x_0\Big/\left(x_0 - x_i\right) \quad \text{and} \quad C_2 = C_0\, x_0\Big/\left(x_0 + x_i\right)$$

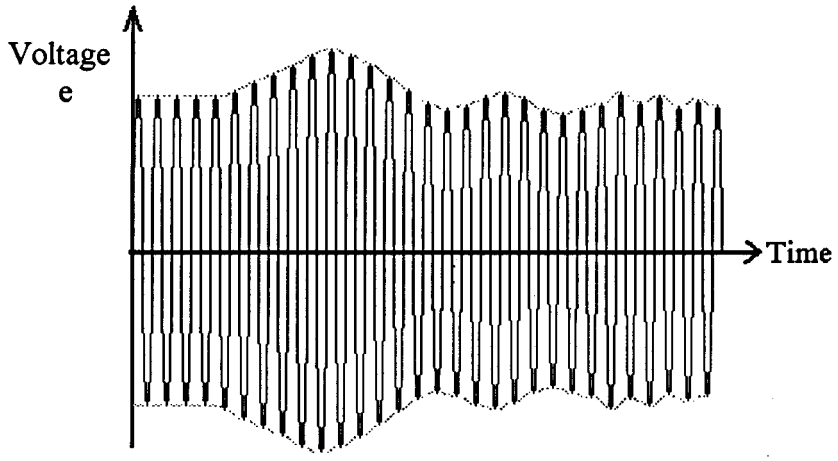


FIGURE 6.36 A amplitude modulated signal. It is possible to configure some sensors to give a amplitude-modulated signals, as in the case of capacitive vibrational displacement sensors. When rectified, the average value of this wave gives the mean separation of the plates. The vibration amplitude around this mean position can be extracted by a demodulator and low-pass filter circuit. The output of the low-pass filter is a direct indication of vibrations.

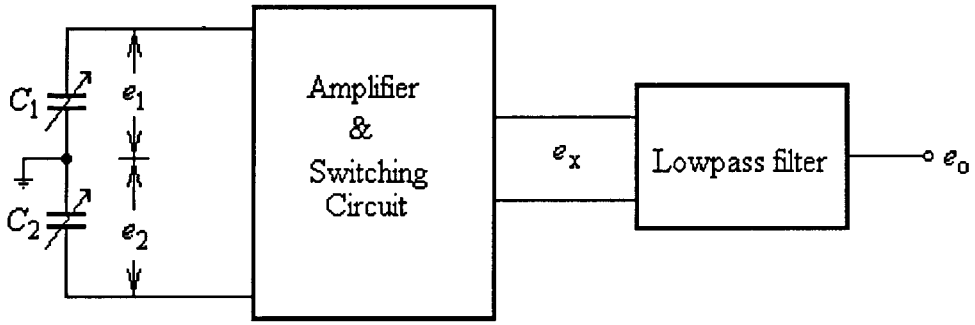


FIGURE 6.37 Block diagram of a square-wave linearization circuit. This is particularly useful for differential capacitance type sensors. The voltages of these two capacitors are made to switch back and forth with a high excitation frequency between excitation voltage and ground. As the relative positions of the plates change due to vibration, the average value of the output voltage becomes positive or negative, depending on the direction of the displacement.

yields

$$e_o = e_{ex}\, x_i / x_0 \tag{6.52}$$

Thus, the output is directly proportional to the variable x_i.

Feedback Linearization

Linearization of a capacitance transducer can also be obtained using a feedback system that adjusts capacitor current amplitude so that it stays constant at a reference value for all displacements. This is accomplished by obtaining a dc signal proportional to capacitor current from a demodulator, comparing this current with the reference current, and adjusting the voltage amplitude of the system excitation oscillator until the two currents agree. If the capacitor current is kept constant irrespective of capacitor motion, then the voltage amplitude is linearly related to x as:

$$e = K\, x_i \tag{6.53}$$

where

$$K = \left| i_c \right| \Big/ \omega\, C_0\, x_0 \tag{6.54}$$

Oscillator Circuits

In many applications, the resultant changes in the capacitance of capacitive transducers can be measured with a suitable ac bridge such as Wein bridge or Schering bridge. However, in a majority of cases, improvised versions of bridges are used as oscillator circuits for capacitive signal processing. The transducer is configured as a part of the oscillatory circuit that causes changes in the frequency of the oscillations. This change in frequency is scaled to be a measure of the magnitude of the physical variable.

As part of the oscillator circuits, the capacitive transducer has excellent frequency response and can measure both static and dynamic phenomena. Its disadvantages include sensitivity to temperature variations and the possibility of erratic or distorted signals due to long lead lengths. Also, the receiving instrumentation may be large and complex, and it often includes a second fixed-frequency oscillator for heterodyning purposes. The difference frequency thus produced can be read by an appropriate output device such as an electronic counter.

References

1. J. P. Bentley, *Principles of Measurement Systems,* 2nd ed., United Kingdom: Longman Scientific and Technical, 1988.
2. E. O. Doebelin, *Measurement Systems: Application and Design,* 4th ed., New York: McGraw-Hill, 1990.
3. J. P. Holman, *Experimental Methods for Engineers,* 5th ed., New York: McGraw-Hill, 1989.
4. F. T. Noth and G. C. M. Meijer, A Low-Cost, Smart Capacitive Position Sensor, IEEE Trans. Instrum. Meas., 41, 1041-1044, 1992.

Appendix to Section 6.3

List of Manufacturers

ANALITE Inc.
24-T Newtown Plaza
Plainview, NY 11803
Tel: (800) 229-3357

FSI/FORK Standards Inc.
668 Western Avenue
Lombard, IL 60148-2097
Tel: (708) 932-9380

Gordon Engineering Corp.
67 Del Mar Drive
Brookfiled, CT 06804
Tel: (203) 775-4501

Hecon Corp.
15-T Meridian Rd.
Eatontown, NJ 07724
Tel: (800) 524-1669

Kistler Instrumentation Corp.
Amherst, NY 14228-2171
Tel: (716) 691-5100
Fax: (716) 691-5226

Locon Sensor Systems, Inc.
1750 S. Eber Road
P.O. Box 789
Holland, OH 43526
Tel: (419) 865-7651
Fax: (419) 865-7756

Rechner Electronic Industries Inc.
8651 Buffalo Avenue, Box 7
Niagara Falls, NY 14304
Tel: (800) 544-4106

RDP Electrosense, Inc.
2216-Dept. B
Pottstown, PA
Tel: (800) 334-5838

6.4 Piezoelectric Transducers and Sensors

Ahmad Safari, Victor F. Janas, Amit Bandyopadhyay, and Andrei Kholkine

Piezoelectricity, discovered in Rochelle salt in 1880 by Jacques and Pierre Curie, is the term used to describe the ability of certain materials to develop an electric charge that is proportional to a direct applied mechanical stress. These materials also show the converse effect; that is, they will deform (strain) proportionally to an applied electric field. Some crystalline materials show piezoelectric behavior due to their unique crystal structure. The lattice structure of a crystal is described by the Bravais unit cell [1]. There are 230 microscopic symmetry types (space groups) in nature, based on the several symmetry elements such as translation, inversion center, mirror plane, or rotation axes. Combinations of these symmetry elements yield the macroscopic symmetry known as point groups. All natural crystals can be grouped into 32 different classes (point groups) based on their symmetry elements. The 32 point groups can be further classified into two subgroups: (1) crystals with a center of symmetry, and (2) crystals with no center of symmetry. The 11 centrosymmetric subgroups do not show piezoelectricity. Of the 21 noncentrosymmetric groups, 20 show the piezoelectric effect along unique directional axes. An important class of piezoelectric materials includes ferroelectrics, in which the piezoelectric effect is closely related to the ferroelectric polarization that can be reversed by the application of sufficiently high electric field [2, 3]. To induce piezoelectric properties in ferroelectric materials, a poling procedure is often required, which consists of the temporary application of a strong electric field. Poling is analogous to the magnetizing of a permanent magnet.

Governing Equations and Coefficients

The phenomenological master equation describing the deformations of an insulating crystal subject to both elastic and electric stress is given by:

$$x_{ij} = s_{ijkl} X_{kl} + d_{mij} E_m + M_{mnij} E_m E_n , \tag{6.55}$$

where x_{ij} are components of elastic strain, s_{ijkl} is the elastic compliance tensor, X_{kl} are the stress components, d_{mij} are the piezoelectric tensor components, M_{mnij} is the electrostrictive tensor, and E_m and E_n are components of the electric field.

Neglecting the second-order effects (electrostriction) and assuming that the material is under no stress ($X_{kl} = 0$), the elastic strain is given by:

$$x_{ij} = d_{mij} E_m \tag{6.56}$$

Equation 6.56 is the mathematical definition of the converse piezoelectric effect, where induced strain is directly proportional to the first power of the field. The thermodynamically equivalent direct piezoelectric effect is given by:

$$P_m = d_{mij} X_{ij} , \tag{6.57}$$

where P_m is the component of electrical polarization. The difference between the direct and converse piezoelectric effect is shown schematically in Figure 6.38. The converse effect describes the actuating function of a piezoelectric, where a controlled electric field accurately changes the shape of a piezoelectric material. The sensing function of a piezoelectric is described by the direct effect, where a controlled stress on a piezoelectric material yields a charge proportional to the stress.

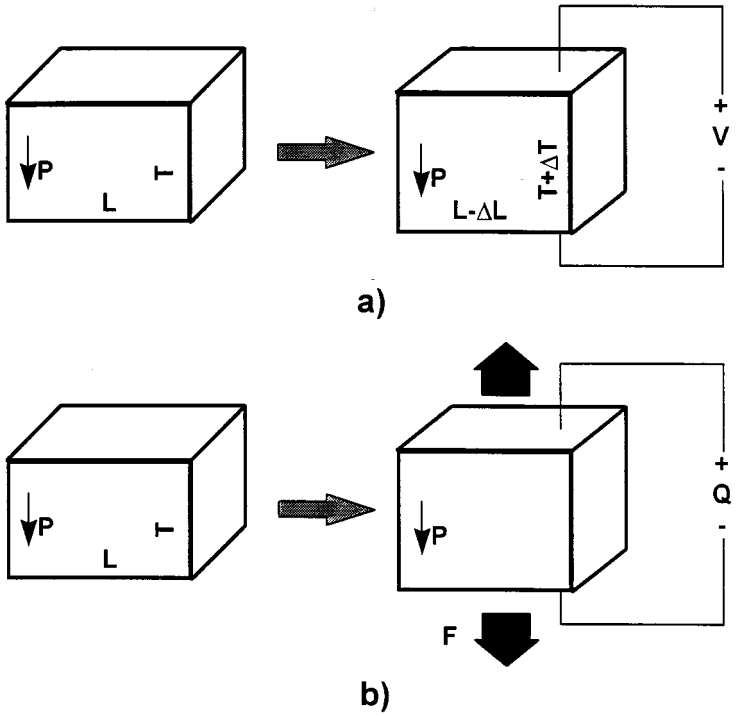


FIGURE 6.38 Schematic representations of the direct and converse piezoelectric effect: (*a*) an electric field applied to the material changes its shape; (*b*) a stress on the material yields a surface charge.

The third rank tensor d_{mij} may be simplified using matrix notation [1] to d_{ij}, a second rank tensor. In this form, d_{ij} is simply known as the piezoelectric charge coefficient, with units of coulombs per newton (CN^{-1}) or meters per volt ($m\,V^{-1}$). Another set of moduli that may be used to characterize piezoelectric materials are the piezoelectric voltage coefficients, g_{ij}, defined in matrix notation as:

$$E_i = g_{ij} X_j, \tag{6.58}$$

where E_i is the component of electric field arising due to the stress X_j. The d_{ij} and g_{ij} coefficients are related by:

$$g_{ij} = d_{ij} \Big/ \left(\varepsilon_0 \varepsilon_{ii}\right), \tag{6.59}$$

where ε_{ii} is the dielectric constant (relative permittivity), and ε_0 is the permittivity of free space (8.854×10^{-12} F m^{-1}). Another key property of piezoelectric materials is their electromechanical coupling coefficient *k*, defined as:

$$k^2 = \text{resulting mechanical energy/input electrical energy} \tag{6.60}$$

or

$$k^2 = \text{resulting electrical energy/input mechanical energy} \tag{6.61}$$

The value of *k* represents the efficiency of the piezoelectric in converting one form of energy to another. Since energy conversion is never complete, the value of k^2 is always less than unity, so *k* is always less than 1.

Two final important parameters for piezoelectric (ferroelectric) materials are their Curie point (T_0) and Curie temperature (T_c). The Curie point is the temperature above which the material loses its ferroelectric and piezoelectric behavior. The Curie temperature is defined by the Curie-Weiss law:

$$\varepsilon = \varepsilon_0 + \frac{C}{T - T_c} \qquad \text{for } T > T_c\,, \tag{6.62}$$

where ε is the dielectric constant of the material, C is the Curie–Weiss constant, and T is the temperature. It represents the temperature where the material tends to have its highest dielectric constant. The Curie temperature is always lower (often within 10°C) than the Curie point.

Piezoelectric Materials

Single Crystals

A number of single-crystal materials have demonstrated piezoelectricity. These materials dominate certain applications, such as frequency-stabilized oscillators in watches and radars, and surface acoustic wave devices in television filters and analog signal correlators. A list of single-crystal piezoelectric materials includes quartz, lithium niobate and lithium tantalate, ammonium dihydrogen sulfate, lithium sulfate monohydrate, and Rochelle salt. Recently, it was discovered that some relaxor-based ferroelectric single crystals of lead zinc niobate and lead magnesium niobate, and their solid solutions with lead titanate possess superior piezoelectric properties when compared to other piezoelectric materials.

Quartz, which is crystalline SiO_2, has a low value of d_{11} ($|2.3 \times 10^{-12}|$ C N^{-1}) [1]. Right-handed quartz develops a positive charge when put under compression, and a negative charge when put under tension. The coupling coefficient k for quartz is also very low, typically around 0.1 [4]. In addition, the dielectric constant ε for quartz is small (~4) [5]. The Curie point, however, is relatively high ($T_0 \sim 573$°C) [5], so quartz is stable for high-temperature applications. Despite the low piezoelectric properties, quartz is very abundant and inexpensive, and has found a strong position in low-cost or high-temperature applications.

Piezoelectric behavior in lithium niobate ($LiNbO_3$) and lithium tantalate ($LiTaO_3$) was first studied in the mid-1960s [2]. Under shear, the d_{15} of $LiNbO_3$ and $LiTaO_3$ are 73 and 26 ($\times 10^{-12}$ C N^{-1}), respectively. Both have ε values of approximately 40. If cut correctly, they have coupling coefficient (k) values of 0.65 and 0.4, respectively. In addition, the Curie points for both are extremely high ($T_0 \sim 1210$°C for $LiNbO_3$, and 620°C for $LiTaO_3$). Both $LiNbO_3$ and $LiTaO_3$ are commonly used in infrared detectors.

Rochelle salt ($KNaC_4H_4O_6 \cdot H_2O$) was first found to be piezoelectric in 1880 [2]. The d_{31} and k_{31} are 275×10^{-12} C N^{-1} and 0.65, respectively. The relative dielectric constant is approximately 350. Rochelle salt has two Curie points (lower $T_0 \sim 18$°C, and upper $T_0 \sim 24$°C). It is highly soluble in water, and is still extensively used in electroacoustic transducers.

Lead zinc niobate, $Pb(Zn_{1/3}Nb_{2/3})O_3$, and lead magnesium niobate, $Pb(Mg_{1/3}Nb_{2/3})O_3$, are typical relaxor materials characterized by the broad frequency-dependent maximum of dielectric constant vs. temperature. The solid solutions of these materials with lead titanate, $PbTiO_3$, were shown to posses excellent piezoelectric properties when oriented along the [001] direction [6]. The piezoelectric charge coefficient d_{33} of 25×10^{-10} C N^{-1}, coupling coefficient k of more than 0.9, and ultrahigh strain of 1.7% were achieved in $Pb(Zn_{1/3}Nb_{2/3})O_3$-$PbTiO_3$ solid solution. These single-crystal relaxor materials are now being intensively investigated and show great promise for future generations of piezoelectric transducers and sensors.

Other types of piezoelectric materials dominate the market for transducers. These materials include piezoelectric ceramics, piezoelectric polymers, and composites of piezoelectric ceramic with inactive polymers. The focus of the remainder of this chapter will be on non-single-crystal piezoelectric materials.

Piezoelectric Ceramics

In polycrystalline ceramics with polar grains, the randomness of the grains, as shown schematically in Figure 6.39(*a*), yields a nonpiezoelectric material. Piezoelectric behavior is induced by "poling" the

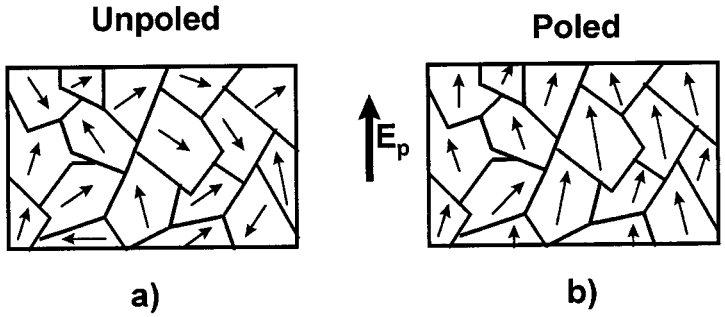


FIGURE 6.39 Schematic of the poling process in piezoelectric ceramics: (*a*) in the absence of an electric field, the domains have random orientation of polarization; (*b*) the polarization within the domains are aligned in the direction of the electric field.

ceramic. By applying a strong dc electric field at a temperature just below the Curie temperature, the spontaneous polarization in each grain gets oriented toward the direction of the applied field. This is schematically shown in Figure 6.39(*b*). Although all of the domains in a ceramic can never be fully aligned along the poling axis due to symmetry limitations, the ceramic ends up with a net polarization along the poling axis.

The largest class of piezoelectric ceramics is made up of mixed oxides containing corner-sharing octahedra of O^{2-} ions. The largest structure type, built with corner-shared oxygen octahedra, is the perovskite family, which is discussed in the following section.

Perovskites

Perovskite is the name given to a group of materials with general formula ABO_3 having the same structure as the mineral calcium titanate ($CaTiO_3$). Piezoelectric ceramics having this structure include barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$), lead zirconate titanate ($PbZr_xTi_{1-x}O_3$, or PZT), lead lanthanum zirconate titanate [$Pb_{1-x}La_x(Zr_yT_{1-y})_{1-x/4}O_3$, or PLZT], and lead magnesium niobate [$PbMg_{1/3}Nb_{2/3}O_3$, or PMN]. Several of these ceramics are discussed below.

The piezoelectric effect in $BaTiO_3$ was discovered in the 1940s [4], and it became the first piezoelectric ceramic developed. It replaced Rochelle salt because it is more stable, has a wider temperature range of operation, and is easily manufacturable. The Curie point, T_0, is about 130°C. Above 130°C, a nonpiezoelectric cubic phase is stable, where the center of positive charge (Ba^{2+} and Ti^{4+}) coincides with the center of the negative charge (O^{2-}) (Figure 6.40(*a*)). When cooled below the Curie point, a tetragonal structure (shown in Figure 6.40(*b*)) develops where the center of positive charge is displaced relative to the O^{2-} ions, leading to the formation of electric dipoles. Barium titanate has a relative dielectric constant ε_{33} of 1400 when unpoled, and 1900 when poled [2, 4]. The d_{15} and d_{33} coefficients of $BaTiO_3$ are 270 and 191×10^{-12} C N^{-1}, respectively. The k for $BaTiO_3$ is approximately 0.5. The large room temperature dielectric constant in barium titanate has led to its wide use in multilayer capacitor applications.

Lead titanate, $PbTiO_3$, first reported to be ferroelectric in 1950 [4], has a similar structure to $BaTiO_3$, but with a significantly higher Curie point ($T_0 = 490$°C). Pure lead titanate is difficult to fabricate in bulk form. When cooled through the Curie point, the grains go through a cubic to tetragonal phase change, leading to large strain and ceramic fracturing. This spontaneous strain can be decreased by the addition of dopants such as Ca, Sr, Ba, Sn, and W. Calcium-doped $PbTiO_3$ [7] has a relative dielectric constant ε_{33} of 200, a d_{33} of 65×10^{-12} C/N, and a k of approximately 0.5. The addition of calcium results in a lowering of the Curie point to 225°C. The main applications of lead titanate are hydrophones and sonobuoys.

Lead zirconate titanate (PZT) is a binary solid solution of $PbZrO_3$ (an antiferroelectric orthorhombic structure) and $PbTiO_3$ (a ferroelectric tetragonal perovskite structure) [2–4]. It has a perovskite structure, with the Zr^{4+} and Ti^{4+} ions occupying the B site of the ABO_3 structure at random. At the morphotropic

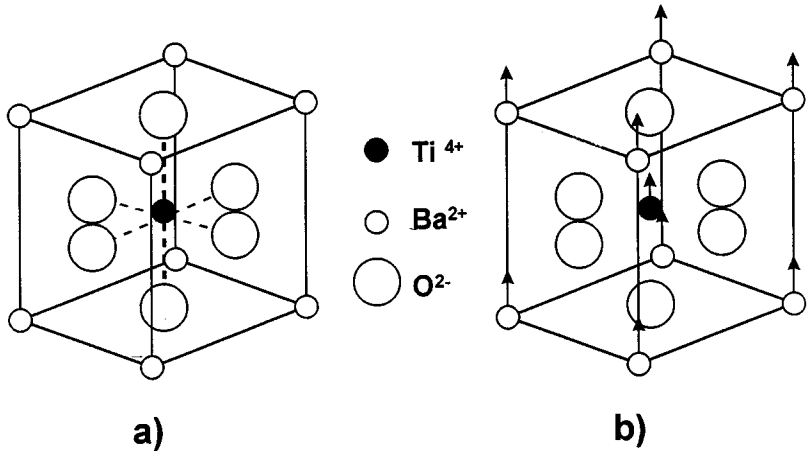


FIGURE 6.40 The crystal structure of $BaTiO_3$: (*a*) above the Curie point, the cell is cubic; (*b*) below the Curie point, the cell is tetragonal with Ba^{2+} and Ti^{4+} ions displaced relative to O^{2-} ions.

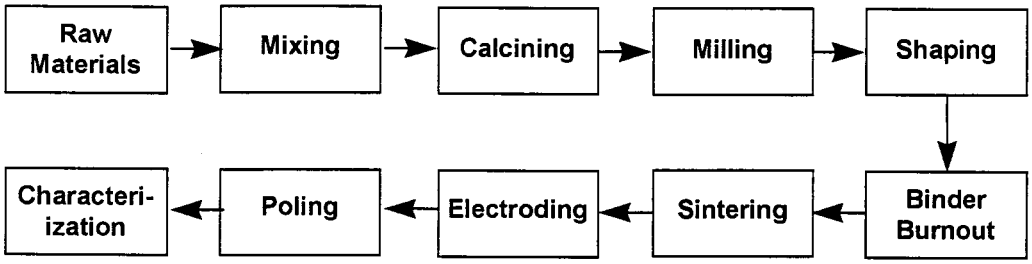


FIGURE 6.41 Flow chart for the processing of piezoelectric ceramics.

phase boundary (MPB) separating the tetragonal and orthorhombic phases, PZT shows excellent piezoelectric properties. At room temperature, the MPB is at a Zr/Ti ratio of 52/48, resulting in a piezoelectric ceramic which is extremely easy to pole. Piezoelectric PZT at the MPB is usually doped by a variety of ions to form what are known as "hard" and "soft" PZTs. Hard PZT is doped with acceptor ions, such as K^+ or Na^+ at the A site, or Fe^{3+}, Al^{3+}, or Mn^{3+} at the B site. This doping lowers the piezoelectric properties, and makes the PZT more difficult to pole or depole. Typical piezoelectric properties of hard PZT include [5, 7]: Curie point, T_0, of 365°C, ε_{33} of 1700–1750 (poled), a piezoelectric charge coefficient d_{33} of 360 to 370×10^{-12} C N^{-1}, and a coupling coefficient of about 0.7. Soft PZT is doped with donor ions such as La^{3+} at the A site, or Nb^{5+} or Sb^{5+} at the B site. It has very high piezoelectric properties, and is easy to pole or depole. Typical piezoelectric properties of soft PZT include [5, 7]: Curie point, T_0, of 210°C, relative dielectric constant ε_{33} of 3200–3400 (poled), a d_{33} of 580 to 600×10^{-12} C N^{-1}, and a coupling coefficient k_{33} of 0.7.

Processing of Piezoelectric Ceramics

The electromechanical properties of piezoelectric ceramics are largely influenced by their processing conditions. Each step of the process must be carefully controlled to yield the best product. Figure 6.41 is a flow chart of a typical oxide manufacturing process for piezoelectric ceramics. The high-purity raw materials are accurately weighed according to their desired ratio, and mechanically or chemically mixed. During the calcination step, the solid phases react to yield the piezoelectric phase. After calcining, the solid mixture is ground into fine particles by milling. Shaping is accomplished by a variety of ceramic

TABLE 6.7 Advantages (+) and Disadvantages (–) of Piezoelectric Ceramics, Polymers and Composites

Parameter	Ceramic	Polymer	Ceramic/Polymer Composite
Acoustic impedance	High (–)	Low (+)	Low (+)
Coupling factor	High (+)	Low (–)	High (+)
Spurious modes	Many (–)	Few (+)	Few (+)
Dielectric constant	High (+)	Low (–)	Medium (+)
Flexibility	Stiff (–)	Flexible (+)	Flexible (+)
Cost	Cheap (+)	Expensive (–)	Medium (+)

Adapted from T. R. Gururaja, *Amer. Ceram. Soc. Bull.*, 73, 50, 1994.

processing techniques, including powder compaction, tape casting, slip casting, or extrusion. During the shaping operation, organic materials are typically added to the ceramic powder to improve its flow and binding characteristics. These organics are removed in a low temperature (500 to 600°C) binder burn-off step.

After burnout, the ceramic structure is sintered to an optimum density at an elevated temperature. For the lead-containing piezoelectric ceramics ($PbTiO_3$, PZT, PLZT), sintering is performed in sealed crucibles with an optimized PbO atmosphere. This is because lead loss occurs in these ceramics above 800°C. As mentioned earlier (Figure 6.39), the randomness of the ceramic grains yields a nonpiezoelectric material. By electroding the ceramic and applying a strong dc electric field at high temperature, the ceramic is poled. At this point, the piezoelectric ceramic is ready for final finishing and characterization.

Piezoelectric Polymers

The piezoelectric behavior of polymers was first reported in 1969 [8]. The behavior results from the crystalline regions formed in these polymers during solidification from the melt. When the polymer is drawn, or stretched, the regions become polar, and can be poled by applying a high electric field. The most widely known piezoelectric polymers are polyvinylidene fluoride [9, 10], also known as PVDF, polyvinylidene fluoride — trifluoroethylene copolymer, or P(VDF-TrFE) [9, 10], and odd-number nylons, such as Nylon-11 [11].

The electromechanical properties of piezoelectric polymers are significantly lower than those of piezoelectric ceramics. The d_{33} values for PVDF and P(VDF-TrFE) are approximately 33 ($\times 10^{-12}$ C N^{-1}), and the dielectric constant ε is in the range 6 to 12 [12, 13]. They both have a coupling coefficient (k) of 0.20, and a Curie point (T_0) of approximately 100°C. For Nylon-11, ε is around 2 [11], while k is approximately 0.11.

Piezoelectric Ceramic/Polymer Composites

As mentioned above, a number of single-crystal, ceramic, and polymer materials exhibit piezoelectric behavior. In addition to the monolithic materials, composites of piezoelectric ceramics with polymers have also been formed. Table 6.7 [14] summarizes the advantages and disadvantages of each type of material. Ceramics are less expensive and easier to fabricate than polymers or composites. They also have relatively high dielectric constants and good electromechanical coupling. However, they have high acoustic impedance, and are therefore a poor acoustic match to water, the media through which it is typically transmitting or receiving a signal. Also, since they are stiff and brittle, monolithic ceramics cannot be formed onto curved surfaces, limiting design flexibility in the transducer. Finally, they have a high degree of noise associated with their resonant modes. Piezoelectric polymers are acoustically well matched to water, are very flexible, and have few spurious modes. However, applications for these polymers are limited by their low electromechanical coupling, low dielectric constant, and high cost of fabrication. Piezoelectric ceramic/polymer composites have shown superior properties when compared to single-phase materials. As shown in Table 6.7, they combine high coupling, low impedance, few spurious modes, and an intermediate dielectric constant. In addition, they are flexible and moderately priced.

TABLE 6.8 Suppliers of Piezoelectric Materials and Sensors

Name	Address	Ceramic	Polymer	Composite
AMP Sensors	950 Forge Ave. Morristown, PA 19403 Phone: (610) 650-1500 Fax: (610) 650-1509		X	
Krautkramer Branson	50 Industrial Park Rd. Lewistown, PA 17044 Phone: (717) 242-0327 Fax: (717) 242-2606			X
Materials Systems, Inc.	531 Great Road Littleton, MA 01460 Phone: (508) 486-0404 Fax: (508) 486-0706			X
Morgan Matroc, Inc.	232 Forbes Rd. Bedford, OH 44146 Phone: (216) 232-8600 Fax: (216) 232-8731	X		
Sensor Technology Ltd.	20 Stewart Rd. P.O. Box 97 Collingwood, Ontario, Canada Phone: +1 (705) 444-1440 Fax: +1 (705) 444-6787	X		
Staveley Sensors, Inc.	91 Prestige Park Circle East Hartford, CT 06108 Phone: (860) 289-5428 Fax: (860) 289-3189			X
Valpey-Fisher Corporation	75 South Street Hopkinton, MA 01748 Phone: (508) 435-6831 Fax: (508) 435-5289		X	
Vermon U.S.A.	6288 SR 103 North Bldg. 37 Lewistown, PA 17044 Phone: (717) 248-6838 Fax: (717) 248-7066	X		
TRS Ceramics, Inc.	2820 E. College Ave. State College, PA 16801 Phone: (814) 238-7485 Fax: (814) 238-7539	X		

Suppliers of Piezoelectric Materials

Table 6.8 lists a number of the suppliers of piezoelectric materials, their addresses, and whether they supply piezoelectric ceramic, polymers, or composites. Most of them tailor the material to specific applications.

Measurements of Piezoelectric Effect

Different means have been proposed to characterize the piezoelectric properties of materials. The resonance technique involves the measurement of the characteristic frequencies when the suitably shaped specimen (usually ceramic) is driven by a sinusoidally varying electric field. To a first approximation, the behavior of a piezoelectric sample close to its fundamental resonance frequency can be represented by an equivalent circuit as shown in Figure 6.42(*a*). The schematic behavior of the reactance of the sample as a function of frequency is represented in Figure 6.42(*b*). By measuring the characteristic frequencies

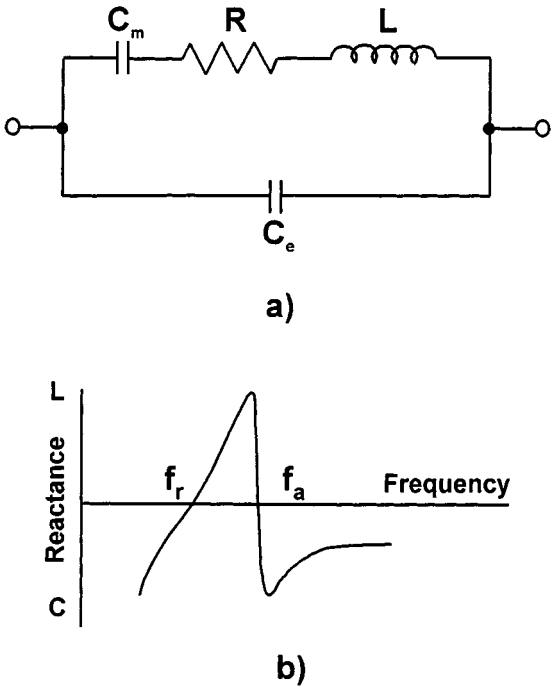


FIGURE 6.42 (*a*) Equivalent circuit of the piezoelectric sample near its fundamental electromechanical resonance (top branch represents the mechanical part and bottom branch represents the electrical part of the circuit); (*b*) electrical reactance of the sample as a function of frequency.

of the sample, the material constants including piezoelectric coefficients can be calculated. The equations used for the calculations of the electromechanical properties are described in the IEEE Standard on piezoelectricity [15]. The simplest example of piezoelectric measurements by resonance technique relates to a piezoelectric ceramic rod (typically 6 mm in diameter and 15 mm long) poled along its length. It can be shown that the coupling coefficient k_{33} is expressed as a function of the series and parallel resonance frequencies, f_s and f_p, respectively:

$$k_{33}^2 = \frac{\pi}{2}\frac{f_s}{f_p}\tan\left(\frac{\pi}{2}\frac{f_p - f_s}{f_p}\right) \tag{6.63}$$

The longitudinal piezoelectric coefficient d_{33} is calculated using k_{33}, elastic compliance s_{33}^E and low-frequency dielectric constant ε_{33}^X:

$$d_{33} = k_{33}\left(\varepsilon_{33}^X s_{33}^E\right)^{1/2} \tag{6.64}$$

Similarly, other electromechanical coupling coefficients and piezoelectric moduli can be derived using different vibration modes of the sample. The disadvantage of the resonance technique is that measurements are limited to the specific frequencies determined by the electromechanical resonance. It is used mostly for the rapid evaluation of the piezoelectric properties of ceramic samples whose dimensions can be easily adjusted for specific resonance conditions.

Subresonance techniques are frequently used to evaluate piezoelectric properties of materials at frequencies much lower than the fundamental resonance frequency of the sample. They include both the measurement of piezoelectric charge under the action of external mechanical force (direct effect) and the

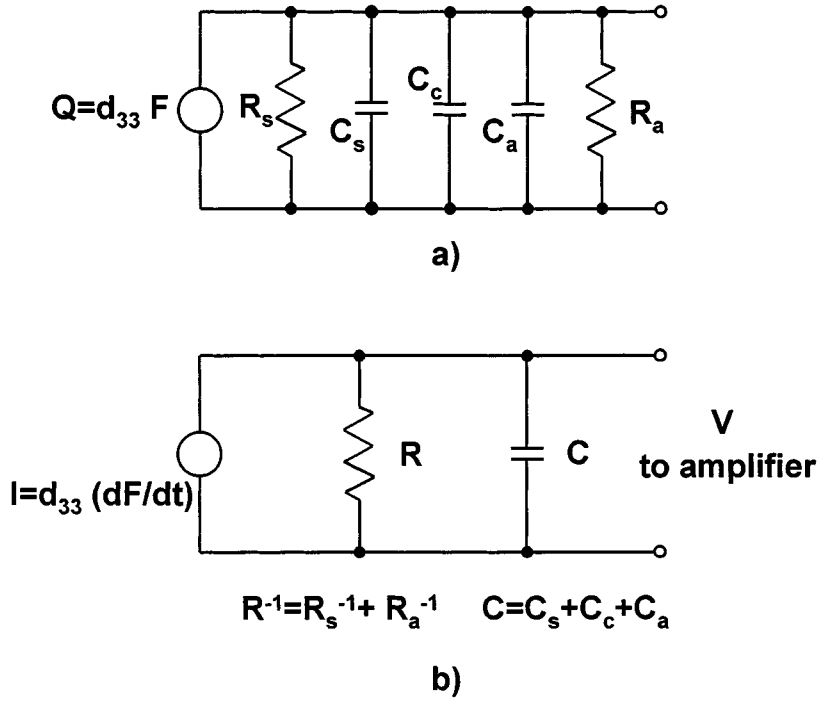


FIGURE 6.43 Full (*a*) and simplified (*b*) equivalent electrical circuits of the piezoelectric sensor connected to the voltage amplifier.

measurement of electric field-induced displacements (converse effect). In the latter case, the displacements are much smaller than in resonance; however, they still can be measured by using strain gages, capacitive sensors, LVDT (linear variable differential transformer) sensors or by optical interferometry [16, 17].

A direct method is widely used to evaluate the sensor capabilities of piezoelectric materials at sufficiently low frequency. The mechanical deformations can be applied in different modes such as thickness expansion, transverse expansion, thickness shear, and face shear to obtain different components of the piezoelectric tensor. In the simplest case, the metal electrodes are placed onto the major surfaces of the piezoelectric transducer normal to its poling direction (direction of ferroelectric polarization) and the mechanical force is applied along this direction (Figure 6.38(*b*)). Thus, the charge is produced on the electrode plates under mechanical loading, which is proportional to the longitudinal piezoelectric coefficient d_{33} of the material. To relate the output voltage of the transducer to the piezoelectric charge, it is necessary to consider the equivalent circuit (Figure 6.43(*a*)). A circuit includes the charge generator, $Q = d_{33}F$, leakage resistor of the transducer, R_s, transducer capacitance, C_s, capacitance of the connecting cables, C_c, and input resistance and capacitance of the amplifier, R_a and C_a, respectively. Here, F denotes the force applied to the transducer (tensile or compressive). All the resistances and capacitances shown in Figure 6.43(*a*) can be combined, as shown in Figure 6.43(*b*). A charge generator can be converted to a current generator, I, according to:

$$I = \frac{dQ}{dt} = d_{33}\frac{dF}{dt} \tag{6.65}$$

Assuming that the amplifier does not draw any current, the output voltage V at a given frequency ω can be calculated:

$$V = \frac{d_{33}F}{C}\frac{j\omega\tau}{1+j\omega\tau}, \tag{6.66}$$

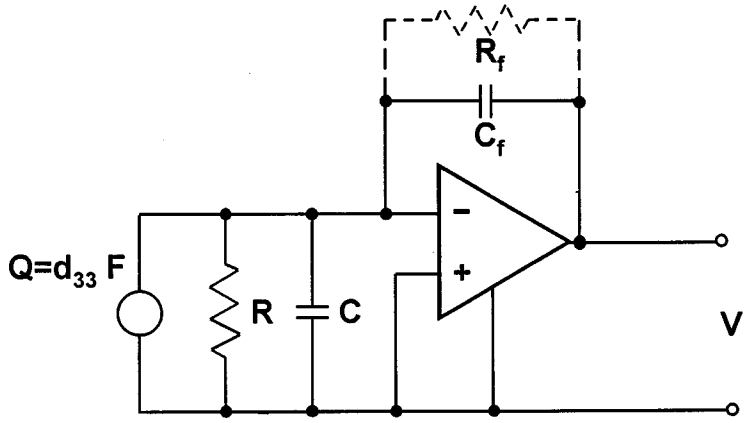


FIGURE 6.44 Equivalent electrical circuit of the piezoelectric sensor connected to the charge amplifier.

where $\tau = RC$ is the time constant that depends on all resistances and capacitances of the circuit. For sufficiently high frequency, the measured response is frequency independent and d_{33} can be easily evaluated from Equation 6.66 if the equivalent capacitance C is known. Since C is determined by the parallel capacitances of the sample, connecting cables, and amplifier (typically not exactly known), the standard capacitance is often added to the circuit, which is much greater than all the capacitances involved. However, according to Equation 6.66, the sensitivity of the circuit is greatly reduced with decreasing C. If τ is not large enough, the low-frequency cut-off does not allow piezoelectric measurements in quasi-static or low-frequency conditions.

To overcome the difficulties of using voltage amplifiers for piezoelectric measurements, a so-called charge amplifier was proposed. The idealized circuit of a charge amplifier connected with the piezoelectric transducer is shown in Figure 6.44. Note that a FET-input operational amplifier is used with a capacitor C_f in the feedback loop. Assuming that the input current and voltage of the operational amplifier are negligible, one can relate the charge on the transducer with the output voltage:

$$V = -Q/C_f = -d_{33}F/C_f \tag{6.67}$$

Equation 6.67 gives frequency-independent response where the output voltage is determined only by the piezoelectric coefficient d_{33} and the known capacitance C_f. Unfortunately, this advantage is difficult to realize since even a small input current of the amplifier will charge the feedback capacitor, leading to saturation of the amplifier. Therefore, a shunt resistor R_f is added to the circuit (dotted line in Figure 6.44), which prevents such charging. If one takes into account the RC circuit of the feedback loop, the output voltage will have the form of Equation 6.66, i.e., becomes frequency dependent. In this case, the time constant τ is determined by the parameters of the feedback loop and does not depend on the capacitance of the transducer, connecting cables, or the input capacitance of the amplifier. This gives an important advantage to the charge amplifier when it is compared to the ordinary voltage amplifier.

Applications

The direct and converse piezoelectric effects in a number of materials have led to their use in electromechanical transducers. Electromechanical transducers convert electrical energy to mechanical energy, and vice versa. These transducers have found applications where they are used in either passive or active modes. In the passive (sensor) mode, the transducer only receives signals. Here, the direct piezoelectric properties of the material are being exploited to obtain a voltage from an external stress. Applications in the passive mode include hydrophones, or underwater listening devices, microphones, phonograph pickups, gas igniters, dynamic strain gages, and vibrational sensors. In the active (actuator) mode, the

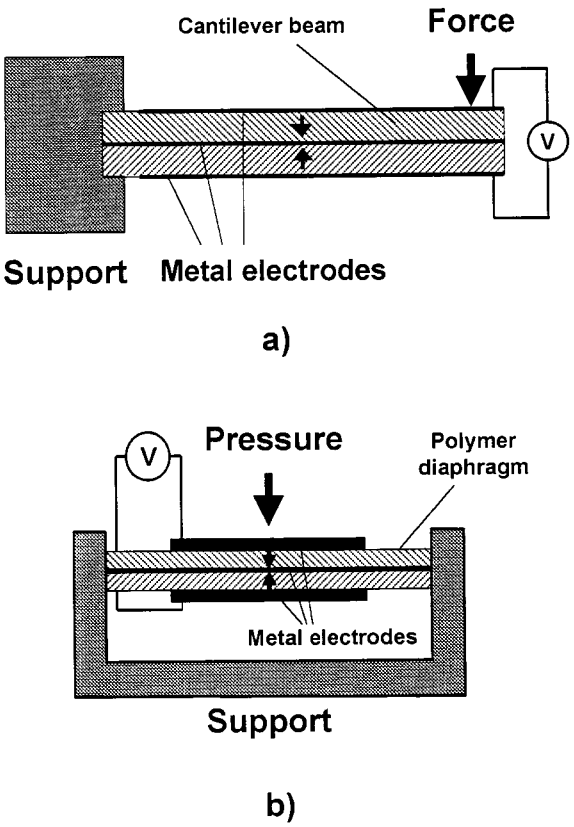


FIGURE 6.45 Schematic designs of the displacement sensor based on piezoelectric ceramic (*a*) and of the pressure sensor based on piezoelectric polymer film (*b*). Arrows indicate the directions of ferroelectric polarization in the piezoelectric material.

transducer, using the converse piezoelectric properties of the material, changes its dimensions and sends an acoustic signal into a medium. Active mode applications include nondestructive evaluation, fish/depth finders, ink jet printers, micropositioners, micropumps, and medical ultrasonic imaging. Often, the same transducer is used for both sensor and actuator functions.

Two examples of piezoelectric sensors are given below. The first example is the ceramic transducer, which relates the deformation of the piezoelectric sensor to the output voltage via direct piezoelectric effect. Piezoceramics have high Young's moduli; therefore, large forces are required to generate strains in the transducer to produce measurable electric response. Compliance of the piezoelectric sensor can be greatly enhanced by making long strips or thin plates of the material and mounting them as cantilevers or diaphragms. Displacement of the cantilever end will result in a beam bending, leading to the mechanical stress in the piezoelectric material and the electric charge on the electrodes. A common configuration of the piezoelectric bender is shown in Figure 6.45(*a*). Two beams poled in opposite directions are cemented together with one common electrode in the middle and two electrodes on the outer surfaces. Bending of such a bimorph will cause the upper beam to stretch and the lower beam to compress, resulting in a piezoelectric charge of the same polarity for two beams connected in series. To the first approximation, the charge Q appearing on the electrodes is proportional to the displacement Δl of the end of the bimorph via Equation 6.68 [18]:

$$Q = \frac{3}{8}\frac{Hw}{L} e_{31} \Delta l \,, \tag{6.68}$$

where H, w, and L are the thickness, the width, and the length of the bimorph, respectively, and e_{31} is the transverse piezoelectric coefficient relating electric polarization and strain in a deformed piezoelectric material. The charge can be measured either by the voltage amplifier (Figure 6.43) or by the charge amplifier (Figure 6.44).

In certain applications, the parameters of piezoelectric sensors can be improved by using ferroelectric polymers instead of single crystals and piezoceramics. Although the electromechanical properties of polymers are inferior to those of piezoelectric ceramics, their low dielectric constant offers the higher voltage response since they possess higher g piezoelectric coefficients. Also, the polymers are more mechanically robust and can be made in the form of thin layers (down to several micrometers). An example using the polymer bimorph as a pressure sensor is shown in Figure 6.45(*b*). A circular diaphragm composed of two oppositely poled polymer films is clamped along its edges to a rigid surround, forming a microphone. The voltage appearing on the electrodes is proportional to the applied pressure p by Equation 6.69 [19]:

$$V = \frac{3}{16}\frac{d_{31}}{\varepsilon_{33}}\frac{D^2}{h}\left(1-\nu\right)p \tag{6.69}$$

where D and h are the diameter and thickness of the diaphragm, respectively, and ν is the Poisson ratio. The high d_{31}/ε_{33} value for polymer sensors is advantageous to obtain higher voltage response. According to Equation 6.66, this advantage can be realized only if the high input impedance amplifier is used in close proximity to the transducer to reduce the influence of the connecting cables.

Defining Terms

Piezoelectric transducer: Device that converts the input electrical energy into mechanical energy and vice versa via piezoelectric effect.

Coupling coefficients: Materials constants that describe an ability of piezoelectric materials to convert electrical energy into mechanical energy and vice versa.

Piezoelectric coefficients: Materials constants that are used to describe the linear coupling between electrical and mechanical parameters of the piezoelectric.

Ferroelectrics: Subgroup of piezoelectric materials possessing a net dipole moment (ferroelectric polarization) that can be reversed by the application of sufficiently high electric field.

Poling: Process of aligning the ferroelectric polarization along a unique (poling) direction.

Piezoelectric composites: Materials containing two or more components with different piezoelectric properties.

Charge amplifier: An operational amplifier used to convert the input charge into output voltage by means of the capacitor in the feedback loop.

References

1. J. F. Nye, *Physical Properties of Crystals*, Oxford: Oxford University Press, 1985.
2. Y. Xu, *Ferroelectric Materials and Their Applications*, Amsterdam: North-Holland, 1991.
3. L. E. Cross, Ferroelectric ceramics: tailoring properties for specific applications, In N. Setter and E. L. Colla (ed.), *Ferroelectric Ceramics: Tutorial Reviews, Theory, Processing, and Applications*, Basel: Birkhauser, 1993.
4. B. Jaffe, W. R. Cook, Jr., and H. Jaffe, *Piezoelectric Ceramics*, Marietta, OH: R. A. N., 1971.
5. *The User's Guide to Ultrasound & Optical Products*, Hopkinton, MA: Valpey-Fisher Corporation, 1996.
6. S.-E. Park and T. R. Shrout, Relaxor based ferroelectric single crystals with high piezoelectric performance, *Proc. of the 8th US-Japan Seminar on Dielectric and Piezoelectric Ceramics*: October 15-18, Plymouth, MA, 1997, 235.

7. *Piezoelectric Products*, Sensor Technology Limited, Collingwood, Ontario, Canada, 1991.
8. H. Kawai, The piezoelectricity of poly(vinylidene fluoride), *Japan. J. Appl. Phys.*, 8, 975, 1969.
9. L. F. Brown, Ferroelectric polymers: Current and future ultrasonic applications, *Proc. 1992 IEEE Ultrasonics Symposium*: IEEE, New York, 1992, 539.
10. T. Furukawa, Recent advances in ferroelectric polymers, *Ferroelectrics*, 104, 229, 1990.
11. L. F. Brown, J. I. Scheinbeim, and B. A. Newman, High frequency dielectric and electromechanical properties of ferroelectric nylons, *Proc.* 1994 *IEEE Ultrasonics Symposium*: IEEE, New York, 1995, 337.
12. *Properties of Raytheon Polyvinylidene Fluoride* (*PVDF*), Raytheon Research Division, Lexington, MA, 1990.
13. *Standard and Custom Piezo Film Components*, Atochem Sensors Inc., Valley Forge, PA, 1991.
14. T. R. Gururaja, Piezoelectric transducers for medical ultrasonic imaging, *Amer. Ceram. Soc. Bull.*, 73, 50, 1994.
15. IEEE Standards on Piezoelectricity, *IEEE Std.* 176, 1978.
16. W. Y. Pan and L. E. Cross, A sensitive double beam laser interferometer for studying high-frequency piezoelectric and electrostrictive strains, *Rev. Sci. Instrum.*, 60, 2701, 1989.
17. A. L. Kholkin, Ch. Wuethrich, D. V. Taylor, and N. Setter, Interferometric measurements of electric field-induced displacements in piezoelectric thin films, *Rev. Sci. Instrum.*, 67, 1935, 1996.
18. A. J. Moulson and J. M. Herbert, *Electroceramics: Materials, Properties, Applications*, London: Chapman and Hall, 1990.
19. J. M. Herbert, *Ferroelectric Transducers and Sensors*, New York: Gordon and Breach, 1982.

6.5 Laser Interferometer Displacement Sensors

Bernhard Günther Zagar

In the past few years, very high precision, numerically controlled machine tools have been developed. To achieve the potential precision of these tools, length and displacement measurements whose resolution exceeds the least significant digit of the tool must be made. The measurement equipment typically would not rely on mechanical scales.

Laser interferometers compare the changes in optical path length to the wavelength of light, which can be chosen from atomic constants that can be determined with very little uncertainty.

In 1983, there was a redefinition of the meter [1] that was previously defined in 1960. The old definition was based on the wavelength of a certain radiation (the krypton-86 standard) that could not be realized to better than 4 parts in 10^9. The new definition, being based on frequency but not related to any particular radiation, opened the way to significant improvements in the precision with which the meter can be realized. As recommended in resolution 2 for the practical realization of the meter, the wavelength in vacuum λ_v of a plane electromagnetic wave of frequency f is $\lambda_v = c/f$, where c is the speed of light in vacuum, $c = 299, 792, 458$ m s^{-1} exactly. This way, the wavelength is related to *frequency* and *time*, which can be measured with the highest precision of all units within the Système International (SI).

In order to be independent of any environmental parameters, the meter is defined using the speed of light in a vacuum. However, interferometers usually must operate in ambient air. Thus, environmental parameters that influence the speed of light in a particular medium (air) will affect and degrade the precision of the measurement.

Three major factors limit the absolute accuracy attainable with laser interferometers operating in ambient air: (1) the uncertainties of the vacuum wavelength, λ_v, of the laser source; (2) the uncertainty of the refractive index of the ambient air; and (3) the least count resolution of the interferometer.

This chapter section is organized as follows. First, some basic laser principles are detailed, including ways to stabilize the vacuum wavelength of the laser. The effect most often used to stabilize lasers in commercial interferometers is the Zeeman effect, which yields relative uncertainties of 10^{-8}.

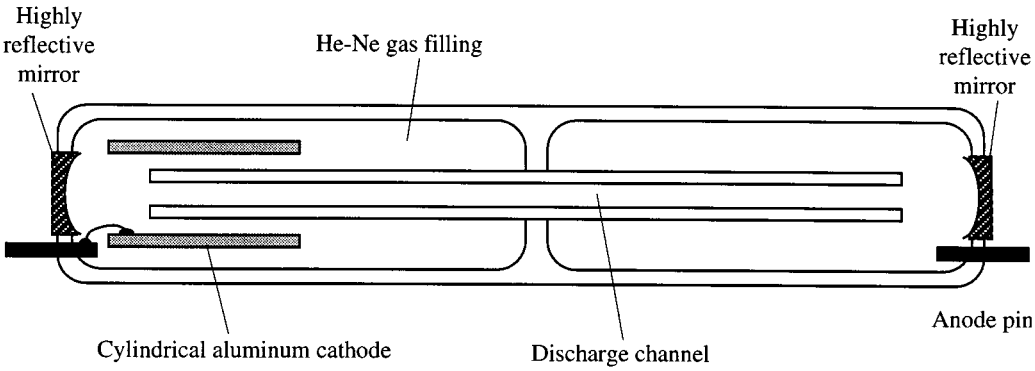


FIGURE 6.46 Schematics of the helium–neon laser (Reprinted with permission of University Science Books [3]).

Second, the refractive index of air as another major factor limiting the attainable accuracy of laser interferometers operated in air is addressed. It is shown that it cannot be determined currently with uncertainty better than 5×10^{-8}.

And finally, the chapter section describes the most widely used Michelson interferometer and two of its variants for long-travel length measurement and gives their resolution.

Helium–Neon Laser

In order to attain the best possible accuracy, great care must be taken to ensure the highest wavelength stability of the light source. Almost all interferometric dimensional gages utilize a helium–neon laser because it has proven reliable, its emitted wavelength is in the visible range at about 633 nm, and it can be stabilized sufficiently well utilizing the Zeeman effect and to an even higher degree with the use of a very well-defined iodine absorption line also at ≈633 nm [1, 2].

The helium–neon laser consists of a discharge tube as shown in Figure 6.46 [3] filled with the single-isotope gases helium (He^3) at a partial pressure of ≈105 Pa and neon (Ne^{20}) with a partial pressure of ≈13 Pa. It is pumped electrically using a voltage on the order of kilovolts with a current of a few milliamperes to excite both helium and neon atoms. Since the helium gas is the majority component, it dominates the discharge properties of the laser tube. Neutral helium atoms collide with free electrons that are accelerated by the axial voltage and become excited and remain in two rather long-lived metastable states. These are close enough to the energy levels of certain excited states of neon atoms so that collisional energy transfer can take place between these two groups of atoms. Excited helium atoms may drop down to the ground state, while simultaneously neon atoms take up almost exactly the same amount of energy. Therefore helium only serves to excite neon atoms, they do not contribute to the emission of light. The excited neon atoms remain in the excited state for a rather long period of time (on the order of 10^{-3} s). They return to lower energetic levels by stimulated emission of highly coherent light. This stimulated emission comes into effect when light emitted by some neon atoms also prompts other atoms to emit. The mirrors of the laser cavity, by reflecting most of the incident light cause the light to traverse multiple paths through the active laser volume, thereby greatly amplifying the light if the cavity length L is an integer multiple m of half the wavelength λ.

$$L = m\frac{\lambda}{2} \tag{6.70}$$

The emitted light is fairly monochromatic, but still has some finite spectral linewidth determined by the random emissions of Ne^{20}.

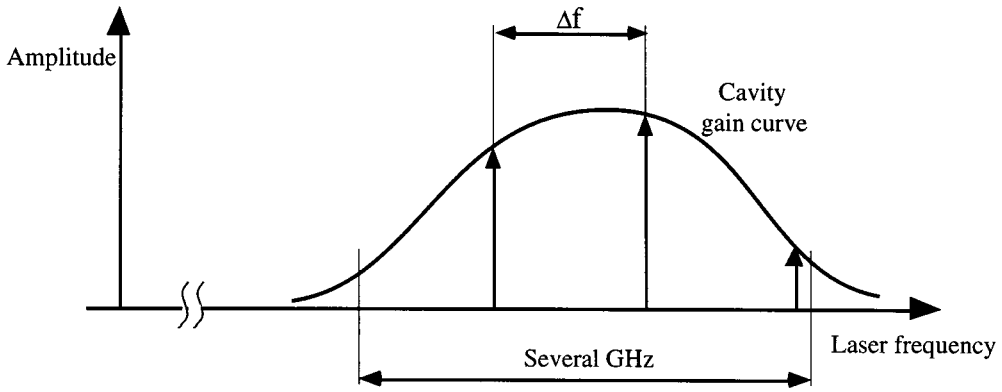


FIGURE 6.47 A He–Ne laser can have multiple resonating modes (shown for ≈20 cm cavity length) (Reprinted with permission of University Science Books [3]).

Brewster angle [4] end windows of the discharge tube transmit light of the proper linear polarization if desired. The end mirrors have to be carefully polished. Their curvature radii have to satisfy the condition for stability. They have wavelength–selective dielectric coatings of very high reflectivity sometimes exceeding 99%.

Unless special precautions are taken, a He–Ne laser will emit several axial modes as shown schematically in Figure 6.47, resulting in a beat frequency that limits the temporal coherence and renders the laser unsuitable for interferometric purposes. Also, due to thermal expansion of the laser tube, the end mirrors will change their relative distance, thereby effectively tuning the wavelength within the linewidth of the gain curve.

Another important property of a high-quality He–Ne laser is its Gaussian cross-sectional profile, which is maintained along a propagating wave, i.e., fundamental lateral mode. It is also a necessary condition for the wavefronts to remain quasiplanar.

Frequency Stabilization of He–Ne Lasers

The resonant frequency (and the wavelength λ) of the laser is determined in part by the distance between the two end mirrors and also by the refractive index n_M of the active medium (the He–Ne mixture). Since the linewidth of the gain profile of the active medium is usually in the gigahertz range, multiple axial modes can resonate in the cavity as is shown in Figure 6.47. The frequency difference Δf between two adjacent longitudinal modes is the *free spectral range* (FSR), which is given by Equation 6.71 and depends only on the cavity length, L, the refractive index of the active medium, n_M, and the speed of light in vacuum, c.

$$\Delta f = \frac{c}{2n_M L} \tag{6.71}$$

Due to thermal expansion of the laser cavity and/or thermally induced change in refractive index of the active medium, all resonating laser modes will move within the envelope of the gain profile. The effort undertaken in stabilizing the laser wavelength or equivalently stabilizing its frequency is aimed at locking the modes with respect to the gain profile and reducing the number of modes resonating simultaneously.

Longitudinal Zeeman Effect

One of the most often used effects in stabilizing the frequency of a He–Ne laser for distance measurements is the Zeeman effect [3, 5].

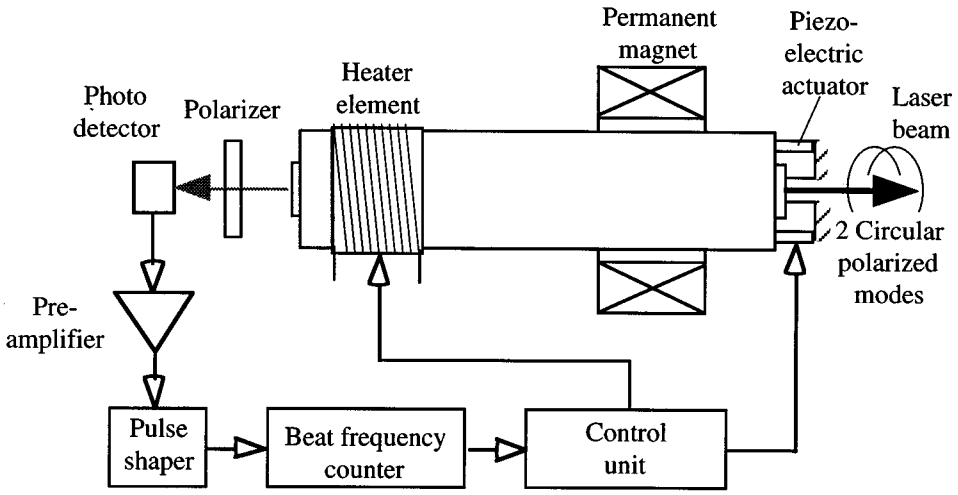


FIGURE 6.48 Laser wavelengths can be stabilized by utilizing the Zeeman-effect (Courtesy of Spindler & Hoyer Inc.).

The tube of the laser whose frequency is to be stabilized is mounted inside an axial permanent magnet as shown in Figure 6.48. Because of a short cavity length chosen, the free spectral range Δf of the laser is large. Therefore, the laser emission appears as a single longitudinal mode when the cavity length is properly controlled. An externally applied magnetic field in longitudinal direction causes the transition frequencies of the neon atoms to split symmetrically to the nominal frequency, the separation in frequency being proportional to the magnetic field strength. This phenomenon is called the *longitudinal Zeeman effect.*

With the applied field strength, the single longitudinal laser mode of the He–Ne laser is split into right and left circularly polarized modes with a frequency difference of typically 1.5 MHz. The frequency difference depends on the laser frequency and exhibits an extreme value when the laser frequency matches exactly the atomic transition frequency. Frequency stabilization is accomplished with control of the cavity length so that the frequency is locked to the extreme value corresponding to the nominal frequency of the laser. In order to determine the actual laser frequency within the gain profile, the cavity length is periodically modulated with a piezoelectric actuator shown in Figure 6.48. The frequency difference is constantly monitored at the minimum and maximum elongation of the modulation cycle. In the case when the extremal values are identical, the difference frequency will assume a maximum and thus the nominal value of the laser frequency is attained. This is achieved by controlling the length of the laser cavity. The cavity length is controlled both by using the thermal expansion of the laser tube, which is in close proximity to an electric heater whose operating current can be varied on a less dynamic scale, and on a very short time basis with a piezoelectric actuator attached to one of the end faces of the cavity [5].

This fast actuator allows for only a limited amplitude, typically on the order of less than a wavelength. Therefore, the thermal expansion must be used to attain greater modulation amplitude. To obtain the highest possible stability of the laser frequency, the frequency differences are measured within a very short time interval, and the determined deviations from the nominal value are fed back in a closed control loop to the electromechanical and thermal positioning devices.

The short- and long-term frequency stability $\Delta f/f$ of such a stabilized laser can be as high as 2×10^{-8}, depending also on the parameters of the feedback loop.

Another phenomenon that exceeds the stabilization attainable with the Zeeman effect by far uses a particular absorption line of an iodine isotope ($^{127}I_2$, transition 11-5, R(127), component *i*) and locks the He–Ne laser [2, 6] in that very well-defined frequency. Its wavelength λ_I = 632.9913981 nm is very close to one of the nominal wavelengths of the He–Ne laser, so this can easily be accomplished; however, the necessary equipment including the sophisticated electronics is rather involved [2]. For this reason, iodine-stabilized He–Ne lasers are only used where very high precision is required.

TABLE 6.9 Refractive Indices n of Various Gaseous Compounds Their Maximum Workplace Concentration (MWC) and the Change in Refractive Index Caused by that Particular Concentration (for $T = 20°C$, $P = 101.315$ hPa, $\lambda = 633$ nm)

Gas	Refractive index $(n-1)10^4$	Concentration for $\Delta n/n = 10^{-7}$ in air (ppm)	MWC mg m^{-3}	$\Delta n/n \times 10^7$ due to MWC
Air	2.72			
Propane	10.3	130	1800	8
Butane	12.9	98	2350	10
Ethanol	8.1	190	1900	5
Ethyl acetate	13.0	97	1400	4
Dimethylketone	10.2	130	2400	9
Octane	23	50	2350	10
Chlorofluorocarbons: e.g., R12	10.3	130	5000	7

Reprinted with permission of VDI Verein Deutscher Ingenieure, G. Wilkening, Kompensation der Luftbrechzahl.

The overall estimated relative uncertainty of such a stabilized laser is $\pm 10^{-9}$ (which results from an estimated relative standard deviation of 3.4×10^{-10} [1]), which makes it suitable for the practical realization of the definition of primary and secondary standards.

Refractive Index of Air

The length scale of a displacement interferometer that is operated in ambient air is given by $\lambda_A = \lambda_V / n_A$, where λ_A is the wavelength of the source in air, λ_V is the wavelength in vacuum and n_A is the refractive index of air. The vacuum wavelength λ_V is related to the speed of light by $\lambda_V = c/f_0$. The constant c, the speed of light in vacuum, was redefined in 1983 [1]. It is now defined to be exactly $c = 299{,}792{,}458$ m s^{-1}.

Thus, in order to measure distances interferometrically, it is essential to know the refractive index of air with an accuracy that is not less than the stability and the degree of certainty of the vacuum wavelength of the light source.

According to Edlén [7, 8] and Owens[9], the refractive index is a function of the atmospheric pressure P, the temperature T, the relative humidity H, or alternatively the partial pressure of the water vapor in air e_s, and the carbon dioxide concentration by volume D. The standard composition by volume of dry air is given by 78.03% N_2, 20.99% O_2, 0.933% Ar, 0.035% CO_2, and some other trace components, mainly noble gases [10]. The only component of the above list that might have a variability is CO_2, which follows a long-term increasing behavior presumably associated with the combustion of fossil fuel [11]. The current value is ≈350 ppm by volume and is increasing by ≈1.4 ppm per year. The CO_2 concentration in air can also change in an industrial environment due to CO_2 emitters and can therefore show significant local variations.

More recently, dependencies of the refractive index on nonnatural gaseous compounds like hydrocarbons in the air have been published [12] as well as corrections to Edlén's formulations [13–15]. Corrections to the refractive index due to those compounds may be necessary for very high-precision measurements in an industrial environment where chemical solvents or oils are in use. In Table 6.9, some nonnatural compounds, their maximum workplace concentrations, and their effect on the refractive index if present at the location of the interferometer are listed.

Jones [16] combined a precise determination of the density of moist air with Edlén's formulation to yield a somewhat simpler representation. For a typical iodine-stabilized He–Ne laser that has a vacuum wavelength of $\lambda_V = 632.9913$ nm, the Jones formulation is given by [17]:

$$n(P,T,H,D) = 1 + A - B \tag{6.72}$$

where

$$A = 78.603\left[1 + 0.540\left(D - 0.0003\right)\right]\frac{P}{TZ} \times 10^{-8}$$

$$B = \left(0.00042066\ f_E\ e_s\ H\right) \times 10^{-8} \tag{6.73}$$

In Equation 6.73, P is the atmospheric pressure in pascals, T is the absolute temperature in kelvin, H is the relative humidity in %, and D is the concentration of CO_2 in percent by volume. There are three additional factors in Jones' formulation that take the nonideal behavior of moist air as compared to an ideal gas into account. They are Z, a compressibility factor that reflects the nonideality of the air–water vapor mixture and which, for air containing reasonable amounts of CO_2 at a temperature between 15°C and 28°C and pressure of between 7×10^4 Pa and 11×10^4 Pa, lies in the range between 0.99949 and 0.99979. f_E is an enhancement factor that expresses the fact that the effective saturation vapor pressure of water in air is greater than the saturation vapor pressure e_s. For the pressure and temperature ranges given above, f_E is bounded between 1.0030 and 1.0046 [16]. e_s is the saturation vapor pressure over a plane surface of pure liquid water and according to Jones is about 1705 Pa at a temperature of 15.0°C and about 3779 Pa for 28.0°C. Tables of Z, f_E and e_s are included in the Appendix of Jones' paper [16].

Table 6.10 gives an overview of the changes in environmental parameters that would cause a relative index change of 10^{-7}.

Edlén [8] and Jones [16] estimate that their empirical expressions for the dependency of the refractive index of air on the listed parameters has an absolute uncertainty of 5×10^{-8}.

Besides this fundamental limitation, there are some practical considerations that must be taken into account regarding the precision with which the environmental parameters can be measured. Estler [17] states that atmospheric pressure P can currently be determined with an uncertainty of ≈2.7 Pa, which can be assumed to be constant for the entire optical path of the interferometer if it is oriented horizontally. Please note that at sea level, the pressure gradient is ≈–13 Pa m^{-1}, resulting in a pressure-induced change in n_A of 3.4×10^{-8} m^{-1} if the measuring equipment is not kept level.

In an exceptionally well-controlled laboratory environment where special care is devoted to keep temperature gradients from affecting the refractive index along the optical path as much as possible, uncertainties of the temperature measurement can be as low as 0.01°C according to [17]. Humidity measured with high accuracy dew-point hygrometers can have uncertainties down to 0.5%. Changes in carbon dioxide concentrations have to be very significant (20% of the natural concentration) to cause a $\Delta n/n$ of 10^{-8}.

Michelson Interferometer

The basis for most interferometers used in interferometric dimensional gages is the classical Michelson interferometer [4] which is shown in Figure 6.49. The coherent monochromatic light of a wavelength-stabilized He–Ne laser is incident onto a beam splitter which splits the light into two equally intense beams (**1**) and (**2**).

TABLE 6.10 Parameters of Standard Air and Their Deviation to Cause a $\Delta n/n$ of 10^{-7}

Parameter	Standard value	Variation for $\Delta n/n = +1 \times 10^{-7}$
Pressure P	101.3 kPa	+37.3 Pa
Temperature T	20.0°C	–0.1°C
Humidity H	40%	–10.0%
CO_2 concentration	350 ppm	+670 ppm

Reprinted with permission of *J. Applied Optics* [17].

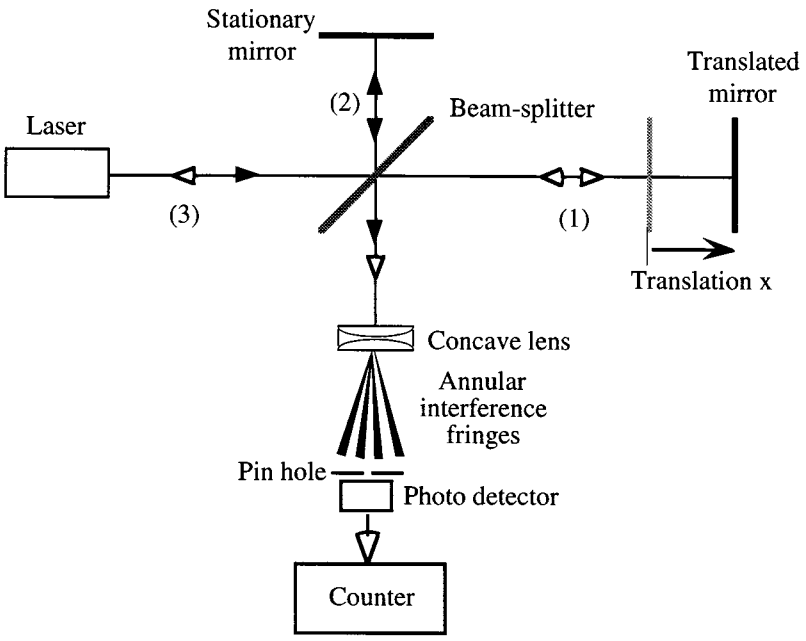


FIGURE 6.49 Schematics of the basic Michelson interferometer.

They are reflected off of both the stationary and the translatable mirror whose displacement x is to be measured, and recombined at the splitter, where they are redirected toward a concave lens. Due to the coherence of the laser light, the wavefronts have a well-defined phase relation with respect to each other. This phase is determined by the difference between the optical path lengths of the two beams in arms **1** and **2**. If this path difference is continuously changed by translating one of the mirrors, a sinusoidal intensity variation can be observed at a fixed location in space behind the lens used to introduce a beam divergence, resulting in an annular fringe pattern. The pinhole is used to define an exact location of observation and the photodetector picks up the varying intensity for further processing. In the most basic signal processing setup, the number of bright and dark cycles are fed into a counter, which then counts changes in optical path length in integer multiples of $\lambda_A/2$. More sophisticated signal processing not only counts cycles but also determines relative phase changes in the sinusoidal varying intensity so that resolutions of $\lambda_A/512$ can ultimately be achieved.

When moving the mirror, one must guarantee a smooth motion without backward jitter of the mirror carriage to avoid double counts of interference fringes. Very high-quality linear bearings (such as air bearings) are necessary to accomplish just that.

As can be seen in Figure 6.49, the light reflected off both mirrors essentially retraces its own path and is at least partially incident onto the active volume of the laser source (**3**), thereby forming an external laser resonator which is able to detune the laser, effectively modulating its output power as well as its wavelength. To avoid this effect, commercial versions of Michelson interferometers employ corner-cube reflectors instead of plane mirrors as well as optical isolators, as shown in Figures 6.50 and 6.51. Some authors, however, report using optical arrangements that utilize this effect in conjunction with laser diodes to realize low-cost, short-travel displacement sensors [18, 19]. These setups will not be discussed here, however.

Two-Frequency Heterodyne Interferometer

Figure 6.50 shows the commercially available two-frequency Michelson interferometer operating with a Zeeman-stabilized He–Ne laser source. This laser emits two longitudinal modes with frequencies f_1 and f_2 that are both circularly polarized in opposite directions. By passing the modes through a quarter-wave

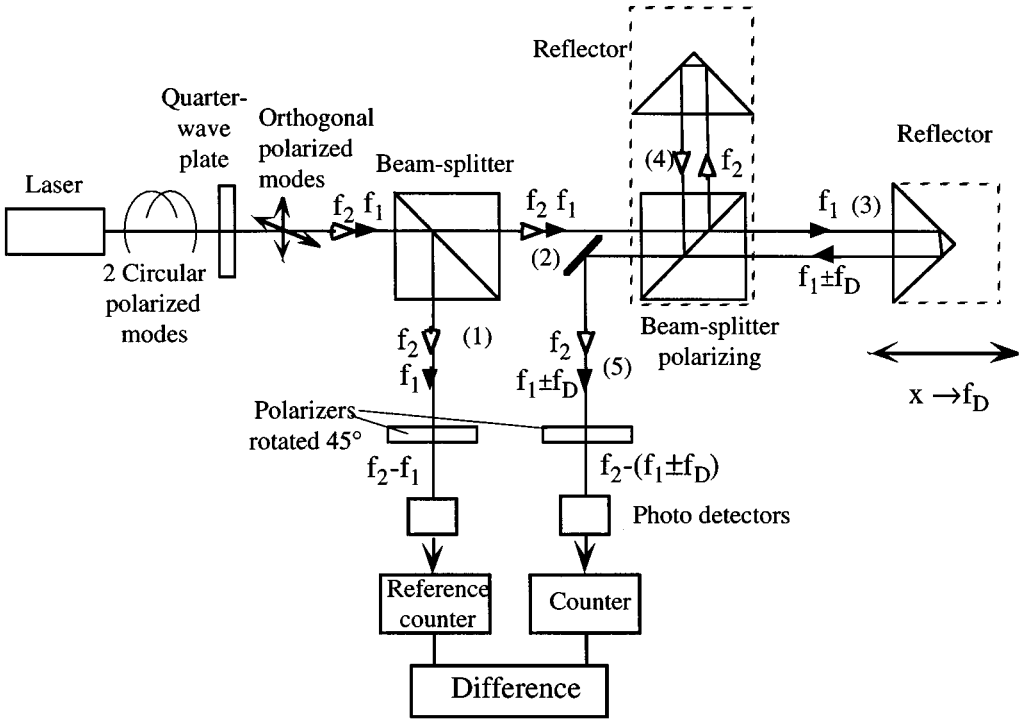


FIGURE 6.50 Two-frequency heterodyne interferometer (Courtesy of Spindler & Hoyer Inc.).

plate, two orthogonal linearly polarized waves are generated. Both are split by a nonpolarizing beam splitter. There is a polarizer located in arm **1** of that splitter, which is rotated by 45° with respect to both polarized waves impingng on it, thus effectively allowing them to interfere behind it yielding a difference frequency of $f_2 - f_1$ that is picked up by a photodetector and counted by a reference counter (frequency difference typically 1.5 MHz).

The orthogonal polarized waves in **2** are further split by a polarizing splitter. Spectral component $f_1 < f_2$ is transmitted into measuring arm **3** and frequency component f_2 is reflected into reference arm **4** of the interferometer. Due to the velocity v of the reflector in arm **3** resulting in a displacement x, the frequency f_1 is Doppler-shifted by f_D (Equation 6.74). Movement of the reflector toward the interferometer results in a positive Doppler frequency $f_D > 0$. After recombining both waves from **3** and **4** in the beam splitter again, they are sent through a polarizer in arm **5** that also is rotated by 45° with respect to the direction of polarization of both recombined waves, thereby allowing them to interfere, yielding a difference frequency of $f_2 - f_1 - f_D$ at the location of the photodetector, which is counted by a second counter. By continuously forming the difference of both counts, the measurand (the displacement of x in multiples of $\lambda_A/2$) is calculated.

With this type of interferometer, the direction of motion is given by the sign of the resulting count. One disadvantage of the two-mode heterodyne interferometer is its limited dynamic range for the velocity v of the reflector moving toward the interferometer, since the Doppler frequency f_D, given by:

$$f_D = \frac{2}{\lambda_A} v \tag{6.74}$$

is bound to be less than the initial frequency difference between f_2 and f_1 for stationary reflectors. Given a typical Zeeman effect-induced frequency difference f_z of 1.5 MHz, the velocity v is therefore bound to be less than:

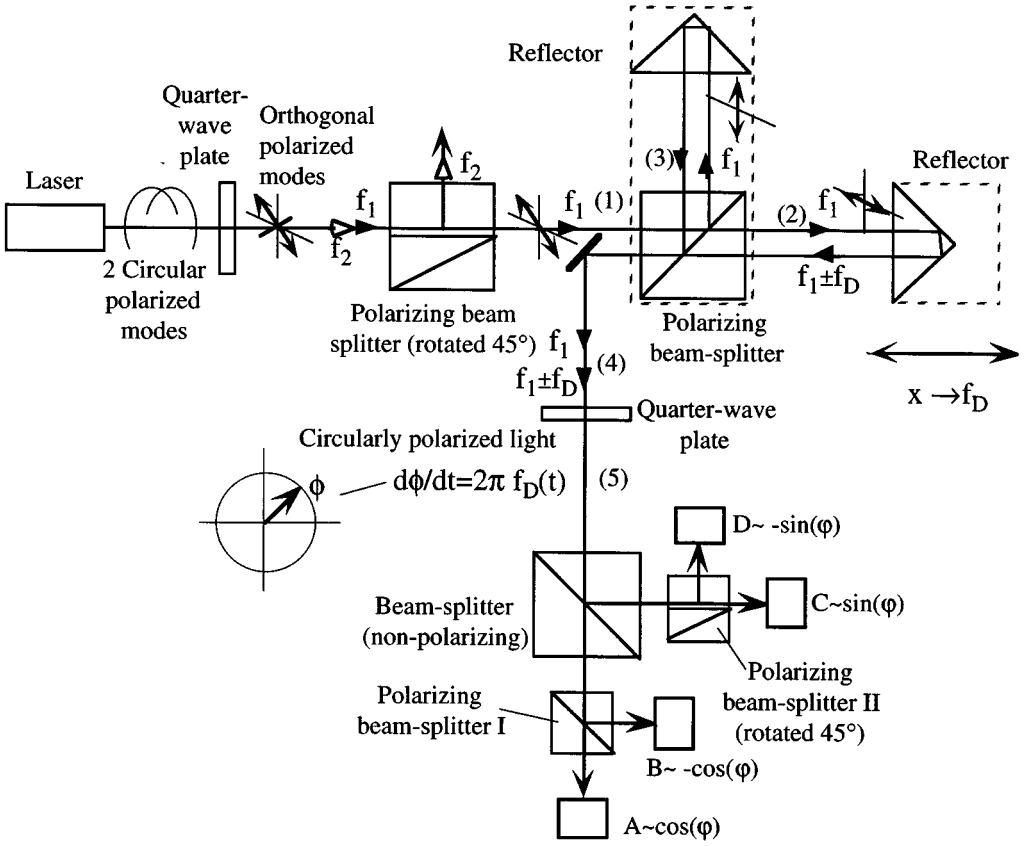


FIGURE 6.51 Single-mode homodyne interferometer (Courtesy of Spindler & Hoyer Inc.).

$$v < \frac{f_z}{2} \lambda_A = 0.474 \text{ m s}^{-1} \tag{6.75}$$

There is no such bound if the reflector is traveling away from the interferometer. By electronically interpolating the output signal of the photodetector, subwavelength resolution can be obtained [22].

Single-Mode Homodyne Interferometer

An interferometer setup that has no limitation on the maximum velocity in the above sense is the single-mode homodyne interferometer shown in Figure 6.51.

As in the two-frequency heterodyne interferometer, a Zeeman effect-stabilized laser source that emits two frequency-displaced circularly polarized axial modes is usually used. After passing through a quarter-wave plate, two orthogonal polarized waves are generated, only one of which (f_1) is further used. The other (f_2) is reflected out of the optical path by an appropriately oriented polarizing beam splitter. The plane of polarization in arm **1** of the interferometer is tilted by 45° with respect to the plane defined by the two arms **2** and **3**. The second polarizing beam splitter will transmit one horizontally oriented component into the measuring arm **2** and reflect a vertically oriented component into the reference arm **3** of the interferometer. The two arms of the interferometer maintain their orthogonal polarizations. The frequency of the wavefront in arm **2** is shifted by the Doppler effect (Equation 6.74) due to the motion of the reflector. The light reflected by the two triple mirrors is recombined in the polarizing beam splitter and redirected by a mirror. Since in this particular optical setup, the polarization states of the two beams in the two arms of the interferometer are orthogonal, there is no interference after the redirecting mirror

in arm **4** as was the case with the basic Michelson interferometer setup (Figure 6.49). After passing a quarter-wave plate at 45°, two opposite circular polarized waves (one with frequency f_1, the other with frequency $f_1 \pm f_D$) are generated and can be described by a rotating phasor (characterized by $\Phi(t)$) with constant amplitude whose rate of rotation is dependent on the Doppler frequency (in arm **5**). Amplitude fluctuations can be observed at the photodetectors A–D after this phasor has passed polarizers, which it does after being split by a nonpolarizing beam splitter.

The output of an interferometer has the general form:

$$I(t) = I_o(t)\frac{1}{2}\left[1+\cos\left(\Phi(t)\right)\right] \tag{6.76}$$

It is desired to infer $\Phi(t)$ from observation of $I(t)$. Note that $I_o(t)$, which is the intensity of the laser, can also fluctuate with time. The problems encountered with this are (1) the ambiguity in the sign of $\Phi(t)$ and (2) the dependence of the calculated phase on the intensity fluctuations due to the aging of the laser and optical components. The first problem stems from the fact that arccos(…) yields two solutions to Equation 6.76:

$$\Phi(t) = \pm\arccos\left[\frac{2I(t)}{I_o(t)} - 1\right] \tag{6.77}$$

The sign ambiguity can be resolved by also generating a $\sin(\Phi(t))$ yielding quadrature signals. In order to do so, a second output of the interferometer of the form:

$$I_2(t) = I_o(t)\frac{1}{2}\left[1+\sin\left(\Phi(t)\right)\right] \tag{6.78}$$

is sought. Equations 6.76 and 6.78 will determine $\Phi(t)$ unambiguously only in the region $[0,2\pi)$, but there is still an ambiguity modulo 2π. The second problem can be dealt with by adding two more outputs of the form:

$$I_3(t) = I_o(t)\frac{1}{2}\left[1-\cos\left(\Phi(t)\right)\right] \tag{6.79}$$

$$I_4(t) = I_o(t)\frac{1}{2}\left[1-\sin\left(\Phi(t)\right)\right] \tag{6.80}$$

Taking Equations 6.76 to 6.79 and 6.78 to 6.80, it is possible to obtain a zero crossing at the linear most sensitive point of inflection of the fringe where the effect of intensity fluctuations on the phase measurement is minimal. The setup of Figure 6.51 attempts to obtain these four outputs. Signal A represents the intensity variations as given in Equation 6.76 and signal B due to the nature of the splitting action is shifted with respect to A by 180° (Equation 6.79). There is another arm to the right of the nonpolarizing beam splitter incorporating the polarizing beam splitter II, which is rotated by 45° with respect to beam splitter I so that the attached detectors C and D are generating the signals defined by Equations 6.78 and 6.80.

Since the Doppler frequency is time dependent according to the velocity v of the measuring reflector, the distance, x, traveled by the reflector up to time T is given by Equation 6.81.

$$x = \int_0^T v(t)\,dt = \int_0^T f_D(t)\frac{\lambda_A}{2}\,dt = \int_0^T \frac{\partial\Phi(t)}{\partial t}\frac{\lambda_A}{4\pi}\,dt \tag{6.81}$$

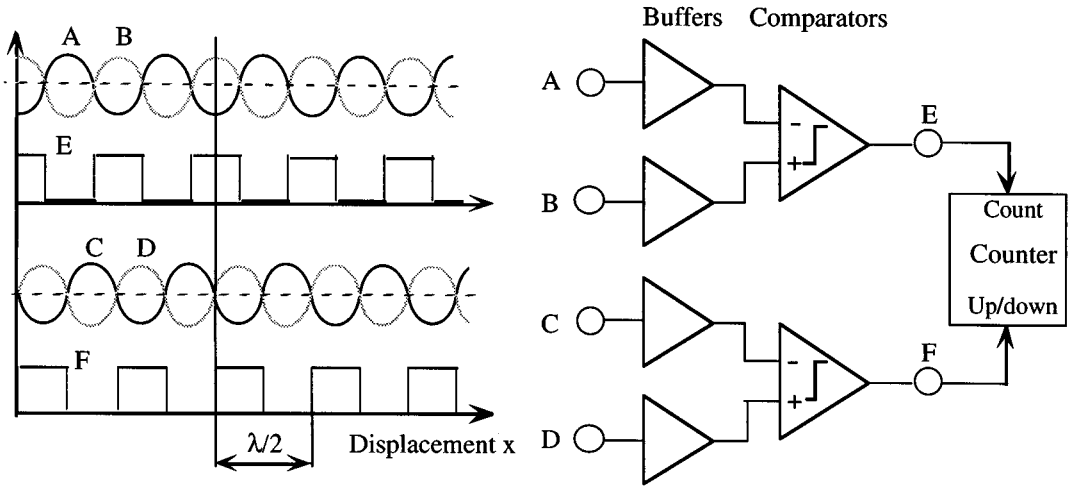


FIGURE 6.52 A simple signal conditioning circuit yielding quadrature signals E and F.

It should be noted that the resolution of this interferometer is limited to $\lambda_A/4$ if no special hardware is used to interpolate between interference fringes.

This particular interferometer is not limited with respect to a maximum unambiguous Doppler frequency.

Interferometer Signal Processing

At the photodetectors ends the optical path of the interferometers. Since the photodetectors are sensitive to light intensity values only the sinusoidal signal caused by motion of the reflector will be superimposed onto a pedestal signal proportional to the mean intensity over time or displacement, respectively. The electronics attached to the photodetectors is aimed to reliably detect zero-crossings of the sinusoidal component of the signals (A through D) even under low contrast conditions by subtracting out the pedestal signal component in the comparators. Low fringe contrast can be the result of small reflector tilt and/or vibration during periods of reflector movement, nonideal interferometer alignment, and imperfections in optical components like unequal splitting of the beam in beam splitters. Figure 6.52 shows a simple conditioning unit.

The main purpose of that set up is to produce digital quadrature signals (**E** and **F** in Figure 6.52) that can most easily be used to perform signal interpolation digitally [20, 21] and also allow for the determination of the direction of movement by using, for example, signal **E** as count signal and **F** as an up/down indicator.

Both comparators compare sinusoidal components having a possibly slowly time-varying dc pedestal amplified by buffer amplifiers. All photodetector signals A through D will be affected by this dc pedestal so that forming their difference in the comparator will yield zero-crossing signals independent of the pedestal. The two comparator units are necessary to generate quadrature signals for direction detection. Furthermore, this electronics allows easily for a fringe interpolation by a factor of 4, bringing the least-count resolution of the interferometer down to ≈160 nm. This interpolation can be done if instead of feeding the signal E into the count input, the exored signal E⊕F is used. In this case, the up/down terminal of the counter circuit needs to be connected to the signal E&F. If an even higher resolution is sought, digital interpolation of these quadrature signals can be performed [20–22].

Conclusions

Using laser interferometers operating with highly stabilized laser sources, relative uncertainties in length measurements as low as 5×10^{-8} can be realized, which makes these kinds of equipment suitable to convey

primary and secondary length standards into industrial measurement labs. Since, for practical purposes, these interferometers need to be operated in ambient air, the uncertainty in the refractive index of air gives the largest single factor limiting the overall precision. Decreasing the uncertainty stemming from that source necessitates the determination of environmental parameters such as relative humidity, temperature along the optical path, and atmospheric pressure with very little error. Laser refractometers able to determine the refractive index of air directly, on the other hand, allow for the direct measurement and compensation of the refractive index. Zeeman effect-stabilized laser sources can reach a relative uncertainty of less than 10^{-8} and thus contribute only a small portion of the overall error. The interferometer itself, depending on the signal processing involved, will have at least count resolution of $\lambda_A/4$ without any optical or electronic interpolation and can have an order of magnitude less uncertainty if high-performance phase meters are used to subdivide the wavelength λ_A. Very high-quality air bearings and a very well-controlled measurement environment are necessary to reach these goals.

Defining Terms

Heterodyne technique: The superposition of two harmonic signals with frequencies f_1 and f_2 in a nonlinear device results in a signal containing both sum and differences frequencies. For interferometric purposes where each individual frequency is in the 10^{14} Hz range, the frequency difference might be low enough to be registered by photodetectors. They also serve as nonlinear devices because their response is to light intensity only, which is proportional to the light amplitude squared.

Interference: A phenomenon that strikingly illustrates the wave nature of light. It occurs when radiation follows more than one path from its source to the point of detection. It may be described as the local departures of the resultant intensity from the law of addition as the point of detection is moved, for the intensity oscillates about the sum of the separate intensities.

Laser: The acronym of *L*ight *A*mplification by *S*timulated *E*mission of *R*adiation. It was originally used to describe light amplification, but has more recently come to mean an optical oscillator.

Refractive index: A number specifying the ratio of the propagation velocities of light in vacuum c to that in a medium. The refractive index for any medium is always somewhat larger than 1. Its value is dependent on the composition of the medium and the wavelength of the incident radiation.

Zeeman effect: An effect that is observed if excited atoms emit their radiation in the presence of a magnetic field. The longitudinal Zeeman effect causes a single emission line to split symmetrically into right and left circularly polarized lines.

References

1. Documents concerning the new definition of the metre, *Metrologia,* 19, 163-177, 1984.
2. E. Jaatinen and N. Brown, A simple external iodine stabilizer applied to 633 nm 612 nm and 543 nm He–Ne lasers, *Metrologia,* 32, 95-101, 1995.
3. A. E. Siegman, *Lasers,* Mill Valley, CA: University Science Books, 1986.
4. W. R. Steel, *Interferometry,* 2nd ed., (Cambridge studies in modern optics), Cambridge, U.K.: Cambridge University Press, 1985.
5. W. R. C. Rowley, The performance of a longitudinal Zeeman-stabilized He–Ne laser (633 nm) with thermal modulation and control, *Meas. Sci. Technol.,* 1, 348-351, 1990.
6. Wolfgang Demtröder, *Laser Spectroscopy: Basic Concepts and Instrumentation,* 2nd ed., Berlin: Springer, 1996.
7. B. Edlén, The dispersion of standard air, *J. Opt. Soc. Amer.,* 43(5), 339-344, 1953.
8. B. Edlén, The refractive index of air, *Metrologia,* 2(2), 71-80, 1966.
9. J. C. Owens, Optical refractive index of air: dependence on pressure, temperature and composition, *Appl. Optics,* 6(1), 51-59, 1967.
10. H. D. Baehr, *Thermodynamik,* 6th ed., Berlin: Springer, 1988.

11. R. Revelle, Carbon dioxide and world climate, *Sci. Amer.,* 247(2), 35, 1982.
12. K. P. Birch, F. Reinboth, R. W. Ward, and G. Wilkening, Evaluation of the effect of variations in the refractive index of air upon the uncertainty of industrial length measurement, *Metrologia,* 30(1), 7-14, 1993.
13. K. P. Birch and M. J. Downs, An updated Edlén equation for the refractive index of air, *Metrologia,* 30, 155-162, 1993.
14. K. P. Birch and M. J. Downs, Corrections to the updated Edlén equation for the refractive index of air, *Metrologia,* 31, 315-316, 1994.
15. P. E. Ciddor, Refractive index of air: new equations for the visible and near infrared, *Appl. Optics,* 35(9), 1566-1573, 1996.
16. F. E. Jones, The refractivity of air, *J. National Bureau of Standards,* 86(1), 27-32, 1981.
17. W. T. Estler, High-accuracy displacement interferometry in air, *J. Appl. Optics,* 24(6), 808-815, 1985.
18. J. A. Smith, U. W. Rathe, and C. P. Burger, Lasers with optical feedback as displacement sensors, *Opt. Eng.,* 34(9), 2802-2810, 1995.
19. N. Takahashi, S. Kakuma, and R. Ohaba, Active heterodyne interferometric displacement measurement using optical feedback effects of laser diodes, *Opt. Eng.,* 35, 802-907, 1996.
20. K. Oka, M. Tsukada, and Y. Ohtsuka, Real-time phase demodulator for optical heterodyne detection processes, *Meas. Sci. Technol.,* 2, 106-110, 1991.
21. J. Waller, X. H. Shi, N. C. Altoveros, J. Howard, B. D. Blackwell, and G. B. Warr, Digital interface for quadrature demodulation of interferometer signals, *Rev. Sci. Instrum.,* 66, 1171-1174, 1995.
22. J. A. Smith and C. P. Burger, Digital phase demodulation in heterodyne sensors, *Opt. Eng.,* 34, 2793-2801, 1995.

Appendix to Section 6.5

In the appendix below some companies are listed that manufacture either complete interferometer systems or major components thereof, such as beam splitters, retroreflectors, refractometers, wavemeters, etc. This list is by no means exhaustive. Furthermore, no price information is included because the system cost is too much dependent on the particular choice of system components.

Companies that produce interferometers or significant components.

Manufacturer	Sub-systems	Complete systems	Manufacturer	Sub-systems	Complete systems
Aerotech Inc. 101 Zeta Drive Pittsburgh, PA 15238 Tel: (412) 963-7470	*	*	Oriel Instruments Inc. 250 Long Beach Blvd. Stratford, CT 06497-0872 Tel: (203) 380-4364	—	*
Burleigh Inc. Burleigh Park Fishers, NY 14453-0755 Tel: (716) 924-9355	—	*	Polytec PI Inc. Auburn, MA 01501 Tel: (508) 832-3456	*	*
Hewlett-Packard Inc. Test & Measurement Customer Business Center, P.O. Box 4026 Englewood, CO 80155-4026 Tel: (800) 829-4444	*	—	Spindler & Hoyer Inc. 459 Fortune Blvd. Milford, MA 01757-1745 Tel: (508) 478-6200	*	*
Melles Griot Inc. 4665 Nautilus Court South Boulder, CO 80301 Tel: (303) 581-0337	*	*	Zygo Corporation Middlefield, CT 06455-0448 Tel: (860) 347-8506	*	*

6.6 Bore Gaging Displacement Sensors

Viktor P. Astakhov

Dimensions are a part of the total specification assigned to parts designed by engineering. However, the engineer in industry is constantly faced with the fact that no two objects in the material world can ever be made exactly the same. The small variations that occur in repetitive production must be considered in the design. To inform the workman how much variation from exact size is permissible, the designer uses a tolerance or limit dimension technique. A *tolerance* is defined as the total permissible variation of size, or the difference between the limits of size. *Limit dimensions* are the maximum and minimum permissible dimensions. Proper tolerancing practice ensures that the finished product functions in its intended manner and operates for its expected life.

Bore Tolerancing

All bore dimensions applied to the drawing, except those specifically labeled as basic, gage, reference, maximum, or minimum, will have an exact tolerance, either applied directly to the dimension or indicated by means of general tolerance notes. For any directly tolerated decimal dimension, the tolerance has the same number of decimal places as the decimal portion of the dimension.

Engineering tolerances may broadly be divided into three groups: (1) *size tolerances* assigned to dimensions such as length, diameter, and angle; (2) *geometric tolerances* used to control a hole shape in the longitudinal and transverse directions; and (3) *positional tolerances* used to control the relative position of mating features. Interested readers may refer to [1, 2].

The ISO system of limits and fits (ISO Recommendation R 286) covers standard tolerances and deviations for sizes up to 3150 mm. The system is based on a series of tolerances graded to suit all classes of work from the finest to the most coarse, along with different types of fits that range from coarse clearance to heavy interference. Here, *fit* is the general term used to signify the range of tightness that may result from the application of a specific combination of tolerances in the design of mating parts.

There are 18 tolerance grades intended to meet the requirements of different classes of parts. These tolerance grades are referred to as ITs and range from IT 01, IT 02 (reserved for the future), and IT 1, to IT 16 (for today's use). In each grade, the tolerance values increase with size according to a formula that relates the value of a given constant to the mean diameter of a particular size range. The system provides 27 different fundamental deviations for sizes up to and including 500 mm, and 14 for larger sizes to give different type of fits ranging from coarse clearance to heavy interference. Interested readers may refer to [3].

Bore Gage Classification and Specification

To measure the above-listed tolerances, modern manufacturing requires the use of gages. A *gage* is defined as a device for investigating the dimensional fitness of a part for specific function. *Gaging* is defined by ANSI as a process of measuring manufactured materials to assure the specified uniformity of size and contour required by industries. Gaging thereby assures the proper functioning and interchangeability of parts; that is, one part will fit in the same place as any similar part and perform the same function, whether the part is for the original assembly or replacement in service.

Bore gages may by classified as follows:

1. Master gages
2. Inspection gages
3. Manufacturer's gages
4. Gages that control dimensions
5. Gages that control various parameters of bore geometry
6. Fixed limit working gages

7. Variable indicating gages
8. Post-process gages
9. In-process gages

Master gages are made to their basic dimensions as accurately as possible and are used for reference, such as for checking or setting inspection of manufacturer's gages. *Inspection gages* are used by inspectors to check the manufactured products. *Manufacturer's gages* are used for inspection of parts during production.

Post-process gages are used for inspecting parts after being manufactured. Basically, this kind of gage accomplishes two things: (1) it controls the dimensions of a product within the prescribed limitations, and (2) it segregates or rejects products that are outside these limits. Post-process gaging with feedback is a technique to improve part accuracy by using the results of part inspection to compensate for repeatable errors in the machine tool path. The process is normally applied to CNC (computer numerically controlled) machines using inspection data to modify the part program, and on tracer machines using the same data to modify the part template.

In-process gages are used for inspecting parts during the machining cycle. In today's manufacturing strategy, in-process gages and data-collection software provide faster feedback on quality. Indeed, the data-collection and distribution aspect of 100% inspection has become as important as the gaging technology itself. Software specifically designed to capture information from multiple gages, measure dozens of products types and sizes, and make it available to both roving inspectors and supervising quality personnel as needed, is quickly becoming part of quality control strategies. In conjunction with computer numerically controlled (CNC) units, in-process gaging can automatically compensate for workpiece misalignment, tool length variations, and errors due to tool wear.

Gages That Control Dimensions

Gages that control dimensions are used to control bore diameter. These gages can be either post-process or in-process gages. Further, these gages can be either *fixed limit gages* or *variable indicating gages.*

A *plug gage* is a fixed limit working bore gage. These inexpensive gages do not actually measure dimensions or geometry. They simply tell the operator whether the bore is oversized or undersized. The actual design of most plug gages is standard, being covered by American Gage Design (AGD) standards. However, there are many cases where a special plug gage must be designed.

A plug gage is usually made up of two members. One member is called the go end, and the other the no-go or not-go end. The gage commonly has two parts: the gaging member, and a handle with the sign, go or no-go, and the gagemaker's tolerance marked on it. There are generally three types of AGD standard plug gages. First is the single-end plug gage (Figure 6.53(*a*)); the second is the double-end (Figure 6.53(*b*)); and the third is the progressive gage (Figure 6.53(*c*)). Interested readers may refer to [4].

Fixed-limit gage tolerance is generally determined from the amount of workpiece tolerance. A 10% rule is generally used for determining the amount of gage tolerance for fixed, limit-type gages. Four classes of gagemakers' tolerances have been established by the American Gage Design Committee and are in general use [4]. These four classes establish maximum variation for any designed gage size. The degree of accuracy needed determines the class of gage to be used. Table 6.11 shows these four classes of gagemakers' tolerances. Referring to Table 6.11, class XX gages are used primarily as master gages and for final close tolerance inspection. Class X gages are used for some types of master gage work and as close tolerance inspection and working gages. Class Y gages are used as inspection and working gages. Class Z are used as working gages where part tolerances are large. Table 6.12 shows the diameter ranges and prices of the plug gages manufactured by the Flexbar Machine Corp.

Variable indicating gages allow the user to inspect some bore parameters and get numbers for charting and statistical process control (commonly abbreviated as SPC). These gages have one primary advantage over fixed gages: they show how much a hole is oversized or undersized. When using a variable indicating gage, a master ring gage to the nominal dimension to be checked must be used to preset the gage to zero. Then, in applying the gage, the variation from zero is read from the dial scale. Figure 6.54 shows industry's

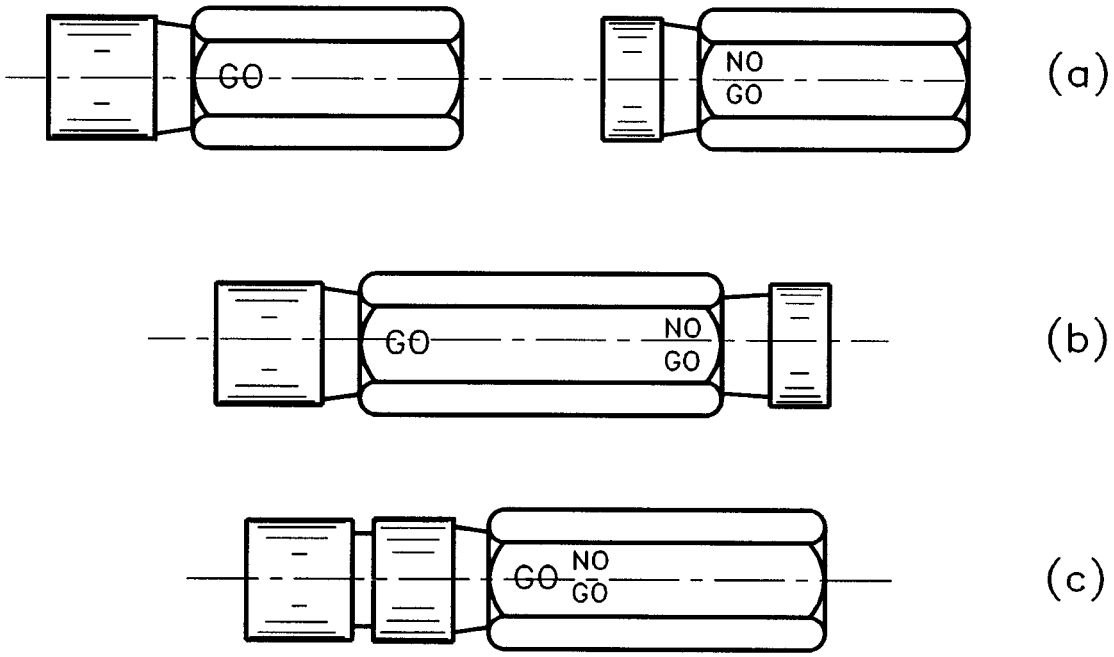


FIGURE 6.53 AGD cylindrical plug gages to inspect the diameter of holes: (*a*) two separate gage members; (*b*) two gage members mounted on single handle with one gage member on each end; (*c*) progressive gage.

TABLE 6.11 Standard Gagemakers' Tolerances

Above	To and including	Class XX	Class X	Class Y	Class Z
0.010 in.	0.825 in.	0.00002 in.	0.00004 in.	0.00007 in.	0.00010 in.
0.254 mm	20.95 mm	0.00051 mm	0.00102 mm	0.00178 mm	0.00254 mm
0.825 in.	1.510 in.	0.00003 in.	0.00006 in.	0.00009 in.	0.00012 in.
20.95 mm	38.35 mm	0.00076 mm	0.00152 mm	0.00229 mm	0.00305 mm
1.510 in.	2.510 in.	0.00004 in.	0.00008 in.	0.00012 in.	0.00016 in.
38.35 mm	63.75 mm	0.00102 mm	0.00203 mm	0.00305 mm	0.00406 mm
2.510 in.	4.510 in.	0.00005 in.	0.00010 in.	0.00015 in.	0.00020 in.
63.75 mm	114.55 mm	0.00127 mm	0.00254 mm	0.00381 mm	0.00508 mm

most popular dial and electronic bore gages. Figure 6.55 shows a set of dial bore gages for the range of 35 mm to 150 mm. Figure 6.56 shows the Intrimic® plus (Brown & Sharpe) internal micrometer. Intrimic® plus provides simple, accurate inside measurement capability. The micrometer features automatic shut-off, electronic memory mode, instantaneous inch/metric conversion, and a standard direct output for SPC applications. Tolerance classification at a glance allows quick, efficient sorting while inspecting. The display shows the inspector, in color, if any dimension is within tolerance (green), out-of-tolerance (red), or reworkable (yellow). Table 6.13 presents the basic ranges and prices for these micrometers.

Gages That Control Geometry

Gages that control various parameters of bore geometry are used for complex comparisons of part shape to an ideal shape. All these gages are post-process gages. Two major categories of geometry gages are in use: gages with manual probe head systems and form measuring machines.

TABLE 6.12 Premium Quality Hardened Steel GO/NO GO Plug Gages by the Flexbar Machine Corp.

		Price ($)			Handle	
Size range	Class	1	2–4	5–10	No	Price ($)
0.01 in. to 0.030 in.	XX	35.45	28.65	22.26	1W	8.00
0.25 mm to 0.762 mm	X	31.15	23.35	17.00		
	Y	28.35	22.15	15.05		
	Z	25.65	20.70	12.55		
0.03 in. to 0.075 in.	XX	19.25	15.30	13.63	1W	8.00
0.762 mm to 1.91 mm	X	15.05	11.70	10.60		
	Y	13.90	10.80	9.40		
	Z	11.40	9.35	8.15		
0.075 in. to 1.80 in.	XX	21.15	17.00	14.80	2W	8.30
1.91 mm to 4.57 mm	X	18.15	14.45	12.85		
	Y	17.00	13.90	11.25		
	Z	14.00	11.10	9.65		
0.180 in. to 0.281 in.	XX	22.00	17.80	15.30	3W	9.00
4.57 mm to 7.14 mm	X	18.90	15.30	13.10		
	Y	17.85	13.55	11.95		
	Z	14.30	11.40	9.90		
0.281 in. to 0.406 in.	XX	24.20	19.50	17.80	4W	9.35
7.17 mm to 10.31 mm	X	21.15	17.00	14.80		
	Y	19.30	15.65	13.80		
	Z	14.90	12.00	10.50		
0.406 in. to 0.510 in.	XX	25.35	20.35	17.80	5W	9.70
10.31 mm to 12.95 mm	X	22.25	17.80	15.80		
	Y	20.45	16.50	14.55		
	Z	16.30	13.15	11.30		
0.510 in. to 0.635 in.	XX	26.70	21.45	18.50	6W	11.40
12.95 mm to 16.13 mm	X	23.70	19.25	16.30		
	Y	21.85	17.60	15.50		
	Z	17.50	14.30	12.55		
0.635 in. to 0.760 in.	XX	28.15	22.60	19.85	7W	13.50
16.13 mm to 19.30 mm	X	25.35	20.05	17.50		
	Y	23.25	18.45	16.40		
	Z	18.65	15.15	13.45		
0.780 in. to 1.010 in.	XX	44.25	34.25	32.80	8W	20.00
19.30 mm to 25.65 mm	X	39.56	29.80	27.30		
	Y	36.30	27.25	24.90		
	Z	32.40	25.65	24.20		

Gages with Manual Probe Head Systems

Geometry gages with manual probe head systems are rapidly becoming common in many high-precision metalworking applications. The simplest form of manual probe head systems in common use is the air plug gage (spindle) (Figure 6.57). Compressed air from the air gage indicating unit is pressed to the plug gage and allowed to escape from two or more jets in the periphery. When the air plug gage is inserted into a hole, the air escaping from the jets is limited by the clearance between the jet faces and the hole. The small changes in clearance, arising when the air plug gage is inserted in successive holes of different sizes, produce changes in the flow rate or back pressure in the circuit. The magnification and datum setting of systems with variable control orifices and zero bleeds is carried out with master holes. Some errors of form that can be detected with air plug gages are: (1) taper, (2) bell mouthing, (3) barreling, (4) ovality, and (5) lobing. Dearborn (Dearborne Gage Company, MI) open-orifice air spindles are available as standard to use in measuring thru, blind, and counterbored holes ranging in diameter from 0.070 in. (2 mm) to 6.000 in. (154 mm).

FIGURE 6.54 Industry's most popular dial and electronic bore gages (Courtesy of The L.S. Starrett Co.).

Another type of geometry gage with manual probe head system is the electronic bore gage. These gages measure bores at various depths to determine conditions such as bellmouth, taper, convexity, or concavity. They are also able to determine out-of-roundness conditions when equipped with a 3-point measuring system. Figure 6.58 shows a TRIOMATIC® electronic bore gage (Brown & Sharpe), and

FIGURE 6.54 (continued)

Table 6.14 presents the basic ranges and prices for these gages. TRIOMATIC® electronic bore gages feature automatic shut-off, electronic memory mode, instantaneous inch/metric conversion, and a standard direct output to a data handling system for SPC and statistical quality control. Mechanically, TRIOMATIC® electronic bore gages use the time-tested, three contact points interchangeable heads. The contact points are spaced 120° apart, ensuring proper centering and alignment that are especially essential for deep holes. The tips of the points are made of tungsten-carbide to resist wear, and extend to the surface of the heard for measuring at the bottom of blind holes or the surface of steps within a hole. Since the tips are connected to a cone-actuated electronic measuring system, these gages are referred to as *electronic gages.*

Form Measuring Instruments

Most modern form measuring instruments stage the workpiece on a turntable and provide a means to position a gage head against the part (Figure 6.59). As the turntable rotates, the gage head measures deviation from the true circle. Those gages where the gage head is supported by a simple, rigid, manual, or motorized stand that does not provide precise control over positioning are capable of performing the following measurements: roundness, concentricity, circular runout, circular flatness, perpendicularity, plane runout, top and bottom face runout, circular parallelism, and coaxiality. Modern fully automatic machines are the most sophisticated measuring instruments. Normally, they are equipped with a Windows™-based, PC-compatible graphical user interface to perform real-time data acquisition and processing. Mitutoyo/MTI Corp. produces a wide range of these machines (Table 6.15). The RA-600 series (Figure 6.60) features an innovative, fully automatic method that enables the machine to perform centering and leveling automatically if any deviation is detected during preliminary measurement. In addition, these machines can measure thickness, squareness, cylindricity, spiral cylindricity, straightness, total

FIGURE 6.54 (continued)

runout, vertical straightness, and vertical parallelism. The machines are supplied with MeasurLink® data acquisition software for Windows™, which allows immediate measurement data analysis and feedback for variable, attribute, and short inspection runs. The software gives the quality control/production manager the ability to create traceability lists of unlimited size. Information such as machine center, operator, materials used, assignable causes, and other relevant data can be stored and attached to measurement values. See Table 6.16 for a lsit of companies that make bore gages.

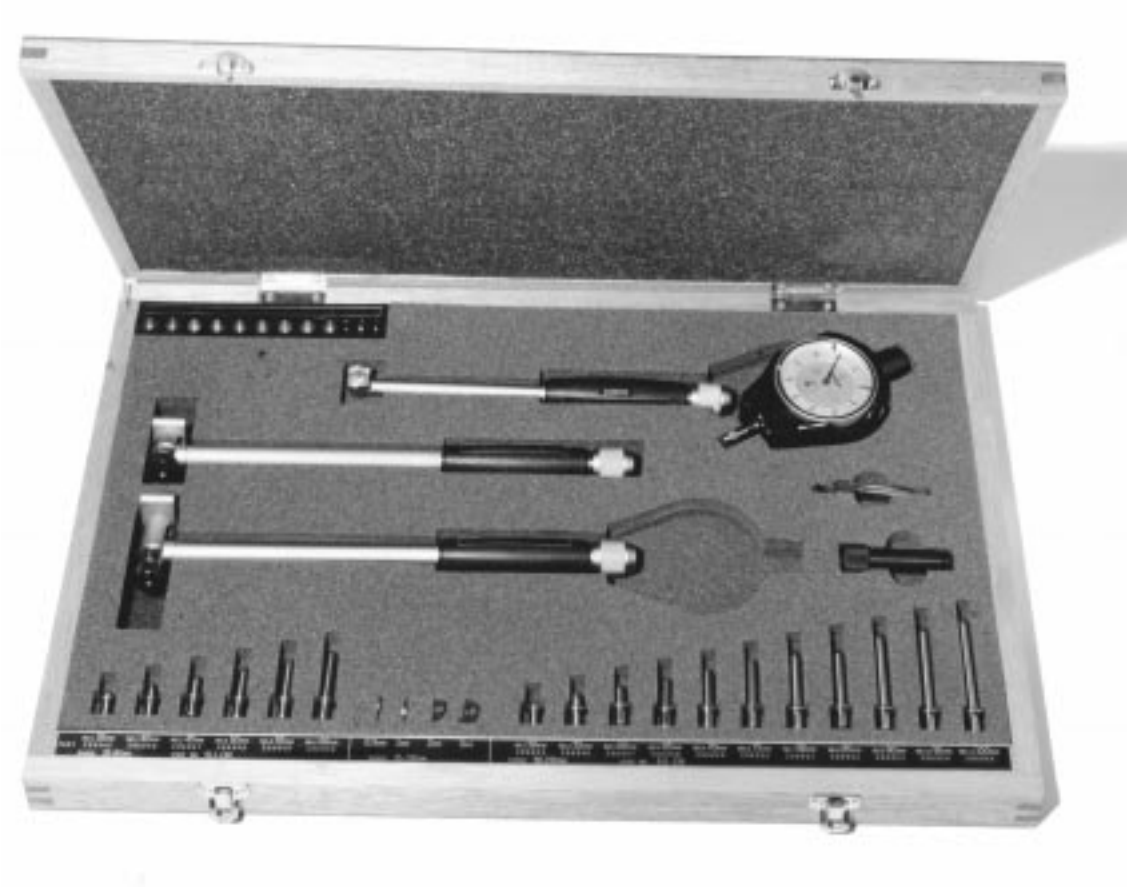


FIGURE 6.55 Set of dial bore gages (Courtesy of MITUTOYO/MTI Corporation).

FIGURE 6.56 Intrimic® plus internal micrometer (Courtesy of Brown & Sharpe Manufacturing Company).

GAGE R AND R Standards

Gage Repeatability and Reproducibility (GAGE R AND R) capability standards have direct implications for parts makers and for gage manufacturers. Repeatability is the ability of an operator using a single gage to obtain the same measurements during a series of tests. Reproducibility is the ability of different

TABLE 6.13 Intrimik® Plus Internal Micrometers by Brown & Sharpe

Range	B&S Tool No.	Price ($)
0.275 in. to 0.350 in. (6–8 mm)	599-290-35	745.40
0.350 in. to 0.425 in. (8–10 mm)	599-290-42	745.40
0.425 in. to 0.500 in. (10–12 mm)	599-290-50	745.40
0.500 in. to 0.600 in. (12–14 mm)	599-290-60	826.70
0.600 in. to 0.700 in. (14–17 mm)	599-290-70	826.70
0.700 in. to 0.800 in. (17–20 mm)	599-290-80	826.70
0.800 in. to 1.0 in. (20–25 mm)	599-290-100	843.60
1.0 in. to 1.2 in. (25–30 mm)	599-290-120	843.60
1.2 in. to 1.4 in. (30–35 mm)	599-290-140	854.30
1.4 in. to 1.6 in. (35–40 mm)	599-290-160	854.30
1.6 in. to 2.0 in. (40–50 mm)	599-290-200	933.10
2.0 in. to 2.4 in. (50–60 mm)	599-290-240	933.10
2.4 in. to 2.8 in. (60–70 mm)	599-290-280	933.10
2.8 in. to 3.2 in. (70–80 mm)	599-290-320	950.20
3.2 in. to 3.6 in. (80–90 mm)	599-290-360	950.20
3.6 in. to 4.0 in. (90–100 mm)	599-290-400	950.20
Intrimik Plus Complete Set #5	599-290-5	3374.60

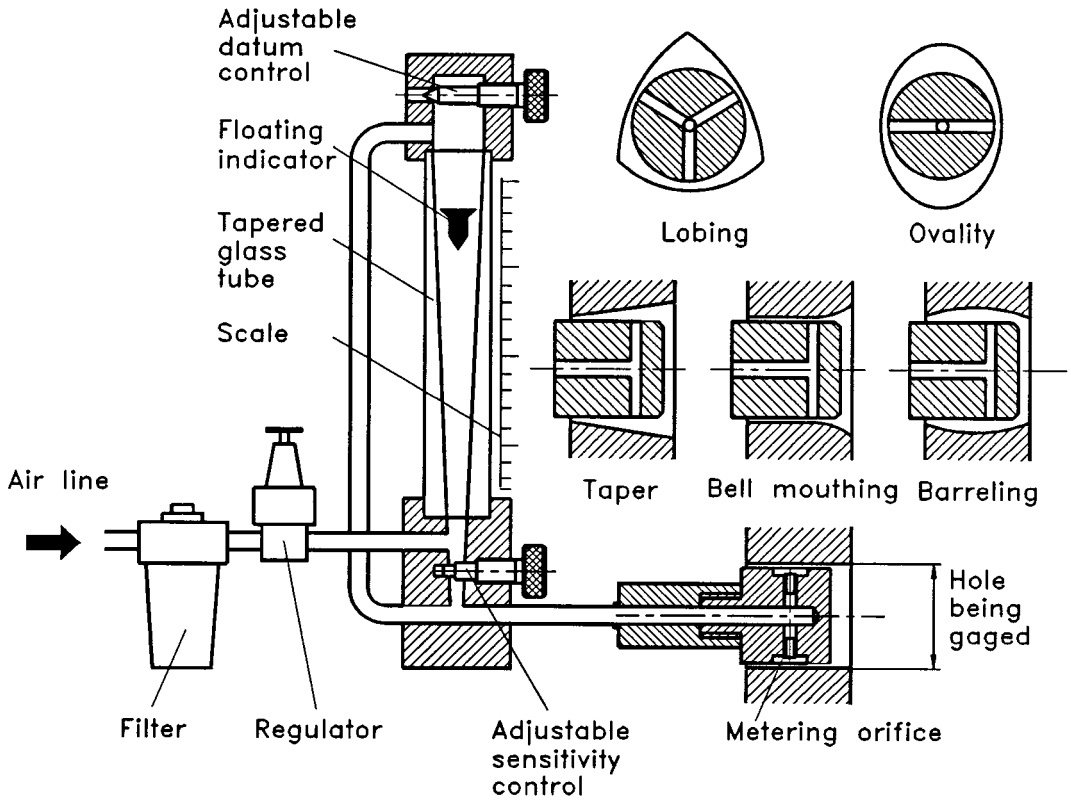


FIGURE 6.57 Air plug gage.

operators to obtain similar results with the same gage. GAGE R AND R blends these two factors together to determine a measuring system's reliability and its suitability for a particular measuring application. For example, a gage design that meets the GAGE R AND R standards for 50 mm (2 in.) bores may be unsatisfactory on 250 mm (10 in.) bores. A gage that meets a tolerance of 2 μm, may not be satisfactory

FIGURE 6.58 TRIOMATIC® electronic bore gage (Courtesy of Brown & Sharpe Manufacturing Company).

TABLE 6.14 TRIOMATIC® Electronic Bore Gages by Brown & Sharpe

Range	B&S Tool No.	Price ($)
0.600 in. to 0.800 in. (12–15 mm)	62-32005	1492.60
0.800 in. to 1.000 in. (20–25 mm)	62-32006	1501.20
1.000 in. to 1.200 in. (25–30 mm)	62-32007	1515.40
1.200 in. to 1.600 in. (30–40 mm)	62-32008	1578.10
1.600 in. to 2.000 in. (40–50 mm)	62-32009	1612.40
2.000 in. to 2.400 in. (50–60 mm)	62-32010	1639.70
2.400 in. to 2.800 in. (60–70 mm)	62-32011	1697.10
2.800 in. to 3.200 in. (70–80 mm)	62-32012	1695.90
3.200 in. to 3.600 in. (80–90 mm)	62-32013	1702.10
3.600 in. to 4.000 in. (90–100 mm)	62-32014	1759.60
4.000 in. to 4.400 in. (100–110 mm)	62-32015	1759.60
4.400 in. to 4.800 in. (110–120 mm)	62-32016	1759.60
TRIOMATIC II Means Sets		
0.600 in. to 1.200 in. (12–30 mm)	62-32001	2161.00
1.200 in. to 2.400 in. (30–60 mm)	62-32002	2663.80
2.400 in. to 3.600 in. (60–90 mm)	62-32003	2509.30
3.600 in. to 4.800 in. (90–120 mm)	62-32004	2752.70

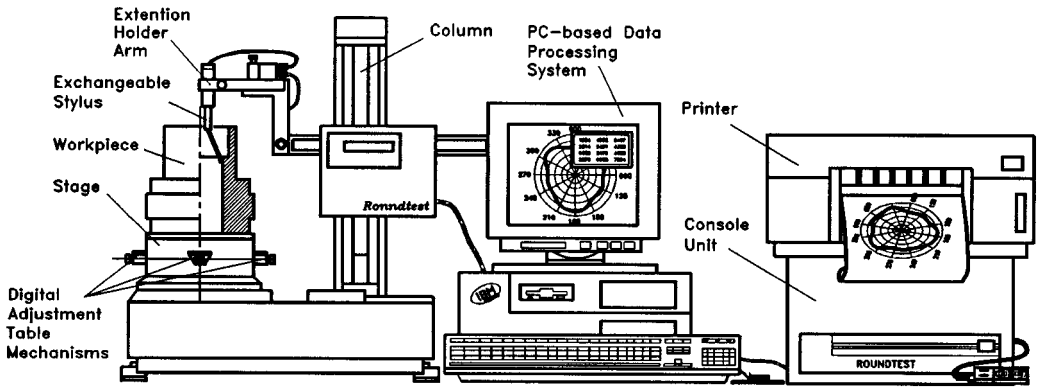


FIGURE 6.59 Form measuring machine.

TABLE 6.15 Rountest Machines by Mitutoyo/MIT Corp.

Model		RA-112	RA- 334	RA-434	RA-661
Measuring range	Max. measuring dia.	11 in. (280 mm)	11.8 in. (300 mm)	11.8 in.(300 mm)	
	Max. measuring height	8.6 in. (220 mm)	27.6 in. (700 mm)	13.8 in. (350 mm)	
	Max. loading dia.	—	21.7 in. (550 mm)	20.4 in. (520 mm)	
	Max. loading capacity	22 lb (10 kg)	66.1 lb (30 kg)	132 lb (60 kg)	
Detector	Range	±0.01 in. (±250 μm)	±0.012 in. (± 300 μm)	±0.012 in. (± 300 μm)	
	Measuring force		7–10 gf	7–10 gf	
Turntable	Rotating accuracy	(1.6 + 0.3H) μinch	(1.6 + 0.6H) μinch	(1.6 + 0.6H) μinch	
		(0.04 + 0.3H) μm	(0.04 + 0.6H) μm	(0.04 + 0.6H) μm	
	Centering adj. range	±0.08 in. (±2 mm)	±0.2 in. (±5 mm)	±0.2 in. (±5 mm)	
	Leveling adj. range		±1°	±1°	
	Rotating speed		6 rpm	2, 4, 6 rpm	
Z-axis column	Straightness	—	—	40 μinch/7.9 in.	8 μinch/8 in.
				1 μm/200 mm	0.2 μm/200 mm
	Parallelism	—	—	120 μinch/7.9 in.	
				80 μinch/13.8 in.	
				3 μm/200 mm	2 μm/200 mm
	Stroke	10 in. (25 mm)		18.9 in. (480 mm)	13.8 in.
Measuring magnifications		100—20,000×		100–50,000×	100–100,000×
Dimensions	Measuring unit	9.9 × 15.9 × 21.5 in.		24.4 × 19.7 × 36.2 in.	28.7 × 23.2 × 62.2 in.
$W \times D \times H$		251 × 404 × 576 mm		620 × 500 × 920 mm	730 × 590 × 1580 mm
	Electric unit	11.4 × 11.8 × 3.6 in.		9.8 × 16.1 × 13 in.	30.7 × 23.3 × 28.8 in.
		290 × 300 × 92 mm		250 × 410 × 330 mm	780 × 592 × 732 mm
Mass	Measuring unit	61.7 lb (28 kg)	276 lb (125 kg)		298 lb (135 kg)
					770 lb (350 kg)
	Electric (analyzer) unit	11 lb (5 kg)		24.2 lb (11 kg)	110 lb (50 kg)
Base price ($)		17,000	45,000	60,000	60,000

FIGURE 6.60 Fully automatic form measuring machine RA-600 (Courtesy of MITUTOYO/MTI Corporation).

at a tolerance of 1 μm. The GAGE R AND R standards will help parts makers identify the best gage for each application and at each tolerance.

The growing acceptance of ISO 9004 as an international quality philosophy is creating significant changes for manufacturers. The most important change that ISO 9004 will create is the need for international GAGE R AND R capability standards and the guarantee of a level playing field. Previously, gage manufacturers and parts makers had many different sets of measuring criteria. Now, in establishing GAGE R AND R capability standards along with ISO 9004 standards, there will be an international set of standards that applies to everyone. A set of measurements from Europe will mean the same as a set from North America or a set from Asia.

The factors that influence GAGE R AND R capability when gaging a bore include:

1. Variation resulting from the bore gage: This variation includes linearity, repeatability, stability, and calibration.
2. Variation resulting from the operation using a bore gage: This variation includes repeatability by an individual operator using the gage and reproducibility by different operators using the same gage.
3. Variation resulting from the production line: Part surface finish, application setup errors, and temperature changes cause this variation.

Dyer company (Dyer, Lancaster, PA 17604) reported the following ranges of these variations for its 830 and 230 series bore gages: linearity ±0.001 mm (±0.000040 in.); repeatability $<\pm$0.00025 mm (±0.000010 in.); stability — bore gages are made from a special alloy steel that resists temperature changes

TABLE 6.16 Companies that Make Bore Gages

Company	Products
Brown & Sharpe Manufacturing Company Precision Park, 200 Frenchtown Road North Kingstown, RI 02852-1700 Tel: (800) 283-3600 Fax: (401) 886-2553	All types of bore gages (from calipers to coordinate measuring machines)
MITUTOYO/MTI Corporation Corporate Headquarters 965 Corporate Blvd. Aurora, IL 60504 Tel: (708) 820-9666 Fax: (708) 820-1393	The world's largest manufacturer of precise measuring instruments; all types of bore gages
Deadborn Gage Company 32300 Ford Road Garden City, MI 48135 Tel: (313) 422-8300 Fax: (313) 422-4445	Air spindles, gaging systems, modular electronic and air-electronic precision gaging instruments
The Dyer Company 1500 McGovernville Road Box 4966 Lancaster, PA 17604 Tel: (800) 631-333 Fax: (717) 569-6721	Dimensional mechanical and electronic bore gages for use in the shop floor
The L.S. Starrett Co. 121 Cresent Street Athol, MA 01331-1915 Tel: (508) 294-3551 Fax: (508) 249-8495	Dimensional mechanical and electronic bore gages, coordinate measuring machines, optical comparators
Flexbar Machine Corporation 250 Gibbs Road Islandia, NY 11722-2697 Tel: (800) 883-5554 Fax: (516) 582-8487	All types of bore gages, including deep-hole gages and comparators
Federal Products Co. 1144 Eddy Street P.O. Box 9400 Providence, RI 02940 Tel: (800) FED-GAGE, (401) 784-3100 Fax: (401) 784-3246 Internet: www.gardnerweb.com/federal/index.html	All types of bore gages: indicator gages, air gages, electronic gaging products, dimensional standards, geometry measurement, laboratory gages
Comtorgage Corporation 58 N.S. Industrial Drive Slatersville, RI 02876 Tel: (401) 765-0900 Fax: (401) 765-2846	Dial indicating — expansion plug system dedicated to one specific size. Customized bore gages
Marposs Corporation Auburn Hills, MI 48326-2954 Tel: (800) 811-0403 Fax: (810) 370-0990	In-process and post-process gaging for grinders and lathes, automatic gaging and special gaging machines
Rank Taylor Hobson Inc. 2100 Golf Road, Suite 350 Rolling Meadows, IL 60008 Tel: (800) 464-7265 Fax: (847) 290-1430	Metrology systems for precision metalworking and high tolerance engineering products
Sterling Mfg. & Engineering Inc. 7539 19th Mile Road Sterling Heights, MI 48314 Tel: (800) 373-0098 Fax: (810) 254-3601	Custom-designed dimensional gages including air, manual, electronic and computerized systems

TABLE 6.16 (continued) Companies that Make Bore Gages

Company	Products
TRAVERS™ TOOL Co. Inc. 128-15 26th Ave. P.O. Box 541 550 Flushing, NY 11354-0108 Tel: (800) 221-0270 Fax: (800) 722-0703	Global Sales Distributer 1 or 2 day delivery at ground rates

or "hand heat." The sealed construction ensures performance reliability even under damp or oily conditions; repeatability by an individual operator — the gages are non-tipping and self-centering in both the axial and radial planes of the bore. The operator can keep his/her hands on or off the gage. The operator has no influence when measuring the bore. The automatic alignment of the bore gage results in a highly reproducible reading.

Defining Terms

Dimension: A numerical value expressed in appropriate units of measure and indicated on drawings with lines, symbols, and notes to define the geometrical characteristics of an object.

Tolerance: The total permissible variation of size, form, or location.

Fit: The general term used to signify the range of tightness which may result from the application of a specific combination of allowances and tolerances in the design of mating parts.

ISO System of Limits and Fits: A standardized system of limits and fits, a group of tolerances considered as corresponding to the same level of accuracy for all basic sizes.

Gage: A device for investigating the dimensional fitness of a part for specified function.

Gaging: A process of measuring manufacturing materials to assure the specified uniformity of size and contour required by industries.

References

1. J. W. Greve (ed.), *Handbook of Industrial Metrology,* ASTM Publications Committee, Englewood Cliffs, NJ: Prentice-Hall, 1967.
2. M. F. Spotts, *Dimensioning and Tolerancing for Quantity Production,* Englewood Cliffs, NJ: Prentice-Hall, 1983.
3. E. R. Friesth, *Metrication for Manufacturing*, New York: Industrial Press Inc., 1978.
4. J. G. Nee, *Fundamentals of Tool Design, 4th ed.,* Dearborn, MI: Society of Manufacturing Engineers, 1998.

Further Information

F. T. Farago, *Handbook of Dimensional Measurement, 3rd ed.*, New York: Industrial Press Inc., 1994, provides extensive definitions of terms, methods, and measuring setups.

J. Dally, I*nstrumentation for Engineering Measurements, 2nd ed.*, New York: Wiley, 1993, provides detailed characteristics of measuring tools and machines.

R. S. Sirohi, *Mechanical Measurements, 3rd ed.,* New York: Wiley, 1991, provides a textbook presentation of the theory of mechanical measurements.

J. D. Meadows, *Geometric Dimensioning and Tolerancing: Applications and Techniques for Use in Design, Manufacturing, and Inspection,* New York: M. Dekker, 1995, presents basic rules and procedures for good dimensioning practice, providing the background that aids in the solution of dimensioning problems as they arise in day-to-day work.

6.7 Time-of-Flight Ultrasonic Displacement Sensors

Teklic Ole Pedersen and Nils Karlsson

Ultrasound is an acoustic wave with a frequency higher than the audible range of the human ear, which is 20 kHz. Ultrasound can be within the audible range for some animals, like dogs, bats, or dolphins. In the years around 1883, Sir Francis Galton performed the first known experiments with whistles generating ultrasound. Many decades later, people started to find ultrasound applications in engineering, medicine, and daily life. The basic principle for the use of ultrasound as a measurement tool is the *time-of-flight technique*. The pulse-echo method is one example. In the pulse-echo method, a pulse of ultrasound is transmitted in a medium. When the pulse reaches an another medium, it is totally or partially reflected, and the elapsed time from emission to detection of the reflected pulse is measured. This time depends on the distance and the velocity of the sound. When sound travels with a known velocity c, the time t elapsed between the outgoing signal and its incoming echo is a measure of the distance d to the object causing the echo.

$$d = \frac{ct}{2} \tag{6.82}$$

Figure 6.61 shows a simple pulse-echo system. The transmitter and the receiver could be the same device, but they are separated for clarity in this figure.

The oscillator generates an electric signal with a typical frequency of 40 kHz. This electric signal is transformed into mechanical vibrations of the same frequency in the transmitter. These vibrations generate sound waves that are reflected by the object. The reflected sound echo causes an electric signal in the receiver. For precise measurements, the speed of sound is a crucial parameter. A typical value in air at 1 atm pressure and room temperature is 343 m s^{-1}, but the speed of sound is influenced by air pressure, air temperature, and the chemical composition of air (water, CO_2, etc.). For example, the speed of sound is proportional to the square root of absolute temperature. Measuring distances in an environment with large temperature gradients can result in erroneously calculated distances. As an advantage, ultrasound waves are robust against other disturbances such as light, smoke, and electromagnetic interference [1–4].

Physical Characteristics of Sound Waves

Sound is a vibration in matter. It propagates as a longitudinal wave, i.e., the displacement in the material is in the direction of the sound wave propagation. A plane wave that propagates in the x direction can be described by

$$\Delta x = A\sin\omega\left(t - \frac{x}{c}\right) \tag{6.83}$$

where A is the amplitude, $\omega = 2\pi f$, f being the frequency of the wave and Δx is the displacement of a particle at time t at the position x.

The velocity of sound depends on the medium in which it propagates. In a homogeneous and isotropic solid, the velocity depends on the density ρ and the modulus of elasticity E according to Equation 6.84.

$$c = \sqrt{\frac{E}{\rho}} \tag{6.84}$$

In a liquid, the velocity depends on the density and the adiabatic compressibility K, Equation 6.85.

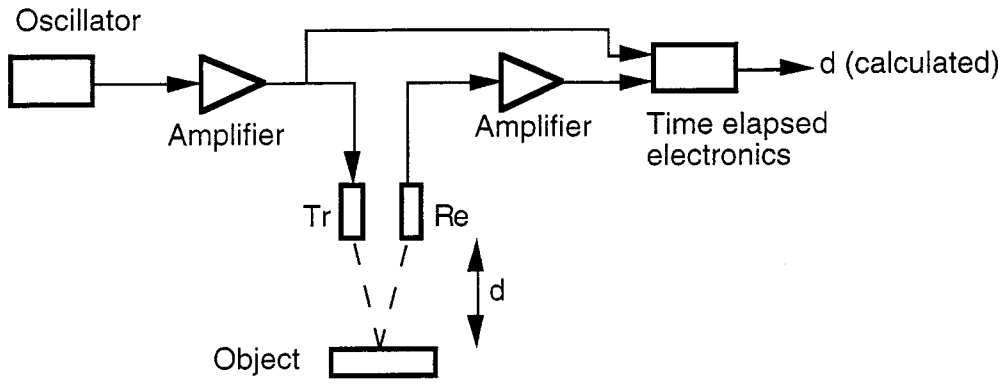


FIGURE 6.61 Principle of a pulse-echo ultrasound system for distance measurements (Tr = transmitter, Re = receiver).

$$c = \sqrt{\frac{1}{K\rho}} \tag{6.85}$$

In gases, the velocity of sound is described by Equation 6.86. Here g represents the ratio of the specific heat at constant pressure (c_p) to the specific heat at constant volume (c_v), p is pressure, R is the universal gas constant, T is the absolute temperature, and M is the molecular weight.

$$c = \sqrt{\frac{gRT}{M}} = \sqrt{\frac{c_p}{c_v}\frac{p}{\rho}} \tag{6.86}$$

An important quantity is the *specific acoustic impedance*. It is, in general, a complex quantity but in the far field (Figure 6.63), the imaginary component disappears, leaving a real quantity. This real quantity is the product of the density ρ and the sound speed c in the medium. This product is called the characteristic impedance R_a (Equation 6.87).

$$R_a = \rho c \tag{6.87}$$

The characteristic impedance is thus independent of the sound frequency.

An acoustic wave has an intensity I (rate of flow of energy per unit area), which can be expressed in watts per square meter (W m^{-2}). A usually unwanted phenomenon arises when the sound wave has to pass from one medium with characteristic impedance R_1 to another medium with characteristic impedance R_2. If R_1 and R_2 have different values, a part of the wave intensity will reflect at the boundary between the two media (see Figure 6.61 and 6.62). The two media are said to be mismatched, or poorly coupled, if a major part of the wave intensity is reflected and a minor part is transmitted. The relative amounts of reflected and transmitted wave intensities can be defined by:

$$\text{Reflection coefficient} = \frac{I_{\text{refl}}}{I_{\text{incident}}} \tag{6.88a}$$

$$\text{Transmission coefficient} = \frac{I_{\text{trans}}}{I_{\text{incident}}} \tag{6.88b}$$

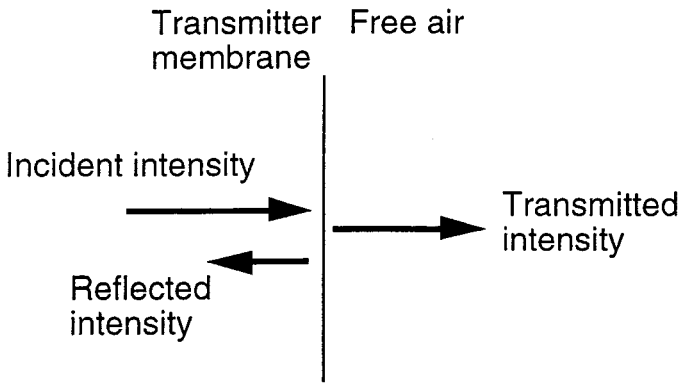


FIGURE 6.62 Reflection and transmission of a sound wave at the interface between media of different characteristic impedances.

It can be shown [1] that these coefficients have simple relations to the previously mentioned characteristic impedances.

$$\text{Reflection coefficient} = \frac{\left(R_1 - R_2\right)^2}{\left(R_1 + R_2\right)^2} \tag{6.89a}$$

$$\text{Transmission coefficient} = \frac{4R_1R_2}{\left(R_1 + R_2\right)^2} \tag{6.89b}$$

The practical importance of the acoustic impedance is realized when the ultrasonic pulse-echo system shown in Figure 6.61 is considered. First, the electric energy is converted into mechanical vibrations of a membrane in the transmitter. Second, the vibrations (the sound wave) have to pass through the boundary between the membrane (usually a solid material) and free air. Because the transmitter membrane and the free air have different characteristic impedances, much of the acoustic intensity is reflected (Figure 6.62).

The transmitted ultrasound in free air will first propagate in a parallel beam (near field of the transducer); but after a distance L, the beam diverges (the far field of the transducer). See Equation 6.90 and Figure 6.63.

$$L \approx \frac{D^2}{4\lambda} \tag{6.90}$$

D is the diameter of the circular transmitter and λ is the wavelength of the ultrasound.

The sound intensity in the near field is complicated due to interference effects of sound originating from different parts of the transducer membrane. In the far field, the intensity is approximately uniform and the beam spread follows:

$$\sin\beta = 1.22\,\frac{\lambda}{D} \tag{6.91}$$

where β is the half lobe angle.

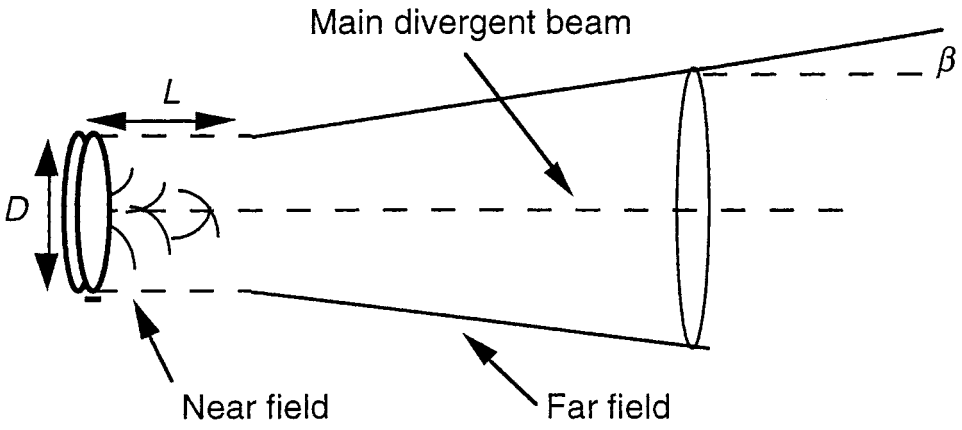


FIGURE 6.63 Illustration of the ultrasound beam in the near field and the far field of the transducer.

To get a narrow beam, the transmitter membrane diameter must be large with respect to the wavelength. High-frequency ultrasound cannot be the general solution as ultrasound of a high frequency is absorbed faster than ultrasound of a low frequency [1–4].

Ultrasound Transducers

Most ultrasound transducers convert electric energy to mechanical energy and vice versa. The most common types of in-air transducers are [5–7]:

1. Mechanical
2. Electromagnetic
3. Piezoelectric
4. Electrostatic
5. Magnetostrictive

The simplest type, mechanical transducers such as whistles and sirens, are used up to approximately 50 kHz. This type works only as a transmitter.

Electromagnetic transducers such as loudspeakers and microphones can be used for ultrasonic wave generation, but they are mainly suited for lower frequencies.

The piezoelectric transducer (Figure 6.64) is more suitable for use in ultrasonics and is quite common. It uses a property of piezoelectric crystals: they change dimensions when they are exposed to an electric field. When an alternating voltage is applied over the piezoelectric material, it changes its dimensions with the frequency of the voltage. The transducer is mainly suited for use at frequencies near the mechanical resonance frequency of the crystal. The piezoelectric transducer can be both a transmitter and a receiver: when a piezolectric material is forced to vibrate by a sound pulse, it generates a voltage. Some natural crystals, such as quartz, are piezoelectric. Ceramics can be polarized to become piezoelectric; so can some polymers like PVDF (polyvinylidene fluoride). Polymers are suitable as transducers in air since their acoustic impedance is low [8–10] compared with other standard piezoelectric materials.

The electrostatic transducer (Figure 6.64) is a plate capacitor with one plate fixed and the other free to vibrate as a membrane. When a voltage is applied between the plates, the electrostatic forces tend to attract or repel the plates relative to each other depending on the polarity of the voltage. This transducer can be used both as a transmitter and a receiver [11].

The magnetostrictive transducer is based on the phenomenon of magnetostriction, which means that the dimensions of a ferromagnetic rod change due to the changes of an externally applied magnetic field. This transducer can also act as both a receiver and a transmitter.

FIGURE 6.64 Ultrasonic transducers piezoelectric (left) and electrostatic (right).

Principles of Time-of-Flight Systems

There are several techniques for ultrasonic range measurements [12–15].

The previously described *pulse echo method* is the simplest one. Usually, this method has a low signal-to-noise ratio (SNR) because of the low transmitted energy due to the short duration of the pulse. Multireflections are detectable.

In the *phase angle method*, the phase angle is measured between the continuous transmitted signal and the continuous received signal and is used as a measure of the distance. The method is relatively insensitive to disturbances. Multireflections are not detectable in a meaningful way. When the distance is longer than one wavelength, another method must be used to monitor the distance.

The *frequency modulation method* uses transmitted signals that are linearly frequency modulated. Thus, detected signals are a delayed replica of the transmitted signal at an earlier frequency. The frequency shift is proportional to the time-of-flight. The method is robust against disturbing signals, and multireflections are detectable.

The *correlation method* (Figure 6.65) determines the cross-correlation function between transmitted and received signals. When the transmitted signal is a random sequence, i.e., white Gaussian noise, the cross-correlation function estimates the impulse response of the system, which, in turn, is a good indicator of all possible time delays. The method is robust against disturbances, and multireflections are detectable.

Industrial acoustic noise can affect the received signals in an ultrasound time-of-flight system. The noise can be generated from leaking compressed air pipes, noisy machines, or other ultrasonic systems. This external noise is not correlated with the relevant echo signals of the transmitted noise and can therefore be eliminated by the use of correlation methods. Disturbances correlated with the relevant echo signal (e.g., unwanted reflections) will not be eliminated by the use of correlation methods.

The impulse response $h(t,t_0)$ is used as a sensitive indicator of time delay between transmitted signal at time t_0 and received signal at time t. The impulse response is given by [14]:

$$h\left(t, t_0\right) = F^{-1}\left[\frac{S_{xy}}{S_{xx}}\right] \tag{6.92}$$

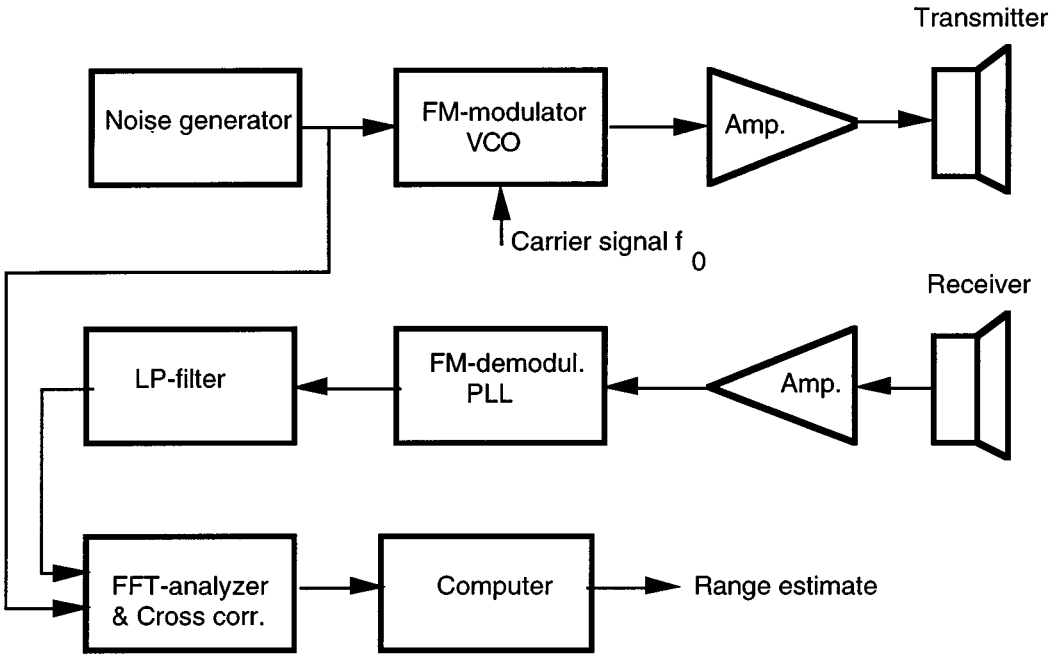


FIGURE 6.65 Diagram of a correlation-based time-of-flight system.

where F^{-1} is the inverse Fourier transform, x is the transmitted signal, y is the received signal, $S_{xy}(f)$ is the cross-spectral density function [the Fourier transform of the cross-correlation function of the transmitted signal $x(t)$ and the received signal $y(t)$] and S_{xx} is the power density function (the Fourier transform of the auto-correlation function of the transmitted signal).

To analyze the transfer channel by data acquisition requires a high sampling rate, theoretically at least two times the highest frequency component in the received signal (and in practice as high as 10 times the highest frequency). One way to reduce the sampling rate is to first convert the signal from its bandpass characteristics around a center frequency f_0 (approx. 50 kHz) to lowpass characteristics from dc to $B/2$, where B is the appropriate bandwidth. The accuracy of the range estimate, and hence the time interval $t - t_0$, can be improved by processing the estimate of the impulse response $h(t - t_0)$ with a curve-fitting (least square) method and digital filtering in a computer. A block diagram of a correlation-based time-of-flight system is shown in Figure 6.65. Further details and complete design examples can be found in the literature [12–15].

Table 6.17 lists some advantages and drawbacks of the described time-of-flight methods.

TABLE 6.17 Advantages and Disadvantages of Time-of-Flight Methods

Method	Main advantage	Main disadvantage
Pulse echo method	Simple	Low signal-to-noise ratio
Phase angle method	Rather insensitive to disturbances	Cannot be used directly at distances longer than the wavelength of the ultrasound
Frequency modulation method	Robust against disturbances; multireflections detectable	Can give ambigous results measurements on long and short distances can give the same result (compare with phase angle method)
Correlation method	Very robust against disturbances	Make relatively high demands on hardware and/or computations

References

1. J. Blitz, *Elements of Acoustics,* London: Butterworth, 1964.
2. L. E. Kinsler, A. R. Frey, A. B. Coppens, and T. V. Sanders, *Fundamentals of Acoustics, 3rd ed.,* New York: John Wiley & Sons, 1982.
3. G. Lindstedt, Borrowing the bat's ear for automation. Ultrasonic measurements in an industrial environment, Dept. of Industrial electrical engineering and automation, Lund Institute of Technology, 1996.
4. G. S. Kino, *Acoustic Waves: Devices, Imaging and Analog Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1987.
5. S. R. Ruocco, Robot Sensors and Transducers, New York: John Wiley & Sons, 1987.
6. P. H. Sydenham and R. Thorn, *Handbook of Measurement Science, Vol. 3, Elements of Change,* New York: John Wiley & Sons, 1992.
7. J. Fraden, *AIP Handbook of Modern Sensors, Physics, Design and Applications,* New York: American Institute of Physics, 1993.
8. H. R. Gallantree, Review of transducer applications of polyvinylidene fluoride, *IEEE Proceedings,* 130, 219–224, 1983.
9. T. T. Wang and J. M. Herbert, *The Applications of Ferroelectric Polymers,* London: Chapman & Hall, 1988.
10. C. Z. Rosen, B. V. Hiremath and R. Newnham, *Piezoelectricity,* New York: American Institute of Physics, 1992.
11. P. Mattila, F. Tsuzuki, H. Väätäjä, and K. Sasaki, Electroacoustic Model for Electrostatic Ultrasonic Transducers with V-Grooved Backplates, in *IEEE Trans. Ultrasonics, Ferroelectrics and Frequency Control,* Vol. 42, No. 1, January, 1995.
12. P. Holmberg, Instrumentation, Measurements and Applied Signal Processing for Industrial Robot Applications, Ph.D. dissertation No. 334, Dept. of Physics and Measurement Technology, Linköping University, 1994.
13. J. A. Kleppe, *Engineering Applications of Acoustics,* Boston, 1989.
14. J. S. Bendat and A. G. Piersol, *Random Data Analysis and Measurement Procedures, 2nd ed.,* New York: John Wiley & Sons, 1986.
15. P. Holmberg, Robust ultrasonic range finder — an FFT analysis, *Meas. Sci. Technol.,* 3, 1025–1037, 1992.

6.8 Optical Encoder Displacement Sensors

J. R. René Mayer

The detection of angular and linear motion is a key function in a multitude of systems such as machine tools, industrial robots, a variety of instruments, computer mice, etc. Although they are one of many techniques capable of such measurements, the ease with which they are interfaced to digital systems has made them very popular.

Optical encoders are used to measure either angular or linear positions. Those used for angular detection are commonly called rotary or shaft encoders, since they usually detect the rotation of a shaft. Optical encoders encompass a variety of devices, all of which use light as the means to transform movement into electrical signals. All devices have two basic building blocks: a main grating and a detection system. It is the position of one with respect to the other that is detected. The main grating represents the measurement standard. For linear measurements, the main grating, commonly called the scale, is one or more sets of parallel lines of constant or specially coded pitch supported by a substrate. Similarly, a rotary encoder has a grating with radial lines on a disk.

Both linear and rotary encoders can, in principle, be absolute or incremental, although in practice, linear absolute encoders employing optical principles are quite uncommon and have drastically limited

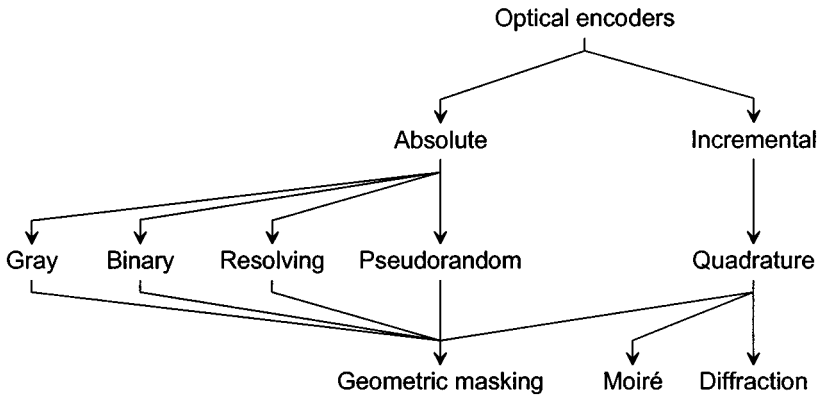


FIGURE 6.66 Classifications of optical encoders based on (1) the nature of the final information provided, (2) the type of signals generated, and (3) the technology used to generate the signals.

performance characteristics (accuracy, resolution, and/or maximum operating speed). Figure 6.66 shows a simplified classification of optical encoders. This classification refers to the nature of the information generated. The incremental encoder detects movement relative to a reference point. As a result, some form of reference signal is usually supplied by the encoder at a fixed position in order to define a reference position. The current position is then incremented (or decremented) as appropriate. Multiple reference marks can also be used, where the distance between successive marks is unique so that as soon as two successive marks have been detected, it becomes possible to establish absolute position from then on. The reference point can also be mechanical. Should power be lost or a signal transmission error occur, then the absolute position is lost and the encoder must return to one or more reference points in order to reset its counters. Unfortunately, a loss of count may not be detected until a reference point is reaccessed. Furthermore, reading errors may accumulate. On the other hand, absolute encoders produce a set of binary signals from which the absolute position can be deduced without the knowledge of the previous motion history. The current position is known right from powering-on. In the case of absolute rotary encoders, single and multiturn devices are available. Multiturn devices use an internal mechanical transmission system to drive a second grating that serves as turn counter.

Most incremental encoders use quadrature signals as output to carry the motion information. Some encoders use one square-wave signal, which is used for position in one direction only. Also, this single square wave can be fed into either a PLC (programmable logic controller) or another electronic interface that converts this signal to a rate or RPM (revolution per minute) for speed indication. However, whenever bidirectional operation is required, quadrature signals are necessary. Quadrature signals come in analog or digital form. The analog form consists simply of a sine and a cosine signal. The number of sinusoidal cycles per unit change of the measured variable (a revolution or 360° for a rotary encoder) determines the basic resolution of the encoder prior to interpolation. The digital form consists of two square-wave trains, 90° (often called electrical degree) out of phase. The 90° phase lag is indispensable in order to detect the motion direction and hence increment or decrement the position counter accordingly. The main optical techniques to generate the quadrature signals are geometric masking, Moiré fringes, and diffraction based. For linear encoders, the basic resolution is related to the distance traveled by the grating in order for the encoder to produce one full quadrature cycle. For rotary encoders, the basic resolution is usually described as the number of quadrature cycles per turn. The resolution of an encoder system can be increased by electronic means. With analog quadrature signals, it is possible to interpolate within each quadrature cycle. The limit of the interpolation factor depends on the quality (mark space, quadrature separation, and jitter) of the basic signals. With square-wave signals, multiplication by a factor of two or four is easily achieved. Increasing the resolution in this manner does not, however, improve the trueness, often called accuracy (or linearity) of the measurement.

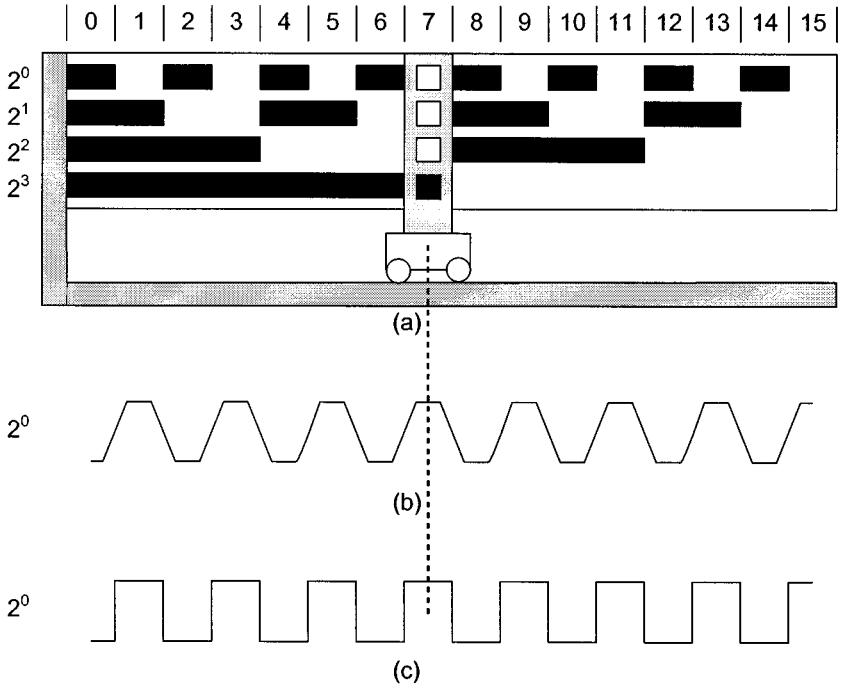


FIGURE 6.67 (*a*) Absolute encoders using a natural binary code of four digits. Four tracks are required. The moving read head has four apertures and is shown in position 7 along the scale. (*b*) The output of the read head aperture corresponding to the least significant track. It represents the proportion of light area covering the aperture. (*c*) The binary digit obtained after squaring the raw output signal.

Absolute encoders are classified according to the type of code used. The main four codes are Gray, binary (usually read by vee-scan detection), optical resolving, and pseudorandom. All absolute encoders use geometric masking to generate the code.

Encoder Signals and Processing Circuitry

Absolute Encoders

Direct Binary

Figure 6.67(*a*) illustrates the concept of an absolute linear optical encoder using a direct binary encoded scale. The fixed scale has n tracks (here n is 4), each providing one bit of a direct binary number. The lowest track (first track from the center of the disk for a rotary encoder) is the most significant digit and has a weight, 2^{n-1} (here 2^3), while the upper track is the least significant digit with a weight 2^0. The track providing the least significant digit has 2^{n-1} cycles of light and dark records, while the most significant track has 2^0 or 1 such cycle. For each track, the moving read head has a readout unit consisting of a light source, a mask, and a photodetector. Figure 6.67(*b*) shows the output from the photodetector, which represents the total intensity of light reaching its surface. As the mask passes over a clear region of the grating, the photodetector output increases, and then decreases. In theory, a truncated triangular wave is obtained, which can easily be converted to a square wave (Figure 6.67(*c*)) by a suitably chosen thresholding level. The result is a high or 1 for a light record and a low or 0 for a dark one. The position, in base 10, corresponding to the reading head position in Figure 6.67 is

$$1\times 2^0 + 1\times 2^1 + 1\times 2^2 + 0\times 2^3 = 7 \tag{6.93}$$

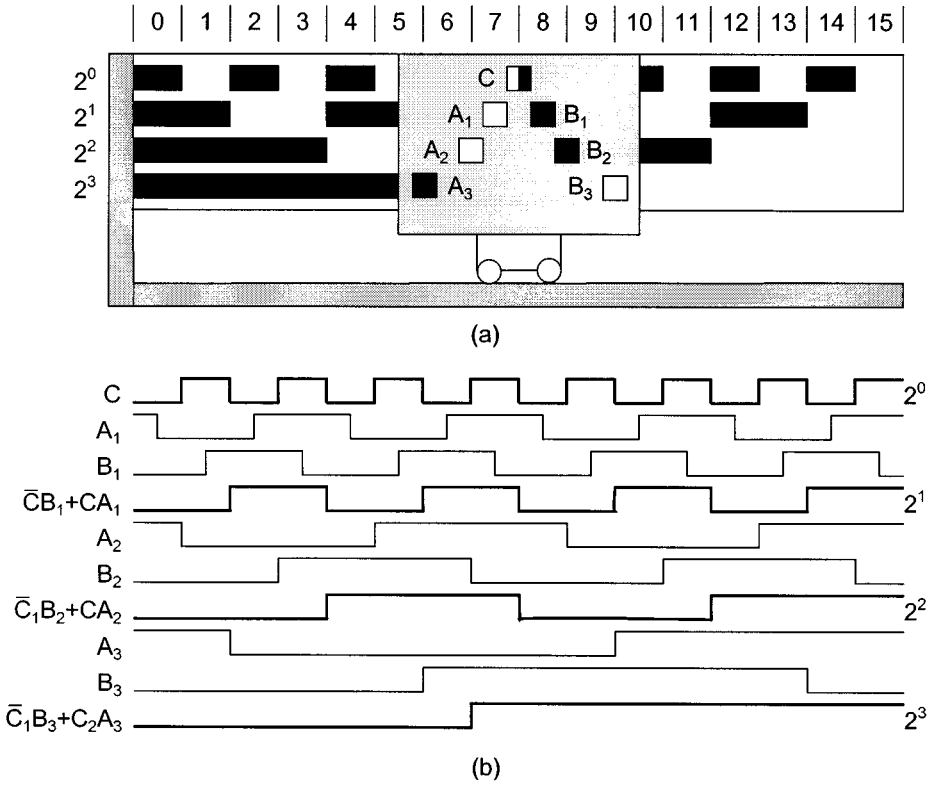


FIGURE 6.68 The vee-scan configuration of reading units in (*a*) removes the ambiguity associated with a natural binary scale. Simple combinational logic is then used to generate the natural binary readout in (*b*).

The code configuration just described is not suitable for practical use because some transitions require that two or more bit values change simultaneously. For example, from position 7 to position 8, all bits change values. Unless the change is simultaneous, an incorrect position readout results at some position. This would require that the scale is geometrically perfect, that the read head be perfectly aligned with the scale, and that the electronics are perfectly adjusted and stable over time. This problem is solved either by the use of a vee-scan detection method or the use of a unit-distance code such as the Gray code.

Vee-scan

The vee-scan method uses a V-shape pattern of readout units that removes the potential reading ambiguity of direct binary scales. Stephens et al. [1] indicate the read points at which transitions are detected in Figure 6.68(*a*). They also describe the conversion of the thresholded output signals to binary code using combinational logic. The primary advantage is that the location tolerance of the transition point of each reading unit need only be ±1/8 of the cycle length for that particular track. For example, ±45° for the most significant track of a rotary encoder disk. Figure 6.68(*b*) shows a direct binary word obtained through logic combinations of the vee-scan readings.

Gray Code

The use of vee-scan requires additional reading heads as well as processing electronics. The Gray code is a unit-distance code and so only one bit of data changes between representations of two consecutive numbers or successive positions. This removes the possibility of ambiguous readout. It has the following advantages: (1) it is easily converted to direct binary code, and (2) the finest tracks are twice the width of equivalent direct binary code tracks. Figure 6.69(*a*) shows a Gray code linear scale. Figure 6.69(*b*) shows a scheme for the conversion from Gray code to binary code and proceeds as follows: (1) the most significant bit (msb) of the binary code equals the msb of the Gray-coded number; (2) add (modulo-2)

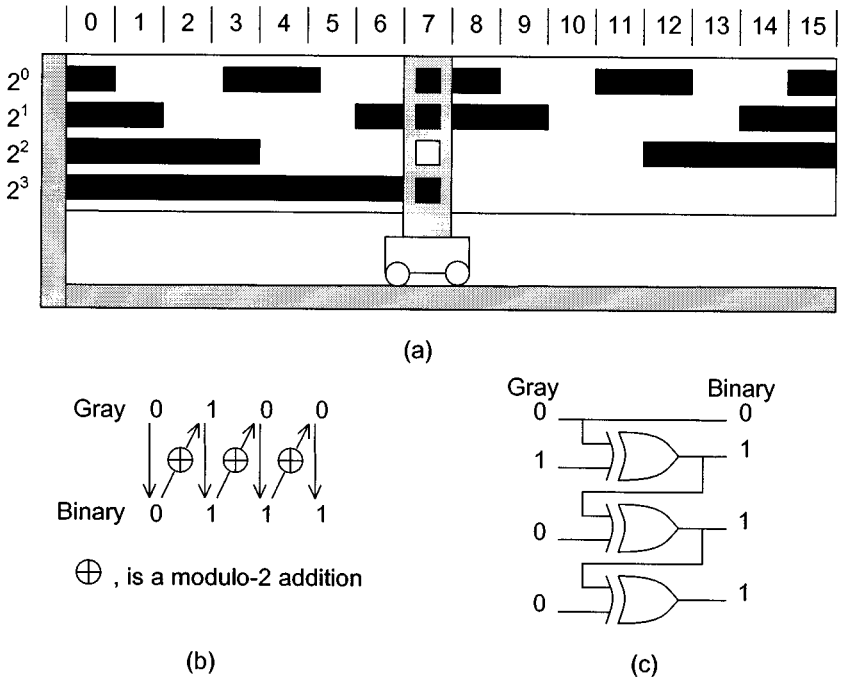


FIGURE 6.69 (*a*) Gray code allows a transition on only one track between each successive position so that no ambiguity arises. A scheme based on modulo-2 additions converts the Gray code to natural binary code in (*b*). Exclusive-ORs implement the conversion in (*c*).

the msb of the binary number to the next significant bit of the Gray-coded number to obtain the next binary bit; and (3) repeat step (2) until all bits of the Gray-coded number have been added modulo-2. The resultant number is the binary equivalent of the Gray-coded number. The modulo-2 addition is equivalent to the action of an exclusive-OR. Figure 6.69(*c*) shows a simple circuit using combinational logic to perform Gray to binary conversion. Sente et al. [2] suggest the use of an external ROM to convert the code where the input coded word is the address and the data is the output coded word. Using a 16-bit ROM for a 12-bit code, Sente et al. [2] suggest using the remaining 4 bits to implement a direction signal. Stephens et al. [1] describe the use of vee-scan with a gray code scale for even better robustness.

Pseudorandom Code

Pseudorandom encoding allows the use of only two tracks to produce an absolute encoder. One track contains the pattern used to identify the current position, while the other is used to synchronize the reading of the encoded track and remove ambiguity problems. A pseudorandom binary sequence (PRBS) is a series of binary records or numbers, generated in such a way that any consecutive series of n digits is unique. Such code is called chain code, and it has the property that the first $n-1$ digits of an n-bit word are identical to the last $n-1$ digits of the previous code word. This allows their partial overlapping on a single track. A PRBS of length 2^n-1 is defined by:

$$\mathrm{XN}\left(j\right)\Big|j=0,1,\ldots,2^{n-1} \tag{6.94}$$

The code can be generated by reading the nth stage of a feedback shift register after j shifts. The register must be initialized so that at least one of the registers is nonzero and the feedback connection implements the formula

$$X\left(0\right)=X\left(n\right)\oplus c\left(n-1\right)X\left(n-1\right)\oplus\ldots\oplus c\left(1\right)X\left(1\right) \tag{6.95}$$

TABLE 6.18 Shift-Register Feedback Connections for Generating Pseudorandom Binary Sequences

n	Length	Direct sequence	Reverse sequence
4	15	1, 4	3, 4
5	31	2, 5	3, 5
6	63	1, 6	5, 6
7	127	3, 7	4, 7
8	255	2, 3, 4, 8	4, 5, 6, 8
9	511	4, 9	5, 9
10	1023	3, 10	7, 10
11	2047	2, 11	9, 11
12	4095	1, 4, 6, 12	6, 8, 11, 12
13	8191	1, 3, 4, 13	9, 10, 12, 13
14	16,383	1, 6, 10, 14	4, 8, 13, 14

where the c coefficients are 0 or 1. The feedback registers for which c is 1 are listed in Table 6.18 for values of n from 4 to 14. The following is a PRBS for $n = 4$ with the pseudocode obtained using all registers set to 1 initially, 111101011001000. For a rotary encoder disk, the 15 sectors would have a 24° width.

Petriu [3, 4] describes a possible configuration of a PRBS disk that uses a PRBS track and a synchronization track (Figure 6.70), together with the processing method to reconstitute the position in natural binary. Table 6.18 gives the reverse feedback configuration. The shift register is initially loaded with the current n-tuple. Then the reverse logic is applied recurrently until the initial sequence of the PRBS is reached. At this point, the n-bit counter represents the value of j. For a rotary encoder $(j * 360)/(2^n - 1)$ is the current angular position; whereas for a linear encoder, the position is $j * P$ where P is the scale record length or pitch. Petriu [4] suggests that in order to allow nonambiguous bidirectional reading, $n + 1$ heads are used on the PRBS track. The synchronization track has a series of 0s and 1s in records of the same width as the PRBS track and in phase. There is a $P/2$ shift, where P is the record's length between the A head on the synchronization track and the $n + 1$ read heads on the PRBS track. The $n + 1$ records are updated on a trigger from the A signal. This ensures that the $n + 1$ heads are closely aligned with the PRBS record mid-position. A second read head called B on the synchronization track is shifted by $P/2$ relative to A. A and B are in quadrature, which allows their simultaneous use to generate a motion direction signal. The correct n-tuple, i.e., the lower or upper subset, is selected on the basis of the moving direction and is then converted to natural binary by reverse feedback. Petriu [4] also suggests a simple means of increasing the resolution by a factor of 2 using some additional electronics. He also proposes a scheme to use an arbitrary (not $2^n - 1$) number of sectors, but this requires a third track and some additional correction electronics to handle the last $n - 1$ records of a disk, since these are no longer PRBS patterns. Tomlinson [5] proposes another method for truncation of the PRBS sequence that does not require a third track. Instead, particular codes were removed by applying additional logic in the direct and reverse feedback logic.

Ross and Taylor [6] and Arsic and Denic [7] suggest ways of reducing the number of reading heads by accumulating readings into a shift register so that a minimum of two heads are sufficient to read the PRBS track. However, on start-up, the correct position is not known until the encoder has moved so that all registers have been updated. This type of encoder is therefore not completely absolute because it does not indicate its correct position on start-up. Finally, Arazi [8] mentions the use of a ROM that stores the translation table, as an alternative to a logic circuit.

Optical Resolving

This method has similarities with its electromagnetic counterpart and depends on the generation of a sine and a cosine signal pair per encoder shaft revolution. The resolution and accuracy of this encoder depend on its ability to generate signals that conform to their ideal waveforms and the resolving power of the electronic circuit responsible for performing the rectangular to polar conversion to produce angular

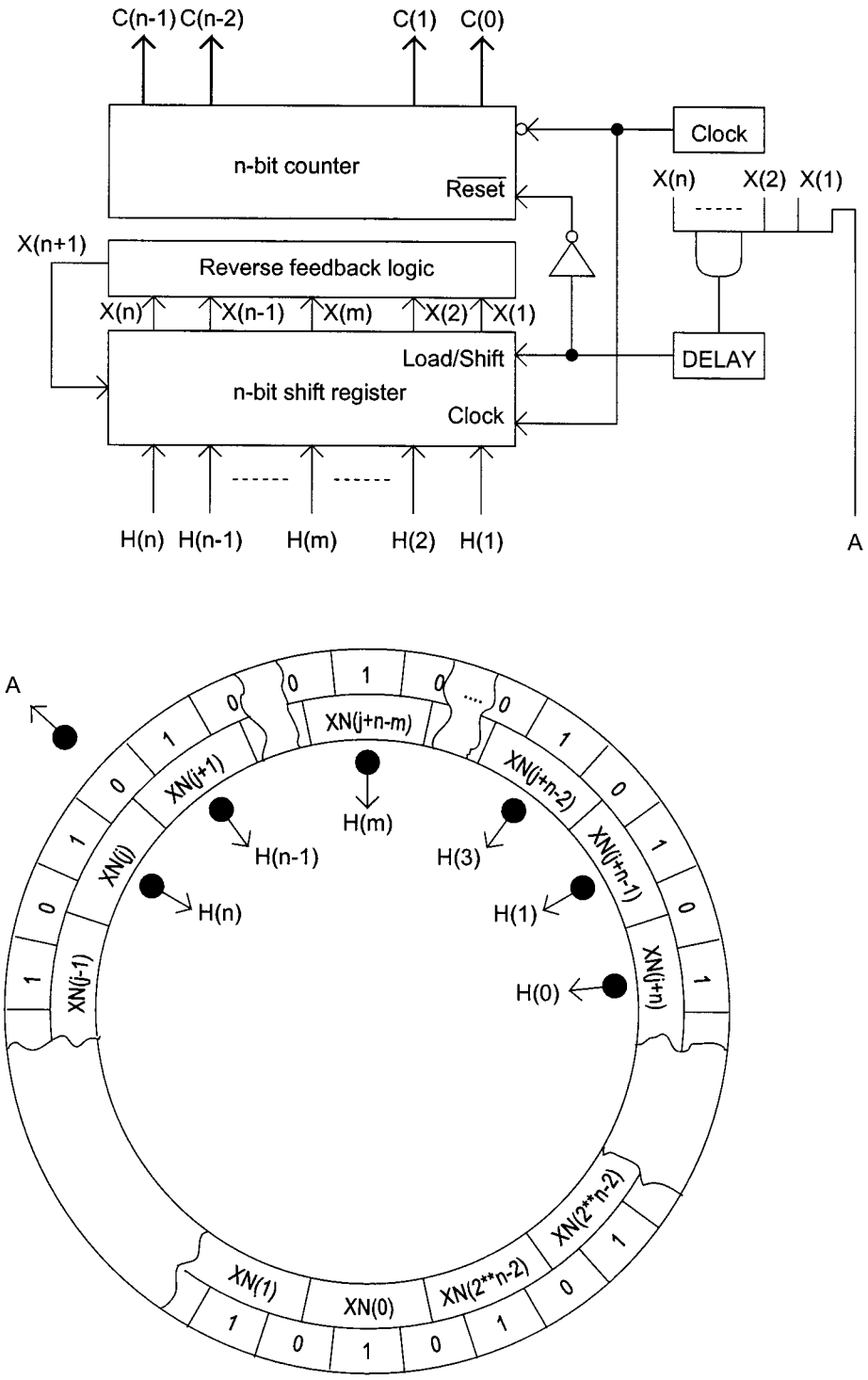


FIGURE 6.70 Pseudorandom shaft encoder with a simple synchronization track to validate the readout from the n-tuple read by the reading heads $H(i)$, $i = n, \ldots, 1$. The circuit is a simplified code conversion to binary based on the reverse feedback logic. The counter counts the number of steps required to return to the initial sequence.

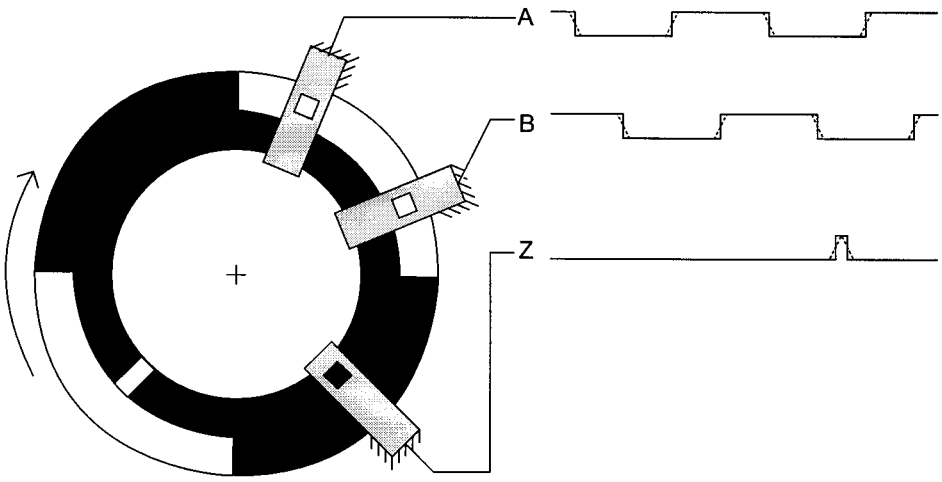


FIGURE 6.71 The lowest possible resolution for a rotary incremental encoder. The outer track provides $n = 2$ cycles of the alternating black and clear records. Two heads A and B displaced by 90° give, after squaring, two square waves in quadrature.

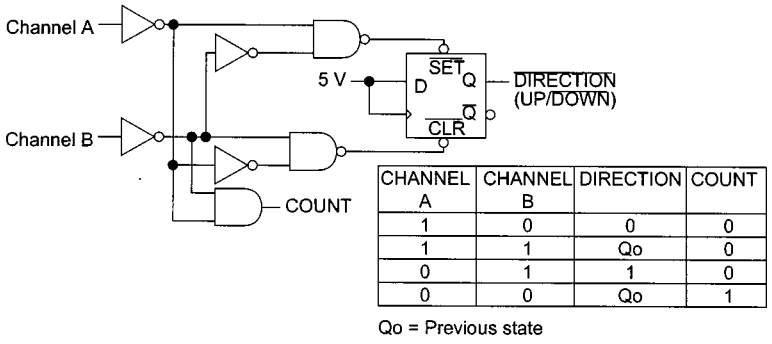
output data. The methods for accomplishing this conversion are similar to those described in the sections on analog quadrature signal interpolation. The code disk of such a device does not incorporate a series of parallel lines of constant pitch, but rather depends on special code tracks having analog profile signal outputs. These can be effected by changing the code tracks' cross-sections or by graduating their optical density around the revolution. In the latter case, a masking reticle need not be employed.

Incremental Encoders Quadrature Signals

Digital Quadrature Signals

Figure 6.71 illustrates the concept of an incremental rotary optical encoder. The shaft-mounted disk has a series of alternating dark and light sectors of equal length and in equal numbers. The dark and light code is detected by a stationary mask with two apertures, A and B, displaced one quarter of a cycle from each other. When a light sector covers a window, a 1 signal is produced, and a 0 results from a dark sector. At a transition, a rising or falling signal occurs. These signals require some pretreatment to square the signals and avoid problems associated with slow movement at the transition positions, resulting in slow rise time and fall time, in the presence of low noise. The resulting cleaned A and B signals are two square waves 90° out of phase and are called quadrature signals. Since the A and B signals have only four possible states, they clearly do not provide a means of distinguishing more than four different locations. As a result, absolute position discrimination is only possible within one quadrature cycle. Instead, the quadrature signals are used to increment or decrement a counter that gives the actual position. The counter is initialized (usually reset) on a *z*-phase signal produced on a separate track; the innermost track in Figure 6.71. It is also possible to have a number of *z*-phase signals over the encoder range with a distance coded variation between the *z*-phase markers. The code can be used to initialize the counter at nonzero positions.

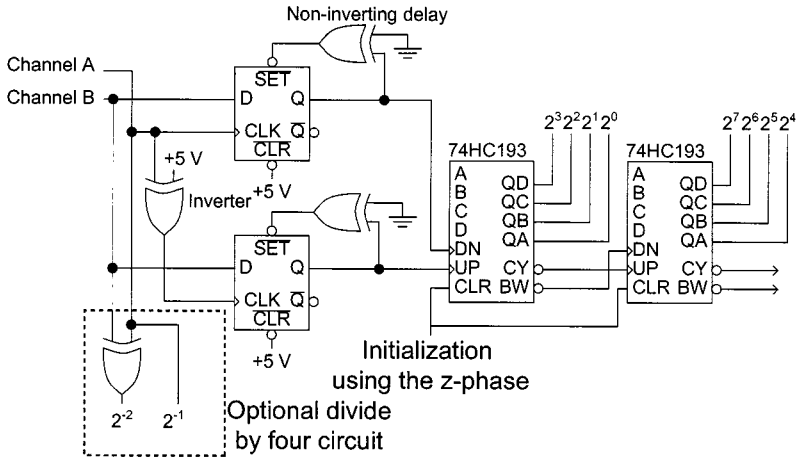
Figure 6.72(*a*) shows a simple circuit from Conner [9] to provide one count per cycle of the quadrature signal. The direction information necessary to decide whether to increment or decrement the counters is also produced. Note that Schmitt triggers provide some amount of hysteresis and redress the falling and rising edges from the encoder. Further circuitry is required to increase the amount of hysteresis in very noisy environments. Additional treatment may be required before feeding these signals to counters. The reason is that if, following a count, there is a small movement in the opposite direction without a reverse count signal being issued followed by a forward motion, then the forward motion produces a second count at the same position. Kuzdrall [10] proposes a more reliable circuit in Figure 6.72(*b*). The



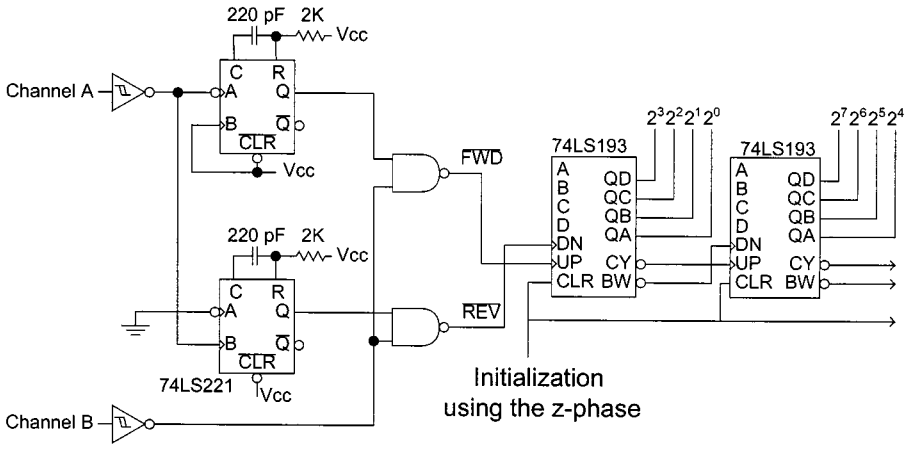
CHANNEL A	CHANNEL B	DIRECTION	COUNT
1	0	0	0
1	1	Qo	0
0	1	1	0
0	0	Qo	1

Qo = Previous state

(a)



(b)



(c)

FIGURE 6.72 Circuit producing a direction and a count signal as described by Conner [9] in (*a*). Caution must be exercised when using this circuit for counting purposes since multiple counts are possible when the encoder oscillates around the A = 0, B = 0 and the A = 0, B = 1 states. Kuzdrall's circuit [10] in (*b*), counts up and down at the same states transition but depending on the direction of movement.

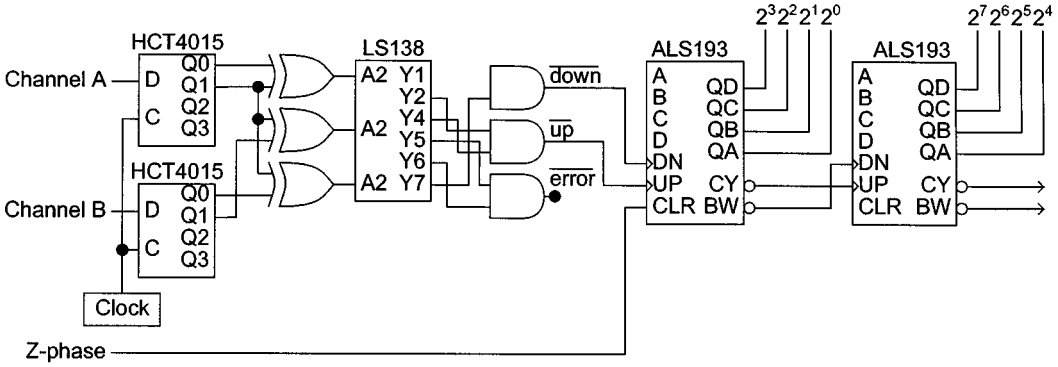


FIGURE 6.73 Divide-by-four circuit producing four counts per cycle of the quadrature signals. Based on Butler's design [14], except for the LS138 demultiplexer that replaces the suggested 4051 multiplexer because of the latter's unequal fall time and rise time. The clocked shift registers generate a 4-bit code to the demultiplexer for each clock cycle. The clock frequency should be 8 times the maximum frequency of the quadrature signals.

circuit drives a 74HC193 counter with a down clock and an up clock. Both flip-flops are normally in the high state. Whenever one of them switches to a low state, it is soon reset by the use of that output as a Set signal. That low output is only possible when phase B is low and there is a transition on A. Depending on the direction of that transition, one of the flip-flops produces the brief low state, causing an appropriate up or down count. The problem associated with small oscillations around a position is reduced since the up and down counts occur at the same encoder position. Venugopal [11] proposes a circuit, in Figure 6.72(*c*), that produces a similar effect. A count can only occur when B is low and there is a transition on A. Depending on the transition direction, one of two monostables triggers and effectively allows the count at the appropriate input of the 74LS193 counter.

In cases where a noisy environment is present, Holle [12] describes a digital filter to clean the A and B signals further, a filter that uses a small number of logic gates to form a 4-bit delay filter. Wigmore [13] proposes other circuits for count and direction detection.

Tables 6.23 and 6.24 list commercial chips and their suppliers' details. Some chips produce count and direction signals or up and down clocking signals. Others also include counters.

The above circuits do not fully exploit the information contained in the quadrature signals since only one of the four edges (or states) within one quadrature cycle is used to count. Figure 6.73 shows a slightly modified version of a divide-by-four counter circuit proposed by Butler [14]. Two 4-bit shift registers, three exclusive-OR gates, and an eight channel demultiplexer derive the up and down count signals. The clocked shift registers generate a 4-bit code to the demultiplexer for each clock cycle. To ensure that no encoder transitions are missed, the clock frequency should be at least 8 *NS*, where *N* is the number of cycles produced by the encoder for each shaft revolutions and *S* is the maximum speed in revolutions per second. The up and down signals can be fed to cascaded 74ALS193 counters. This circuit analyzes the current and previous states of the A and B channels to either count up, count down, or issue an error flag when an improper sequence occurs.

Kuzdrall [10] proposes to view a quadrature signal cycle as a Gray code so that the two least significant bits of the count are obtained by a Gray code to binary code conversion as in Figure 6.72(*b*). Phase A generates bit 1 of the natural binary position, and phase B is exclusive-ORed with phase A to produce the least significant bit. The counter provides the remaining bits of the natural binary position. Marty [15] proposes a state machine that stores both the actual A and B values and the previous values using D-type flip-flops (in a similar way to Butler) to form a hexadecimal number. In all, eight different hex-digits are generated, four for each direction of motion. A 4-line to 16-line decoder then feeds into two 4-input NAND gates to produce an up or down count signal. As for Butler, this last circuit is not dependent on propagation delays as with Kuzdrall's circuit.

Analog Quadrature Signals

Interpolation by resistor network

Some optical encoders deliver analog quadrature signals in the form of a sin(θ) signal, A, and a cos(θ) signal, B, where θ is the phase (electrical degree) within one cycle of the quadrature signal. θ does not equal shaft angle but is related to it: 360 electrical degrees of θ corresponds to 360/N mechanical degrees, where N is the number of analog quadrature signal cycles per shaft revolution. Indeed, θ still exists for linear encoders of this type. Although these signals can be squared and fed to a divide-by-four counter, they can also be processed directly to generate a finer resolution. The main techniques are (1) multiple phase shifted signals, (2) lookup table, and (3) arctangent processor.

The multiple phase-shifted signals method relies solely on electronics to increase the frequency of the final digital quadrature signals by an integer amount. Benzaid et al. [16] propose the circuit of Figure 6.74(*a*), which has been designed in this particular case for a fourfold frequency increase. This can then be followed by a digital divide-by-four circuit (not shown). The A and B signals are combined to produce an additional six phase-shifted signals, three of which are shifted by π/8, π/4, and 3π/8, respectively, from the A signal, and three others that are shifted similarly from the B signal. The result is a total of 8 available sinusoidal signals, phase-shifted by $\alpha = i\,\pi/8$ with $i = 0, 1, \ldots, 7$. This can be generalized, saying that for an m-fold increase in resolution, $2m$ signals that are phase-shifted by $\alpha i\,\pi/2m$ with $i = 0, 1, \ldots, 2m-1$ are required. The phase-shifted sinusoidal signals are then squared using TTL converters. Figure 6.74(*b*) shows the resulting square waves. The addition, modulo-2, of the signals for even values of i gives A′, and similarly for the signals with odd values of i gives B′. The modulo-2 sum is performed via exclusive-ORs.

The vector additions of the initial A and B signals result in phase-shifted signals. The weights of A and B are calculated from the trigonometric relations:

$$\sin(\theta + \alpha) = \cos(\alpha)\,\sin(\theta) + \sin(\alpha)\,\cos(\theta) \tag{6.96}$$

and

$$\cos(\theta + \alpha) = -\sin(\alpha)\,\sin(\theta) + \cos(\alpha)\,\cos(\theta) \tag{6.97}$$

where sin(θ) = A and cos(θ) = B.

Thus,

$$\sin(\theta + \alpha) = \cos(\alpha)\,A + \sin(\alpha)\,B \tag{6.98}$$

and

$$\cos(\theta + \alpha) = -\sin(\alpha)\,A + \cos(\alpha)\,B \tag{6.99}$$

Note that the amplitudes of the phase-shifted signals are not critical, since it is their zero-crossing points that produce the square-wave transition. Also, in the circuit of Figure 6.74(*a*), the weights were slightly modified to simplify the implementation, which results in a small variation of the duty cycle of A′ and B′.

The phase-shifted signals can alternatively be produced using voltage dividers and Schmitt triggers, where the divider resistors are in the ratio tan(α) [17]. As m increases, the precision of the weights become more stringent and the speed of the electronics processing the high-frequency square signals might limit the upper value of m possible with this method.

Interpolation by Sampling and Numerical Processing

A number of methods digitize the analog quadrature signals in order to perform a digital interpolation within one cycle of the quadrature signals. These techniques permit an even higher interpolation. The signals are periodically sampled in sample-and-hold circuitry and digitized by an analog to digital converter (ADC). One technique uses an interpolation table, while the other performs arctangent calculations.

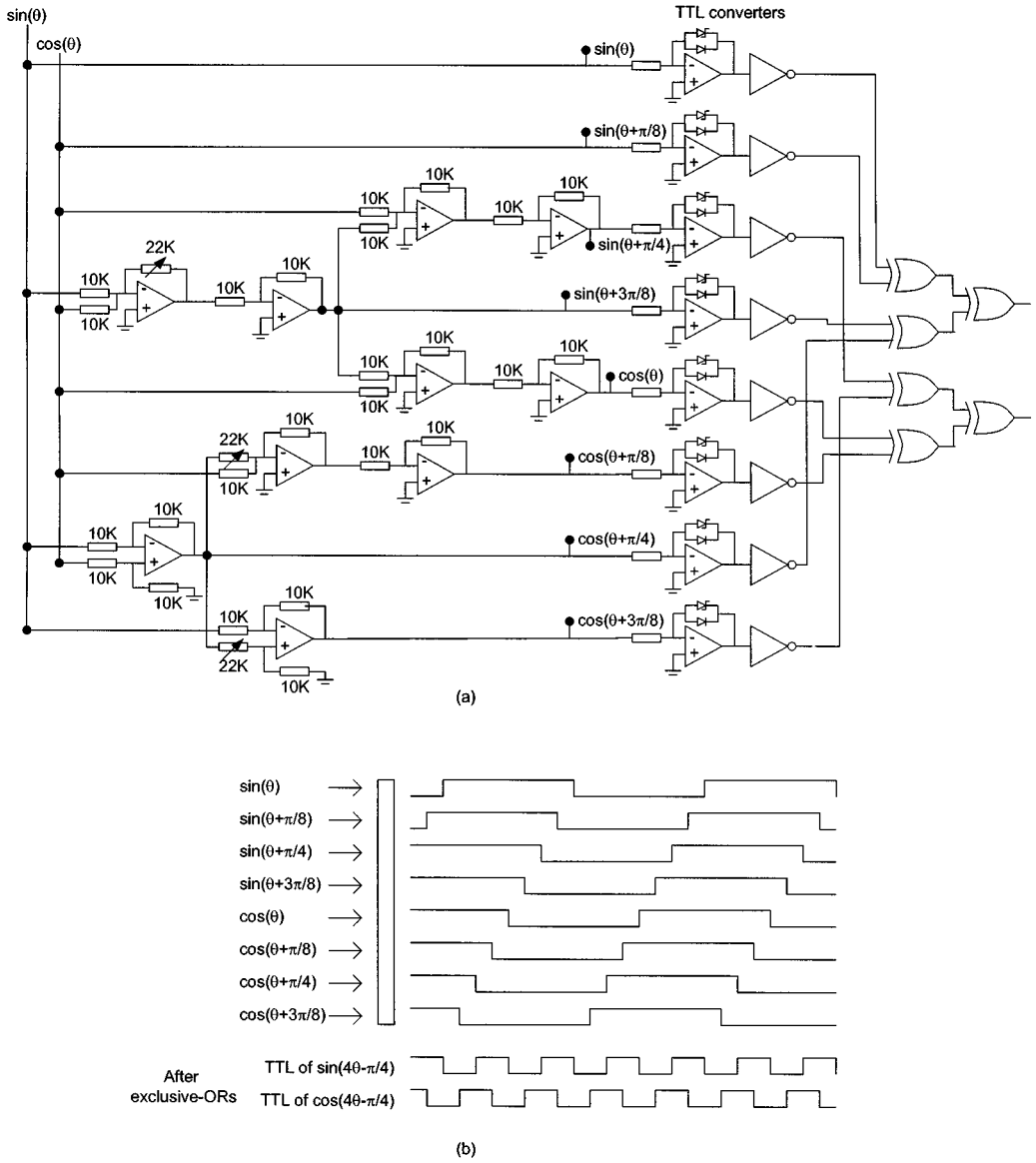


FIGURE 6.74 Interpolation circuit by Benzaid et al. [16]. The sine and cosine signals from the encoder are vector-combined in (*a*) to produce various phase-shifted signals. Following TTL conversion, they are exclusive-ORed. The squared phase-shifted signals and the final result are in (*b*).

Hagiwara [18] uses an m-bit ADC to digitize the analog A and B signals into the binary numbers Da and Db. Figure 6.75(*a*) shows how these binary numbers are used as the addresses of a grid, here built for $m = 3$ as a 2^m by 2^m grid. Hagiwara then associates a phase angle with the center of each cell of the matrix using simple arctangent calculations. The angle is calculated with respect to the center of the grid. This phase angle is then associated with one of $2^n - 1$ phase codes using:

$$\text{Phase code} = \text{integer part of}\left(2^n \theta / 2\pi\right) \tag{6.100}$$

as shown in Figure 6.75(*b*).

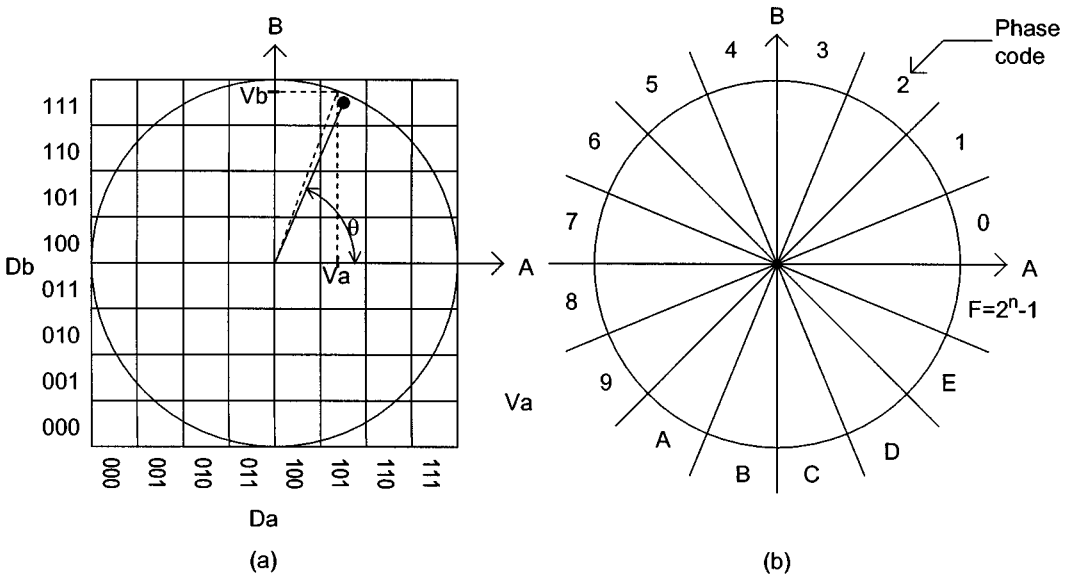


FIGURE 6.75 Hagiwara [18] proposes a two-dimensional look-up table that associates a quantized phase value to a set of digitized values of the quadrature signals. The phase code associated with a grid address may be adjusted to compensate for known errors in the quadrature signals.

The proposed circuit includes the means of performing the accumulation of interpolated values to deliver the encoder absolute position. A modification to this circuit is also proposed that can compensate the phase code for known inaccuracies in the encoder signals with respect to the actual encoder position.

For the highest possible level of interpolation, Mayer [19] describes an arctangent calculator method. It uses a microprocessor to perform high-level trigonometric calculations on the digitized analog signals. As this process is performed, the analog signals are continuously squared and fed to a divide-by-four counter. The two information sources are combined to produce the encoder's position with a very high level of resolution. Note, however, that although the resulting theoretical resolution is limited only by the ADC, in practice it is the quality of the analog signals in relation to the physical position being measured that will limit the precision obtained.

Encoding Principles

Optical encoders use one of three techniques to generate the electrical signals from the relative movement of the grating and the reading heads. They are (1) geometric masking, (2) Moiré effects, and (3) laser interference. Note that absolute encoders mainly use geometric masking. Geometric masking relies on geometric optics theory and considers light as traveling in a straight line. However, as the grating period reduces and the resolution of the encoder increases, Moiré fringe effects are observed and used to produce the signals. Although diffraction effects then become nonnegligible, their influence can be controlled by careful design. Finally, for very high resolution, diffraction effects are directly exploited to perform the measurements.

Geometric Masking

Geometric masking is applied to absolute and incremental encoders. The electromagnetic field associated with the propagation of visible light is characterized by very rapid oscillations (frequencies of the order of 10^{14} s^{-1}). It may therefore be expected that a good first-order approximation to the propagation laws of light is obtained by neglecting the wavelength of light. In this case, diffraction phenomena may be ignored and light may be thought to propagate in a straight line. Geometry can then be used to analyze

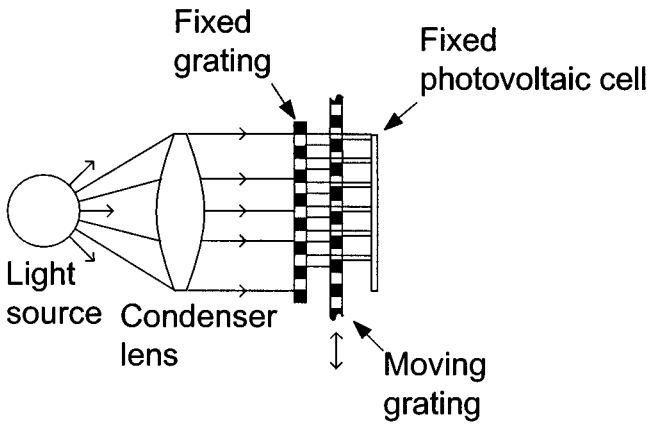


FIGURE 6.76 The reading head contains a light source that can be an incandescent bulb or a light emitting diode. The light may require a condenser in order to collimate it. This redirects the light rays from the source so that they travel perpendicular to the gratings. The stationary index grating structures the light beam into bands that then cross the moving grating. Depending on the relative position of the fixed and the moving gratings, more or less light reaches the photodetector. The detector produces a signal proportional to the total amount of light on its sensitive surface.

the behavior of light. The approximation is valid whenever light rays propagate through an encoder grating with fairly coarse pitch (say, more than 10 μm). Figure 6.76 shows a portion of an encoder scale being scanned by a reading head with multiple slits in order to send more light onto the photosensitive area. The light source is collimated (rays are made parallel to each other after having been emitted in multiple directions from a small source), passes through the stationary grating or index grating, and propagates through the moving grating. In the case of an incremental encoder, a second head would read from the same moving grating but be displaced by $n + 1/4$ pitch in order to produce a second signal with a phase of 90°. When the slits of the two gratings are aligned, a maximum amount of light reaches the detector. Similarly, a minimum amount of light is transmitted when the gratings are out of phase by 180°. Depending on the design of the grating, such as the duty cycle of the transmitting and opaque sectors, various cyclic signal patterns are obtained. Generally, it is noteworthy that optical encoders can also be used as a reflective as opposed to transmissive design. Also, in some absolute optical encoders, the encoding principle does not rely on a primary grating, per se, or the detection system does not necessarily incorporate any masking element.

Moiré Fringes

Moiré fringe methods are primarily associated with incremental encoders. Moiré fringes are observed when light passes through two similar periodic patterns brought close to each other and with their line pattern nearly parallel. Figure 6.77 shows linear and radial gratings producing Moiré fringes. Take a linear grating with sinusoidal amplitude transmittance as proposed by Gasvik [20]:

$$f(x, y) = a + a\cos(2\pi x/P) \tag{6.101}$$

where P is the grating period, a is the amplitude, and x is measured perpendicularly to the grating lines. It is also possible to represent a square wave type grating using a Fourier series in the case of a radial grating and a Fourier integral for a linear grating. When two linear gratings, a and b, are laid in contact, the resulting transmittance, f_c, is the product of their individual transmittances f_a and f_b:

$$f_c = f_a \times f_b \tag{6.102}$$

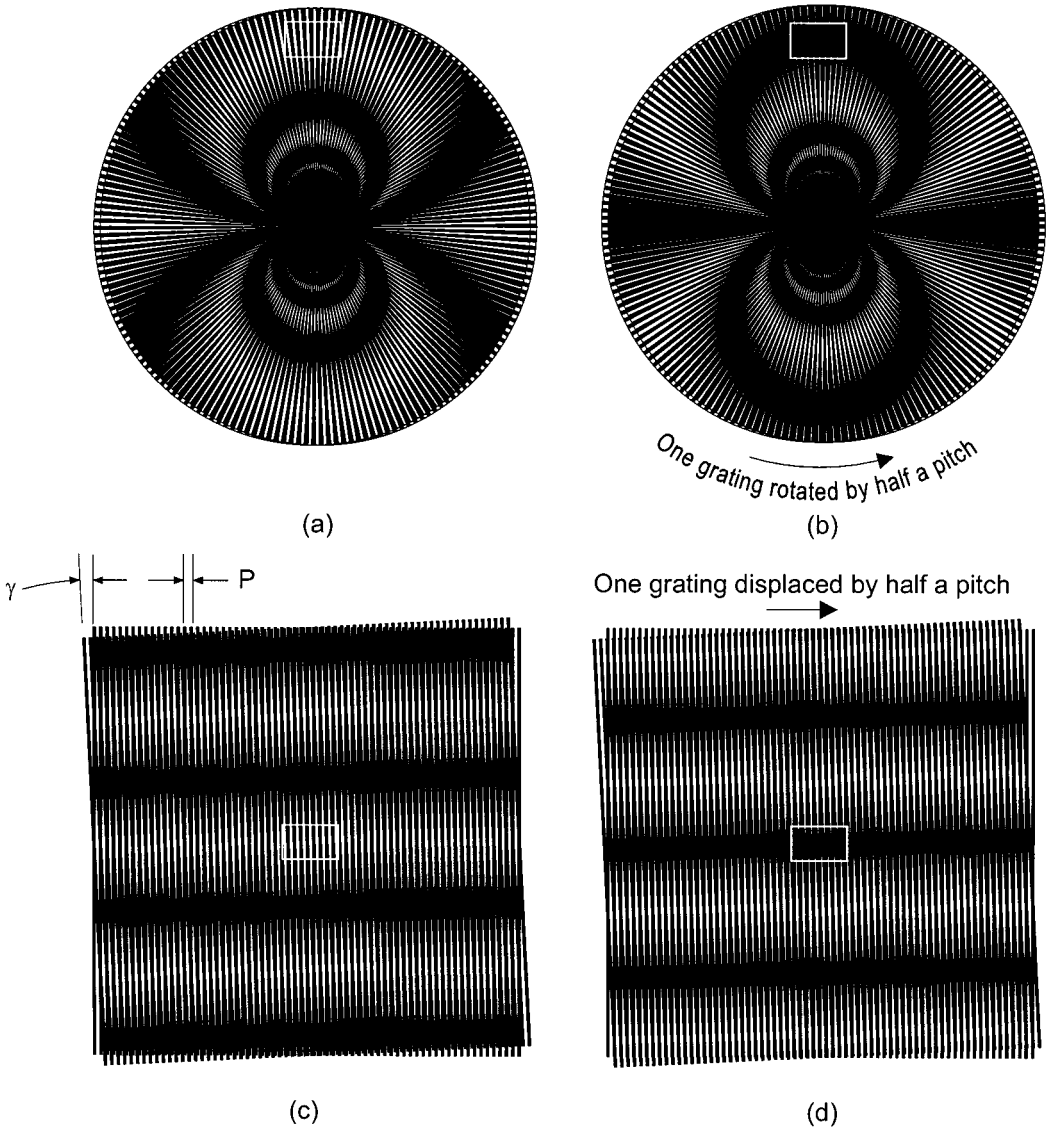


FIGURE 6.77 Moiré fringes produced when two gratings of equal pitch are overimposed. In (*a*), two radial gratings with 180 cycles have their centers displaced horizontally by approximately half of the grating pitch (as measured on the perimeter). This could be due to encoder assembly errors. Ditchburn [21] explains that the fringes are always circular and pass by the two centers. In (*b*), the same displacement was followed by a rotation of half a pitch as a result of the normal movement of the encoder disk. This causes a phase change of 180° of the fringe pattern at any fixed observation point, as represented by the rectangular observation region. The width of the fringes depends on the offset displacement between the two grating centers. If they coincide, the whole pattern changes from a dark to a relatively bright fringe. In (*c*), two linear gratings are rotated by one tenth (in radian) of their grating pitch. This could be due to encoder assembly or mounting errors. As a result, Moiré fringes occur at 90° to the grating and with a period 10 times larger. The normal displacement of one grating by half of the grating pitch causes a phase change of 180° of the Moiré fringe pattern at any fixed observation point, as represented by the rectangular observation region.

In the case when a misalignment γ exists between the two gratings, the separation of the Moiré fringe is d with:

$$d = P \Big/ \left[2 \sin\left(\gamma/2\right)\right] \tag{6.103}$$

and for small γ,

$$d = P/\gamma \tag{6.104}$$

It follows that when the two gratings are moved relative to each other by a distance x' in the direction perpendicular to their lines, then the phase $\varphi(x')$ of the Moiré fringes at a stationary location changes by:

$$\varphi\left(x'\right) = 2\pi x'/P \tag{6.105}$$

In effect, every time x' equals P, the entire Moiré fringe pattern undergoes one complete cycle of fluctuation. If the two gratings are adjusted to near parallelism, then the fringes are "fluffed out" (see Burch [22]) to an infinite width. Under such conditions, there is no fringe motion to be observed and the entire region simply goes from brightness to darkness. These effects are mathematically predicted using a convolution integral.

$$f_c\left(x'\right) = \int_{x_1}^{x_2} f_a\left(x\right) f_b\left(x - x'\right) \Big/ dx \tag{6.106}$$

with x_1 to x_2 representing the region covered by the photodetector. If $g_a(k)$ and $g_b(k)$ are the exponential Fourier transforms of f_a and f_b, then the photodetector response f_c will have the Fourier transform $g_c(k)$,

$$g_c\left(k\right) = g_a\left(k\right) g_b\left(-k\right) \tag{6.107}$$

Moiré effects can be obtained by transmission or by reflection at the grating.

It must be kept in mind, however, that as the grating pitch reduces, diffraction effects become significant; careful design and adjustment of the gap between the stationary and moving grating and also lighting and detection arrangements become critical. Since a large number of lines of the gratings are used to generate the photodetector response, small local imperfections of the grating pitch are averaged out, resulting in an improved measurement accuracy.

Diffraction-Based Encoders

Diffraction-based encoders are used for the highest levels of precision. They successfully exploit diffraction effects and are referred to as physical optics encoders as opposed to geometrical optics encoders. Figure 6.78 shows diffraction effects observed when coherent light encounters a pattern of slits [20]. The slit is a few wavelengths wide, so strong diffraction effects are observed. Each slit may be thought of as acting as a coherent light source producing circular wavefronts. Along certain directions, portions of these wavefronts are in phase with each other. These directions are given by:

$$\sin\theta_m = m\lambda \Big/ P \tag{6.108}$$

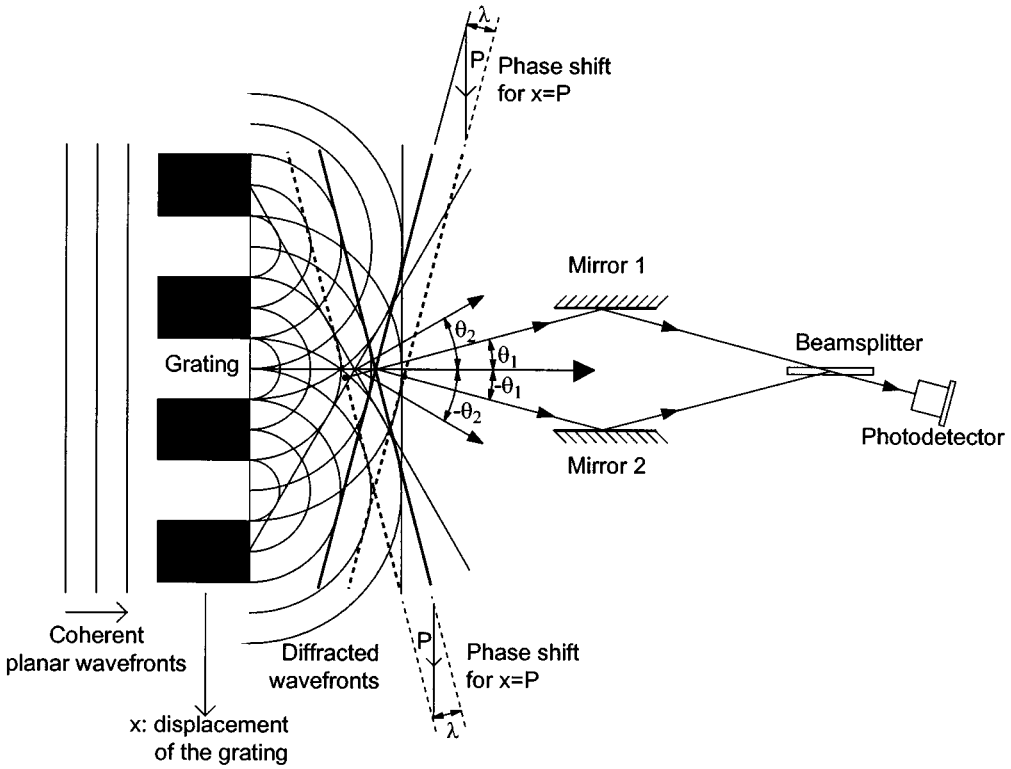


FIGURE 6.78 The grating diffracts the coherent planar wavefront coming from the left into a series of circular wavefronts. The coherent circular wavefronts are in phase along certain directions resulting in diffracted planar wavefronts. These planar wavefronts correspond to the common tangents of the circles. For example, the third innermost circle of the uppermost slit is in phase with the fourth innermost circle of the middle slit and also with the fifth innermost circle of the lowest slit. This diffracted wavefront corresponds to the first order diffraction ($m = 1$) and makes an angle θ_1 with the initial planar wavefront direction. A similar order wavefront ($m = -1$) has a direction $-\theta_1$. These two wavefronts are initially in phase. A displacement of the grating downward by x causes the $m = 1$ wavefront to move by a distance $x\lambda/P$ and the $m = -1$ wavefront by $-x\lambda/P$, thus causing a relative phase shift between the two wavefronts of $2x\lambda/P$. This relative phase shift produces a movement of interference fringes at the photodetector.

where m is a positive or negative integer and is the order of diffraction, and P is the diffraction grating period. Suppose now that the slits are moving by a distance x as shown in Figure 6.78. Then the phase of the wavefront at a stationary location will increase for m positive and decrease for m negative by $2\pi mx/P$. When the slit pattern has moved by one pattern cycle, then the two wavefronts will have developed a relative phase change of $2m$ cycles. The two phase-shifted wavefronts may be recombined to produce interference fringes.

Rotary Encoders

The Canon laser rotary encoder uses an optical configuration that generates four cycles of interference fringes per cycle of the diffraction grating. This is achieved by splitting the original wavefront in two and interrogating the grating at two diametrically opposed locations. Reflecting both diffracted beams back through the grating results in a further doubling of the resolution. The two beams are finally recombined for interference fringe counting. Furthermore, as explained by Nishimura and Ishizuka [23], using diametrically opposed portions of the disk attenuates the effect of eccentricity. An encoder with an external diameter of 36 mm produces 81,000 analog quadrature cycles per revolution. The grating has a pitch of approximately 5 μm and the laser light has a wavelength of 780 nm.

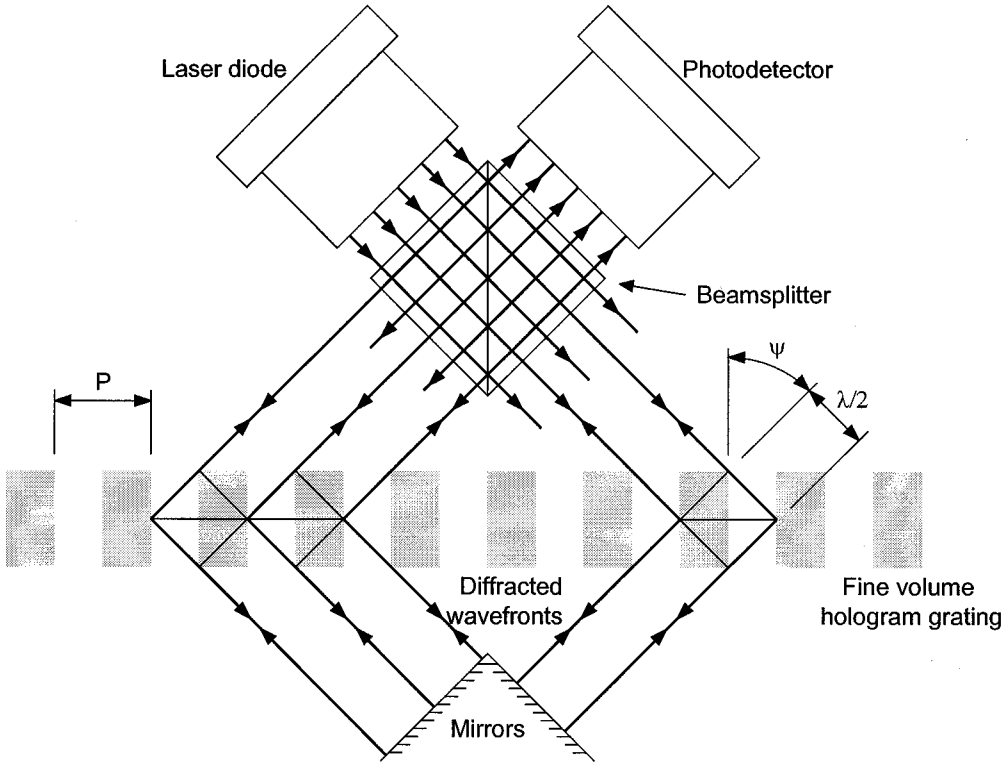


FIGURE 6.79 The coherent planar wavefront of the laser diode is split into two at the beam splitter. Both beams can be regarded as being partially reflected at the grating by minute mirrors separated by a pitch, *P*. Maximum reflection is achieved when $2P \sin \psi = \lambda$. Each reflected portion of the incident beam is phase shifted by one λ with respect to that reflected by an adjacent mirror so that the reflected wavefront remains coherent. The reflected wavefronts are truly reflected at the orthogonal mirrors for a second pass through the grating, after which they are recombined by the splitter to interfere at the photodetector. A displacement of the grating by *x* along its axis causes one beam path to be shortened by $2x\lambda/P$, while the other is lengthened by the same amount. A grating motion by *P* causes four fringe cycles.

Linear Encoders

Sony markets a laser linear encoder with a volume hologram grating of 0.55 μm pitch [24]. Figure 6.79 gives some insight into the principle of operation of this type of encoder. In theory, for maximum intensity of the reflected light beam through the hologram grating, the path length difference from successive plane mirrors must be equal to λ [20]. The Sony Laserscale encoder uses a semiconductor laser as the coherent light source. The beam splitter produces two beams to interrogate the grating at two separate locations. The beams are then diffracted by the hologram grating, followed by a reflection at the mirrors and pass a second time through the grating. They are finally recombined at the photodetector to produce interference fringes. Because the two beams are diffracted by the grating in opposite directions, and because they pass twice through the grating, four signal cycles are obtained when the grating moves by one pitch.

The Heidenhain company produces both linear and rotary encoders of very high resolution, also using the principles of diffraction. Their approach, unlike those previously mentioned, uses an index grating and a reflecting scale grating [17, 25].

The Renishaw company uses a fine reflective grating of 20 μm pitch that diffuses the light from an infrared light-emitting diode. In order to avoid the problems caused by diffraction effects, the index

TABLE 6.19 Commercial Optical Rotary Incremental Encoders

Manufacturer	Model No.	Output type	Counts[a]	Resolution[b]	Price[c]
BEI	H25	Square wave	2540	50,800	$340
Canon	K1	Sine wave	81,000	1,296,000	$2700
Canon	X-1M	Sine wave	225,000	18,000,000	$14,000
Dynamics	25	Sine wave	3000	60,000	$290
Dynamics	35	Sine wave	9000	360,000	$1150
Gurley	911	Square wave	1800	144,000	$1300
Gurley	920	Square wave	4500	144,000	$500
Gurley	835	Square wave	11,250	360,000	$1500
Heidenhain	ROD 426	Square wave	50 to 10,000	200 to 40,000	$370
Heidehain	ROD 905	Analog	36,000	0.035 arcsec	$12,600
Lucas Ledex	LD20	Square wave	100	—	$180
Lucas Ledex	LD20	Square wave	1000	—	$195
Lucas Ledex	DG60L	Square wave	5000	—	$245
Renco	RM21	Square wave	2048	—	$176
TR Electronic	IE58	Square wave	10,000	—	$428

[a] Number of quadrature cycles per revolution without electronic divide-by-four or interpolation.

[b] Unless otherwise specified, is the number of counts per revolution with electronic interpolation, either internal or external to the encoder, supplied by the manufacturer as an option.

[c] Based on orders of one unit. Must not be used to compare products since many other characteristics, not listed in this table, determine the price.

grating of a conventional Moiré approach would require a very small gap between the index and main gratings. Instead, the index grating is located at a distance of 2.5 mm. The index grating is then able to diffract the diffused light from the main grating and to image it 2.5 mm further. There, a fringe pattern of the Moiré type is produced and photoelectrically analyzed [26].

Components and Technology

The choice of an encoder demands careful consideration of a number of factors, such as: (1) the required resolution, repeatability, and accuracy (linearity); (2) the maximum and minimum operating speeds; (3) the environmental conditions: temperature range, relative humidity (condensing or not condensing), contaminants such as dust, water, oil, etc.; (4) minimum friction torque (or force) acceptable; (5) maximum inertia acceptable; (6) available space; (7) position to be known immediately after a power loss; (8) range in degrees (or mm); (9) mounting and shaft loading; and (10) price.

Rotary encoders also sometimes employ reflective tapes attached to, or markings directly etched into, the surface of a drum or spindle to be read from the side. In special cases, tapes or engravings of this type can be made to conform to the surface of an elliptical cross-section cam or other irregular contour. Cylindrical primary gratings employing transmissive readout have also been produced.

The main gratings (scale) come in a wide variety of materials, both for the substrate and for the marking. Flexible scales based on a metal tape substrate are also available from Renishaw, allowing very long (tens of meters) continuous reading of linear motion. Tables 6.19 through 6.21 list a number of encoders currently available on the market. The list is not exhaustive in terms of suppliers and does not cover all the models of the suppliers listed. Table 6.22 gives the address and telephone numbers of the suppliers. Tables 6.23 and 6.24 list some suppliers of quadrature decoding circuits.

TABLE 6.20 Commercial Optical Rotary Absolute Encoders

Manufacturer	Model Number	Steps per turn	No. of turn	Price[a]
BEI	M25	65,536	1	$2130
BEI	MT40	512	16	$1240
BEI	MT40	65,530	512	$5000
Gurley	25/04S	131,072	1	$1900
Heidenhain	ROC 424	4096	4096	
Lucas Ledex	AG60E	360 or 512	1	$486
Lucas Ledex	AG661	4096	4096	$1260
TR Electronic	CE65[b]	8192	4096	$1408

[a] Based on orders of one unit. Must not be used to compare products since many other characteristics, not listed in this table, determine the price.
[b] Programmable output.

TABLE 6.21 Commercial Optical Linear Incremental Encoders

Manufacturer	Model No.	Output type	Pitch[a]	Resolution[b]	Length (mm)	Price[c](length)
Canon	ML-16+	Sine wave	1.6 μm	0.4 μm	To 300	$1525 (50 mm)
Canon	ML-08+	Sine wave	0.8 μm	0.2 μm	To 150	$3100
Gurley	LE18	Square wave	20 μm	0.1 μm	To 1500	$750 (1000 mm)
Gurley	LE25	Square wave	20 μm	0.1 μm	To 3000	$800 (1000 mm)
Heidenhain	LS603	Sine wave	20 μm	5 μm	To 3040	$932 (1020 mm)
Heidenhain	LIP401	Sine wave	2 μm	0.005 μm	To 420	$4000 (100 mm)
Renishaw	RG2	RS422A	20 μm	0.5 μm	To 60,000	$640 + $360/1000 mm
Sony	BS75A-30NS	Square wave	0.14 μm	0.05 μm	30	$2628

[a] Period of the quadrature cycle without electronic divide-by-four or interpolation.
[b] With electronic interpolation supplied by the manufacturer.
[c] Based on orders of one unit. Must not be used to compare products since many other characteristics, not listed in this table, determine the price.

TABLE 6.22 Companies that Make Optical Encoders

BEI Sensors and Motion Systems Company
Encoder Systems Division
13100 Telfair Avenue
Sylmar, CA
Tel: (848) 341-6161

Canon USA Inc.
Components Division
New York Headquarters :
One Canon Plaza
Lake Success, NY 11042
Tel: (516) 488-6700

DR. JOHANNES HEIDDENHAIN GmbH
DR.-Johannes-Heidenhain-Strasse 5
D83301 Traunreut, Deutschland
Tel: (08669)31-0

Gurley Precision Instruments Inc.
514 Fulton Street
Troy, NY 12181-0088
Tel: (518) 272-6300

Ledex Products
Lucas Control Systems Products
801 Scholz Drive
P.O. Box 427
Vandalia, OH 45377-0427
Tel: (513) 454-2345

Renco Encoders Inc.
26 Coromar Drive
Goleta, CA 93117
Tel: (805) 968-1525

Renishaw plc,
Transducer Systems Division
Old Town, Wotton-under-Edge
Gloucestershiire GL12 7DH
United Kingdom
Tel: +44 1453 844302

TR Electronic GmbH
Eglishalde 6
Postfach 1552
D-7218 Trossingen
Germany
Tel: 0 74 25/228-0

Sony Magnescale Inc.
Toyo Building, 9-17
Nishigotanda 3-chome
Shinagawa-ku, Tokyo
141 Japan
Tel: (03)-3490-9481

TABLE 6.23 Commercial Digital Quadrature Signal Decoder Circuits

Manufacturer	Model No.	Output	Decoding factor	Counter	Price
Hewlett Packard	HCTL-2000	Count	12-bit	×4	$12.75
Hewlett Packard	HCTL-2016	Count	16-bit	×4	$12.75
Hewlett Packard	HCTL-2020	Count	16-bit & cascade o/p	×4	$14.55
U.S. Digital Corp.	LS7083	Up and Down clock		×1 or ×4	$3.05
U.S. Digital Corp.	LS7084	Count and direction		×1 or ×4	$3.60

TABLE 6.24 Companies that Make Divide-by-Four Decoders

U.S. Digital Corporation
3800 N.E. 68th Street, Suite A3
Vancouver, WA 98661-1353
Tel: (360) 696-2468

Hewlett-Packard Company
Direct Marketing Organization
5301 Stevens Creek Boulevard
P.O. Box 58059, MS 51LSJ
Santa Clara, CA 95052-8059
Tel: (408) 246-4300

References

1. P. E. Stephens and G. G. Davies, New developments in optical shaft-angle encoder design, *Marconi Rev.,* 46 (228), 26-42, 1983.
2. P. Sente and H. Buyse, From smart sensors to smart actuators: application of digital encoders for position and speed measurements in numerical control systems, *Measurement,* 15(1), 25-32, 1995.
3. E. M. Petriu, Absolute-type position transducers using a pseudorandom encoding, *IEEE Trans. Instrum. Meas.,* IM-36, 950-955, 1987.
4. E. M. Petriu, Scanning method for absolute pseudorandom position encoders, *Electron. Lett.,* 24, 1236-1237, 1988.
5. G. H. Tomlinson, Absolute-type shaft encoder using shift register sequences, *Electron. Lett.,* 23, 398-400, 1987.
6. J. N. Ross and P. A. Taylor, Incremental digital position encoder with error detection and correction, *Electron. Lett.,* 25, 1436-1437, 1989.
7. M. Arsic and D. Denic, New pseudorandom code reading method applied to position encoders, *Electron. Lett.,* 29, 893-894, 1993.
8. B. Arazi, Position recovery using binary sequences, *Electron. Lett.,* 20, 61-62, 1984.
9. D. Conner, Long-lived devices offer high resolution, *EDN,* 35 (9), 57-64, 1990.
10. J. A. Kuzdrall, Build an error-free encoder interface, *Electron. Design,* September 17, 81-86, 1992.
11. P. Venugopal, Reflective optical SMT module reduces encoder size, *Power Conversion and Intelligent Motion,* 21(5), 60-62, 1995.
12. S. Holle, Incremental encoder basics, *Sensors,* 7(4), 22-30, 1990.
13. T. Wigmore, Optical shaft encoder from sharp, *Elektor Electron.,* 15(169), 60-62, 1989.
14. M. M. Butler, Simplified multiplier improves standard shaft encoder, *Electronics,* November 20, 128-129, 1980.
15. B. Marty, Design a robust quadrature encoder, *Electron. Design,* June 24, 71-72, 74, 76, 1993.
16. O. Benzaid and B. M. Bird, Interpolation techniques for incremental encoders, *Proc. 23rd Int. Intelligent Motion Conf.,* Jun 22-24, 165-172, 1993.
17. Heidenhain General Catalog, Dr. Johannes Heidenhain GmbH, DR.-Johannes-Heidenhain-Strasse 5, D83301 Traunreut, Deutschland, November 1993, 8.

18. N. Hagiwara, Y. Suzuki, and H. Murase, A method of improving the resolution and accuracy of rotary encoders using a code compensation technique, *IEEE Trans. Instrum. Meas.*, 41(1), 98-101, 1992.
19. J. R. R. Mayer, High-resolution of rotary encoder analog quadrature signals, *IEEE Trans. Instrum. Meas.,* 43(3), 494-498.
20. K. J. Gasvik, *Optical Metrology*, New York: John Wiley & Sons, 1987.
21. R. W. Ditchburn, *Light Volume 1,* New York: Academic Press, 1976.
22. J. M. Burch, The metrological applications of diffraction gratings, in E. Wolf (Eds.) *Progress in Optics, Volume II*, Amsterdam: North-Holland Publishing, 1963.
23. T. Nishimura and K. Ishizuka (From Canon, Inc., Tokyo), Laser Rotary Encoders, Motion: Official Journal of the Electronic Motion Control Association, September/October 1986, Reprint obtained from Canon.
24. Anonymous from Sony Magnescale America Inc., Hologram technology goes to work, *Machine Design,* January 12 1995.
25. Anonymous from Heidenhain, Encoding systems Vorsprung durch Heidenhain, *Engineering Materials and Design,* September 1989 :53-54.
26. Jim Henshaw of Renishaw, Linear encoder offers superior flexibility, *Design Engineering,* September 1995.

6.9 Magnetic Displacement Sensors

David S. Nyce

Several types of linear and angular displacement measuring devices rely on electromagnetic fields, and the magnetic properties of materials, in the operation of their basic sensing elements. Some may not commonly be referred to as magnetic sensors, but are instead named according to their specific sensing technique. Magnetic sensors presented here use a permanent magnet, or an ac or dc powered electromagnet. Together with various materials used to sense the magnetic field, the combination is arranged to obtain a response indicating angular or linear displacement. The sensor is either caused to operate by a magnetic field, or the properties of the sensor are derived from the use of a magnetic field. Types of magnetic sensors presented in this section include magnetostrictive, magnetoresistive, Hall effect, and magnetic encoders. Some versions of synchro/resolvers and related sensors meet those requirements, but are included in Section 6.10 and thus will not be included in this section. Inductive proximity sensors measure displacement over a very limited range, and are covered in Section 6.2. LVDTs meet these requirements, but are also in Section 6.2.

An important aspect of magnetic sensors is that they utilize a noncontact sensing element. There is no mechanical connection or linkage between the stationary members and the movable members of the sensor. In some devices that sense a position magnet or core, the sensor can even be designed to allow removal of the magnet or core from the sensitive element, when readings are not required. Noncontact implies that the lifetime of the sensing element is not limited to a finite number of cycles by friction-induced wear. This is important in some industrial machinery. Sensors presented here utilize noncontact sensing techniques.

Displacement refers to a change in position, rather than an absolute position. In common industrial practice, however, displacement sensors are typically labeled as either incremental or absolute. An incremental sensor indicates the amount of change between the present location and a previous location. If the information that describes the current location is lost, due to power loss or other disturbance, the system must be reset. During the reset, the sensor must be in a reference position. Magnetic encoders can be designed as either incremental or absolute reading. Optical encoders, inductosyns, and synchro/resolvers are types of displacement sensors that can be designed as either incremental or absolute reading, but are covered in other chapters.

Most displacement position sensors described in this section are absolute reading. They supply a reading of distance or angle from a fixed datum, rather than from a previous position. Consecutive readings can be subtracted to give an incremental indication. An absolute sensor indicates the current position without the need for knowledge of the previous position. It never needs to be reset to a reference location in order derive the measured location. Absolute reading displacement sensors are also commonly called *position sensors*.

Magnetic sensor types will be described here based on the technology employed, rather than the application. Relative usefulness for making linear or angular measurements will be indicated for each type of sensor.

Noncontact magnetic sensor technology for displacement measurement includes magnetostrictive, magnetoresistive, Hall effect, and magnetic encoders.

Magnetic Field Terminology: Defining Terms

Magnetic field intensity (*H*), or magnetizing force: The force that drives the generation of magnetic flux in a material. H is measured in A m^{-1}.

Magnetic flux density (*B*): The amount of magnetic flux resulting from the applied magnetizing force. B is measured in N/(A-m).

Magnetic permeability (μ): The ability of a material to support magnetic lines of flux. The μ of a material is the product of the relative permeability of that material and the permeability of free space. The relative permeability of most nonferrous materials is near unity. In free space, magnetic flux density is related to magnetic field intensity by the formula:

$$B = \mu_0 H$$

where μ_0 is the permeability of free space, having the value $4\pi \times 10^{-7}$ H m^{-1}. In other materials, the magnetic flux density at a point is related to the magnetic intensity at the same point by:

$$B = \mu H$$

where

$$\mu = \mu_0 \mu_r$$

and μ_r is the relative permeability [1].

Hysteresis: A phenomenon in which the state of a system does not reversibly follow changes in an external parameter [2]. In a displacement sensor, it is the difference in output readings obtained at a given point when approaching that point from upscale and downscale readings. Figure 6.80 is a typical output vs. input graph.

Magnetic hysteresis: Depicted in the hysteresis loop, Figure 6.81. When a ferromagnetic material is placed in an alternating magnetic field, the flux density (B) lags behind the magnetizing force (H) that causes it. The area under the hysteresis loop is the hysteresis loss per cycle, and is high for permanent magnets and low for high permeability, low-loss magnetic materials [3].

Magnetic saturation: The upper limit to the ability of ferromagnetic materials to carry flux.

Magnetization curve: Shows the amount of magnetizing force needed for a ferromagnetic material to become saturated. It is a graph with B as the ordinate and H as the abscissa (also known as the B–H curve). A magnetization curve for a specific material would look the same as Figure 6.81, with the addition of calibration marks and the curve adjusted to describe the characteristic of that material.

Magnetostrictive Sensors

A magnetostrictive displacement sensor uses a ferromagnetic element to detect the location of a position magnet that is displaced along its length. The position magnet is attached to a member whose position

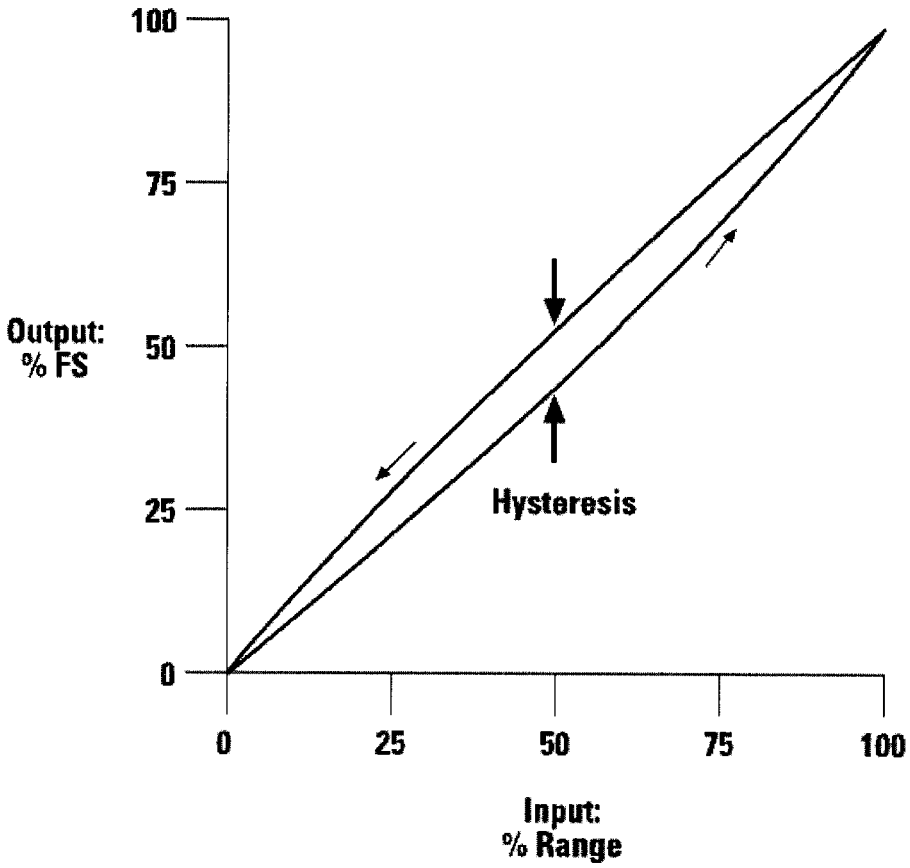


FIGURE 6.80 Hysteresis: output vs. input.

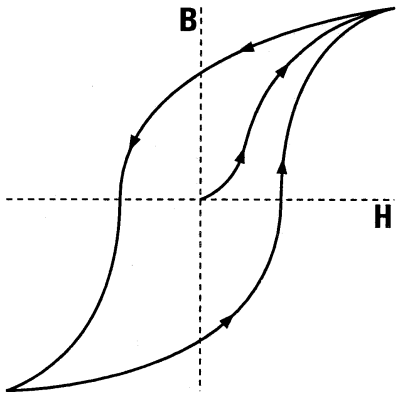


FIGURE 6.81 Magnetic hysteresis.

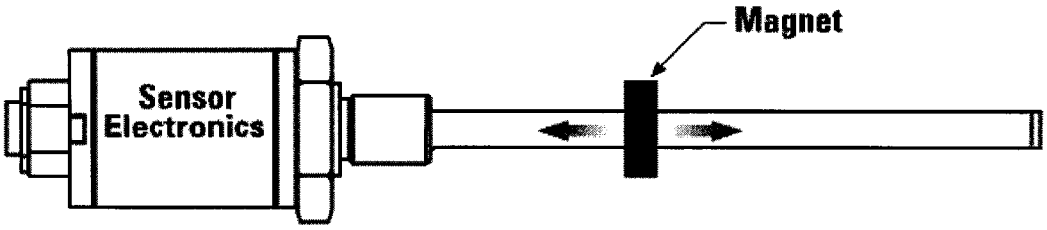


FIGURE 6.82 Magnetostrictive sensor with position magnet.

is to be sensed, and the sensor body remains stationary, see Figure 6.82. The position magnet moves along the measuring area without contacting the sensing element.

Ferromagnetic materials such as iron and nickel display the property called *magnetostriction*. Application of a magnetic field to these materials causes a strain in the crystal structure, resulting in a change in the size and shape of the material. A material exhibiting positive magnetostriction will expand when magnetized. Conversely, with negative magnetostriction, the material contracts when magnetized [4].

The ferromagnetic materials used in magnetostrictive displacement sensors are transition metals, such as iron, nickel, and cobalt. In these metals, the 3*d* electron shell is not completely filled, which allows the formation of a magnetic moment (i.e., the shells closer to the nucleus are complete, and they do not contribute to the magnetic moment). As electron spins are rotated by a magnetic field, coupling between the electron spin and the electron orbit causes electron energies to change. The crystal strains so that electrons at the surface can relax to states of lower energy [5].

This physical response of a ferromagnetic material is due to the presence of magnetic moments, and can be understood by considering the material as a collection of tiny permanent magnets, called *domains*. Each domain consists of many atoms. When a material is not magnetized, the domains are randomly arranged. However, when the material is magnetized, the domains are oriented with their axes approximately parallel to each other. Interaction of an external magnetic field with the domains causes the magnetostrictive effect. See Figure 6.83. This effect can be optimized by controlling the ordering of domains through alloy selection, thermal annealing, cold working, and magnetic field strength.

While application of a magnetic field causes the physical strain, as described above, the reverse is also true: exerting stress causes the magnetic properties (permeability, susceptibility) to change. This is called the *Villari effect*.

In magnetostrictive sensors, uniform distortions of length, as shown in Figure 6.83, offer limited usefulness. Usually, the magnetization is rotated with a small field to induce a local distortion, using the

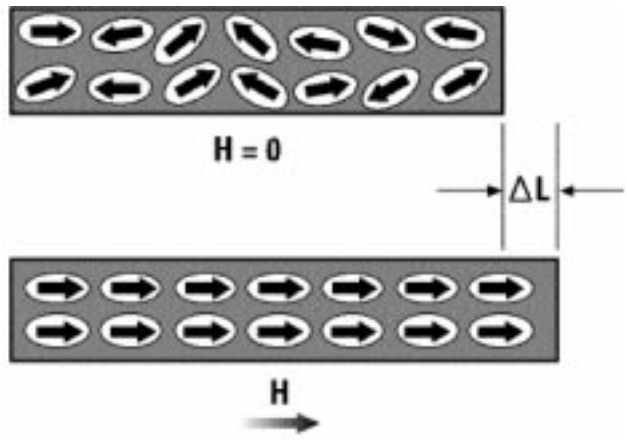


FIGURE 6.83 Magnetic domains: alignment with magnetic field, "*H*", causes dimensional changes.

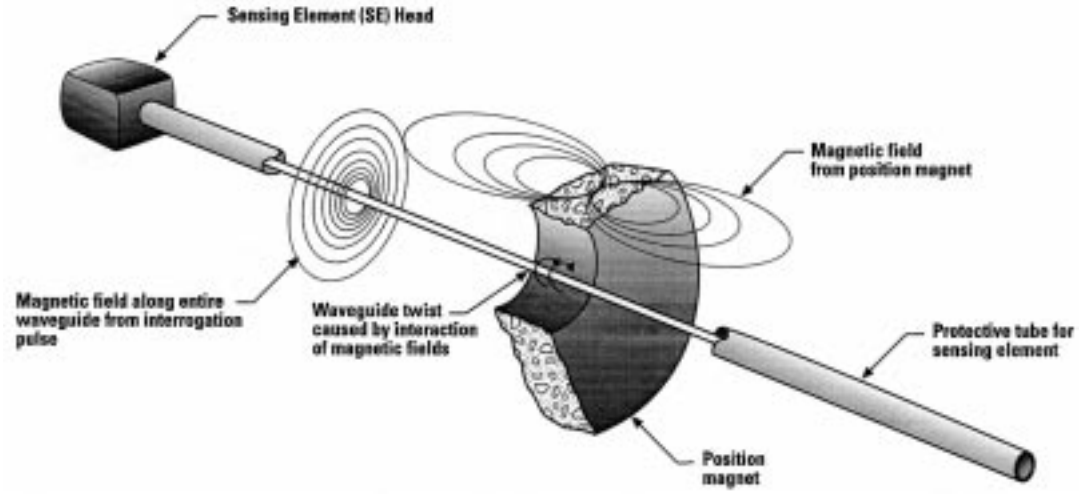


FIGURE 6.84 Operation of magnetostrictive position sensor.

Wiedemann effect. This is a mechanical torsion that occurs at a point along a magnetostrictive wire when an electric current is passed through the wire while it is subjected to an axial magnetic field. The torsion occurs at the location of the axial magnetic field, which is usually provided by a small permanent magnet called the position magnet.

In a displacement sensor, a ferromagnetic wire or tube called the waveguide is used as the sensing element, see Figure 6.84. The sensor measures the distance between the position magnet and the pickup. To start a measurement, a current pulse I (called the interrogation pulse), is applied to the waveguide. This causes a magnetic field to instantly surround it along its full length.

In a magnetostrictive position sensor, the current is a pulse of approximately 1 to 2 μs duration. A torsional mechanical wave is launched at the location of the position magnet due to the Wiedemann effect. Portions of this wave travel both toward and away from the pickup. The wave traveling along the waveguide toward the pickup is detected when it arrives at the pickup. The time measurement between application of the current pulse (launching of the torsion wave at the position magnet) until its detection by the pickup represents the location of the position magnet. The speed of the wave is typically about 3000 m s^{-1}. The portion of the wave traveling away from the pickup could act as an interfering signal after it is reflected from the waveguide tip. So instead, it is damped by a damping element when it reaches the end of the waveguide opposite the pickup. Damping is usually accomplished by attaching one of various configurations of elastomeric materials to the end of the waveguide. The end of the waveguide within the damping element is unusable for position determination, and therefore called the "dead zone."

The time measurement can be buffered and used directly as the sensor output, or it can be conditioned inside the sensor to provide various output types, including analog voltage or current, pulse width modulation, CANbus, SSI, HART, Profibus, etc. Magnetostrictive position sensors can be made as short as 1 cm long or up to more than 30 m long. Resolution of those produced by MTS Systems Corp. is as fine as 1 μm. Temperature coefficients of 2 to 5 ppm $°C^{-1}$ can be achieved. The sensors are inherently stable, since the measurement relies on the physical properties of the waveguide material. Longer sensors become very cost effective because the same electronics package can drive sensors of varying length; only the waveguide and its packaging are increased in length to make the sensor longer.

The magnetostrictive wire can be straight for a linear sensor, or shaped to provide curved or rotary measurements. Curved sensors are often used to measure angular or nonlinear motion in industrial applications, although rotary magnetostrictive sensors are not yet very popular.

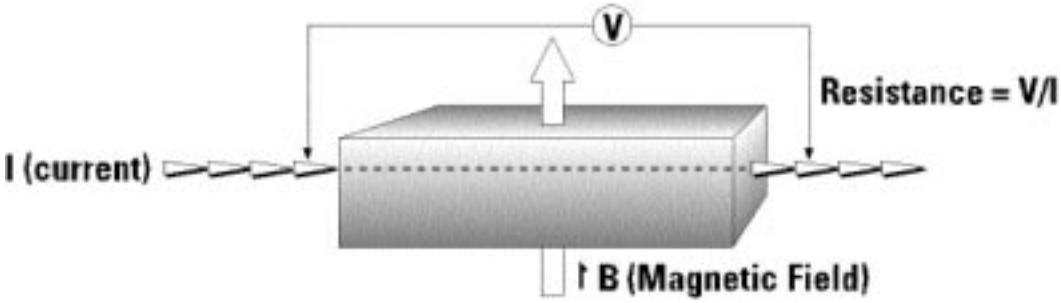


FIGURE 6.85 Magnetoresistance.

Magnetoresistive Sensors

In most magnetic materials, electrical resistance decreases when a magnetic field is applied and the magnetization is *perpendicular* to the current flow (a current will be flowing any time electrical resistance is measured) (see Figure 6.85). The resistance decreases as the magnetic flux density increases, until the material reaches magnetic saturation. The rate of resistance decrease is less as the material nears saturation. The amount of resistance change is on the order of about 1% at room temperature (0.3% in iron, 2% in nickel). When the magnetic field is *parallel* to the current, the resistance increases with increasing magnetic field strength. Sensitivity is greatest when the magnetic field is perpendicular to the current flow. These are properties of the phenomenon called *magnetoresistance* (MR). The MR effect is due to 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, and a reduction in the effective cross-sectional area of the conductor as a result of the carriers being crowded to one side [6].

When a position magnet is brought close to a single MR sensing element, the resistance change is maximum as the magnet passes over the approximate center of the element and then reduces until the magnet is past the element. The resistance changes according to:

$$\text{Resistivity} = \text{Voltage} \Big/ \left(\text{carrier density} \times \text{carrier velocity}\right) \tag{6.111}$$

By using multiple MR elements arranged along a line, a longer displacement measuring device can be fashioned. The signals from the string of sensors are decoded to find which elements are being affected by the magnet. Then the individual readings are used to determine the magnet position more precisely. Relatively high-performance sensors can be manufactured. Temperature sensitivity of the MR elements needs to be compensated, and longer sensors contain many individual sensing elements. Because of this, longer sensors become more difficult to manufacture, and are expensive.

Anisotropic MR materials are capable of resistance changes in the range of 1% or 2%. The MR of a conductor body can be increased by making it a composite of two or more layers of materials having different levels of magnetoresistance. Multilayered structures of exotic materials (sometimes more than 10 layers) have enabled development of materials that exhibit much greater magnetoresistive effect, and saturate at larger applied fields. This has been named Giant MagnetoResistance (GMR). Some commercial sensors based on GMR are currently available. The GMR elements can be arranged in a four-element bridge connection for greater sensitivity. In this arrangement, two of the elements are shielded from the applied magnetic field. The other two elements are sensitive to the applied field. Sensitivity can also be increased by incorporating flux concentrators on the sensitive elements. In a bridge connection, the output voltage can vary by more than 5% of the supply voltage [7]. Rotary sensors can be constructed by attaching a pole piece to a rotating shaft. One or more permanent magnets and the pole piece are arranged to cause the magnetic field around the MR element to change with angular displacement.

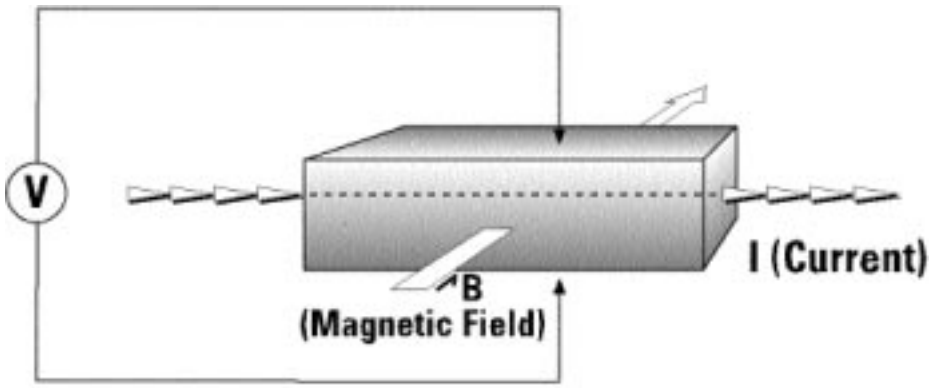


FIGURE 6.86 Hall effect.

Further research is being conducted on MR materials to improve the sensitivity by lowering the strength of magnetic field needed, and increasing the amount of resistance change. The next higher level of MR performance is being called Colossal MagnetoResistance (CMR). CMR is not yet practical for industrial sensors because of severe limitations on the operating temperature range.

Although MR, GMR, and CMR are limited for use in displacement sensors at this time by cost, temperature, and fabrication constraints, much research is in progress. Maybe Humongous Magneto-Resistance (HMR) is next?

Hall Effect Sensors

The Hall effect is a property exhibited in a conductor affected by a magnetic field. A voltage potential V_H, called the Hall voltage, appears across the conductor when a magnetic field is applied at right angles to the current flow. Its direction is perpendicular to both the magnetic field and current. The magnitude of the Hall voltage is proportional to both the magnetic flux density and the current. The magnetic field causes a gradient of carrier concentration across the conductor. The larger number of carriers on one side of the conductor, compared to the other side, causes the voltage potential V_H. A pictorial representation is shown in Figure 6.86. The amplitude of the voltage varies with the current and magnetic field according to: [8]

$$V_H = K_H \beta I / z \tag{6.112}$$

where V_H = Hall voltage
K_H = Hall constant
β = magnetic flux density
I = current flowing through the conductor
z = thickness of the conductor

Sensors utilizing the Hall effect typically are constructed of semiconductor material, giving the advantage of allowing conditioning electronics to be deposited right on the same material. Either *p*- or *n*-type semiconductor material can be used, with the associated polarity of current flow. The greatest output is achieved with a large Hall constant, which requires high carrier mobility. Low resistivity will limit thermal noise voltage, for a more useful signal-to-noise ratio (SNR). These conditions are optimized using an *n*-type semiconductor [6].

A displacement sensor can be made with a Hall sensing element and a movable magnet, with an output proportional to the distance between the two. Two magnets can be arranged with one Hall sensor as in Figure 6.87 to yield a near-zero field intensity when the sensor is equidistant between the magnets. These

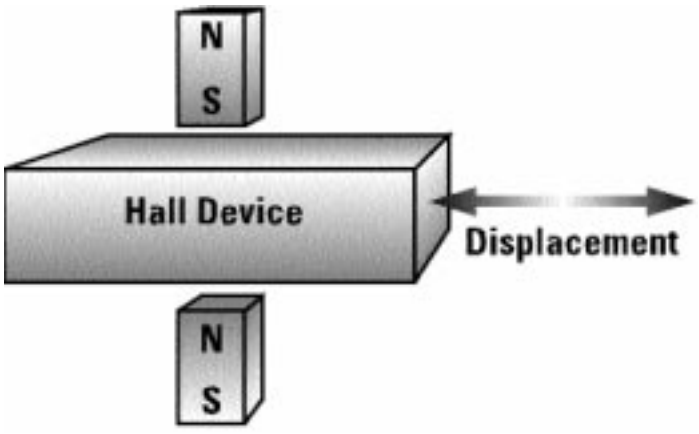


FIGURE 6.87 Two magnet hall sensor.

single Hall effect device configurations have a very limited linear range. Longer range displacement sensors can be built using multiple Hall sensors spaced along a carrier substrate. A magnet is moved along in close proximity to the carrier. As the magnet approaches and then moves away from each Hall element, the respective sensors will have increasing or decreasing outputs. The output from the battery of sensors is derived by reading the individual outputs of the sensors closest to the magnet, and also decoding those particular sensors being read. This method can produce relatively high-performance displacement sensors of up to several meters long. Longer sensors become increasingly more difficult to produce and are expensive because of the large number of sensors being multiplexed. Rotary, as well as linear displacement, sensors can be produced by mechanical arrangement of the sensing elements to cause magnetic field variation with the desired angular or linear input.

Magnetic Encoders

Magnetic encoders use a strip or disk of magnetic media onto which digital information is stored. This information is recorded at the location it describes, and is in the form of a collection of magnetized and nonmagnetized areas. A magnetic encoder includes this sensing element, as well as one or more read heads, electronics, and a mechanical enclosure with input shaft and bushings. The input shaft moves in and out for a linear sensor. It has wipes to prevent ingestion of foreign material, and bushings designed to accept side-loading. An angular sensor has a shaft that rotates, and includes bushings to withstand thrust and side-loading. The encoded media is implemented as either a strip in a linear sensor, or as a disk in an angular sensor.

As a read head passes above the encoded area, it picks up the magnetic variations and reads the position information. The information, digital ones and zeroes, will usually be encoded in several parallel tracks to represent the binary digits of the position information. A standard binary code presents a problem for encoders in that some numbers require the changing of several of the bits at one time to indicate a single increment of the number represented. If all the changing bits are not perfectly aligned with each other, instantaneous erroneous readings will result. To avoid this problem, a special adaptation of the binary code called "Gray code" is used. See Table 6.25. A single increment of the number represented causes a change of only 1 bit in the Gray code.

The read head incorporates a ferromagnetic core wound with input and output windings. A read pulse is applied to the input winding, and information is read on the output winding. If the core is above a magnetized area of the magnetic media, the core becomes saturated, no output pulse is generated, and a logic 0 results [9]. If the core is above a nonmagnetized area when the read pulse is applied, an output pulse occurs and produces a logic 1. Another arrangement that is practical for angular, but not linear encoders, uses a ring-shaped multipole permanent magnet. The magnet is rotated past a pair of sensors

TABLE 6.25 Gray Code

$Base_{10}$ number	"Natural" binary	Gray code	Binary Coded Decimal (BCD)	
			tens	units
0	0000	0000	0000	0000
1	0001	0001	0000	0001
2	0010	0011	0000	0010
3	0011	0010	0000	0011
4	0100	0110	0000	0100
5	0101	0111	0000	0101
6	0110	0101	0000	0110
7	0111	0100	0000	0111
8	1000	1100	0000	1000
9	1001	1101	0000	1001
10	1010	1111	0001	0000

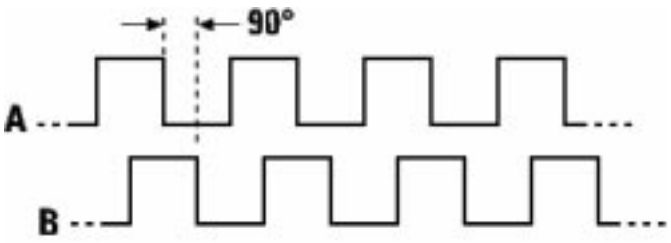


FIGURE 6.88 Quadrature output.

to yield an incremental reading with sine and cosine outputs (called "quadrature" output. See Figure 6.88). The waveforms can be square, sinusoidal, or triangular. A and B outputs are used to indicate the displacement and the direction of the displacement. The number of transitions or "counts" is proportional to the displacement magnitude. The direction of displacement (i.e., + or –) can be found by comparing the two phases. For example, in Figure 6.88, at the time the A phase changes from a logic 0 to a 1, the status of the B phase will indicate the direction of travel. A logic 0 on the B phase could equal the positive direction; a logic 1 could equal the negative direction.

Magnetic encoders can be incremental or absolute. In an incremental configuration, equally spaced pulses encoded on the magnetic media are read from one or more tracks. The pulses are collected by an up/down counter, and the counter output represents the position. Quadrature outputs can be coded to tell the direction of displacement, as described above. The zero position is set by resetting the counter.

Absolute magnetic encoders have the digital code representing the position encoded directly at that position. No counter is needed. The Gray code can be interpreted to yield the position in engineering units. Nonlinear coding, such as sine or cosine, is sometimes used. Table 6.26 provides a list of sensors and manufacturers.

References

1. O. Esbach, *Handbook of Engineering Fundamentals,* New York: John Wiley & Sons, 1975, 957.
2. R. Lerner and G. Trigg, *Encyclopedia of Physics,* New York: VCH Publishers, 1990, 529.
3. P. Neelakanta, *Handbook of Electromagnetic Materials,* Boca Raton, FL: CRC Press, 1995, 333.
4. D. S. Nyce, Magnetostriction-based linear position sensors, *Sensors,* 11(4), 22, 1994.
5. R. Philippe, *Electrical and Magnetic Properties of Materials,* Norwood, MA: Artech House, 1988.
6. H. Burke, *Handbook of Magnetic Phenomena,* New York: Van Nostrand Reinhold, 1986.

TABLE 6.26 Sensors and Manufacturers

Technology	Manufacturers	Description	Price
Magnetostrictive	MTS Systems Corp. Cary, NC & Germany	Lengths to 20 m; 2 μm resolution; CAN, SSI, Profibus, HART	$150–$3000
	Balluff Germany	Lengths to 3.5 m, 20 μm resolution no standard interfaces in head	$400–$2300
Magnetoresistive	Nonvolatile Electronics Eden Prairie, MN	GMR sensors with flux concentrator and shield	$2.50–$6.00
	Midori America Fullerton, CA	Rotary MR sensors Linear MR up to 30 mm	$64–$500 $67–$200
Hall Effect	Optec Technology, Inc. Carrollton, TX	Linear position	$5–$50
	Spectec Emigrant, MT	Standard and custom sensors	Approx. $90
Magnetic encoder	Heidenhain Schaumburg, IL	Rotary and linear encoders	$300–$2000
	Sony Precision Technology America Orange, CA	Rotary and linear encoders	$100–$2000

7. Nonvolatile Electronics Inc. NVSB series datasheet. March 1996.
8. J. R. Carstens, *Electrical Sensors and Transducers,* Englewood Cliffs, NJ: Regents/Prentice-Hall, 1992, p. 125.
9. H. Norton, *Handbook of Transducers,* Englewood Cliffs, NJ: Prentice-Hall, 1989, 106-112.

Further Information

B. D. Cullity, *Introduction to Magnetic Materials,* Reading, MA: Addison-Wesley, 1972.
D. Craik, *Magnetism Principles and Applications,* New York: John Wiley & Sons, 1995.
P. Lorrain and D. Corson, *Electromagnetic Fields and Waves,* San Francisco: W.H. Freeman, 1962.
R. Boll, *Soft Magnetic Materials,* London: Heyden & Son, 1977.
H. Olson, *Dynamical Analogies,* New York: D. Van Nostrand, 1943.
D. Askeland, *The Science and Engineering of Materials,* Boston: PWS-Kent Publishing, 1989.
R. Rose, L. Shepard, and J. Wulff, *The Structure and Properties of Materials,* New York: John Wiley & Sons, 1966.
J. Shackelford, *Introduction to Materials Science for Engineers,* New York: Macmillan, 1985.
D. Jiles, *Introduction to Magnetism and Magnetic Materials,* London: Chapman and Hall, 1991.
F. Mazda, *Electronics Engineer's Reference Book, 6th ed.,* London: Butterworth, 1989.
E. Herceg, *Handbook of Measurement and Control,* New Jersey: Schaevitz Engineering, 1976.

6.10 Synchro/Resolver Displacement Sensors

Robert M. Hyatt, Jr. and David Dayton

Most electromagnetic position transducers are based on transformer technology. Transformers work by exciting the primary winding with a continuously changing voltage and inducing a voltage in the secondary winding by subjecting it to the changing magnetic field set up by the primary. They are ac-only devices, which make all electromagnetically coupled position sensors ac transformer coupled. They are inductive by nature, consisting of wound coils. By varying the amount of coupling from the primary (excited) winding of a transformer to the secondary (coupled) winding with respect to either linear or rotary displacement, an analog signal can be generated that represents the displacement. This coupling variation is accomplished by moving either one of the windings or a core element that provides a flux path between the two windings.

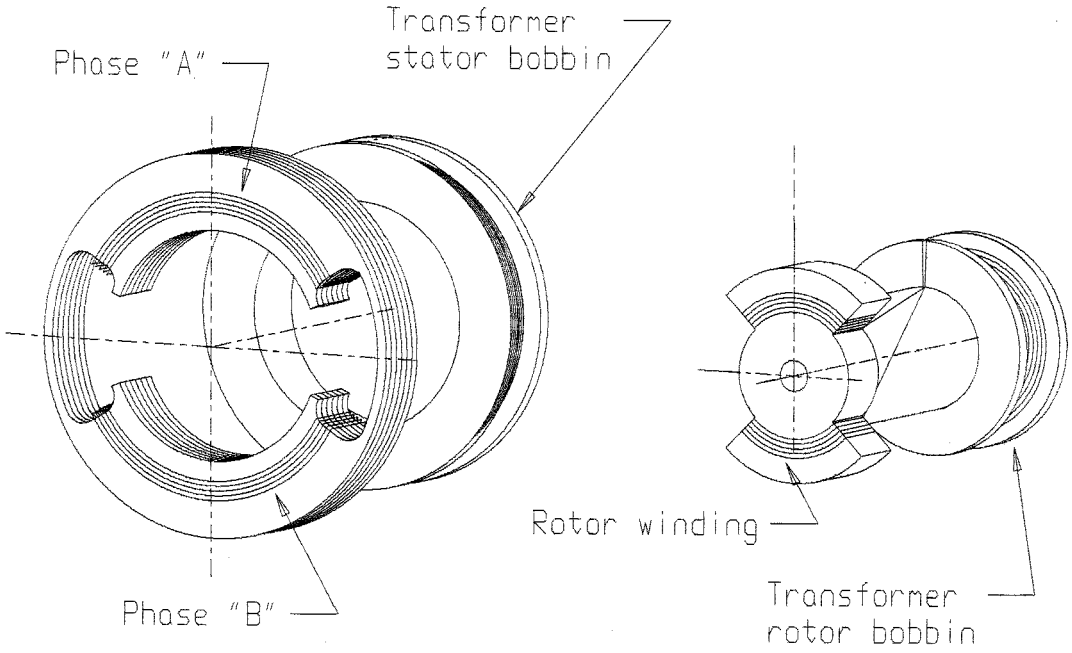


FIGURE 6.89 The induction potentiometer has windings on the rotor and the stator.

One of the simplest forms of electromagnetic position transducers is the LVDT, which is described in Section 6.2 on inductive sensors. If the displacement of the core in an LVDT-type unit is changed from linear to rotary, the device becomes an RVDT (rotary variable differential transformer).

Induction Potentiometers

The component designer can "boost" the output, increase accuracy, and achieve a slightly greater angular range if windings are placed on the rotor as shown in Figure 6.89. The disadvantages to this method are (1) additional windings, (2) more physical space required, (3) greater variation over temperature, and (4) greater phase shift due to the additional windings.

The advantage of the induction pot design is greater sensitivity (more volts per degree), resulting in better signal to noise and higher accuracy in most cases.

Resolvers

If the two-slot lamination in the stator stack shown in Figure 6.89 is replaced by a multislot lamination (see Figure 6.90), and two sets of windings are designed in concentric coil sets and distributed in each quadrant of the laminated stack; a close approximation to a sine wave can be generated on one of the secondary windings and a close approximation to a cosine wave can be generated on the other set of windings. Rotary transformers of this design are called *resolvers*. Using a multislot rotor lamination and distributing the windings in the rotor, the sine-cosine waveforms can be improved even further.

A resolver effectively amplitude modulates the ac excitation signal placed on the rotor windings in proportion to the sine and the cosine of the angle of mechanical rotation. This sine-cosine electrical output information measured across the stator windings may be used for position and velocity data. In this manner, the resolver is an analog trigonometric function generator. Most resolvers have two primary windings that are located at right angles to each other in the stator, and two secondary windings also at right angles to each other, located on the rotor (see Figure 6.91).

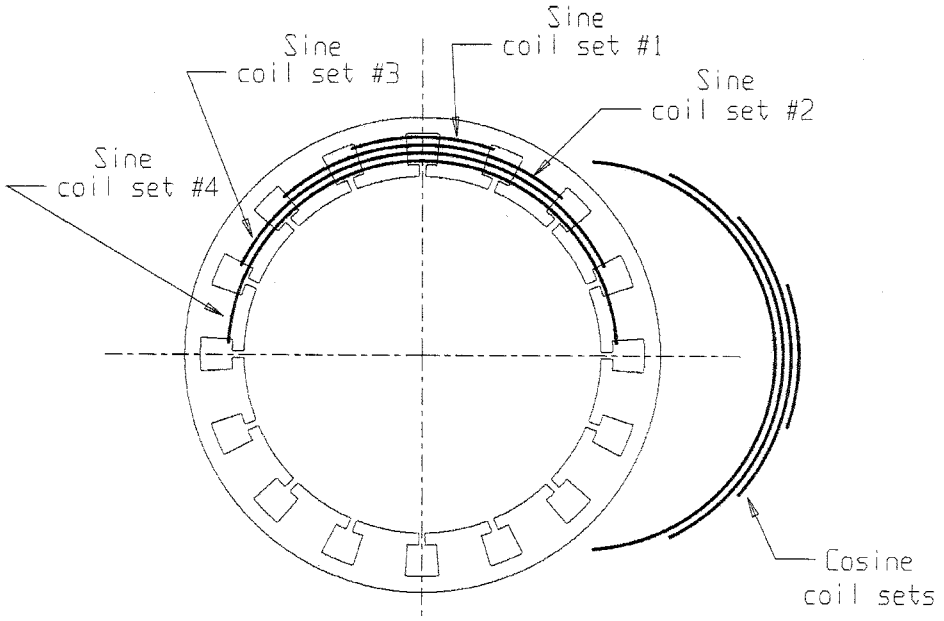


FIGURE 6.90 The resolver stator has distributed coil windings on a 16-slot lamination to generate a sine wave.

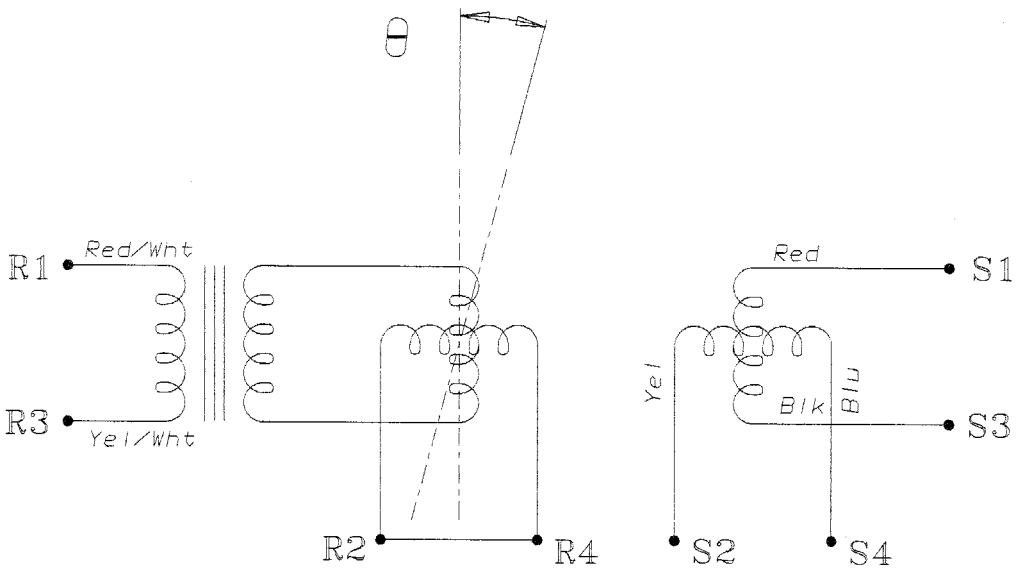


FIGURE 6.91 A brushless resolver modulates the ac excitation on the rotor by the rotation angle.

If the rotor winding (R1-R3) is excited with the rated input voltage (see Figure 6.92), the amplitude of the output winding of the stator (S2-S4) will be proportional to the sine of the rotor angle θ, and the amplitude of the output of the second stator winding (S1-S3) will be proportional to cosine θ. (See Figure 6.93.) This is commonly called the "control transmitter" mode and is used with most "state-of-the-art" resolver to digital converters.

In the control transmitter mode, electrical zero may be defined as the position of the rotor with respect to the stator at which there is minimum voltage across S2-S4 when the rotor winding R1-R3 is excited

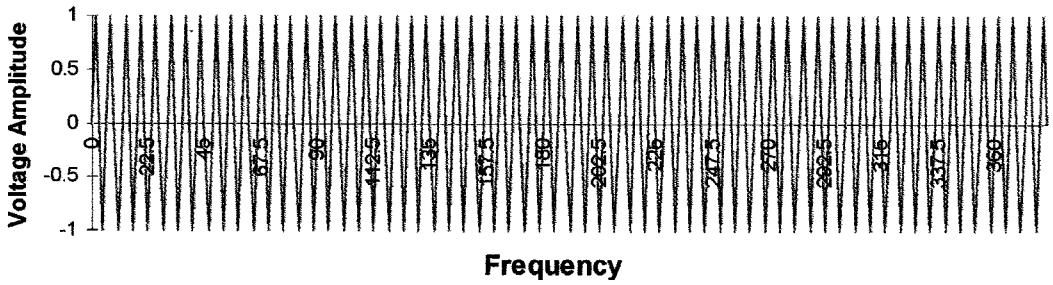


FIGURE 6.92 The resolver rotor winding is excited with the rated input voltage.

with rated voltage. Nulls will occur across S2-S4 at the 0° and 180° positions, and will occur across S1-S3 at the 90° and 270° positions.

If the stator winding S1-S3 is excited with the rated input voltage and stator winding S2-S4 is excited with the rated input voltage electrically shifted by exactly 90°, then the output sensed on the rotor winding R1-R3 does not vary with rotor rotation in amplitude or frequency from the input reference signal. It is the sum of both inputs. It does, however, vary in time phase from the rated input by the angle of the shaft from a referenced "zero" point (see Figure 6.94). This is a "phase analog" output and the device is termed a "control transformer." By measuring the time difference between the zero crossing of the reference voltage waveform and the output voltage waveform, the phase angle (which is the physical angular displacement of the output shaft) can be calculated.

Because the resolver is an analog device and the outputs are continuous through 360°, the theoretical resolution of a resolver is infinite. There are, however, ambiguities in output voltages caused by inherent variations in the transformation of the voltage from primary to secondary through 360° of rotation. These ambiguities result in inaccuracy when determining the true angular position. The types of error signals that are found in resolvers are shown in Figure 6.95.

As a rule, the larger the diameter of the stator laminations, the better the accuracy and the higher the absolute resolution of the device. This is a function of the number of magnetic poles that can be fit into the device, which is a direct function of the number of slots in the stator and rotor laminations. With multispeed units (see the section on multispeeds below), the resolution increases as a multiple of the speeds. For most angular excursions, the multispeed resolver can exceed the positioning accuracy capability of any other component in its size, weight, and price range.

Operating Parameters and Specifications for Resolvers

There are seven functional parameters that define the operation of a resolver in the analog mode. These are (1) accuracy, (2) operating voltage amplitude, (3) operating frequency, (4) phase shift of the output voltage from the referenced input voltage, (5) maximum allowable current draw, (6) the transformation ratio of output voltage over input voltage, and (7) the null voltage. Although impedance controls the functional parameters, it is transparent to the user. The lamination and coil design are usually developed to minimize null voltage and input current, and the impedance is a direct fallout of the inherent design of the resolver. The following procedure can be used to measure the seven values for most resolvers.

Equipment Needed for Testing Resolvers

A mechanical index stand that can position the shaft of the resolver to an angular accuracy that is an order of magnitude greater than the specified accuracy of the resolver.

An ac signal generator capable of up to 24 Vrms at 10 kHz.

A phase angle voltmeter (PAV) capable of measuring "in phase" and "quadrature" voltage components for determining the transformation ratio and the null voltage as well as the phase angle between the output voltage and the reference input voltage.

A 1 Ω resistor used to measure input current with the PAV.

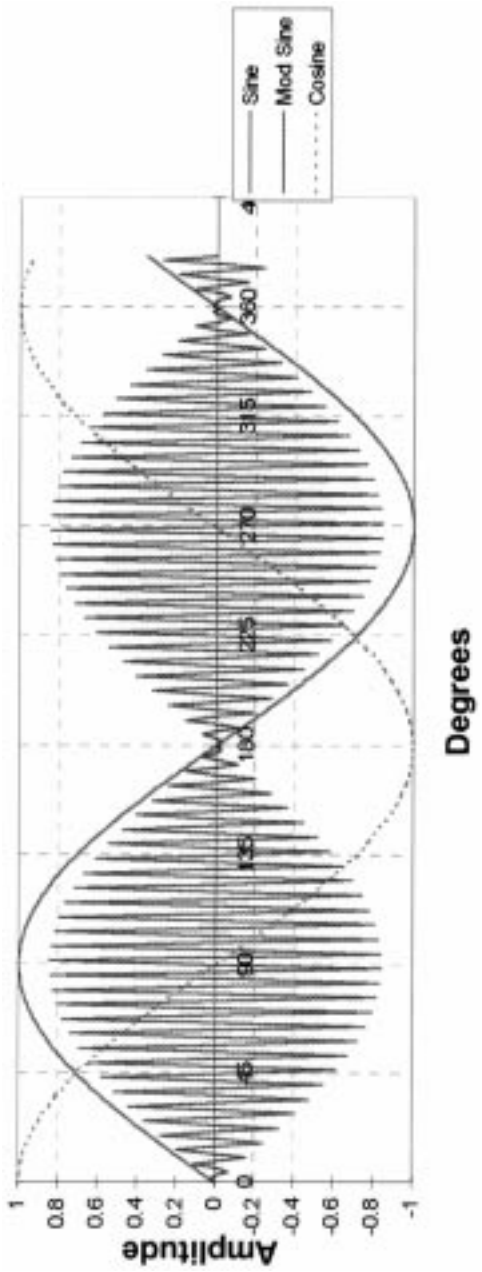


FIGURE 6.93 The single-speed resolver stator output is the sine or cosine of the angle.

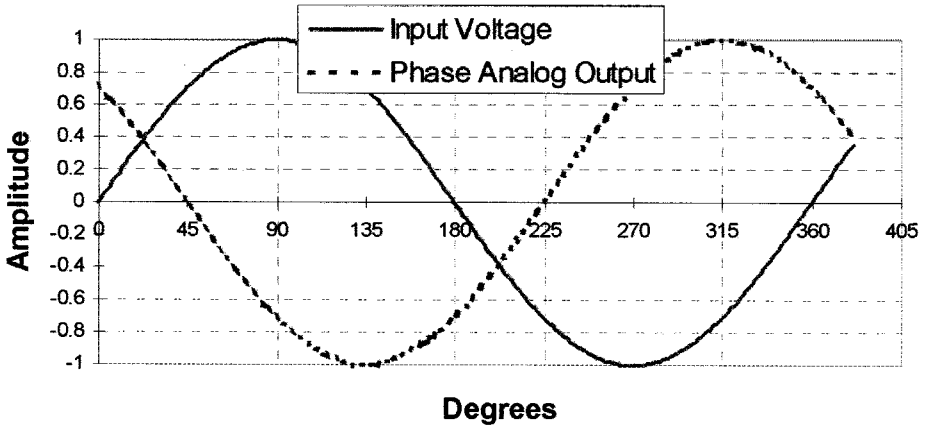


FIGURE 6.94 Resolver stator windings excited at 0° and 90° yield a phase shift with rotor rotation. Phase analog output at 225° vs. input voltage.

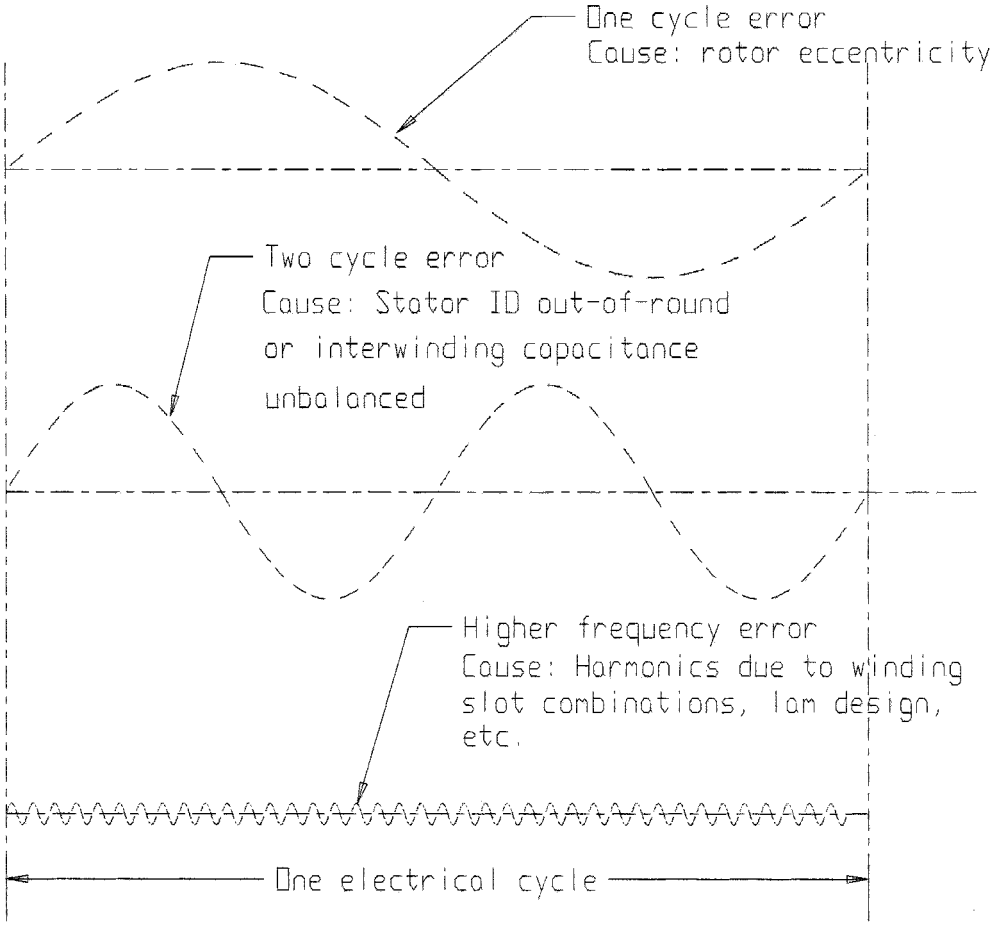


FIGURE 6.95 There are several causes of errors in resolvers and synchros.

An angle position indicator (API) used in conjunction with the index stand to measure the accuracy of the resolver.

1. Using the above equipment with the index stand in the 0° position, mount the resolver on the index stand and lock the resolver shaft in place.
2. Place the 1 Ω resistor in series with the R1 (red/wht) lead of the resolver.
3. Connect the resolver rotor leads, R1 (red/wht) & R3 (yel/wht) (see Figure 6.91) to the sine wave signal generator and set the voltage and frequency to the designed values for the resolver.
4. Connect the terminals from the sine wave signal generator to the reference input of the phase angle voltmeter (PAV).
5. Connect the resolver output leads S2 (yel) and S4 (blu) to the PAV input terminals and place the PAV in the total output voltage mode.
6. Rotate the index stand to the 0° position.
7. Turn the resolver housing while monitoring the total output voltage on the PAV until a minimum voltage reading is obtained on the PAV. Record this value. This is the *null voltage* of the resolver.
8. Turn the index stand to the 90° position.
9. Change the output display on the PAV to show the *phase angle* between the reference voltage and the voltage on leads S2-S4. Record this phase angle reading.
10. Change the output display on the PAV to show the *total output voltage* of the resolver.
11. Record this voltage.
12. Calculate the *transformation ratio* by dividing the reading in step 11 by the input reference voltage amplitude.
13. Connect the input leads of the PAV across the 1 Ω resistor and record the total voltage on the PAV display. Since this reading is across a 1 Ω resistor, it is also the *total current*.
14. Disconnect the resolver and the signal generator from the PAV.
15. Connect the terminals from the sine wave signal generator to the reference input of the angle position indicator (API).
16. Connect the stator leads S1 (red), S3 (blk), S2 (yel), and S4 (blu) to the API S1, S3, S2, and S4 inputs, respectively.
17. Check the *accuracy* of the resolver by recording the display on the API every 20° through 360° of mechanical rotation on the index stand. Record all values.

Resolver Benefits

Designers of motion control systems today have a variety of technologies from which to choose feedback devices. The popularity of brushless dc motors has emphasized the need for rotor position information to commutate windings as a key application in sensor products. Encoders are widely used due to the ease of interface with today's drivers but have some inherent performance limitations in temperature range, shock and vibration handling, and contamination resistance. The resolver is a much more rugged device due to its construction and materials, and can be mated with modular R to D converters that will meet all performance requirements and work reliably in the roughest of industrial and aerospace environments.

The excitation signal E_x into the resolver is converted by transformer coupling into sine and cosine (quadrature) outputs which equal $E_x \sin\theta$ and $E_x \cos\theta$. The resolver-to-digital converter (shown in Figure 6.98) calculates the angle:

$$\Theta = \arctan\left(\frac{E_x \sin\Theta}{E_x \cos\Theta}\right) \tag{6.113}$$

In this ratiometric format, the output of the sine winding is divided by the output of the cosine winding, and any injected noise whose magnitude is approximately equivalent on both windings is canceled. This provides an inherent noise rejection feature that is beneficial to the resolver user. This feature also results in a large degree of temperature compensation.

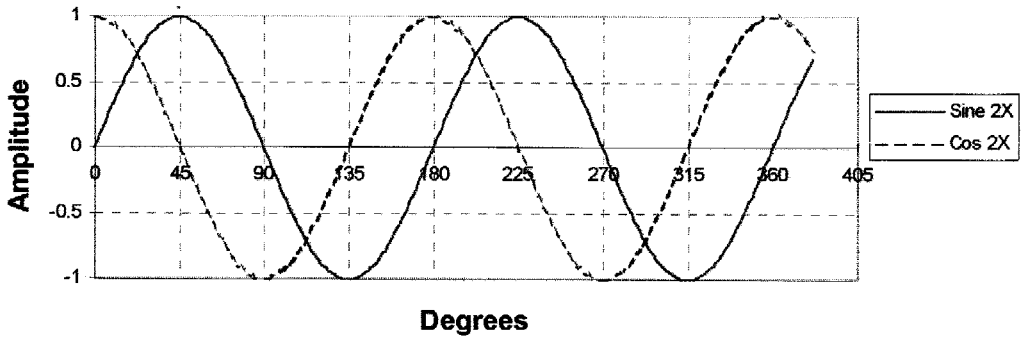


FIGURE 6.96 A two-speed resolver yields two electrical cycles for one rotation.

Multispeed Units

The relationship for multispeeds is that the speed (2×, 3×, etc.) designates how many full sinusoidal cycles the resolver output electrically completes in 360° of mechanical rotation. The 2× electrical output is such that the full sinusoidal cycle for a 2× resolver occurs in 180° instead of 360°. A 2× resolver output is shown in Figure 6.96. A full 3× cycle is completed in 120°. The number of speeds selected for use is a function of the system requirements. Increasing the number of magnetic poles in the rotor and stator creates multispeed units. Each speed has several winding and slot combinations. The optimum combination is selected by the resolver designer based on system demands.

Applications

Resolvers are often used in conjunction with motors, and because of their inherent similarity of design (copper windings on iron lamination stacks), their environmental resistance is quite similar. They are ideal to design into industrial applications where dust and airborne liquids can obscure optical encoder signals. NC machines, coil winders, presses, and positioning tables are uses where resolvers excel. The resolver's inherent resistance to shock and vibration makes it uniquely suited to moving platforms, and their reliability under these conditions lends a welcome hand to the designers of robots, gantries, and automotive transfer lines.

Heat sensitivity is always a problem for motion control systems designers. Resolvers used for sensing the position of valves in high-temperature applications such as aircraft engines, petrochemical refining, and chemical processing have continually proven their reliability.

Moving devices to precise positions with smooth and accurate control can be a real challenge in the electromagnetic noise environment of the modern industrial facility. Emitted and conducted EMI from adjacent equipment, and input voltage variations with unwanted current spikes on input power lines can rob digital systems of their signal integrity. The analog resolver continues to function without information loss or signal interruption. Digitizing the signal can be done at a remote interface under more controlled conditions than on the factory floor. Only robust materials can perform well in harsh environments.

Synchros

As long ago as World War II, synchros were used in analog positioning systems to provide data and to control the physical position of mechanical devices such as radar antennae, indicator needles on instrumentation, and fire control mechanisms in military equipment. The term "synchro" defines an electromagnetic position transducer that has a set of three phase output windings that are electrically and mechanically spaced by 120° instead of the 90° spacing found in a resolver. In the rotor primary mode, the synchro is excited by a single-phase ac signal on the rotor. As the rotor moves 360°, the three amplitude modulated sine waves on the three phases of the output have a discrete set of amplitudes for each angular position. By interpreting these amplitudes, a table can be established to decode the exact rotary position.

In most applications, resolvers have replaced synchros because of the sophistication of the resolver-to-digital converters that are commercially available. Working with a sine and cosine is simpler and requires less conversion and decoding than using three 120° spaced signals. If conversion of a synchro output is desired in resolver format, a device known as a Scott "T" transformer can be used for conversion. In most synchro-to-digital processors, the first step is to convert the signal to a resolver format with a Scott "T" device.

A Modular Solution

The brushless resolver is a self-contained feedback device that, unlike optical encoders, provides an analog signal with infinite resolution. Not only can the output signal be converted to precise digital position information, but it also provides an accurate velocity signal, thus eliminating the need for using a separate tachometer. Reliability is enhanced using the same resolver for speed feedback and commutation. Piece part count can be reduced and the complexity of using Hall-effect devices for timing signals for commutation can be eliminated.

A modular approach allows the designer to easily select a single or multispeed resolver and appropriate electronics that will meet almost any desired level of resolution and accuracy. The resolvers are designed in the most commonly used frame sizes: 8, 11, 15, and 21. Housed models feature high-quality, motor-grade ball bearings. Heavy-duty industrial grade units are enclosed in rugged black painted aluminum housings with either flange, face, or servo-type mounting, and utilize MS-style connectors.

The Sensible Design Alternative for Shaft Angle Encoding

The requirement for velocity and position feedback plays an important role in today's motion control systems. With the development of low-cost monolithic resolver-to-digital converters, a resolver-based system provides design engineers with the building blocks to handle a wide variety of applications. A resolver's small size, rugged design, and the ability to provide a very high degree of accuracy under severe conditions, make this the ideal transducer for absolute position sensing. These devices are also well suited for use in extremely hostile environments such as continuous mechanical shock and vibration, humidity, oil mist, coolants, and solvents. Absolute position sensing vs. incremental position sensing is a necessity when working in an environment where there is the possibility of power loss. Whenever power is supplied to an absolute system, it is capable of reading its position immediately; this eliminates the need for a "go home" or reference starting point.

Resolver-to-Digital Converters

A monolithic resolver-to-digital converter requires only six external passive components to set the bandwidth and maximum tracking rate. The bandwidth controls how quickly the converter will react to a large change in position on the resolver output. The converter can also be programmed to provide either 10, 12, 14, or 16 bits of parallel data. A resolver- based system can provide high dynamic capability and high resolution for today's motion control systems where precision feedback for both position and velocity is required.

Closed Loop Feedback

In a typical closed loop servo model as in Figure 6.97, the position sensor plays an important role by constantly updating the position and velocity information. Selection of a machine control strategy will often be based on performance, total application cost, and technology comfort. The accuracy of the system is determined by the smallest resolution of the position-sensing device. A resolver-to-digital converter in the 16-bit mode has 2^{16} (65,536) counts per revolution, which is equivalent to a resolution of 20 arc seconds. The overall accuracy of the resolver-to-digital converter is ±2.3 arc minutes. An accuracy specification defines the maximum error in achieving a desired position. System accuracy must be smaller than the tolerance on the desired measurement. An important feature of the resolver-to-digital converter

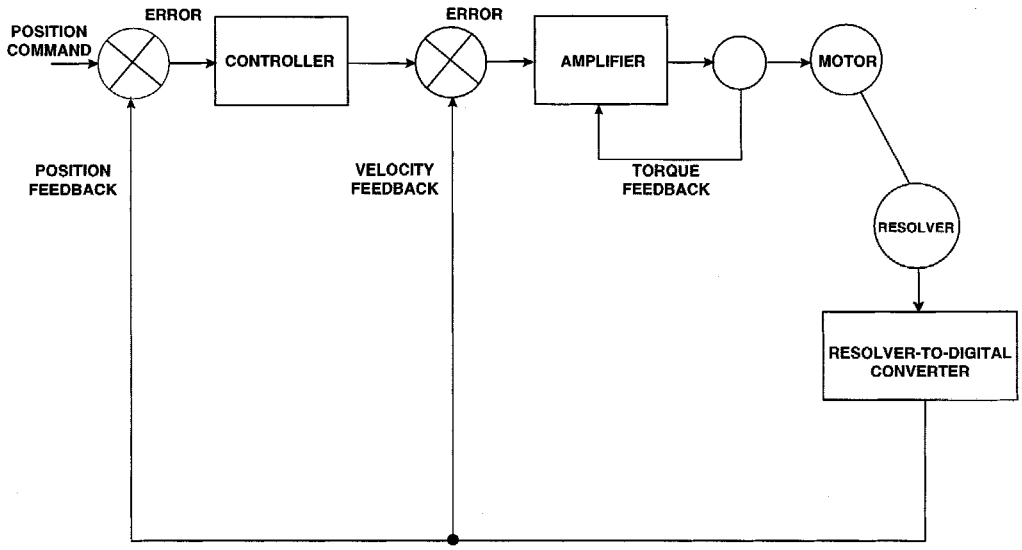


FIGURE 6.97 A closed-loop servo model uses a resolver-to-digital converter.

is repeatability. With a repeatability specification of ±1 LSB (least significant bit) in the 16-bit mode, this provides an accurate measurement when determining position from point to point. For example, moving from point A to point B and back to point A, the converter in the 16-bit mode will be accurate within 20 arc seconds of the original position. The error curve of a resolver-to-digital converter is repeatable within ±1 LSB. The combination of high precision resolvers (±20 arc seconds) with a resolver-to-digital converter provides accurate absolute position information for precision feedback for motion control.

Type II Servo Loop

The motor speed is monitored using the velocity output signal generated by the resolver-to-digital converter. This signal is a dc voltage proportional to the rate of speed, positive for increasing angles and negative for decreasing angles, with a typical linearity specification of 0.25% and a typical reversal error of 0.75%. The error processing is performed using the industry standard technique for type II tracking, resolver-to-digital converters (see Figure 6.98).

The dc error is integrated, yielding a velocity voltage that drives a voltage-controlled oscillator (VCO). This VCO is an incremental integrator (constant voltage input to position rate output) that together with the velocity integrator, forms a type II critically damped, servo feedback loop. This information allows the motor to maintain constant speeds under varying loads when it is interfaced with a programmable logic controller (PLC). The PLC-based architecture is used for I/O intensive control applications. The PLC provides a low-cost option for those developers familiar with its ladder logic programming language. Integration of the motion, I/O, operator's interface, and communication are usually supported through additional cards that are plugged into the backplane.

Applications

Specific applications require unique profiles to control the speed and acceleration of the motor to perform the task at hand. By reducing the accelerations and decelerations that occur during each operation, it is possible to lower the cost and use more efficient motors. Industrial applications include the following:

Ballscrew positioning
Motor commutation
Robotics positioning

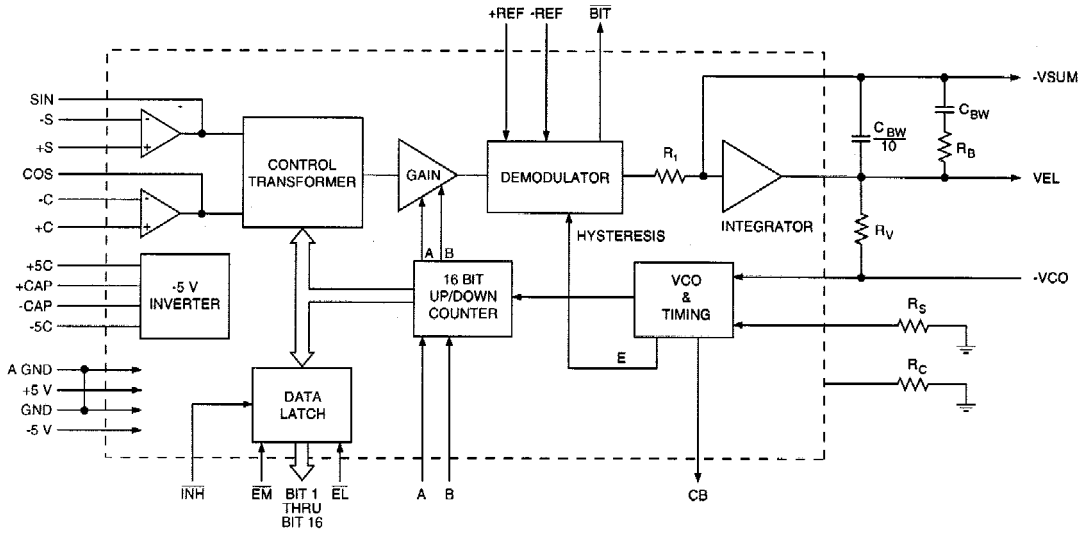


FIGURE 6.98 Error processing uses type II tracking resolver-to-digital converters.

Machine vision systems
X–Y tables
Component insertion
Remote video controls
Web guides
Pick and place machines

Resolver-to-Digital Conversion

For a resolver-to-digital converter, the resolver information is presented to a solid-state resolver conditioner that reduces the signal amplitude to 2 V rms sine and cosine; the amplitude of one being proportional to the sine of θ (the angle to be digitized), and the amplitude of the other being proportional to the cosine of θ. (The amplitudes referred to are, of course, the carrier amplitudes at the reference frequency, i.e., the cosine wave is actually cos θ cos ω*t*; but the carrier term, cos ω*t*, will be ignored in this discussion because it will be removed in the demodulator, and at any rate contains no data). A quadrant selector circuit in the control transformer enables selection of the quadrant in which θ lies, and automatically sets the polarities of the sine θ and cos θ appropriately, for computational significance. The sin θ, cos θ outputs of the quadrant selector are then fed to the sine and cosine multipliers, also contained in the control transformer. These multipliers are digitally programmed resistive networks. The transfer function of each of these networks is determined by a digital input (which switches in proportioned resistors), so that the instantaneous value of the output is the product of the instantaneous value of the analog input and the sine (or cosine) of the digitally encoded angle. If the instantaneous value of the analog input of the sine multiplier is cos θ, and the digitally encoded "word" presented to the sine multiplier is ϕ, then the output code is cos θ sin ϕ. Thus, the two outputs of the multipliers are

$$\text{From the sine multiplier: } \cos\theta \sin\phi$$

$$\text{From the cosine multiplier: } \sin\theta \cos\phi$$

These outputs are fed to an operational subtractor, at the differencing junction shown, so that the input fed to the demodulator is

$$\sin\theta \cos\phi - \cos\theta \sin\phi = \sin(\theta - \phi) \tag{6.114}$$

The right-hand side of this trigonometric identity indicates that the differencing-junction output represents a carrier-frequency sine wave with an amplitude proportional to the sine of the difference between θ (the angle to be digitized) and ϕ (the angle stored in digital form in the up/down counter). This point is the ac error signal brought out as (e). The demodulator is also presented with the reference voltage, which has been isolated from the reference source, and appropriately scaled, by the reference conditioner. The output of the demodulator is then, an analog dc level, proportional to sin $(\theta - \phi)$, in other words, to the sine of the "error" between the actual angular position of the resolver and the digitally encoded angle, ϕ, which is the output of the counter. This point dc error is sometimes brought out as (D) while an addition of a threshold detector will give a built-in-test (BIT) flag. When the ac error signal exceeds 100 LSBs, the BIT flag will indicate a tracking error. This angular error signal is then fed into the error processor and VCO. This circuit consists essentially of an analog integrator whose output (the time integral to the error) controls the frequency of a voltage-controlled oscillator (VCO). The VCO produces clock pulses that are counted by the up/down counter. The "sense" of the error (ϕ too high or ϕ too low) is determined by the polarity of (ϕ), and is used to generate a control counter signal (U), which determines whether the counter increments upward or downward. Finally, note that the up/down counter, like any counter, is functionally an incremental integrator; therefore, the tracking converter constitutes in itself a closed-loop servomechanism (continuously attempting to null the error to zero) with two integrators in series. This called a "Type II" servo loop, which has decided advantages over Type 1 or Type 0 loops. In order to appreciate the value of a Type II servo behavior of this tracking converter, consider first the shaft is not moving. Ignoring inaccuracies, drifts, and the inevitable quantizing error, the error should be zero ($\theta = \phi$), and the digital output represents the true shaft angle of the resolver. Now, start the resolver shaft moving, and allow it to accelerate uniformly, from $d\theta/dt = 0$ to $d\theta/dt = V$. During the acceleration, an error will develop because the converter cannot instantaneously respond to the change of angular velocity. However, since the VCO is controlled by an integrator, the output of which is the integral of the error, the greater the lag (between θ and ϕ), the faster the counter will be called on to catch up. When the velocity becomes constant at V, the VCO will have settled to a rate of counting that exactly corresponds to the rate of change in θ per unit time and instantaneously $\theta = \phi$. Therefore, $d\phi/dt$ will always track $d\theta/dt$ without a velocity or position error. the only error will be momentary (transient) error, during acceleration or deceleration. Furthermore, the information produced by the tracking converter is always "fresh," being continually updated, and always available at the output of the counter. Since $d\theta/dt$ tracks the input velocity it can be brought out as velocity, a dc voltage proportional to the rate of rotation, which is of sufficient linearity in modern converters to eliminate the need for a tachometer in many systems.

Bandwidth Optimization

When using a low-cost monolithic converter for position and velocity feedback, it is important to understand the dynamic response for a changing input. When considering what bandwidth to set the converter, several parameters must be taken into consideration. The ability to track step responses and accelerations will determine what bandwidth to select. The lower the bandwidth of the resolver-to-digital converter, the greater the noise immunity; high frequency noise will be rejected. The relationship between the maximum tracking rate and bandwidth determines the settling time for small and large steps. For a small step input, the bandwidth determines the converter settling time. When one has a large step, the maximum velocity slew rate and bandwidth together, determine the settling time.

Encoder Emulation

Today's resolver-to-digital converters also have the ability to emulate the output of an optical incremental encoder. By providing the outputs A, B, and Zero Index, the encoder can be replaced with a resolver and resolver-to-digital converter without changing the existing interface hardware.

Determining Position Lag Error Due to Acceleration

As the bandwidth and the maximum tracking rate are varied, one can determine the acceleration constant (K_a) and large step settling time.

EXAMPLE:

Resolution:	16 bit
Bandwidth:	100 Hz
Reference:	1000 Hz
Max tracking:	10 rps

$$\mathrm{BW} = \frac{\sqrt{2}A}{\pi} \tag{6.115}$$

If the bandwidth = 100 Hz

A = 222

$K_a = A^2$

K_a = 49,284° s^{-2}

The lag in degrees during an acceleration is:

$$\frac{Acceleration}{K_a} \tag{6.116}$$

EXAMPLE:

Acceleration = 19,000° s^{-2}

K_a = 49,284° s^{-2}

LAG = 19,000/49,284 = 0.38°

In 16 bit 1 LSB = 0.0055°

Acceleration (1 LSB lag) = $K_a \times 0.0055$ (16 bit)

49,284 × 0.0055 = 270° s^{-2}

Large Step Settling Time

To determine the settling time for a large step response (179°), one must take into account the maximum tracking rate and bandwidth.

EXAMPLE:

A 179° step with 100 Hz BW and 10 rps max tracking

Max tracking at 10 rps = 3600° s^{-1}

179/3600 = 49 ms

Then one must add the settling time due to bandwidth limitations; this is approximately 11 time constants (16-bit mode).

Time Constants

Resolution	# of counts/rotation	# of time constants
10	1024	7
12	4096	8
14	16384	10
16	65536	11

Time constant	= 1/A
A	= 222
1/A	= 4.5 ms
11 time constants	= 45 ms

TABLE 6.27 List of Resolver and Synchro Suppliers

Company	Location	Phone number	Types of resolvers
API Harowe	West Chester, PA	(800) 566-5274	Brushless frameless, housed, & heavy-duty units
Admotec, Inc.	Norwich, VT	(802) 649-5800	Rotasyn solid rotor resolvers
Neotech, Inc.	Hatfield, PA	(215) 822-5520	Housed brush & brushless units
Vernitron Corp.	San Diego, CA	(800) 777-3393	Brushless, segment, brushed, & housed units
Servo Systems	Montville, NJ	(973) 335-1007	Brushed & brushless units
American Electronics	Fullerton, CA	(714) 871-3020	Housed units
Computer Conversions	East Northport, NY	(516) 261-3300	Explosion proof & specialty units
Poltron Corp.	Gaithersburg, MD	(301) 208-6597	Resolvers
Tamagawa Trading Co.	Tokyo, Japan	011-81-37-383-175	Brushless frameless & housed units
MPC Products	Skokie, IL	(800) 323-4302	Housed resolvers
Transicoil, Inc.	Trooper, PA	(800) 323-7115	Housed brushless & brushed resolvers
Litton Poly-Scientific	Blacksburg, VA	(800) 336-2112	Brushed and brushless resolvers
Kearfott Guidance & Navigation Corp.	Wayne, NJ	(973) 785-6000	Brushed & brushless resolvers
Novatronics, Inc.	Stratford, Ontario, Canada	(519) 271-3880	Resolvers
Muirhead Vactric	Lake Zurich, IL	(847) 726-0270	Resolvers

Therefore, the approximate settling time for a large step would be 94 ms. This is an approximation. Synchros and resolvers are used in a wide variety of dynamic conditions. Understanding how the converter reacts to these input changes will allow one to optimize the bandwidth and maximum tracking rate for each application. Table 6.27 provides a list of resolver and synchro suppliers.

Further Information

Synchro/Resolver Conversion Handbook, 4th ed., Bohemia NY: ILC Data Device Corp., 1994.
Analog Devices, Inc., *Analog-Digital Conversion Handbook, 3rd ed.,* Englewood Cliffs, NJ: Prentice-Hall, 1986.
Synchro & Resolver Conversion, East Molesey, UK: Memory Devices Ltd., 1980.

6.11 Optical Fiber Displacement Sensors

Richard O. Claus, Vikram Bhatia, and Anbo Wang

The objective of this section is to present a rigorous theoretical and experimental analysis of short gage length optical fiber sensors for measurement of cyclical strain on or in materials. Four different types of sensors are evaluated systematically on the basis of various performance criteria such as strain resolution, dynamic range, cross-sensitivity to other ambient perturbations, simplicity of fabrication, and complexity of demodulation process. The sensing methods that would be investigated include well-established technologies, (e.g., fiber Bragg gratings), and rapidly evolving measurement techniques such as long-period gratings. Other than the grating-based sensors, two popular versions of Fabry–Perot interferometric sensors (intrinsic and extrinsic) will be evaluated for their suitability. A theoretical study of the cross-sensitivities of these sensors to an arbitrary combination of strain vectors and temperature, similar to that proposed by Sirkis in his SPIE paper in 1991 [1], will be developed.

The outline of this section is as follows. The principle of operation and fabrication process of each of the four sensors are discussed separately. Sensitivity to strain and other simultaneous perturbations such as temperature are analyzed. The overall cost and performance of a sensing technique depend heavily on the signal demodulation process. The detection schemes for all four sensors are discussed and compared

on the basis of their complexity. Finally, a theoretical analysis of the cross-sensitivities of the four sensing schemes is presented and their performances are compared.

Strain measurements using optical fiber sensors in both embedded and surface-mounted configurations have been reported by researchers in the past [2]. Fiber optic sensors are small in size, immune to electromagnetic interference, and can be easily integrated with existing optical fiber communication links. Such sensors can typically be easily multiplexed, resulting in distributed networks that can be used for health monitoring of integrated, high-performance materials and structures. Optical fiber sensors for strain measurements should possess certain important characteristics. These sensors should either be insensitive to ambient fluctuations in temperature and pressure, or should have demodulation techniques that compensate for changes in the output signal due to the undesired perturbations. In the embedded configuration, the sensors for axial strain measurements should have minimum cross-sensitivity to other strain states. The sensor signal should itself be simple and easy to demodulate. Nonlinearities in the output demand expensive decoding procedures or require precalibrating the sensor. The sensor should ideally provide an absolute and real-time strain measurement in a form that can be easily processed. For environments where large strain magnitudes are expected, the sensor should have a large dynamic range while at the same time maintaining the desired sensitivity. A discussion of each of the four sensing schemes individually, along with their relative merits and demerits, follows.

Extrinsic Fabry–Perot Interferometric Sensor

The extrinsic Fabry–Perot interferometric (EFPI) sensor, proposed by Murphy et al., is one of the most popular fiber optic sensors used for applications in health monitoring of smart materials and structures [3]. As the name suggests, the EFPI is an interferometric sensor in which the detected intensity is modulated by the parameter under measurement. The simplest configuration of an EFPI is shown in Figure 6.99.

The EFPI system consists of a single-mode laser diode that illuminates a Fabry–Perot cavity through a fused biconical tapered coupler. The cavity is formed between an input single-mode fiber and a reflecting single-mode or multimode fiber. Since the cavity is external to the lead-in/lead-out fiber, the EFPI sensor is independent of transverse strain and small ambient temperature fluctuations. The input fiber and the reflecting fiber are aligned using a hollow-core silica fiber. For uncoated fiber ends, a 4% Fresnel reflection results at both ends. The first reflection, R_1, called the reference reflection, is independent of the applied perturbation. The second reflection, R_2, termed the sensing reflection, is dependent on the length of the cavity, d, which in turn is modulated by the applied perturbation. These two reflections interfere (provided $2d < L_c$, the laser diode's coherence length), and the intensity I at the detector varies as a function of the cavity length:

$$I = I_0 \cos\left(\frac{4\pi}{\lambda} d\right) \tag{6.117}$$

where, I_0 is the maximum value of the output intensity and λ is the laser diode center wavelength.

The typical EFPI transfer function curve is shown in Figure 6.100. Small perturbations that result in operation around the quiescent-point or Q-point of the sensor lead to a linear variation in output intensity. A fringe in the output signal is defined as the change in intensity from a maximum to a maximum or from a minimum to a minimum. Each fringe corresponds to a change in the cavity length by one half of the operating wavelength, λ. The change in the cavity length, Δd, is then employed to calculate the strain ε using the expression:

$$\varepsilon = \frac{\Delta d}{L} \tag{6.118}$$

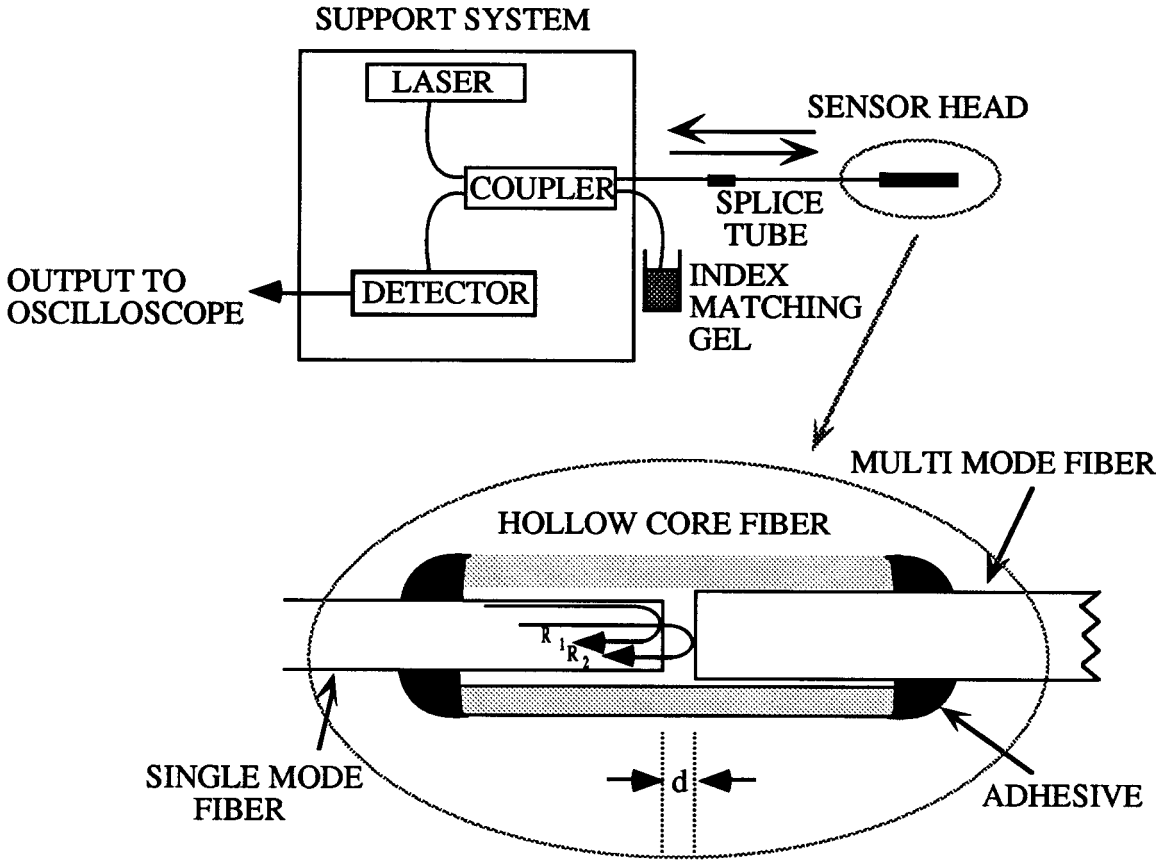


FIGURE 6.99 A simple configuration of an extrinsic Fabry–Perot interferometric (EFPI) sensing system.

where, L is defined as the gage length of the sensor and is typically the distance between two points where the input and reflecting fibers are bonded to the hollow-core fiber. Matching of the two reflection signal amplitudes allows good fringe visibility in the output signal.

The EFPI sensor has been extensively used for measuring fatigue loading on F-15 aircraft wings, detection of crack formation and propagation in civil structures, and cure and lifetime monitoring in concrete and composite specimens [2, 4]. The temperature insensitivity of this sensor makes it attractive for a large number of applications. The EFPI sensor is capable of measuring sub-Angstrom displacements with strain resolution better than 1 microstrain and a dynamic range greater than 10,000 με. Although the change in output intensity of the EFPI is nonlinear corresponding to the magnitude of the parameter being measured, for small perturbations its operation can be limited to that around the Q-point of the transfer function curve. Moreover, the large bandwidth available with this sensor simplifies the measurement of highly cyclical strain. The EFPI sensor is capable of providing single-ended operation and is hence suitable for applications where access to the test area is limited. The sensor requires simple and inexpensive fabrication equipment and an assembly time of less than 10 min. Additionally, since the cavity is external to the fibers, transverse strain components that tend to influence intrinsic sensors through the Poisson's effect have negligible effect on the EFPI sensor output. The sensitivity to only axial strain and insensitivity to input polarization state have made the EFPI sensor the most preferred fiber optic sensor for embedded applications [1]. Thus, overall, the EFPI sensing system is very well suited to measurement of small magnitudes of cyclical strain.

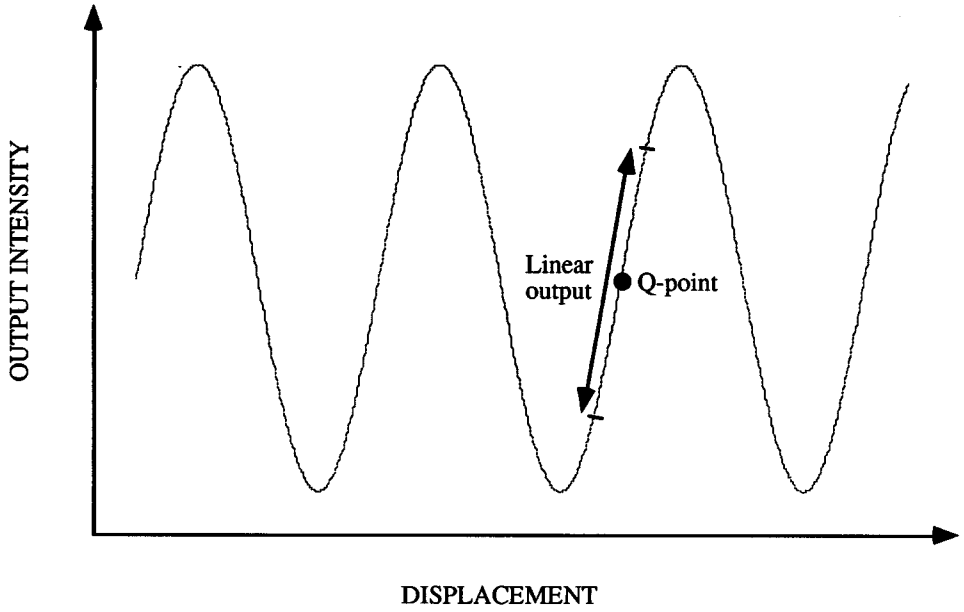


FIGURE 6.100 A typical EFPI transfer function curve.

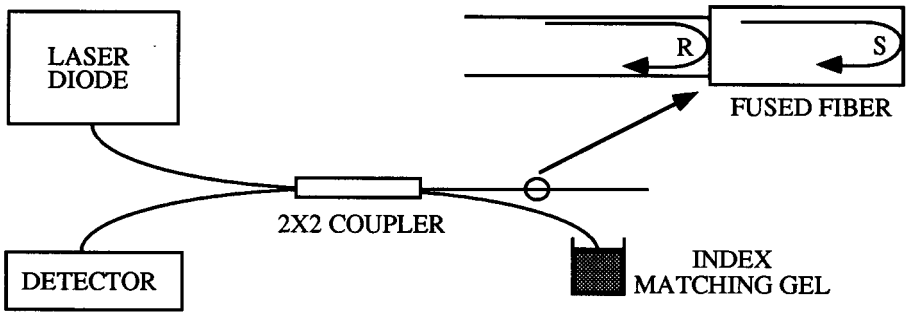


FIGURE 6.101 An intrinsic Fabry-Perot interferometric sensor (IFPI).

Although a version of the EFPI sensor that provides absolute output has been demonstrated, it lacks the bandwidth typically desired during the measurement of cyclical strain [5]. We have also recently proposed a small cavity length/high finesse EFPI sensor for measurement of small perturbations [6]. This configuration has a simple output that can be demodulated using an optical filter/photodetector combination.

Intrinsic Fabry–Perot Interferometric Sensor

The intrinsic Fabry–Perot interferometric (IFPI) sensor is similar in operation to its extrinsic counterpart but significant differences exist in the configurations of the two sensors [7]. The basic IFPI sensor is shown in Figure 6.101. An optically isolated laser diode is used as the optical source to one of the input arms of a bidirectional 2×2 coupler. The Fabry–Perot cavity is formed by fusing a small length of a single-mode fiber to one of the output legs of the coupler. As shown in Figure 6.101, the reference (R) and sensing (S) reflections interfere at the detector face to provide a sinusoidal intensity variation. The cavity can also be obtained by introducing two Fresnel reflectors — discontinuities in refractive index — along the length of a single fiber. Photosensitivity in germanosilicate fibers has been used in the past to fabricate broadband reflectors that enclose an IFPI cavity [8]. Since the cavity is formed within an optical

fiber, changes in the refractive index of the fiber due to the applied perturbation can significantly alter the phase of the sensing signal, *S*. Thus, the intrinsic cavity results in the sensor being sensitive to ambient temperature fluctuations and all states of strain.

The IFPI sensor, like all other interferometric signals, has a nonlinear output that complicates the measurement of large magnitude strain. This can again be overcome by operating the sensor in the linear regime around the Q-point of the sinusoidal transfer function curve. The main limitation of the IFPI strain sensor is that the photoelastic effect-induced change in index of refraction results in a nonlinear relationship between the applied perturbation and the change in cavity length. In fact, for most IFPI sensors, the change in propagation constant of the fundamental mode dominates the change in cavity length. Thus, IFPIs are highly susceptible to temperature changes and transverse strain components [1]. In embedded applications, the sensitivity to all the strain components can result in erroneous outputs. The fabrication process of an IFPI strain sensor is more complicated than that of the EFPI sensor since the sensing cavity must be formed within the optical fiber by some special procedure. The strain resolution of the IFPIs is also expected to be around 1 με with an operating range greater than 10,000 με. IFPI sensors also suffer from drift in the output signal due to variations in the polarization state of the input light.

Thus, the preliminary analysis shows that the extrinsic version of the Fabry–Perot optical fiber sensor seems to have an overall advantage over its intrinsic version. The extrinsic sensor has negligible cross-sensitivity to temperature and transverse strain. Although the strain sensitivity, dynamic range, and bandwidth of the two sensors are comparable, the IFPIs can be expensive and cumbersome to fabricate due to the intrinsic nature of the sensing cavity.

The extrinsic and intrinsic Fabry–Perot interferometric sensors possess nonlinear sinusoidal outputs that complicate the signal processing at the detection end. Although intensity-based sensors have a simple output variation, they suffer from limited sensitivity to strain or other perturbations of interest. Grating-based sensors have recently become popular as transducers that provide wavelength-encoded output signals that can typically be easily demodulated to derive information about the perturbation under investigation. The advantages and drawbacks of Bragg grating sensing technology are discussed first. The basic operating mechanism of the Bragg grating-based strain sensor is elucidated and the expressions for strain resolution is obtained. These sensors are then compared to the recently developed long-period gratings in terms of fabrication process, cross-sensitivity to other parameters, and simplicity of signal demodulation.

Fiber Bragg Grating Sensor

The phenomenon of photosensitivity in optical fibers was discovered by Hill and co-workers in 1978 [9]. It was found that permanent refractive index changes could be induced in fibers by exposing the germanium-doped core to intense light at 488 or 514 nm. The sinusoidal modulation of index of refraction in the core due to the spatial variation in the writing beam gives rise to a refractive index grating that can be used to couple the energy in the fundamental guided mode to various guided and lossy modes. Later Meltz et al. proposed that photosensitivity is more efficient if the fiber is side-exposed to fringe pattern at wavelengths close to the absorption wavelength (242 nm) of the germanium defects in the fiber [10]. The side-writing process simplified the fabrication of Bragg gratings, and these devices have recently emerged as highly versatile components for communication and sensing systems. Recently, loading the fibers with hydrogen has been reported to result in two orders of magnitude higher index change in germanosilicate fibers [11].

Principle of Operation

Bragg gratings are based on the phase-matching condition between spatial modes propagating in optical fibers. This phase-matching condition is given by:

$$k_g + k_c = k_B \tag{6.119}$$

where, k_g, k_c, and k_B are, respectively, the wave-vectors of the coupled guided mode, the resulting coupling mode, and the grating. For a first-order interaction, $k_B = 2\pi/\Lambda$, where Λ is the grating periodicity. Since it is customary to use propagation constants while dealing with optical fiber modes, this condition reduces to the widely used equation for mode coupling due to a periodic perturbation:

$$\Delta\beta = \frac{2\pi}{\Lambda} \tag{6.120}$$

where, $\Delta\beta$ is the difference in the propagation constants of the two modes involved in mode coupling (both assumed to travel in the same direction).

Fiber Bragg gratings (FBGs) involve the coupling of the forward-propagating fundamental LP_{01} optical fiber waveguide propagation mode to the reverse-propagating LP_{01} mode [12]. Consider a single mode fiber with β_{01} and $-\beta_{01}$ as the propagation constant of the forward- and reverse-propagating fundamental LP_{01} modes. To satisfy the phase-matching condition,

$$\Delta\beta = \beta_{01} - \left(-\beta_{01}\right) = \frac{2\pi}{\Lambda} \tag{6.121}$$

where, $\beta_{01} = 2\pi n_{eff}/\lambda$ (n_{eff} is the effective index of the fundamental mode and λ is the free-space wavelength). Equation 6.121 reduces to [12]:

$$\lambda_B = 2\Lambda n_{eff} \tag{6.122}$$

where λ_B is termed the Bragg wavelength. The Bragg wavelength is the wavelength at which the forward-propagating LP_{01} mode couples to the reverse-propagating LP_{01} mode. This coupling is wavelength dependent since the propagation constants of the two modes are a function of the wavelength. Hence, if an FBG is interrogated by a broadband optical source, the wavelength at which phase-matching occurs is found to be reflected back. This wavelength is a function of the grating periodicity (Λ) and the effective index (n_{eff}) of the fundamental mode (Equation 6.122). Since strain and temperature effects can modulate both these parameters, the Bragg wavelength shifts with these external perturbations. This spectral shift is utilized to fabricate FBGs for sensing applications.

Figure 6.102 shows the mode coupling mechanism in fiber Bragg gratings using the β-plot. Since the difference in propagation constants ($\Delta\beta$) between the modes involved in coupling is large, Equation 6.120 reveals that only a small value of periodicity, Λ, is needed to induce this mode coupling. Typically for telecommunication applications, the value of λ_B is around 1.5 μm. From Equation 6.122, Λ is determined to be 0.5 μm (for $n_{eff} = 1.5$). Due to the small periodicities (of the order of 1 μm), FBGs are classified as short-period gratings.

Fabrication Techniques

Fiber Bragg gratings have commonly been manufactured using two side-exposure techniques: the interferometric method and the phase mask method. The interferometric method, depicted in Figure 6.103

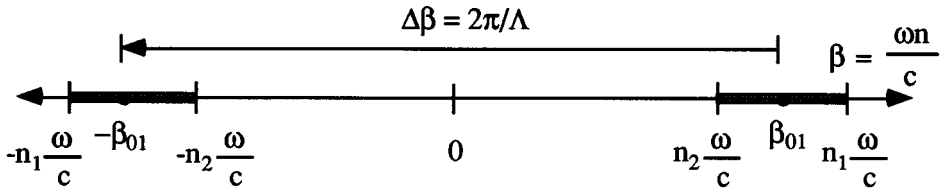


FIGURE 6.102 Mode coupling mechanism in a fiber Bragg grating.

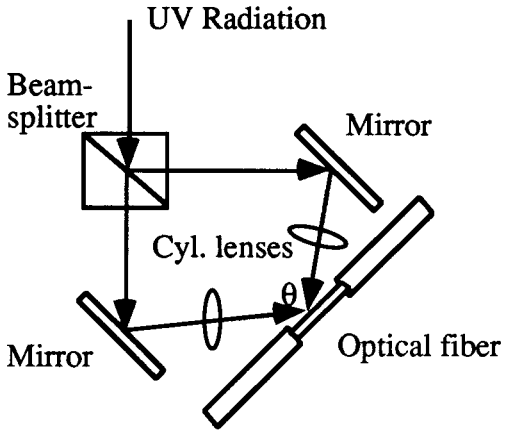


FIGURE 6.103 The interferometric fiber Bragg grating.

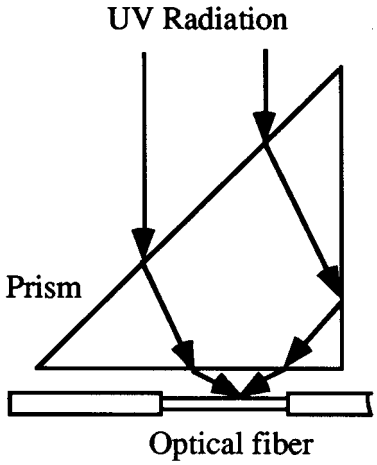


FIGURE 6.104 The novel interferometer technique.

comprises a UV beam at 244 or 248 nm spilt in two equal parts by a beam splitter [10]. The two beams are then focused on a portion of Ge-doped fiber (whose protective coating has been removed) using cylindrical lenses, and the periodicity of the resulting interference pattern and, hence, the Bragg wavelength are varied by altering the mutual angle, θ. The limitation of this method is that any relative vibration of the pairs of mirrors and lenses can lead to the degradation of the quality of the final grating and, hence, the entire system has a stringent stability requirement. To overcome this drawback, Kashyap et al. have proposed the novel interferometer technique where the path difference between the interfering UV beams is produced by the propagation through a right-angled prism (Figure 6.104) [12]. This technique is inherently stable because both beams are perturbed similarly by any prism vibration.

The phase mask technique has recently gained popularity as an efficient holographic side-writing procedure for grating fabrication [13]. In this method, as shown in Figure 6.105, an incident UV beam is diffracted into –1, 0, and +1 orders by a relief grating generated on a silica plate by e-beam exposure and plasma etching. The two first diffraction orders undergo total internal reflection at the glass/air interface of a rectangular prism and interfere on the bare fiber surface placed directly behind the mask. This technique is wavelength specific since the periodicity of the resulting two-beam interference pattern is uniquely determined by the diffraction angle of –1 and +1 orders and, thus, the properties of the phase

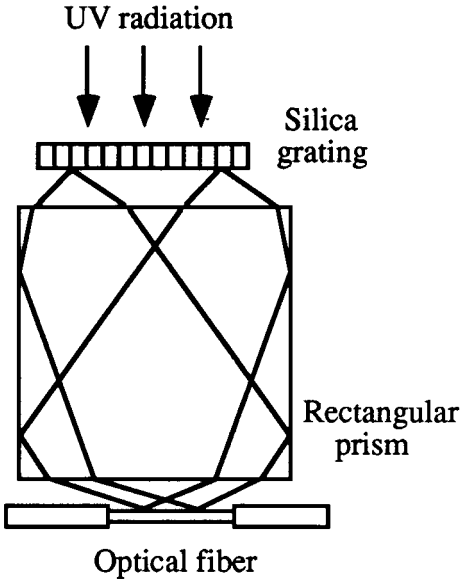


FIGURE 6.105 The phase mask technique for grating fabrication.

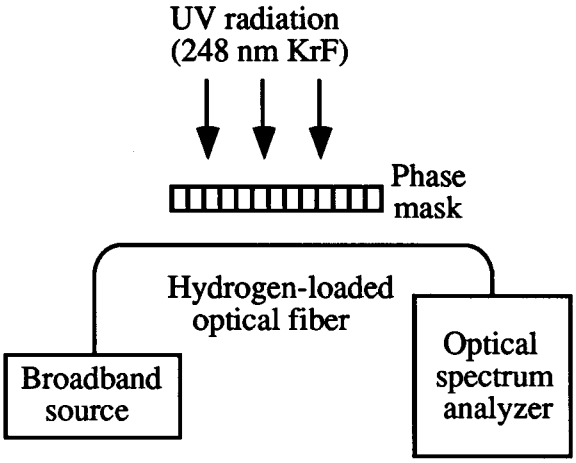


FIGURE 6.106 Diagram of the setup for monitoring the growth of the grating in transmission during fabrication.

mask. Obviously, different phase masks are required for fabrication of gratings at different Bragg wavelengths. The setup for actively monitoring the growth of the grating in transmission during fabrication is shown in Figure 6.106.

Bragg Grating Sensors

From Equation 6.122 we see that a change in the value of n_{eff} and/or Λ can cause the Bragg wavelength, λ, to shift. This fractional change in the resonance wavelength, $\Delta\lambda/\lambda$, is given by the expression:

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta\Lambda}{\Lambda} + \frac{\Delta n_{\text{eff}}}{n_{\text{eff}}} \tag{6.123}$$

where, $\Delta\Lambda/\Lambda$ and $\Delta n_{\text{eff}}/n_{\text{eff}}$ are the fractional changes in the periodicity and the effective index, respectively. The relative magnitudes of the two changes depend on the type of perturbation the grating is subjected to; for most applications, the effect due to change in effective index is the dominating mechanism.

Any axial strain, ε, applied to the grating changes the periodicity and the effective index and results in a shift in the Bragg wavelength, given by:

$$\frac{1}{\lambda}\frac{\Delta\lambda}{\varepsilon} = \frac{1}{\Lambda}\frac{\Delta\Lambda}{\varepsilon} + \frac{1}{n_{\text{eff}}}\frac{\Delta n_{\text{eff}}}{\varepsilon} \tag{6.124}$$

The first term on the right-hand side is unity, while the second term has its origin in the photoelastic effect. An axial strain on the fiber serves to change the refractive index of both the core and the cladding. This results in the variation in the value of the effective index of glass. The photoelastic or strain-optic coefficient that relates the change in index of refraction due to mechanical displacement is about –0.27. Thus, the variation in n_{eff} and Λ due to strain have contrasting effects on the Bragg peak. The fractional change in the Bragg wavelength due to axial strain is 0.73 ε or 73% of the applied strain. At 1550 and 1300 nm, the shifts in the resonance wavelength are 11 nm/%ε and 9 nm/%ε, respectively. With temperature, a FBG at 1500 nm shifts by 1.6 nm for every 100°C rise in temperature [9].

Limitations of Bragg Grating Strain Sensors

The major limitation of Bragg grating sensors is the complex and expensive fabrication technique. Although side-writing is commonly used to manufacture these gratings, the requirement of expensive phase masks increases the cost of the sensing system. In the interferometric technique, stability of the setup is a critical factor in obtaining high-quality gratings. Since index changes of the order of 10^{-3} are required to fabricate these gratings, laser pulses of high energy levels are necessary. This might reduce laser operating lifetime and lead to increased maintenance expense. Additionally, introducing hydrogen or deuterium into the fiber allows increased index modulation as a result of the irridiation process.

The second major limitation of Bragg gratings is their limited bandwidth. The typical value of the full-width at half maximum (FWHM) is between 0.1 and 1 nm. Although higher bandwidths potentially can be obtained by chirping the index or periodicity along the grating length, this adds to the cost of the grating fabrication. The limited bandwidth requires high-resolution spectrum analyzers to monitor the grating spectrum. Kersey et al. have proposed an unbalanced Mach–Zender interferometer to detect the perturbation-induced wavelength shift [14]. Two unequal arms of the Mach–Zender interferometer are excited by the backreflection from a Bragg grating sensor element. Any change in the input optical wavelength modulates the phase difference between the two arms and results in a time-varying sinusoidal intensity at the output. This interference signal can be related to the shift in the Bragg peak and, hence, the magnitude of the perturbation can be obtained. Recently, modal interferometers have also been proposed to demodulate the output of a Bragg grating sensor [15]. The unbalanced interferometers are also susceptible to external perturbations and hence need to be isolated from the parameter under investigation. Moreover, the nonlinear output might require fringe counting equipment, which can be complex and expensive. Additionally, a change in the perturbation polarity at the maxima or minima of the transfer function curve will not be detected by this demodulation scheme. To overcome this limitation, two unbalanced interferometers can be employed for dynamic measurements.

The cross-sensitivity to temperature fluctuations leads to erroneous strain measurements in applications where the ambient temperature has a temporal variation. Thus, a reference grating that measures the temperature change must be utilized to compensate for the output of the strain sensor. Recently, temperature-independent sensing has been demonstrated using chirped gratings written in tapered optical fibers [16].

Last, the sensitivity of fiber Bragg grating strain sensors might not be adequate for certain applications. This sensitivity of the sensor depends on the minimum detectable wavelength shift at the detection end. Although excellent wavelength resolution can be obtained with unbalanced interferometric detection

techniques, standard spectrum analyzers typically provide a resolution of 0.1 nm. At 1300 nm, this minimum detectable change in wavelength corresponds to a strain resolution of 111 με. Hence, in applications where strain smaller than 100 με is anticipated, Bragg grating sensors might not be practical. The dynamic range of strain measurement can be as much as 15,000 με.

Long-Period Grating Sensor

This section discusses the use of novel long-period gratings as strain sensing devices. The principle of operation of these gratings, their fabrication process, preliminary strain tests, demodulation process, and cross-sensitivity to ambient temperature are analyzed.

Principle of Operation

Long-period gratings that couple the fundamental guided mode to different guided modes have been demonstrated in the past [17, 18]. Gratings with longer periodicities that involve coupling of a guided mode to forward-propagating cladding modes were recently proposed by Vengsarkar et al. [19, 20]. As stated previously, fiber gratings satisfy the Bragg phase-matching condition between the guided and cladding or radiation modes or, another guided mode. This wavelength-dependent phase-matching condition is given by:

$$\beta_{01} - \beta = \Delta\beta = \frac{2\pi}{\Lambda} \tag{6.125}$$

where Λ is the periodicity of the grating, β_{01} and β are the propagation constants of the fundamental guided mode and the mode to which coupling occurs, respectively.

For conventional fiber Bragg gratings, the coupling of the forward propagating LP_{01} mode occurs to the reverse propagating LP_{01} mode ($\beta = -\beta_{01}$). Since $\Delta\beta$ is large in this case (Figure 6.107(*a*)), the grating periodicity is small, typically of the order of 1 μm. Unblazed long-period gratings having index variations parallel to the long axis of the fiber couple the fundamental mode to the discrete and circularly-symmetric, forward-propagating cladding modes ($\beta = \beta^n$), resulting in smaller values of $\Delta\beta$ (Figure 6.107(*b*)) and

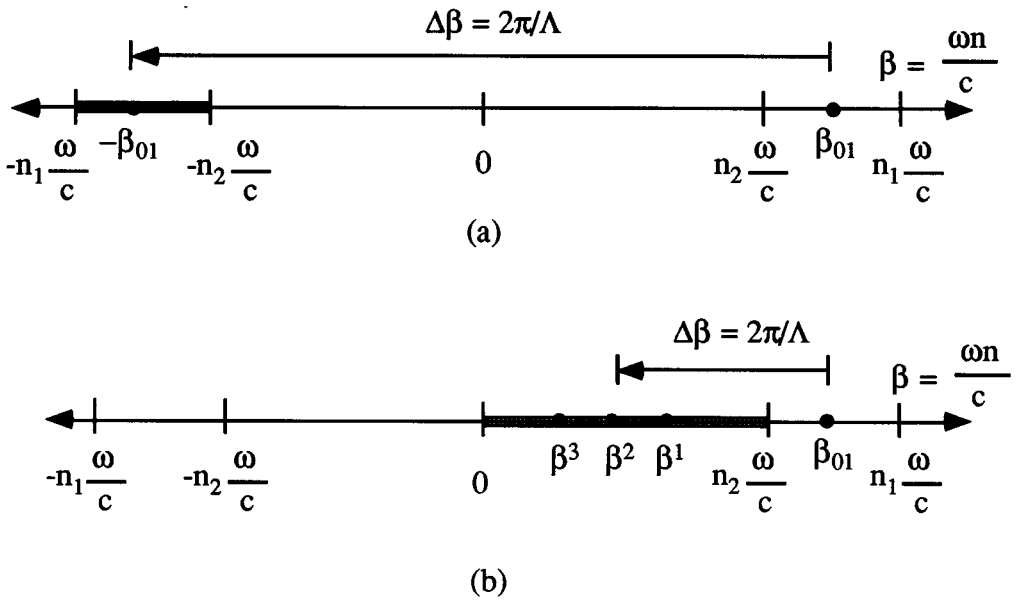


FIGURE 6.107

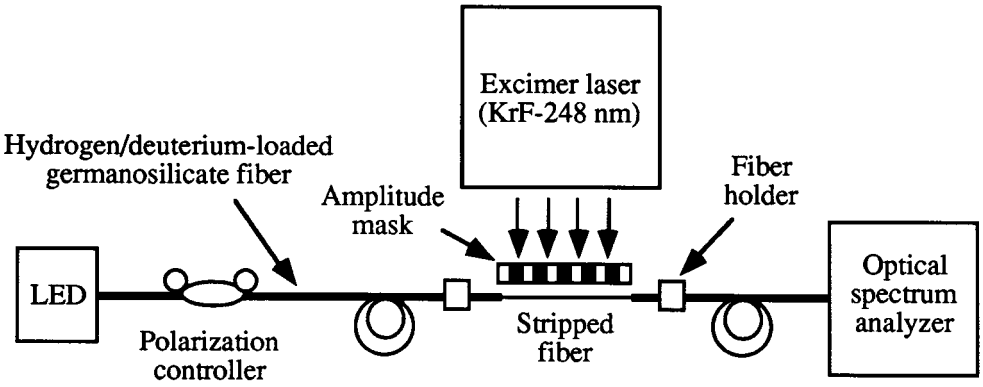


FIGURE 6.108 Setup used to fabricate long-period gratings.

hence periodicities ranging in hundreds of micrometers [19]. The cladding modes attenuate rapidly as they propagate along the length of the fiber due to the lossy cladding-coating interface and bends in the fiber. Since $\Delta\beta$ is discrete and a function of the wavelength, this coupling to the cladding modes is highly selective, leading to a wavelength-dependent loss. As a result, any modulation of the core and cladding guiding properties modifies the spectral response of long-period gratings, and this phenomenon can be utilized for sensing purposes. Moreover, since the cladding modes interact with the fiber jacket or any other material surrounding the cladding, changes in the properties of these ambient materials can also be detected.

Fabrication Procedure

To fabricate long-period gratings, hydrogen-loaded (3.4 mol%) germanosilicate fibers are exposed to 248 nm UV radiation from a KrF excimer laser, through a chrome-plated amplitude mask possessing a periodic rectangular transmittance function. Figure 6.108 shows the setup used to fabricate the gratings. The laser was pulsed at 20 Hz with a 8 ns pulse duration. The typical writing times for an energy of 100 mJ cm^{-2} $pulse^{-1}$ and a 2.5 cm exposed length vary between 6 to 15 min for different fibers. The coupling wavelength, λ_P, shifts to higher values during exposure, due to the photoinduced enhancement of the refractive index of the fiber core and the resulting increase in β_{01}. After writing, the gratings are annealed at 150°C for 10 h to remove the unreacted hydrogen. This high-temperature annealing causes λ_p to move to shorter wavelengths due to the decay of UV-induced defects and diffusion of molecular hydrogen from the fiber. Figure 6.109 depicts the typical transmittance of a grating. Various attenuation bands correspond to coupling to discrete cladding modes of different orders. A number of gratings can be fabricated at the same time by placing more than one fiber behind the amplitude mask. Moreover, the stability requirements during the writing process are not as severe as those for short-period Bragg gratings.

For coupling to the highest-order cladding-mode, the maximum isolation (loss in transmission intensity) is typically in the 5 to 20 dB range on wavelengths, depending on fiber parameters, duration of UV exposure, and mask periodicity. The desired fundamental coupling wavelength can easily be varied using inexpensive amplitude masks of different periodicities. The insertion loss, polarization-mode dispersion, backreflection, and polarization-dependent loss of a typical grating are 0.2 dB, 0.01 ps, –80 dB, and 0.02 dB, respectively. The negligible polarization sensitivity and backreflection of these devices eliminate the need for expensive polarizers and isolators.

Preliminary experiments were performed to examine the strain sensitivity of long-period gratings written in different fibers [21, 22]. Gratings were fabricated in four different types of fibers: standard dispersion-shifted fiber (DSF), standard 1550 nm fiber, and conventional 980 and 1050 nm single-mode fibers, which for the sake of brevity are referred to as fibers A, B, C, and D, respectively. The strain sensitivity of gratings written in different fibers was determined by axially straining the gratings between

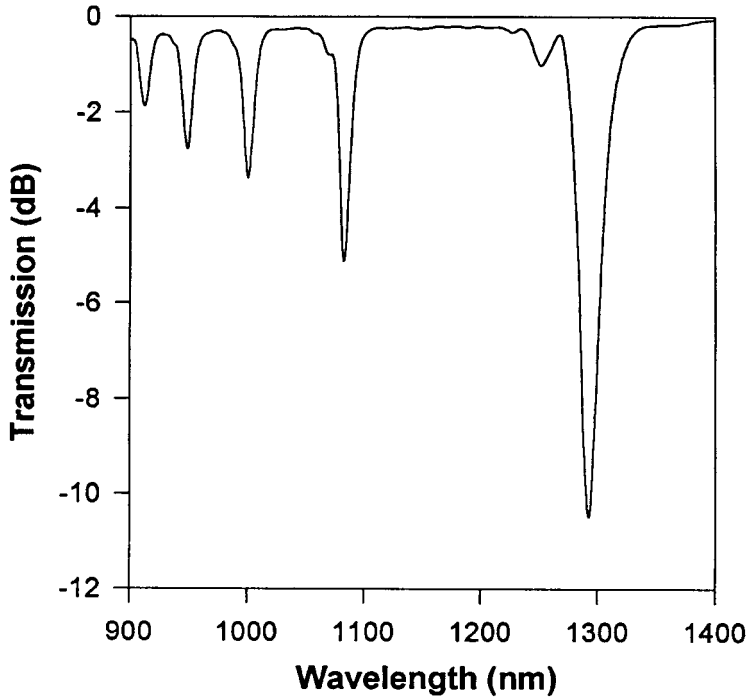


FIGURE 6.109 Typical transmission of a grating.

two longitudinally separated translation stages. The shift in the peak loss wavelength of the grating in fiber D as a function of the applied strain is depicted in Figure 6.110, along with that for a Bragg grating (about 9 nm $\%\varepsilon^{-1}$, at 1300 nm) [9]. The strain coefficients of wavelength shift (β) for fibers A, B, C, and D are shown in Table 6.28. Fiber D has a coefficient 15.2 nm $\%\varepsilon^{-1}$, which gives it a strain-induced shift that is 50% larger than that for a conventional Bragg grating. The strain resolution of this fiber for a 0.1 nm detectable wavelength shift is 65.75 $\mu\varepsilon$.

The demodulation scheme of a sensor determines the overall simplicity and sensitivity of the sensing system. Short-period Bragg grating sensors were shown to possess signal processing techniques that are complex and expensive to implement. A simple demodulation method to extract information from long-period gratings is possible. The wide bandwidth of the resonance bands enables the wavelength shift due to the external perturbation to be converted into an intensity variation that can be easily detected.

Figure 6.111 shows the shift induced by strain in a grating written in fiber C. The increase in the loss at 1317 nm is about 1.6 dB. A laser diode centered at 1317 nm was used as the optical source, and the change in transmitted intensity was monitored as a function of applied strain. The transmitted intensity is plotted in Figure 6.112 for three different trials. The repeatability of the experiment demonstrates the feasibility of using this simple scheme to utilize the high sensitivity of long-period gratings. The transmission of a laser diode centered on the slope of the grating spectrum on either side of the resonance wavelength can be used as a measure of the applied perturbation. A simple detector and amplifier combination at the output can be used to determine the transmission through the detector. On the other hand, a broadband source can also be used to interrogate the grating. At the output, an optical bandpass filter can be used to transmit only a fixed bandwidth of the signal to the detector. The bandpass filter should again be centered on either side of the peak loss band of the resonance band. These schemes are easy to implement, and unlike conventional Bragg gratings, the requirement of complex and expensive interferometric demodulation schemes is not necessary [22].

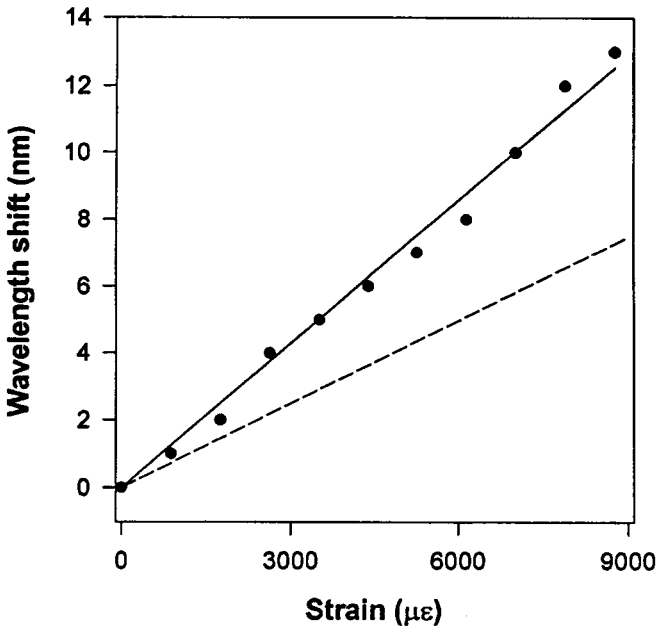


FIGURE 6.110 Shift in peak loss wavelength as a function of the applied strain.

TABLE 6.28 Strain Sensitivity of Long-Period Gratings Written in Four Different Types of Fibers

Type of fiber	Strain sensitivity (nm %ε^{-1})
A — Standard dispersion-shifted fiber (DSF)	–7.27
B — Standard 1550 nm communication fiber	4.73
C — Converntional 980 nm single-mode fiber	4.29
D — Conventional 1060 nm single-mode fiber	15.21

Note: The values correspond to the shift in the highest order resonance wavelength.

Temperature Sensitivity of Long-Period Gratings

Gratings written in different fibers were also tested for their cross-sensitivity to temperature [22]. The temperature coefficients of wavelength shift for different fibers are shown in Table 6.29. The temperature sensitivity of a fiber Bragg grating is 0.014 nm °C^{-1}. Hence, the temperature sensitivity of a long-period grating is typically an order of magnitude higher than that of a Bragg grating. This large cross-sensitivity to ambient temperature can degrade the strain sensing performance of the system unless the output signal is adequately compensated. Multiparameter sensing using long-period gratings has been proposed to obtain precise strain measurements in environments with temperature fluctuations [21].

In summary, long-period grating sensors are highly versatile. These sensors can easily be used in conjunction with simple and inexpensive detection techniques. Experimental results prove that these methods can be used effectively without sacrificing the enhanced resolution of the sensors. Long-period grating sensors are insensitive to the input polarization and do not require coherent optical sources. The cross-sensitivity to temperature is a major concern while using these gratings for strain measurements.

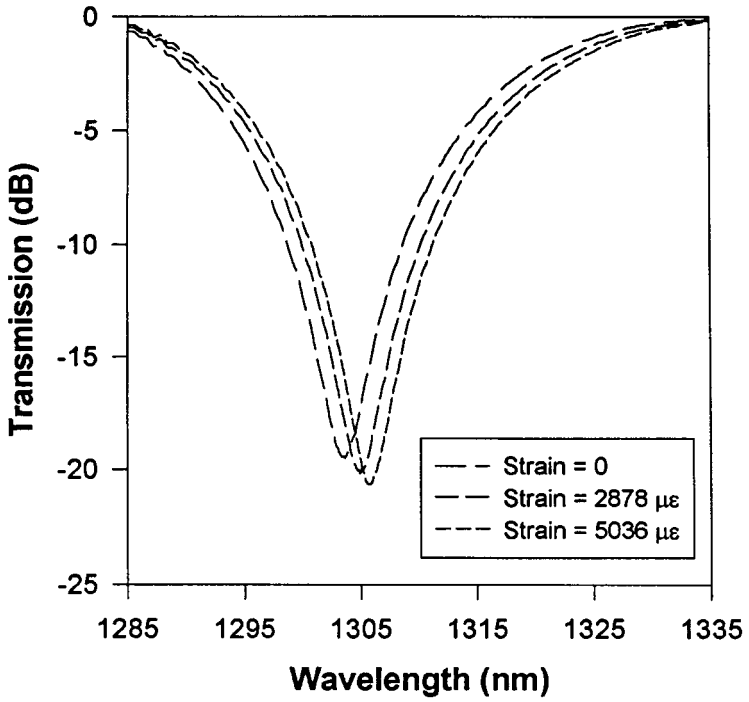


FIGURE 6.111 The shift induced by strain in a grating written in fiber C.

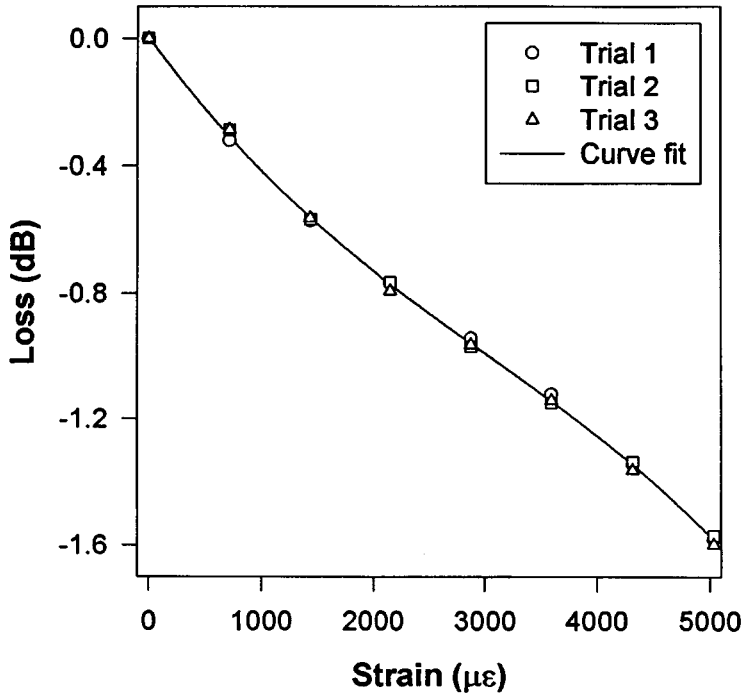


FIGURE 6.112 Plot of the change in transmitted intensity as a function of strain, for three different trials.

Table 6.29 Temperature Sensitivity of Long-Period Gratings Written in Four Different Types of Fibers

Type of fiber	Temperature sensitivity (nm °C^{-1})
A — Standard dispersion-shifted fiber (DSF)	0.062
B — Standard 1550 nm communication fiber	0.058
C — Converntional 980 nm single-mode fiber	0.154
D — Conventional 1060 nm single-mode fiber	0.111

Note: The values correspond to the shift in the highest order resonance wavelength.

Comparison of Sensing Schemes

Based on the above results, the interferometric sensors have a high sensitivity and bandwidth, but are limited by the nonlinearity in their output signals. Conversely, intrinsic sensors are susceptible to ambient temperature changes while the grating-based sensors are simpler to multiplex. Each may be used in specific applications.

Conclusion

We have investigated the performance of four different interferometric and grating-based sensors. This analysis was based on the sensor head fabrication and cost, signal processing, cross-sensitivity to temperature, resolution, and operating range. The relative merits and demerits of the various sensing schemes were also discussed.

References

1. J. Sirkis, Phase-strain-temperature model for structurally embedded interferometric optical fiber strain sensors with applications, *Fiber Optic Smart Structures and Skins IV*, SPIE, Vol. 1588, 1991.
2. R. O. Claus, M. F. Gunther, A. Wang, and K. A. Murphy, Extrinsic Fabry-Perot sensor for strain and crack opening displacement measurements from –200 to 900°C, *J. Smart Materials and Structures*, 1, 237-242, 1992.
3. K. A. Murphy, M. F. Gunther, A. M. Vengsarkar, and R. O. Claus, Fabry-Perot fiber optic sensors in full-scale fatigue testing on an F-15 aircraft, *Appl. Optics,* 31, 431-433, 1991.
4. V. Bhatia, C. A. Schmid, K. A. Murphy, R. O. Claus, T. A. Tran, J. A. Greene, and M. S. Miller, Optical fiber sensing technique for edge-induced and internal delamination detection in composites, *J. Smart Materials Structures*, 4, 164-169, 1995.
5. V. Bhatia, M. J. de Vries, K. A. Murphy, R. O. Claus, T. A. Tran, and J. A. Greene, Extrinsic Fabry-Perot interferometers for absolute measurements, *Fiberoptic Product News*, 9(Dec.), 12-13, 1994.
6. V. Bhatia, M. B. Sen, K. A. Murphy, and R. O. Claus, Wavelength-tracked white light interferometry for highly sensitive strain and temperature measurements, *Electron. Lett.*, 32, 247-249, 1996.
7. C. E. Lee and H. F. Taylor, Fiber-optic Fabry-Perot temperature sensor using a low-coherence light source, *J. Lightwave Technol.*, 9, 129-134, 1991.
8. J. A. Greene, T. A. Tran, K. A. Murphy, A. J. Plante, V. Bhatia, M. B. Sen, and R. O. Claus, Photoinduced Fresnel reflectors for point-wise and distributed sensing applications, *Proc. Conf. Smart Structures and Materials*, *SPIE'95,* paper 2444-05, February 1995.
9. K. O. Hill, Y. Fuijii, D. C. Johnson, and B. S. Kawasaki, Photosensitivity in optical fiber waveguides: applications to reflection filter fabrication, *Appl. Phys. Lett.*, 32, 647, 1978.
10. G. Meltz, W. W. Morey, and W. H. Glenn, Formation of Bragg gratings in optical fibers by transverse holographic method, *Optics Lett.*, 14, 823, 1989.

11. P. J. Lemaire, A. M. Vengsarkar, W. A. Reed, V. Mizrahi, and K. S. Kranz, Refractive index changes in optical fibers sensitized with molecular hydrogen, in *Proc. Conf. Optical Fiber Communications, OFC'94,* Technical Digest, paper TuL1, 47, 1994.
12. R. Kashyap, Photosensitive optical fibers: devices and applications, *Optical Fiber Technol.*, 1**,** 17-34, 1994.
13. D. Z. Anderson, V. Mizrahi, T. Ergodan, and A. E. White, Phase-mask method for volume manufacturing of fiber phase gratings, in *Proc. Conf. Optical Fiber Communication*, post-deadline paper PD16, 1993, p. 68.
14. A. D. Kersey and T. A. Berkoff, Fiber-optic Bragg-grating differential-temperature sensor, *IEEE Photonics Technol. Lett.*, 4, 1183-1185, 1992.
15. V. Bhatia, M. B. Sen, K. A. Murphy, A. Wang, R. O. Claus, M. E. Jones, J. L. Grace, and J. A. Greene, Demodulation of wavelength-encoded optical fiber sensor signals using fiber modal interferometers, *SPIE Photonics East*, Philadelphia, PA, paper 2594-09, October 1995.
16. M. G. Xu, L. Dong, L. Reekie, J. A. Tucknott, and J. L. Cruz, Chirped fiber gratings for temperature-independent strain sensing, in *Proc. First OSA Topical Meet. Photosensitivity and Quadratic Nonlinearity in Glass Waveguides: Fundamentals and Applications*, paper PMB2, 1995.
17. K. O. Hill, B. Malo, K. Vineberg, F. Bilodeau, D. Johnson, and I. Skinner, Efficient mode-conversion in telecommunication fiber using externally written gratings, *Electron. Lett.,* 26, 1270-1272, 1990.
18. F. Bilodeau, K. O. Hill, B. Malo, D. Johnson, and I. Skinner, Efficient narrowband $LP_{01} \leftrightarrow LP_{02}$ mode convertors fabricated in photosensitive fiber: spectral response, *Electron. Lett.,* 27, 682-684, 1991.
19. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, J. E. Sipe, and T. E. Ergodan, Long-period fiber gratings as band-rejection filters, *Proc. Conf. Optical Fiber Communications, OFC '95,* post-deadline paper, PD4-2, 1995.
20. A. M. Vengsarkar, P. J. Lemaire, J. B. Judkins, V. Bhatia, J. E. Sipe, and T. E. Ergodan, Long-period fiber gratings as band-rejection filters, *J. Lightwave Technol.*, 14, 58-65, 1996.
21. V. Bhatia, M. B. Burford, K. A. Murphy, and A. M. Vengsarkar, Long-period fiber grating sensors, *Proc. Conf. Optical Fiber Communication,* paper ThP1, February 1996.
22. V. Bhatia and A. M. Vengsarkar, Optical fiber long-period grating sensors, *Optics Lett.*, 21, 692-694, 1996.
23. C. D. Butter and G. B. Hocker, Fiber optics strain gage, *Appl. Optics*, 17, 2867-2869, 1978.
24. J. S. Sirkis and H. W. Haslach, Interferometric strain measurement by arbitrarily configured, surface mounted, optical fiber, *J. Lightwave Technol.*, 8, 1497-1503, 1990.

6.12 Optical Beam Deflection Sensing

Grover C. Wetsel

Measurements of the intensity of the light reflected and transmitted by a sample have been sources of information concerning the structure of matter for over a century. In recent decades, it has been found that measurement of the position of an optical beam that has scattered from a sample is an important and versatile means of characterizing materials and the motion of devices. Surely, the availability of a well-collimated beam from a laser has been crucial in the development of techniques and applications of *optical beam deflection* (OBD) sensing; however, the development and ready availability of various types of *position sensing detectors* (PSDs) have also been important factors. Optical beam deflection may be caused, for example, by propagation of a laser beam through a refractive-index gradient or by reflection from a displaced surface. A PSD provides an electronic signal that is a function of the laser beam position on the detector.

In this section, applications of optical beam deflection sensing are reviewed, the theories of operation of the three most common types of OBD sensors are developed, and typical operational characteristics of the devices are presented. The advantages and disadvantages of the various PSDs are also discussed.

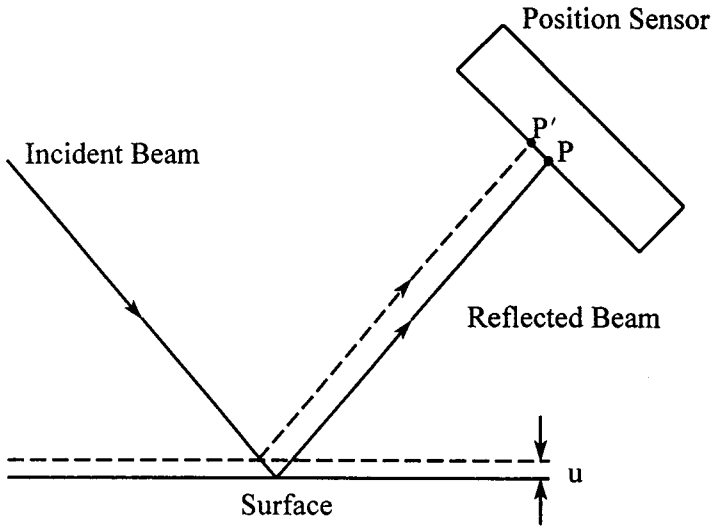


FIGURE 6.113 A schematic diagram of the basic optical-beam-deflection (OBD) sensing configuration.

A schematic diagram of the basic OBD sensing configuration is illustrated in Figure 6.113. In this case, the displacement, u, of the surface causes the position of the reflected beam on the PSD to move from point P to point P′; the positional change produces a change in the output voltage of the PSD. The output voltage, V, of the PSD electronics can be calibrated in terms of the actual displacement, u, by measuring V versus u for known displacements.

OBD sensing has been used in a variety of applications, including photothermal optical beam deflection (PTOBD) spectroscopy [1], absolute measurement of optical attenuation [2], PTOBD imaging of surface and subsurface structure [3], photothermal displacement spectroscopy [4], atomic-force microscopy [5], and materials characterization [6]. It has also been used as an uncomplicated, sensitive, and accurate method of measurement of surface motion for scanning tunneling microscope scanner transducers [7] and ultrasonic transducer imaging [8].

Theory

The three basic types of devices for OBD sensing are (1) a photodetector behind a sharply edged screen (a knife edge); (2) a small array of photodetectors separated by relatively small, insensitive areas (bicell, quadcell); and (3) a continuous solid-state position sensor (one or two dimensional). Sensing characteristics of a device are determined by the effect of optical beam displacement on the photodetector power distribution. Since laser beams are commonly used in OBD sensing, the analysis involves the assumption that the spatial distribution of the intensity (I) in the plane perpendicular to the direction of wave propagation is axially symmetric with a Gaussian radial variation.

Knife-Edge Photodetector

The essential features of a PSD are represented by a photodetector shadowed by a semi-infinite knife edge, $y < 0$, as illustrated in Figure 6.114. As can be anticipated from the symmetry of the arrangement and proved mathematically, the maximal deflection sensitivity occurs when the undeflected beam is centered on the knife edge. The intensity of the light reaching the photodetector due to the displacement (u) of the center of the beam is given in the reference frame of the displaced beam by:

$$I\left(r'\right) = \frac{aP}{\pi} e^{-ar'^2} \tag{6.126}$$

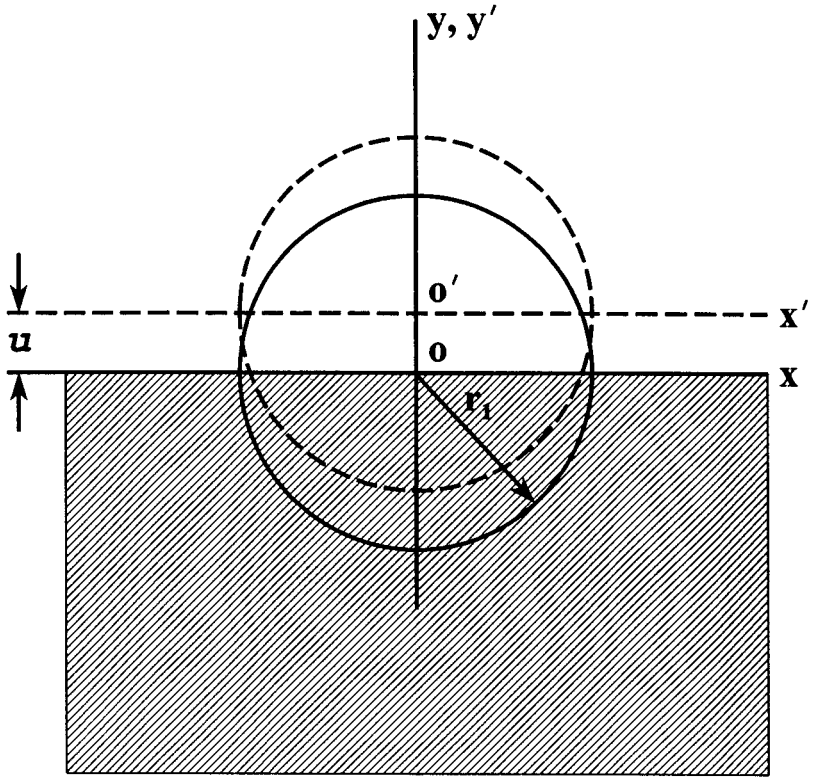


FIGURE 6.114 Essential features of a position-sending detector (PSD), as represented by a photodetector shadowed by a semiinfinite knife edge.

where P is the total incident beam power, $a = 2/r_1^2$, r_1 is the Gaussian beam radius, and $r'^2 = x'^2 + y'^2$. In terms of the coordinates (x,y) of the undeflected beam, the rectangular coordinates of the deflected beam are $x' = x$ and $y' = y - u$. The power (P_d) on the detector is thus given by:

$$P_d = \frac{aP}{\pi} \int_0^\infty e^{-a(y-u)^2}\,dy \int_{-\infty}^{\infty} e^{-ax^2}\,dx = \frac{P}{2}\left[1 + erf\left(\sqrt{2}\,\frac{u}{r_1}\right)\right] \tag{6.127}$$

where *erf* is the error function. One can see by inspection of Equation 6.127 that the essential characteristics of this position sensor are determined by u/r_1. The normalized response, P_d/P, is shown in Figure 6.115 as a function of u/r_1. When $u = r_1$, then $P_d = 97.7\%\ P$.

The deflection sensitivity is given by the slope of Equation 6.127,

$$\frac{dP_d}{du} = \sqrt{\frac{2}{\pi}}\,\frac{P}{r_1}\,e^{-2\left(\frac{u}{r_1}\right)^2}, \quad with \left(\frac{dP_d}{du}\right)_{max} = \left(\frac{dP_d}{du}\right)_{u=0} = \sqrt{\frac{2}{\pi}}\,\frac{P}{r_1} \tag{6.128}$$

Define the small-signal position sensor sensitivity (units of m^{-1}):

$$\alpha_{KE} \equiv \frac{1}{P}\left(\frac{dP_d}{du}\right)_{u=0} = \frac{1}{r_1}\sqrt{\frac{2}{\pi}} \tag{6.129}$$

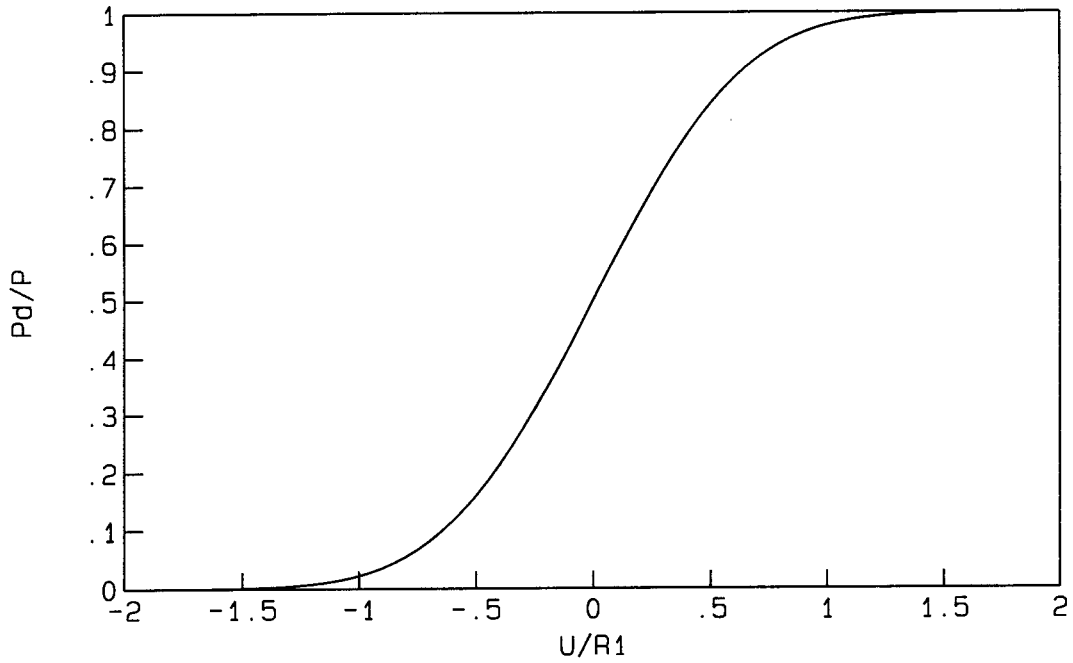


FIGURE 6.115 The normalized response P_d/P as a function of u/r_1 (u = displacement of center of beam; r_1 = Gaussian beam radius; P_d = power on the detector; P = total incident beam power).

The optical power reaching the photodetector for small signals is then given by:

$$P_d \cong \frac{P}{2} + \left(\frac{dP_d}{du}\right)_{u=0} u = \frac{P}{2} + \alpha u P = \frac{P}{2}\left[1 + 2\alpha u\right] \tag{6.130}$$

The photodetector signal will be linear in displacement if $u \le 0.387 r_1$.

The photocurrent is:

$$I = KP_d \cong \frac{KP}{2}\left[1 + 2\alpha u\right] \tag{6.131}$$

where K is the photodetector responsivity in A/W. The position sensor voltage is obtained using a transimpedance amplifier with gain Z:

$$V = KZP_d \cong \frac{KZP}{2}\left[1 + 2\alpha u\right] \tag{6.132}$$

Bicell Detector

The deflection of a Gaussian beam initially centered in the insensitive gap of a bicell detector is illustrated in Figure 6.116. The power incident on the upper half of the bicell is given by:

$$P_2 = \frac{aP}{\pi}\int_{y_2}^{\infty} e^{-a(y-u)^2} dy \int_{-\infty}^{\infty} e^{-ax^2} dx = \frac{P}{2}\left[1 - erf\left(\frac{\sqrt{2}}{r_1}\left(y_2 - u\right)\right)\right] \tag{6.133}$$

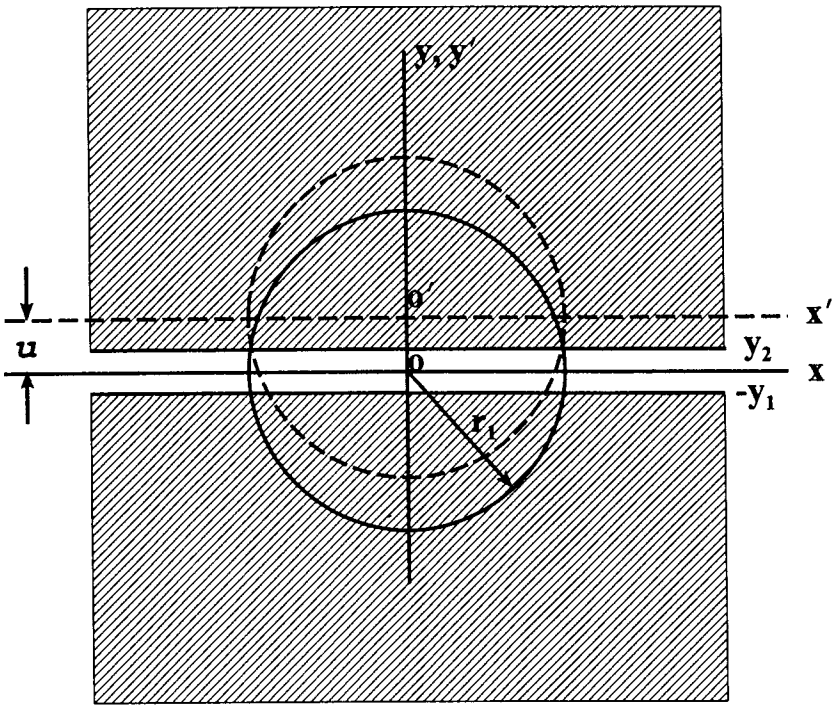


FIGURE 6.116 Deflection of a Gaussian beam initially centered in the insensitive gap of a bicell detector.

The power incident on the lower half of the bicell is given by:

$$P_1 = \frac{aP}{\pi}\int_{-\infty}^{-y_1} e^{-a(y-u)^2}\,dy \int_{-\infty}^{\infty} e^{-ax^2}\,dx = \frac{P}{2}\left[1 - erf\left(\frac{\sqrt{2}}{r_1}(y_1+u)\right)\right] \tag{6.134}$$

The photocurrent from each detector of the bicell is converted to voltage by identical transimpedance amplifiers: $V_2 = KZP_2$ and $V_1 = KZP_1$; a difference amplifier is then used to obtain the bicell signal voltage:

$$V = V_2 - V_1 = KZ(P_2 - P_1) = \frac{KZP}{2}\left[erf\left(\frac{\sqrt{2}}{r_1}(y_1+u)\right) - erf\left(\frac{\sqrt{2}}{r_1}(y_2-u)\right)\right] \tag{6.135}$$

The normalized response, $2V/(KZP)$, is shown in Figure 6.117 as a function of u/r_1 for $y_1 = y_2 = r_1/10$.

Suppose that the beam is centered in the gap, $y_1 = y_2$; then, for small displacements, one obtains:

$$V \cong 2\sqrt{\frac{2}{\pi}}\,\frac{KZPu}{r_1}\,e^{-2(y_1\ r_1)^2} \tag{6.136}$$

The small-signal sensitivity is:

$$\alpha_{BC} \equiv \frac{1}{KZP}\left(\frac{dV}{du}\right)_{u=0} = 2\sqrt{\frac{2}{\pi}}\,\frac{e^{-2(y_1\ r_1)^2}}{r_1} \tag{6.137}$$

This quantity is optimized when $r_1 = 2y_1$, and the optimal sensitivity is $0.484/y_1$.

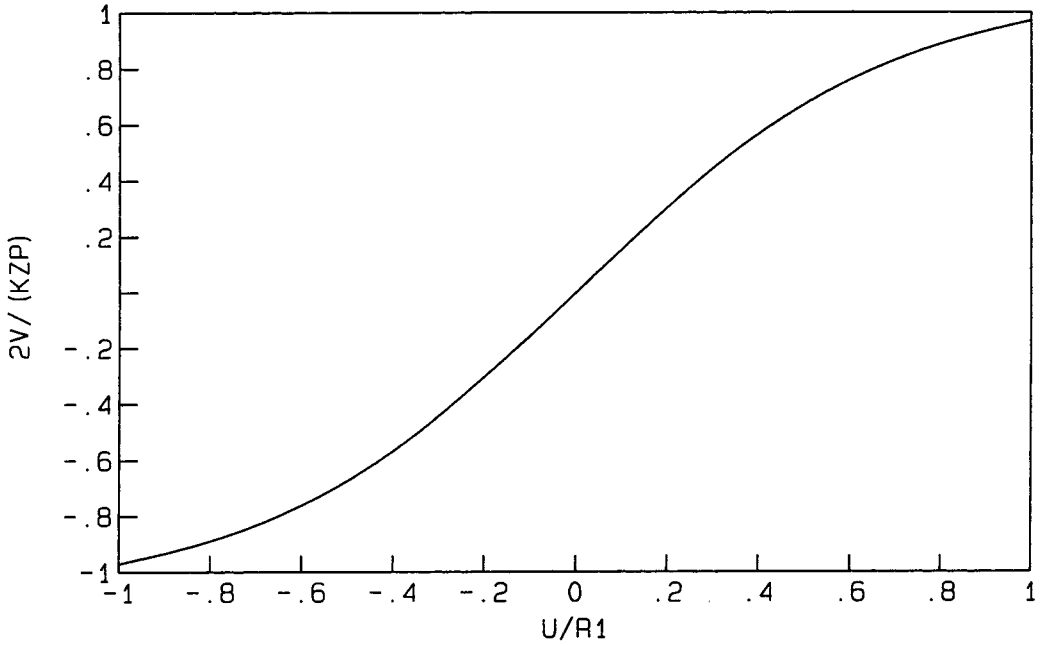


FIGURE 6.117 The normalized response $2V/KZP$ in a bicell detector as a function of u/r_1 for $y_1 = y_2 = r_1/10$.

Continuous Position Sensor

The position information in a continuous position sensor (also known as a lateral-effect photodiode) is derived from the divided path taken by photon-generated electrons to two back electrodes on the device. For a homogeneous device, the current to an electrode depends only on the distance of the centroid of the light beam from that electrode; the currents would be equal in an ideal device when the beam is located at its electrical center.

Consider the analysis of a one-dimensional continuous PSD. The current signal from each electrode is converted to a voltage signal by a transimpedance amplifier with gain, Z. Operational amplifiers are then used to provide the sum signal, $V_s = KPZA_s$, and the difference signal, $V_d = KPZA_d\alpha u$, where A_s and A_d are the sum and difference amplifier gains, respectively. The position sensor sensitivity, α, is then given by:

$$\alpha \equiv \frac{V_d A_s}{u V_s A_d} \tag{6.138}$$

which is determined by measuring V_d and V_s as a function of u.

Characterization of PSDs

The PSD characteristics presented here were measured by mounting the device to a translation stage with an optical encoder driven by a piezoelectric motor [9]; the position accuracy was ±0.1 μm. An appropriately attenuated He–Ne laser beam was directed at normal incidence to the PSD as it was translated past the beam. The PSD sum and difference voltages were measured with GP–IB digital voltmeters as a function of displacement; the PSD displacement and voltage measurements were computer controlled.

The operation of a knife-edge PSD can be evaluated using the signal from one side of a bicell as well as from a photodetector behind a knife edge. The transimpedance amplifier output voltage corresponding

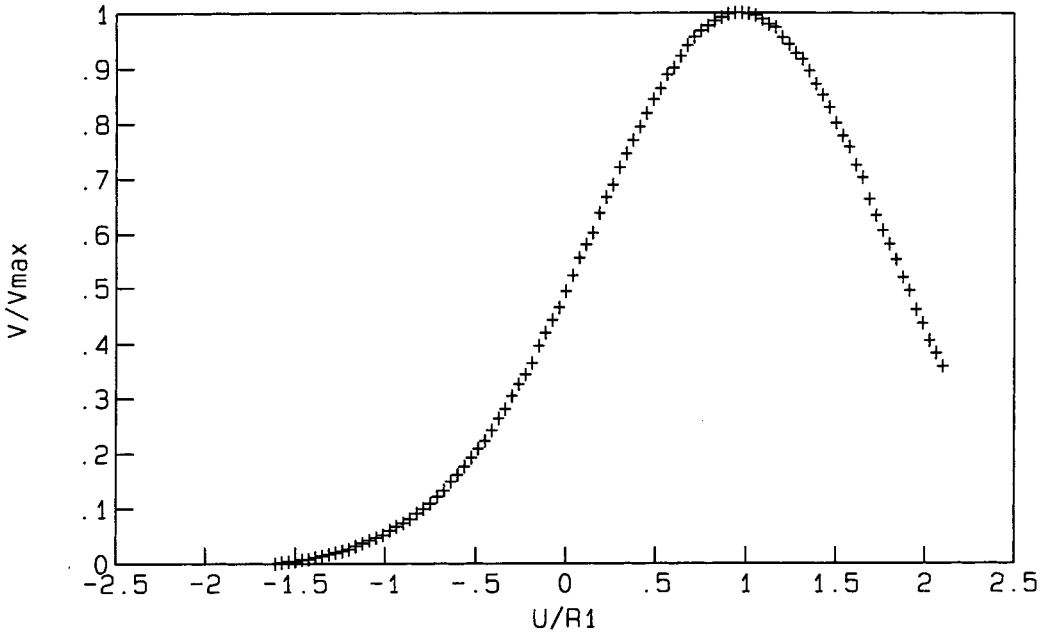


FIGURE 6.118 Transimpedance amplifier output voltage corresponding to the current signal from one cell of a United Detector Technology (UDT) SPOT2D bicell as a function of displacement. Data are plotted as normalized cell voltage vs. u/r_1.

to the current signal from one cell of a United Detector Technology [10] (UDT) SPOT2D bicell is shown as a function of displacement in Figure 6.118. The data are plotted as normalized cell voltage as a function of u/r_1. Since the signal from one cell of a bicell is equivalent to a knife-edge photodetector, the data can be compared with Figure 6.115. The data are in reasonable agreement for $u/r_1 < 1$ and a best fit allows the inference of $r_1 = 0.4$ mm; the disagreement for $u/r_1 \geq 1$ corresponds to the laser beam partially moving off the outside edge of the cell.

The operation of a bicell PSD is evaluated by the data of Figure 6.119 which shows the difference voltage as a function of displacement for the UDT SPOT2D. The deviations from the theoretical curve of figure 6.117 for $u \leq 0.4$ mm and $u \geq 1.1$ mm are due to the beam moving off the outside edges of the bicell. The linear region of the PSD is centered at the optimal quiescent point ($V_1 - V_2 = 0$); a least-squares fit of a straight line gives a slope of $A_{12} = 4.88$ mV μm^{-1} for this device. For the data of Figures 6.118 and 6.119, $P = 0.37$ mW and $Z = 5$ kΩ; thus, the experimental value of the responsivity is determined to be $K_{exp} = 0.65$ A W^{-1}. The nominal gap size of the SPOT2D bicell is such that $y_1 = 63.5$ μm; thus, using this value of y_1 and $r_1 = 0.4$ mm in Equation 6.137, the calculated bicell sensitivity is 3.8×10^3 m^{-1}. The experimental value of sensitivity is calculated from A_{12}, K_{exp}, P, and Z to be 4.1×10^3 m^{-1}.

Evaluation of the operation of the x-axis of a UDT SC10 two-axis continuous PSD is illustrated in Figures 6.120 through 6.122. The PSD was oriented such that the y-axis was vertical; thus, ideally, V_{dy} is constant as the PSD is translated, except near the edges of the 10 mm × 10 mm device. As shown in Figure 6.120, V_{sx} is virtually constant except near the edges of the device; this means that the optical responsivity (K) does not greatly vary with the laser beam position on this detector. Variation of V_{dx} with displacement is shown in Figure 6.121; this PSD has a broad linear range. The linearity of continuous PSDs can often be improved by using the ratio, V_{dx}/V_{sx}, as shown in Figure 6.122; the sensitivity of this PSD is determined from a least-squares fit of the linear part of this characteristic to be 183 m^{-1}.

A newer class of PSDs is represented by the Sitek [11] 1L10 single-axis continuous PSD, which is typically more linear. An example of the V_d/V_s versus u characteristic is shown in Figure 6.123. The sensitivity is determined from a least-squares fit of the linear part of the characteristic to be 307 m^{-1}.

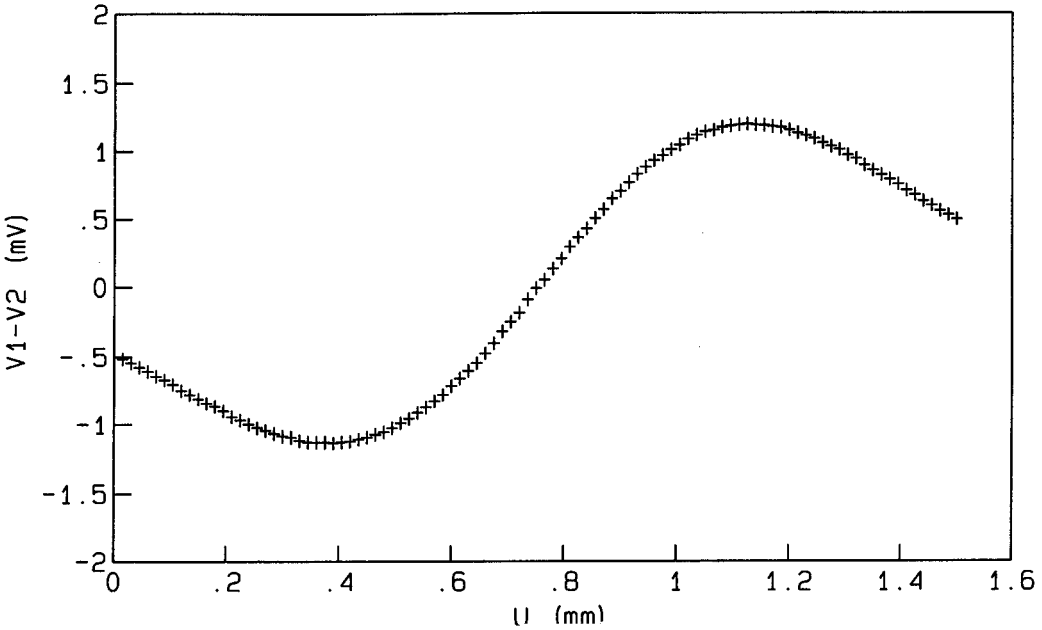


FIGURE 6.119 Difference voltage as a function of displacement for the United Detector Technology (UDT) SPOT2D bicell.

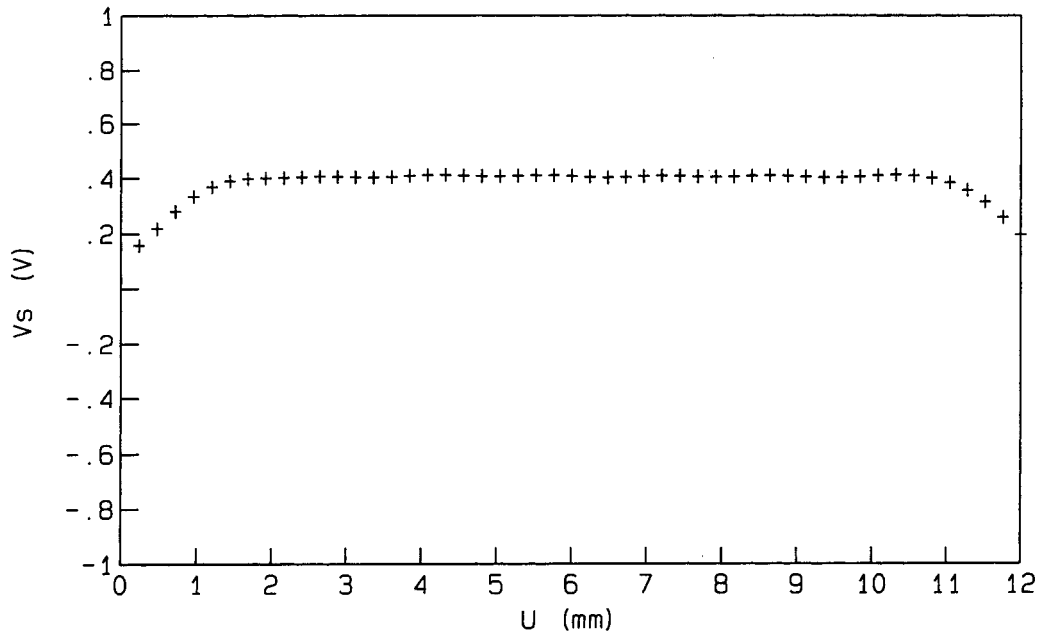


FIGURE 6.120 Sum voltage vs. displacement for x axis of UDT SCIO continuous PSD.

Summary

A PSD composed of a knife-edge photodetector is simple, has a sensitivity that depends only on the radius of the laser beam, and has a response time determined by the photodetector. Thus, for fast rise-time detection of small displacements, this type of PSD is to be preferred.

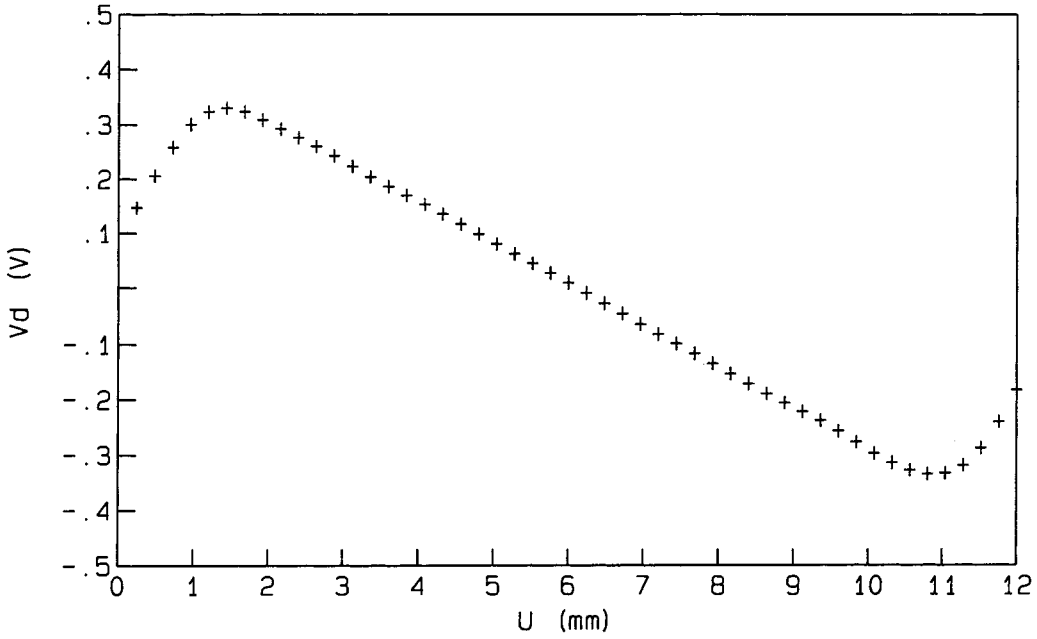


FIGURE 6.121 Difference voltage vs. displacement for x axis of UDT SCIO continuous PSD.

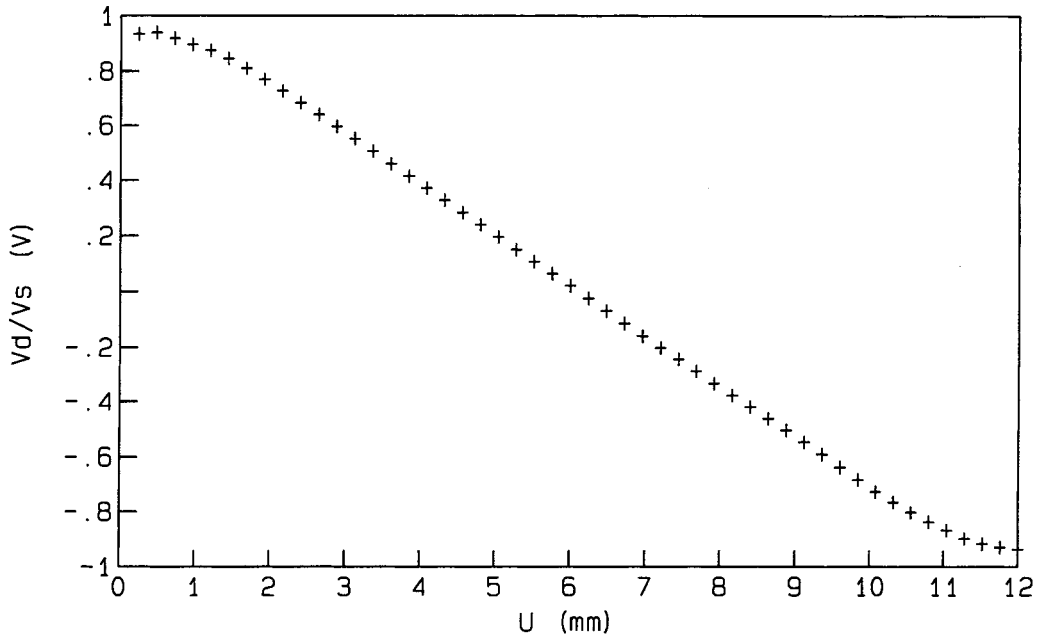


FIGURE 6.122 Ratio of difference voltage to sum voltage vs. displacement for x axis of UDT SCIO continuous PSD.

The maximal sensitivity of a bicell is about 20% greater than that of a knife-edge photodetector. It has the advantages of small size and economy. The nominal risetime of the UDT SPOT2D is of the order of 10 ns. Furthermore, two-axis detection can be readily obtained with a quadcell, with an increase in risetime due to the increased capacitance of the detector.

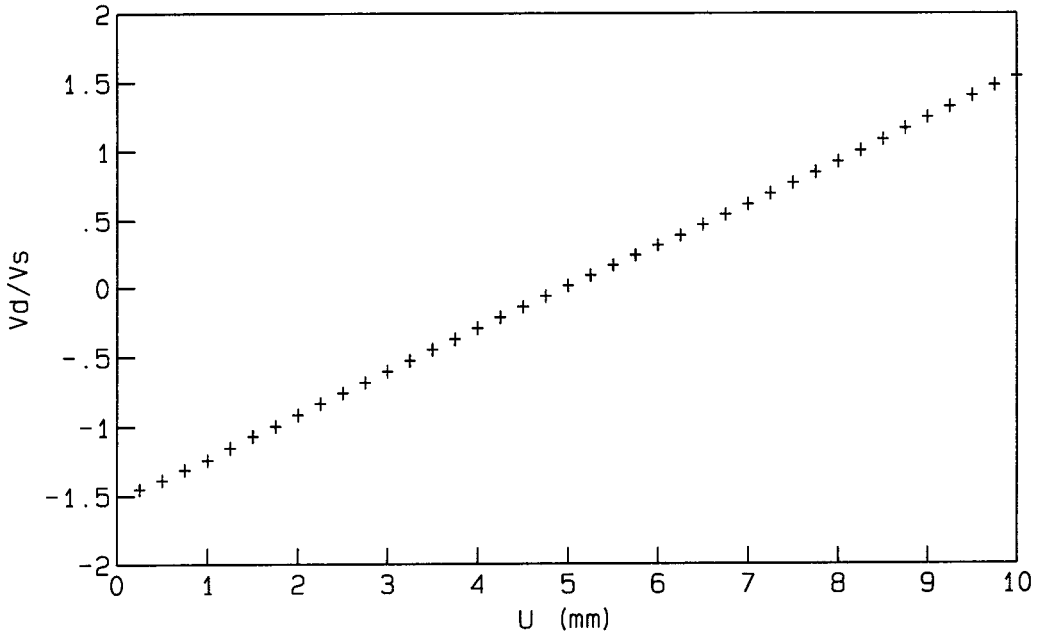


FIGURE 6.123 Ratio of difference voltage to sum voltage vs. displacement for Sitek 1L10 continuous PSD.

The disadvantages due to small laser beam sizes and small displacements are overcome by the continuous PSDs. Since these devices are typically much larger (available in several-inch diameters), they typically have longer risetimes than the other PSDs. However, the Sitek 1L10, with a 10 mm linear active range has a measured upper half-power frequency of 3 MHz.

In applications where the output signal from a PSD must be linearly proportional to the displacement of the beam, analog-divider operational amplifiers to obtain V_d/V_s in real time are used to extend the range of linearity of the device. Unfortunately, the frequency response of these amplifiers are often the frequency-response-limiting factors of the PSD system. In cases where high-frequency response is important, V_d alone can often be used if care is taken to operate in the linear range of the device. For large static displacements that are of the order of the size of the detector, V_d and V_s can be recorded with computer-controlled data acquisition, the calibration characteristic numerically fitted to a polynomial, and then any voltage from the detector can be related to beam position.

The noise limitations in OBD sensing are due to the laser, the nature of the reflecting surface, and the PSD. Lasers with good amplitude stability are to be preferred, but this is not an important contribution to noise when V_d/V_s is used to infer displacement. Laser beam-pointing stability, on the other hand, is important. If the reflecting surface is that of a typical solid, then negligible noise is introduced on reflection; this may not be true for a reflector such as a pellicle, where Brownian motion of the surface may be significant. The noise limitations of the PSD are the usual ones associated with the photodetector and the amplifiers.

References

1. A. C. Boccara, D. Fournier, and J. Badoz, *Appl. Phys. Lett.,* 30, 933, 1983.
2. G. C. Wetsel, Jr. and S. A. Stotts, *Appl. Phys. Lett.,* 42, 931, 1983.
3. e.g.: D. Fournier and A. C. Boccara, *Scanned Image Microscopy,* E. A. Ash, Ed., London: Academic Press, 1980, 347-351; J. C. Murphy and L. C. Aaamodt, *Appl. Phys. Lett.,* 39, 519, 1981; G. C. Wetsel, Jr. and F. A. McDonald, *Appl. Phys. Lett.,* 41, 926, 1982.
4. M. A. Olmstead, S. Kohn, N. M. Amer, D. Fournier, and A. C. Boccara, *Appl. Phys. A,* 132, 68, 1983.

5. G. Meyer and N. M. Amer, *Appl. Phys. Lett.,* 53, 1045, 1988.
6. J. C. Murphy and G. C. Wetsel, Jr., *Mater. Evaluation,* 44, 1224, 1986.
7. G. C. Wetsel, Jr., S. E. McBride, R. J. Warmack, and B. Van de Sande, *Appl. Phys. Lett.,* 55, 528, 1989.
8. S. E. McBride and G. C. Wetsel, Jr., Surface-displacement imaging using optical beam deflection, *Review of Progress in Quantitative Nondestructive Evaluation,* Vol. 9A, D. O. Thompson and D. E. Chimenti, (Eds.), New York: Plenum, 1990, 909-916.
9. Burleigh Instruments, Inc., Fishers, NY 14453.
10. United Detector Technology, 12525 Chadron Ave., Hawthorne, CA 90250.
11. On-Trak Photonics Inc., 20321 Lake Forest Dr., Lake Forest, CA 92630.