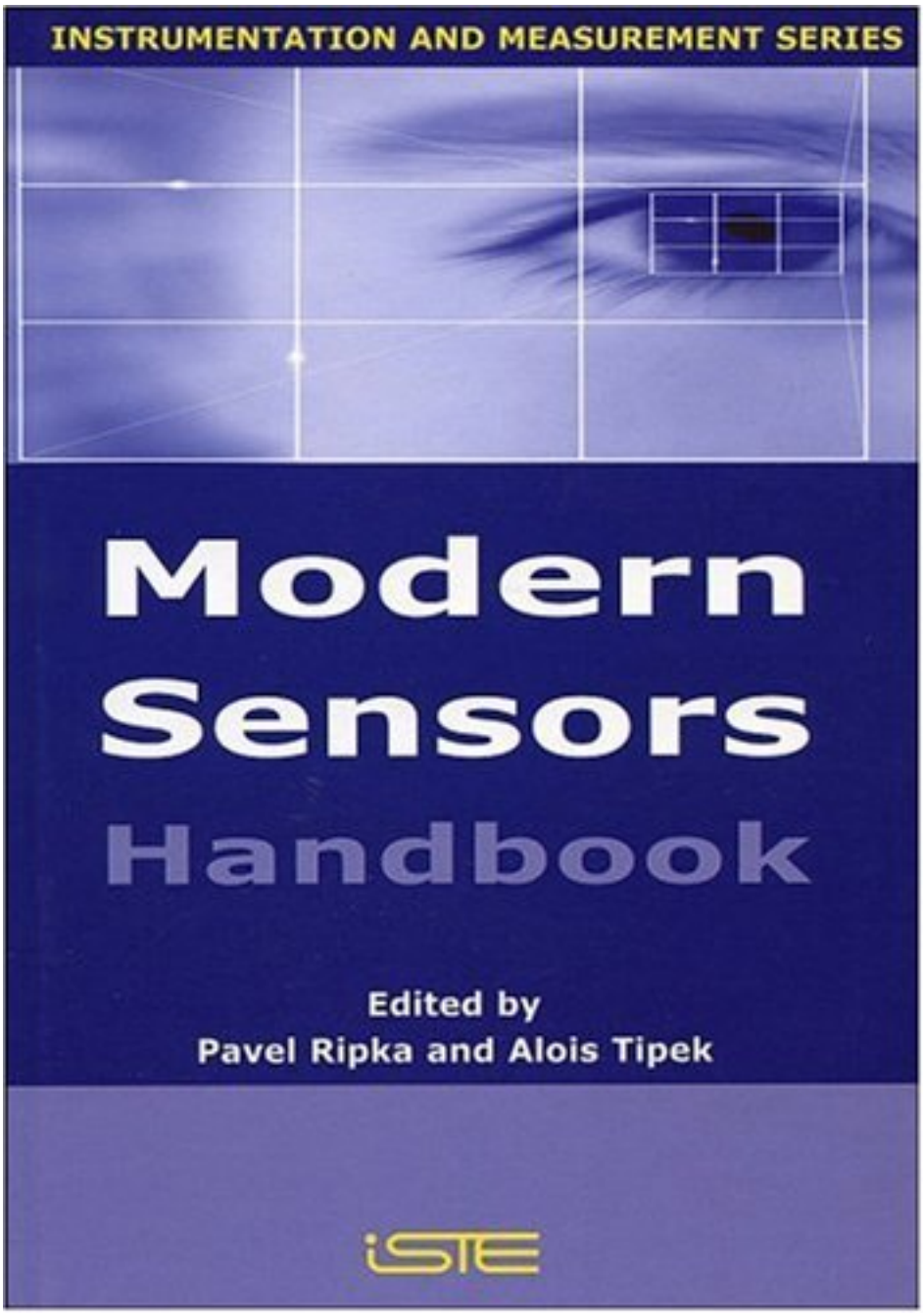
Modern Sensors Handbook



Modern Sensors Handbook

Edited by
Pavel Ripka
Alois Tipek

iSTE

First published in Great Britain and the United States in 2007 by ISTE Ltd

ISTE Ltd
6 Fitzroy Square
London W1T 5DX
UK

ISTE USA
4308 Patrice Road
Newport Beach, CA 92663
USA

www.iste.co.uk

Library of Congress Cataloging-in-Publication Data

Modern sensors handbook/edited by Pavel Ripka, Alois Tipek.
p. cm.
ISBN 978-1-905209-66-8
1. Detectors--Handbooks, manuals, etc. I. Ripka, Pavel. II. Tipek, Alois.
TA165.M585 2007
681'.2--dc22

2007003344

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 13: 978-1-905209-66-8

Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire.

Table of Contents

Chapter 1. Pressure Sensors 1
André MIGEON and Anne-Elisabeth LENEL

1.1. Introduction 1
1.2. Pressure 2
1.2.1. Pressure as a physical quantity 2
1.2.1.1. Static pressure 2
1.2.1.2. Units 3
1.2.2. Absolute, relative and differential sensors 3
1.2.3. Fluid physical properties 5
1.2.3.1. Liquids 5
1.2.3.2. Gases 5
1.2.3.3. Sensor pneumatic connection influence 6
1.3. Pressure ranges 6
1.3.1. Vacuum and ultra-vacuum 6
1.3.2. Middle range pressure 8
1.3.3. High pressure 10
1.4. Main physical principles 10
1.4.1. The sensing device 11
1.4.2. Sensors with elastic element 13
1.4.2.1. Conversion by resistance variation 13
1.4.2.2. Conversion by capacitance variation 21
1.4.2.3. Conversion by inductance variation 26
1.4.2.4. Conversion by piezoelectric effect 27
1.4.2.5. Conversion by oscillators 30
1.4.2.6. Optical conversion 38
1.4.2.7. Servo controlled sensors with balance of force 40
1.4.3. Vacuum sensors 41
1.4.3.1. Ionization pressure sensors 41
1.4.3.2. Heating effect sensors 42

1.5. Calibration: pressure standards 43
1.5.1. Low pressure standard 43
1.5.2. High pressure standard 43
1.6. Choosing a pressure sensor 45
1.7. References 45
1.8. Other pressure sensor manufacturers 46
1.9. Bibliography 46

Chapter 2. Optical Sensors 49
Stanislav ĎAĎO and Jan FISCHER

2.1. Optical waveguides and fibers 49
2.2. Light sources and detectors 51
2.2.1. Light sources 51
2.2.1.1. Semiconductor sources of light 51
2.2.1.2. Laser diodes 53
2.2.2. Light detectors 54
2.2.2.1. Photoresistors 54
2.2.2.2. Photodiodes 54
2.2.2.3. Phototransistor 57
2.2.2.4. Position sensitive photo-detectors (PSD) 57
2.2.2.5. Charged coupled device image sensors 59
2.3. Sensors of position and movement 62
2.3.1. Position sensors using the principle of triangulation 62
2.3.2. Incremental sensors of position or displacement 63
2.3.2.1. General principles 63
2.3.2.2. Linear incremental encoder 63
2.3.2.3. Optical sensors of displacement with absolute encoding disk . . 65
2.3.2.4. Sensors with pseudorandom coding 65
2.3.3. Photoelectric switches 66
2.3.3.1. Through beam PES 66
2.3.3.2. Diffuse reflective PES 67
2.3.3.3. Retro-reflective PES 68
2.3.3.4. PES for detection of colors or color marks 70
2.4. Optical sensors of dimensions 71
2.4.1. Dimensional gauge with scanned beam 71
2.5. Optical sensors of pressure and force 73
2.5.1. Pressure sensor using the optical resonator 73
2.6. Optical fiber sensors 74
2.6.1. Introduction and classification of sensors with optical fibers 74
2.6.2. Optical fiber sensors with amplitude modulation 75
2.6.3. Sensor with wavelength modulation 77

2.6.4. Optical sensors with phase modulation 78
2.6.5. Perspective of optical fiber sensors 78
2.7. Optical chemical sensors 78
2.7.1. Introduction 78
2.7.2. Chemical sensors based on the absorbency measurement 79
2.7.3. Turbidity sensors 80
2.8. Bibliography 81
2.8.1. Books 81
2.8.2. Physical background – websites 82

Chapter 3. Flow Sensors 83
R. MEYLAERS, F. PEETERS, M. PEETERMANS and L. INDESTEEGE

3.1. Introduction 83
3.1.1. Volume flow and mass flow 83
3.1.2. Influences on the flow 85
3.1.3. Bernoulli equation 86
3.2. Flow measurements based on the principle of difference in pressure . . 88
3.2.1. The Pitot and Prandtl tube 89
3.2.1.1. Principle 89
3.2.1.2. Practical set-up 91
3.2.1.3. Characteristics 93
3.2.2. The orifice plate 93
3.2.2.1. Principle 93
3.2.2.2. Practical installation 95
3.2.3. The flow nozzle 98
3.2.4. The Venturi tube 99
3.2.5. The Dall tube 99
3.2.6. General guidelines for a correct reading 100
3.3. Flow measurements based on variable passage 101
3.3.1. The float flow meter (rotameter) 101
3.3.1.1. Principle 101
3.3.1.2. Characteristics 103
3.3.2. Target flow meter 103
3.3.2.1. Principle 103
3.3.2.2. Characteristics 104
3.4. Turbine flow meter 104
3.4.1. Principle 104
3.4.2. Practical installation 106
3.4.3. Characteristics 107
3.5. The mechanical flow meter (positive displacement) 108
3.5.1. Principle 108
3.5.2. Characteristics 110

3.6. Magnetic flow meter 110
3.6.1. Principle 110
3.6.2. Construction of the measuring instrument 112
3.6.3. Practical installation 113
3.6.4. Characteristics 115
3.7. The vortex flow meter 116
3.7.1. Principle 116
3.7.2. Construction of the vortex flow meter 117
3.7.3. Practical installation 120
3.7.4. Characteristics 121
3.8. Ultrasonic flow meter 122
3.8.1. Principle 122
3.8.2. Practical installation 125
3.8.3. Characteristics 125
3.9. Coriolis mass flow meters 126
3.9.1. Principle 126
3.9.2. Applications 128
3.9.3. Practical installation 129
3.9.4. Characteristics 129
3.10. Flow measurements for solid substances 129
3.10.1. Flow measurement of solids by means of an impact plate 130
3.10.2. Flow measurement of solids based on the weighing method 132
3.10.3. Capacitive flow measurement of solid substances 133
3.10.4. Detection of solid substances using microwaves 134
3.11. Flow measurement for open channels with weirs 135
3.12. Choice and comparison of flow measurements 137
3.13. Bibliography 137
3.14. Website references 137

Chapter 4. Intelligent Sensors and Sensor Networks 141
Jiří NOVAK

4.1. Introduction 141
4.2. Intelligent sensors 142
4.2.1. Sensors and transducers 143
4.2.1.1. Variable voltage or current source 143
4.2.1.2. Variable resistance 143
4.2.1.3. Variable impedance or mutual impedance 144
4.2.1.4. Charge generator 144
4.2.2. Signal conditioning (SC) 144
4.2.2.1. Amplification and signal conversion 145
4.2.2.2. Sensor insulation 145
4.2.2.3. Filtration 145

4.2.2.4. Detection . . . 145
4.2.2.5. Correction of non-linearity . . . 145
4.2.2.6. Correction of influence of disturbing quantities . . . 146
4.2.2.7. Sensor excitation . . . 146
4.2.3. A/D conversion . . . 146
4.2.3.1. SAR converters . . . 146
4.2.3.2. Sigma-delta modulator converters . . . 147
4.2.3.3. Flash (pipelined flash) converters . . . 147
4.2.4. Data processing . . . 147
4.2.5. Human-machine interface . . . 148
4.2.6. Communication interface . . . 148
4.2.6.1. IEEE 1451 . . . 148
4.2.7. Industrial examples . . . 149
4.2.7.1. Micronas HAL805 Hall sensor . . . 149
4.2.7.2. Yokogawa DPharp family of pressure sensors . . . 150
4.3. Sensor networks and interfaces . . . 151
4.3.1. Centralized and distributed industrial systems . . . 152
4.3.2. Hierarchical structure of distributed communication . . . 154
4.3.3. Data communication basics . . . 155
4.3.3.1. Open Systems Interconnection (OSI) model . . . 155
4.3.3.2. Physical layer . . . 157
4.3.3.3. Data link layer . . . 160
4.3.3.4. Network layer . . . 163
4.3.3.5. Transport layer . . . 164
4.3.3.6. Session layer . . . 164
4.3.3.7. Presentation layer . . . 164
4.3.3.8. Application layer . . . 164
4.3.3.9. Data distribution models . . . 165
4.3.4. Simple sensor interfaces . . . 166
4.3.4.1. Analog interfaces . . . 166
4.3.4.2. Digital interfaces . . . 167
4.3.5. Sensor networks . . . 171
4.3.5.1. AS-Interface . . . 171
4.3.5.2. CAN (Controller Area Network) and CANopen . . . 173
4.3.5.3. HART (Highway Addressable Remote Transducer) . . . 180
4.3.5.4. Foundation Fieldbus (FF) . . . 181
4.3.5.5. Interbus . . . 184
4.3.5.6. M-Bus . . . 186
4.3.5.7. Profibus . . . 188
4.3.5.8. Other standards . . . 190
4.3.6. Wireless sensor networks . . . 190
4.3.6.1. IEEE 802.15.4 . . . 190
4.3.6.2. ZigBee . . . 191

4.3.6.3. IEEE 802.15.4 and ZigBee summary 192
4.3.6.4. Other wireless standards 192

Chapter 5. Accelerometers and Inclinometers 193
André MIGEON and Anne-Elisabeth LENEL

5.1. Introduction 193
5.2. Acceleration 194
5.2.1. Physical quantity 194
5.2.2. Application to velocity and position measurements 198
5.2.3. Application to position measurements 199
5.2.4. The inclinometers 200
5.3. Application ranges 201
5.3.1. Static and low-frequency acceleration 201
5.3.2. Vibrations 202
5.3.3. Shocks 203
5.3.4. Inclination 204
5.4. Main models of accelerometers 205
5.4.1. Piezoelectric accelerometers 206
5.4.1.1. General principle 208
5.4.1.2. Accelerometers with compression 208
5.4.1.3. Shear-mode accelerometers 209
5.4.1.4. Features and limits of these accelerometers 209
5.4.2. Piezoresistive accelerometers 213
5.4.2.1. General principle 213
5.4.2.2. Silicon semiconductor strain gauges 213
5.4.2.3. Features and limits of these accelerometers 217
5.4.3. Accelerometers with resonators 219
5.4.3.1. Principle 219
5.4.3.2. Features and limits of these accelerometers 220
5.4.4. Capacitive accelerometers 221
5.4.4.1. Principle 221
5.4.4.2. Features and limits of these accelerometers 224
5.4.5. Potentiometric accelerometers 224
5.4.5.1. Principle 224
5.4.5.2. Features and limits of these accelerometers 225
5.4.6. Optical detection accelerometers 226
5.4.6.1. Principle 226
5.4.6.2. Features and limits of these accelerometers 226
5.4.7. Magnetic detection accelerometers 227
5.4.7.1. Principle 227
5.4.7.2. Features and limits of these accelerometers 228

5.4.8. Servo accelerometers with controlled displacement 229
5.4.8.1. Principle 229
5.4.8.2. Servo accelerometers with balance of torque 229
5.4.8.3. Servo accelerometers with balance of force 230
5.4.8.4. Features and limits of these accelerometers 231
5.5. The signal processing associated with accelerometers 231
5.6. Manufacturing process 232
5.6.1. The monolithic processes 232
5.6.1.1. CMOS (Complementary MOS) – BICMOS standard (Bipolar Technology and MOS) 233
5.6.1.2. CMOS – BICMOS standard + back etching 233
5.6.1.3. Above IC 233
5.6.1.4. Specific process 234
5.6.2. Hybrid process 234
5.6.3. Packaging 234
5.7. The calibrations 235
5.7.1. Inclinometers and accelerometers with range lower than 1 g 235
5.7.2. Acceleration range higher than 1 g. 235
5.8. Examples of accelerometers and inclinometers 236
5.9. List of manufacturers of accelerometers 242
5.10. References 243
5.11. Bibliography 244

Chapter 6. Chemical Sensors and Biosensors 245
Gillian McMAHON

6.1. Introduction 245
6.2. What is involved in developing a sensor? 249
6.2.1. Molecular recognition 250
6.2.2. Immobilization of host molecules 252
6.2.3. Transduction of signal 253
6.3. Electrochemical sensors 253
6.3.1. Amperometric and voltammetric sensors 254
6.3.1.1. Cyclic voltammetry 256
6.3.1.2. Hydrodynamic amperometry 257
6.3.2. Potentiometric sensors 258
6.3.2.1. Ion-selective electrodes 259
6.3.2.2. Coated-wire electrodes and polymer-membrane electrodes . . . 260
6.3.2.3. Potentiometric sensor arrays 262
6.3.3. Resistance, conductance and impedance sensors 263
6.4. Optical sensors 265
6.4.1. Methods of detection 265
6.4.1.1. Evanescent wave sensors 266

6.4.2. Reagent-mediated sensors . . . 268
6.5. Acoustic (mass) sensors . . . 269
6.5.1. Quartz crystal microbalance sensors . . . 270
6.5.2. Sensor arrays . . . 272
6.6. Biosensors . . . 274
6.6.1. Affinity biosensors . . . 275
6.6.1.1. Electrochemical transduction . . . 275
6.6.1.2. Piezoelectric transduction . . . 276
6.6.1.3. SPR biosensors . . . 278
6.6.1.4. Proteomics . . . 283
6.6.1.5. IAsys biosensor . . . 283
6.6.1.6. Miniature TI-SPR sensor . . . 284
6.6.2. Catalytic biosensors . . . 285
6.6.2.1. Electrochemical transduction . . . 286
6.6.2.2. Calorimetric transduction . . . 290
6.7. Future trends . . . 290
6.7.1. Microanalytical instruments as sensors . . . 291
6.7.1.1. Design considerations . . . 292
6.7.1.2. On-chip chromatographic and electrophoretic separations . . . 294
6.7.2. Autonomous sensing devices . . . 298
6.7.3. Sub-micron dimensioned sensors . . . 298
6.7.3.1. Microamperometric sensors . . . 298
6.7.3.2. Microelectrodes in biological systems . . . 299
6.8. Conclusions . . . 301
6.9. References . . . 302

Chapter 7. Level, Position and Distance . . . 305
Stanislav ĎAĎO and G. HARTUNG

7.1. Introduction . . . 305
7.1.1. Classification of LPD sensors . . . 305
7.2. Resistive LPD sensors . . . 306
7.2.1. Potentiometer . . . 306
7.2.2. Angular position measurement . . . 307
7.2.3. Draw wire sensors . . . 308
7.2.4. Inclination detectors . . . 308
7.2.5. Application of potentiometers . . . 309
7.3. Inductive LPD sensors . . . 309
7.3.1. Linear variable differential transformers . . . 310
7.3.2. Inductosyns . . . 311
7.3.3. Resolvers . . . 312
7.3.4. Selsyn . . . 313
7.3.5. Inductive sensors of angular velocity . . . 313

7.3.6. Eddy current distance sensors 314
7.4. Magnetic LPD sensors 315
7.4.1. Magnetic field sensors 315
7.4.2. Reed switches 316
7.4.3. Hall sensors 316
7.4.4. Semiconductor magnetoresistors 317
7.4.5. Wiegand wire 318
7.4.6. Magnetostrictive sensor 318
7.5. Capacitive LPD sensors 319
7.5.1. Introduction 319
7.5.2. Signal conditioning circuits for capacitive sensors 320
7.5.3. Using capacitive sensors 321
7.6. Optical LPD sensors 323
7.6.1. Introduction 323
7.6.2. Photo-electric switches (PES) 323
7.6.3. LPD sensors based on triangulation 327
7.6.4. Optical encoders 328
7.6.4.1. Incremental sensors 328
7.6.4.2. Absolute encoders 329
7.6.4.3. Gray code 330
7.6.5. Interferometry 330
7.6.6. Optical LPD sensors based on travel time (time-of-fly) measurement 331
7.6.7. Image-based measurement-machine vision, videometry 332
7.6.7.1. Introduction 332
7.6.7.2. Light sheet method 332
7.7. Ultrasonic sensors 333
7.7.1. Introduction 333
7.7.2. Travel time principle 334
7.7.3. Doppler effect 334
7.8. Microwave distance sensors (radar) 335
7.8.1. Introduction 335
7.8.2. Microwave sensors based on FMCW 336
7.8.3. Properties of microwave sensors 337
7.9 Level measurement 337
7.9.1. Introduction 337
7.9.2. Detection limits 338
7.9.2.1. Capacitive level switch 338
7.9.2.2. Ultrasonic switch 338
7.9.2.3. Vibrational switch 338
7.9.2.4. Conductive sensors 338
7.9.2.5. Floating switch 338
7.9.2.6. Fiber optics level switches 339

7.9.3. Continuous level measurement 339
7.9.3.1. Principles of measurement 339
7.9.3.2. Capacitive sensors 339
7.9.3.3. Ultrasonic sensors 341
7.9.3.4. Microwave sensors (radar) 342
7.9.3.5. Pressure difference (hydrostatic) sensors 342
7.10. Conclusions and trends 343
7.11. References 343
7.12. Online references 344

Chapter 8. Temperature Sensors 347
F. PEETERS, M. PEETERMANS and L. INDESTEEGE

8.1. Introduction 347
8.2. Thermal measuring techniques 348
8.2.1. Heat and temperature 348
8.2.2. Static and dynamic readings 348
8.2.3. Time constant and response time 349
8.2.4. Thermal units 349
8.2.5. Thermal equilibrium 350
8.2.6. Temperature measuring options 354
8.2.7. Quality of a measurement 355
8.3. Physical or direct temperature measurement 355
8.3.1. Glass thermometer 355
8.3.2. Liquid filled expansion thermometers 356
8.3.3. Gas filled expansion thermometer or pressure thermometer detector 358
8.3.4. Vapor-pressure systems 359
8.3.5. Bimetallic thermometer 361
8.4. Thermoelectric measurements (thermocouples) 363
8.4.1. Measuring principle: thermoelectricity 363
8.4.2. Thermoelectric laws 364
8.4.3. Practical temperature measurement with thermocouples 367
8.4.4. Technological realizations of thermocouples 371
8.4.5. Applications 374
8.4.6. Parallel and series connections of thermocouples 375
8.5. Resistance temperature detectors (RTDs) 377
8.5.1. Principle 377
8.5.2. Used materials and construction 379
8.5.3. Applications 380
8.6. Thermistors 382
8.6.1. Principle 382
8.6.2. Thermistor technology 383
8.6.3. Application 384

8.7. Monolithic temperature sensors (IC sensor). 384
8.8. Pyrometers 385
8.8.1. Introduction 385
8.8.2. Basic principles of pyrometry 386
8.8.3. Measurement possibilities for pyrometers 387
8.8.4. Implementation and construction of pyrometers. 389
8.9. References 391
8.10 Bibliography. 391

Chapter 9. Solid State Gyroscopes and Navigation 395
André MIGEON and Anne-Elisabeth LENEL
9.1. Introduction. 395
9.2. The angular rate 396
9.2.1. Definition of rate gyro 399
9.2.1.1. Comparison between a gyroscope and angular rate meter (gyrometer) 399
9.2.2. Use of rate sensors 401
9.3. Different ranges of rate gyro. 401
9.3.1. Control of trajectory 402
9.3.2. Piloting and stabilization 402
9.3.3. Guidance 402
9.3.4. Navigation 402
9.4. Main models of rate gyro. 404
9.4.1. Rotary gyrometers 404
9.4.2. Vibrating gyrometers. 404
9.4.2.1. Gyrometers with elementary or coupled bars 406
9.4.2.2. Gyrometers with a tuning fork 409
9.4.2.3. Gyrometers with coplanar interdigitated comb fingers 411
9.4.2.4. Gyrometers with vibrating shell and cylinder 414
9.4.2.5. Gyrometers with vibrating disk. 417
9.4.2.6. Gyroscopes with vibrating ring. 418
9.4.3. Optical gyrometers 420
9.4.3.1. Ring laser gyrometers 420
9.4.3.2. Fiber optic gyrometers (FOG) 421
9.4.4. Other original principles. 426
9.5. Calibration of rate sensors 426
9.6. General features of the gyrometers 427
9.7. The main manufacturers 429
9.8. References 430
9.9. Bibliography 431

Chapter 10. Magnetic Sensors 433
S. RIPKA and Pavel RIPKA

10.1. Introduction 433
10.2. Hall sensors 434
10.2.1. The Hall effect 435
10.2.2. New types of Hall sensors 437
10.2.3.1. High mobility InSb Hall elements 437
10.2.3.2. Integrated Hall sensors 437
10.3. AMR sensors 439
10.3.1. Operating principles of the AMR effect 439
10.3.1.1. Geometrical linearization of the AMR 441
10.3.2. Measuring configuration of the AMR 443
10.3.3. Flipping 444
10.3.4. Magnetic feedback 446
10.4. GMR sensors 447
10.4.1. Physical mechanism 450
10.4.2. Spin valves 450
10.4.3. Sandwiches and multilayers 453
10.4.3.1. Temperature characteristics 453
10.4.3.2. Cross-field error 453
10.4.3.3. Unpinned sandwich 453
10.4.3.4. GMR multilayer 454
10.4.4. SDT sensors 454
10.4.5. Linear GMR sensors 454
10.4.5.1. Bipolar response using biasing coils 456
10.4.5.2. GMR gradiometer 456
10.4.6. Rotational GMR sensors 456
10.5. Induction and fluxgate sensors 457
10.5.1. Induction coil sensors 458
10.5.2. Fluxgate sensors 459
10.5.2.1. Core shapes of fluxgates 461
10.5.2.2. Double-rod sensors 461
10.5.2.3. Ring-core sensors 461
10.5.2.4. Race-track sensors 461
10.5.2.5. Principles of fluxgate magnetometers 462
10.6. Other magnetic field sensors 463
10.6.1. Resonance sensors 463
10.6.1.1. Magnetic sensors based on electron spin resonance (ESR) . . . 464
10.6.1.2. Overhauser magnetometers 465
10.7. Magnetic position sensors 465
10.7.1. Sensors using permanent magnets 465
10.7.1.1. Induction position sensors 465

10.7.2. Eddy current sensors 466
10.7.3. Linear and rotational transformers 467
10.7.3.1. Linear transformer sensors. 467
10.7.3.2. Rotation transformer sensors 468
10.7.4. Magnetostrictive position sensors 469
10.7.5. Proximity switches 469
10.7.5.1. Reed contacts 470
10.7.5.2. Wiegand sensors. 470
10.8. Contactless current sensors 471
10.8.1. Hall current sensors 472
10.8.2. Magnetoresistive current sensors 472
10.8.3. AC and DC transformers. 472
10.8.4. Current clamps. 472
10.9. References. 473

Chapter 11. New Technologies and Materials. 477
A. TIPEK, P. RIPKA and E. HULICIUS, with contributions from
A. HOSPODKOVÁ and P. NEUŽIL

11.1. Introduction: MEMS. 477
11.2. Materials. 480
11.2.1. Passive materials 480
11.2.2. Active materials 481
11.2.3. Silicon. 481
11.2.4. Other semiconductors 483
11.2.5. Plastics 484
11.2.6. Metals 486
11.2.7. Ceramics 486
11.2.8. Glass. 486
11.3. Silicon planar IC technology 487
11.3.1. The substrate: crystal growth 488
11.3.2. Diffusion and ion implantation 488
11.3.3. Oxidation 489
11.3.4. Lithography and etching 489
11.3.5. Deposition of materials. 490
11.3.6. Metallization and wire bonding 490
11.3.7. Passivation and encapsulation. 491
11.4. Deposition technologies. 491
11.4.1. Introduction 491
11.4.2. Chemical reactions 492
11.4.3. Physical reactions 495
11.4.4. Epitaxial techniques for semiconductor device preparation 498

11.5. Etching processes 500
11.5.1. Wet etching/micromachining 501
11.5.2. Dry etching/micromachining 502
11.6. 3-D microfabrication techniques. 503
11.6.1. LIGA 504
11.6.2. Laser assisted etching (LAE) 504
11.6.3. Photo-forming and stereo lithography 505
11.6.4. Microelectrodischarging (MEDM and WEDG) 506
11.6.5. Microdrip fabrication 507
11.6.6. Manufacturing using scanning probe microscopes and electron microscopes. 508
11.6.7. Handling of micro particles with laser tweezers 508
11.6.8. Atomic manipulation 509
11.7. References. 510

List of Authors 513

Index 515

Chapter 1

Pressure Sensors

1.1. Introduction

Together with temperature, pressure is one of the most important physical quantities in our environment. Pressure is a significant parameter in such varied disciplines as thermodynamics, aerodynamics, acoustics, fluid mechanics, soil mechanics and biophysics. As an example of important industrial applications of pressure measurement we may consider power engineering. Hydroelectric, thermal, nuclear, wind and other plants generating mechanical, thermal or electrical energy require the constant monitoring and control of pressures: overpressure could cause the deterioration of enclosures or drains and cause very significant damage.

As a significant parameter, pressure enters into the control and operation of manufacturing units that are automated or operated by human operators. Pressure measurement is also used in robotics, either directly in controls or indirectly as a substitute for touch (artificial skin for example), for pattern recognition or for determining strength of grip. All these activities require instrument chains in which the first element is the pressure sensor, delivering data relating to the pressure of compressed air, gas, vapor, oil or other fluids, determining the correct operation of machines or systems.

The variety of mentioned applications demands a great diversity of sensors. This diversity also derives from the fact that pressure covers a very wide range from ultra-high vacuums to ultra-high pressures. It can be expressed as an absolute value (compared to vacuum) or as a relative value (compared to atmospheric pressure); it

---

Chapter written by André MIGEON and Anne-Elisabeth LENEL.

can also represent a difference between two pressures or relate to various media and fluids whose physical characteristics (e.g. temperature) or chemical characteristics (e.g. risk of corrosion) are very varied. Pressure units are summarized in Table 1.1.

1.2. Pressure

In what follows, we will consider the different physical characteristics necessary to understand pressure sensors: pressure as a physical quantity, and various sensor models with absolute, relative or differential pressure sensors. We will take a brief look at the physical properties of fluids.

1.2.1. *Pressure as a physical quantity*

1.2.1.1. *Static pressure*

From a phenomenological point of view, pressure, p, as a macroscopic parameter is defined starting with element of force $d\vec{F}$, exerted perpendicularly on an element of surface $d\vec{A}$ of the wall, by the fluid contained in the container:

$$p = dF / dA \tag{1.1}$$

The element of force $d\vec{F}$ caused by pressure p is perpendicular to the element of surface $d\vec{A}$.

For pressure p inside the fluid with free surface we may write:

$$p = p_0 + \rho gh \tag{1.2}$$

p_0: atmospheric pressure

ρgh: hydrostatic pressure

ρ: density

g: acceleration of gravity at the place of measurement

h: distance from the free surface

1.2.1.2. *Units*

	pascal (Pa)	**bar** (bar)	**atmosphere (Atm)**	**Comments**
1 pascal	1	10^{-5}	$9.869\ 10^{-6}$	Standard International Unit
1 bar	10^{5}	1	$9.869\ 10^{-1}$	1 Bar is standard atmospheric pressure
1 kg/cm^2	$9.8039\ 10^{4}$	$9.803\ 10^{-1}$	$9.86\ 10^{-1}$	Old Unit
1 atmosphere	$1.013\ 25\ 10^{5}$	1.0133	1	Normal Atmospheric Pressure
1 cm of water	98.04	$9.80\ 10^{-4}$	$9.68\ 10^{-4}$	
1 mm of Hg	$1.33\ 10^{2}$	$1.333\ 10^{-3}$	$1.316\ 10^{-3}$	For an Hg density of 13.59593 kg/ dm^3. 1 mmHg is also called Torr
1 inch Hg	$3{,}386\ 10^{3}$	$3{,}386\ 10^{-2}$	$3{,}342\ 10^{-2}$	
1 psi	$6.890\ 10^{3}$	$6.89\ 10^{-2}$	$6.89\ 10^{-2}$	Pound per Square Inch

Table 1.1. *Units of pressure*

1.2.2. *Absolute, relative and differential sensors*

An absolute pressure sensor measures static, dynamic or total pressure with reference to a vacuum (see Figure 1.1).

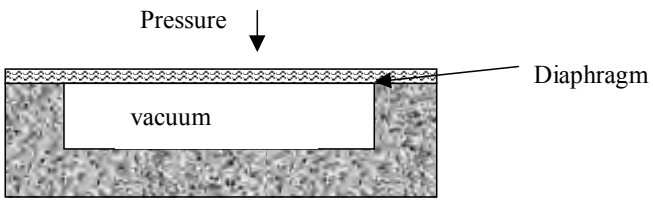


Figure 1.1. *Absolute pressure sensor*

A relative pressure sensor measures static, dynamic or total pressure with reference to ambient atmospheric pressure (Figure 1.2).

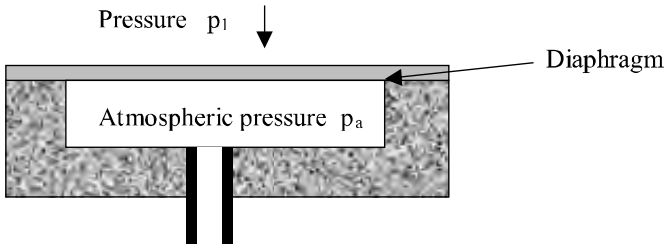


Figure 1.2. *Relative pressure sensor*

A sealed relative pressure sensor measures static, dynamic or total pressure with reference to ambient atmospheric pressure, sealed at the time of manufacture of the sensor (see Figure 1.3).

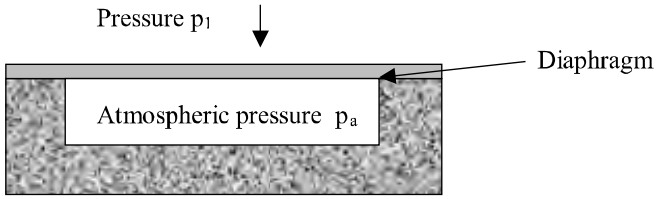


Figure 1.3. *Sealed Relative pressure sensor*

A differential pressure sensor measures a static, dynamic or total pressure with reference to an unspecified variable pressure p_2 (Figure 1.4).

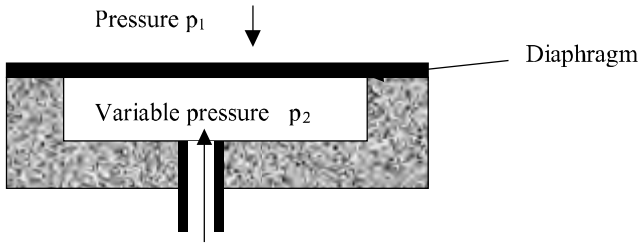


Figure 1.4. *Differential pressure sensor*

1.2.3. *Fluid physical properties*

In static fluids, the pressure force F is exerted on the surface originates only from the random kinetic energy of molecules. In dynamic fluids force F originates from the random and directed kinetic energy of the molecules.

We generally distinguish between two main fluid families: gases and liquids.

1.2.3.1. *Liquids*

The total pressure is the sum of the static pressure, the pressure due to external forces and the dynamic pressure. This has the same value in all points for a fluid moving horizontally (incompressible, negligible viscosity, like liquids), following Bernoulli's theorem:

$$p_t = p_s + p_d = p_s + \frac{1}{2}\rho v^2 \tag{1.3}$$

with:

p_t: total pressure

p_s: static pressure

p_d: dynamic pressure

v: local velocity

ρ: density

1.2.3.2. *Gases*

The pressure of a gas in a tank is the force exerted by gas on the walls of the tank per unit of area. When a tank contains a mixture of gases, we can define a partial pressure for each of them. The sum of the partial pressures is equal to the total pressure. The equation of an ideal gas is:

$$pV = nk_BT \tag{1.4}$$

p: pressure

n: number of molecules

T: temperature

V: volume

k_B: Boltzmann constant

According to the kinetic theory, the molecules of a gas are driven in a continual and random manner and bump into each other. The trajectory of a molecule between two shocks is a right-hand side segment traversed at constant speed and the direction of a segment after a shock has no correlation with the direction of the segment before the shock. The trajectory of a molecule is therefore a broken line, the average value l of the length of its segments being the free mean course.

When the gas is contained in an enclosure, the molecules also have collisions with the walls and the pressure that they exert on them results from the average effect of these collisions.

A *vacuum* is often characterized by the *Knudsen number:*

$$K = \lambda / l \tag{1.5}$$

K: Knudsen number

λ: mean free course

l: enclosure dimension

1.2.3.3. *Sensor pneumatic connection influence*

When measuring pressure with very slow changes in stationary fluids, there are no problems except that the connection must be leak-proof and free of contaminating material. When the fluid is moving (even when its pressure stays constant) and/or the pressure is changing relatively fast, the dynamic response of the tube connection in the sensor can significantly influence the pressure seen by the sensor in amplitude and phase.

1.3. Pressure ranges

1.3.1. *Vacuum and ultra-vacuum*

The term vacuum gauges refers to sensors for the measurement of gas pressure much lower than normal atmospheric pressure. The interesting parameter is the average number of molecules contained pre unit of volume. Traditionally, four pressure ranges are used in setting up the scale of a vacuum (Table 1.2).

	Primary vacuum (or rough)	**Intermediate vacuum (or medium)**	**High vacuum (or advanced)**	**Ultra-vacuum**
Approximate range of pressure	10^5 to 10^2 Pa	10^2 to 10^{-1} Pa	10^{-1} to 10^{-5} Pa	< 10^{-5}Pa
	10^3 to 1 mbar	1 to 10^{-3} mbar	10^{-3} to 10^{-7} mbar	< 10^{-7} mbar
Number of molecules per cm^3	10^{19} to 10^{16}	10^{16} to 10^{13}	10^{13} to 10^9	< 10^9
Free mean course	10^{-6} to 10^{-3}cm	10^{-3} to 1 cm	1 to 10^4 cm	> 10^4 cm
Mode of flow	rolling (or viscous)	intermediate	molecular	–

Table 1.2. *Various vacuum fields*

Various vacuum gauges

The main principle of primary vacuum measurement derives from a heating effect. For high vacuum cases, the principle uses the property of ionization. Vacuum gauges fit into three principal groups according to their physical effect (Table 1.3):

– mechanical effect gauges: the sensing element becomes deformed under the influence of pressure;

– heating effect gauges: the sensing element is a heated element whose temperature depends on the surrounding pressure;

– gauges using an electrical characteristic of a gas: measurement relates directly to the gas. The molecules are counted by counting the number of ions they provide for an electrical current.

Type of conversion	Physical principle	Applications
Mechanical	Bourdon tube gauge	Very low cost, low accuracy. Recommended for static installations.
	Active diaphragm gauge	Low cost, low accuracy, fast, durable. Recommended for rough vacuum.
	Piezoresistive diaphragm gauge	Industrial vacuum measurements especially for vacuum safety systems.
Electrical (see section 1.4.3.1)	Ionization	Recommended for vacuum measurements up to 10^{-8} Pa in relatively protected environments like clean rooms or laboratories. Relatively bulky.
Thermal (see section 1.4.3.2)	Pirani gauge	Particularly sensitive gauges recommended for very deep vacuum measurements in relatively protected environments. Fragile.
	Thermocouple gauge	Low precision, small size.

Table 1.3. *Different conversion types used for vacuum measurement*

1.3.2. *Middle range pressure*

Average pressures usually lie in the range 10^2 Pa to 10^8 Pa. This pressure range occurs in the majority of industrial applications. All these principles first transfer pressure into mechanical deformation and/or stress that is then measured (see section 1.4). Table 1.4 shows the different types of conversion into electric signals used in mid-range pressures.

Type of conversion	Physical principle	Applications
Resistance variation (see section 1.4.2.1)	Piezoresistive diffused (strain) gauge	Low cost, compact. Many general applications like altimetry, barometry, process monitoring, safety, etc.
	Gauge with taut wire	Laboratory instrumentation.
	Gauge with manganin wire	Extrusion presses.
	Potentiometer for low pressure	Limited lifespan and hysteresis. It is used especially for low pressure measurements (a few bars) in static barometry.
	Extensometric foil strain gauge	Many specific applications which require limited quantities of sensors.
	Gauges with deposited film	Rugged sensors used in harsh environments such as aerospace, transport, energy (liquid gases).
Capacitance variation (see section 1.4.2.2)	Standard capacitive pressure sensor	Very wide range sensors. Pressure measurements in a harsh mechanical or thermal environment.
Inductance variation (see section 1.4.2.3)	Capacitive film sensors Sensor with Electret effect	In-wall pressure measurements in a harsh mechanical or thermal environment. Analysis of very fast pressure variations.
	Inductance and mutual inductance	Sensitive to vibrations.
Electromechanical oscillator (see section 1.4.2.5)	Oscillator with quartz	Excellent stability but requires numerical corrections; onboard anemobarometry on aircraft: digital avionics.
	Oscillator with vibrating tube or blade	Secondary or transfer standards, anemobarometry onboard aircraft.
Optical (see section 1.4.2.6)	Photo electricity	Laboratory measurements.
	Optical fiber	Remote measurement in harsh environments, e.g.: oil industry, energy industry, refineries, engines, etc.
Servo controlled sensors (see section 1.4.2.7)	Balance of force	Precision laboratory measurements. Anemobarometric measurements. Precision measurements of static pressures.

Table 1.4. *Different types of conversion used for mid-range pressures*

1.3.3. *High pressure*

The field of high and very high pressures relates to the pressures beyond 10^8 Pa. The measured fluids are almost exclusively liquids. When making these measurements, the physical principles of conversion into electric signals are the same as those used for measurements of average pressures, but the sensing element and the packaging of these sensors are very specific.

Table 1.5 explains the different types of conversion which are used for the measurement of high pressure.

Type of conversion	Physical principle	Applications
Piezoelectric effect (see section 1.4.2.4)	Piezoelectricity	Measurements of high pressures in instrumentation on test benches or production machine tools. For dynamic measurements (response time close to the millisecond). Measured pressure in injection molds.
Electromechanical oscillator (see section 1.4.2.5)	Surface waves	High absolute precision and excellent stability but require numerical corrections, hence their use in systems including microprocessors.

Table 1.5. *Types of conversion for high pressure measurement*

1.4. Main physical principles

Initially it must be noted that it is not easy to measure pressure directly from its action on the properties of a particular material. The sensitivity obtained in this case is extremely low and the performance poor. The only advantage is the very low cost. Therefore, the great majority of pressure sensors are “composite sensors” (Figure 1.5).

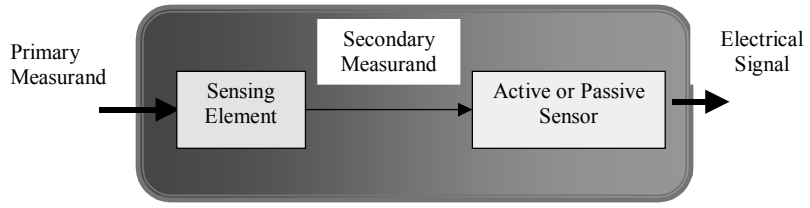


Figure 1.5. *Principle of a composite sensor*

The sensing element is the device which ensures initial translation of the pressure (primary measurand) into another non-electric physical quantity, the secondary measurand. The latter is translated by another sensor into an electric signal.

1.4.1. *The sensing device*

In the case of pressure p, the sensing element is designed to generally provide:

– a deformation and then a displacement;

– a force;

– a strain.

Typically, the most widely used sensing element is the welded diaphragm with effective section S which can be planar, corrugated, cylindrical or a more complex geometric form according to the pressure range or the fluid under consideration (see Table 1.6).

o Embedded diaphragm
o Piston with spring
o Corrugated diaphragm
o Open manometric cell
o Closed manometric cell
o Biconical cell
o Bellows
o Bourdon tube
o Helical twisted tube
o One-eyed tube

Table 1.6. *Examples of sensing elements*

The difficulty with pressure sensors lies primarily in choosing the best compromise between:

– Price.

– Performance.

– Production technology.

– Used materials.

Microelectronic technology adapted to micro systems allows bold, highly integrated and very economic designs. In addition, the progress made in the quality of materials, and the increasing power of data processing, allows the simplification of the geometry of the sensing element. Thus, most pressure sensors today use cylindrical or planar sensing elements (diaphragm).

The materials most often used for the production of sensing elements include the following:

o Stainless steel 17-4 pH
o Stainless steel 316
o Hastelloy
o Monel
o Inconel
o Titanium
o Ni Span C
o Quartz
o Silicon
o Sapphire

Table 1.7. *Examples of constructional materials for sensing elements*

The different geometries of sensing elements are summarized in Figure 1.6.

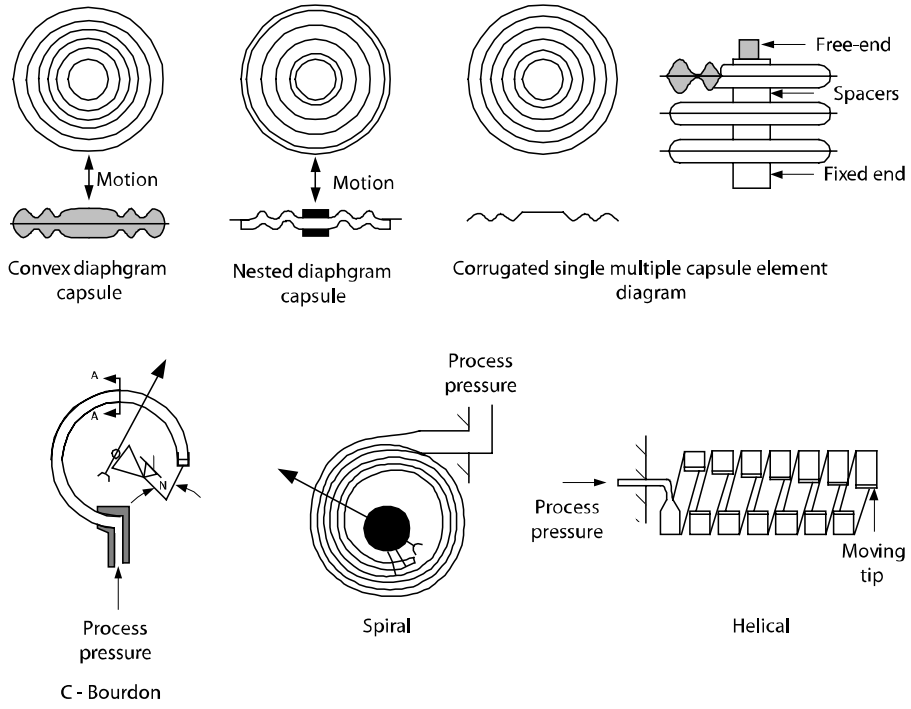


Figure 1.6. *Different sensing element geometry [1]*

1.4.2. *Sensors with elastic element*

1.4.2.1. *Conversion by resistance variation*

1.4.2.1.1. Potentiometer

The wiper of a potentiometer is connected to a diaphragm, a Bourdon tube or cell so that the deformation of this sensing element causes a displacement of the wiper (Figure 1.7).

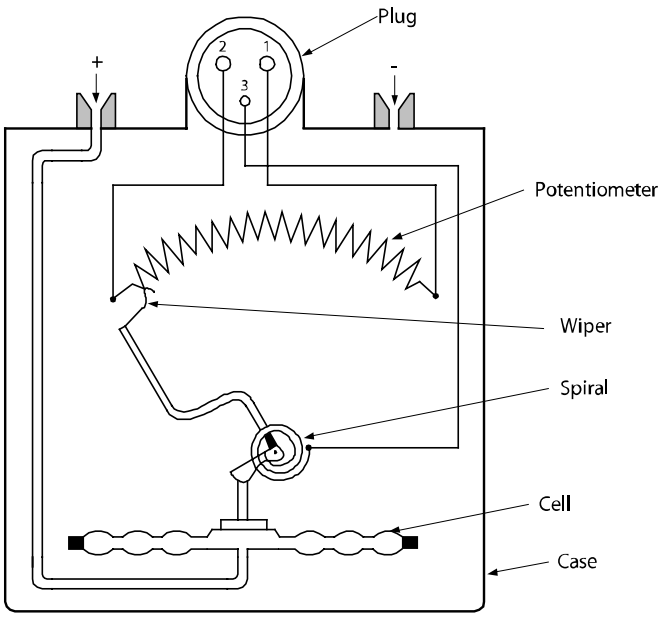


Figure 1.7. *Differential pressure sensor with a potentiometer [2] SFIM*

For an unloaded potentiometer with total resistance R_n, supplied with a source of voltage V_s, voltage V_m between the wiper and one of its ends is:

$$V_m = V_s.R(x)/R_n \tag{1.6}$$

where

R(x): resistance between the wiper and the end of the potentiometer

R_n: total resistance

V_s: supply voltage

V_m: voltage between the wiper and one of its ends

If there is proportionality between:

– pressure p to be measured and deformation of the sensing element;

– deformation of the sensing element and displacement x of the wiper;

– displacement of the wiper and the resistance R(x);

then we may write:

$$V_m = k. V_s .p \tag{1.7}$$

where k is a characteristic constant of the device.

Table 1.8 indicates the advantages and disadvantages of such a principle:

ADVANTAGES	DISADVANTAGES
– high output signal level (no need for amplifier) – low cost – technological robustness – adaptable to many applications	– high hysteresis – sensitive to vibrations – moving contact: wear, contact resistance

Table 1.8. *Advantages and disadvantages of potentiometers*

1.4.2.1.2. Metal strain gauges

Foil-type (piezoresistive) strain gauges are still very widely used. A resistive grid is created on foil glued to the sensing element. The measured pressure induces deformation, which causes change of resistance. If four such sensors are properly connected in a Wheatstone Bridge, temperature compensation and increase of sensitivity are achieved (Figure 1.8). The inner gauges measure tangential strain, while the outer gauges measure radial stress, which has opposite polarity.

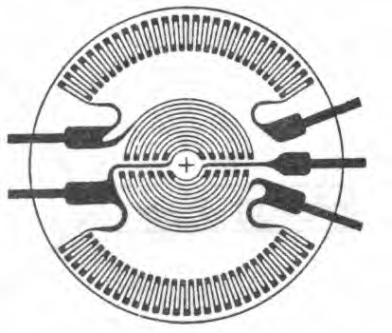


Figure 1.8. *Metal strain gauge for pressure sensors*

Table 1.9 indicates the advantages and disadvantages of such sensors.

ADVANTAGES	DISADVANTAGES
– resistant to vibrations – low cost – great adaptability with various technology applications – simple to implement	– problems attaching the foil gauges to the sensing element – low gauge factor

Table 1.9. *Advantages and disadvantages of sensors with foil strain gauge*

1.4.2.1.3. Gauges with deposited film

To overcome the problems involved in the support of the gauge and its attachment that are causes of instability, we directly deposit a resistive layer on the wall of the sensing element. This deposition is carried out either by sputtering (several techniques can be used) to obtain "thin layer" gauges or by screen-printing to obtain "thick layer" gauges (see Table 1.10).

Technology	Gauge K factor	Long-term stability
Metallic thin layers	2 to 4	excellent
Resistive thick layers	10 to 20	very good
Semiconductor	100	poor

Table 1.10. *Characteristics of gauge and stability of various technologies*

Table 1.11 indicates the advantages and disadvantages of such sensors.

ADVANTAGES	DISADVANTAGES
– resistant to vibrations – good stability – low cost – relatively simple technology	– sensitive to electric overloads – average integration potential, especially for thick layer gauges

Table 1.11. *Advantages and disadvantages of sensors with deposited screen*

1.4.2.1.4. Gauges with diffused piezoresistors

These gauges use microelectronics technologies directly, allowing the use of silicon as sensing element. The sensing element is made of single-crystalline silicon. A piezoresistor (type N) is created by diffusion of dopands onto a specific region of a type P silicon substrate. The PN junction also forms a diode. This sensor has the advantage of having very high sensitivity and good miniaturization.

The first gauges of this type had significant thermal drifts and were limited at high temperatures (> 125°C). Much progress has been made by introducing insulating layers between the gauge and the substrate, while preserving a full single-crystal structure. The process uses silicon substrates of the SIMOX type (see Figure 1.9).

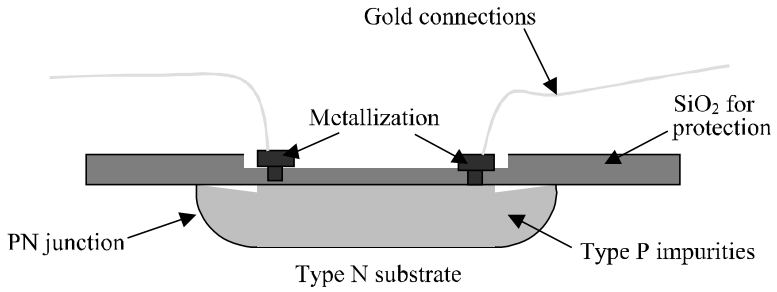


Figure 1.9. *Gauge with diffused piezoresistors*

ADVANTAGES	DISADVANTAGES
– high gauge factor – high potential for integration: possibility of producing a diaphragm 1 mm in size – cost reduced by mass production	– operation temperature limited to approx. 120°C

Table 1.12. *Advantages and disadvantages of gauges with diffused piezoresistors*

1.4.2.1.5. Taut wire gauges

A taut wire is secured between the sensing element (diaphragm) and a rigid support (the sensor case). When the sensing element is exposed to pressure, the wire resistance changes proportionally.

These sensors have good sensitivity, but they are rarely used because of their brittleness and high sensitivity to vibration and shock.

1.4.2.1.6. Industrial examples

SERIES 9 model from KELLER

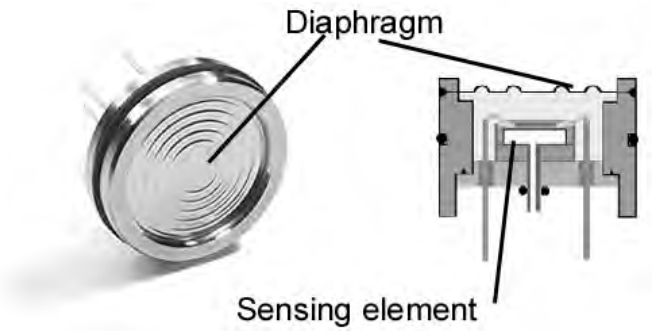


Figure 1.10. *A SERIES 9 piezoresistive sensor model from KELLER [3]*

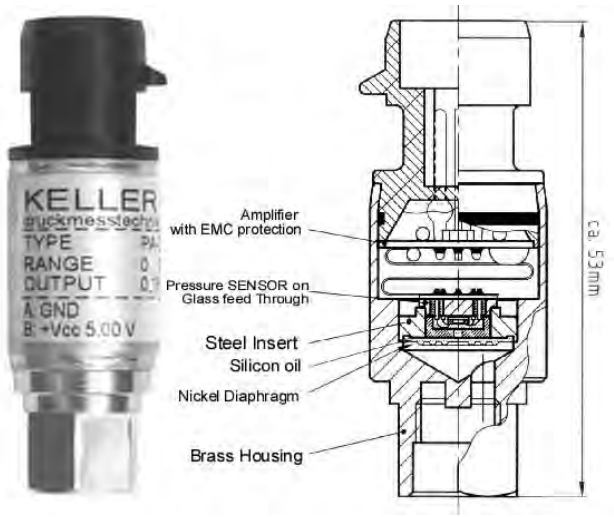


Figure 1.11. *A SERIES 9 piezoresistive sensor model from KELLER [3]*

Piezoresistive OEM **Pressure Transducers Series 9**	**KELLER**
The Series 9 pressure sensor is the most economic version for pressure ranges of from 100 mbar to 200 bar. The standard version is supplied with connecting pins (leadouts are fitted only on request) and the serial number is not engraved. Typical Applications: altitude measurement, aviation electronics, meteorology, servo controls, robotics, hydraulics, sanitary and pharmaceutical engineering, underground mining, injection engineering, etc.	
SPECIFICATIONS (at 4 mA excitation)	
Pressure ranges (FS) Linearity Stability Operating temperature range Temperature-coefficients of: – zero (without comp.) – sensitivity	0.2...200 bar (abs./rel.) typ. < 0.5% FS max. < 1.0% FS 0.5 mV typ 2 mV max –10…80°C (optionally) < 0.1 mV/°C max < 0.2 mV/°C < 0.01%/°C max < 0.03%/°C

Table 1.13. *A SERIES 9 piezoresistive sensor model from KELLER [3]*

SERIES 5 model TAB from KELLER

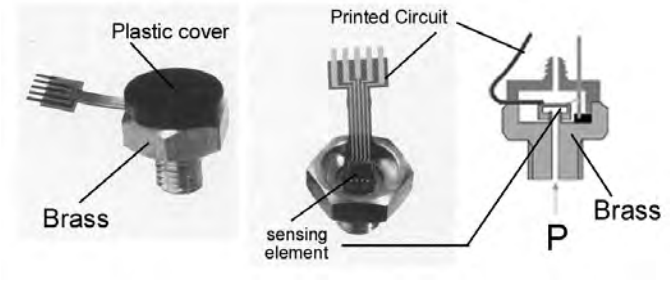


Figure 1.12. *A SERIES 5 piezoresistive sensor model TAB from KELLER (picture by www.keller-druck.com [3])*

Series 5 TAB (Standard) *OEM gauge pressure sensor* *Low cost sensor*	**KELLER**
This pressure sensor has been developed for high volume OEM applications. The pressure media acts on the rear side of the silicon chip, with the diffused strain gauges on the front. The sensor can be used for wet and aggressive media. The piezoresistive sensor is bonded to the brass housing and pressure port. A flexible printed circuit "TAB" (Tape Automated Bonding) connects the resistors on the chip and is sandwiched between the housing and the plastic cover. The free-ends of the TAB can be soldered directly onto a printed circuit or may be extended by wires. The plastic cover (optionally with tube connection) protects the sensor element and holds the flexible print. The transducers are fully tested for function and sensitivity, and (optionally) compensated for thermal zero.	
SPECIFICATIONS (at 4 mA excitation)	
Pressure ranges (FS)	1-20 bar
Linearity	0.25% FS typ. 0.5 mV typ.
Stability	2 mV max.
Operating temperature range	–10…80°C (optionally)
Storage temperature	–20…100°C
Temperature-coefficients of:	
– zero (without comp.)	0.05 mV/K typ. 0.2 mV/K max.
– sensitivity	0.01%/K typ. 0.02%/K max.

Table 1.14. *A SERIES 5 piezoresistive sensor model TAB from KELLER [3]*

1.4.2.2. *Conversion by capacitance variation*

Usually simpler in their principle, pressure sensors using conversion by capacitance variation are relatively robust. One of the capacitor electrodes is connected to a sensing element such as a diaphragm. The variable parameter can be effective area A of the plates as a linear function of the displacement ΔX. More often the variable is the distance d. There are many production geometries based on this principle – Figures 1.13 and 1.14 show one example.

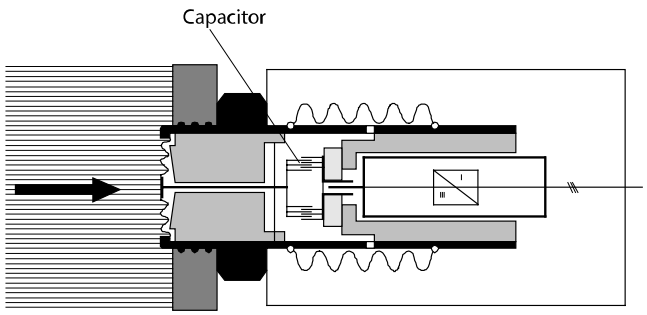


Figure 1.13. *Pressure sensor with variable effective area (after VEGA [4])*

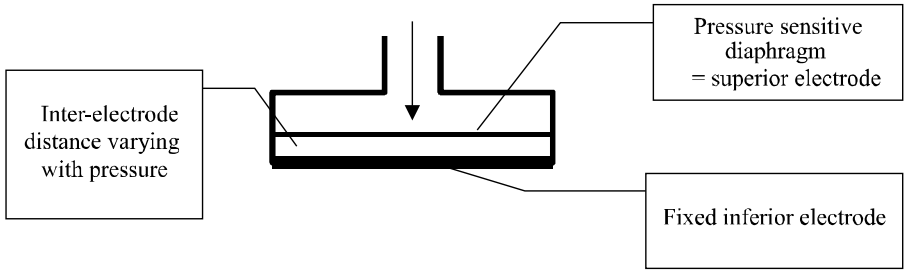


Figure 1.14. *Diagram of a pressure sensor with capacitance conversion*

1.4.2.2.1. Standard capacitive pressure sensors

Capacitance pressure transducers were originally developed for measuring vacuums. Figure 1.15 shows a traditional bridge circuit for capacitance pressure sensor.

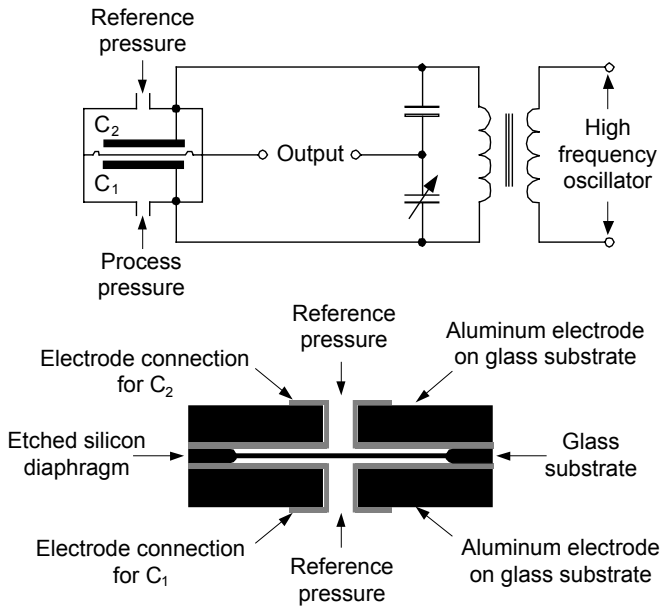


Figure 1.15. *Capacitance-based pressure cell [1]*

In microsensors, the diaphragm is usually micro-machined monocrystalline silicon. The capacitive transducer can be either an absolute gauge or a differential pressure transducer.

Capacitance pressure transducers may cover pressure ranges from high vacuum to 70 MPa. They have smaller drift compared to strain-gauge transducers.

ADVANTAGES	DISADVANTAGES
– compactness	– sensitive to stray capacitance
– low drift	– sensitive to vibrations
– high bandwidth	

Table 1.15. *Advantages and disadvantages of standard capacitive pressure sensors*

1.4.2.2.2. Capacitance thin – film sensors

These sensors use changes in ε: the relative permittivity of the dielectric placed between the two electrodes. Very thin capacitance pressure microsensors with solid dielectric or gas dielectric (approx. 80 μm) were developed by ONERA (France). These sensors are intended for dynamic pressure measurement, i.e. sudden changes of pressure (Figure 1.16).

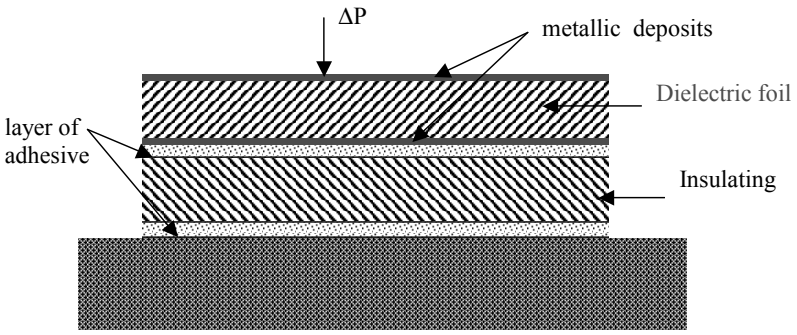


Figure 1.16. *Principle of a capacitance thin – film (pellicular) sensor*

ADVANTAGES	DISADVANTAGES
– compactness – resistant to vibrations – high bandwidth: 50 to 200 kHz	– sensitive to temperature

Table 1.16. *Advantages and disadvantages of pellicular sensors*

To avoid the use of an external power supply, we can use a diaphragm preserving a constant electric polarization (electret effect). The electret effect is also used in microphones, which are in fact sensitive pressure sensors.

1.4.2.2.3. Industrial example

Model PTA 427 analog barometer from VAISALA

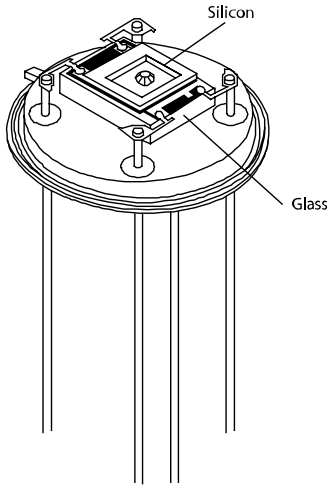


Figure 1.17. *Barocap silicon capacitance pressure sensor (VAISALA [5])*

Barocap analog barometer	***VAISALA***
Capacitance pressure sensor based on silicon membrane is bonded to glass substrate. Linearity of +/-0.3 hPa from 800 to 1,060 hPa. A wide pressure range version adjustable from 600 to 1,060 hPa has also been developed. The temperature dependence +/-0.02 hPa/°C at 1,000 hPa. Total accuracy +/-0.5 hPa at room temperature and +/-2.0 hPa over the temperature range from –20 to +60°C. When a very accurate measurement over a wide pressure and temperature range is required, additional temperature compensation must be used. The PTA 427 is used in many industrial and medical applications.	

Table 1.17. *Model PTA 427 analog barometer from VAISALA [5]*

MODEL P165 from KAVLICO

Pressure Transducer MODEL P165	***KAVLICO***
The *P165 pressure* transducer utilizes *ceramic capacitive sensing technology*. Critical operating parameters such as zero, span and linearity are assured through computer controlled laser trimming of the hybrid circuits which are joined to the ceramic capacitive sensor under tightly monitored conditions. The P165 is available in standard pressure ranges of 0-15 to 0-1,000 psia/psig with custom ranges optional. The sensor provides an amplified voltage output while typically operating on 5 Vdc. Linearity 0.5%, hysteresis and repeatability 0.05%. Temperature coefficient of sensitivity: 0.02%/K. The transducer has EMI protection and can be fitted with a plated carbon steel, brass or stainless steel housing. The sensor can withstand high overpressures up to 5x the rated pressure. Operating temperature span is –40°C to +125°C. A variety of O-ring seal materials are available in order to conform to process media requirements. *Typical applications* include: diesel engines, spark ignition, natural gas, CNG and propane engines, industrial compressors, refrigeration and HVAC systems, depth and level measurements, hydraulic fluid pressure and general industrial pressure monitoring.	

Table 1.18. *Model P165, capacitive sensor from KAVLICO [6]*

Figure 1.18. *Model P165, capacitive sensor from KAVLICO [6]*

1.4.2.3. *Conversion by inductance variation*

These pressure sensors use a variation of the reluctance of a magnetic circuit, by changing one or several of its air-gaps. It is also possible to obtain a variation of the reluctance of a magnetic circuit by using the magnetic properties of the sensing element material. Their linearity can be improved with differential transformers.

The signal translates the amplitude and the direction of the displacement of the core. The core is linked to a diaphragm, a capsule or bellows exposed to pressure or a pressure difference.

Figure 1.19 shows the most popular configuration which uses an LVDT position sensor. The capsule, on which the pressure is exerted, drives a moving core that varies the inductive coupling between the LVDT transformer primary and secondary winding. Table 1.19 shows the advantages and disadvantages of such sensors:

ADVANTAGES	DISADVANTAGES
– very good resolution	– sensitive to vibrations and shocks
– good stability	– sensitive to large magnetic field
– economic	
– large output signal	

Table 1.19. *Advantages and disadvantages of sensors with inductance variation*

1.4.2.3.1. Industrial example

Model P3000 Series – LVDT - designed for Very Low Pressure measurement from Schaevitz.

Parameters:

– combined nonlinearity, hysteresis and non-repeatability: 0.5% FS

– thermal effects (combined offset and hysteresis): 0.02%/K

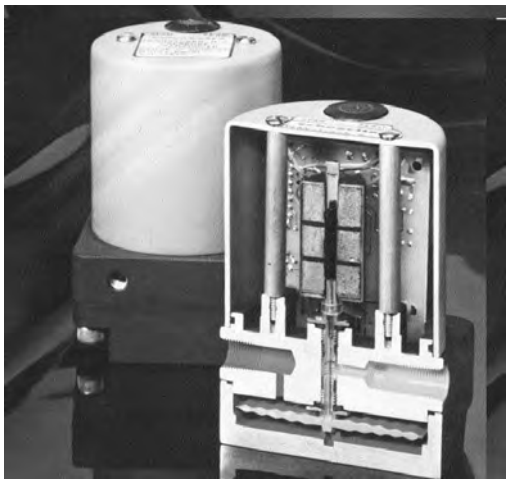


Figure 1.19. *Photograph: Inside of the model P3000 series from Schaevitz [7]*

1.4.2.4. *Conversion by piezoelectric effect*

The piezoelectric structures used as sensing element directly transform the strain, produced by the applied force F, into an electric charge q. These sensors are used for measuring pressure changes in time but not for the static pressure as the electric signal is produced only when a stress is changing.

Thus, a small plate cut from a quartz crystal, perpendicular to one of its three electrical axes, provided with metal electrodes, develops dielectric polarization by compression or extension resulting in the appearance of a charge q on the electrodes. The surface of the disks or plates is determined according to the acceptable maximum strain, depending on the nature of the sensor material (quartz, PVDF, Barium titanate, seignette salt).

However, the applicable ultimate strain depends primarily on the quality of contact between the crystal and electrodes. To this end, parallelism of the faces must be ensured to within 10 μm and flatness to within 1 μm. Only the optical polishing and neat grinding of surfaces will remove the irregularities capable of strain concentration, possibly exceeding the breaking load.

The tubular form makes it possible to increase the load by simplifying the mode of association of the elements. The tube, like a bi-strip, is formed by the association of two elements of opposite polarity compared to its symmetry plane. Tubular structures are, in particular, usable for the production of pressure sensors cooled by water circulation in contact with the metallization of the crystal and the diaphragm.

The pressure transmission is ensured by a rigid metallic component also used for attaching the diaphragm. This piece is extended by a stem, which, with a strong return spring, applies an initial tension or pre-strain improving linearity. Using this initial tension, we can also measure pressures lower than atmospheric pressure (see Figure 1.20).

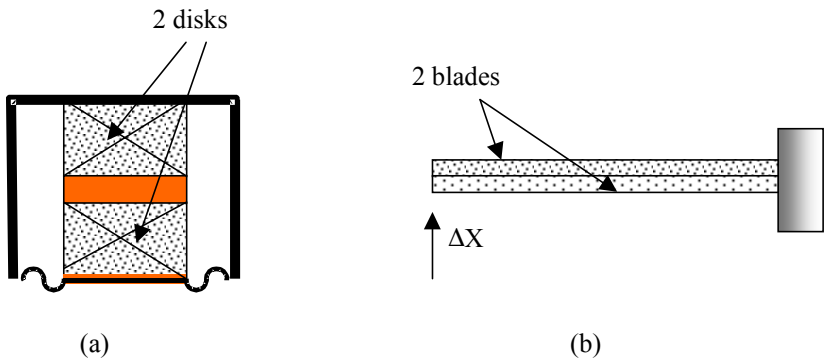


Figure 1.20. *Piezoelectric principle a) disks b) bi-strip*

Piezoelectric sensors can be quite easily miniaturized to a few millimeters.

Table 1.20 indicates the advantages and disadvantages of such sensors.

ADVANTAGES	DISADVANTAGES
– large bandwidth	– high sensitivity to temperature
– possible miniaturization	– processing of low-level signals is necessary
– not very sensitive to accelerations	– need for special connecting cable for dynamic measurement
	– cannot measure static pressure

Table 1.20. *Advantages and disadvantages of piezoelectric sensors*

1.4.2.4.1. Industrial example

Model 111A22 General Purpose ICP® Probe from PCB Piezotronics

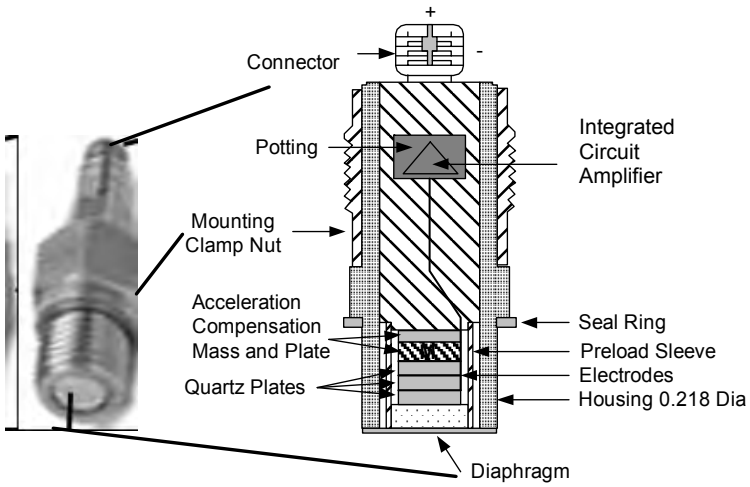


Figure 1.21. *Model 111A22 General Purpose ICP® Probe from PCB Piezotronics [8]*

Model: 111A22 General Purpose ICP® Probe	**PCB Piezotronics**
General purpose quartz pressure sensors are designed for dynamic measurements of compression, combustion, explosion, pulsation, cavitation, blast, pneumatic, hydraulic, fluid and other such pressures. These pressure sensors, structured with naturally piezoelectric, stable quartz sensing element, are well suited to measuring rapidly changing pressure fluctuations over a wide amplitude and frequency range.	
Sensitivity	0.145 ±0.015 mV/kPa
Dynamic Range (for 5V output)	34,475 kPa
Maximum Pressure	103,425 kPa
Low Frequency (–5%)	0.001 Hz
Resonant Frequency	> 400 kHz
Rise Time	< 1 µs
Discharge Time Constant	> 500 sec
Operating Temperature Range	–73 to +135°C
Sensing Element	Quartz
Height	35.05 mm

Table 1.21. *Specifications of Model 111A22 General Purpose ICP® Probe from PCB Piezotronics [8]*

1.4.2.5. *Conversion by Oscillators*

This type of sensor contains a vibrating element. The frequency of its vibrations depends above all on the forces which are applied to it. According to its value, the compressive or tensile force is applied directly or indirectly on the vibrating element. Table 1.22 describes the advantages and disadvantages of such sensors.

ADVANTAGES	DISADVANTAGES
– high precision	– poor linearity, hence need for associated digital processing
– facilitates conversion into numerical form	– sensitivity to temperature

Table 1.22. *Advantages and disadvantages of oscillators*

1.4.2.5.1. Oscillator with vibrating blade or cylinder

There are two ways for the sensing element to be exposed to the pressure to be measured:

a) *The sensing element is the vibrating element:* this is the case for a vibrating tube which is in fact a one-eyed tube.

b) *The sensing element is connected to the vibrating element:* this is the case for a steel string, fork or blade which vibrates when tensioned between a fixed point of the case on the one hand, and a diaphragm or bellows on the other hand (Figure 1.22).

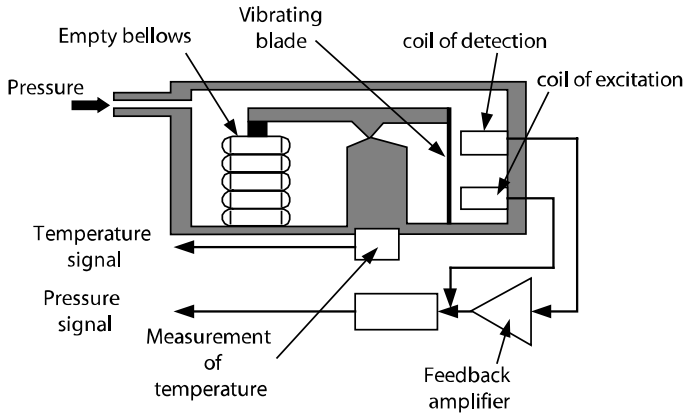


Figure 1.22. *Oscillator with vibrating blade or cylinder*

The vibrations are maintained thanks to two coils: the detection coil and the excitation coil. The detection coil supplies a voltage induced by the vibrating element which is made of ferromagnetic material. This voltage is amplified and supplies the excitation coil. The frequency f of the mechanical vibrations depends on:

– the shape and dimensions of the vibrating element;

– the physical properties of material used (e.g. density ρ and modulus of elasticity);

– the forces which are applied to it.

In the case of the vibrating string, we have:

$$f = \frac{1}{2l}\sqrt{\frac{F}{\rho s}} \tag{1.8}$$

ρ : density

s: cross-sectional area

l: length

F: applied force

f: frequency

The mathematical model associated with vibrating tube oscillators is given by:

$$p = A\,(f - fo) + B\,(f - fo)^2 + C\,(f - fo)^3 \tag{1.9}$$

f_0: frequency of vibration for zero pressure

f: frequency of vibration for measured pressure p

A, B, C are 3 characteristic constants of the sensor

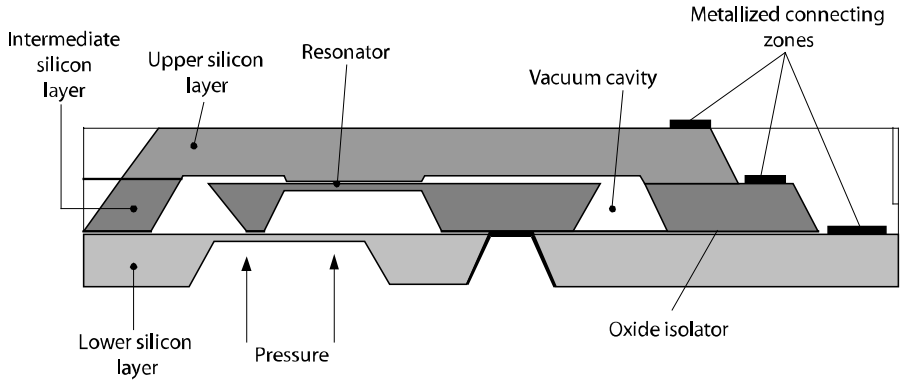


Figure 1.23. *Oscillator with silicon vibrating blade*

The advantages and disadvantages of such sensors are indicated in Table 1.23 below.

ADVANTAGES	DISADVANTAGES
– information carried by the frequency – high output signal – excellent repeatability – excellent resolution – excellent precision – low cost	– large size – limited bandwidth

Table 1.23. *Advantages and disadvantages of sensors with vibrating elements*

The resonant principle is used by THALES in a microsensor with a vibrating blade connected to a silicon diaphragm (Figure 1.23). In this case the excitation and the detection of the vibration of the blade are obtained by an electrostatic field.

1.4.2.5.2. Quartz oscillator

Another principle of measurement uses the influence of a force on the resonant frequency of quartz crystal. In such sensors, force is applied to the edge of a thin

quartz disk with two metal contacts. Technical values are summarized in Table 1.24 below.

Characteristics	Value
measuring range	0-1 bar
linearity error	lower than ±0.025% FS
hysteresis error	lower than ±0.025% FS
repeatability	lower than ±0.025% FS
drift of the zero	0.009% FS per °C
drift of the sensitivity	0.009% FS per °C

Table 1.24. *Technical values of pressure sensor with oscillating quartz*

Many aeronautics sensors use the latter principle. For example, Figures 1.24 and 1.25 show an absolute pressure sensor manufactured by THALES. High vacuum in the sensor body is used as the zero reference. The pressure applied to the bellows, whose external face is in the vacuum, develops a force. This force is transmitted to a quartz blade by an articulated arm. This arm is force balanced and so the centre of gravity of the whole system is kept in the geometric center, which eliminates nearly all the forces due to vibrations and accelerations and their impact is therefore reduced to less than 0.0008% FS per g.

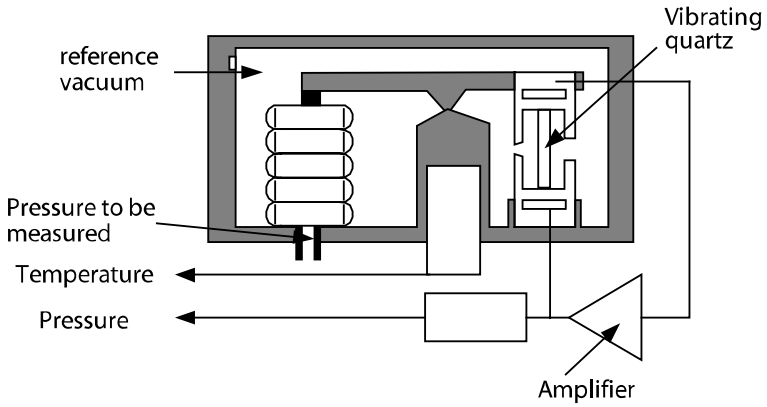


Figure 1.24. *Principle of a sensor with oscillating quartz*

The axial compression of quartz decreases its resonant frequency. Its value is 40 kHz with zero pressure and approximately 36 kHz for pressure corresponding to the nominal range of the sensor. The oscillation frequency f relates to the pressure p by:

$$p = A\,(f_o - f) - B\,(f_o - f)^2 \tag{1.10}$$

f_0: the oscillation frequency for p = 0

A, B are characteristic coefficients of the crystal, bellows, and arm.

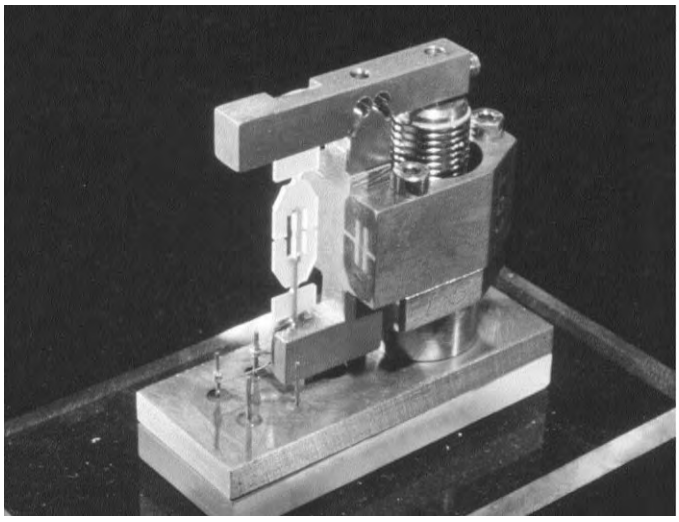


Figure 1.25. *Mechanism of quartz pressure sensor P51 [9]*
Courtesy of THALES AVIONICS ®M.CROUZET/P.DARPHIN

By using a 10 MHz clock and a microprocessor, for example, we determine the duration of 1,000 periods of the quartz oscillation, which makes it possible to obtain a resolution of approximately 0.003% of the full scale in 25 milliseconds. The ends of the quartz blade form a mechanical filter which eliminates any transfer of energy from the blade towards the structure. By this means and, therefore, owing to the fact that the blade is in the vacuum, the vibration damping is minimized. The repeatability and the hysteresis are 0.005% FS (Full Scale). The drift according to temperature is 0.0002% FS per °C for zero and 0.0014% of the value per °C for sensitivity. Table 1.25 indicates the advantages and disadvantages of such sensors:

ADVANTAGES	DISADVANTAGES
– frequency output	– sensitive to vibrations
– high level of output signal	– complicated construction
– excellent stability and resolution	– high price
– repeatability	

Table 1.25. *Advantages and disadvantages of pressure sensor with oscillating quartz*

1.4.2.5.3. SAW pressure sensors

Another type of electromechanical oscillator usable for measuring of pressure is based on the propagation of elastic waves on the surface of a piezoelectric (usually quartz) substrate. The propagation of an elastic wave allows the realization of a delay line with a time delay T:

$$T = \frac{l}{V} \tag{1.11}$$

T: the delay time

l: the distance between transmitter and receiver of the wave,

V: the propagation velocity of the wave

The insertion of the delay line in feedback of an amplifier makes it possible to constitute a sinusoidal oscillator whose frequency f is:

$$f = n\frac{1}{T} = n\frac{V}{l} \tag{1.12}$$

f: the frequency

n: a whole number determined by the dimensions of the substrate and the non-linearities of the amplifier

A pressure sensor is produced by constructing its diaphragm from a quartz blade on which the delay line is deposited (Figure 1.26). Resonant pressure sensors may

be powered and read remotely by RF signal. This can be exploited e.g. for monitoring tire pressure.

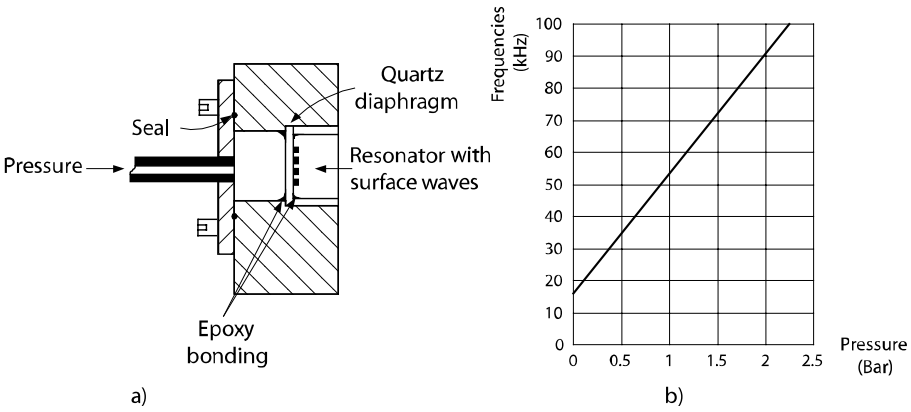


Figure 1.26. *Pressure Sensor with conversion by surface waves [9]*
Courtesy of THALES AVIONICS ®M.CROUZET/P.DARPHIN

Table 1.26 indicates the advantages and disadvantages of such sensors.

ADVANTAGES	DISADVANTAGES
– frequency output	– sensitive to temperature
– high level of output signal	– complicated construction
– average performance	
– can be wireless	

Table 1.26. *Advantages and disadvantages of pressure sensors with conversion by surface waves*

1.4.2.5.4. Industrial example

Model RPT series and sensing element in silicon document DRUCK

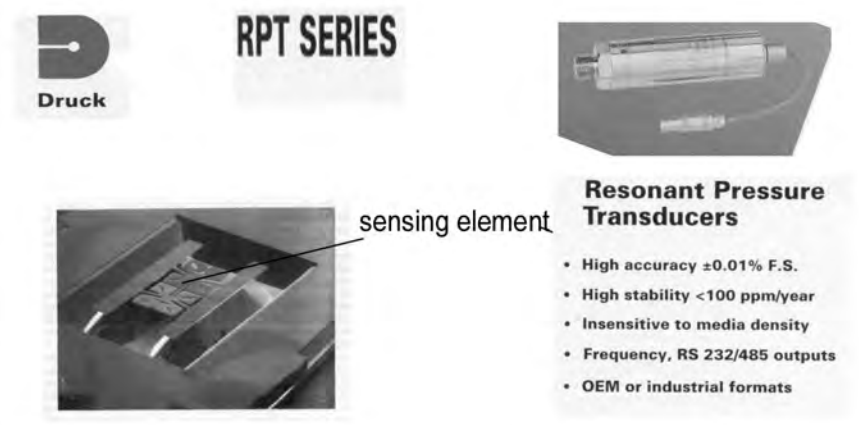


Figure 1.27. *Inside of model RPT series and sensing element in silicon document Courtesy of DRUCK picture by www.keller-druck.com*

RPT Series Resonant pressure transducer	DRUCK France
SPECIFICATIONS	– Ranges from 750 mbar to 1,150 mbar – High accuracy ±0.01% FS – High stability < 100 ppm/year – Frequency, RS232/485 outputs – Insensitive to media density – OEM or industrial formats
The use of resonant structures in pressure transducers has long been recognized as producing highly stable sensors. DRUCK have developed this technology to produce a series of Resonant Pressure Transducers (RPT) using silicon to give high accuracy and stability with low manufacturing costs.	

Table 1.27. *Specifications of model RPT series from DRUCK [10]*

Pressure sensor with vibrating resonant beam principle P90 from THALES

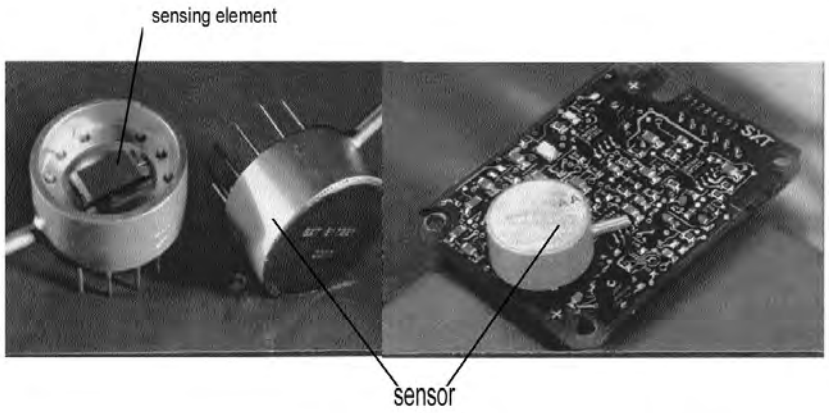


Figure 1.28. *Photograph of inside of pressure sensor with vibrating resonant beam principle model P90 from THALES AVIONICS ®M.CROUZET/P.DARPHIN [9]*

Model P90 Pressure Sensor	**THALES**
– silicon bulk micromachining sensing element	
– vibrating resonant beam principle	
– typical range: 1,500 hPa – 2,900 hPa	
– smart sensor P90: overall accuracy 1.10^{-4} FS fully compensated	
– avionics air data applications	

Table 1.28. *Specifications of pressure sensor with vibrating resonant beam principle model P90 from THALES [9]*

1.4.2.6. *Optical conversion*

The displacement or the deformation of the sensing element can be transformed into a variation of light intensity. The light modulated in this way is received by a photodiode either directly or by means of a light guide (optical fiber, for example).

ADVANTAGES	DISADVANTAGES
– good resistance to mechanically and electrically harsh environments – remote electronics	– very expensive – low accuracy

Table 1.29. *Advantages and disadvantages of pressure sensors with optical conversion*

1.4.2.6.1. Industrial example

MODEL PSI Glow from OPTRAND

Fiber-optic pressure sensor MODEL PSI Glow	OPTRAND
The PSI glow from *Optrand Inc.*, Plymouth, MI, integrates a 1.7 mm dia. sensor with a fully functional glow plug for combustion pressure monitoring of diesel engines without head modification or loss of glow plug functionality. Mounted in the glow plug sealing surface, the sensor's optical fibers are positioned in front of a flexing metal diaphragm directly exposed to the combustion chamber. Reflected light intensity is proportional to the pressure-induced deflections of the diaphragm. The temperature compensated version offers accuracy in the range of 1-2% FS. The sensor requires neither water nor air cooling.	

Table 1.30. *Fiber-optic pressure sensor MODEL PSI Glow from OPTRAND [11]*

Figure 1.29. *Photograph of PSI Glow from OPTRAND [11]*

1.4.2.7. *Servo controlled sensors with balance of force*

Principle

In the servo controlled pressure sensor, the measurement signal is amplified and then used to generate a force F' balancing the measured force F applied to the sensing element (Figure 1.30). The sensor signal is amplified, demodulated and applied to an electrodynamic actuator. The current, passing through the coil of the actuator, is the sensor output. The compensation force F' is given by:

$$F' = B.l.I. \tag{1.13}$$

B: the induction of the permanent magnet of the actuator

l: the length of the wire of the movable return coil

I: current passing through the actuator coil

When the balance is reached, $F = F'$ and thus $p_s = a \, . \, I$ where a is construction constant.

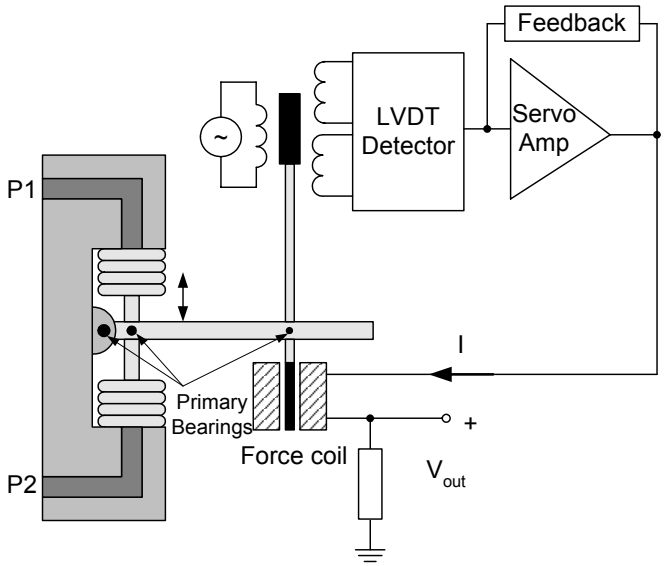


Figure 1.30. *Pressure sensor with balance of force*

Thanks to servo control, the deformation of the sensing element is very small (less than 0.2 μm) resulting in a negligible hysteresis error. The advantages and disadvantages of such sensors are detailed in Table 1.31.

ADVANTAGES	DISADVANTAGES
– very low hysteresis error (5.10^{-6} to 2.10^{-4}FS) – good linearity error (0.01% FS) – excellent precision (10^{-4} FS) – excellent stability in time	– very difficult construction – expensive – sensitive to shocks and vibrations

Table 1.31. *Advantages and disadvantages of servo controlled pressure sensors*

1.4.3. *Vacuum sensors*

1.4.3.1. *Ionization pressure sensors*

Ionization detectors have been available since 1916. They measure vacuum by making use of the current carried by ions formed in the gas by the impact of electrons. A plate brought to a slightly negative potential compared to the filament is associated with a grid whose potential is approximately 250 V (Figure 1.31). The electrons emitted by the heated filament are recovered by the grid while the ions, whose amount I is proportional to the level of vacuum, are recovered by the plate giving a current in the galvanometer.

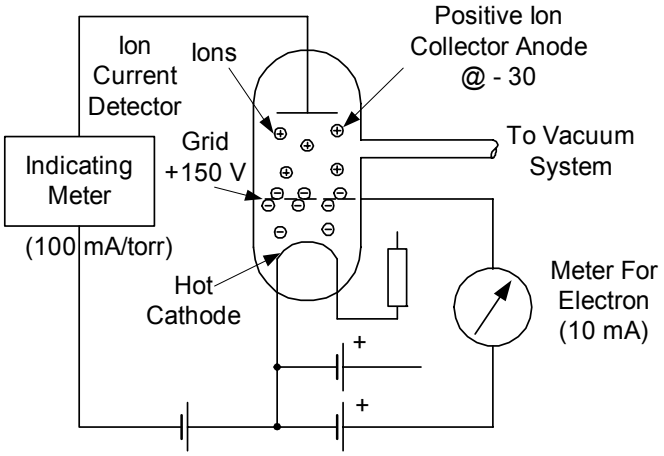


Figure 1.31. *Ionization detector [10]*

Most of these sensors measure vacuum in the range of 5.10^{-9} Pa. Modern vacuum system may be made entirely of metal. One argument in favor of this is that glass decomposes during routine degassing, producing spurious sodium ions and other forms of contamination. Nevertheless, glass gauges are still the most popular hot cathode sensors for the time being. Table 1.32 indicates the advantages and disadvantages of such sensors.

ADVANTAGES	DISADVANTAGES
– vacuum measurement	– fragile construction – many parasitic effects

Table 1.32. *Advantages and disadvantages of ionization pressure sensors*

1.4.3.2. *Heating effect sensors*

The idea of assessing the quality of a vacuum by exploiting the thermal conductivity of gases came from Kundt and Warburg in ca. 1875. The work of Pirani (1906) gave rise to the pressure gauge which bears his name and is one of the principles still most widely used for primary vacuum measurement (less than a few thousandths of Pascal).

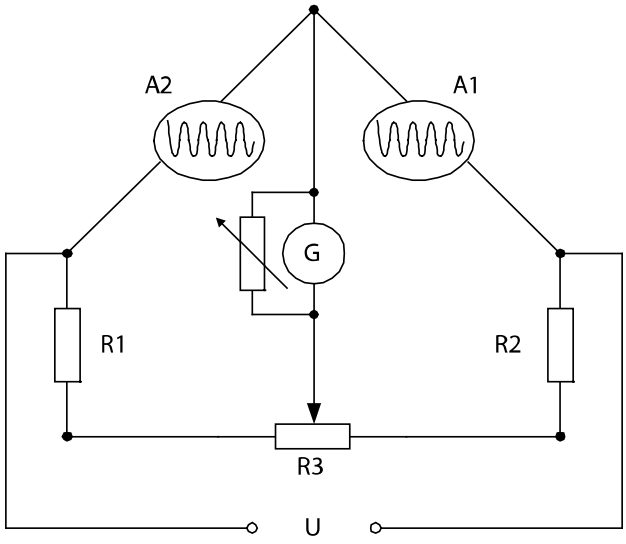


Figure 1.32. *Assembly of Pirani gauge*

The most widely employed assembly resulting from the above is illustrated in Figure 1.32, in which *A2* is a resistive filament identical to *A1*, but is maintained in a bulb sealed in a very effective vacuum, while *A1* is placed in the vacuum to be measured. The Wheatstone Bridge assembly makes it possible to correct the zero drift due mainly to the ambient temperature.

1.5. Calibration: pressure standards

1.5.1. *Low pressure standard*

This pressure standard gauge uses two tanks, one fixed and one mobile, connected by a flexible tube. The mobile tank is raised by means of a precise screw-nut mechanism to a height such that the mass of the column of mercury is balanced by the difference in pressure applied to each tank (Figure 1.33). Achievable resolution is 10^{-2} Pa, precision is $\pm 10^{-2}$ Pa.

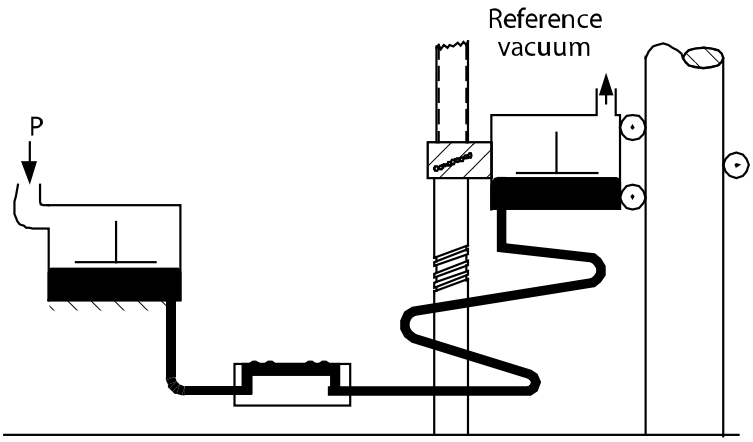


Figure 1.33. *Principle of standard manometer [12]. Document courtesy of Schwien*

1.5.2. *High pressure standard*

A dead-weight calibrator consists of vertical piston and cylinder defining an effective section A_e. The weight m acts on the piston (Figure 1.34).

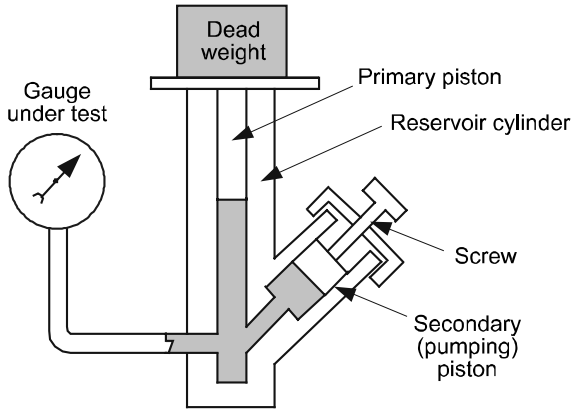


Figure 1.34. *Schematic diagram of a dead-weight calibrator*

If the upper part is in vacuum, the value of the pressure p maintaining the piston in balance is given by the following formula:

$$p = mg/A_e \tag{1.14}$$

g: gravity

m: total mass of the mobile part (piston and masses)

A_e: effective section

Figure 1.35 shows one type of pressure standard.

Figure 1.35. *Pressure standard [14]*

1.6. Choosing a pressure sensor

The choice of pressure sensors faces a great number of determining factors such as:

– precision;

– frequency response;

– sensitivity to the external quantities;

– calibration period;

– size;

– safety;

– reliability;

– lifespan;

– compatibility with the working environment;

– cost.

Generally it is necessary to consider not only the sensor itself, but the whole measuring system:

– the packaging;

– the signal processing electronics;

– the transmission lines;

– the interface protocols.

It is thus important for the user to be familiar with the technology and the design of the sensor as well as the influence of side-effects on the output signal.

All pressure sensors are invasive. It is important to check that the installation of the sensor does not disturb the phenomenon to be measured and does not degrade the safety or the total reliability of the system.

1.7. References

[1] THE PRESSURE, STRAIN and FORCE HANDBOOK, Omega Press, 2000, http://www.omega.com

[2] SFIM SAGEM FRANCE http://www.sagemavionics.com

[3] KELLER AG für Druckmesstechnik http://www.keller-druck.com

[4] VEGA Technique S.A. http://www.vega-technique.fr

[5] VAISALA SA http://www.vaisala.com

[6] KAVLICO CORPORATION, USA http://www.kavlico.com

[7] SCHAEVITZ http://www.schaevitz.com

[8] PCB PIEZOTRONICS INC. http://www.pcb.com

[9] THALES/SEXTANT AVIONIQUE www.thalesgroup.com/aerospace

[10] DRUCK LIMITED http://www.druck.com

[11] OPTRAND INC. http://www.optrand.com

[12] SCHWIEN ENGINEERING, INC. http://www.schwien.com

[13] BNM, Bulletin du Bureau National de Métrologie, October 1987, no. 70, pp. 1-47

[14] DESGRANGES & HUOT http://www.dh-budenberg.com

1.8. Other pressure sensors manufacturers

MOTOROLA Sensor http://www.motorola.com

KRISTAL Instrumente AG http://www.kristal.ch

ENDEVCO http://www.endevco.com

SENSONOR http://sensonor.com

KISTLER http://www.kistler.ch

SENSONOR http://sensonor.com

LETI Laboratoire d'Electronique et de Technologie de l'Information CEA/Grenoble FRANCE http://www-leti.cea.fr

GE novasensor (formerly Lucas NovaSensor) http://www.gesensing.com

1.9. Bibliography

1. Asch G.: *Les capteurs en instrumentation industrielle*, Dunod, 5th ed., 1998.

2. Baltes H., Göpel W., Hesse J.: *Sensors Update – Vol. 1 Sensor Technology – Applications – Markets*, 1996.

3. Campbell S.A. and Lewerenz H.J.: *Semiconductor Micromachining Vol. 1: Fundamental Electrochemistry and Physics*, Lavoisier, 1998.

4. Campbell S.A. and Lewerenz H.J.: *Semiconductor Micromachining Vol. 2: Techniques and Industrial Applications*, Lavoisier, 1998.

5. Greenwood J. and Wray T.: High accuracy pressure measurement with a silicon resonant sensor, *Sensors and Actuators*, Druck Limited, Fir Tree Lane, Groby, Leicester, UK, 1993, pp.37–38.

6. Guiffard B. *et al.*: Effects of Fluorine-Oxygen Substitution on the Dielectric and Electromechanical Properties of Lead Zirconate Titanate Ceramics, *Journal of Applied Physics*, Vol. 86, 1999.

7. Jacobsen E. *et al.*: A Dynamically Compensed Smart Sensor System – Sensors Motorola, Inc., May 1996.

8. Pierson J.: *The Art of Practical and Precise Strain Based Measurement*, 2nd ed., 1999.

9. Kaiser A.: Micromachined Electromechanical Devices for Integrated Wireless Communication Transceivers – The IST MELODICT Project, In mstnews 2/01, pp. 8–11.

10. Mandle J., Lefort O., Migeon A.: A New Macromachined Silicon High Accuracy Pressure Sensor, THALES, *Sensors and Actuators*, A 46-47, 1995, pp. 129–132.

11. Middelhoek S.: Celebration of the tenth transducers conference: the past, present and future of transducer research and development, *Sensors and Actuators*, A: Physical 2000, 82:1-3:2-23.

12. *The Pressure, Strain and Force Handbook*, Omega Press, 1996.

13. Paroscientific Inc.: Product Bulletin: Fiber Optic Pressure Transducer.

14. Guo S., Guo J., Ko W.H.: A Monolithically Integrated Surface Micromachined Touch Mode Capacitive Pressure Sensor, *Sensors and Actuators*, 80, 2000, pp. 224–232.

15. Simmons J.D.: NIST Calibration Services Users Guide, National Institute of Standards and Technology, Gaithersburg MD. USA NIST Special Publication 250 Rev., Oct. 1991.

16. Hecht E.: PHYSIQUE – Translation from 1st ed. by T. Becherrawy, revision by J. Martin, ITP Deboeck University s.a., 1999.

17. Alwang W.G.: Sensors A Comprehensive Survey in *Pressure Sensors*, ed. by Bau H.H., de Rooij N.F., Kloeck B., Vol. 7 (1994), pp. 513–554.

18. Crowe C.T., Elger D.F., Roberson J.A.: *Engineering Fluid Mechanics*, 7th ed., John Wiley & Sons, 2001.

19. Lide David R.: *CRC Handbook of Chemistry and Physics*, 79th ed., CRC Press, Boca Raton, FL, 1998.

20. Wells A.F.: *Structural Inorganic Chemistry*, 5th ed., Clarendon Press, Oxford, 1990.

21. ADEMIS – LETI/CEA.G: Guide des microtechnologies et des microsystemes, May 1997, pp. 1–191.

22. Bicking R.E.: *Fundamentals of Pressure Sensor Technology*, Honeywell Micro Switch, 1998.

Chapter 2

Optical Sensors

Introduction

Optical sensors are measuring devices in which a measured quantity is converted to an optical and, subsequently, an electrical signal by means of an optoelectronic transducer ([10]). Optical sensors belong to the class of contactless methods of measurement eliminating backward influence of a measuring device on an object of measurement.

As output signals from optical sensors are of an electronic nature and the methods of their further conditioning are generally known, the main attention will be devoted to the optical part of a sensor.

2.1. Optical waveguides and fibers

The simplest structure of an optical fiber used in optical sensors consists of a circular core with a cylindrical coating layer ([33]).

The index of refraction of a core, n_j, is always larger than that of coating, n_p. The zero loss propagation of light along the fiber requires the fulfillment of conditions for the total internal reflection, which are reached for angles of incidences larger than a certain value called a critical angle.

Chapter written by Stanislav ĎAĎO and Jan FISCHER.

For internal reflection on the boundary between the core and the coating (Figure 2.1) the value of the critical angle may be found from the formula

$$\sin \Phi_{3c} = \frac{n_p}{n_j} \tag{2.1}$$

The light introduced into the fiber with a certain angle will continue to reflect off the walls of the fiber and thus can travel long distances in the fiber.

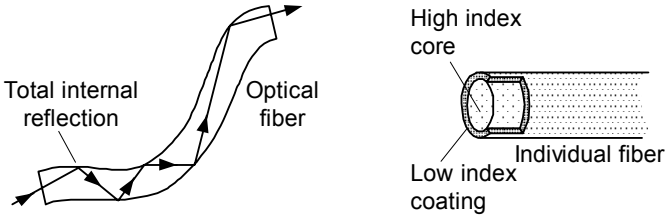


Figure 2.1. *a) Principle of the optical fiber b) Fiber with step index change*

The maximum value of the angle of incidence Φ_{1c} (Figure 2.2) limits the space of acceptance of rays impinging on the front end of a fiber (*acceptance angle, acceptance cone*).

The *sine* of the space angle of acceptance cone Φ_{1c} is usually expressed by means of a so-called *numerical aperture* (NA) i.e.

$$\mathrm{NA} = \sqrt{n_j^2 - n_p^2} = \sin \Phi_{1c} \tag{2.2}$$

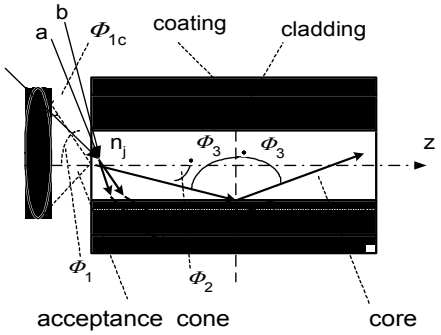


Figure 2.2. *Basic properties of optical fiber a) acceptance cone*

Construction of optical fibers for application in sensors

Optical fibers for sensors operating in the visible light region are made from glass or polymer materials (plastic).

	Glass	Plastic
ambient temperature	450°C	70°C/150°C
bending capability	low	good
length adjustable	not possible	easy
mechanical stability	low	good
total length	> 10 m	< 10 m

Table 2.1. *Selected properties of glass and plastic optical fibers*

For light channeling in the near and far infrared range, hollow tubes, which are highly polished inside and coated with reflective metals, are generally used. The light propagation in a hollow tube waveguide is based on mirror (specular) reflection while fibers use the effect of total reflection.

The basic properties of glass and plastic fibers are compared in Table 2.1.

2.2. Light sources and detectors

2.2.1. *Light sources*

From the variety of existing light sources (incandescent lamps, halogen lamps, etc.) only semiconductor sources ([8],[9],[19]) will be briefly described.

2.2.1.1. *Semiconductor sources of light*

Solid-state light sources such as light emitting diodes (LEDs) can produce light by means of electroluminescence. Under specific conditions coherent light can be produced in *laser diodes*.

Other technologies such as the Texas Instruments' *micromirror devices*, called "*digital light processors*", belong to the category of electronically-controlled light

sources important in measuring tasks working with *structured light* (machine vision, videometry).

Light Emitting Diode structure

LEDs are p-n junction devices constructed of junctions producing IR or visible light (e.g. gallium arsenide – GaAs, but not Si or Ge). The junction in an LED is forward biased and when electrons cross the junction from the n- to the p-type material, the electron–hole recombination process produces some photons in the IR or visible range in a process called electroluminescence.

An LED is a directional light source, with the maximum emitted power in the direction perpendicular to the emitting surface (most of the energy is emitted within 20º of the direction of maximum light). Most of the LEDs include plastic lenses to spread the light for a greater angle of visibility.

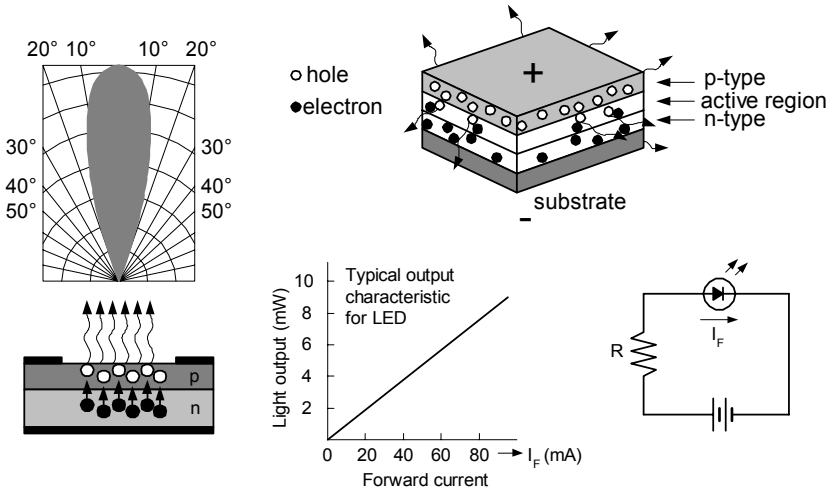


Figure 2.3. *The LED diode – radiation pattern, structure and typical output characteristics*

One way to construct an LED is to deposit three semiconductor layers on a substrate (Figure 2.3). Between p-type and n-type semiconductor layers, an active region emits light when an electron and hole recombine. If the p-n combination is a diode and when the diode is forward biased, holes from the p-type material and electrons from the n-type material are both driven into the active region. In this particular design, the layers of the LED emit light all the way around the layered structure. The LED structure is placed in a tiny reflective cup so that the light from the active layer will be reflected toward the desired exit direction.

2.2.1.2. *Laser diodes*

Laser action (with the resultant monochromatic and coherent light output) can be achieved in a p-n junction formed by two doped GaSe layers. The two ends of the structure need to be optically flat and parallel, with one end mirrored and the other partially reflective. The length of the junction must be precisely related to the wavelength of the light to be emitted.

The junction is forward biased and the recombination process produces light as in the LED (incoherent). Above a certain current level the majority of the charge carriers are in high energy states (population inversion). Population inversion (Figure 2.5) leads to stimulated emission, the photons moving parallel to the junction initiate a laser action.

The photons produced by stimulated emission form a standing wave (constructive interference) in the resonator. The distance between the parallel highly reflective mirrors of the optical resonator is matched to the wavelength of the laser radiation and only a very small part of it is coupled out of the resonator.

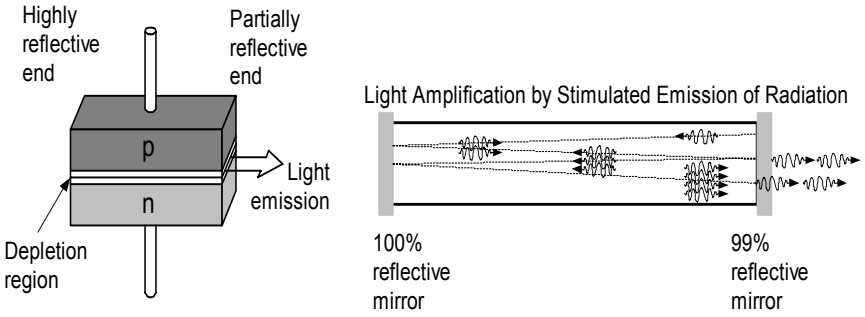


Figure 2.4. *The principle of laser and laser diode*

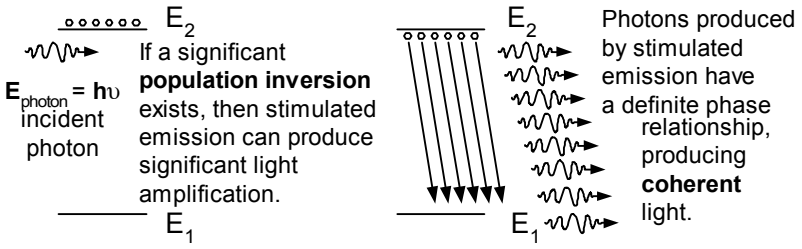


Figure 2.5. *The process of population inversion*

2.2.2. *Light detectors*

2.2.2.1. *Photoresistors*

Photoresistors are devices whose resistance changes upon light entering the surface. The most common materials for their fabrication are cadmium-based materials (CdSe, CdS, CdTe), working in the visible range of the spectrum (400 nm to 700 nm). In the infrared range (1.4 μm to 3 μm) lead-based materials prevail (PbS, PbSe, PbTe). For the range from 3 μm up to 1 mm indium-based materials (InSb, InAs) and doped Si and Ge are suitable.

Analysis shows that a single photon releases about 900 electrons for conduction making a photoresistor work as a photomultiplier and therefore is a very sensitive device.

The time response of a photoresistors is usually slow (fractions of a second) and they find applications in light switches (street lamps switching), automatic headlight dimmers in cars, flame detection, measurement of density of toner in photocopying machines, etc.).

2.2.2.2. *Photodiodes*

If a p-n junction of a photodiode is forward biased and is exposed to light of a proper wavelength, the current increase will be very small with respect to a dark current. If a junction is reverse biased, the current will increase quite noticeably. Impinging photons create electron–hole pairs. Correspondingly, the created holes flow to the negative terminal meaning that a photocurrent flows in the network. The voltage-to-current response of a typical photodiode is shown in Figure 2.6.

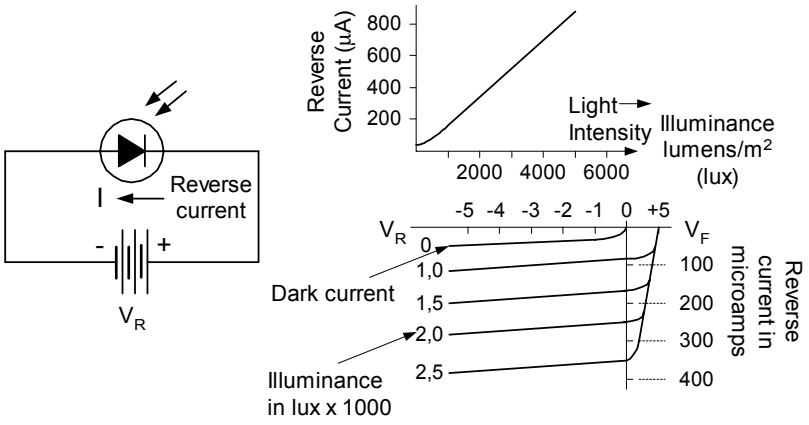


Figure 2.6. *The typical voltage-current (V/A) characteristic of the photodiode*

There are two general operating modes for a photodiode: the photoconductive (PC) and the photovoltaic (PV).

Photoconductive mode of operation

In the PC mode the diode is reverse biased by a voltage less than zero (Figure 2.6) and the diode operates in the third quadrant of the V/A characteristic. The current is almost linearly proportional to the light intensity but instead of the current the voltage drop on the load resistor LR connected in series is measured. For small LR the response to the step of the light intensity corresponds to bandwidths of hundreds of MHz (when the junction is polarized in reverse direction its capacity decreases).

Photovoltaic mode of operation

A typical photodiode (Figure 2.7) is composed of layers P+ N (p-n junction) and layer N+ placed close to the contact electrode. When the light is absorbed in the area of the p-n junction, the electrical field in the depletion region causes a drift of holes to domain P, and electrons to domain N. As a result of this drift movement a positive voltage appears on the anode of the diode. The electron–hole pairs are generated in the P+ domain with high concentration and thus will not contribute substantially to the photocurrent.

Eventually the electron–hole pairs generated by absorption in the depth are too far from the depletion region and soon they will be recombined and lost for the signal current.

Due to the absorption of photons in the N domain, the concentration of the holes increases. The change in the concentration of the electrons by absorption is negligible compared to the already high concentration of them in the N domain. As a result the holes diffuse towards the depletion region and even if the speed of diffusion is small compared to the speed of the drift, they reach the depletion region without being lost by recombination. Once they reach the region their movement (drift) is accelerated by the electric field of the depletion region. Because the speed of diffusion is much smaller than the speed of the drift movement and the reaction of photodiode is slower with respect to the case, when carriers would have been generated in the depletion region.

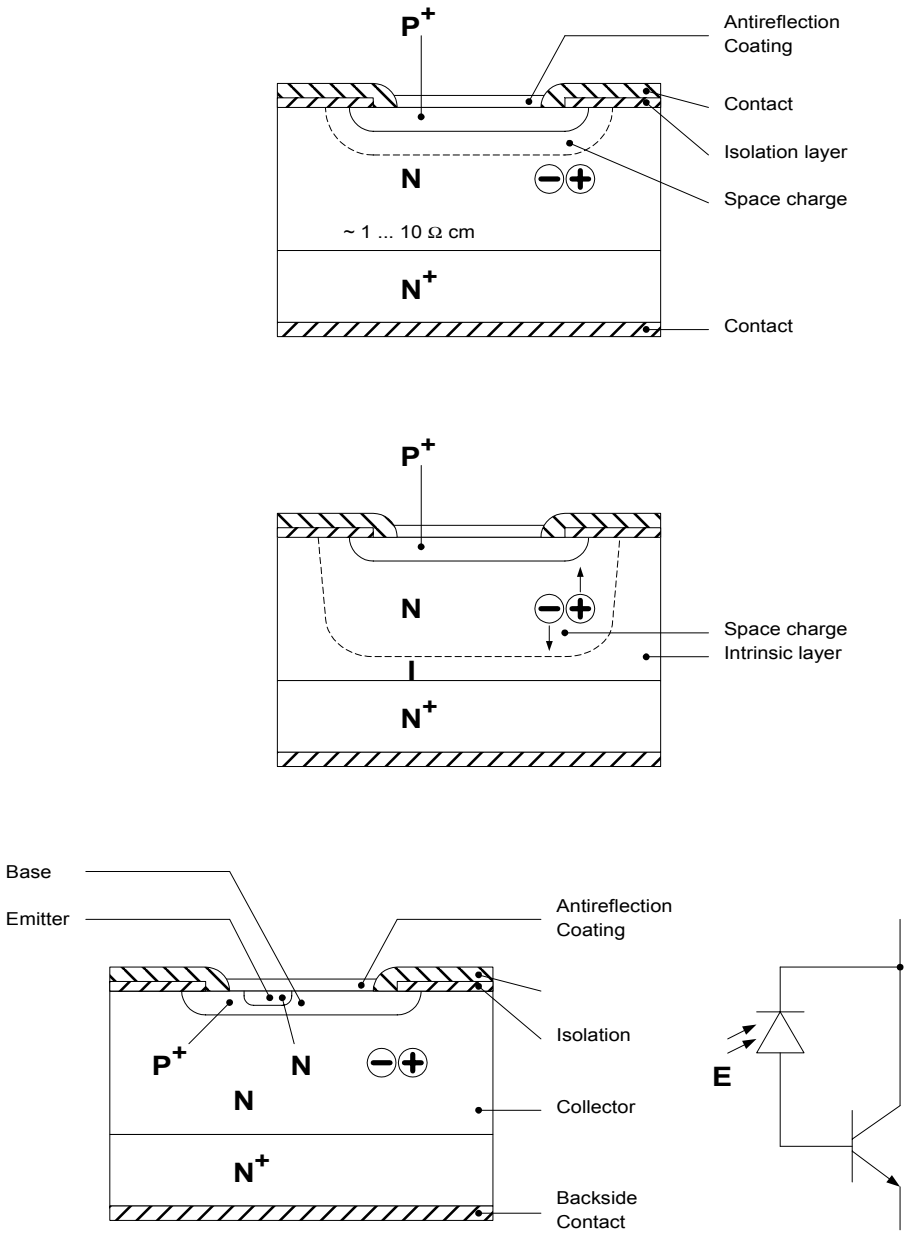


Figure 2.7. *The structure of the photodiode, PIN photodiode and phototransistor*

When a photodiode operates at the load with zero resistance (short-circuit), the intensity of the current is directly proportional to the light intensity.

On the other hand, if the voltage across the diode is measured, voltage is then approximately a logarithmic function of the illumination and reaches a typical value of 0.3 to 0.5 V.

As no bias is applied to the diode in the photovoltaic operating mode, there is no dark current, so there is only thermal noise present. This allows much better sensitivities at low light levels. However, the speed response is worse due to an increase in the junction capacity and a response to longer wavelengths is also reduced.

PIN photodiode

The PIN photodiode (Figure 2.7) uses an extra high-resistance layer, I, between the p and n layers to improve response time. The light penetrates through the thin P+ layer to the intrinsic layer, I, where it is absorbed. The speed of movement of the generated charges is high due to the electrical field of the depletion region. The response time of the PIN photodiodes is of the order of several nanoseconds.

2.2.2.3. *Phototransistor*

A phototransistor (Figure 2.7) operates as a combination of a reverse biased photodiode and a conventional transistor. The light incident on the collector-base junction generates electron–hole pairs. Electrons from the base and the collector region flow toward the positive voltage source (for an *npn* transistor) and they are returned to the collector through the emitter, where they are pulled into the collector by the electric field. The photon-induced base current is then amplified as in a conventional transistor, which makes the phototransistor a very sensitive light detector. The V-A characteristics of the phototransistor differ from those of a conventional transistor only in the fact that now the role of the base current is illumination.

2.2.2.4. *Position Sensitive photo-Detectors (PSD)*

Position sensitive photodetectors or “light potentiometers” are designed for applications, where a measured quantity is converted to a position of the light beam ([13]). The substance of a PSD sensor is the generation of the electron–hole pairs in an intrinsic layer of a large-scale PIN photodiode caused by an incident light on its frontal surface.

A photo-current arising in this way is divided into two parts by the resistive layer made from the P-type semiconductor (Figure 2.8b). The middle part of the sensor

(intrinsic layer -I) is manufactured from silicon with a large resistivity. The electric field corresponding to the depletion region on the junctions, PI and NI, evokes a shift of the holes towards layer P and the electrons towards layer N+. The electron–hole pairs generated at the point of incidence act as the current source with intensity I_o, then currents for the left, I_A and right, I_B electrode are given as:

$$I_A = I_o \frac{R_L - R_x}{R_L}; \quad I_B = I_o \frac{R_x}{R_L} \tag{2.3}$$

These relations are valid for the homogenous distribution of the resistivity of layer P (Figure 2.8).

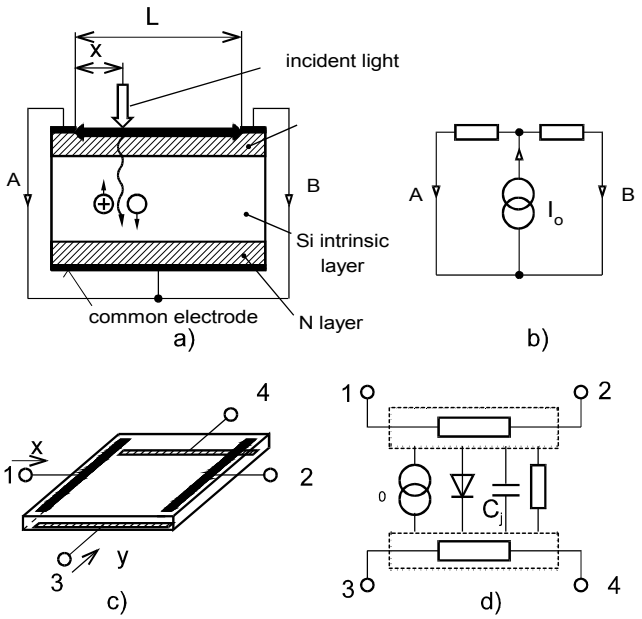


Figure 2.8. *Position Sensitive photo-Detector a) structure b) equivalent circuit c) 2D PSD d) equivalent circuit with spread resistance and capacitance*

In order to avoid the influence of the light intensity (current I_o), a ratiometric measuring circuit is used. The ratiometric circuit should evaluate the relation

$$\frac{I_A - I_B}{I_A + I_B} = \frac{I_A - I_B}{I_o} = 1 - 2\frac{x}{L} \tag{2.4}$$

The dependence on the light intensity is eliminated provided that all currents I_0, I_A, I_B are directly proportional to the light intensity.

A two-dimensional PSD sensor as shown in Figure 2.8c has two homogenous resistive layers located on both the top side (P) and the bottom side (N). Both layers having the length of 2L are provided by the pairs of electrodes (1 and 2, 3 and 4). The shift of the light trace from the center of the sensor's plane by x to the right and by y upwards cause the division of currents on the top layer to the parts I_1, I_2 and to the parts I_3, I_4 in the bottom layer. The co-ordinates of displacement x, y may be found from the ratio of currents I_1/I_2 and I_3/I_4.

2.2.2.5. *Charged coupled device image sensors*

The CCD image sensor is an array of photosensitive elements (photoelectric converters) configured either in lines (1D or line CCD) or in a matrix (2D devices).

The electron–hole couples generated by a known mechanism of photoelectric conversion are separated and moved to the appropriate electrodes by the electric field. The electric field arises from the application of a positive voltage on the upper electrode of the capacitor. The positive voltage creates a space underneath with a very low concentration of electrons (depletion region), attracting electrons and for this reason is often designated as the charge well.

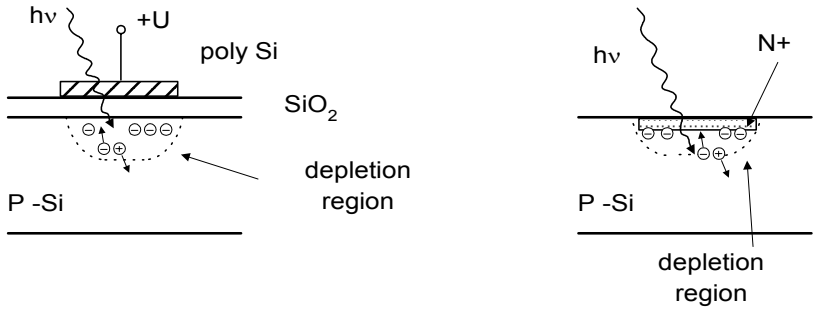


Figure 2.9. *MOS capacitor and PN junction as light to charge converters in CCD image sensors*

The electric field can be also generated by the depletion region of the PN+ junction as shown in Figure 2.9.

The charges accumulated during the exposure to incident light (charge packets) can be transferred from the detectors by means of an *analog shift register*, which can

easily be achieved using a MOS capacitor array. The accumulated charges are, after exposure, transferred by means of gates to the individual cells of the register.

The simplified structure of the analog shift register is depicted in Figure 2.10. Three sets of electrodes are used for the control of the charge packet transfer by means of three series of mutually delayed clock pulses (three-phase control). The voltage on the electrodes forms an electrical field region underneath each electrode known as a "potential or charge well". By means of the clock pulses the wells containing the charge packets are shifted out of register by the "bucket brigade" method.

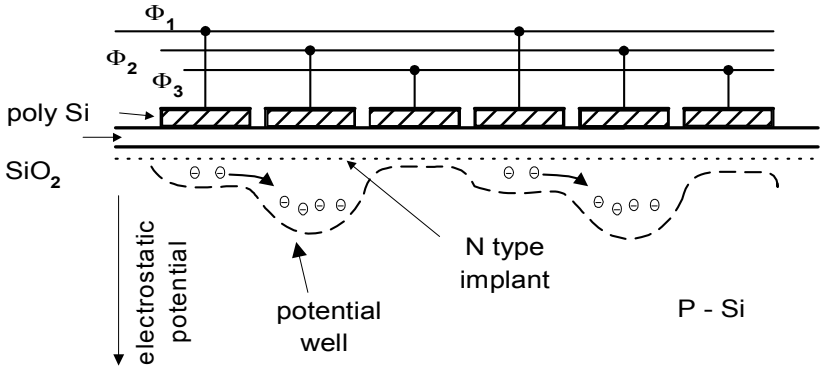


Figure 2.10. *Three phase analog CCD shift register*

One-dimensional CCD image sensors

A one-dimensional (line) CCD image sensor (Figure 2.11) comprises an array (a row) of photoelectric converters. The charge accumulated in each converter corresponds to the reflected light (or brightness) from the elementary area of the observed object (picture element – or pixel). The line CCD sensor then contains a charge image of the corresponding part of the object. The number of pixels in the line sensor (a length) generally varies between 128 and 12,000 pixels.

The charge image is moved to the analog shift register by means of transfer gates. The voltage pulses (clock pulses) generate a depletion region or charge well underneath each electrode. By sequential application of the clock pulses the charge wells move towards the output, where the charge packets are converted to voltage pulses. When this reading-out process is over, a new transfer of charge packets to the register begins.

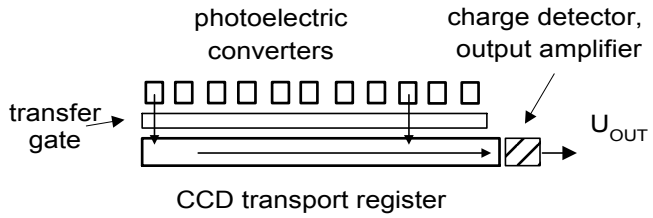


Figure 2.11. *CCD line sensor*

The color CCD sensors comprise three units and three interference based filters for the primary colors, R, G, B.

Two-dimensional CCD image sensors (2D-CCD sensors)

A 2D sensor contains an array of pixels organized into rows and columns in a similar way to a TV picture. The transfer of the charge packets from the individual pixels is most often done by so-called interline transfer which uses the group of vertical registers coupled to a horizontal register (Figure 2.12). After the exposure time the charge packets from the pixels are transferred to the corresponding cells of the vertical analog CCD shift registers. The charge packets in the bottom register cells are then moved to the corresponding cells of the horizontal register and the read out is similar to that of the line sensor. After the horizontal register is emptied, a new set of charge packets is moved from the vertical to the horizontal register and the read-out process is repeated until all vertical register's cells are not empty.

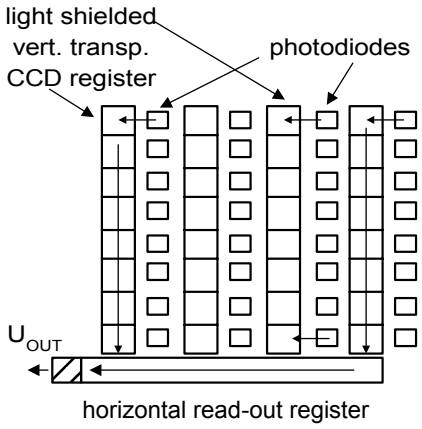


Figure 2.12. *Interline transfer CCD image sensor*

2D CCD sensors form the core of TV cameras using the most common video standards, CCIR (Europe) and RS 170 (USA). According to the CCIR standard the TV camera works in an *interlaced* mode of operation. In this first mode the camera records the pixels corresponding to the *odd* lines and then to the *even* lines of the picture, i.e. a full frame image is composed of two half-frames. Each half-frame contains 312.5 TV lines, however only 287.5 of them are available for image information, the rest serving for synchronization and other controlling purposes.

According to the CCIR standard the line and half-frame frequencies are 15,625 Hz and 50 Hz respectively, with a full-frame readout time of 40 ms. This may limit the application of the camera for rapidly changing scenes.

The voltage corresponding to the brightness of individual pixels is usually 1V with an impedance of 75 Ω.

2.3. Sensors of position and movement

2.3.1. *Position sensors using the principle of triangulation*

The configuration of the sensor designed for a measuring distance, y, of an object (target) from the light based on triangulation is depicted in Figure 2.13.

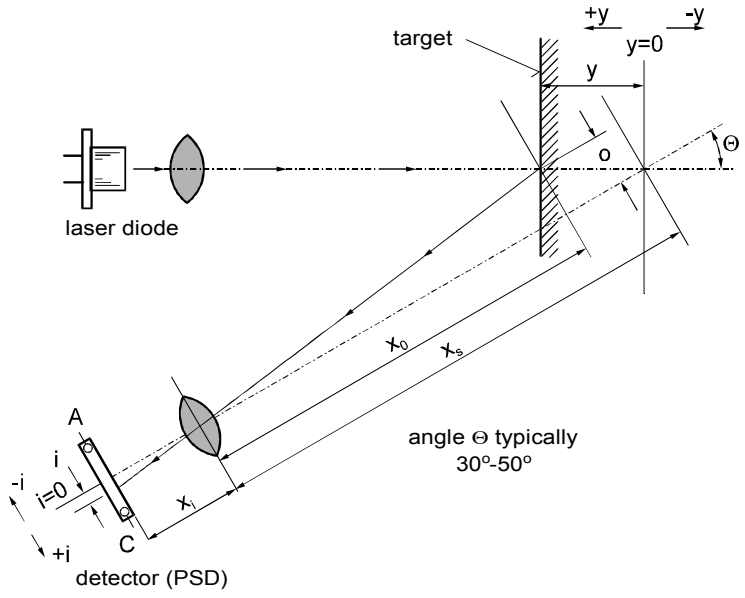


Figure 2.13. *The triangulation sensor with PSD for the distance measurement*

A narrow beam of light (angle < 2° – laser diode or LED) strikes the target and is reflected back towards the PSD sensor. The intensity of a received beam greatly depends on the reflective properties of the target. Nevertheless, the accuracy of the measurement is not dependent on the intensity of the received light.

As the surface moves within the spot, the image moves laterally on the single-axis PSD and the output voltage of the PSD depends linearly on the target displacement.

When the target surface is in its initial position y = 0 (stand-off distance), the detector's optical axis intersects the axis of the beam projected from the light source, placing the image of the light spot at i = 0 (center of the sensor) and the output signal is zero. The detector optics are focused in its initial position. If X_s is the distance of the target from the detector lens and X_i is the image distance, then according to the thin lens equation:

$$\frac{1}{f} = \frac{1}{X_i} + \frac{1}{X_s} \tag{2.5}$$

Instead of PSD sensors any sensor for measuring the position of a light spot, e.g. CCD line sensors, can easily be implemented for the triangulation measurement.

2.3.2. *Incremental sensors of position or displacement*

2.3.2.1. *General principles*

Incremental sensors of linear displacement (position) employ a grating principle, where a moving mask (usually fabricated in the form of a disk) has transparent and opaque sections ([12], [13], [17], web [1], web [2], web [12], web [14]). When the opaque section of the disk breaks the light beam the detector indicates logical zero, and when the light passes through a transparent section, the detector indicates logical one. After each displacement by one (pitch) *step* an impulse signal on the output of the detector is generated. A step, λ, is a distance between two successive marks. In order to measure the angular displacement, the disks with marks close to the circumference are used. An electronic counter counts the output pulses. A pulse generated when a reference mark crosses the light beam resets the counter.

2.3.2.2. *Linear incremental encoder*

A typical set-up for the measurement of linear displacement is depicted in Figure 2.14. A set of transparent marks is created on a glass rule or disk in the case of angular displacement. A pattern of marks with the same shape is located on a grid

placed close to the rule. The optical system of the sensor has three channels, each composed of the light source (usually LEDs), lenses, grids and two photodiodes connected in anti-parallel mode. Two channels deliver the input signals for the logic circuits detecting the displacement and a third channel is used to process signal corresponding to the reference mark (the origin of the scale).

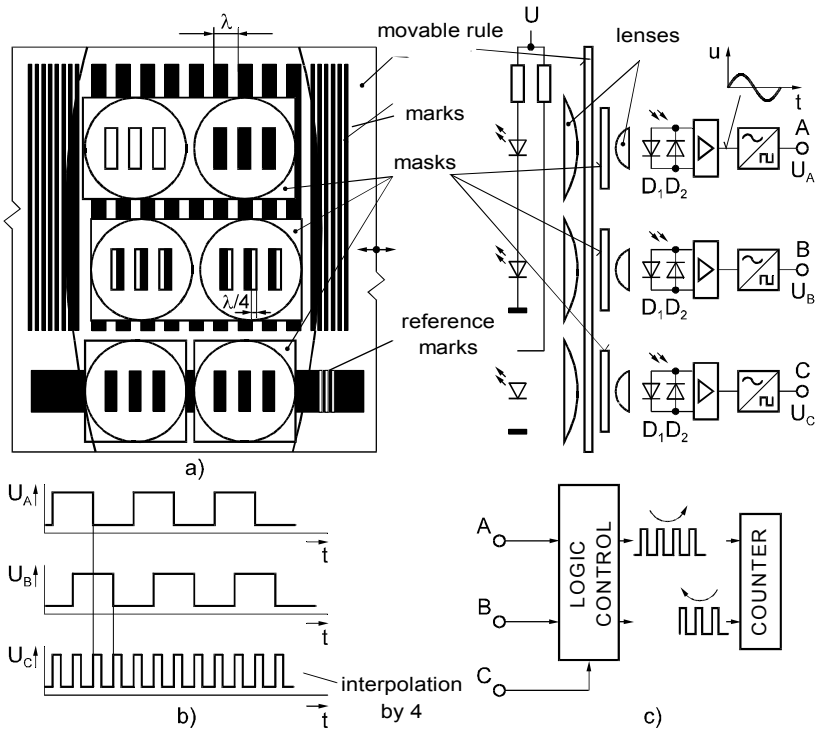


Figure 2.14. *Typical architecture of incremental encoder*

The two photodiodes in the receiving part of the sensor are mutually shifted by a distance of $n\lambda/2 + \lambda/4$ so that due to their anti-parallel connection, the waveform of the output signal has an almost sinusoidal shape. The sense of movement is found from signals U_A and U_B, in channels A and B respectively, providing that the grid in channel B is displaced by $n\lambda + \lambda/4$ with respect to the marks on the rule or disk.

By counting the pulses derived from the edges of the output signals, the resolution is four times increased.

In order to increase the resolution, the processing of signals, U_A and U_B, by an *interpolation procedure* can be implemented. The resolution can reach 0.05μm for linear or 0.00005° for angular measurements.

2.3.2.3. *Optical sensors of displacement with absolute encoding disk*

In this type of sensor the pattern of marks (optical or magnetic nature) is arranged in concentric circular traces on the surface of the disk. The rotation of the disk causes the interruption of the light beam between the linear light source or a group of LEDs placed on one side of the disk and the set of receivers (photodiodes) located on the opposite side. The encoding disk, having *n* traces, is able to distinguish between 2^n different angle positions. Information about the position is immediately available and thus a reference mark is not necessary. An absolute measurement is the most valuable advantage of sensors with coded disks.

The coded disk functions properly only on exact mutual alignment of the n-light sources and receivers. Hazardous states can occur in transition between code 01.... 1 and 10...0 when bits in each of the *n* traces change. The *unit distance code*, i.e. the code in which two neighboring combinations of bits differ only by one bit (e.g. Gray code) can avoid the problems (see Figure 2.15).

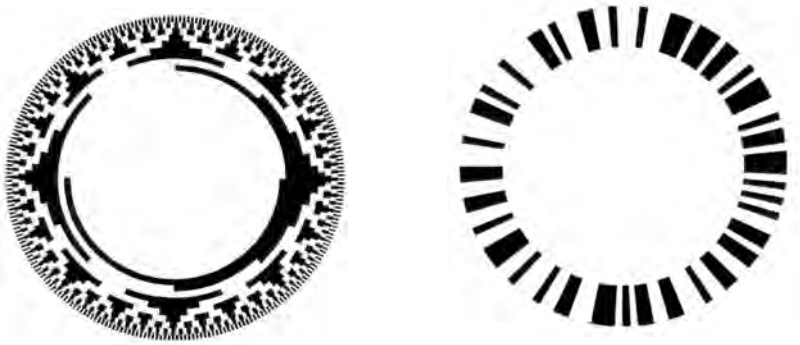
The coded disks with a diameter of around 10 cm could have marks placed in 17 traces and consequently could have a resolution of 2^{17} = 131,072 positions.

2.3.2.4. *Sensors with pseudorandom coding*

Sensors of this type have only one trace of marks with randomly chosen width [13]. The difference between the regular and randomly coded disk is depicted in Figure 2.15.

An array of receivers mutually displaced by one step of resolution is used for reading data from pseudorandom disks. For a resolution of 10 bits an array of 10 receivers reads a segment with an angle of 22.2°. The exact shape of the marks is not prescribed, but it should obey certain rules, e.g. for a disk with a resolution of 10 bits the average width of the mark is equal to 8.4375°.

The second reading possibility uses a set of receivers distributed over the whole trace. In this case receivers generate unit distance code ([13] – sensor ACE™).



a) b)

Figure 2.15. *The coded disks with resolution of 10 bits with a) Gray code b) pseudorandom coding*

2.3.3. *Photoelectric switches*

Photo-Electric Switches (PES) are optical sensors whose output signal has only two states assigned to the presence or absence of a certain feature in the measured object ([30]). Each PES sensor contains at least two basic components: a system *transmitter* (light source and optical parts) and *receiver* (optical components and photoelectric detector).

The basic types of PES are given by the mutual arrangement of the transmitter, the measured object and the receiver and may be divided into three groups [16]:

– Through beam.

– Diffuse reflective.

– Retro-reflective.

2.3.3.1. *Through beam PES*

These are used for the detection of opaque objects causing total or partial interruption of the light beam between the transmitter and receiver (Figure 2.16). The light source and receiver are in two separate housings and both need a power supply. The sensing distance (active zone) is the stretch between the light source and the receiver, without dead zones.

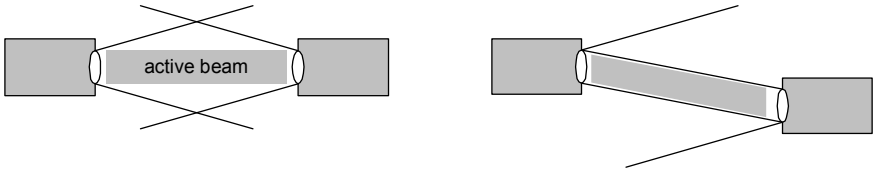


Figure 2.16. *Through beam PES a) principle and active zone b) divergence of beams allows the setting of active zone out of optical axis*

Advantages of through beam PES are the long permissible distances between the source and the receiver, the definite beam diameter, the precise positioning and the applicability of all kinds of opaque objects. The beam divergence supports an easy adjustment of the light source and the receiver for reliable function. The disadvantages of through beam PES are the installation expenses caused by the necessity of two power supplies (two active parts) and the possible difficulties, which could arise in the case where an object under detection is not accessible from both sides.

2.3.3.2. *Diffuse reflective PES*

The sensing distance depends on the reflectivity and the size of the object (Figure 2.17). There is a difference between ON and OFF points when moving the object from or to the sensor (differential travel). This difference also depends on the reflectivity of the target's surface; for "black" objects, it is the distance between points, when the PES is switched ON and OFF, which is smaller than for white objects.

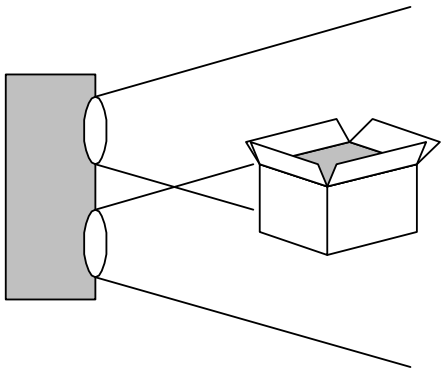


Figure 2.17. *Schematic principle of PES with diffusive reflection*

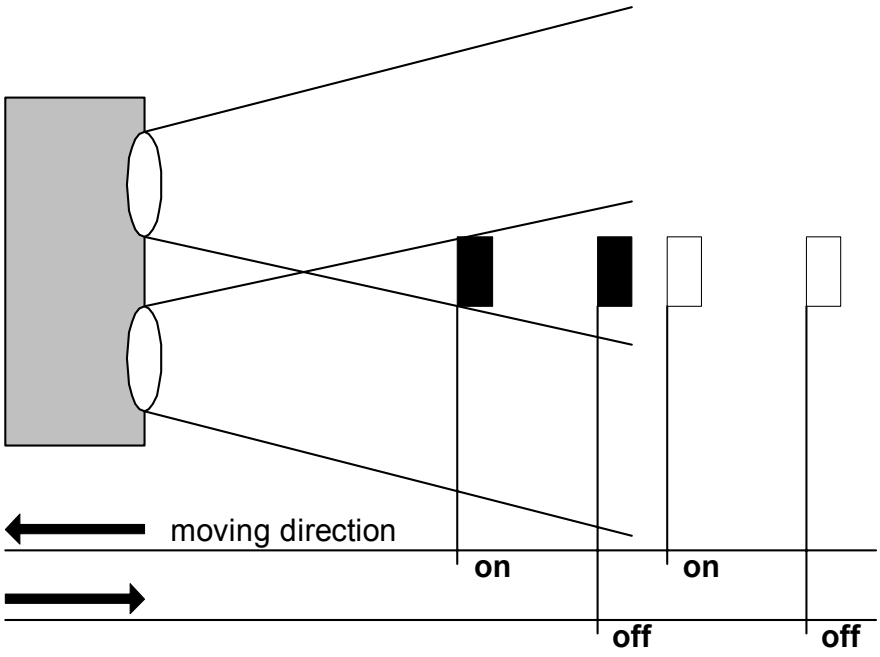


Figure 2.18. *Hysteresis and black and white error of diffusive PES*

The advantages of PES with diffusive reflection are the low installation expenses, the sensor is mounted on only one side and the adjusting sensor is relatively easy. On the other hand, proper function strongly depends on the optical properties of the object, the switching point depends on the position of the object and the sensing distance is relatively short.

2.3.3.3. *Retro-reflective PES*

The principle of these is similar to through beam PES; however, the light source and the receiver are located in one housing and a reflector is placed at the end of the active zone.

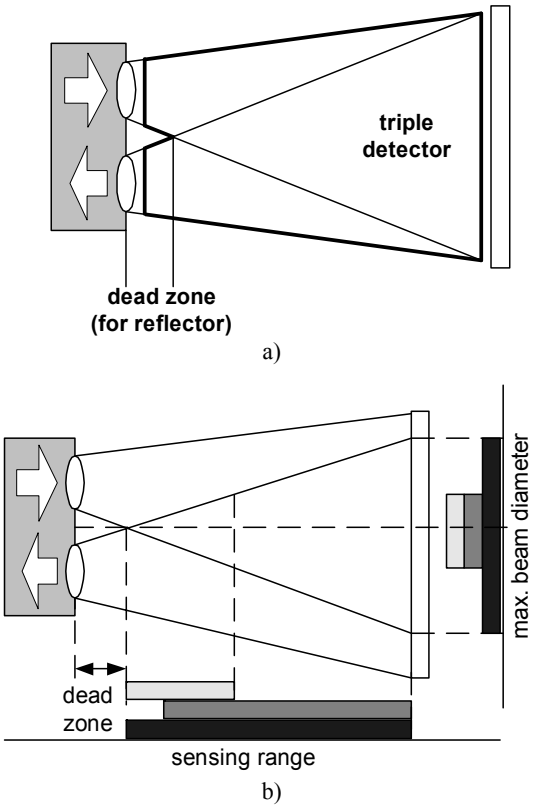


Figure 2.19. *a) Principle of retro-reflective photo-electric switches*
b) Dependence of dead zone on beam diameter and distance

The performance of a sensor depends on the properties of reflector; the *triple reflector* (Figure 2.19) is the most popular for these applications. Due to a divergence in the field of view of the source and detector, a dead zone is shown near the sensor. The active beam diameter depends on the size of the reflector and the distance from the sensor. The *sensing distance* is the stretch between the sensor and the reflector, which can be increased by using a larger reflector.

The main advantages of retro-reflective PES are low production cost, low installation expenses and flexibility of adjustment. However, the retro-reflective sensors are still comprised of two separate parts, where their "switching point" depends on the position of the object, i.e. the presence of the same object may not be detected in a certain position with respect to the reflector or the housing of the sensor (Figure 2.20).

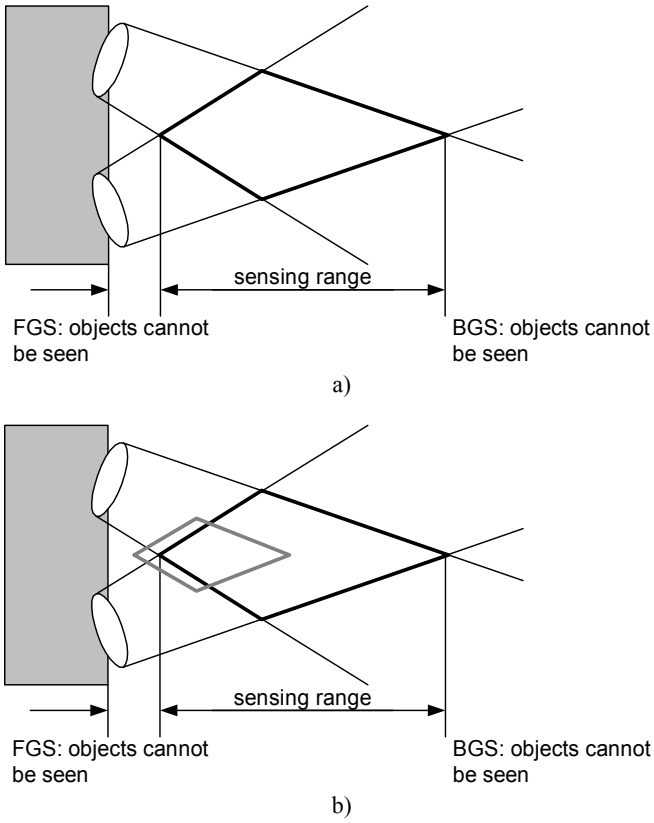


Figure 2.20. *Active zone setting a) principle*
b) change of active zone by optical axis deflection

2.3.3.4. *PES for detection of colors or color marks*

Most often for color detection an analytical approach is used, based on the decomposition of light from a white light source (halogen lamp) by interference filters to three basic colors which are then brought by an optical fiber (for example) to the surface of the object (Figure 2.21). The reflected light on the three primary colors is evaluated by three photodiodes with a maximum of the spectral sensitivity at the corresponding wavelengths. The weighted average of the output signals from the photodiodes corresponds to the measured color on the surface. The weighting coefficients are chosen from the triangle of primary colors. The halogen light source is used because its spectrum is close to white as it operates at a higher temperature. Thus, according to Planck's law, the content of energy in the visible spectrum is large enough.

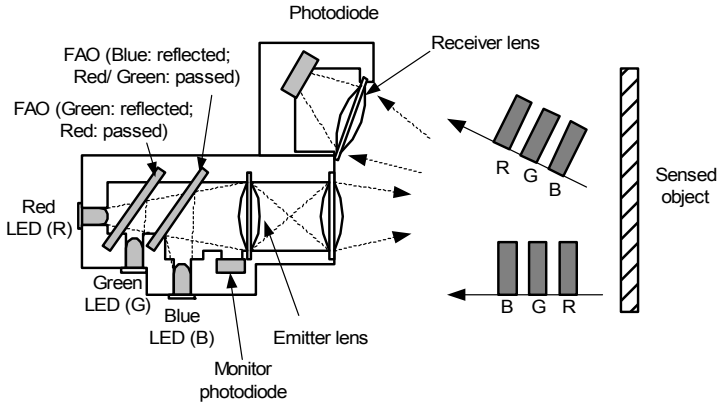


Figure 2.21. *Detector of colors on intensity principle*

2.4. Optical sensors of dimensions

2.4.1. *Dimensional gauge with scanned beam*

Sensors of this type are used in cases when a measurement should not depend on the position of the measured object. In a typical set-up [12] as shown in Figure 2.22 the laser beam is scanned by a five-sided rotating prism.

A special collimating lens produces parallel rays, which sweep through the workspace at a linear rate proportional to the prism's rotational speed. Thus, the position within the workspace of the cylindrical target (whose diameter is to be measured) is not critical.

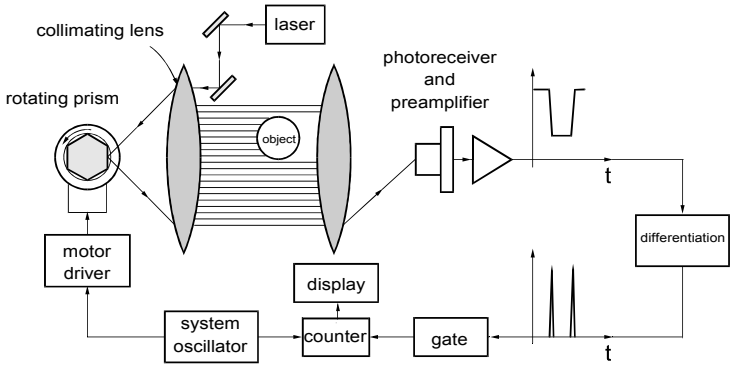


Figure 2.22. *Laser dimensional gauge using a shadowing principle*

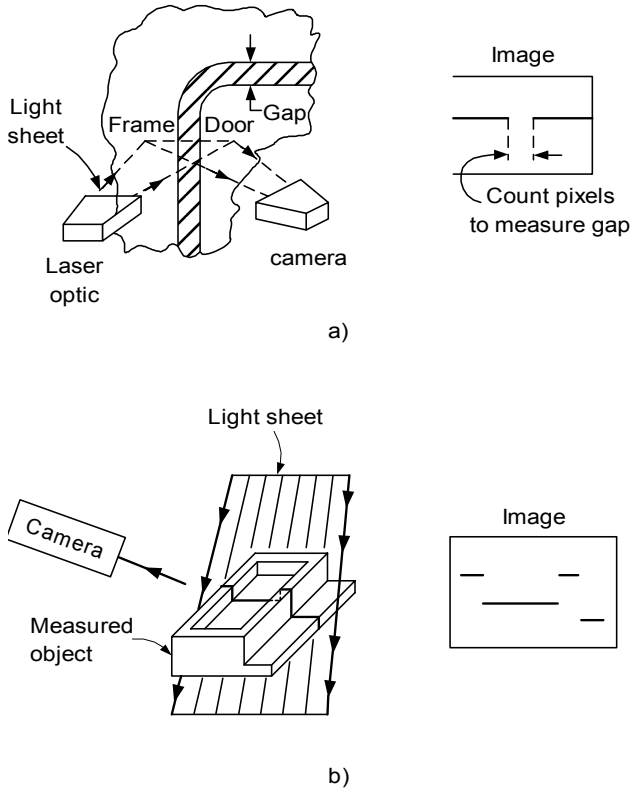


Figure 2.23. *Light sheet application*

Dimension measurement by means of the light sheet

Projecting a sheet of light (created by a rotating or oscillating mirror or by cylindrical lenses) onto a target object produces a line of reflected light, which follows the contours of the object [8],[9],[19]. The light sheet optically "cuts through" the measured body creating a cross-sectional image, which can be recorded by a suitable camera and analyzed in order to obtain the desired part dimensions (Figure 2.23a).

The image of the reflected line is then processed in order to obtain useful information about the geometry of the object. Several such sensors can be stationed around the periphery of an automobile door to check for proper fit during assembly by measuring the gap.

The motion of the target object through the light sheet can be used to produce successive slices, which reveal the shape and size of the entire object. When one-dimensional arrays of sensors are used, two-dimensional images can be formed by scanning or by use of the sensed object's own motion, as along a conveyer line.

Dimension measurement with digitalized video signal

One of the simplest methods of digitalized video signal processing is a binarization. In this approach each sample of the video signal is compared with a selected threshold illumination value and a digital value of 0 or 1 is assigned depending on whether the signal is greater or less than the threshold value. The image now consists only of "white" and "black" pixels. An example of a dimension measurement based on binarization is shown in Figure 2.24. By counting the number of black pixels the measured dimension (gap) can be found easily [12].

When the image suffers from various forms of degradation (background noise, blurring lines or boundaries, poor contrast etc.) improvements may be achieved by using techniques such as contrast stretching, edge detection, etc.

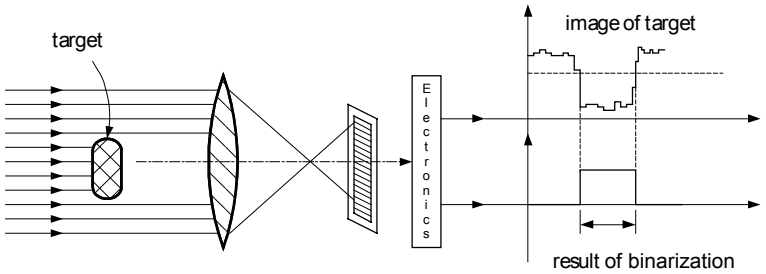


Figure 2.24. *Dimensions measurement using binarization of video signal*

2.5. Optical sensors of pressure and force

2.5.1. *Pressure sensor using the optical resonator*

The sensor (Figure 2.25) consists of the following essential components; a passive optical pressure chip with a membrane etched in silicon, a light emitting diode (LED) and a detector chip [13]. The pressure chip is composed of two electrodes forming an optical cavity working as a Fabry-Perot optical resonator-interferometer [2] measuring the deflection of the membrane. A back-etched, single-crystal diaphragm on a silicon chip is covered with a thin metallic layer and a glass plate with a metallic layer on its backside.

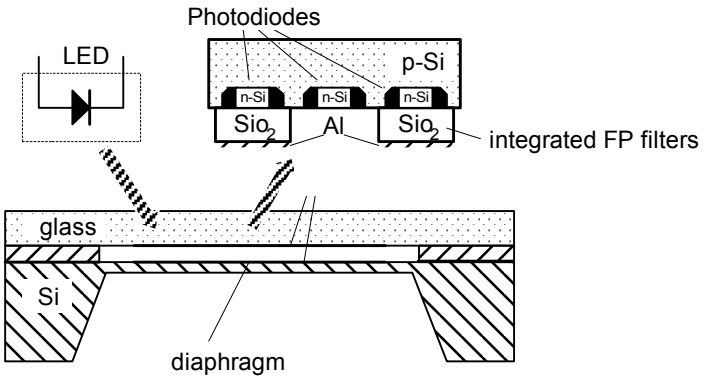


Figure 2.25. *The sensor of pressure with optical resonator*

The glass is separated from the silicon chip by two spacers at a distance, *d*. Two metallic layers form a variable-gap Fabry-Perot resonator with a pressure sensitive movable mirror (on the membrane) and a plain–parallel, stationary-fixed, half-transparent mirror (on the glass). The detector chip contains three p-n junction photodiodes. Two of them are covered with integrated optical Fabry-Perot (interference) filters of slightly different thicknesses. The filters behave as two resonance circuits with slightly different resonance characteristics.

The sensor's operational principle is based on the measurement of a wavelength modulation of the reflected and transmitted light depending on the width of the resonator cavity. The integrated FP filters behave as difference frequency demodulators, when the frequency increases the output of one filter increases and the output signal of the other decreases. The photodiode without the FP filter monitors the total light intensity arriving at the detector.

2.6. Optical fiber sensors

2.6.1. *Introduction and classification of sensors with optical fibers*

The propagation of light in an optical wave-guide or optical fiber may be affected by many physical quantities. There are only a few quantities, which cannot be directly or indirectly measured by means of optical fibers.

The optical fiber sensors can be divided to two groups:

Extrinsic sensors, where the optical fiber serves only to guide the light from the source to the place of interaction with a measured quantity and then to an optoelectronic sensor. Extrinsic sensors are often used in a contact-less measurement of the temperature-pyrometry to guide the thermal radiation (optical cable with hundreds of fibers) from the surface of the measured object to the detector. Optical fibers used in a cable must be made from materials exhibiting low attenuation, at least in near infrared field of spectrum.

Intrinsic sensors are based on direct impact of the measured quantity on the guiding of light propagation in an optical fiber. The measured quantity affects:

– the attenuation of light between the source and the detector (sensors with an amplitude modulation);

– the parameters influencing the velocity of light propagation in the fiber (sensors with a phase modulation);

– the parameters influencing the wavelength or reflected light in an optical fiber.

2.6.2. *Optical fiber sensors with amplitude modulation*

Sensors with deformation of fiber

The mechanical deformation of the fiber affects the light propagation as a consequence of the changes in the fiber core and the cladding geometry and changes in the refractive index caused by the mechanical stress. The deformation along the axis (longitudinal) and perpendicular to the axis (transversal) has a different effect on the attenuation of light in the fiber. These effects could be used for measurement of force at high temperature or in presence of aggressive materials.

The optical fibers fabricated from silicone rubber (SROF) which can be stretched up to 100% of their original length, modulating the light intensity almost linearly, may be used for force, strain and displacement measurement.

Optical liquid level detector

A liquid level detector [13] utilizes the difference between the refractive indices of air and the measured liquid. When the sensor is above the liquid level, the light at the output of the u-shaped fiber (Figure 2.26) is strongest. When sensitive regions of the fiber located near the bends (the smallest curvature) touch liquid, the condition for total reflection is not fulfilled and part of the light propagates beyond the fiber. The shape of the fiber “expels” liquid droplets when the probe is elevated above the level. The repeatability error of a probe with a diameter of 5 mm is about 0.5 mm.

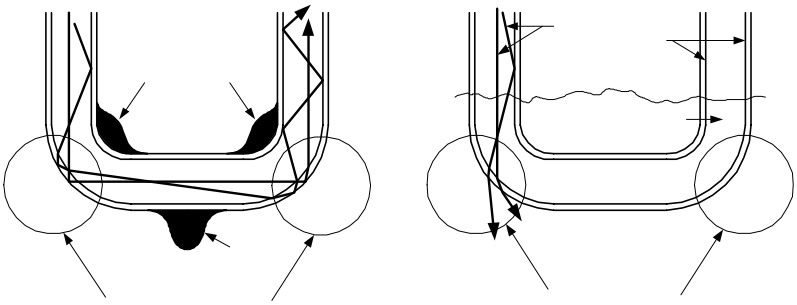


Figure 2.26. *Fiber optic liquid level sensor a) sensor above the liquid b) sensor immersed*

Reflective optical fiber sensors of displacement

In reflective sensors of displacement the light from the source propagates through an optical fiber towards the object with a surface having the properties of diffusive reflection (e.g. the membrane of a microphone or a pressure sensor). The transfer characteristic of a typical reflective sensor (Figure2.27) exhibit a rising part, which corresponds to a gradual overlapping of the conical traces of light emitted from the source, P_0, and the acceptance cone of the receiver fiber, P_1. Further displacement of the membrane will cause a gradual decrease according to the inverse square law.

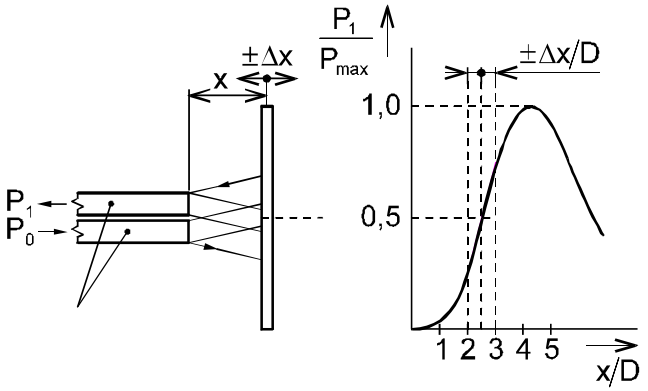


Figure 2.27. *The transfer characteristic of the reflective sensor with optic fiber*

For a small displacement measurement the initial displacement should be set to a point of half of the maximum, where the transfer function is nearly linear.

2.6.3. *Sensor with wavelength modulation*

Fiber Bragg Grating sensors (FBG)

The intensity of the reflected light beam propagating through the optical fiber with periodic changes in refractive index (diffraction grating) reaches maximum value when the condition of Bragg's law is fulfilled, i.e.

$$\lambda_B = 2n_{eff}\Lambda$$

where

λ_B ... is the wavelength of reflected radiation (Bragg wavelength);

n_{eff} ... refraction index of fiber;

Λ ... grating period (pitch).

The basic principle of an FBG-based sensor system (Figure 2.28) lies in the monitoring of the wavelength shift of the returned Bragg-signal, as a function of the measurand causing changes in Λ or n_{eff} (e.g. strain, temperature, force and pressure).

Sensor systems involving gratings such as these usually work by injecting light from a spectrally broadband source into the fiber, with the result that the grating reflects a narrow spectral component at the Bragg wavelength, or in transmission this component is missing from the observed spectrum.

Using a series of FBGs embedded into or attached to structures to record the changes in their dynamic behavior, the monitoring of structural integrity can be performed.

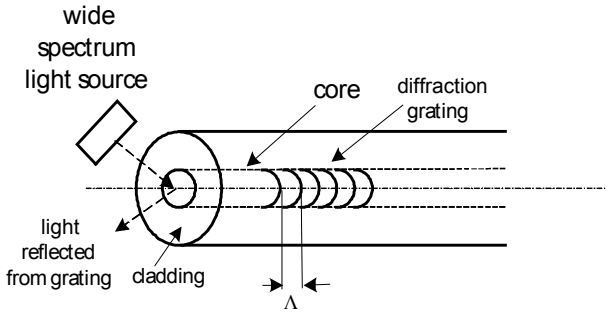


Figure 2.28. *Principle of Fiber Bragg Grating strain sensors (FBG)*

2.6.4. *Optical sensors with phase modulation*

Interferometers with optical fibers

Sensors based on a phase-modulated optical fiber arrangement corresponding to the Mach-Zehnder interferometer are the most popular for practical applications. In this interferometer the laser beam is in the first coupling element of two parts. One part is coupled to the active fiber, where the measured quantity causes changes in the fiber parameters and the second part is coupled to the reference fiber, which is not influenced by the measured quantity. Both active and reference fibers which are joined by the second coupler will interfere.

The results of constructive and destructive interference are then observed on the output of the coupler and transformed into an electrical quantity by some kind of photoelectric detector.

One of the most important optical fiber sensors operating with phase modulation is the solid-state gyroscope utilizing the Sagnac phenomenon (see Chapter 9).

2.6.5. *Perspective of optical fiber sensors*

Optical fiber sensors have a number of advantages over conventional electrical strain gauges. The most significant advantages are that:

– they are immune to electromagnetic fields;

– they have the ability to take many measurement points along a single fiber – greatly improving the ease at which sensors can be multiplexed; and

– they can be embedded within or bonded to structures without the risk of de-bonding during operation.

2.7. Optical chemical sensors

2.7.1. *Introduction*

These sensors are based on the interaction of electromagnetic radiation with matter. The result of this interaction is the alternation of some properties of radiation, like the intensity, the polarization and the velocity of light in the medium.

2.7.2. *Chemical sensors based on the absorbency measurement*

Optical chemical sensors can be designed and built in many ways, which are limited only by the designer's imagination. An example of a simple optical chemical sensor [13], [31] is the CO_2 sensor in Figure 2.29. The sensor consists of two chambers, which are illuminated by a common LED. Each chamber has metallized surfaces for better internal reflectivity. The left chamber has slots covered with a gas-permeable membrane. The slots allow CO_2 to diffuse into the chamber. The bottom part of the chamber is made from glass. Both wafers, A and B, form optical waveguides. The test chamber is filled with a reagent, while the reference chamber is empty. The sample part of the sensor monitors the optical absorbency of a pH indicator in a dilute solution.

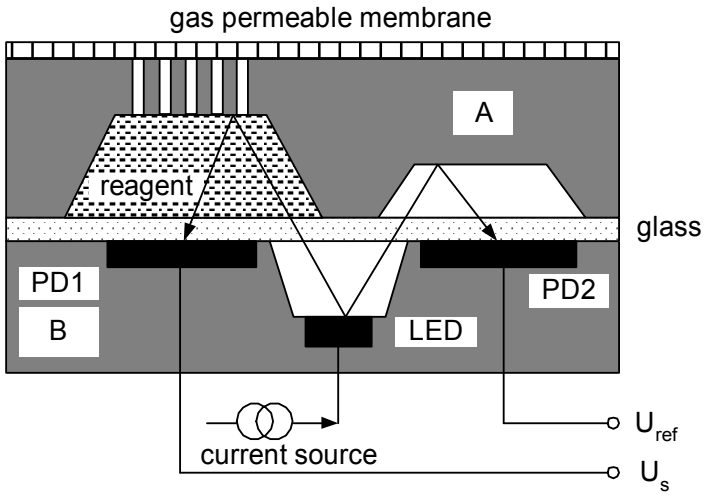


Figure 2.29. *Optical sensor of CO_2*

Ambient CO_2 equilibrates with a reagent and produces a change in the pH of the solution. The solution contains a pH indicator, *Chlorophenol Red*, which exhibits a sharp, nearly linear change in optical absorbency at 560 nm from pH 5 to pH 7. The changes in optical absorbency vary nearly linearly with the logarithm of partial pressure for CO_2.

The LED, common for both halves of the sensor, transmits light through the pH sensitive reagent to an active photodiode (PD1). The reference photodiode (PD2) is used to eliminate variations in light intensity.

2.7.3. *Turbidity sensors*

The purpose of turbidity measurement is to find the concentration of undissolved particles in the liquid on basis of light scattering. The scattering of light depends upon the ratio of the particle size and the wavelength of light. For particles less than 1/10 of the wavelength, scattering is uniform in all directions. With increasing particle size, scattering in the direction of the light propagation prevails. For a measurement in the water service, a particle with a diameter greater than 45 μm is considered undissolved.

Other factors affecting turbidity are the color and the particle form, the difference between the index of refraction of the liquid particle and the layout of the optical system, i.e. the measurement in the direction and at an angle of 90º with respect to the propagation. The majority of turbidity sensors start from measuring the luminous flux at an angle of 90º to the direction of propagation. Very good repeatability is ensured for a configuration (web [17]) with four modulated light beams (Figure 2.30). In the first phase source L_1 is switched on, the sensor of luminous intensity S_2 reads the transition and sensor S_1 reads the scattered light. The second phase (about 0.5s after first one) proceeds similarly with source L_2 switched on. Four independent results are acquired and both sensors alternate acting as a measurement of the transit and the scattered radiation. The microcomputer works out resultant turbidity by ratiometric measurement algorithm. This procedure excludes errors associated with variations in the sensitivity of the sensors and the intensity of light source. Errors due to different color particles and solutions are also reduced.

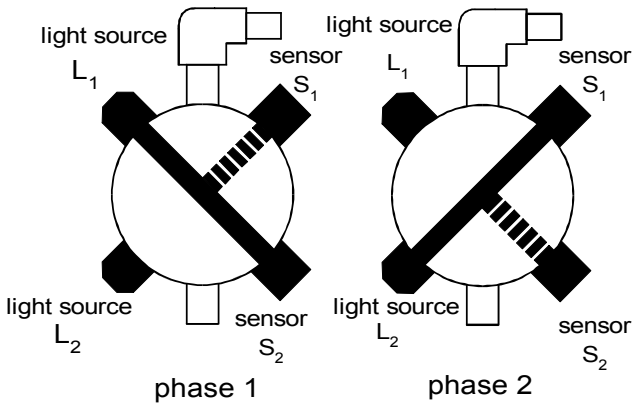


Figure 2.30. *The principle of the turbidity measurement*

The turbidity is often expressed in terms of the FTU (Formazin Turbidity Unit), since the suitable matter for the calibration is formazin (suspension created by polymerization of hexamethylentetranine and sulphate hydrazine).

2.8. Bibliography

2.8.1. Books

Optics-physical principles

[1] Meyer-Arendt, Jurgen R.: *Introduction to Classical and Modern Optics*, 2nd ed., Prentice Hall, 1984.

[2] Hecht, E.: *Optics*, 2nd ed., Addison Wesley, Reading, MA, 1987.

[3] Guenther, R.: *Modern Optics*, John Wiley 1990.

[4] Pedrotti, F.L., Pedrotti, L.S.: *Introduction to Optics*, Prentice Hall, 1987.

[5] Young, H.: *University Physics*, 8th ed., Addison-Wesley, 1992.

[6] Minnaert, M.G.J.: *Light and Color in the Outdoors*, Springer-Verlag, 1993.

[7] Jones, E.R., Childers, R.L.: *Contemporary College Physics*, 2nd ed., Addison-Wesley, 1993.

Light sources and photodetectors

[8] Waynant, R.W., Ediger, M.N. [eds]: *Electro-optics Handbook*, McGraw-Hill, 1994.

[9] Hewlett-Packard: *Optoelectronics Designer's Catalog*, Hewllet-Packard 1993.

Optical sensors – monographs

[10] Wagner, E., Dändliker, R., Spenner, K. [eds]: *Sensors*: A Comprehensive Survey, Volume 6, *Optical Sensors*, December 1991.

Books including topics related to "optical sensors"

[11] De Silva, C.: *Control Sensors and Actuators*, Prentice Hall, New Jersey, 1989.

[12] Doebelin, E.O.: *Measurement Systems: Application and Design*, 4th ed., McGraw-Hill, New York, 1990.

[13] Fraden, J.: *Handbook of Modern Sensors*, American Institute of Physics Press, Woodbury, New York, 1997.

[14] Kovacs, G.T.A.: *Micromachined Transducers Sourcebook*, McGraw-Hill, New York, 1998.

[15] Hentschel, C.: *Fiber Optics Handbook*, Hewlett-Packard GmbH, Boeblingen Instruments Division, October 1983.

[16] Heijden, F.: *Image Based Measurement Systems*, John Wiley, New York, 1994.

[17] Norton, H.N.: *Handbook of Transducers*, Prentice-Hall, Inc. 1989.

[19] Texas Instruments Inc.: *Intelligent Opto-sensors Data Book*, Texas Instruments, 1995, www.ti.com/dlp.

[20] Rutledge, G.J.: *An Introduction to Gray Scale Machine Vision*, Vision 1985, Machine Vision Association of SME, Deaborn, Mich., 1985.

[21] Drain, L.E.: *The Laser Doppler Technique*, Wiley-Interscience, New York, 1980.

Magazines publishing articles related to Optical sensors

[22] Sensors, The Journal of Applied Technology, An Advanstar Communication, Inc., One Phoenix Mill Lane, Suite 401, Peterborough, NH 03458, ww.sensorsmag.com.

[23] M&C – Measurement and Control, Measurement and Data Corporation, Editorial Office, 2994, West Liberty Ave., Pittsburgh, Pa. 15216.

2.8.2. *Physical background – websites*

[33] http://hyperphysics.phy-astr.gsu.edu.

Chapter 3

Flow Sensors

3.1. Introduction

Measuring the flow of liquids, gases, steam or solids is important both for the processing industry and for occasional readings. In some processes inaccurate flow-rate measurement can make the difference between profit and loss. In other cases inaccurate or erroneous flow measurements can have serious or even disastrous consequences.

3.1.1. *Volume flow and mass flow*

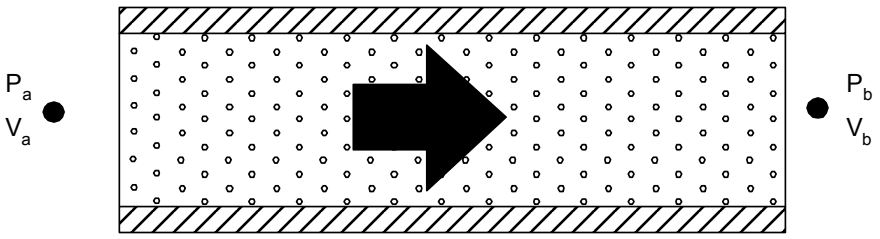


Figure 3.1. *Flow in a pipe*

Chapter written by R. MEYLAERS, F. PEETERS, M. PEETERMANS and L. INDESTEEGE.

The flow (Q) is defined as the amount of a substance that passes a certain point or a certain section during one time unit.

We make a distinction between volume-flow measurements and mass-flow measurements (Q_V and Q_m).

Many principles of *volume-flow measurements* use the following formula:

$$Q_V = v.A \tag{3.1}$$

where

– Q_V: volume flow [m^3/s]

– v: mean velocity [m/s]

– A: cross-sectional area [m^2]

The flow is determined by measuring the velocity or the change in kinetic energy of the medium. The velocity depends on the difference in pressure on a pipe or covering (Figure 3.1). This pressure difference pushes the medium through the pipe or covering. Because the pipe's diameter is known, the average velocity is a measure of the flow.

The International System unit for volume flow is m^3/s, but often m^3/h or l/h are used.

We note that in practice the flow is usually expressed as a volume flow and less as a mass flow.

When dealing with incompressible substances like liquids and solids, a simple relationship exists between the volume flow and the *mass flow*:

$$m = \rho.V \tag{3.2}$$

where

– m: mass [kg]

– ρ: density [kg/m^3]

– V: volume [m^3]

For a certain amount of a compressible substance (e.g. gas or steam), every temperature and every pressure corresponds to a different volume. That is why the

operating conditions (p,T) are always mentioned when dealing with the volume flow of gas or steam, or why the flow is reduced to its standard conditions. This means that the flow is given at p = 1 bar and T = 273 K instead of at the operating conditions. The conversion is performed with the help of the general gas law:

$$\frac{p_0 \cdot V_0}{T_0} = \frac{p.V}{T} \tag{3.3}$$

3.1.2. *Influences on the flow*

Other factors that affect the flow are the viscosity and the friction of the medium as it moves through the pipe. The output of flow meters strongly depends on the dimensionless unit, the *Reynolds number*. This is defined as the proportion between the force of inertia and the force of friction and can be derived as follows:

$$ReD = v \cdot D \cdot \rho / \eta = v \cdot D / \nu \tag{3.4}$$

where

– ReD: Reynolds number [1]

– v: velocity [m/s]

– D: diameter of the pipe [m]

– η: dynamic viscosity (friction coefficient) [Pa·s]

– ν: kinematic viscosity [m2/s]

The velocity, the density of the liquid and the diameter of the pipe determine the force of inertia, where a frictional force is experienced at the solid-liquid interface. The diameter of the pipe and the density remain constant for most applications.

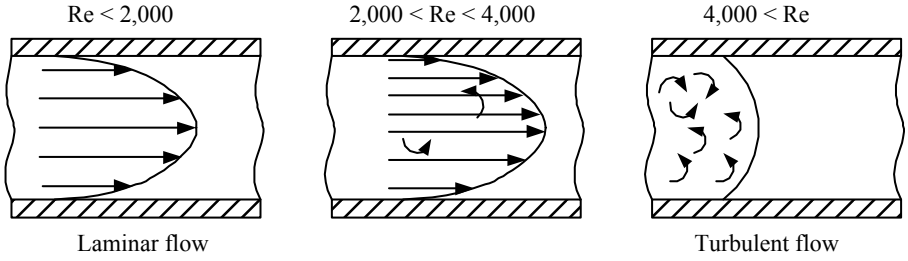


Figure 3.2. *Laminar versus turbulent flow*

At low speeds or with high viscosity, the Re-number is low and liquid runs smoothly in layers, with the highest speed at the center of the pipe and the lower speeds at the sides, where the force of friction blocks the liquid. This kind of flow is called *laminar flow*. The Re-number remains lower than 2,000. The typically parabolic change in velocity characterizes laminar flow (Figure 3.2).

However, most applications deal with *turbulent flow*, with Re-values above 4,000. Turbulent flow originates at high speeds and low viscosity. The flow turns into turbulence, with the same average velocity over the complete diameter of the pipe. The higher the Re-number, the steadier the velocity distribution.

There is a transition zone between laminar and turbulent flow ($2{,}000 < Re < 4{,}000$), which depends upon the pipe and on the velocity, i.e. the flow can be either laminar, turbulent or mixed laminar-turbulent. Knowing the type of flow for the application will later be used to determine the type of flow meter required.

A large diversity of media and a variety of measuring conditions including pressure, temperature, viscosity, etc., has resulted in a large market supply of flow meters and in the development of flow measuring principles. As such, it is a matter of deciding which type of flow meter is suitable to a particular application.

Based on the measured quantity and on the medium we can distinguish between:

– velocity measurements of liquids, gases and steam;

– volume-flow measurements of liquids, gases and steam;

– mass-flow measurements of liquids, gases and steam;

– flow measurements of solids;

– flow measurements of liquids in open channels;

– quantitative measurements.

3.1.3. *Bernoulli equation*

Several flow-measuring principles are based on the *Bernoulli equation*. Figure 3.3 shows a pipe through which an incompressible fluid runs. The velocity and the diameter of the pipe are such that we are dealing with a turbulent flow ($Re > 4{,}000$).

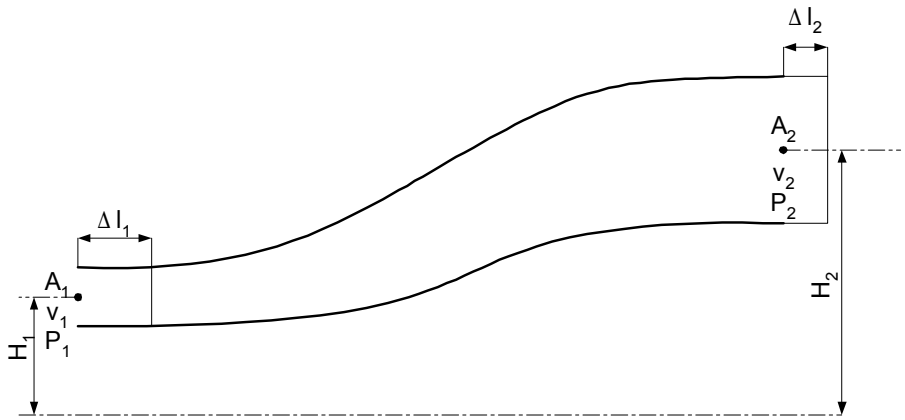


Figure 3.3. *Bernoulli equation*

Because the fluid is incompressible, the mass that enters the pipe through A_1 during one time unit must be the same as the mass that leaves the pipe through A_2 per time unit:

$$\rho \cdot A_1 \cdot v_1 = \rho \cdot A_2 \cdot v_2 \quad \text{or} \quad A_1 \cdot v_1 = A_2 \cdot v_2 \quad \text{or} \quad A \cdot v = \text{constant} \tag{3.5}$$

In addition, the total amount of energy (potential and kinetic) is a constant.

These are continuity equations. They are always valid for an incompressible fluid, because otherwise the fluid would accumulate or dilute in the pipe.

These equations lead to the more general formula:

$$p + \rho.g.H + \frac{1}{2}.\rho.v^2 = \text{const.} \tag{3.6}$$

This is called the *Bernoulli equation.*

The term p + ρ.g.H represents *static pressure*, in which ρ.g.H is the contribution of the hydrostatic pressure.

The term ½. $\rho.v^2$ represents *dynamic pressure*.

From the continuity equation and the Bernoulli equation we can conclude the following: at the position where the velocity is highest, the static pressure must be at its smallest, and vice versa. The Bernoulli equation thus gives us the relation between the velocity of a fluid in a pipe and the corresponding static pressure.

When we are dealing with a laminar flow profile, the flow law of Poiseuille can be applied. For a cylindrical pipe with radius R, a difference in pressure Δp comes into existence over a length l:

$$Q_v = \pi.R^4.\Delta P/8.\eta.l \tag{3.7}$$

with η: dynamic viscosity

Thus, especially for flow measurements that are based on the principle of difference in pressure, it is important to check whether the flow profile is laminar or turbulent.

3.2. Flow measurements based on the principle of difference in pressure

This is the oldest principle, and it is based on the fact that matter only flows in a certain direction when there is a moving force, i.e. a potential, that differs at every point of the flow. In many cases this potential is the pressure that declines from the highest to the lowest value, at which the decline in pressure is evenly spread over the complete distance of the flow.

However, local declining pressure elements can be positioned in the path of a flow by using constrictions, originating in a difference in pressure in front of and behind the element (see (3.6)). These elements are called primary elements. The most common are:

– the Prandtl tube;

– the diaphragm or orifice plate;

– the Venturi tube;

– the Dall tube.

Flow measurement, using the principle of difference in pressure is standardized by ISO 5167 and DIN 1952 (Figure 3.4). This contains geometry, configuration and calculating methods.

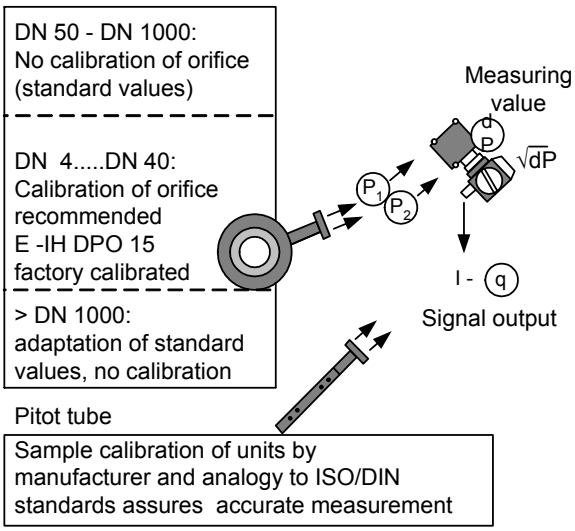


Figure 3.4. *Standardization DIN 1952*

3.2.1. *The Pitot and Prandtl tube*

3.2.1.1. *Principle*

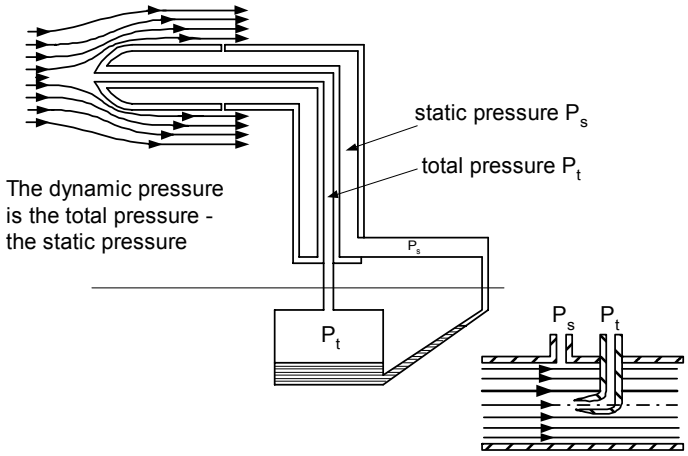


Figure 3.5. *Pitot tube (double-walled construction as Prandtl tube on the left, construction with two separate tubes on the right)*

The Pitot tube is used to measure the rate of flow of a liquid, steam or gas stream (Figure 3.5).

The principle is based on the measurement of the pressure difference between two surfaces, of which one is perpendicular to the direction of the flow and the other runs parallel. At this last surface the flow is almost undisturbed, which means that here we can measure the actual pressure of the medium, the so-called *static pressure* P_s.

When the surface is perpendicular to the direction of the flow, the liquid that collides with it is slowed down until it stands still. At this stagnation point a higher pressure is measured, the *total pressure* P_t. The increase in pressure or the *dynamic pressure* $P_d = P_t - P_s$ is derived from the kinetic energy of the liquid.

It can be calculated using the Bernoulli equation:

$$P_d = \rho . \frac{v^2}{2} \tag{3.8}$$

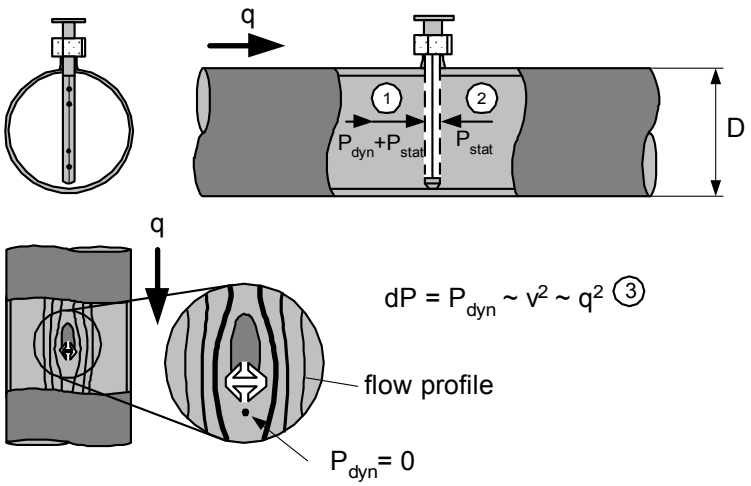


Figure 3.6. *Pitot tube with several measuring points*

Or it can be converted to velocity and measured pressure:

$$v = \sqrt{\frac{2.(P_t - P_s)}{\rho}} \tag{3.9}$$

Note that the velocity depends on the density of the fluid, and therefore depends on the pressure and temperature.

We also see that the velocity depends quadratically on the measured difference in pressure. When a linear output signal is desired, an additional conversion needs to be performed.

The flow can be determined by multiplying this velocity with the flow surface:

$$Qv = v.A \tag{3.10}$$

Of course, this is only valid when the velocity is as high everywhere on the flow surface, or when the flow is turbulent (Re-number sufficiently high).

To reduce the influence of the flow profile, a Pitot tube with several measuring points can be used (Figure 3.6). The advantage is that an average pressure difference is measured (average velocity).

3.2.1.2. *Practical set-up*

Depending on the fluid, the Pitot tube needs to be positioned differently in the pipe. In addition, the position of the pressure gauge depends on the fluid (Figures 3.7, 3.8 and 3.9):

– vertical pipe: horizontal Pitot tube;

– horizontal pipe and liquid: vertical Pitot tube, pressure gauge under the pipe;

– horizontal pipe and gas: vertical Pitot tube, pressure gauge above the pipe;

– horizontal pipe and steam: horizontal Pitot tube, pressure gauge under the pipe.

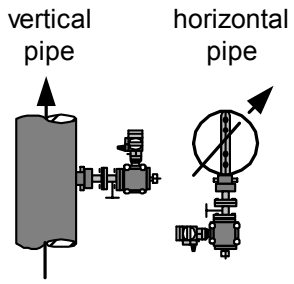


Figure 3.7. *Fluid = liquid*

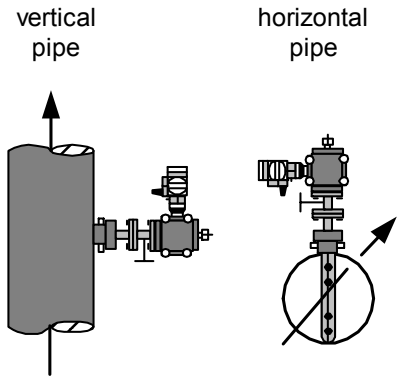


Figure 3.8. *Fluid = gas*

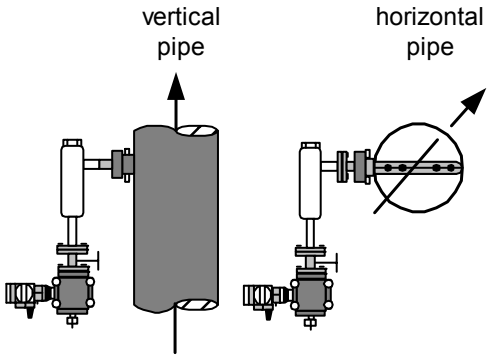


Figure 3.9. *Fluid = steam*

When measuring steam, condensation barrels need to be placed between the measuring element (Pitot tube) and the pressure gauge (e.g. ΔP-cell). This is to ensure that the hot steam does not touch the membrane of the pressure gauge, because that could damage the gauge.

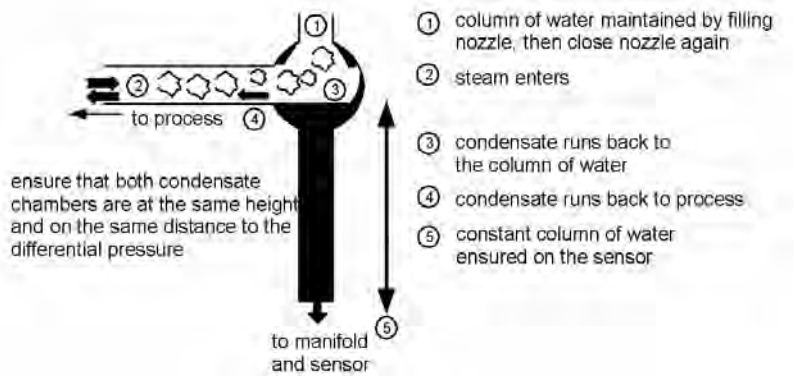


Figure 3.10. *Principle of a condensation barrel*

3.2.1.3. *Characteristics*

Advantages:

– low installation costs, even for already existing installations;

– very low pressure drop.

Applications:

– measuring the velocity of air in channels that are not equipped with permanent measuring apparatus (portable reading instrument);

– measuring on-site to detect errors;

– determining the velocity profile and the flow profile in a pipe;

– suitable for larger pipe diameters: DN200 – DN12 000 (however, it is also used for smaller diameters).

3.2.2. *The orifice plate*

3.2.2.1. *Principle*

Without any doubt the orifice plate is the most utilized flow meter, mainly because of its simplicity, its low cost and the many years of experience of using it.

Figure 3.11 demonstrates a typical set-up of the orifice plate.

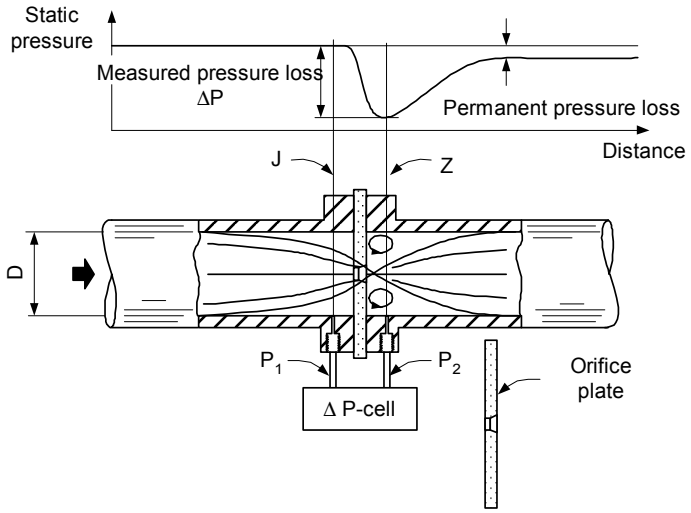


Figure 3.11. *The orifice plate*

Calculating the flow

When we consider the turbulent flow of a one-dimensional, incompressible liquid without heat or work changes, we can describe the flow Q_v as follows:

$$Q = A_1 \cdot v_1 = A_2 \cdot v_2$$

$$A_1^2 \cdot v_1^2 = A_2^2 \cdot v_2^2$$

$$v_1^2 = \left(\frac{A_2}{A_1}\right)^2 \cdot v_2^2$$

For a horizontal pipe the Bernoulli equation is reduced to:

$$Q = A_2 \cdot v_2 = \frac{A_2}{\sqrt{1 - \left(\frac{A_2}{A_1}\right)^2}} \cdot \sqrt{\frac{2 \cdot (p_1 - p_2)}{\rho}} \tag{3.11}$$

with:

– A_1, A_2: the cross-section of the "flow" where p_1 and p_2 are [m^2]

– ρ: the density [kg/m^3]

– p_1, p_2: the static pressure [Pa]

The formula shows that to measure Q_v, the values of A_1, A_2 and ρ need to be known and that P_1 and P_2 must be measured. We also see that, similar to the Pitot tube, the measurement depends on the density of the fluid.

3.2.2.2. *Practical installation*

The bores (drainage points) for pressure readings can be put at three different positions: at the flanges, at the vena contracta and in the pipe itself (Figure 3.12). The connection at the flanges is predominantly used:

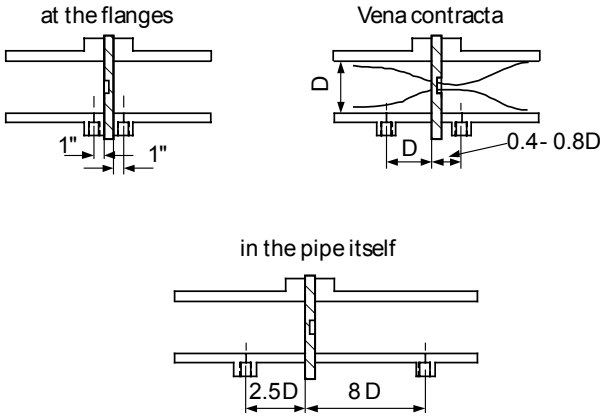


Figure 3.12. *Orifice plate: pressure connections*

– Depending on the fluid, the bores (drainage points) that connect the measuring element with the pressure gauge, need to be positioned differently (Figures 3.13, 3.14 and 3.15).

– Horizontal pipe and liquid: horizontal drains or at the bottom (if there are no sinkable particles in the liquid).

– Horizontal pipe and gas: drains at the top to avoid accumulation of condensation.

– Horizontal pipe and steam: horizontal drains and condensation barrels (Figure 3.15) in the connection to the pressure gauge.

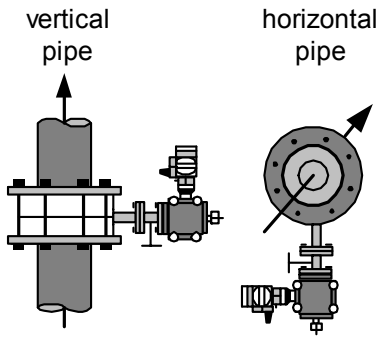


Figure 3.13. *Fluid = liquid*

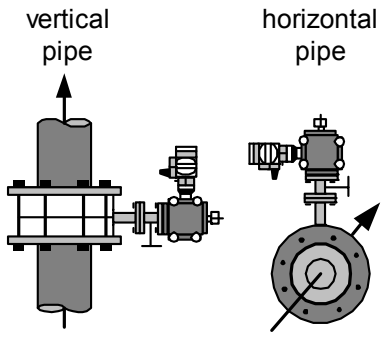


Figure 3.14. *Fluid = gas*

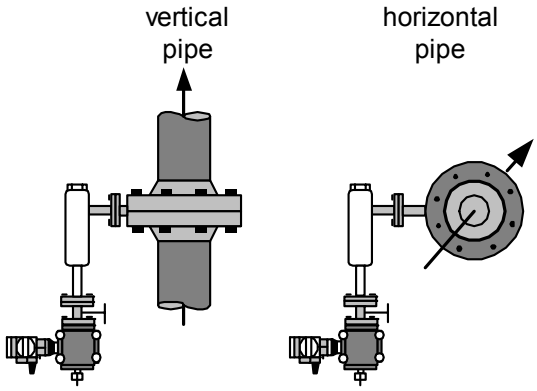


Figure 3.15. *Fluid = steam*

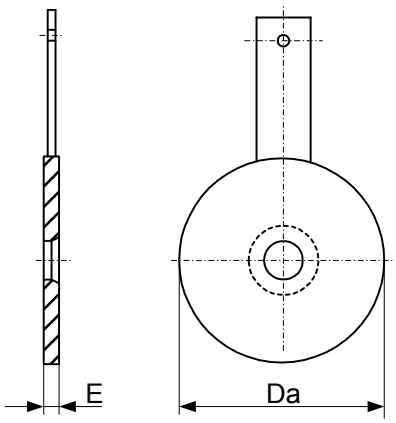


Figure 3.16. *Traditional measuring orifice*

– The orifice plates are either concentric, eccentric or segmental. Eccentric plates are applied for rapidly condensing gases, so that drainage becomes redundant. The opening is situated at the bottom of the pipe, where the condensation passes the orifice. Segmental orifice plates are used for saturated steam, oil containing water particles and liquids containing solid particles (Figures 3.16 and 3.17).

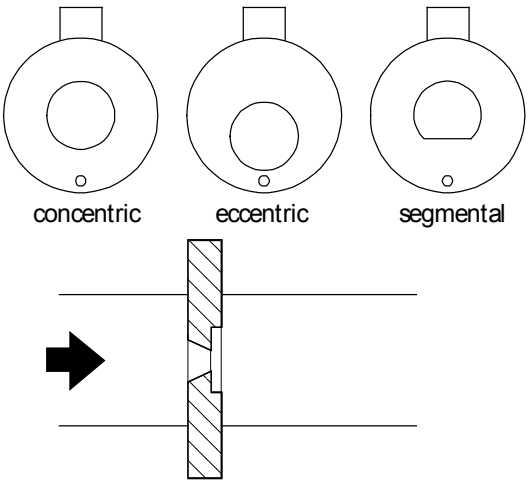


Figure 3.17. *Realizations of orifice plates*

– The square-edged orifice plate is used when the Re-number is smaller than 10,000. As mentioned before, the pressure loss coefficient will vary too much when using a standard metering orifice.

– The orifice plate is usually made of stainless steel, the opening is turned sharp and at a right angle and must be free of dirt. The diameter ratio β = d/D preferably lies between 0.3 and 0.7. The small valve opening allows the passage of gas pockets with fluids, and is situated at the top. With gases it allows the passage of condensations and is situated at the bottom of the pipe.

– When using a metering orifice, it is important to be careful that the decline in pressure is not smothered under the vapor pressure of the liquid. Vaporization would take place and the reading would be unreliable.

– To obtain an evenly distributed flow profile, a straight pipe in front of and behind the orifice plate is essential. These equalization pipes depend on the β-value, the installation and the kind of flange used, and are usually chosen by the manufacturers define them. As a rule of thumb this applies to a straight pipe of 10 to 15 times diameter, D, in front of the flange and 5 to 10 times behind the flange.

3.2.3. *The flow nozzle*

The flow nozzle (Figure 3.18) is a variation on the orifice plate and is used to measure larger flows. For the same flow and the same β-proportion as with a metering orifice, the pressure difference will be smaller, or for the same β-value and pressure difference the flow nozzle will allow for approximately 65% more flow than the orifice plate. The construction is more robust and has a better resistance against corrosion. The installation is less critical than the orifice plate; usually more attention is paid to the correct composition and smooth sides. For the connections of measuring pipes and equalization pipes, the same rules apply as for metering orifices.

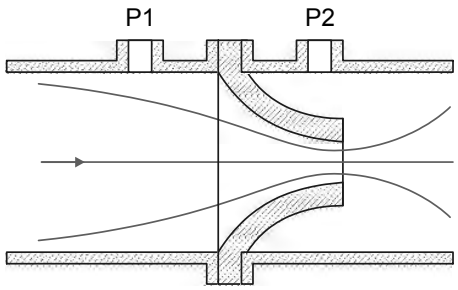


Figure 3.18. *Flow nozzle*

3.2.4. *The Venturi tube*

The Venturi tube (Figure 3.19) is also a variation on the orifice plate and it is applied when a small permanent pressure loss is a priority. It is also used when the liquid or gas contains many solid particles and for liquids with a high level of viscosity. A measurement with a Venturi tube is less influenced by changes in viscosity. The Venturi tube is a rather expensive instrument.

The design of the Venturi tube connects best to the flow profile of a substance, and only a minimum of turbulence occurs, which is beneficial in terms of the permanent loss of pressure.

The Venturi tube consists of a gradually narrowing part, AB, a cylindrical narrowing part, BC, and the diffuser, CE. The length of BC equals the diameter of the pressure connection increased by a length of 0.2 to 0.4 times the diameter of the pipe. Half the opening angle of the diffuser, CE, must be smaller than 150°.

As the opening angle of AB and CE diminishes, the permanent pressure loss decreases.

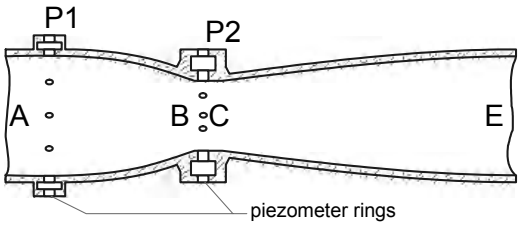


Figure 3.19. *Venturi tube*

3.2.5. *The Dall tube*

The Dall tube (Figure 3.20) is used when the permanent loss of pressure needs to be even smaller than with the Venturi tube. An additional advantage is the shorter length of the Dall tube for the same diameter. Venturi and Dall tubes are rarely used. This is due to the long built-in length and the high costs. Because of the small permanent pressure loss, these tubes are used for larger flows, such as waste water pipes and sewers, for which the standard metering orifice does not qualify. The additional charges of the instrument are recovered over a longer period of time by the energy savings resulting from the elimination of high-pressure losses. Individual calibration is required for each of these instruments, because the different designs (compared to an orifice plate) do not allow accurate reproduction.

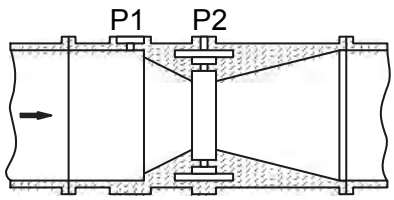


Figure 3.20. *Dall tube*

3.2.6. *General guidelines for a correct reading*

The process and the pipes must comply with certain rules for a sufficiently accurate measurement:

– Completely filled pipes.

– No interfering objects or marks in the tube.

– Relatively constant process conditions (pressure and temperature).

– A homogenous fluid.

– Adequate equalization pipe (straight pipe in front of and behind the measuring element, to equalize the flow).

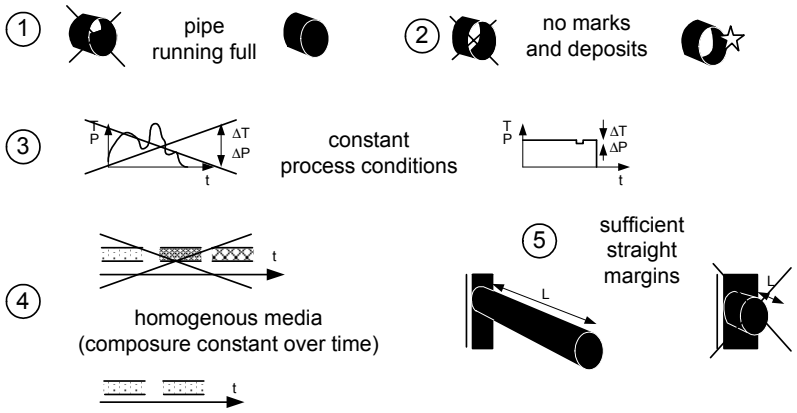


Figure 3.21. *General conditions for an accurate measurement*

In Table 3.1 specific numbers for flow measurement with the Pitot tube, the orifice plate, the flow nozzle and the Venturi tube are given. The different principles can be compared with each other using these numbers.

	Orifice plate	Pitot tube	Nozzle	Venturi
Pipe DN	4 to 2,00	25 to 12,00	50 to 600	100 to 1,00
Pipe form	round	round or square	round	round
Sensitivity to abrasion	sharp plate edges wear, but easy to replace	lower	low	low
Sensitivity to dirt	can gather upstream	lower	low	lowest
Min. Re	2,800	4,000	200,000	40,000
Accuracy at constant density	1%	1.5%	1.5%	2%
Straight length	10 to 16 D	30 to 50% lower than orifice	10 to 16 D	50% lower than orifice
Price	low	lowest	average	high

Table 3.1. *Specifics of flow measurement with several pressure difference sensors*

3.3. Flow measurements based on variable passage

3.3.1. *The float flow meter (rotameter)*

3.3.1.1. *Principle*

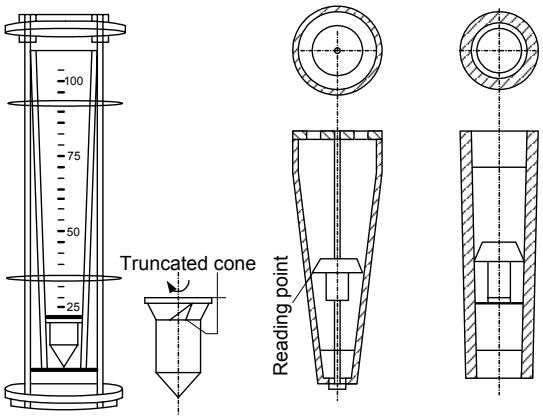


Figure 3.22. *The float flow meter or rotameter*

The float flow meter has a variable passage surface and a nearly constant decline in pressure (Figure 3.22). The variable rotaflow meter consists of a tube, the inside of which is a truncated cone. A float is located in the tube, which is forced upward by the liquid or the gas to be measured and kept balanced at a certain level. The design and the material of the float differ with each manufacturer and depend on the properties of the substance, the flow, the flow profile and the pressure in the pipe. To avoid toppling the float in the conical tube, it is either led along a central rod or an oblique notch is cut to give the float a continuous rotation.

The conical tube is usually made of strong glass to allow a direct reading. When higher pressures occur, a metal tube can be used, from which the position of the float is detected by means of a magnetic field or a movement reader.

The flow of a liquid through a rotaflow meter is given by:

$$Q_v = k_1(ah + bh^2)\sqrt{\frac{\rho_2 - \rho_1}{\rho_1}} = C_d \cdot (A_1 - A_2)\sqrt{\frac{2 \cdot g \cdot V}{A} \cdot (\frac{\rho_2}{\rho_1} - 1)} \qquad (3.12)$$

where

– ρ_1: density of the liquid

– ρ_2: density of the float

– A_2: surface area of the float

– V_2: volume of the float

– k_2: coefficient caused by the type of the float

– C_d: the frictional loss coefficient

– A_1-A_2: the surface area of the ring between the float and tube

A_2 is variable; all the other factors remain constant for a particular instrument and the same liquid.

The formula demonstrates that the reading depends on the *density* of the liquid. However, when we make the float with a material of which the density is twice the density of the liquid, we can ignore the density changes of the liquid.

The frictional loss coefficient, C_d, is influenced by the viscosity of the substance to be measured. As long as the flow is laminar, the value of C_d will change greatly. A turbulent flow results when the value of C_d is constant. That is why the type of flow also defines the design of the float.

3.3.1.2. *Characteristics*

– The visible float also controls the operation.

– Quasi-linear scale.

– No need for equalization pipes.

– Small and constant pressure loss.

– Vertical installation.

– Not equipped for non-transparent substances.

– Dependent on the decline in pressure over the reading instrument and on the temperature of the fluid.

3.3.2. *Target flow meter*

The target flow meter was originally designed to be used where other flow meters are inadequate, as with highly viscous liquids, at extreme temperatures, gas flows at high speed.

3.3.2.1. *Principle*

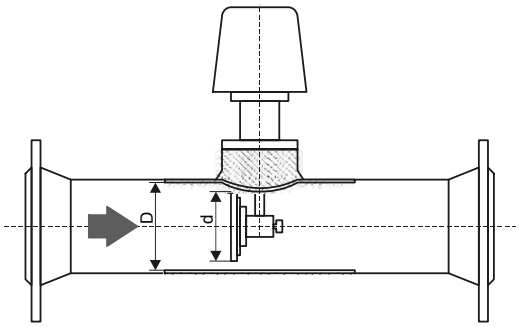


Figure 3.23. *Target flow meter*

The flow meter consists of a target, at which the liquid or the gas exerts a force. We can calculate the flow in function of the exerted force:

$$Q = C_d . \frac{\pi}{4} . (D^2 - d^2) . \sqrt{\frac{2.F}{K.A.\rho}} = C' . \frac{D^2 - d^2}{d} \sqrt{\frac{2.F}{\rho}} \qquad (3.13)$$

$$A = \frac{\pi}{4} d^2$$

where

– K: proportionality constant

– v: velocity of the fluid

– ρ: density of the fluid

– A: target surface area

From the equation it is shown that the flow is proportional to the force that the liquid exerts on the target. To obtain a linear transfer, we need a square-root extractor.

A dynamometer or a movement converter can perform the conversion to an electrical or pneumatic signal. At the assembly of a target flow meter, the same precautions should be taken as for the metering orifice.

3.3.2.2. *Characteristics*

– Well-adapted for viscous substances.

– Relatively inexpensive for extreme measurements (high viscosity).

– Long equalization pipes are required.

– Low rate of accuracy (5% FS).

– Individual calibration is required for each application and each substance.

3.4. Turbine flow meter

3.4.1. *Principle*

This flow meter owes its name to the turbine that is positioned in the flow line (Figure 3.24). The velocity of the substance to be measured makes the turbine rotate at a speed that with a minimum of friction is proportional to the velocity and thus to the flow of the fluid.

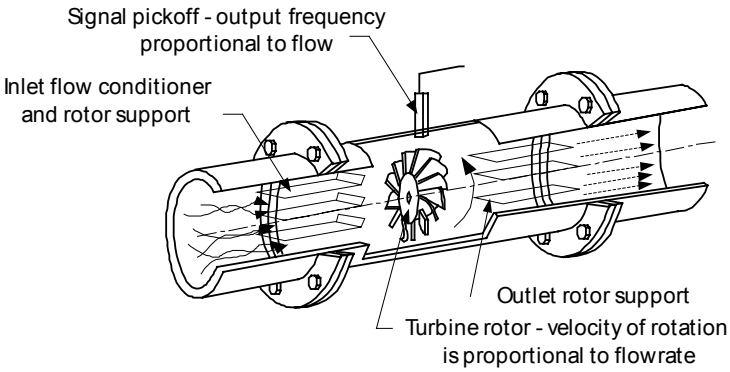


Figure 3.24. *Turbine flow meter*

A reader or a "pick-off" is installed perpendicular to the rotor. We distinguish two types of pick-off assemblies: magnetic pick-off and no drag pick-off.

The magnetic pick-off consists of a permanent magnet with a coil; the coil is wrapped around the magnet. When the turbine rotor blades cut through the magnetic field, an alternating current, the frequency of which is proportional to the flow, is induced in the coil.

The no drag pick-off consists of an oscillator that transmits a high-frequency carrier wave to the coil of the pick-off. The rotation of the turbine modulates the carrier wave depending on the velocity of the rotor. In both cases pulses proportional to the flow are produced.

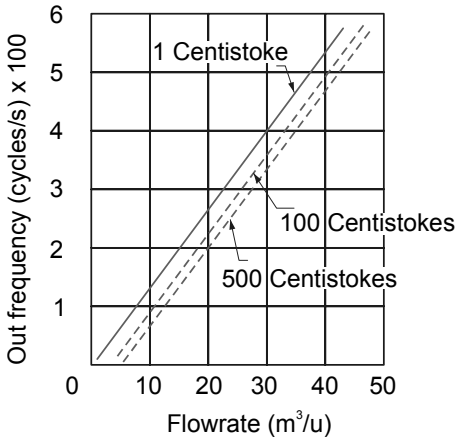


Figure 3.25. *Output frequency and viscosity*

Every turbine flow meter is characterized by the *K-factor*, a coefficient that for a specific flow of a particular fluid shows how many pulses per liter are generated by the flow meter.

The K-factor can be calculated from:

$$K = \frac{60.f}{Q_v} \tag{3.14}$$

where

– Q_V: the flow [liter/minute]

– f: frequency [pulse/s]

– K: K-factor [pulse/liter]

depending on the fluid viscosity. The typical characteristics of the turbine flowmeter are shown in Figure 3.25.

3.4.2. *Practical installation*

– The turbine flow meter requires equalization pipes in front of and behind the instrument. The length of the horizontal pipes (Figure 3.26) that are assembled in front of and behind the rotor depends on the flow conditions. As a rule of thumb an equalization pipe of 15D upstream and one of 5D downstream is postulated.

– The accuracy of the instrument is strongly influenced by the workmanship of the blades and the friction of the rotor against its axis.

– Inertia of the rotor can greatly influence the response time, especially when dealing with gases.

– To perform the digital-to-analog conversion, a frequency-to-voltage transformer can be used, which transforms the pulses into a standard electrical signal.

– To stop any pollution that might block the turbine, a filter can be placed in front of the turbine flow meter if desired.

– Providing a bypass is efficient for continuous processes as it allows the replacement and cleaning of the flow meter (and of the filter) without interrupting the process.

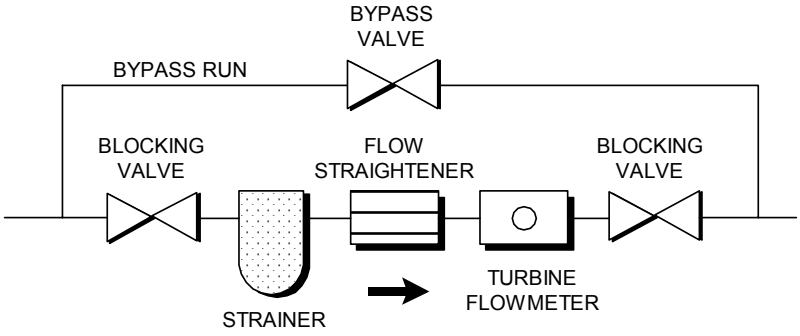


Figure 3.26. *Installation of the turbine flow meter*

3.4.3. *Characteristics*

– An extensive selection of ranges is available, for gases as well as for liquids.

– A high level of accuracy, (0.2-0.3%), is attainable under specific circumstances because of the digital output.

– Very sensitive to wear, especially with highly polluted substances and at high speeds.

– High reproducibility, but linear only in a limited area, which reduces the measuring range.

– The paddlewheel flow meter is a variation (Figures 3.27 and 3.28).

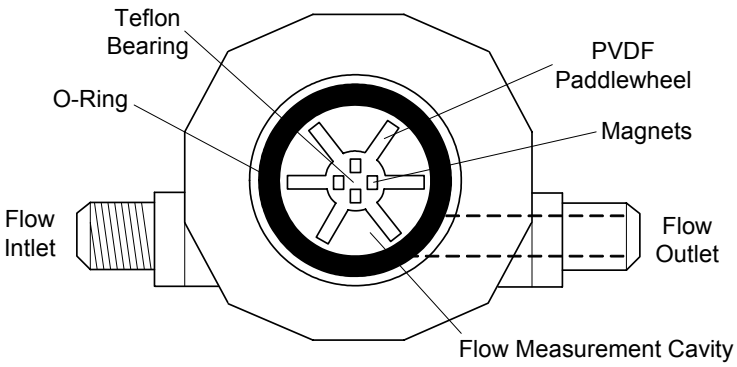


Figure 3.27. *Paddlewheel flow meter*

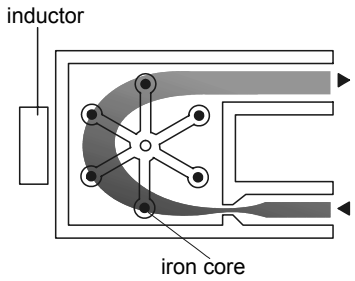


Figure 3.28. *Paddlewheel flow meter*

3.5. The mechanical flow meter (positive displacement)

3.5.1. *Principle*

The positive displacement mechanical flow meter is based on the overpressure of the flow at the inlet of the flow meter. This overpressure sets the rotating chambers of the flow meter in motion, which diverts a portion of the volume from the entry point to the outlet (reverse pump principle).

This type of measuring instrument is normally used to measure the total volume rather than the flow. A simple mechanical meter with reductions counts the total amount of rotations that are proportional to the flow or to the volume that runs through the pipe.

The accuracy level is about 1.5%, while the permanent decline in pressure at a maximum flow is 0.5bar at most.

Most common types are the oval cogwheel meter and the annular piston meter.

The oval cogwheel meter

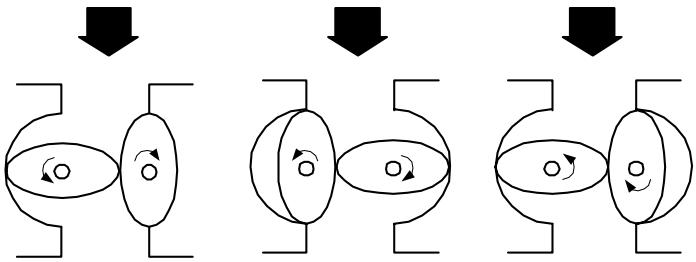


Figure 3.29. *The oval cogwheel meter*

The oval cogwheel meter consists of a housing and two cogged, elliptical wheels, of which one is equipped with an outgoing axis. The wheels set a volume of liquid in motion which equals the volume of the measuring chamber multiplied by the number of revolutions of the outgoing axis. The outgoing axis is connected with a revolution/mechanical counter to read the flow/total volume. The cogs on the oval wheels hitch tightly into each other, which minimizes possible leaks. The high couple at the axis gives the revolution counter an ideal start-up characteristic and causes minimal pressure loss. The measuring instrument does not depend on the viscosity and the temperature of the fluid. The housing, oval wheels, axes and sealing are available in several materials depending on the application.

The annular piston meter

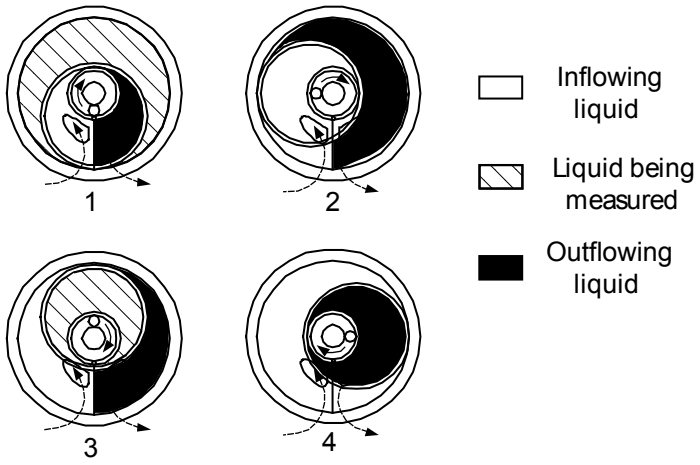


Figure 3.30. *The annular piston meter*

The annular piston meter consists of a stainless steel covering and a piston.

The working principle can be explained by means of the following four phases:

– The liquid enters the inner chamber of the piston through the inlet, forcing the piston in the direction of the arrow at a speed that is proportional to the originated differential pressure.

– The liquid volume on the outside of the piston moves and leaves the flow meter through the outlet. The last volume of liquid enters the inner chamber through the outlet.

– The inner chamber of the piston is completely closed. The liquid stream in the outer chamber makes the piston rotate.

– The volume in the piston escapes through the outlet.

The housing can be opened for cleaning or for replacing the piston. Available piston materials include Teflon, aluminum, bronze, titanium, hard rubber, polypropylene and glass fiber.

3.5.2. *Characteristics*

The piston meter is suitable for measuring small volumes. The accuracy is not dependent on the viscosity (going from fuel oil to pastes).

Other characteristics are:

– High level of accuracy.

– Little dependence on viscosity.

– Lifespan depends on the measured substance (solid particles, temperature differences).

– Permanent pressure loss over meters and filters.

– Sensitive to overloading.

– Blocked pipes in case of a mechanical fault.

– Relatively expensive.

– Used for aggressive substances, pulsated flow.

– Extensive selection available.

3.6. Magnetic flow meter

3.6.1. *Principle*

The principle of the magnetic flow meter is based on Faraday's law:

$$U_e = B.L.v \tag{3.15}$$

where

– B: magnetic field (flux density) [Tesla]

– L: distance between the electrodes [m]

– v: velocity of the flow [m/s]

When an electrically conductive medium (that contains free electrons or ions) moves in a magnetic field, an electric field and thus a potential difference appears in this conductor. The electric field is perpendicular to the magnetic field and perpendicular to the direction of the movement (left hand rule). The potential difference is proportional to the power of the magnetic field and to the velocity of the movement.

It is important to note that the generated voltage does not depend on parameters such as pressure, temperature, viscosity, conductivity, etc. Only a minimal level of conductivity is required to give this signal a (very small) minimal power. This signal is proportional to the volume flow.

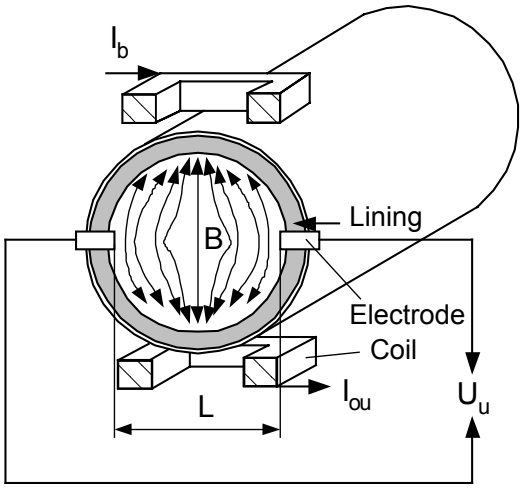


Figure 3.31. *Magnetic flow meter*

The traditional magnetic flow meter consists of two units: the measuring probe (Figure 3.31) and electronic converter-amplifier that transforms this mV signal into one or more standard analog or digital signals. More and more models, in which these two units are combined together (e.g. E+H Pulsmag, Picomag) are appearing on the market.

3.6.2. *Construction of the measuring instrument*

The measuring probe consists of two electrodes made of a non-magnetic material and is positioned at the inside covered with an electric insulating layer. There are several possibilities for this insulating layer:

– Natural rubber or neoprene is inexpensive but it has only a limited strength at higher temperatures. Rubber is mainly used for water applications.

– PTFE (Teflon) is a good-insulation material, but it is difficult to affix to steel, which means that it cannot be used in vacuum applications. PTFE is permeable by particular liquids, which can cause short circuits and corrosion.

– PFA (Per Fluor Alkoxy): liquid penetration is less common for this material, yet it is not dimensionally stable. Therefore it has to be provided with a (RVS) steel strengthening mat. PFA is more expensive than PTFE but it can be used in vacuum.

– Ceramic materials are non-permeable, dimensionally stable, wear-resistant and can resist very high temperatures. The disadvantages of these materials are the mechanical vulnerability and the high manufacturing costs. Temperature shocks and mechanical tensions in the material caused by an inaccurate set-up can cause bursting or breaking of the probe.

The electromagnetic coils are situated above and below the probe. The outside covering of the coils is made of carbon steel, which results in the concentration of the magnetic field inside the meter. In this way there are no leakage fields outside the meter. A disadvantage is that the meter becomes quite heavy. The two electrodes are in the direction perpendicular to the axis of the field coils. The electrodes are fixed to the lining and are in contact with the liquid. The basic material for these electrodes is stainless steel, although other materials can be used, such as titanium, hastelloy-c or platinum.

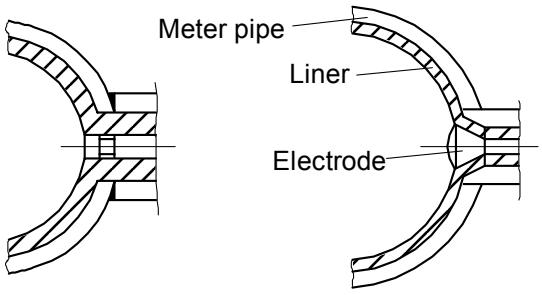


Figure 3.32. *Electrodes*

The magnet coils generate a magnetic field that depends on the form of the field excitation signal. There are several types of flow meters including:

– Electromagnetic flow meters with DC field excitation: only applicable with liquid metals (be aware of electrolysis). Hardly ever used.

– Electromagnetic flow meters with sinusoidal field excitation: the alternating field is produced with the aid of an alternating current with the same frequency as the mains voltage. The disadvantage of these applications is that the electronic noise causes the zero point to drift after a certain time. It is essential to manually readjust the zero point at regular time intervals.

– Electromagnetic flow meters with a switched DC field: here the converter feeds the magnetic coils with a switched DC (low frequency square wave) current. Because the converter is provided with the necessary intelligence, it can independently control the zero point, so that the zero point is stable. This method has lower power consumption (5-25W).

– Electromagnetic flow meters with capacitive detection: here the electrodes, which are in contact with the medium in standard models, are replaced by capacitive plates that are casted in the lining and that function as the electrodes of a capacitor. For this model the minimum required conductivity is 100 times lower than for the model with the contact electrodes.

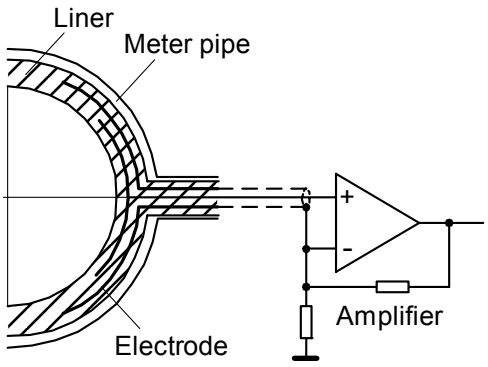


Figure 3.33. *Capacitive detection*

3.6.3. *Practical installation*

The installation of the flow sensor may occur in every position, as long as the measuring instrument is completely filled with the fluid (also the adjustment of the

zero point requires a complete filling). Several methods are available to obtain this complete filling (Figures 3.34 and 3.35).

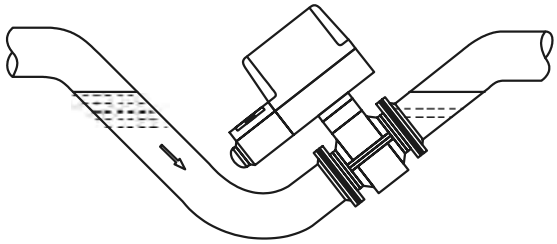


Figure 3.34. *Solution for a completely filled pipe*

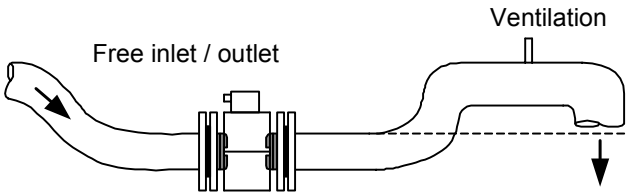


Figure 3.35. *Completely filled pipe for free outlet*

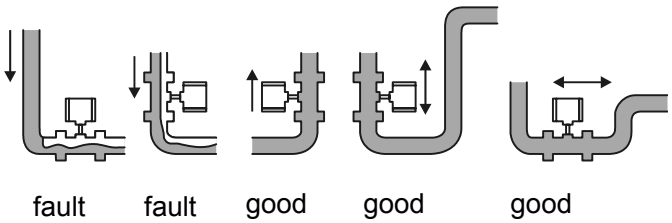


Figure 3.36. *Position of the flow meter in the pipe*

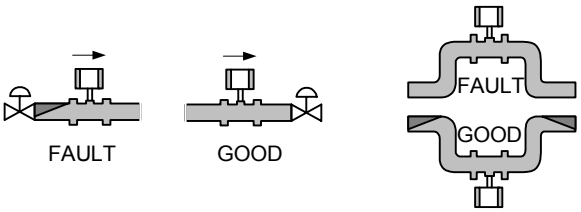


Figure 3.37. *Position of the flow meter for gas pockets*

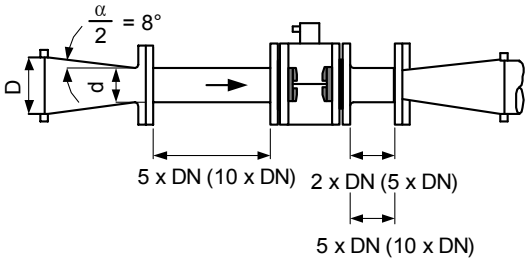


Figure 3.38. *Reduction of the pipe diameter and equalization pipe*

If the fluid contains solid particles or fats, the magnetic flow meter is best positioned vertically. When placing it horizontally, the heavier particles will precipitate and the lighter particles will come up. This can cause pollution that influences the magnetic field and consequently also the measurement. Positioning electrodes vertically in a horizontal flow meter is absolutely forbidden. Pollution of the electrodes will render a reading totally erroneous (Figures 3.36, 3.37 and 3.38).

Theoretically the direction of the flow is not important, as long as the correct electric connection is used (most manufacturers, however, will indicate a favorable flow direction).

At a low rate of flow (< 1 m/s) the desired accuracy cannot be obtained. The velocity can be increased by reducing the pipe diameter (max. angle = 8°, Figure 3.38).

The flow profile is not very important for the magnetic flow meter. In practice, however, most suppliers recommend an equalization pipe that is 5 times the pipe's diameter (Figure 3.38).

The grounding is of vital importance. This is very important because the voltage on the electrodes amounts to only a few mV. Friction of the liquid against the pipe can cause static voltage.

3.6.4. *Characteristics*

– Completely obstruction-free and therefore no pressure loss.

– High level of accuracy, usually better than 1% FS as long as the rate of flow is sufficiently high (>1 m/s). The minimum and maximum measurable velocities are

0.3 and 12 m/s. As a general rule it is recommended to choose a velocity of 2 to 3 m/s at full scale (reduction of the pipe if necessary).

– Wide span, good linearity.

– Measuring principle not dependent on pressure, temperature or viscosity.

– No equalization pipes (theoretically, but recommended in practice).

– Expensive because of electronics.

– Minimum conductivity required (5μS/cm) and therefore only practically used when dealing with liquids (at capacitive detection the required conductivity is 0,05μS/cm).

– Not dependent on the flow profile when using a measuring instrument with characterized, weighted field.

– No mechanically moving components, maintenance-free.

– Ideal for polluted liquids.

– Possibility to clean on the spot, together with the pipes.

3.7. The vortex flow meter

3.7.1. *Principle*

The vortex flow meter uses the phenomenon of vortex shedding that occurs when a fluid (steam, gas or liquid) flows along a non-streamlined object which is called the bluff body (Figure 3.39). The cylindrical streamline of the fluid is not capable of following the outlines of the bluff body and downstream some vortices come into being in the so-called Karman vortex street. The vortices separate themselves alternately at each side of the bluff body, with a frequency proportional to the average velocity of the fluid running through the pipe.

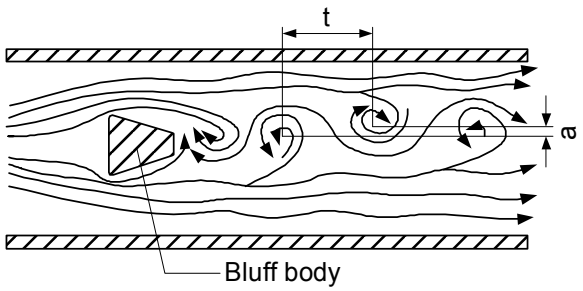


Figure 3.39. *Vortex flow meter*

Sensors, like thermistors, piezoelectric or capacitive cells and ultrasonic measuring elements, detect the resulting velocity or pressure pulses that are generated by the shedding of the vortices. The corresponding output signal comes from local electronics. Vortex flow meters coming from different manufacturers distinguish themselves by the shape of the bluff body, the type of sensor, the place of detection and the electronics.

Vortex flow meters have been used in industrial flow measurements since 1970. Bluff bodies with a better signal-to-noise ratio were developed and sensors were notably improved. Many technicians consider the vortex flow meter as the future solution for measuring non-conductive fluids and as the replacement of the metering orifice and other Δp measurements.

3.7.2. *Construction of the vortex flow meter*

The vortex flow meter consists of the following components:

– The bluff body: the vortex is generated at the back of the bluff body (Figure 3.40). The linearity of the proportion of v/f (where the velocity is proportional to the vortex frequency) depends on the shape and the dimensions of the bluff body.

– Round bluff body: the original bluff bodies were cylindrical. The shedding point of the vortex fluctuated upwards and downwards with the rate of flow. Because of this the frequency was not proportional to the velocity.

– Delta bluff body: many tests have revealed that the linearity of the delta shape is very good. The vortex shedding angle is outlined clearly. Pressure variations, viscosity or other process parameters do not affect the level of accuracy.

– Double bluff body: this body is obtained when the manufacturer connects the sensor to the bluff body. The secondary section moves and the Karman vortex street is transformed into a twisted movement. Another possibility is to place two bluff bodies after each other. In this case the permanent pressure loss is doubled, but a stronger vortex is generated (hydraulic amplification). This means that fewer complex sensors and amplifiers can be used.

– Rectangular bluff body: when instruments are equipped with these bodies, the linearity fluctuates greatly with the process parameters.

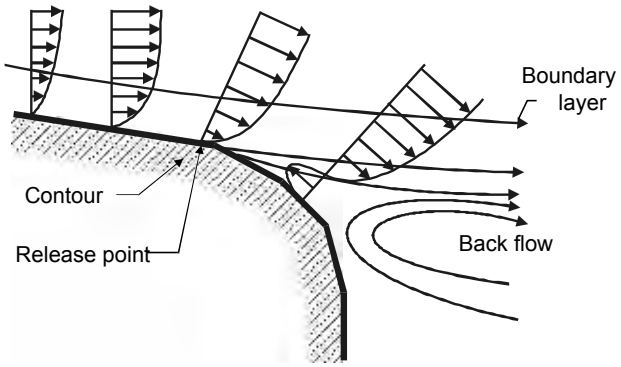


Figure 3.40. *Vortex flow meter: bluff body*

– Sensors: there are different ways to measure the vortex frequency. At the moment there is still no sensor available that is independent of all process parameters.

– Thermistors: a decreasing or increasing temperature causes the low or high pressure on the body. A disadvantage is that these sensors, depending on their position, are sensitive to pollution. In addition, the large time constant of the thermistor is a disadvantage. The thermistor is normally used when working with pure gases and liquids.

– Pressure sensors: the shedding of the vortex causes a pressure fluctuation on the membrane. This fluctuation can be measured with a capacitive or piezoelectric pressure cell. Problems might occur if the temperature exceeds 150°C or if the membrane breaks. Depending on the construction of the sensor, a large pressure range might be required to measure the pressure difference. This can result in poor sensitivity in the lower range of the flow meter. Pressure sensors are usually used for liquids, gases and steam at low pressure.

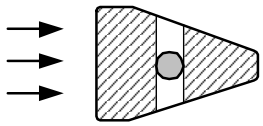


Figure 3.41. *Mechanical sensors*

– Mechanical sensors: a simple method by which a lateral boring connects both sides of the body. Inside the boring, a sphere or a disk oscillates from side to side with a deviation of approximately 0.2 mm. A magnetic reader detects the movement.

Problems might occur when pollution causes obstructions. When dealing with saturated steam, the mechanical movement can be slowed down by condensed steam. Applications can be found for warm water, steam and liquids at low temperatures.

– Strain gauges: pressure differences can set the bluff body in motion with a movement of about 10 μm. Inside the body a stick is placed to which strain gauges are attached (Figure 3.42). The dynamic movement of the element changes the resistance of the strain gauge. The temperature range is limited to below 120°C. Large diameters are sensitive to external vibrations because the mass of the body has increased.

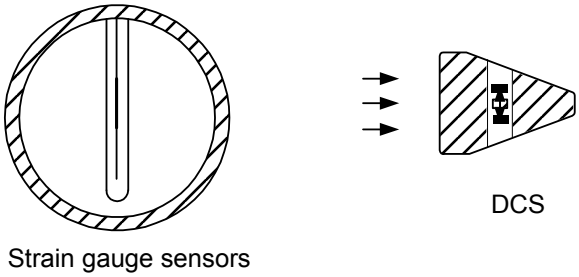


Figure 3.42. *Bluff bodies with strain gauges*

Piezoelements: instead of elastic strips, piezoelectric elements can be applied. A second crystal can be attached to detect external vibrations exclusively. The signal difference is merely caused by the vortex shedding. The correct functioning of the piezoelectric sensor is guaranteed between –40 and +300°C as long as the temperature variations do not occur too quickly (> 100°C during 0.5s).

Differentially connected capacitor (DCC): this system is almost identical to the instrument with piezocrystals and elastic strips. Only one sensor is used for all models. Lateral borings transmit pressure pulses from the vortex to the sensor. Because the sensor is situated inside the bluff body, it is protected from heavy pressure shocks in the flow. The construction of the sensor is very sturdy and can be considered “solid state”. Two stable electrodes are placed in the bluff body; a vibration electrode (“tongue”) is centered between these electrodes. The two stable electrodes and the tip of the axis form two identical capacitors. The pressure waves of the vortex set the tongue into motion and change the capacitances. A preamplifier measures this change. This design allows temperatures from –200 up to +400°C. This sensor is applied for overheated steam, gases, liquids and cryogen gases.

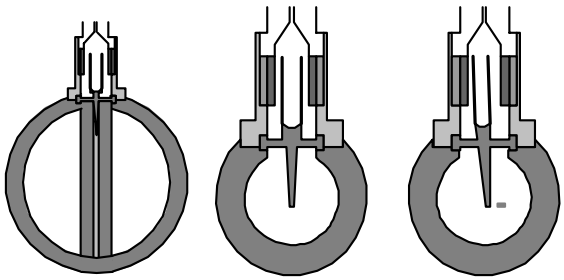


Figure 3.43. *Vibrating DCC*

An ultrasonic transmitter and receiver are placed behind the bluff body. The vortices modulate the sound wave, the frequency of the modulated wave is proportional to the velocity of the fluid. A correct alignment is required to avoid standing sound waves influencing the reading. Sources of sound other than the ultrasonic transmitter render the measurement unreliable. This measuring method is used for liquids as well as for gases.

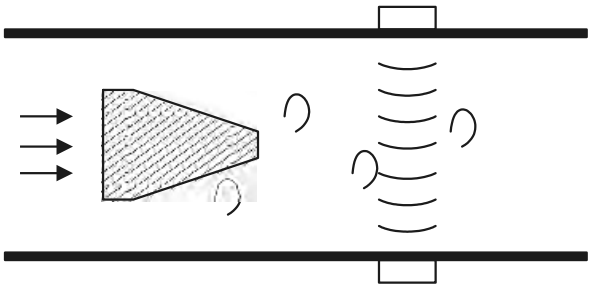


Figure 3.44. *Ultrasonic sensor*

3.7.3. *Practical installation*

– Vortex flow meters can be placed both horizontally and vertically. In a vertical position, the flow should preferably run from bottom to top. Furthermore, we should comply with some rules (as with the electromagnetic flow meter).

– Completely filled pipes are required (as with the electromagnetic flow meter).

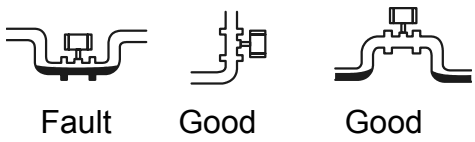


Figure 3.45. *Position of the flow meter for gases, with a chance of condensation*

– The vortex flow meter only works well if the flow profile is completely developed and undisturbed. Due to this, long equalization pipes are necessary: in front of the meter a minimum of 5D; behind the instrument, 3D. However, some manufacturers even advise 20D in front of and 10D behind the meter.

– The measuring instrument should be correctly aligned with the pipe.

– The installation should be performed at a location with little vibration, if necessary the pipes should be supported in front of and behind the meter.

– When in addition to the flow measurement, temperature and pressure readings also need to be performed, the pressure gauge must be placed one diameter in front of the meter, and the thermometer at least 5 pipe diameters behind the meter.

The following occurrences can heavily influence the output of the reading instrument:

– Shape shifting of the bluff body as a result of corrosion.

– Corrosion of the pipe in which the meter is placed.

– Contamination of the bluff body.

– Hydraulic vibrations of pipe and instrument.

– Incorrect alignment of the flange gaskets.

– Incompletely filled measuring instrument.

3.7.4. *Characteristics*

Low installation costs.

– Wide dynamic span.

– The minimum measurable velocity depends on the type of sensor.

– The maximum velocity of the fluid is allowed approximately 7.5 m/s for liquids and approximately 75 m/s for gases and steam.

– High level of accuracy at Re-number > 10,000 (1% for liquid, 1.5% for gases).

– Limited linearity at Re-number < 10,000.

– Good stability and drifting.

– Linearity does not depend on density, viscosity and pressure.

– Valid for gases, liquids and steam.

– Small permanent pressure loss.

– Almost no moving parts, little or no maintenance.

– Limited use at high viscosity and large pressure pulses.

– Good flow profile is required for a correct functioning of the vortex flow meter.

– Resistant to temperature shocks of 100°C/s.

– Depending on the type of sensor, it is insensitive to vibrations up to 1 g (1-500 Hz).

Specific applications

Steam:

– Steam boilers: in the main pipes (overheated steam).

– Chemistry: measuring saturated steam to heat reaction vessels.

– Industrial processes: heating and cooling systems.

Gases:

– Purification plant: flow measurement of methane.

– Industry: measuring compressed air.

– Cryogen gases: liquid nitrogen at temperatures down to –200°C.

Liquids:

– Non-conductive liquids: distilled or demineralized water, glycol, etc.

– Low-viscous hydrocarbon: gasoline, diesel, hydraulic oil, etc.

3.8. Ultrasonic flow meter

3.8.1. *Principle*

The working principle of the ultrasonic flow meter is based on the transmission of sound waves in an acoustic transparent medium. One or more ultrasonic

transmitting-receiving pairs are constructed in or on the pipe, diametrically against each other. The first pair is placed slightly more downstream than the second pair, so they make a certain angle with the pipe longitudinally.

We distinguish two measuring instruments, depending on the measuring principle:

– Ultrasonic flow meters based on the time-of-flight (transit time) principle.

– Ultrasonic flow meters based on the Doppler effect.

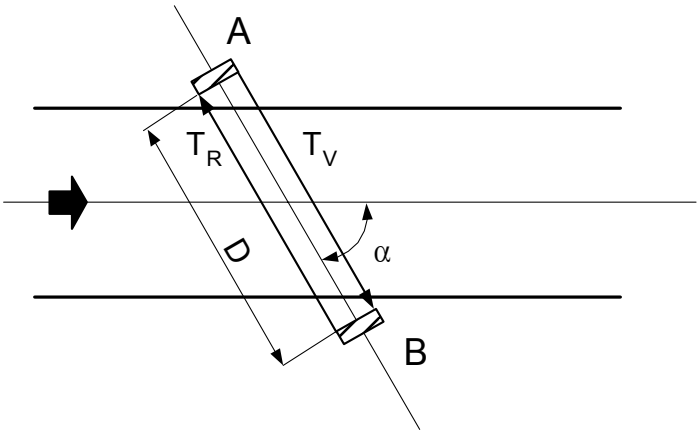


Figure 3.46. *Time-of-flight principle*

Measuring the flow following the execution time principle

Because the sound velocity adds up vectorially with the velocity of the flow, and because transmitter-receiver B is situated downstream with respect to A, the sound wave train from A to B will arrive sooner than the train from B to A (Figure 3.46). This implies that the execution time from A to B is shorter than that from B to A, Tr > Tv (if velocity > 0).

The difference in execution time provides the average flow-rate v_{gem}:

$$v_{gem} = \frac{\Delta t.(\frac{D}{t_v})^2}{2.D.\cos\theta} = \frac{\Delta t.D}{2.(t_v)^2.\cos\theta} \tag{3.16}$$

Measurements based on the Doppler effect

This older technique (discovered in 1842 by Christiaan Doppler) has become well known mostly due to its application in the so-called "clamp-on" meters. The Doppler effect occurs with sound as well as electromagnetic waves. When a source or receiver moves in a wave medium, the frequency at the receiver will differ from the frequency at the transmitter. The frequency increases with a movement towards the source and it decreases with a movement away from the source. This is caused by the constant velocity of the wave in the medium. If all velocities in the same direction are counted positively, we can describe the Doppler effect as follows:

$$f_w = \frac{c - v_w}{c - v_b} \cdot f_b \tag{3.17}$$

where

– f_w: observed frequency for movement

– f_b: frequency of the source in rest

– c: transmission velocity in the medium

– v_w: velocity of the observer with respect to the medium

– v_b: velocity of the source with respect to the medium

During the measurement the source (transmitting crystal) and the observer (receiving crystal) are fixed and the intermediate substance, the fluid, is moving. The transmitted bundle is only detected if dispersed by moving fluid particles. These particles can be solids or small gas pockets (Figure 3.47). The Doppler technique only works in liquids that contain enough solids or gas pockets. These are normally considered "difficult" substances that damage "normal" flow meters.

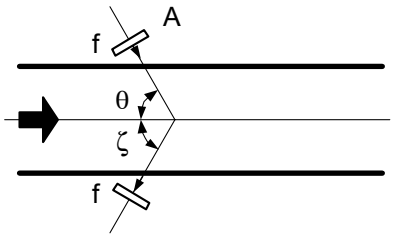


Figure 3.47. *Transmitting crystals A and B*

The Doppler frequency shift is:

$$\Delta f \approx \pm f_s \cdot \frac{v}{c} \cdot (\cos\theta + \cos\zeta) \tag{3.18}$$

The ± sign indicates the direction of the velocities, thus also the position of the sensors.

Note t: Δf is proportional to v. So for a very sensitive measurement:

– f_s needs to be as large as possible (limited by attenuating);

– cosθ + cosζ needs to be as large as possible, which means that θ and ζ need to be as small as possible.

3.8.2. *Practical installation*

In general acoustic flow meters need no special requirements regarding installation on or in the process pipe:

– A vibration-free location is recommended especially when applying the Doppler type flowmeter, as vibrations cause false signals which may fool the electronics.

– Similar to most flow meters, the measuring pipe must be completely filled with the fluid (emptying when measuring gases, Figure 3.45); see also section 3.7.3.

– A well-developed flow profile is absolutely required for a reliable and accurate measurement. That is why equalization pipes (10 ÷ 20) D in front of and 5 D behind the meter are recommended to obtain the given level of accuracy (2%).

– Pilot valves closely behind the flow meter negatively influence the measurement, especially when cavitation or supersonic velocities occur.

3.8.3. *Characteristics*

– No pressure loss in the pipe.

–It is possible to measure without making contact with the fluid (“clamp-on” realizations).

– Only useful for liquids that are acoustically transparent.

– A small, but not excessive amount of contamination of the liquid is necessary for the Doppler effect. The time-of-flight principle needs as little pollution as possible.

– Difficult for small diameters, especially when using the time-of-flight principle.

– For the time-of-flight difference meter the turndown can amount to 1:1,000 and an accuracy level from 1% to 2.5% is possible.

– For the Doppler type meter the turndown can amount to 1:3,000 with an accuracy level from 2% to 5%.

– Individual calibration is needed for every medium.

– Using the Doppler effect, the reading depends on the flow profile.

– Accuracy levels up to 1% are possible (not when using clamp-on realizations).

– At present it is also possible to ultrasonically measure the flow of gases, steam and even high-temperature steam.

3.9. Coriolis mass-flow meters

In many flow-measuring applications the mass flow is much more important than the volume flow. The operational range of an airplane for instance is determined by the mass of the liquid, and not the volume. Therefore, flow meters that are used to fill up airplanes better indicate mass instead of volume. Mass-flow readings are the most important measurements in the chemical industry. The mass in an invariable quantity, while volume depends on temperature and pressure. Only the most common Coriolis mass-flow meters are presented.

3.9.1. *Principle*

One or two U-shaped pipes are set to vibrate at their natural frequency (with an amplitude of a few mm at approx. 80 Hz) by means of a magnet coil that is fixed at the bend (Figure 3.48).

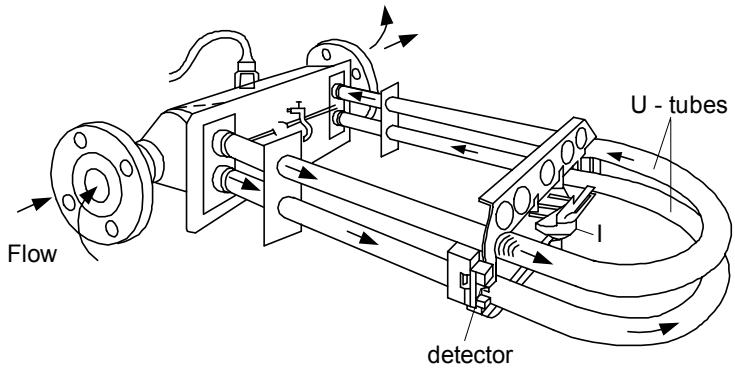


Figure 3.48. *Coriolis mass-flow meter*

The passing medium is forced to take on the vertical movement of the vibrations, increasingly in the direction of the bend, decreasingly past the bend, in the direction of the outlet of the meter. The forces that the moving liquid transfers to the pipe are called Coriolis forces. Although they are very small, they are strong enough to twist the tube so that the two parts (normally symmetrical with zero flow) pass the proximity sensors with time difference Δt. The mass flow Q_m can be calculated from

$$Q = \frac{k_s}{8.r^2}.\Delta t \tag{3.19}$$

where

– k_s: spring constant

– r: radius of the pipe.

Nowadays there are more and more Coriolis flow meters with straight pipes. The theory and working method are identical to that of the U-shaped pipes but the instruments become smaller and there is less decline in pressure over the flow meter. The two straight pipes of the flow meter are set to vibrate by an electromagnet. When a flow runs through these vibrating pipes, a wave motion starts. The frequency of this wave motion (measured by sensors fixed to the tubes) is a measure of the mass flow (Figures 3.49 and 3.50).

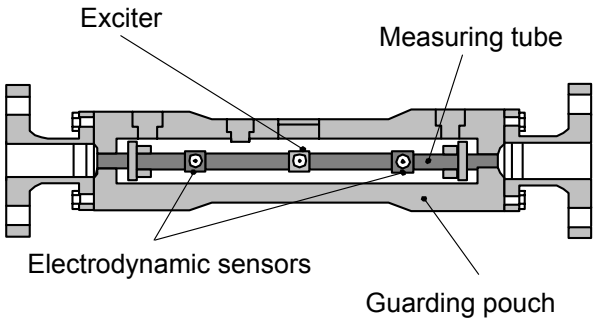


Figure 3.49. *Coriolis flow meter with straight pipes*

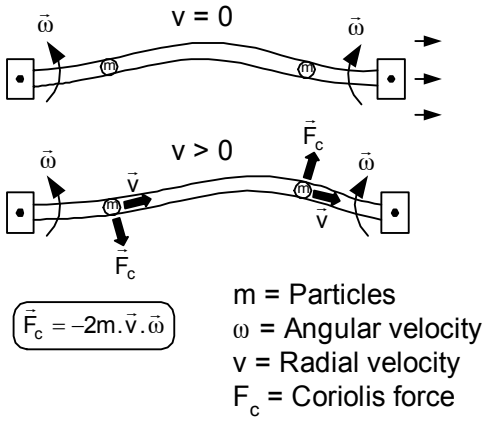


Figure 3.50. *Coriolis flow meter with straight pipe: principle*

3.9.2. *Applications*

Because of its capacity to measure from less than 0.05 kg/min to more than 10,000 kg/min, with an accuracy level of 0.4% FS the instrument we described before and similar models have found their way to many industries like the food industry, chemical industry, petrochemical industry, synthetic material industry and pharmaceutical industry.

The capacity to measure masses directly in only one instrument makes this meter a versatile instrument in many applications. A reduced selection of applications

shows the extended working space: paint, slack, bunker C, CO_2, milk, cream, soup, adhesives, acids, natural gas.

3.9.3. *Practical installation*

– Completely filled pipes are required, even for the zero point adjustment (see section 3.7.3).

– Best used in a vertical pipe with an upward flow.

– Sufficient support of the pipes in front of and behind the meter is needed.

3.9.4. *Characteristics*

– Direct indication of the mass flow.

– Almost static reading.

– High level of accuracy.

– No equalization pipes needed.

– Almost no influence on the flow by pressure losses in the meter.

– Complex electronics, expensive to buy.

– Sensitive to vibrations.

– High installation costs.

3.10. Flow measurements for solid substances

The previously discussed flow meters are only suitable for measuring liquids and/or gases. Some processes, however, also require flow measurements of solids, i.e. bulk goods, powders, granules, scales etc.

Some applications in which flow measurements of solids are essential are:

– The production of chipboard: wood chips with a bulk weight of 0.1-0.2 kg/m^3 are controlled in a fixed proportion to glue.

– Measurement of fragments of glass.

– Batch measurement: from roasted coffee beans to the many different packing machines.

– Fly ash measurement for the production of cement.

– Production of spaghetti dough: the different ingredients are measured with the solids flow meter. When the preset value is reached, a PLC switches off the supply of the concerned ingredient and the next ingredient is switched on. In this way the correct recipe is put into the blender.

– The production of compound feed.

To continuously measure the flow of solids, three common principles exist:

– Flow measuring with a measuring plate.

– Weight measuring on a conveyor belt.

– Capacitive measurement in pneumatic transport systems.

– Schenck also manufactures a Coriolis flowmeter for solid particles.

To detect the flow of a solid substance, different types of proximity switches, depending on the installation, are applicable. The mostly commonly used method is with microwaves.

3.10.1. *Flow measurement of solids by means of an impact plate*

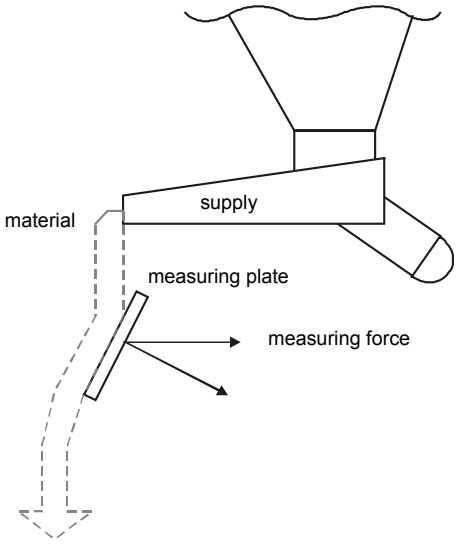


Figure 3.51. *Flow measurement by means of an impact plate*

The measurement is based on the impulse transfer of a free-falling substance to a plate. The basic principle is Newton's second law of motion:

$$F.\Delta t = \Delta m.\Delta v \tag{3.20}$$

If we assume that the starting velocity equals zero, the horizontal force is (Figure 3.52):

$$F_m = Q_m.\sqrt{h}.\sin\alpha.\sin\gamma.100 \tag{3.21}$$

where

– F_m: the horizontal force in grams, that originates from the impact on the measuring plate

– Q_m: the mass flow [t/h]

– h: the height of fall [m] (the starting velocity should be approximately zero)

– α: angle of the measuring plate

– γ: angle of impact

The collision force is converted into a horizontal displacement of 0.8 to 2.4 mm, which is measured by a position sensor.

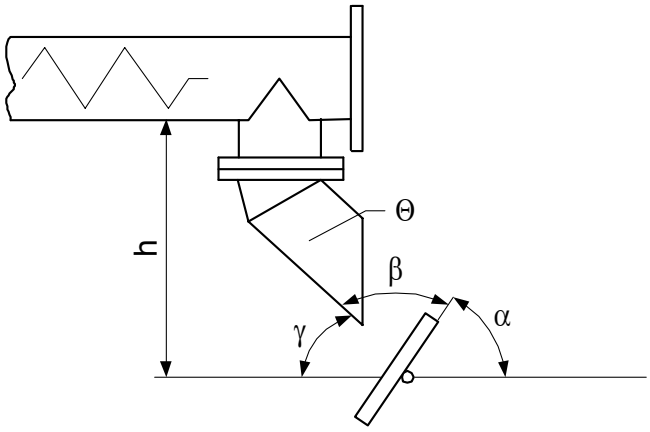


Figure 3.52. *Flow measurement by means of an impact measuring plate: distribution of forces*

The measuring system has a quick response time and so it is often used to adjust proportions in continuous mixing processes. As a result of its high flexibility it can be built into any conveyance device.

The measuring system provides the following information:

– The flow [kg/h] or [t/h].

– The total passed quantity, [kg] or [t], on a total counter.

Characteristics

– Simple.

– Accurate, up to 1% FS.

– Stable.

– Not dependent on pollution of the measuring plate, because only the horizontal force is measured.

– Working temperatures from –40 to +50°C.

– Mechanical wear.

– Specific weight must remain constant.

3.10.2. *Flow measurement of solids based on the weighing method*

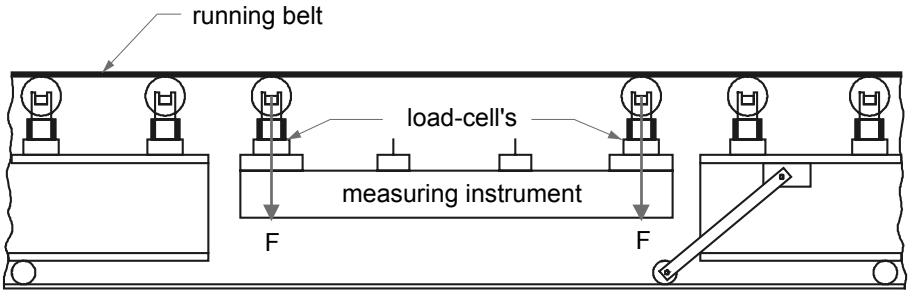


Figure 3.53. *Flow measurement: weighing method*

This flow meter continuously measures the flow and the quantity of solid substances that are transported on conveyor belts (Figure 3.53). The product on the belt causes a weight on the measuring instrument that is set up between the existing running belt. The weight affects the load cell, where it is converted into an electrical

standard signal. The processing unit, which is operated by a microprocessor, shows, depending on the choice, the quantity in kg/h or %, of the already counted quantity, the velocity of the conveyor belt or the weight on the belt. The measuring system is accurate to 1% of the measured value. An angle of inclination of the belt up to 20° is allowed.

3.10.3. *Capacitive flow measurement of solid substances*

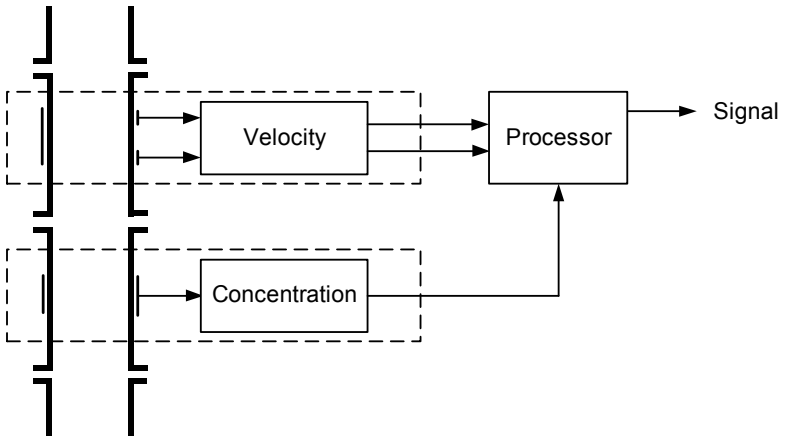


Figure 3.54. *Capacitive flow measurement of solid substances*

This measuring method is applied to the inline measurement of the average transport speed and the solids concentration of pneumatically transported solid substances, such as coal, cement, grains, synthetic granules, etc. (Figure 3.54). The flow is calculated with the help of both measuring signals and the known, constant properties of the solids.

The total measurement is performed with two instruments, which can also be used independently. The first instrument measures the velocity capacitively; the second the concentration. The μ-processor calculates both signals into a solid flow. The measuring tube consists of two capacitors; the movement of the substance causes capacitance variations in both capacitors. This happens with a specific delay, from which the velocity of the substance can be calculated. The concentration measuring tube on the other hand consists of only one capacitor. This sensor renders the concentration variation of the two-phase flow gas/solid.

Characteristics

– No mechanical components.

– Reproducibility up to 2%.

– A new calibration is needed when the product properties (i.e. dielectric constant) change.

– Response time of 1s.

– Range from 0.01 to 60 m/s.

3.10.4. *Detection of solid substances using microwaves*

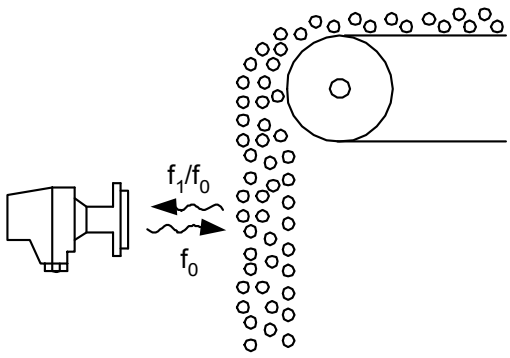


Figure 3.55. *Flow detection using microwaves*

The flow/no flow detector works with microwaves, using the Doppler effect. The sensor transmits a microwave; this signal is reflected by the solid substance and received again by the sensor (Figure 3.55).

Because microwaves penetrate non-conductive materials, the detection can occur without contact, for instance, through synthetic pipes, wooden panels, sight-glass or basalt coverings. Neither smoke, mist, sounds, light nor air turbulence influence the sensor. The working frequency is 24,125 GHz, which is situated within the internationally released ISM band. The radiation intensity is less than 0.15 mW/cm^2.

Characteristics

– Contact-less detection, no mechanical wear.

– The working does not depend on material properties.

– Adjustable sensitivity and time retardation are desirable.

– Be careful with a vibrating installation.

– Preference to a fail-to-safe adjustment.

3.11. Flow measurement for open channels with weirs

This type of flow meter is based on the decline in height.

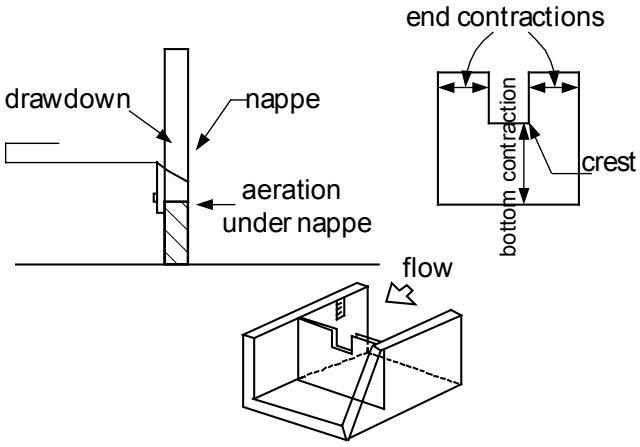


Figure 3.56. *The weir*

When the flow profile of a free water flow is confined by a local vertically positioned protrusion above the bottom, we call this a weir (Figure 3.56). In other words, a dam is made in a liquid stream, and because of this the liquid undergoes resistance and starts rising. This rise is a measure of the flow at that moment.

To measure the flow this weir must have a *sharp crown*, which means that the overflowing stream can only linearly touch the vertical protrusion.

Besides, it must be a *complete weir*, which means that the downward level cannot influence the flow.

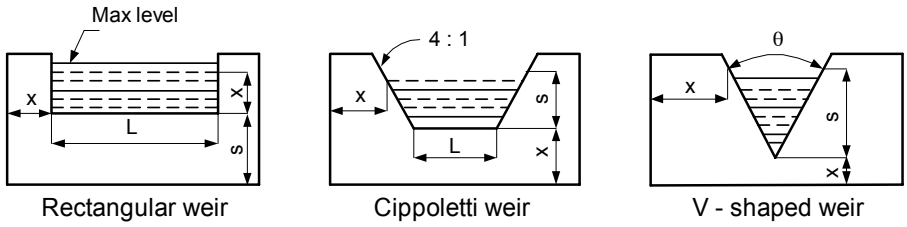


Figure 3.57. *Weir shapes*

Different kinds of weirs exist; they usually have simple geometrical shapes, such as a rectangular or triangular weir (Figure 3.57):

– For a rectangular weir:

$$Q = 3.33.(L - 0.2.H).H^{1.5} \tag{3.22}$$

– For a Cippoletti weir with:

$$Q = 3.367.L.H^{1.5} \tag{3.23}$$

– A V-shaped weir:

$$Q = 2.48.\tan(\frac{1}{2}.\theta.H^{2.5}) \tag{3.24}$$

where

– θ: angle "V" [degrees]

– H: the height of the liquid [feet]

– L: the length of the bottom [feet]

We notice that the instream speed of the liquid (usually water) is reduced to 6 cm/s for a calm water surface and an even distribution. This is clearly a laminar flow. Measuring the height normally occurs in an ultrasonic or hydrostatic manner, depending on the liquid, or with the float method.

3.12. Choice and comparison of flow measurements

When choosing and implementing a flow meter, the following issues should be taken into account:

– The nature of the product: gas, steam or liquid (conductive or non-conductive).

– Mass-flow or volume-flow measurement.

– Characteristics of the installation: pipe diameter, size of the meter, capacity.

– Choice of material (corrosion, erosion).

– Maximum allowed temperature and pressure.

– Connections, output signals, accessories.

– Performance: accuracy, reproducibility, decline in pressure.

– Price: purchase, installation costs, maintenance, lifespan.

3.13. Bibliography

1. John, H.: *Low Reynolds Number Hydrodynamics*, Martinus Nijhoff, 1983.

2. Wolfgang, R., Bergeles, G.: *Engineering Turbulence Modelling and Experiments 3*, Elsevier Science, 1996.

3. Borer J.: *Instrumentation and Control for the Process Industries*, Elsevier Applied Science Publishers, 1985.

4. Doeblin Ernest, O.: *Measurement systems: Application and Design*, McGraw-Hill International Book Company, 1991.

5. Endress + Hauser technical documentation.

6. Hoffman, K. *Eine Einführung in die Technik des Messens mit Dehnungsmebstreifen*, HSM, Darmstadt, 1987.

7. ISA: *Process Instrumentation Terminology: Appendix A*, ISA, 1979.

8. Johnson Curtis, D.: *Process control instrumentation technology*, Wiley and Sons, New York, 2nd ed., 1982.

9. WIKA: *Handbook of Pressure Measurement, with Resilient Elements*, Gottlob Volkhärdtsche Druckerei, Anorback, 1981.

3.14. Website references

http://www.flowmetermanufacturers.com
http://www.endress.com

http://www.emcoflow.com
http://www.omega.com
http://www.davisontheweb.com
http://www.barnant.com

Turbines

http://www.onicon.com
http://www.sponsler.com
http://www.jlcinternational.com
http://www.awcompany.com

Mechanical

http://www.badgermeter.com
http://www.awcompany.com

Electro-magnetic

http://www.marsh-mcbirney.com
http://www.chemtec.com

Ultrasonic

http://www.alaxa.nl
http://www.dynasonics.com

Mass-flow (liquids)

http://www.emersonprocess.com/micromotion/default.html

Mass-flow (gases)

http://www.kurz-instruments.com

Weirs

http://www.mjk-automation.nl
http://www.marsh-mcbirney.com

Reynolds number

http://www.wtb.tue.nl/woc/wet/Reynolds.htm

Bernoulli's equation

http://theory.uwinnipeg.ca/mod_tech/node68.html

Vortex

http://www.flowmeterdirectory.com

Pitot tube

http://www.grc.nasa.gov/WWW/K-12/airplane/pitot.html
http://www.svce.ac.in/~msubbu/FM-WebBook/Unit-III/PitotTube.htm

Venturi tube

http://www.ce.utexas.edu/prof/kinnas/319LAB/Applets/Venturi/venturi.html
http://pages.prodigy.net/bderoes/calculus/epsilon_delta/venturiTube.html

Chapter 4

Intelligent Sensors and Sensor Networks

4.1. Introduction

Ordinary sensors, as described in other chapters of this book, were used for a long time and are still used in many applications. Thanks to the explosive progress in microelectronics in the final three decades of the 20th century, sensor technology moved to a completely new level of quality. The functionality of the ordinary sensors has been expanded in many ways and a new group of so-called *intelligent* sensors has appeared, providing a number of additional properties. They include higher accuracy, better immunity to environmental conditions, application flexibility and especially the possibility of easy integration into the industrial distributed systems.

In contrast to a common centralized approach, nowadays control, measurement and data acquisition systems are widely based on the distribution of a system's tasks among a number of nodes (e.g. intelligent sensors, actuators, controllers), interconnected using an appropriate communication network. It is not enough to acquire the information; it has to be transferred from the sensors to the place where the decisions are made and results have possibly to be transferred back. The behavior of the communication network between the nodes of a distributed system has a significant influence on overall system performance. Maximum node distance, communication speed, response times, determining of data transfers, resolving of transmission errors and a number of other parameters should be taken into account when the suitable communication network standard for a particular application is being selected. Communication networks used to interconnect intelligent sensors

---

Chapter written by Jirí NOVAK.

and (sometimes) actuators with controllers are often called sensor networks or industrial distributed systems.

4.2. Intelligent sensors

An intelligent sensor is a system usually consisting of a chain of analog and digital blocks, each of which provides a specific function. Sensors or transducers of particular physical quantities described in other chapters constitute only small but important part. Figure 4.1 shows the structure of an intelligent sensor. Of course, not every intelligent sensor contains all the presented blocks; it depends on sensor type and functionality.

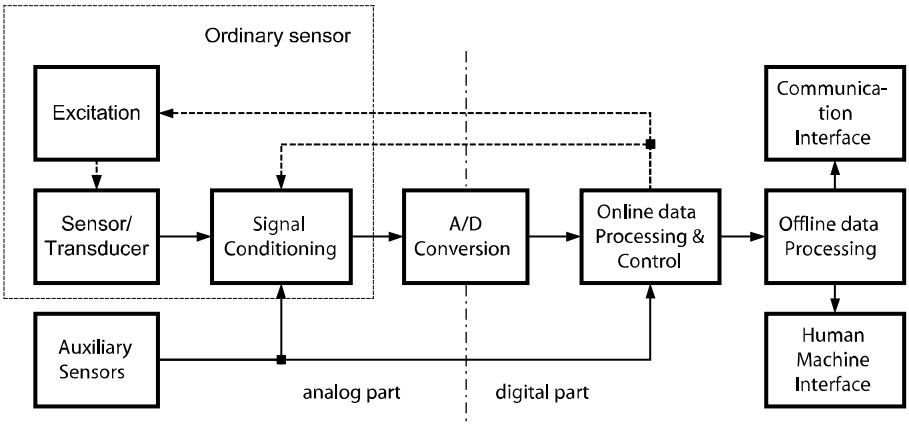


Figure 4.1. *Intelligent sensor structure*

The sensor or transducer usually provides some electric quantity output (e.g. voltage, current or impedance), which depends on the value of measured physical quantity (either electric or non-electric). Many sensor types require external excitation to work, usually electric, but sometimes also magnetic or mechanic excitation. A signal conditioning block may provide amplification, filtering, non-linearity correction and similar functions. Usually they are implemented using analog circuits, but digital implementation (especially of specific functions like non-linearity correction or environmental influence compensation) becomes more frequent. An A/D converter block converts analog signal into digital values, suitable for processing by microprocessor or other digital devices. The following blocks are implemented either in digital hardware (ASIC, PLD) or in software (running on the microcontroller). The online data processing and control block provides part of the signal conditioning, and for some sensor types it controls excitation parameters or

analog signal conditioning processes (e.g. synchronous detection). The offline data processing block receives instantaneous values of measured quantities. It is more focused on data storage, report generation, trend evaluation, etc. The last two blocks are well known to users; the HMI (*Human Machine Interface*) allows local sensor control and data output, and the communication block provides an interconnection to the system controller via the distributed system. Some intelligent (as well as ordinary) sensors also contain an auxiliary sensor that measures other physical quantity interfering with the main sensor output. A typical example is an auxiliary temperature sensor used to compensate the unwanted temperature influence.

4.2.1. *Sensors and transducers*

This book contains a large number of sensors and transducers for different physical quantities. The electric quantities at sensors' outputs that are used for further processing within the signal-conditioning block can be divided into several groups regardless of the original physical quantity which was measured. The way the sensor output signal is processed in the signal-conditioning block is often similar within a particular group. The most important examples are listed below.

4.2.1.1. *Variable voltage or current source*

The most common sensor with a voltage output is probably a thermocouple, which is based on thermoelectric laws (see Chapter 8) and used to measure temperature.

A photodiode in photovoltaic mode loaded by zero impedance is a current source controlled by an ambient luminance (see Chapter 2).

The Hall sensor used to measure the magnetic field also produces a voltage at its output, which is proportional to the measured field strength (see Chapter 10).

An induction sensor of angular velocity (see Chapter 7) based on Faraday's law produces alternating voltage at the output.

4.2.1.2. *Variable resistance*

A potentiometer (see Chapter 7), which is part of many sensor types (e.g. position or pressure), is a basic example.

Resistance temperature detectors (RTD) and thermistors (PTC or NTC) are both examples of the sensors (see Chapter 8) whose resistance depends on temperature.

Piezoresistive strain gauges (see Chapter 1) are used in many sensors, e.g. pressure or distance.

An anisotropic magnetoresistor's (AMR, see Chapter 10) resistance depends on applied magnetic field strength.

4.2.1.3. *Variable impedance or mutual impedance*

The measured impedance usually predominantly displays either a capacitive or inductive nature.

The capacitance is measured with capacitive level sensors or small displacement capacitive sensors (see Chapter 7).

A similar principle is used for capacitive pressure sensors (see Chapter 1).

Many position and distance sensors are based on inductance measurements, namely with the eddy current proximity sensor (see Chapter 7).

A large number of position and distance sensors are based on mutual inductance, e.g. LVDT (*Linear Variable Differential Transformer*) for linear and RVDT (*Rotary Variable Differential Transformer*) for angular measurements.

4.2.1.4. *Charge generator*

Piezoelectric accelerometers (see Chapter 5) are typical representatives of this output type. Another is a pyroelectric sensor of infrared emission, on which PIR motion detectors are based.

4.2.2. *Signal conditioning (SC)*

A signal-conditioning block is mainly used to extract information about the measured quantity from the sensor output signal and to match it to the input of the following block – an A/D converter. It typically implements some (or all) of the following functions – amplification and signal conversion, sensor insulation, filtration, detection, non-linearity correction, and environmental influences correction. Sometimes selected functions (especially the non-linearity and environmental corrections, more rarely the filtration or detection) are implemented later in the chain within the data processing block. Often the SC block also contains the circuits that allow sensor and/or sensor connection diagnostics. This feature is necessary for sensors whose failures could cause a large amount of physical damage or even risk personal safety.

4.2.2.1. *Amplification and signal conversion*

Amplification and signal conversion is nearly always used, as direct sensor output signal magnitudes are usually low. In case the output electric quantity of the sensor is not a voltage, voltage signal conversion is usually applied before further processing.

4.2.2.2. *Sensor insulation*

Galvanic insulation is often required to avoid ground loop currents that introduce errors in the measuring chain. Insulation also provides a barrier against interfering voltage spikes coming from the examined technology that may damage electronic circuits. Galvanic insulation of the sensor is also required if there is a higher voltage difference between the technology the sensor is mounted on and the sensor itself.

4.2.2.3. *Filtration*

Filtration is a very important part of the *SC* block and can be found in nearly every intelligent sensor. The sensor output signal that carries information about the measured quantity is often distorted by a number of noise sources that can be (at least partially) removed using filtering. A low-pass, high-pass, band-pass, stopband, notch filter or a combination of these is applied depending on the specific situation. In conjunction with sampling and A/D conversion, anti-aliasing filters are often applied. They are usually higher order low-pass filters that are used so that the output signal satisfies the sampling theorem.

4.2.2.4. *Detection*

The higher frequency signal envelope sometimes carries information about the measured physical quantity, e.g. in some SC circuits for the eddy current distance sensor (see Chapter 7). Detection is then used to extract the envelope amplitude. Often a synchronous detection is used in order to increase the signal to noise ratio, e.g. in SC circuits for optical sensors (see Chapter 2) or LVDT displacement sensor (see Chapter 7). Many sensors provide information encoded into the phase shift between the excitation and output signals. Synchronous detection is used here as well, providing 0° and 90° components.

4.2.2.5. *Correction of non-linearity*

Dependence between the measured physical quantity and the sensor output value is often non-linear. This is not very practical, but in many cases the non-linearity is known and the correction can be used. It can be applied either before or after the A/D conversion. The first case is more usual for simpler (non-intelligent) sensors with an analog output. The second is typical for microprocessor-equipped sensors, where either an analytic equation form or a correction table is implemented.

4.2.2.6. *Correction of influence of disturbing quantities*

Unfortunately, the sensor output signal value is not dependent only on the measured physical quantity. Other physical quantities (e.g. temperature) often have an indispensable influence on the sensor output. This influence can sometimes be decreased by suitable arrangement of sensors (e.g. into the bridge), in which the sensitivity to the measured quantity is twice (or even four times) higher whilst the sensitivity to disturbing quantity is nearly suppressed. Of course, such an arrangement is not always possible and auxiliary sensors are used to measure the interfering physical quantities. A correction is then applied as in the case of non-linearity. Again there are two typical arrangements; in the first the auxiliary sensor is a part of the analog SC circuit, in the second the correction is applied after the A/D conversion.

4.2.2.7. *Sensor excitation*

Additionally, many sensors need excitation. The excitation block, which is separated in Figure 4.1, is often considered a part of the SC block. Excitation is always required for sensors that do not have a natural source (for example, the thermocouple has). Either a DC voltage or current sources (usually for sensors with resistive output) or an AC voltage or current sources (usually for sensors with a common impedance output) are used for excitation. Sometimes the DC excitation is not applied continuously and is controlled to avoid a self-heating of resistive temperature sensors, for example. AC excitation is often used as a reference for synchronous detection.

4.2.3. *A/D Conversion*

Today intelligent sensors usually use one of three A/D converter types. They are a SAR (*Successive Approximation Register*) converter, a converter with a Sigma/Delta modulator, and a Flash (or pipelined Flash) converter.

4.2.3.1. *SAR converters*

This type of A/D converter is probably the most common in intelligent sensor applications. It provides resolution from 8 to 16 bits and typical conversion time is several microseconds. They are available as standalone chips or (very often) integrated within single-chip micro controllers, ASIC or SOC designs. Sometimes they allow the user to choose between a higher resolution with longer conversion time or vice versa.

They are used especially for conversion of output signal from sensors of dynamic processes with a high frequency limit at about 100 kHz (or several

hundreds). The piezoelectric accelerometers (see Chapter 5) are the typical sensors providing this type of output.

4.2.3.2. *Sigma-delta modulator converters*

This converter type provides the highest resolution of all mentioned here, up to 24 bits. The conversion rate as well as the conversion resolution depends on the modulator over-sampling and output filter decimation factor. The higher decimation provides higher resolution but a lower conversion rate and vice versa. The converters are available as standalone ICs or integrated in ASIC or SOC designs.

These converters are used for high resolution DC or low frequency sensor outputs, e.g. for temperature or static pressure sensors.

4.2.3.3. *Flash (pipelined flash) converters*

Flash converters are based on parallel comparison of an analog input voltage by a number of comparators (e.g. 255 comparators for 8-bit resolution). They are very fast – the conversion time is several nanoseconds, and usually available with an 8-bit resolution. To reduce the number of comparators, a pipelined flash architecture is used that uses a cascade of 3-5 bit flash converters to receive a higher resolution (up to 16 bits), but there is a higher latency (waiting time).

They are used especially in optical application, e.g. for CCD (see Chapter 2) or ultrasonic sensor data processing.

4.2.4. *Data processing*

Taking into account the block diagram in Figure 4.1, data processing consists of two parts. Online data processing can rather be considered as a part of signal conditioning, because it usually caries out a function that belongs to that block, but in the digital domain. Sometimes it is used to control the analog SC circuits or the sensor excitation. As it works in real-time, its implementation has to be fast enough to be able to process the data and to control the whole measurement process.

Offline data processing then provides more functions which are useful for the user, such as the data storage, averaging, extreme value searching, boundary crossing detection, future trend evaluation and so on. These functions are usually only included in sensors with either human or digital communication interfaces.

4.2.5. *Human Machine Interface*

Although modern intelligent sensors are nearly always equipped with a communication interface and a local user interface may seem to be of no use, it is often provided by manufacturers. The local interface does, however, enable the sensors to be read even in case of a malfunction of the communication system and also assures that a human operator can monitor the output.

Usually only basic information is available about the measured physical quantity – typically it is an instantaneous value. Sometimes users can select which value has to be shown (e.g. maximum, minimum, average) or some measurement parameters can be set (e.g. emergency limits). Critical settings are often protected by password or can be locked and unlocked remotely.

4.2.6. *Communication interface*

The communication interface is not only the point (connector) where the distributed system (e.g. the sensor network) is connected. It also consists of a complex stack of hardware and software layers that ensure interoperability with the rest of the system. There are many different standards of sensor network; each suiting a particular application class or focusing on a particular branch of industry (see section 4.3).

4.2.6.1. *IEEE 1451*

Work began in 1994 to standardize these interfaces under IEEE 1451. Focusing on the intelligent sensor/actuator internal modularity its aim was to define a common interface to access services of any distributed network standard. The basic principle of the standard is a division of an intelligent sensor/actuator into two parts that are connected using a standard interface. The first part (in 1451 terminology called NCAP – *Network Capable Application Processor*) provides a standard transducer interface to a communication network. Appropriate NCAP implementation will allow easy connection of transducers into arbitrarily distributed systems. The second part is a transducer providing standardized services and equipped by a TEDS (*Transducer Electronic Data Sheet*). A TEDS contains data describing particular sensors, acquisition channels used to process information and calibration data.

IEEE 1451.0 defines commands and operations between the NCAP and the transducer. They are independent of physical layer interface implementation. IEEE 1451.1 defines an object model of smart transducers, an NCAP model and communication models between them. IEEE 1451.2 defines a standard interface TII (*Transducer Independent Interface*) between a STIM (*Smart Transducer Interface*

Module), which contains the transducer, and an NCAP. IEEE 1451.3 defines a multi-drop transducer to NCAP interface by means of the TBIM (*Transducer Bus Interface Module*) allowing easy connection of more transducers to one NCAP. IEEE 1451.4 is focused on implementation of TEDS into mixed-mode (analog output) transducers by MMI (*Mixed Mode Interface*). IEEE 1451.5 defines an NCAP to transducer interface and TEDS for wireless transducers. Finally, the IEEE 1451.6 defines this interface for CANopen (see section 4.3.5.2) connected transducers.

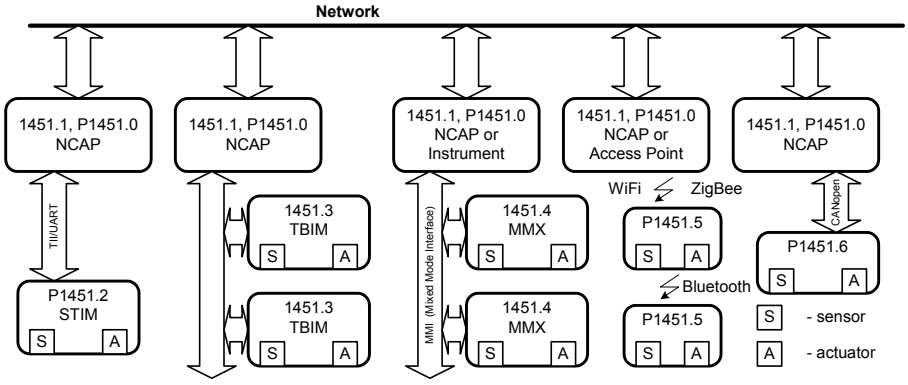


Figure 4.2. *IEEE 1451 standard configuration variants*

Particular standard parts allow different variants of NCAP and transducer connections, as described above and shown in Figure 4.2 (parts marked with P are not yet approved). Nevertheless, a lot of work remains for this standard to become widely, if ever, used. Until now only part 4 is quite widely used by sensor manufacturers.

4.2.7. *Industrial examples*

Two intelligent sensor examples are shown here. The first is the rather unusual Hall sensor; the second is a fully equipped resonance pressure sensor.

4.2.7.1. *Micronas HAL805 Hall sensor*

Micronas offers several types of programmable Hall sensors (see Chapter 10) either with digital (switch) or linear output characteristics. A block diagram of the Hall sensor HAL805 with linear output is shown in Figure 4.3.

The sensor is designed to be used for distance or angle measurement application. The magnetic field is sensed by the Hall sensor and the sensor output is digitized. It is processed in the DSP block using the parameters pre-programmed in the EEPROM. The analog output voltage is then generated using a D/A converter. The parameters' lock block allows the content of EEPROM to be locked and not changed again.

Particular sensor parameters (sensitivity, zero field output voltage, output voltage range, magnetic material characteristics and so on) can be digitally programmed using the power supply pin; responses are available at the output. Frequency range is DC to 2 kHz. HAL810 sensor type provides similar functions but with PWM output that suits direct microprocessor connection.

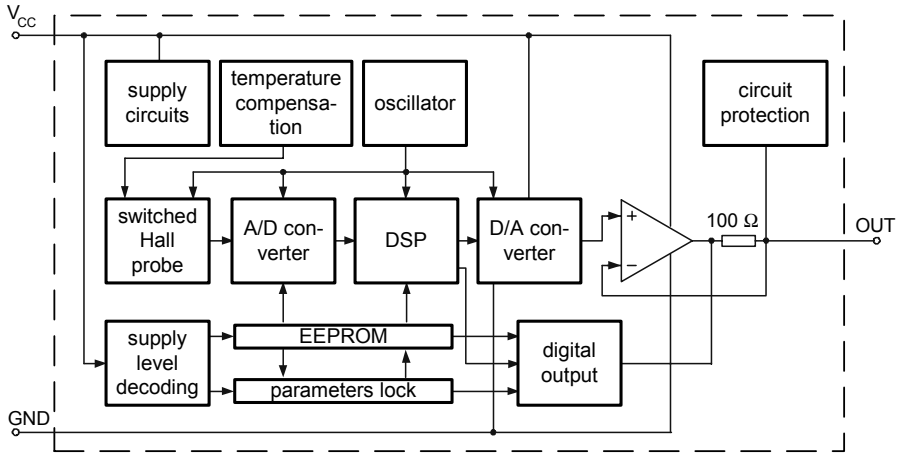


Figure 4.3. *HAL805 block diagram*

The sensor is available in an industrial temperature range and provides over-voltage, reverse-voltage and ESD protection.

4.2.7.2. *Yokogawa DPharp family of pressure sensors*

DPharp (*high-accuracy resonating pressure*) family of sensors is manufactured using MEMS technology. The block diagram of an EJA series intelligent sensor is shown in Figure 4.4.

The sensor is based on the principle described in Chapter 1. The capsule contains the sensor, excitation circuit, auxiliary temperature sensor and EEPROM containing the sensor specific parameters. The converter block then measures the sensor output

frequency, provides digital corrections and final data processing. LCD is available as an option for monitoring by humans and the HART or Foundation Fieldbus communication interfaces are also available.

The sensor provides high linearity, long-term stability and high overpressure immunity. The zero point, high and low range values can be set either locally or through a communication interface. Many other parameters, including the sensor information, pressure unit selection, automatic ranging, diagnostics status, and internal error logging are available only via the communication interface.

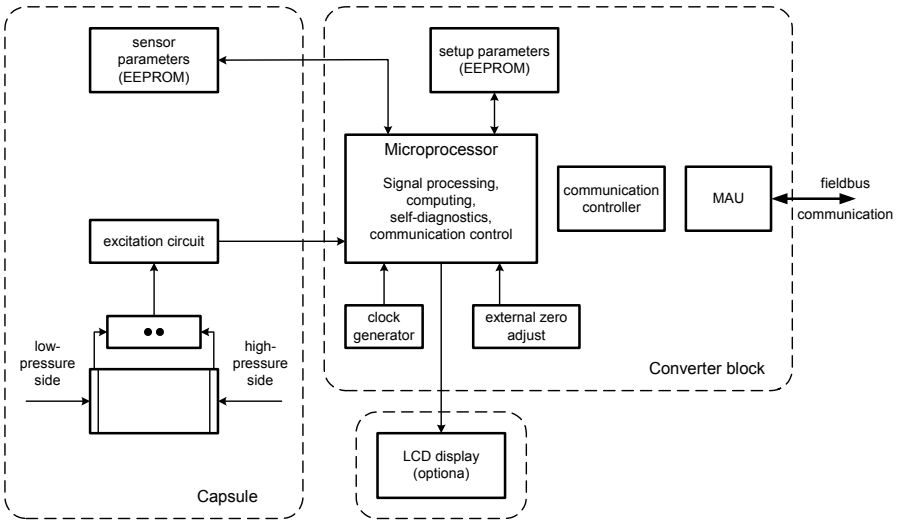


Figure 4.4. *Block diagram of EJA series pressure sensor*

A new EJX series of pressure sensors uses an enhanced two-resonator arrangement, which allows both the differential and static pressure to be measured within one device. The feature is suitable for example for the flow measurement.

4.3. Sensor networks and interfaces

Whether we use an ordinary or an intelligent sensor we still have one problem to solve. The information from the sensor is available at the sensor location and not at the place where it is needed. Overtime many different methods of data transmission have been used to transfer information from the sensor to the controller or any other device that uses it. The principles and concrete data transmission standards that are widely used in industrial practice today are described in this chapter.

4.3.1. *Centralized and distributed industrial systems*

Early electronic industrial systems were built as centralized systems. Here all the sensors and actuators were usually connected to one controller using a star topology interconnection as shown in Figure 4.5. Each partial connection is dedicated to a single sensor or actuator. A central controller acquires all the sensor data. The data are then all processed at one place and commands are calculated and sent to the actuators. The sensors as well as the actuators are usually "dumb" devices, without local knowledge. Either the voltage or current value is typically used to carry information from the sensors to the controller and back to the actuators.

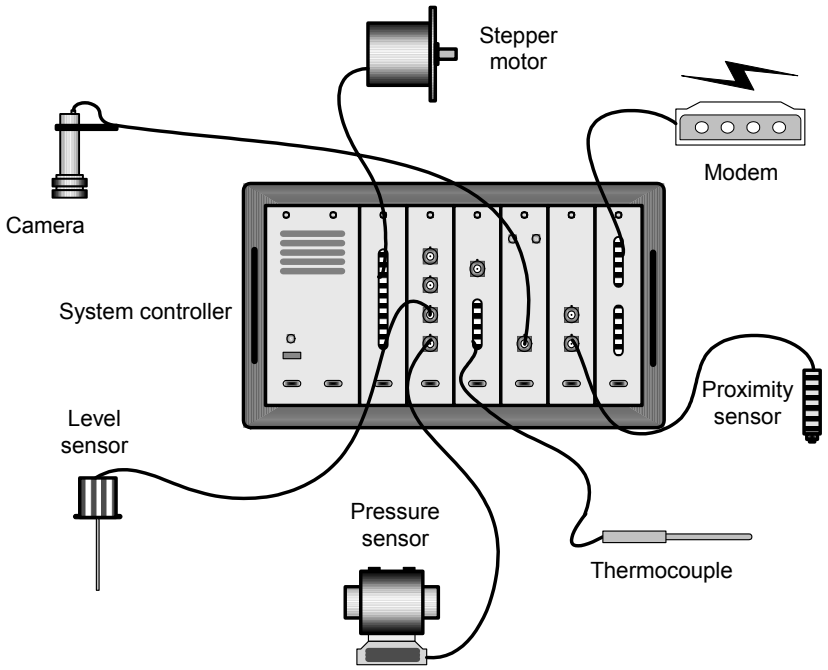


Figure 4.5. *Example of centralized system structure*

This concept has some advantages and disadvantages compared to the distributed system concept, which is described below. The biggest advantage is in uninterrupted and concurrent access to information from all sensors as well the ability to command the actuators at any time. This is enabled by a dedicated communication infrastructure between each sensor or actuator and controller. Generally only ordinary sensors are used, which are cheaper than their intelligent counterparts.

The greatest disadvantage is the high sensitivity of analog data transmission paths to external interferences often found in an industrial environment. If the useful signal and the interference occupy the same frequency band, it is difficult to separate them. The cable attenuation introduces errors for voltage transfer; cable inductance limits the frequency range for current transfer. Star wiring is complex and provides low flexibility. If a new sensor or actuator has to be added, a new cable must be installed, which is not always an easy task. Additionally, if the central controller malfunctions the whole system is functionless. Backing up the controller can solve this problem, but brings additional costs.

Since the late 1980s and especially in the 1990s the distributed systems have started to become popular. In such a system specific intelligent sensors, actuators and controllers create a network using digital communication, as shown in Figure 4.6.

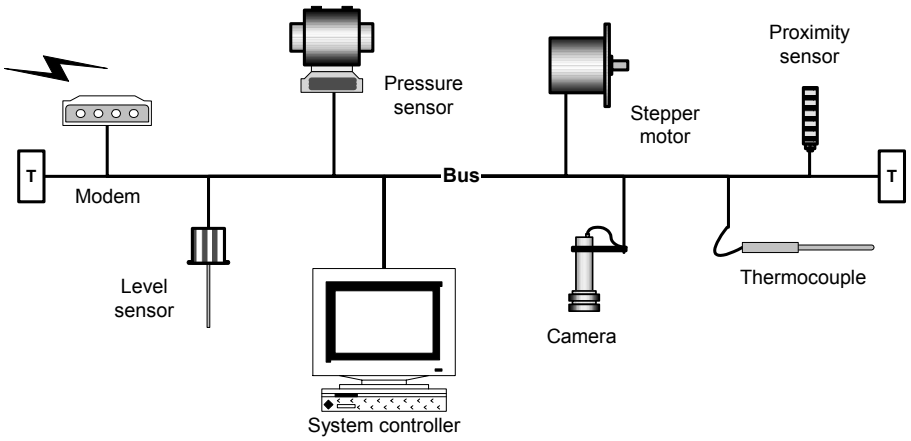


Figure 4.6. *Example of distributed system structure*

The main advantages of this approach are high immunity of data transmission, high cabling flexibility and a variety of communication possibilities using the same physical infrastructure. Information transfer from the intelligent sensor to the system controller (if there is dedicated controller within the system) and back to actuators is digital, allowing data integrity checking and retransmission in case of error. A new node can easily be added to the distributed system, as the physical communication infrastructure is usually shared by all of them. Digital communication also enables the transmission of more than one value, e.g. instantaneous and average or other physical quantities. Moreover diagnostic information can be provided. Specific node

malfunction does not usually stop the whole system; additionally it is easy and relatively cheap to back-up critical nodes.

Of course, there are also some disadvantages. The most important is probably a low immunity to communication channel (cable) breakage, especially for some physical arrangements (see section 4.3.3.2.4). Another issue that must be addressed in control algorithms is a discrete-time availability of sensor data as well as the access to the actuators, because of the physical medium is shared (see section 4.3.3.3.1). Finally, intelligent sensors or actuators are usually more expensive.

4.3.2. *Hierarchical structure of distributed communication*

Communication hierarchy in manufacturing automation systems is usually represented by a communication pyramid. Its form slightly depends on the purpose for which it is intended. In Figure 4.7 a hierarchy of distributed communication is shown.

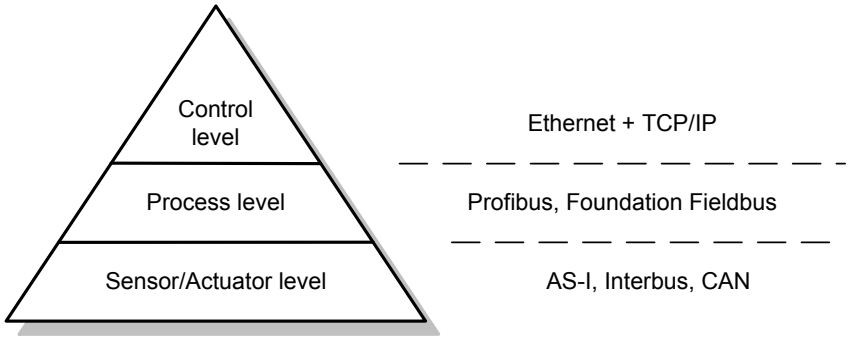


Figure 4.7. *Pyramid of distributed communication*

At the lowest level the high number I/Os, short frames and high-speed communication with low latency is typical to satisfy the real-time requirements. At the process level more complex protocols are necessary to provide communication support for distributed process control. Real-time requirements at this level are not as important. The highest level usually serves for supervision and management, although it can also ensure some high level control.

The assignment of standards to particular hierarchy levels must not be understood as a set rule. It always depends on the application as to which standard best suits its requirements. The sensor networks, that is networks that connect the

intelligent sensors (actuators and controllers), typically belong to the lower layers. Nevertheless, the intelligent sensors can be equipped with an interface that suits the highest hierarchy level.

4.3.3. *Data communication basics*

Modern distributed communication standards are mostly based on an approach introduced by the ISO/OSI (*Open Systems Interconnection*) model. The OSI-based model is used in common distributed system standards – the sensor networks as well as computer networks.

4.3.3.1. *Open Systems Interconnection (OSI) model*

The model defines 7 protocol layers, each of them providing a specific function. Particular layer protocol implementations together form a so-called protocol stack shown in Figure 4.8.

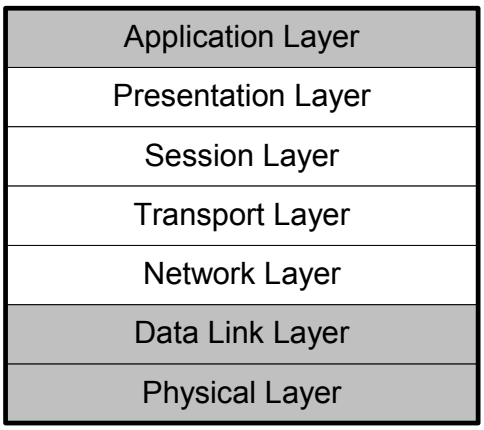


Figure 4.8. *ISO/OSI model*

Each protocol layer provides its services only to the above-adjacent protocol layer and uses only the services of below-adjacent protocol layer. This approach ensures the implementation of lower layers is transparent for higher layers. For example, it allows the network layer implementation to work over different implementations of the data link and physical layers – IP-based networks are the best-known example.

Each protocol layer is implemented in separate instances within particular nodes of the distributed system. An inter-node communication creates virtual communication channels (VC) among the instances of the same protocol layer that are shown in Figure 4.9. At the physical layer both the virtual and physical channel are the same.

The data transferred via the virtual channels are usually referred to as PDU (*Protocol Data Unit*). Physically, of course, the PDUs are sent using the lower layer service and the receiving protocol layer instance obtains them from the lower layer as well. At each protocol layer the PDU consists of two basic parts – the data content (often called the payload) and protocol layer control information (CI). The whole PDU of the protocol layer N is only the payload at protocol layer N-1.

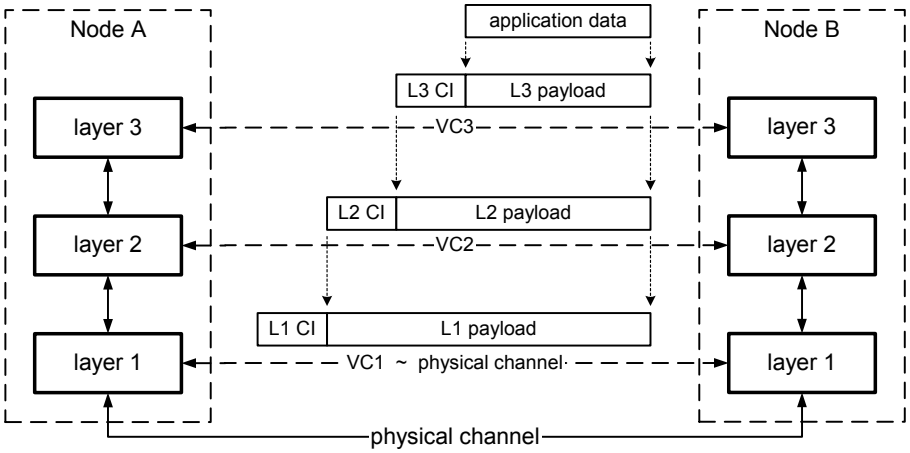


Figure 4.9. *Virtual channels between protocol layer instances*

In small distributed systems (like sensor networks) usually only the function provided by the grey layers (Figure 4.8) is required, while all 7 layers are required for complex computer networks, for example. In the next chapters the function of specific layers is described, particularly focusing on physical, data link and application layers.

4.3.3.2. *Physical layer*

Physical layer protocol defines a physical medium (e.g. copper wire, fiber, wireless) and its parameters, signaling type and parameters (signaling levels), communication speed, mechanical parameters (e.g. connectors), and functional

parameters (e.g. a particular signals sense). It receives a bit sequence from the data link layer and is responsible for its correct transmission and reception, including the bit (or symbol) synchronization. Physical layer protocol creates a basis for upper layers and its behavior significantly influences upper layer implementation. It is present as a basis in all distributed system standards – including sensor networks. In the next chapter the most important physical layer parameters will be described and their influence on higher layers will be explained.

4.3.3.2.1. Baseband and RF band

Baseband signaling is usually considered to contain frequency components from DC to a maximum limit. As many physical channels are not able to carry a DC component (e.g. all wireless channels) or if channel sharing (see section 4.3.3.2.2) is required, the modulation techniques are used to translate the required bandwidth in a frequency spectrum. Figure 4.10 shows examples of power spectrum density envelopes for both baseband and RF band communication.

Most sensor networks use baseband communication, but some are based on RF band communication, e.g. HART (see section 4.3.5.3). Often RF band communication is used for PLC (*Power Line Communication*) physical layer implementations. For example, the LON standard offers this possibility.

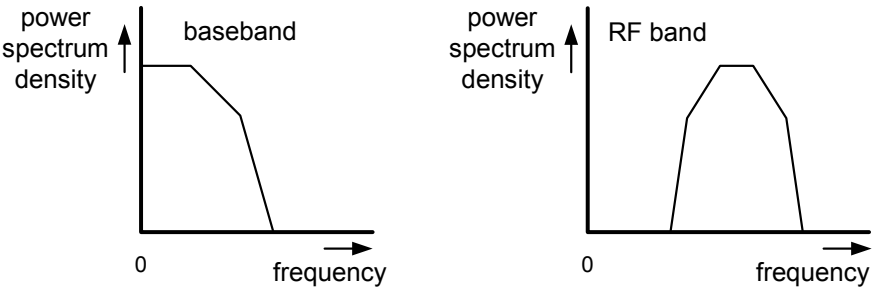


Figure 4.10. *Baseband and RF band frequency spectra*

Wireless standards also have to use RF band communication, as their frequency ranges are limited and regulated.

4.3.3.2.2. Channel capacity sharing

Often one physical channel is used for communication among many nodes. A channel sharing mechanism is therefore required to create a set of logic channels over the single physical channel. Three basic mechanisms are used at physical layer

protocol – TDMA (*Time Division Multiple Access*), FDMA (*Frequency Division Multiple Access*) and CDMA (*Code Division Multiple Access*).

In TDMA systems the whole physical channel capacity is alternately used for particular logic channels. A static assignment schema of physical channel to logic channels is used – each logic channel can access the physical channel for a constant amount of time. Static channel assignment does not suit the industrial distributed system's needs. The MAC methods, belonging to the link layer and described in section 4.3.3.3.1, solve this issue.

FDMA divides the available physical channel frequency bandwidth in a set of frequency-limited sub-channels. Each sub-channel is continuously available for one logic channel. If FDMA is used, RF band communication techniques are necessary. Wireless communication standards always take part in FDMA in wireless communication channels.

In CDMA systems all logic channels may simultaneously share the same frequency band. Particular logic channels are separated by orthogonal pseudo-random sequences used to spread the bandwidth of the original channel signal. The logic channel reception is based on cross correlation between the received signal and the respective pseudorandom sequence. Orthogonality of sequences used for different channels ensures that the cross correlation between any two of them is zero. CDMA is used in wireless network standards (see section 4.3.6).

4.3.3.2.3. Data flow direction

From this point of view the communication channels can be divided to either simplex or duplex. In an area of industrial systems the simplex channel is able to transmit data in just one direction an example of this being an analog current loop (see section 4.3.4.1). The duplex channel allows data transmission in both directions. If the communication can run simultaneously in both directions, we call it a full-duplex channel. If either one or the opposite direction may be used alternately, we call it a half-duplex channel. EIA-232 (section 4.3.4.2.1) is an example of a full-duplex channel, whilst most of the standards described in section 4.3.5 are based on half-duplex channels.

4.3.3.2.4. Physical topologies

Physical topology describes the manner in which the physical network nodes are connected together by physical layer channels. There are several basic topologies shown that are used by most of the systems described below.

A bus topology, shown in Figure 4.11A, is often used in sensor networks standards (e.g. Profibus, CAN, M-bus). Its advantages include short (and low cost)

cabling, flexible configuration changes and independence of any central element. Its main disadvantage is a susceptibility to cable breakage, which divides the network into two separated parts. Bus segment length limit, necessity of termination, limited stubs length and more difficult maintenance and repair are further disadvantages.

A star topology, shown in Figure 4.11B, is more typical for small computer networks. Each node is connected to another via a central element (often called hub). Its advantages are easy structure modification without system shutdown, high cabling redundancy (a cable breakage affects only one node) and easy maintenance and repair. The disadvantages are dependence on the central element and high length of cabling.

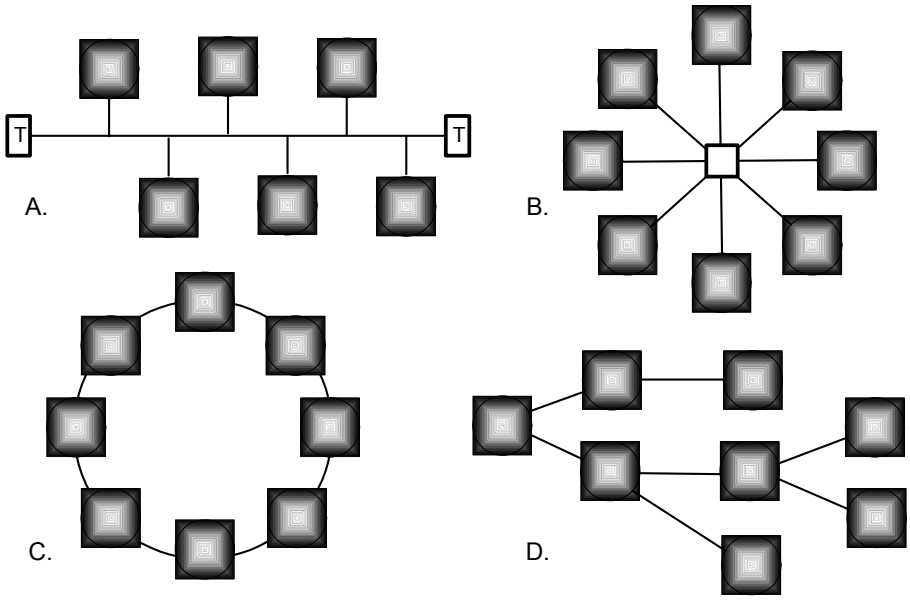


Figure 4.11. *Basic distributed system topologies*

A ring topology, shown in Figure 4.11C, is rather seldomly used in sensor network standards (e.g. Interbus). The advantages are signal regeneration in each node and possibility of addressing nodes by their position. The main disadvantage is a susceptibility to cable breakage, which stops the whole network. In addition, the system has to be shutdown for adding or removing nodes.

A tree topology, shown in Figure 4.11D, is also seldomly used in sensor network standards (e.g. AS-I). The main advantages are minimum cabling length and easy reconfiguration (adding or removing nodes). The disadvantages are limited cabling length (depends on particular standard) and higher susceptibility to cable breakages (depends on position in tree).

It is important to distinguish between physical topologies described here and logical topologies that are sometimes mentioned. The term logic topology describes communication at a higher protocol layer and is described and explained in the section 4.3.3.3.

4.3.3.3. *Data link layer*

Data link protocol layer is used to define the way in which data are transferred within the network. It provides functions and procedures to create, control and release data links between nodes. Its implementation is usually divided into two sublayers – the upper LLC (*Logical Link Control*) and lower MAC (*Media Access Control*). In industrial distributed systems LLC usually implements the frame composition and decomposition, addressing and error checking, while MAC is focused on the implementation of access methods to the physical medium.

4.3.3.3.1. MAC control methods

The physical medium is often shared among all (or many) network nodes. Simultaneous access (a try to transmit) of more than one node leads to collision, which usually destroys the information content. To avoid such collisions several different methods are used that try to coordinate the transmission of particular network nodes. They are divided into two basic groups – deterministic and random. Deterministic methods guarantee the media access within some predefined finite time interval whilst random methods do not. There are many media access methods and variants; additionally, the wireless networks require a different approach than their wired counterparts. This is the reason that we are focusing only on the most important methods.

Master-Slave is the simplest and most widely used deterministic media access method. A dedicated Master node periodically polls Slave nodes allowing them transmission. Data are transferred only between the Master and particular Slaves – direct communication between two Slaves is not possible. This feature and the dependence on a dedicated Master are the main disadvantages of this method. Profibus, AS-I, HART, and many other standards use this method.

Delegated Token method also relies on a dedicated node usually called the Bus Arbiter. It transmits special frames allowing particular nodes to transmit data frames. All other nodes may receive these frames simultaneously (message oriented

addressing is used). Dependence on a dedicated Bus Arbiter is disadvantageous. This deterministic method is used e.g. in Foundation Fieldbus (see section 4.3.5.4) or in an LIN (*Local Interconnect Network*) focused on the automotive industry.

Token Passing is another deterministic media access method that by contrast uses a distributed algorithm. A special token is passed among nodes, which serves as permission for transmission. If the token holder has no more data to transmit or the maximum time to hold the token is over, it has to pass the token to the next node in the chain. This algorithm does not depend on any dedicated node function; on the other hand it may take some time to establish communication at network start-up or after configuration changes. It is used for example by Profibus Master nodes (see section 4.3.5.7) or in some computer network standards.

Most of the random access methods today rely on variants of CSMA (*Carrier Sense Multiple Access*). All the nodes are generally equal and have the same right to start transmission. If a node wants to transmit data, it waits until the medium is free (no communication in progress). Then it can start transmitting either immediately or after some (usually random) time interval. It is clear that two or more nodes can still start transmitting simultaneously and the collision then occurs on the physical medium. Particular variants of the CSMA method provide different solutions. For us the most important are CSMA/CD (with *Collision Detection*), CSMA/CR (with *Collision Resolution*) and CSMA/CA (with *Collision Avoidance*).

Popular IEEE 802.3 networks are based on the CSMA/CD medium access method. If a collision is detected, each participating node waits for a random time interval (its distribution depends on the number of frame transmission repetitions), until the medium is free and then tries transmission again.

CSMA/CR access method is implemented in CAN (see section 4.3.5.2). It is based on special physical layer behavior where collision does not destroy the content of transmitted frames (more exactly the content of the frame with the highest priority). Nodes that transmit frames with lower priority have to wait until the bus is free and try transmitting again.

CSMA/CA method is mostly used in wireless networks, where some nodes may not detect the transmission of currently transmitting nodes and therefore they are not able to detect collisions. If the channel free condition is detected, the node still waits for a predefined time interval and if the channel is free again, it starts transmitting. Often a channel reservation mechanism is implemented to further decrease collision probability. Special frames are used for the channel reservation request and acknowledge. They contain a time interval for which channel will be occupied if request is acknowledged. All nodes receive them and consider the channel is not free during the acknowledged time interval even if they do not receive anything.

4.3.3.3.2. Data link layer addressing

Addressing defines the way in which the frames are passed from the source to the destination. There are two basic addressing schemas – node-oriented and message-oriented addressing.

Node-oriented addressing is generally used in industrial distributed systems. A unique address is assigned to each node (often called MAC address, which is a well known term from computer networks), which is used to identify either the source or the destination of the data link layer frame. Some addresses may be reserved for broadcast or multicast services. All frames transmitted by a node have the same source address. Particular nodes receive only frames with their destination or respective broadcast or multicast addresses. Profibus, AS-Interface, HART, IEEE 802.3 and others use this addressing schema.

Interbus (see section 4.3.5.5) uses a special type of node-oriented addressing. An individual node address is defined by its order in the physical ring. The data link layer frame does not contain any node addresses and each Slave node uses a dedicated part of the frame (defined by the order in ring) for the data exchange with the Master node.

Message-oriented addressing is usually used in systems where data link frames are broadcast into the network. Each message is identified by a unique address (often called an identifier), which defines its content at the data link layer. Particular nodes can usually transmit frames with different addresses depending on the data content. The frame address does not identify a destination node; all nodes interested in particular data may receive the frame simultaneously. This principle is best known from the CAN and CAN-based standards (see section 4.3.5.2); partially it is used in Foundation Fieldbus.

4.3.3.3.3. Error control mechanisms

As mentioned in section 4.3.1, digital transmission allows easy identification of data transmission errors. Data link layer frames are always supplied with redundant information that is used for error control – usually error detection; error correction methods are not used in industrial distributed systems. This information field is often called FCS (*Frame Check Sum* or *Frame Check Sequence* – with respect to the algorithm used for their calculation). The algorithms which are used most often are checksum and CRC (*Cyclic Redundancy Check*).

Checksum is usually calculated as an arithmetic sum of frame bytes modulo 256 resulting in the 8-bit value CS (*Checksum*). The CS value is then transmitted as the last (or nearly the last) byte of the frame. The receiver calculates incoming data modulo 256 in the same way and compares the calculated CS with the received one.

If they are the same, the data are assumed to have been correctly received. Alternatively a difference (256 – CS) is transmitted; in this case the receiver-evaluated sum modulo 256 including the FCS value should be zero (it can easily be detected) for correct data transmission.

CRC is usually calculated as a reminder of division between the frame data field, interpreted as a binary polynomial D(*x*) and a so-called generating (or key) polynomial G(x) with degree ***n***. The maximum reminder's degree is then ***n*** – 1, which means it is ***n*** bits long. Many different polynomials are used in practice – CRC-16, CRC-CCIT and CRC-32 are the best known. In addition, the algorithm and its implementation have many variants. The basic one simply divides the input polynomial D(*x*) by the generating polynomial G(x) and the reminder R(x) is sent as FCS. At the receiver side the R(x) is computed again and compared with the received FCS value. If both values are the same the data are assumed to have been received correctly. This implementation is used for example in CAN standard (see section 4.3.5.2). Often the variant is used where the input polynomial D(x) is at first multiplied by x^n (this means that ***n*** zero bits is added) to form D'(x). D'(x) is then divided by G(x) and R'(x) is obtained and sent as FCS. At the receiver side the sum of D'(x) + R'(x) is used as an input and is divided by G(x). In this case the R(x) should always be zero (easily detected at the receiver side).

CRC is generally better at detecting errors in the transferred data. It is easy to see that if two data bit values change (the first bit in one byte from 0x02 to 0x03, the second bit in other byte from 0x05 to 0x04) the checksum is the same and an error is not detected. For CRC algorithms different data can produce the same CRC value, but the number of required bit changes (the so-called Hamming distance of code words) is usually higher and depends on the selected generating polynomial.

If the receiver detects an error, it ignores the frame and does not pass its data to the upper protocol layer. Then it either requests that the transmitter repeat the transmission of the same frame or does nothing and relies on upper layer protocols.

4.3.3.4. *Network layer*

Network layer protocols determine routing from the source to the destination node. They also provide a network layer addressing schema and its translation into data link (MAC) addresses. There are two basic routing mechanisms – circuit switching or packet switching.

In a circuit switching network initially the dedicated interconnection between the source and destination nodes is established. All packets are then routed via this path. Finally, the connection is closed and network resources are released. PSTN (*Public*

Switched Telephone Network) or ATM (*Asynchronous Transfer Mode*) based networks are examples of circuit switching networks.

In a packet switching network specific packets are routed individually. For each packet an optimal routing strategy is searched and used. This means that depending on the utilization of particular sub-networks at that moment each packet can be routed different way, resulting for example in changes to the ordering of packets. All IP-based networks use packet switching.

In sensor networks the internetworking is seldom used and most standards don't define the network layer. Meter bus (see section 4.3.5.6) is an exception.

4.3.3.5. *Transport layer*

The transport layer is used to provide the service for the upper layers that hides the lower layers characteristics and imperfections. It handles error detection and recovery, received packet reordering (to get the original transmission order), and large data fragmentation. It is generally not needed in sensor networks, as some of its functions are moved either to the data link layer (especially the error control mechanism) or the application layer.

4.3.3.6. *Session layer*

The session layer allows two nodes of a distributed system to establish, use and finally end a session. It is usually not needed in sensor networks or its functions are moved to the application layer, e.g. for remote sensors connected through PSTN or GSM networks via modem. In these cases the session has to be established by the modem before communication can take place.

4.3.3.7. *Presentation layer*

Presentation layer protocols are responsible for data translation into and from a common intermediate format. They can provide for example little and big endian conversion, character set translation or data compression. In sensor networks its functions are (if necessary) included in an application layer.

4.3.3.8. *Application layer*

Application layer protocols and services provide the upper-most interface for application programs. The purpose (with the support of lower layers) is to hide the network structure and resource distribution from the application program. The application then uses network resources (e.g. variables in other nodes) in a similar or the same manner as the local ones.

4.3.3.9. *Data distribution models*

A data distribution model defines the information flow within the system. Four basic service primitives are used to describe communication – request, indication, response and confirmation. Their usage is shown on the Client-Server and Publisher-Subscriber models that are the most important.

The Client-Server model describes communication between two network entities (nodes). The node role as a Client or Server is not assigned forever. The same node may be a Client for one data transfer and Server for second one. Figure 4.12 shows a description of confirmed communication service based on a Client-Server model.

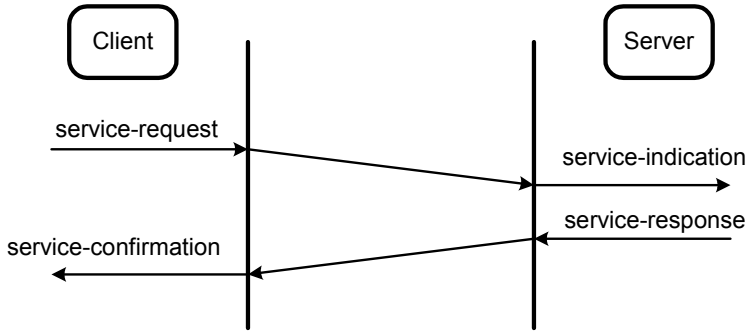


Figure 4.12. *Confirmed communication service in Client-Server model*

The node that needs a service (Client) provided by another node (Server) uses a service-request primitive. The request is transferred through the network and the Server node receives a service-indication – the information about the service that it is asked to provide. After it is processed (either with success or not) the Server uses the service-response primitive to send the result back through the network to the Client. When delivered, the service-confirmation primitive is invoked, which passes the result to the Client. For unconfirmed service only the two first service primitives are used and no information is transferred back.

The Producer-Consumer model describes data flow from one node (Producer) to many nodes (Consumers). The information is broadcast to the network and Consumers receive it as required. Services are mostly defined as unconfirmed, as a Producer may not know the number of Consumers. An example is given in Figure 4.13.

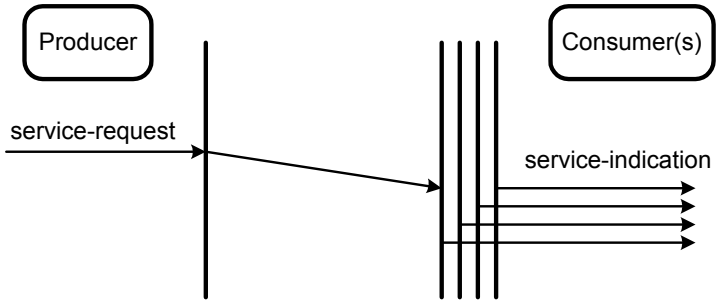


Figure 4.13. *Unconfirmed communication service in Producer-Consumer model*

Different models may be used at specific OSI protocol layers. For example, the CANopen standard (see section 4.3.5.2) uses a Producer-Consumer model in the data link layer (CAN) and both of these models in the application layer.

4.3.4. *Simple sensor interfaces*

Sensors that are targeted to the centralized systems or that are used as standalone in tiny applications are always equipped with some kind of simple output, either the analog or digital. In case of digital communication they do not implement the full protocol stack as described in section 4.3.2. Usually a simple set of application level commands is used to read the measured values or to set internal sensor parameters. Command set is usually dependent on the manufacturer or even on the product and does not comply with any standard.

First it is necessary to explain the terms analog and digital interface that are used in the next two sections. An analog interface uses an instantaneous value of some electric quantity to carry out the instantaneous (or other) measurement of a physical quantity. The values of the electric quantity used are continuous. A digital interface, on the other hand, carries digitized data (after the A/D conversion) and instantaneous values of a selected electric quantity are only used to distinguish between logic levels (or, more generally, between information symbols). In both cases the measured value can either be analog (e.g. temperature) or digital (e.g. a switch state).

4.3.4.1. *Analog interfaces*

The most common analog interfaces use either the voltage or current values as information carriers. Generally any voltage range can be used, but in practice a range of 0-10 V is the most common for analog voltage output. Current outputs

usually use either the 0-20 mA or 4-20 mA ranges. The latter also provides the loop break detection feature in case the current falls below 4 mA. Sometimes the sensors with 4-20 mA current loop are powered via the interface.

Although the voltage and current are most common, other electric quantities can be used as well, e.g. frequency, duty cycle or pulse count.

Digital sensors (level switches, proximity sensors) often use miscellaneous output voltage levels (5 V, 12 V, 24 V or 48 V) according to the technology in which they are intended to be used.

4.3.4.2. *Digital interfaces*

A digital interface can generally be serial or parallel. Although there are some sensors with a parallel digital interface (e.g. the absolute angular position sensor described in Chapter 7), most digital sensor interfaces are serial. They are usually (but not always) based on UART (*Universal Asynchronous Receiver Transmitter*) character format, which is shown in Figure 4.14. It requires the physical layer to distinguish between two logic levels – log.0 and log.1.

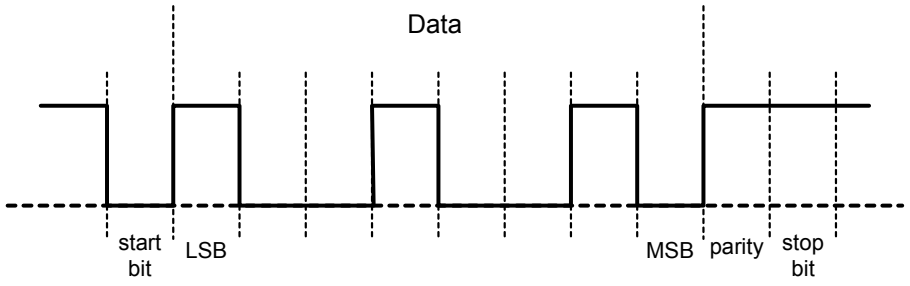


Figure 4.14. *UART character structure*

Interface idle state is represented by log.1. Character transmission starts with log.0 level bit, which is called start bit. Then the data bits follow (usually 8, but other numbers from 5 to 9 are also possible). The LSB (*least significant bit*) is transmitted first, the MSB (*most significant bit*) last. As a consequence, the data content of the character shown in Figure 4.14, interpreted as an 8-bit hexadecimal number, is 0x49. After the data bits there is an optional parity (either odd or even) bit, followed by one or two stop bits of log.1 level. Information is coded into the data bits; usually a sequence of characters is used. The start bit is used to synchronize the receiver, the parity bit serves for error detection purposes and the

stop bit (-s) separate particular characters in sequence. There are several widely used standards that exploit this data transmission format.

4.3.4.2.1. EIA-232

This standard is also known as ITU-T V.24 or under the older title RS-232. It is a rather complex standard defining point-to-point communication. There are two device types defined – DTE (*Data Terminal Equipment,* usually a computer) and DCE (*Data Circuit-terminating Equipment*, usually a modem). The standard was primarily designed to connect DTE equipment to DCE equipment for communication via PSTN (*Public Switched Telecom Network*), but it was quickly adopted for other applications. It provides many different configurations, allowing both the synchronous and asynchronous communication, data flow control and so on. The configuration well known from IBM PC compatible computers, which is probably the most widely used EIA-232 version, is described below.

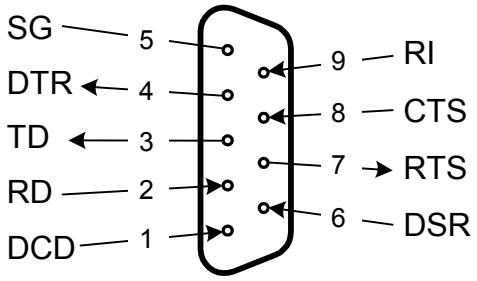


Figure 4.15. *EIA232 interface on 9-pin D-Sub connector*

As the information about the standard is easily available from many Internet resources, only a brief description of the PC interface version is provided here. It consists of a full duplex communication channel equipped with data flow control and further signals, as shown in Figure 4.15. Signals TD (*Transmitted Data*), RD (*Received Data*) and SG (*Signal Ground*) form the full duplex communication path. Signals RTS (*Request To Send*) and CTS (*Clear To Send*) may provide data flow control mechanism. The DTR (*DTE Ready*) and DSR (*DCE Ready*) signal are used for readiness confirmation. The RLSD (*Received Line Signal Detect*, an older name is DCD – *Data Carrier Detect*) and RI (*Ring Indicator*) signals indicate carrier signal presence and ring signal detection at the modem input.

Only the first five (or even three) signals are usually used in communication within an industrial environment, as the rest of them do not have any useful functions. Some cable configurations became de facto standards for this purpose,

e.g. the null-modem or three-wire cables. Unfortunately, there are examples where industrial equipment manufacturers use these signals in a specific and non-standard manner.

All signals use single-ended voltage signaling. The voltage range from +3 V to +25 V represents log.0, the range from –3 V to –25 V represents log. 1. The maximum recommended signaling rate is 20 kb/s, but higher rates are often used (e.g. the PC implementation limit is 115.2 kb/s). The maximum recommended cable length is up to 15 m. The actual cable length that can be used in practice depends on the signaling rate and environmental conditions (external interference).

4.3.4.2.2. EIA-423

Unlike EIA-232, this standard defines only the electrical characteristics of drivers and receivers. It is intended for point-to-point or multi-drop communication with one transmitter and up to 10 receivers. Again it is a single-ended (unbalanced) communication with a maximum signaling rate of 100 kb/s (depending on cable length). The voltage range (driver side) from +3.6 V to +6 V represents log.0, the range from –3.6 V to –6 V represents log.1. Maximum recommended cable length is up to 1,200 m. EIA-423 does not define connector or particular signal usage, for example EIA-449 or proprietary arrangements can be used. This standard is not often used in an industrial environment.

4.3.4.2.3. EIA-422

EIA 422 is similar to EIA 423 standard. It is again intended for point-to-point or multi-drop communication, but it uses a differential (balanced) signaling. It again defines only electrical parameters of drivers and receivers; neither the cable nor connector is defined. Maximum signaling rate is about 10 Mb/s (depending on cable length); maximum cable length is 1,200 m. The differential voltage range (driver side) from +2 V to +6 V represents log.0, the range from –2 V to –6 V represents log. 1. The standard is widely used in an industrial environment, usually with a proprietary connector arrangement. Either the simplex (one 2-wire line) or full-duplex (two 2-wire lines) communication is possible.

4.3.4.2.4. EIA-485

Like the previous standard this defines the electrical characteristics of drivers and receivers for point-to-point, multi-drop as well as for multipoint communication using differential (balanced) signaling. The main difference in comparison to the EIA-422 is a support for multipoint communication. It consists of extended common mode range, protection in case of driver contention and higher drive current. Up to 32 so-called unit loads (usually each unit load equals one EIA-485 transceiver, but half or even less unit load transceivers are also available) can be connected to the

single bus, providing a half-duplex communication channel. Signaling rate, logic level ranges and maximum cable length are the same like for EIA-422. EIA-485 standard is widely used in an industrial environment either in proprietary application or as a part of the physical layer definition in many distributed system standards.

The signaling rate of all previously mentioned standards (except of EIA-232) depends on actual cable parameters, particularly its length. Figure 4.16 shows this dependence approximately for terminated 24AWG copper twisted-pair cable with 50 pF/m capacitance. Actual data rate reached in concrete application depends also on the level of external electromagnetic interference, but for the common industrial environment presented dependence can be used as a first approximation.

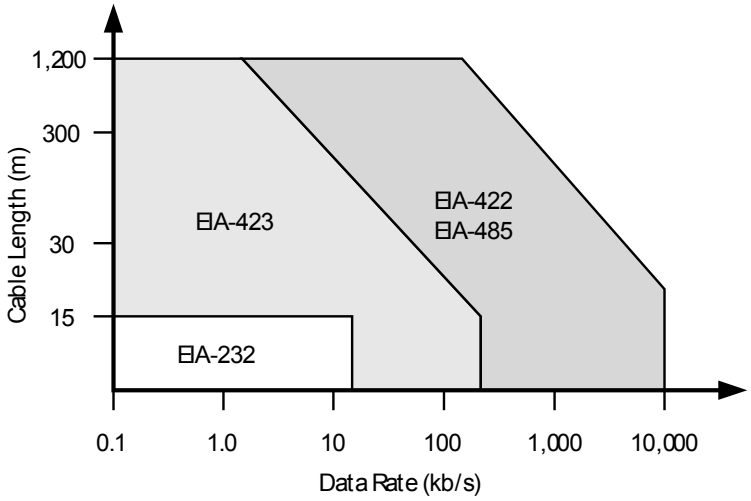


Figure 4.16. *Maximum signaling rate vs. cable length*

Thanks to the high CMRR (*Common Mode Rejection Ratio*) the balanced communication standards (EIA-422, EIA-485) provide higher immunity to external interference which is typical for an industrial environment. Any unbalance introduced either by interconnection or the communicating nodes leads to the decreased immunity and higher bit-error rates.

4.3.4.2.5. Digital current loop

The current loop need not only be used to transfer analog information (see section 4.3.4.1) but digital communication can be carried out as well. Logic levels are implemented by the loop current values – log.0 if current is not present (or if it is

low) and log.1 if current is present. Usually 20 mA current loops are used today. Signaling rate depends on the cable length. Maximum signaling rates are usually in the order of 10 kb/s, for very low speeds (e.g. 300 b/s) the cable length may reach several kilometers.

There is no common standard available for digital current loop communication.

4.3.5. *Sensor networks*

In this section the most widely used wired sensor network standards are briefly introduced. Their description is focused on the physical and data link layer protocols, as they are the most important at sensor/actuator level. The standards that are more often used as sensor networks are described in more detail. CANopen standard application protocol is also described in detail in order to provide an example of more complex application layer protocols.

4.3.5.1. *AS-Interface*

Also called AS-I (*Actuator Sensor Interface*), this standard is one of the simplest from both an implementation and exploitation point of view. It is a typical sensor/actuator bus with a short cycle time. It is particularly suited for interconnection of simple digital and analog I/Os. An important advantage is the use of a single 2-wire unshielded cable for both the communication and power supply distribution. As with most sensor networks the AS-I specification defines only the physical, data link and application protocol layers.

4.3.5.1.1. AS-I physical layer protocols

The most important feature of the AS-I physical layer is the communication and power distribution over the same line. It offers free physical topology (no termination is required) but the total segment length (including all branches and stubs) is limited to 100 m. For larger networks up to two repeaters may be used that increase the maximum system length to 300 m. The system consists of three component types – Master, Power Supply and arbitrary number of Slaves, as shown in Figure 4.17.

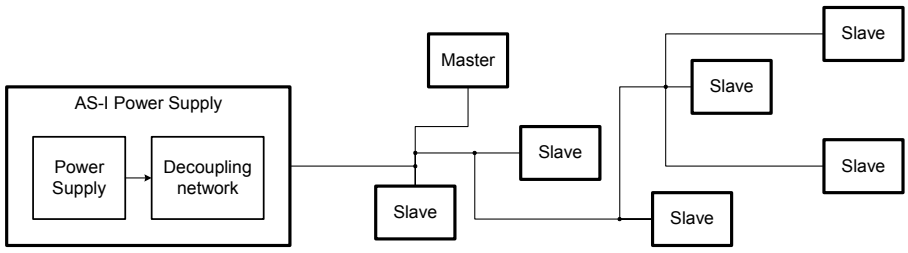


Figure 4.17. *Possible AS-Interface system structure*

The power supply provides power for other network nodes (usually up to 8 A, 24 V nominal) and it also provides the decoupling network for data communication. It can simply be imagined as an inductance in series with the power source output. Data to be transmitted are Manchester encoded at first. Specific network nodes then transmit using supply current modulation. Voltage spikes therefore appear on the network and the receivers detect them. This method is called APM (*Alternating Pulse Modulation*). Up to 62 Slaves can be connected into one system; the node's current consumption is limited to 200 mA, but overall current consumption is limited either by power supply current limit or by voltage drop on the cable. Physical bit-rate is 167 kb/s.

4.3.5.1.2. AS-I data link layer protocols

AS-I implements the Master-Slave medium access algorithm. There are two addressing variants – the original allows addressing for up to 31 Slaves, the new one up to 62 Slaves. The Master cyclically polls Slaves; the cycle time is about 5 ms for 31 Slaves. The structure of request and acknowledge frames is shown in Figure 4.18.

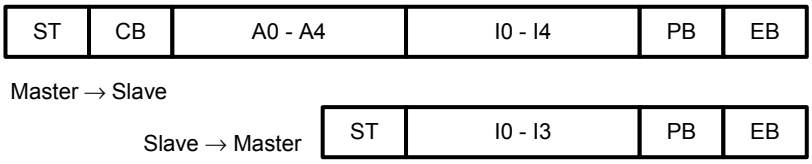


Figure 4.18. *AS-I data link layer frames*

Each frame begins with a start bit (ST), which is always log.0. The control bit (CB) in the Master frame distinguishes between the control (CB = log.1) and data/parameters (CB = log.0) request. Five bits (A0 – A4) are used to address the Slave node; five bits (I0 – I5) carry the information. The even parity bit (PB) is used

for error detection. End bit (EB) is always log.1. The new addressing schema allowing up to 62 Slaves within the system also uses bit I3 as an address bit.

Error checking mechanisms use the start, end, and parity bit, but the constant frame length and APM coding rules are also checked. The corresponding Hamming distance is equal to 4. If any error is detected, the frame is considered wrong and after the timeout its transmission is repeated (in case of Slave response the Master must first request retransmission).

4.3.5.1.3. AS-I application layer protocols

Application messages are transferred by means of transactions. For short messages (up to 4 bits) so-called single transactions are used, consisting of one frame in each direction. For longer messages combined transactions are used, consisting of two or more data link layer frames in sequence. In particular the Slaves with analog sensors or actuators require more than 4-bit values to be transferred. Different combined transactions are defined to transfer particular data types. Slave profiles are defined that standardize support for particular transaction types, addressing schema, parameters usage and others. Older versions of AS-I standard have provided only single transactions and analog I/Os were not supported.

Besides the standard process data exchange application protocols are available for identification of Slaves, Slave address assignment and diagnostics.

4.3.5.1.4. AS-I summary

– Free physical topology, maximum segment length 100 m (300 m with repeaters), power and communication over single 2-wire cable.

– Up to 31 Slaves (for the new addressing schema 62), cycle time 5 ms (for 31 Slaves), preferably short data exchange (up to 4 bits).

– Easy exploitation, Slave replacement in running system possible.

– Easy and standardized integration with other systems, e.g. Profibus or CANopen.

4.3.5.2. *CAN (Controller Area Network) and CANopen*

CAN is used in miscellaneous industrial applications at the sensor/actuator and process levels (see section 4.3.2). It comes from the automotive industry and up to now it has been widely used for communication among Electronic Control Units (ECUs) in vehicles. Thanks to this fact CAN hardware is cheap and widely available. CAN is just a data link layer protocol and number of incompatible physical and application protocols exist. This section focuses on those that are used in industrial automation systems.

4.3.5.2.1. CAN physical layer protocols

So-called high-speed CAN is mostly used in industrial applications. It is based on a balanced bus with a line impedance of 120 Ω. Baseband signaling uses two bus states, dominant and recessive, as required by the data link layer. There is an important rule for the bus state: if all nodes transmit the recessive value, the bus state is recessive; if one or more nodes transmit the dominant value, the bus state is dominant regardless of number of nodes transmitting the recessive value. The dominant state is equal to log.0 and it is signaled by a differential voltage of at least 2 V on the bus; the recessive state is equal to log.1 and it is signaled by zero differential voltage on the bus. Maximum bus length depends on selected bit-rate – for a maximum 1 Mb/s bit-rate the bus length may be up to 40 m, for 50 kb/s a total bus length of 1,600 m is available.

4.3.5.2.2. CAN data link layer protocols

CAN uses a message-oriented addressing schema as described in section 4.3.3.3.2. The CAN data frame structure is shown in Figure 4.19. It begins with the frame identifier, which provides the information about the frame content. In CAN it also defines the frame priority – the lower the identifier, the higher the priority. CSMA/CR (*Carrier Sense Multiple Access with Collision Resolution*) medium access method is used. All nodes have the same right to start transmitting after the idle state is detected on the bus (predefined number of recessive bits is received).

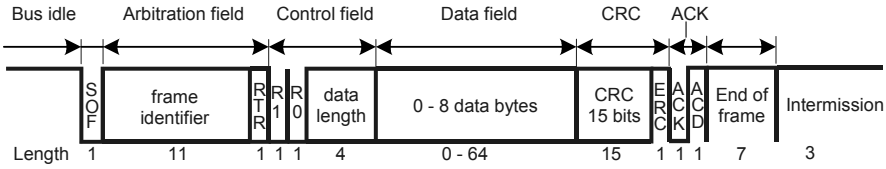


Figure 4.19. *CAN data frame structure*

The collision that possibly happens, when two or more nodes start transmitting at the same time, is resolved in the following manner. Each frame starts with the start-of-frame bit, which is always dominant, followed by the so-called Arbitration field. When two or more nodes start transmitting simultaneously, the resulting bus state is received back and the transmitting node receiving the dominant level while sending the recessive one stops transmitting. As the identifiers are unique, i.e. only one node can transmit the frame with one particular identifier, the colliding frames differ at least in one identifier bit and thus the collision is resolved. Thanks to the physical layer feature the collision is non-destructive and the winning node continues to transmit the following fields (control, data and CRC).

Before the frame is passed to the physical layer a bit-stuffing procedure takes place. During it a stuffing bit of opposite logic level is inserted after each sequence of 5 consecutive bits of the same logic level. This mechanism ensures enough changes in a bit stream for the receivers to keep synchronization. At the receivers the stuffing bits are removed using the inverse algorithm.

As the frames are broadcast to the network, all nodes may receive them simultaneously creating on efficient use of available bandwidth. To keep the data consistency within the whole network, an error signaling method is used ensuring that either all nodes receive the frame or no node does. If any node detects an error, it transmits a so-called error frame consisting of 6 consecutive dominant bits. For all other nodes it means a bit-stuffing rule violation and the frame being received is considered wrong. There are several frame error checking mechanisms – CRC, bit-stuffing, frame format check, transmission monitoring (transmitter receives back the transmitted bit values, if there is a difference it again transmits an error frame), and frame acknowledgement check. Receivers should acknowledge the transmitted frame at ACK bit position. ACK is transmitted as recessive by the transmitting node and the receivers transmit a dominant value for this bit only. If the transmitting node does not receive a dominant value of the ACK bit back, it immediately sends an error frame.

4.3.5.2.3. CAN application layer protocols

As already mentioned several incompatible application protocols exist above the CAN data link layer. They are for example DeviceNet, SDS (*Smart Distributed System*) or CANopen. This section is focused on the latter, as it is probably the most widely used in Europe.

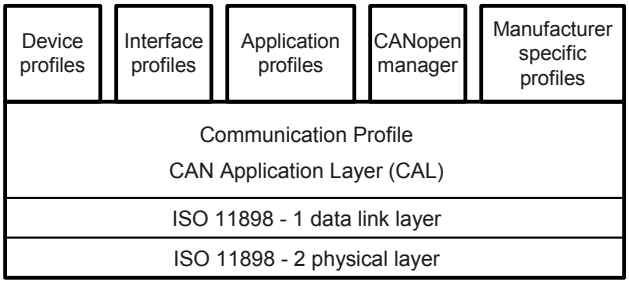


Figure 4.20. *CANopen standard structure*

The CANopen standard was developed by the CiA (*CAN in Automation*) Association. Its structure is rather complex – Figure 4.20 provides a simplified view. The lowest physical and data link layers were described above.

CAL provides a basis for CANopen implementation, but it is a standalone standard not included in CANopen. It defines two protocol groups. The first is a CMS (*CAN Message Specification*), which specifies data types, encoding rules and protocols for variable access. The second group, consisting of NMT (*Network Management*), DBT (*Distributor*) and LMT (*Layer Management*), provides services allowing network configuration and management.

CMS defines abstract objects that are described by a set of attributes (e.g. the name, user type, priority, data type, access type, class an so on). There are three object types – variables, events and domains. The CMS variable can be either basic or multiplexed. A multiplexed variable means a set of variables defined by a single name, which members are accessible using a multiplexer – a pointer to the set. Variable objects are intended to transfer process data of defined types. CMS event objects can be used to support information transfer about asynchronous events, like input limit violation or actuator function failure. They allow additional information to be transferred besides the fact that the event has occurred. Finally, the domains are used for large unstructured data transfers. Similar to variables domain classes can be also either basic or multiplexed. CMS also specifies the services and protocols allowing access to the objects as well as means the requests and responses are coded into CAN frames.

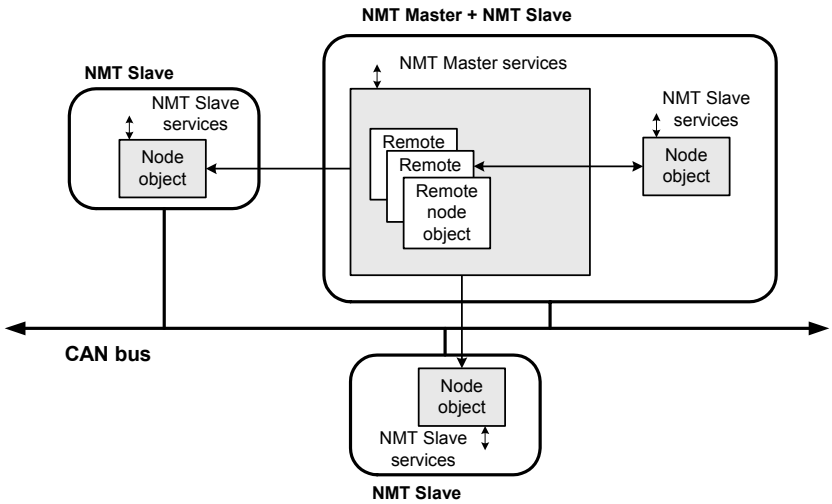


Figure 4.21. *Structure of NMT objects*

NMT provides services and protocols for network parameterization, start-up and management including reporting communication failures. Within NMT the Master-

Slave structure is used, as shown in Figure 4.21. A basic NMT data structure at the Master site is a *network object* containing *remote node objects* – one for each network node. The remote node object mirrors *node objects* that reside in particular network nodes. NMT provides services for controlled network start-up during which the parameters can be downloaded into particular nodes. It also enables the detection and control of communication problems. Each node in the network is identified by an NMT address (in range 1-255, address 0 is used for NMT broadcast) called the Node-ID.

DBT services and protocols may be called-up during network start-up, for example, the CAN frame identifiers are assigned by DBT for frames that are used for CMS communication. These protocols are seldom implemented in CANopen nodes and an alternative setting (allowed by the standard) is provided by manufactures.

LMT serves for particular node identification (each node is defined by a unique LMT address, provided by manufacturer), NMT address assignment and physical layer parameter (e.g. the bit-rate) setting. Again, they are often replaced by alternative manufacturer-dependent settings.

CANopen *communication profile* residing at the top of the CAL is the first and most important part of CANopen standard. Each CANopen node is built around the data structure called *Object Directory*. Node logical structure is shown in Figure 4.22.

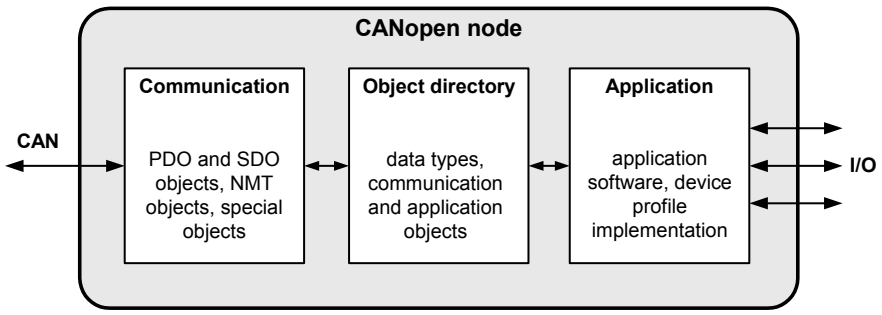


Figure 4.22. *Logical structure of a CANopen node*

The object directory contains definitions of all objects that describe the node behavior from both the application and communication point of view. Specific objects in the directory are accessible using a 16-bit index. For example, in index range 0x0001-0x009f the data types are defined, range 0x1000-0x1fff is used for

objects defined by the communication profile, range 0x6000-0x9fff is used for objects defined by device profiles and so on. Objects in the object directory can be accessed using the CMS multiplexed domain access protocol. An object can be further structured (e.g. array or record). Access to object elements is available through an 8-bit sub-index. For example, at index 0x1018 there is an *Identity object*, which contains manufacturer identification, device type, device version and serial number, each value at a separate sub-index.

The communication profile defines two basic communication object types – PDO (*Process Data Object*) and SDO (*Service Data Object*). PDOs are used during normal network operation to transfer process data. Two objects in the object directory define each PDO. The first object (PDO Communication parameters) defines, for example, the frame identifier (identifier of the CAN frame used to transfer the data), inhibit time (minimum time period in which the PDO can be transmitted again), and transmission type (defines under which condition the PDO is transmitted). The second object (PDO Mapping) allows the user to define which application data is transferred in the PDO. PDO mapping can be simply explained by the following example. Let us take a CANopen device with two digital 8-bit outputs. A dedicated object in the object directory describes each output. If the node receives the CAN frame (it should contain 1 data byte) whose identifier is the same as that specified in PDO communication parameters, the node application software looks to the related PDO Mapping object and copies the frame data at the appropriate 8-bit output. PDO Mapping allows data to be transferred for more than one application object in a single PDO. CMS *stored event* protocols are used for PDOs.

SDOs are used to directly access objects in the object directory. SDO communication takes place especially during the network start-up, where the object directory content may be modified (e.g. PDO mapping is defined). CMS *multiplexed domain* protocols are used for SDOs. SDO objects use low priority identifiers and are always transmitted asynchronously (explained below).

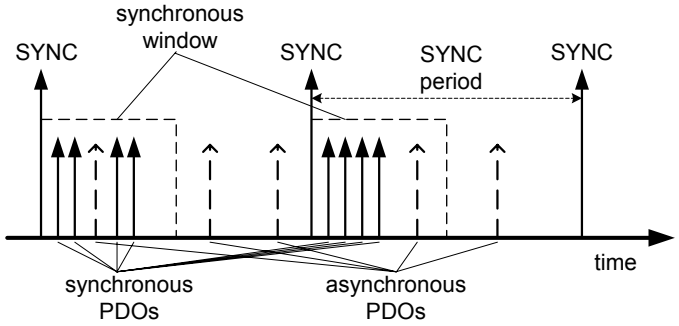


Figure 4.23. *Data transfer synchronization in CANopen network*

Communication profile also defines communication objects that provide special functionality. The most important is a SYNC object, which provides synchronization of data transfers in the network. In Figure 4.23 the principle is clearly visible.

The SYNC object is transmitted periodically to the network. If a PDO is defined as synchronous, it may be transmitted only within a specified time interval (synchronous window) after the SYNC is received. SYNC parameters are defined in the object directory at indexes 0x1005 (SYNC frame identifier), 0x1006 (SYNC period), and 0x1007 (synchronous window length).

The TIMESTAMP object can be used to distribute time information (time of day format) to the network nodes. If more precise time synchronization is necessary for an application, the *High Resolution Synchronization Protocol* can be used, which allows time synchronization down-to 10 μs.

EMERGENCY objects are defined for each node to report device problems or even failure. They use high priority frame identifiers and in the data field there is a standardized error code and error register content.

CANopen *device profiles* are defined in order to ensure compatibility of products from different manufacturers. A node produced by manufacturer A can be replaced by node produced by manufacturer B without any change in the rest of network if they are built according to the same device profile. This statement is not always valid, as profiles allow implementation of optional and even manufacturer-specific features, but for mandatory and optional (if implemented) features it is true.

Computer-aided configuration is supported by EDS (*Electronic Data Sheet*) description files. They provide a general object directory description (template) of particular device types. In design software the EDS file is transformed into the DCF (Device Configuration File) format – one for each device.

4.3.5.2.4. CAN and CANopen Summary

– 120 Ω bus terminated on both ends, maximum bus length depends on selected bit-rate (40 m for 1 Mb/s, 1,600 m for 50 kb/s), differential signaling with dominant and recessive states .

– CSMA/CR medium access method, prioritized frame broadcasting, short waiting time for high priority frame transmission, low cost controllers.

– Very high data integrity, reliable error checking, short error recovery.

– Up to 255 network nodes (NMT limit), protocols for variables, asynchronous events and unstructured domain transfer.

– Nodes described by the object directory, standard communication objects defined by communication profile, device profiles ensure easy device replacement.

– Standardized device description by means of EDS.

4.3.5.3. *HART (Highway Addressable Remote Transducer)*

HART is a digital extension of analog 4–20 mA current loop (see section 4.3.4.1). While keeping the analog current output it allows more complex digital communication with intelligent sensors. It is widely used especially in the field of process instrumentation, where successive replacement of old current loop sensors with new HART sensors is easy and cost effective.

4.3.5.3.1. HART physical layer protocols

RF-band communication is used for digital data transfer by means of FSK (*Frequency Shift Keying*) modulation, which is superimposed on an analog current loop signal. As the mean value of the FSK signal is zero, digital communication does not influence the analog one. Bell 202 modulation standard is used providing 1,200 b/s bit-rate. The Master transmits using voltage modulation and the Slave (-s) using current modulation. More HART sensors may be connected in parallel; in this case only digital communication is possible. A twisted pair cable (shielded or unshielded) is usually used.

4.3.5.3.2. HART data link layer protocols

A Master-Slave medium access method is used together with half-duplex communication. Up to two Master nodes may be connected – usually the permanently installed primary Master and temporarily connected secondary Master (handheld controller). Masters use token-passing to gain control. Data link layer frames (shown in Figure 4.24) consist of UART characters (see section 4.3.4.2).

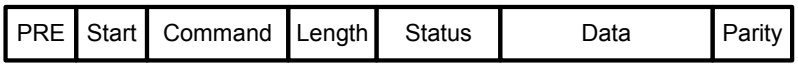


Figure 4.24. *HART data link frame*

A preamble is used to synchronize the receiving node. The *start* character defines the frame format and direction (from Master or from Slave). Address, command, frame length and data fields are self-explanatory. The status field is used only in frames transmitted by Slaves and it contains information about either communication or command relative errors. HART Slave response time is about 400 ms, which means the data update frequency is relatively low. A special burst

mode may be used, in which the Slave transmits data continuously. The update period is then slightly shorter.

4.3.5.3.3. HART application layer protocols

HART commands can be divided into three groups. *Universal* command implementation is mandatory for all HART devices. They identify the device, measured physical quantity and its range, read the measured values and so on. Optional *Common Practice* commands allow, for example, loop current calibration, sensor range alteration or sensor calibration. *Device-specific* commands form the last group. The functions of Host (Master node) are also standardized. *Host conformance classes* are defined that describe particular implementation functionality.

HART uses DDL (*Device Description Language*) to describe device features. DDL provides modeling application data and command description. Its use is optional but most HART devices are accompanied with DDL description.

4.3.5.3.4. HART summary

– 4-20 mA current loop extension with RF band digital communication, 1,200 b/s, cable length up to 3,000 m.

– Master-Slave medium access control, 400 ms transaction time.

– Standardized application commands, standardized device description language, no device profiles.

– Widely used in practice, large number of HART devices available.

4.3.5.4. *Foundation Fieldbus (FF)*

In addition to the communication infrastructure the Foundation Fieldbus partially defines application behavior based on implementation of *function blocks*. Two variants are defined; H1 for the low-speed process communication and HSE (High Speed Ethernet) for the upper control level. The following description of physical and data link layers is focused on H1, as HSE is built upon a standard TCP/IP protocol stack. FF is standardized by IEC61518 or EN50170. An FF communication stack is generally divided into three parts:

– Physical layer.

– Communication stack (common name for OSI data link and application protocol layers).

– User application (layer not defined by OSI model).

4.3.5.4.1. FF physical layer protocols

The H1 physical layer protocol is based on IEC61158. Bit-oriented synchronous transfer with Manchester encoding and special bit sequences for frame delimiters are used. A single pair cable may be used for simultaneous data transmission and power distribution; bus or tree topologies are supported with a maximum segment length of 1.9 km. A communication speed of 31.25 kb/s is used. Particular nodes use a supply current modulation (±10 mA) into the 50 Ω load, resulting in maximum 1 V_{pp} on top of the power supply voltage. Power supply voltage range is 9-32 V. A maximum of 32 nodes per segment is allowed, but the actual number depends on cable type and power consumption of the nodes.

4.3.5.4.2. FF data link layer protocols

At the data link layer FF uses the delegated token method (see section 4.3.3.3.1). A dedicated device called LAS (*Link Active Scheduler*) is used to manage the medium access. Generally the FF devices are either *link Master* devices or *basic* devices. *Link master* device is able to play an LAS role whereas *basic* devices cannot.

Regular scheduled communication uses a delegated token method. The LAS contains a list of all scheduled data frames and their transmission periods. It sequentially issues a CD (*Compel Data*) frame, which requires transmission of a relative DT (*Data*) frame. The DT frame is then published on the bus and all subscribers receive it simultaneously. The control loop data is typically transferred in this way.

Besides the scheduled communication, particular devices are allowed to send frames after receiving a PT (*Pass Token*) frame from the LAS. When a device such as this finishes transmitting or after a timeout it returns the right to transmit back to the LAS by a final flag in DT frame or, if it has no data to transfer, by RT (*Return Token*) frame. Devices that respond correctly on PT frames are kept in a so-called *live list*.

Using the PN (*Probe Node*) frame the LAS checks whether a new node has been connected to the network. If yes, it responds using a PR (*Probe Response*) frame and LAS adds it to the *live list*. The *live list* is then broadcast to all *link master* devices.

Time synchronization is made available using TD (*Time Distribution*) frames. Data link layer scheduling as well as application block execution is based on this time data.

For the HSE FF variant there is no LAS in the network and messages are sent immediately. The new device types are defined for the HSE variant:

– Linking device is used to interconnect H1 and HSE networks. It maps application layer messages between protocol stacks.

– Ethernet device executes *function blocks* (explained in the following section) and optionally provides standard I/Os.

– Gateway device connects HSE to other distributed system standards, like Profibus or Modbus.

– Host device is a non-HSE device which is able to communicate with HSE devices, e.g. operator workstations

4.3.5.4.3. FF application and user layer protocols

The application layer uses the VFD (*Virtual Field Device*) concept with object directory and so on. It can be divided into two sublayers called FAS (*Fieldbus Access Sublayer*) and FMS (*Fieldbus Message Specification*). They provide following services:

– Virtual device management.

– Object directory access.

– Process variable and domain access.

– Event processing.

– Program invocation services.

The highest (user) layer is standardized by means of *blocks*. Particular blocks represent different application functions. There are three basic block types:

– *Resource blocks* are used for device description (manufacturer, type, serial number and so on). There is only one resource block in a device.

– *Function blocks* provide the application behavior; their inputs and outputs are virtually connected over the network. More function blocks may reside in one device. More than 20 standard function blocks are defined, e.g. AI (*Analog Input*), DO (*Discrete Output*), PID (*Proportional/Integral/Derivative*) or TMR (*Timer*). User specific function blocks are implemented by means of FFB (*Flexible Function Block*).

– *Transducer blocks* are used to decouple a function block from the local inputs (sensors) and outputs (actuators).

Block execution is synchronized by LAS in H1 and by synchronization messages in HSE networks.

To ensure easy device interoperability and in system integration DD (Device Description) files are used that describe VFD objects. A DDS (*Device Description Services*) library is available for DD file reading.

4.3.5.4.4. FF summary

– H1 variant with 31.25 kb/s, cable length up to 1,900 m, power supply over communication line, 32 devices per segment (12 with bus power).

– HSE variant based on IEEE 802.3 and TCP/IP.

– Delegated token MAC method, scheduled and unscheduled communication.

– Application protocols based on Virtual Device Concept and FMS.

– User layer defined by blocks, standard function and transducer blocks define application behavior, device description files allow easy integration.

4.3.5.5. *Interbus*

As with AS-I, an Interbus is a typical sensor/actuator system optimized for fast transfers of short data. It is a very interesting Fieldbus as it uses several specific features.

4.3.5.5.1. Interbus physical layer protocols

Interbus topology is a ring consisting of a chain of point-to-point connections between nodes. Each node contains a receiver which receives data from the previous node in a ring and a transmitter which transmits it to the next node in a ring. Nevertheless, the Interbus cable contains wires for both the forward and backward direction and therefore the system structure (see Figure 4.25) may look like a star or even a tree.

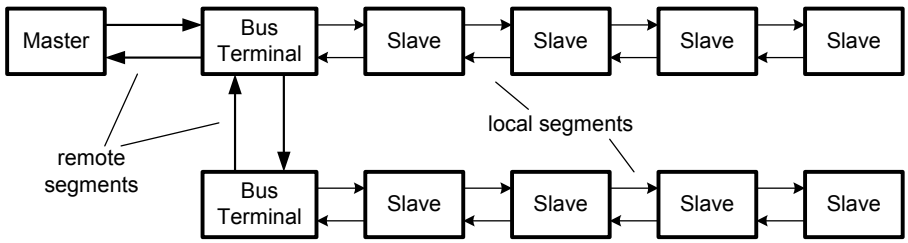


Figure 4.25. *Interbus system structure*

The Interbus loop consists of two types of segments. Local segment length is limited to up to 10 m and the maximum distance between nodes is 1.5 m. They use

TTL level signaling. Local segment cable is used to supply power to the nodes using dedicated wires. Remote segments are based on either the EIA-485 standard (see section 4.3.4.2.4) or on a fiber. For EIA-485 the maximum distance between remote devices is 400 m and total system length can be up to 13 km; for fibers the distance between remote devices depends on fiber type. Bus Terminals are used to transform the remote and local segment physical layers. The 500 kb/s bit-rate is the same for both segment types.

4.3.5.5.2. Interbus data link layer protocols

A Master-Slave medium access control method with a single summing frame (see Figure 4.26) is used. Specific Slaves are addressed by their order in the ring. Master transmits the frame, then it is passed through particular Slaves along the ring and finally it comes back to the Master.

A 16-bit *Loopcheck* field is used to identify beginning of the frame and also for Master to check when received back. Particular data fields (*IO 1 – I/O N*) are used for communication with relative Slaves. Within one circle the new data are delivered from the Master to all Slaves and replaced by responses from Slaves to the Master.

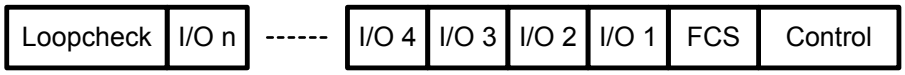


Figure 4.26. *Interbus data link layer frame structure*

A 16-bit FCS (*Frame Check Sequence*) is used for error checking and a *Control* field is used for synchronization and error reporting. Data fields for particular Slaves may have different length; the mechanism is available to find and identify devices.

4.3.5.5.3. Interbus application layer protocols

At the application layer two service types are available – unacknowledged services for fast process I/O with nearly no overhead, and acknowledged services. The first is implemented by means of PDC (*Process Data Channel*); the latter are based on the FMS (*Fieldbus Message Specification*) and VFD (*Virtual Field Device*) concept and implemented by a PCP (*Peripherals Communication Protocol*) channel.

Device profiles are available for particular device types like sensors/actuators (this is a basic profile that all devices should conform to), encoders, process and robot controllers and others. The PCP channel is used to invoke services, e.g. to access object directory variables.

4.3.5.5.4. Interbus summary

– Ring topology with local and remote segments, 500 kb/s bit-rate, maximum system size 13 km.

– Master-Slave access control with addressing by a node position in a ring, very short cycle time, low protocol overhead.

– PDC channel for fast process communication and PCP channel for message-based application layer services, device profiles are defined.

– Widely used for control systems, e.g. in the car manufacturing industry.

4.3.5.6. *M-Bus*

M-Bus (Meter-Bus) is designed for data acquisition from different types of utility meters (water, gas, electricity or heat) in households. It is specialized particularly for this purpose and does not suit other applications. It allows data to be acquired from a large number of nodes (hundreds to thousands) with a high degree of security and at a high distance.

4.3.5.6.1. M-bus physical layer protocols

M-Bus system consists of one Master node, which is usually called the *Repeater* and at most 250 Slaves (within one segment). The repeater also provides power supply to Slave nodes. Segment length is limited to 1,000 m (for 300 b/s, 350 m for 9,600 b/s).

Physical signaling differs by communication direction. From Master to Slaves the line voltage changes are used (the Master is a supply power source). Log.1 is signaled as 36 V at the Master output, while log.0 is signaled as 24 V level. In the opposite direction (Slave to Master) Slave current consumption is used. Log.1 is represented by current consumption up to 1.5 mA and for log.0 the current consumption increases by 11 – 20 mA. As the number of Slaves in a segment may vary, the voltage and current changes are evaluated, not the absolute values.

4.3.5.6.2. M-bus data link layer protocols

The Master-Slave bus access protocol is used with node-oriented addressing. Four data link frames are defined for different purposes; the frame format used to transfer data is shown in Figure 4.27.

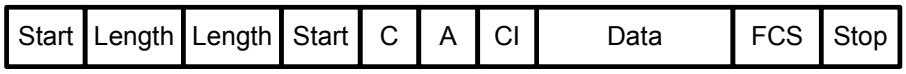


Figure 4.27. *M-Bus control/long frame format*

All frame fields (except the data) are one character long (8 data bits). The Start field defines frame type; here it is 0x68. Next a doubled frame length field is used, containing data length + 3 for C, A and CI fields. Then the Start field is repeated followed by the C (*Control*) field. It either defines control codes or it allows repeated frame receipt detection and data flow control. The A (Address) field serves as a Slave address. Addresses 0–250 may be used, address 253 is used for network layer addressing, addresses 254 and 255 for broadcast and 251 and 252 are reserved for future use. The CI (*Control Information*) field is used by the application layer protocol. Only the long frame contains data, followed by the FCS (*Frame Check Sequence*) in the control frame, where a simple check sum is used. The frame ends with the stop field (code 0x16).

4.3.5.6.3. M-bus network layer protocols

The network layer is used to extend the number of Slaves in system over 250. The physical address defined by ID number, Manufacturer, Version and Medium is used to assign address 253 to a selected Slave. After the required data exchange the address 253 may be reassigned to another Slave.

4.3.5.6.4. M-bus application layer protocols

EN1434-3 is a basis for M-Bus application protocols. Most often the application data are transferred using a Variable Data Structure within a data field of long frame. Variable Data Structure formatted data are shown in Figure 4.28.

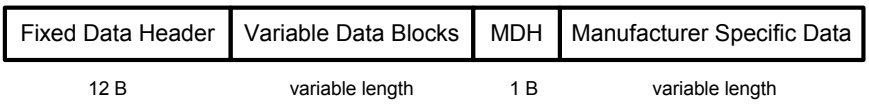


Figure 4.28. *Application data in Variable Data Structure format*

Any number of values may be transferred; only the maximum long frame length limits it. The fixed *Data Header* field contains meter identification (ID number, manufacturer code, version and so on). It is followed by a variable number of data blocks. In each data block a specific measured value is transferred, specified by value type (e.g. instantaneous, minimum), measured quantity (e.g. volume) and its unit (e.g. m^3), data type (integer, floating point) and value coding type (BCD, binary). Besides these standardized data blocks a *Manufacturer Specific Data* may also be transferred, indicated by an MDH (*Manufacturer Data Header*). Transferred data may be protected against intentional changes using a method similar to a digital signature.

4.3.5.6.5. M-Bus summary

– Low-speed (300–9,600 b/s) UART-based communication, power distribution over communication line, bus physical topology, segment length 1,000 m

– Master-Slave bus access method, up to 250 Slaves in segment

– Network layer implementation allows increasing number of Slaves in system

– Application layer is focused on utility meters reading and control

4.3.5.7. *Profibus*

Profibus (Process Fieldbus) is one of most widely used distributed systems, particularly in Europe. It is not a single standard, but a whole family of standards, which is supported by international standards EN 50170, EN 50254 and IEC 64158. Altogether there are three basic Profibus variants:

– Profibus DP (*Decentralized Peripherals*) is the most widely used variant focusing on the factory automation area.

– Profibus FMS is a Fieldbus Message Specification-based standard for higher-level communication. It uses a VFD concept. It is especially targeted for intelligent controller communication and it is rarely used.

– Profibus PA (*Process Automation*) is based on DP functions and is focused on automation of processes.

The standard structure and particular variants are shown in Figure 4.29.

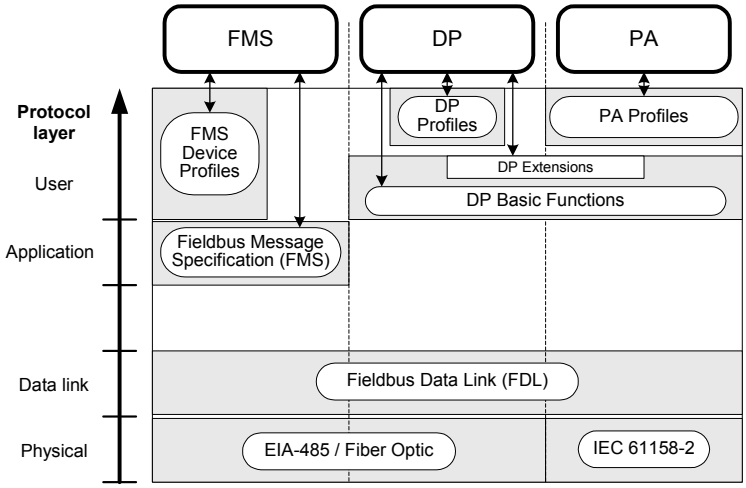


Figure 4.29. *Profibus standard structure*

4.3.5.7.1. Profibus physical layer protocols

There are three possible physical layers. The PA variant uses the same physical layer implementation according to IEC 61158-2 like FF. It is described in section 4.3.5.4.1. DP and FMS use the same physical layer – either the EIA-485, described in section 4.3.4.2.4 or a fiber optic. If EIA-485 is used, a bit-rate up to 12 Mb/s may be used for 100 m segment length. For bit-rate 93.75 kb/s the length limit is 1.2 km. Three repeaters are allowed providing a total bus length of 4.8 km.

4.3.5.7.2. Profibus data link layer protocols

The data link layer interface is the same for all three Profibus variants. It uses the Master-Slave medium access control method, which (usually in large systems) can be combined with a token-passing method among the Master nodes. Asynchronous UART-based communication is used and four frame types are defined. As an example the structure of the variable length data frame is shown in Figure 4.30.

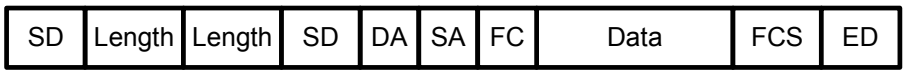


Figure 4.30. *Profibus variable length data frame*

SD is a start delimiter, Length field is for frame length, DA and SA are destination and source addresses, FC is a frame control byte, FCS is a frame checksum and ED is an end delimiter.

4.3.5.7.3. Profibus application and user layer protocols

At the application layer the DP variant is most widely used, providing cyclic and acyclic message transfer between the Master node and its Slaves. In a new protocol version (DPV2) a publisher-subscriber model is implemented for direct Slave-to-Slave communication. Application profiles may be used, for example, for drive or safety applications. In a PA variant device profiles are available allowing easy system integration and node replacement. The FMS variant is based on a VFD concept and FMS services are available. Device profiles are also available in FMS.

DP and FMS variants can share the same physical network infrastructure. The standardized interface between DP and PA variants is defined.

Device description GSD files generally describe device communication features. The EDDL (*Electronic Device Description Language*) is intended for application feature description, which is provided by means of EDD files.

4.3.5.7.4. Profibus summary

– DP, PA and FMS variants, IEC 61158-2, EIA-485 or fiber optic physical layers.

– Master-Slave medium access control, more Masters are possible (bus access control via token-passing).

– Common data link protocol, UART-based communication, up to 12 Mb/s.

– Simple DP application layer with profiles, PA and FMD device profiles.

– GSD and EDD files for device description.

4.3.5.8. *Other standards*

There are some other industrial distributed system standards besides those described above. For example, Modbus, P-Net or WorldFIP are used for industrial automation and LON or KNX are intended for building automation.

Ethernet (or more precisely IEEE 802.3 100BaseTX) and TCP/IP-based protocols are also more often used in an industrial environment. New features (power over Ethernet, rugged design, deterministic protocols) are coming in use. Many industrial systems have Ethernet-based counterparts (e.g. FF H1 – FF HSE, DeviceNet – Ethernet/IP).

4.3.6. *Wireless sensor networks*

Wireless sensor network are becoming more and more popular because of their flexibility and ease of use. On the other hand users meet a new class of problems when implementing wireless solutions, especially the time-variable network throughput (depending on physical network arrangement, ambient environment changes, radio interference level and other factors) and a compromise between device access period and life cycle for battery-powered sensors. There are a number of either standardized or proprietary (e.g. www.dustnetworks.com) wireless sensor solutions. Their communication parameters differ depending on the application area on which they are focused.

4.3.6.1. *IEEE 802.15.4*

IEEE 802.15.4 standard is a physical and data link layer standard for low-speed "personal" networks (WPAN – *Wireless Personal Area Network*). It is also optimized for battery-powered devices, allowing up to several years life without battery exchange.

Three frequency bands are available – 2.4 GHz (16 channels, 250 kb/s, worldwide), 915 MHz (10 channels, 40 kb/s, USA and Australia) and 868 MHz (1 channel, 20 kb/s, Europe). In 2.4 GHz band O-QPSK (Offset-QPSK) modulation is used, BPSK is used for other bands.

Two basic device types are defined. FFD (*Full Function Device*) provides full protocol implementation and can play any role (coordinator or device). RFD (*Reduced Function Device*) may only play a device role – their implementation is simplified and requires minimum resources. RFD cannot communicate with other RFDs directly. Thus, they can only be found at a network border. Two network topologies are possible – the star and the tree, but only tree leaves may be RFD devices. There is always at least one coordinator (PAN coordinator) in the network, in complex networks more coordinators (but a single PAN coordinator) are possible.

The CDMA/CA medium access method is used (see section 4.3.3.3.1), but particular time slots (GTS – *Guaranteed Time Slots*) may be reserved for contention free (device reserved) transmission.

Together four data link layer frames are defined – beacon frame (provides synchronization and control), data frame, acknowledge frame and MAC control frame. The latter allows association to and disassociation from a network, data transmission requests, GTS requests and so on.

Data encryption using AES (*Advanced Encryption Standard*) is possible, as is the limited node access base on ACL (*Access Control List*) and frame integrity checking by adding a MIC (*Message Integrity Code*).

4.3.6.2. *ZigBee*

ZigBee is a network and application layer protocols set residing at the top of IEEE 802.15.4. It was developed by ZigBee Alliance (www.zigbee.org).

At the network layer three device types are distinguished – the ZC (*ZigBee Coordinator*), which is just one in a network, ZR (*ZigBee Router*), which must be a FFD according to IEEE 802.15.4, and ZED (*ZigBee End Device*), which may be IEEE 802.15.4 RFD. The services are provided for node location, network organization and message routing. Star, tree and mesh network topologies are available.

At an application layer the APS (*Application Sublayer*), ZDO (*ZigBee Device Object*) and application framework create a basis for application profiles defined by the standard. User defined profiles (*Network Specific*) are also possible.

ZigBee is not widely used yet. However, it provides an interesting platform for wireless sensor networks.

4.3.6.3. *IEEE 802.15.4 and ZigBee summary*

– 3 ISM bands with 27 channels, bit-rate 250 kb/s, 40kb/s and 20 kb/s.

– CSMA/CA.

– 25 m indoor, 100 m outdoor typical nodes distance.

– Star, tree and mesh topologies, very large node number within a network.

– Network protocol layer allows message routing.

– Application protocols and profiles provide device interoperability.

– Long battery life systems can be built.

4.3.6.4. *Other wireless standards*

There are other wireless standards that can be used to communicate with sensors. The most widely known is Wi-Fi (IEEE 802.11), the wireless computer network with a very high communication throughput (more than 50 Mb/s) for distances up to several kilometers but also the high power consumption. Bluetooth (IEEE 802.14.1) is another PAN standard allowing data transmission up to 100 m with a physical bit-rate up to 1 Mb/s. Battery power is possible for Bluetooth equipped sensors, but the life cycle is much shorter than for ZigBee (actually it depends on transmission duty cycle). In addition, a Z-Wave standard (www.z-wavealliance.org) targets home automation and sensor networks applications.

Chapter 5

Accelerometers and Inclinometers

5.1. Introduction

Accelerometers have existed for several decades and they are always in constant evolution because they influence the performance of many devices in a strategic way. In the past 15 years in particular, thanks to micro technologies, there has been enormous progress in precision, linearity, stability and also size and electric consumption of the sensors.

There is a great diversity of applications of accelerometers in various fields like automotive, aeronautics, instrumentation, medical devices and automation. We will limit this module to the presentation of relevant physical principles and their associated technologies, and we will present some examples of applications.

The term "accelerometer" is in general used for a device which measures linear (not angular) acceleration.

Absolute accelerometer

This device measures the inertial force exerted on the seismic mass. It is attached to the measured object and does not need a reference.

Relative accelerometer

This device measures the distance between the measured object and reference point. The reference point should be stable or moving with constant speed.

Chapter written by André MIGEON and Anne-Elisabeth LENEL.

Acceleration is then given by double differentiation of this distance. Relative accelerometers are mainly used to measure vibrations from a distant stable point (e.g. by laser vibrometers).

This chapter covers absolute accelerometers. Acceleration sensors can be classified according to the physical principle they use:

– a direct measurement of a force (piezoelectric sensor, sensor with force balance); or

– an indirect measurement, by means of displacement or deformation of a sensing element.

We can also classify these sensors by referring to the phenomena they intended to analyze. The useful frequency band of these phenomena then determines the type of suitable sensor taking into account the required precision.

5.2. Acceleration

5.2.1. *Physical quantity*

Acceleration a can be obtained via inertial force F on a mass m subjected to acceleration a of the moving object:

$$F = -ma \tag{5.1}$$

F = inertial force

m = mass

a = acceleration

Table 5.1 presents the different units for acceleration.

Acceleration	USI	meter per square second	m/s^2	Acceleration of a moving object whose velocity varies by 1 meter per second, in 1 second.
	other	gal	Gal	10^{-2} m/s^2 this unit is used in geophysics.
Angular acceleration	USI	radian per square second	rad/s^2	Angular acceleration of a moving object which is rotating around a fixed axis with an angular velocity that varies by 1 radian per second, in 1 second.

Table 5.1. *Acceleration units (from [1])*

A characteristic of accelerometers is that the measured acceleration is directive and is applied to the sensing element through the case of the sensor.

This characteristic has two main consequences:

– Firstly it produces geometrical engineering demands: the mechanical interfaces between the fixing plane and the sensing element of the accelerometer have to be evaluated precisely.

– Secondly, and this is an advantage, it facilitates the adaptation of the accelerometer to a large measuring range thanks to adaptation of the used seismic mass.

The inertial force can be measured either through strain (if deformation is minimum) or through the deformation of elastic element. The principles used in accelerometers are described in Table 5.2.

Secondary Measurand	Types of Accelerometers
force → strain	– piezoelectric accelerometers – piezoresistive accelerometers – resonators
force → displacement	– potentiometric accelerometers – capacitive accelerometers – inductive accelerometers – servo controlled accelerometers – optical accelerometers

Table 5.2. *The different families of accelerometers*

Seismic mass m is suspended by a spring inside the case of the accelerometer.

When the accelerometer is subjected to acceleration a, a relative displacement X of the seismic mass is produced by inertia and detected by a secondary sensor which delivers an electric signal. Figure 5.1 illustrates this principle.

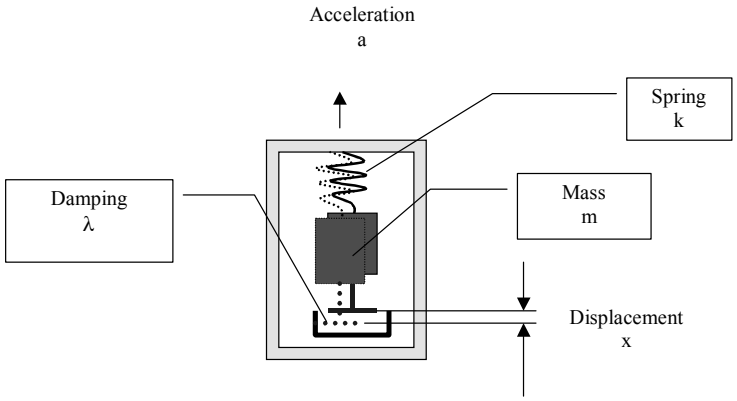


Figure 5.1. *Accelerometer principle*

The equation of the movement is given by equation (5.2)

$$ma = m\left(\frac{d^2x}{dt^2}\right) + \lambda\left(\frac{dx}{dt}\right) + kx \tag{5.2}$$

where:

k = stiffness of spring

λ = damping coefficient

a = acceleration

x = displacement

t = time

m = mass

The damping coefficient of the moving element is due to the mechanical losses in the spring and viscosity of the ambient medium. In a stable state, the relationship between displacement x and acceleration a is:

$$\frac{x}{a} = \frac{m}{k} \tag{5.3}$$

k = the stiffness of the spring

a = acceleration

x = displacement

m = mass

The sensitivity of the accelerometer x/a is proportional to (m/k). The resonance frequency of the system is given by equation (5.4)

$$f_r = \frac{1}{2\pi}\left(\frac{m}{k}\right)^{-\frac{1}{2}} \tag{5.4}$$

k = stiffness of spring

fr = resonance frequency

m = mass

It immediately appears that high sensitivity involves very low resonance frequency. This resonance frequency limits the operational frequency field of the acceleration sensor: therefore, it is necessary to find the best compromise to suit each user. In addition, the consideration of its value leads naturally to the above mentioned classification of the various accelerometers.

The curve below (Figure 5.2) illustrates this phenomenon and shows the importance of damping coefficient λ.

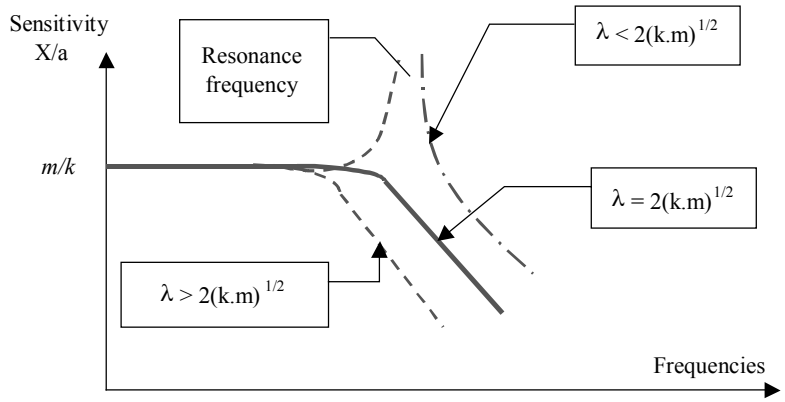


Figure 5.2. *Influence of frequency on sensitivity (from [2])*

The condition to obtain optimal frequency response and avoid deterioration of the accelerometer when resonance occurs is given by equation (5.5)

$$\lambda \geq 2\sqrt{km} \tag{5.5}$$

where:

k = stiffness of spring

m = mass

λ = damping coefficient

This principle is applied to all accelerometers.

5.2.2. *Application to velocity and position measurements*

Accelerometers (generally linked to gyrometers) are frequently used to determine the speed and the position of various vehicles (planes, ships, cars, robots, etc.). In these configurations we refer to inertial sensors or inertial multisensors.

Inertial navigation systems (INS) integrate several inertial sensors of the same or different natures.

Gimbaled navigation systems

Systems with a platform or UJ (universal joint) have a core controlled by the local geographical trihedral (Northern, Is and vertical) and sometimes an inertial reference frame. In this case, the platform preserves a fixed orientation compared to a reference frame related to the stars (terrestrial, inertial or Galilean) thanks to the universal joints.

Gyroscopes are then used as zero detectors and the carrier attitude is given using the angles measured at the levels of the UJ axes between the inertial core and the carrier.

Strapdown navigational system

Systems with strapped or linked components appear as a result of the progress made:

– on the one hand, by gyroscopes with very stable scale factor;

– and on the other in electronics.

They correspond to devices (without platforms) directly fixed on the carrier. This became possible with the arrival of the gyrolaser, and processors with sufficient speed and power, and great memory capacity.

The calculator permanently maintains a virtual inertial trihedral using gyrometric information. Accelerometric information is then projected into this virtual trihedral to determine linear velocities and displacements of the carrier.

An inertial measuring unit (IMU) gathers data from the inertial sensors and performs the calculations necessary to supply information about angular rotation and linear acceleration. Generally, this is in the form of angle increments and/or linear velocity increments.

Intuitively, by the integration of angular velocities measured by gyrometers it is possible to determine the attitude of a moving object knowing its initial attitude. The attitude of the moving object can in fact be represented by a matrix of passage from a reference frame related to the moving object to an inertial reference frame.

5.2.3. *Application to position measurements*

An inertial central is a system that includes an IMU integrating:

– three channel orthogonal gyrometric measurements;

– three channel orthogonal accelerometric measurements;

– and a means of calculation to determine at all times the attitude and position of the moving object.

The position of the moving object is characterized by:

– longitude (L);

– latitude (g);

– and altitude (Z).

The position can be obtained directly by velocity integration.

The disadvantage of this method is the inability to ensure correctly the passage of the pole. In this case, we use the attitude of an intermediate platform compared to the terrestrial reference frame. Figure 5.3 shows the principle of an inertial central.

Three gyrometers and three accelerometers collect the data of the angular velocity and the acceleration of the moving object. By successive integration we

have access to the complete data of the position of the moving object (x,y,z). Finally, by comparison with the original position, it is possible to determine attitude, speed and position of the moving object.

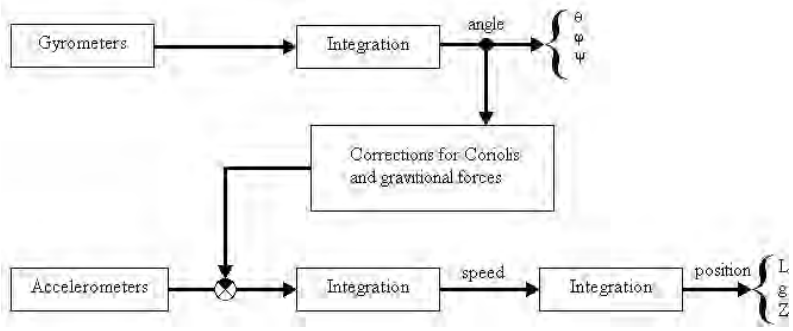


Figure 5.3. *Principle of an inertial central from M2A technologies*

5.2.4. *The inclinometers*

Inclinometry corresponds to the measurement of an angle which determines the position of an object in three dimensions. Inclinometry consists of determining angles of rotation α β γ of the trihedral O' X' Y' Z' related to the object compared to the trihedral OXYZ of the reference related to the ground. The absolute direction of reference (vertical) is given by the direction of the field of terrestrial gravity at the point of measurement (see Figure 5.4).

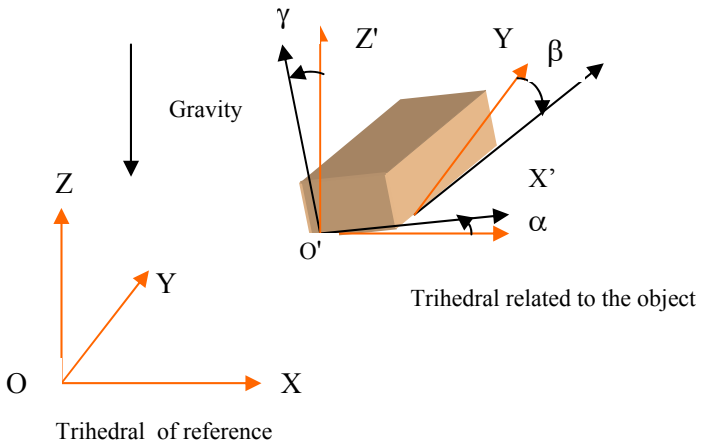


Figure 5.4. *Inclination*

It is possible to measure a relative inclination angle by choosing its own reference direction, in this case we refer to angular measurement but no longer to inclination.

The inclination measurement of a very bulky object (more than several meters high) is generally achieved by level measurement.

Inclinometers differ from accelerometers by:

– their limited measuring range (< ± 1 g);

– their high sensitivity;

– and their low bandwidth.

Consequently, many accelerometers can be used as inclinometers.

A precision inclinometer comprises an inertial mass around a torsion bar and controlled in position by an inductive torque actuator which current is proportional to the applied acceleration. The detector of position can be capacitive, optical or different.

For small angles, measured acceleration is comparable to the angle of inclination. For wide measuring ranges, an arc-sinus conversion is necessary to determine the inclination.

5.3. Application ranges

Depending on acceleration levels and frequency ranges, it is necessary to distinguish the various application ranges. Some typical application examples are given in the tables below.

5.3.1. *Static and low-frequency acceleration*

This application group ranges from DC to 50 Hz in frequency and from 0 to approx. 10 g in amplitude. High precision is usually required.

FIELDS	APPLICATIONS
POSITION CONTROL	Military machines, cars, railways, stabilization of platforms
INCLINOMETRY	On-board instrumentation: aeronautics, vehicles, building machines, tools
INSTRUMENTATION	Test benches, vehicle tests, swell studies, process monitoring, transport
METROLOGY/CALIBRATION	Accelerometer calibration benches: centrifugal machines, control accessories on various measuring benches
NAVIGATION, GUIDANCE, PILOTING	Planes, boats, military and civil terrestrial vehicles, robots
SEISMICS	Vulcanology, oil research (geophones), monitoring for houses and road building
SECURITY SYSTEMS	Automobile, armament, nuclear power, home automation, aeronautics
MEDICAL	Measurement of tremor

Table 5.3. *Application of DC accelerometers*

5.3.2. *Vibrations*

Vibration frequencies range from approx. 7 Hz to 10 kHz, with amplitudes up to 100 g. The sensors need not measure DC acceleration.

FIELDS	APPLICATIONS
MODAL ANALYSIS	Research and Design: automotive, aeronautics, electric household appliances, house and road construction
CAR INDUSTRY	Engines, noise
METROLOGY	Calibration benches, test benches
STRUCTURE MONITORING	Industrial machines, tool monitoring, damping system monitoring, house construction
TRANSPORT	Railways, aeronautical, military

Table 5.4. *Applications of vibration sensors*

5.3.3. *Shocks*

Sensors for measuring mechanical shocks should have a frequency range from 500 Hz to 100 kHz and a full-scale range up to 100,000 g.

FIELDS	APPLICATIONS
CAR INDUSTRY	Crash tests, airbags, engines
INSTRUMENTATION	Packaging, transported products
METROLOGY	Calibration and test benches
MILITARY	Various release systems (alarms, firing, protections, etc.)
SAFETY AND MONITORING SYSTEMS	Transport, cargo monitoring, structure and fatigue monitoring, house automation

Table 5.5. *Applications of shock sensors*

5.3.4. *Inclination*

Inclination is in fact the measurement of components of the gravitational acceleration. The required range is therefore ± 1 g, and sensors should measure from DC.

FIELDS	APPLICATIONS
BUILDING AND ROADS	Monitoring of structures (buildings, bridges), machines and crane safety
MILITARY	Gun mounting control, various vehicles, roll and pitch control
PETROLEUM DRILLING	Drill rig monitoring, drill path control
INSTRUMENTATION	Tools, machine tool monitoring, control processing
CAR INDUSTRY	Frame stabilization, robotics

Table 5.6. *Application of inclination sensors*

Table 5.7 summarizes the various fields of application and best-suited types of sensors.

	USES	TYPES OF ACCELEROMETERS
1	– Acceleration of moving object having a certain mass such as aircraft, missiles, terrestrial or maritime vehicles – Frequencies (0-50 Hz) – Low acceleration	– Servo controlled accelerometers – Accelerometers with measurement of displacement (inductive, capacitive, with potentiometers, optics) – Strain gauge accelerometers – Accelerometers with contact or threshold, although generally of average precision, are included in this first sensor category intended for measurement of moving object center of gravity movement
2	– Vibratory acceleration – For rigid structures or significant masses – Around 100 Hz – Measuring a continuous or pseudo-continuous acceleration with a satisfactory damping	– Accelerometers with variable inductance – Metallic or generally piezoresistive strain gauges
3	– Vibratory acceleration – Around 10,000 Hz	– Piezoresistive or piezoelectric sensors
4	– Measurement of shocks – Pulsated accelerations – Up to 100,000 Hz	– Sensors with a bandwidth extending from low to high frequencies

Table 5.7. *Different uses and required accelerometers*

5.4. Main models of accelerometers

This part explains the different conversion principles used by the main accelerometers.

Table 5.8 classifies the different principles according to the frequency range to be measured.

Principles of detection	Recommended range of frequencies (Hz)						
	0,1	**1**	**10**	**100**	**1,000**	**10,000**	**100,000**
With Foucault current							
With resonator							
Servo controlled (electrodynamic)							
Electromagnetic							
Electrostatic							
Optical							
Piezoelectric (quartz or ceramics)							
Piezotransistor							
Capacitive bridge							
Bridge of piezoresistive gauges							
Bridge of resistive gauges							

Table 5.8. *Classification of the different principles of accelerometers*

5.4.1. *Piezoelectric accelerometers*

Piezoelectricity is defined as the electric polarization of certain crystals caused by a mechanical strain. The piezoelectric materials are sensitive to compressive linear stress and shear.

The piezoelectric materials can be divided into two categories: crystals and artificially polarized ferroelectric ceramics containing barium titanate and lead zirconate. The choice of material depends on the working environment and the measurement to be carried out.

Natural crystals

The most widely used crystals are quartz and tourmaline, and sometimes Bismuth Germanium Oxidizes (BGO). Quartz can be found in its natural state but it is preferable to produce it artificially for accelerometers. On the other hand, tourmaline is used in its natural form.

These crystals are perfect for measuring low frequencies, and for environments with temperature fluctuations.

Ferroelectric crystals

These are polycrystalline dielectric materials known as piezoceramics. These materials require the application of a continuous electric field to be polarized. The most widely used are lead zirconate titanate and bismuth titanate. They have high sensitivity and can be used in a wider frequency band than natural crystals.

A comparison of some piezoelectric materials is detailed in Table 5.9.

	Materials	**Piezoelectric coefficients** 10^{-12} **C/N**	**Maximum temperature** **°C**	**Resonance frequency** **at 50 pC/g in Hz**
Natural crystals	Quartz	2.2	250	7,000
	Tourmaline	1.8	600	7,000
	BGO	22	350	8,000
Ferroelectric crystals	Bism. titanate	20	500	15,000
	Zirconate Pb	280	260	25,000
Conclusion: natural crystals are less sensitive to temperature variations but have a lower piezoelectric coefficient.				

Table 5.9. *Piezoelectric coefficients for various materials (from [2])*

5.4.1.1. *General principle*

Figure 5.5 shows the general principle of a piezoelectric accelerometer.

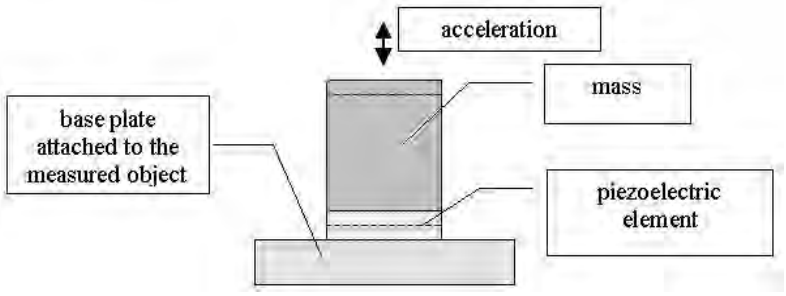


Figure 5.5. *General principle of piezoelectric accelerometers*

In these types of accelerometer devices, the piezoelectric element is placed in such way that when the unit is in vibration, a mass applies a force proportional to acceleration to the piezoelectric element. There are two types of assembly:

– the type with linear stress (compression);

– the type with shear.

5.4.1.2. *Accelerometers with compression*

Figure 5.6 shows a cross-section of a typical accelerometer with spring. It is often necessary to prestress the piezoelectric element in order to extend the range for both directions of acceleration.

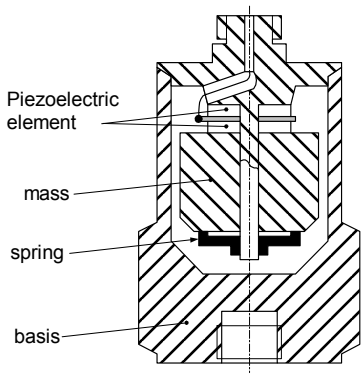


Figure 5.6. *Piezoelectric accelerometer with axial compression*

5.4.1.3. *Shear-mode accelerometers*

These accelerometers consist of annular piezoelectric material (or a stack of small plates) – Figure 5.7. The advantage of the shear-mode accelerometers is better resistance to parasitic influences such as temperature changes.

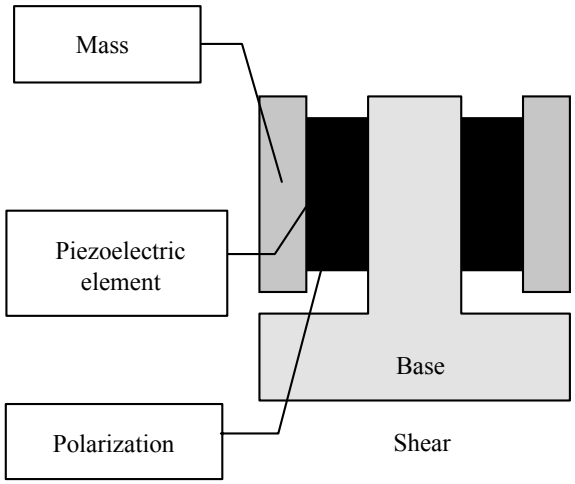


Figure 5.7. *Principle of the shear-mode accelerometer*

5.4.1.4. *Features and limits of these accelerometers*

The accelerometer must be assembled with its principal axis of sensitivity aligned with the required direction of measurement. The mounting surface must be rigid without flexible elements or shock absorbers. Small differences in the place of assembly can incur great errors.

There are 3 principal types of assembly:

– Fixing with screws.

– Fixing by magnet.

– Adhesive fixing.

The best assembly is to use electrically insulating screws to avoid ground loops.

Influence of the environment

1. *High and low temperatures*

The piezoelectric accelerometer is able to measure vibrations over a wide thermal range.

High temperature: at 250°C, piezoelectric ceramics start to depolarize and then their sensitivity is permanently changed. Therefore, for higher temperatures we have to use special accelerometers.

Low temperature: the low limit temperature of most accelerometers is specified as –74°C.

2. *Thermal fluctuations*

Even low variations of ambient temperature generate a low frequency noise signal. We can reduce this influence by using a shearing accelerometer or by protecting it with a light thermal insulator.

3. *Moisture*

The accelerometers are hermetically sealed inside a welded case to protect them from moisture.

4. *Noise due to connection cables*

Errors in vibration measurements are often due to bad assembly and to the layout of the cable connecting the accelerometer to the amplifier. This induces noise signals in the connection cable. These disturbances come mainly from 3 sources:

– triboelectric noise;

– electromagnetic noise;

– ground loops.

Table 5.10 below indicates various cases and applied solutions.

Influence of connection cables		
Noise	**Explanation of mechanism**	**Solutions**
Triboelectric noise (a particularly disturbing effect during the measurement of low vibratory levels)	Noise is induced in the accelerometer cable by mechanical movement of the cable itself. When a coaxial cable is subjected to flexion, compression or tension, the shielding inside the cable is separated temporarily from dielectric material at certain points; local capacity variations then appear.	– Use a special coaxial cable minimizing the noise – Fix or stick the cable as close as possible to the accelerometer
Electromagnetic noise	Noise is induced in cables located near: – rotating machines (with turning systems) – magnetic fields or intense radioelectric fields.	– Using double shielded cable – Routing cable far from the electromagnetic sources – In more serious cases, it is necessary to use a bipolar accelerometer and a differential preamplifier.
Ground loops	Usually, earth connection of the measuring instrument case is made through the machine which is grounded. However, in some industrial cases, the machine is not grounded.	Solutions to “break” the ground loop: – insulating the accelerometer electrically, – using reliable connections – choosing a charge preamplifier equipped with a floating input.

Table 5.10. *Influence of connection cables on shearing accelerometer performance*

The piezoelectric accelerometer has capacitive impedance, and generally it cannot be connected to the circuit having resistive input impedance. Indeed the discharge of the capacity would be too fast to allow the exploitation of the signal

and the tension collected would be sensitive to the erratic variations of the sum of the capacitances of the assembly and in particular of the connecting cables.

The device to be used in this case is the charge amplifier which delivers a voltage proportional to the charge and independent of the capacitance of the sensor and connecting cables (Figure 5.8).

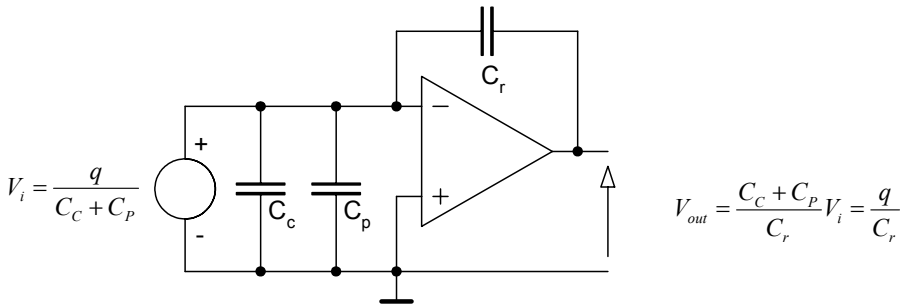


Figure 5.8. *Connection with a piezoelectric accelerometer*

The connecting cable quality associated with the charge amplifier strongly influences the obtained results. In general, sensor suppliers offer cables and signal conditioners suited for their products.

The main advantages and disadvantages of piezoelectric accelerometers are summarized in Table 5.11.

Piezoelectric Accelerometers	
Advantages	**Disadvantages**
– robust – compact – high reliability – generally very light – very large bandwidth (from a few Hz to several tens of kHz)	– operates only in dynamic mode (cannot measure constant acceleration) – detector output signal is high impedance, hence the need for a specific connection between the detector and the electronics signal processing which increases the measurement cost – high sensitivity to temperature

Table 5.11. *Main advantages and disadvantages of piezoelectric accelerometers*

5.4.2. *Piezoresistive accelerometers*

5.4.2.1. *General principle*

A seismic mass is placed on an elastic return blade equipped with two or four piezoresistive gauges in a Wheatstone Bridge. The blade flexion is translated into gauged deformation. These gauges enable conversion of the acceleration into an electric quantity, since the received signal is proportional to the acceleration of the moving object (see Figure 5.9 below).

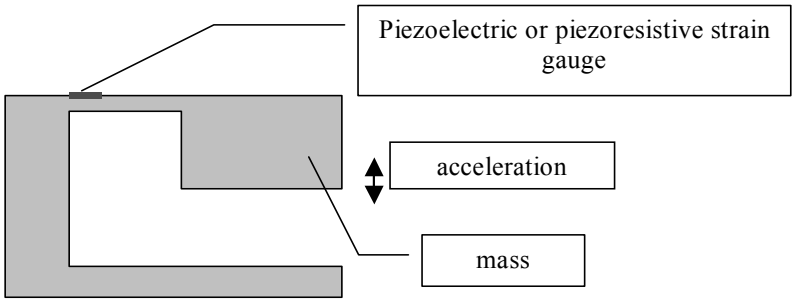


Figure 5.9. *Principle of piezoresistive accelerometers*

5.4.2.2. *Silicon semiconductor strain gauges*

The resistivity variation depends on material, resistivity, doping level, type of doping agent and the crystallographic direction in which the material is machined, and the resistivity itself is given by the concentration of the doping agent. The gauge factor of silicon varies as follows:

– [+100 to + 175] for type P;

– [–100 to –140] for type N.

The important parameters are:

– gauge factor K;

– temperature coefficient of resistance;

– temperature coefficient of gauge factor.

Gauge factor K is initially determined by doping level but also depends on temperature.

Figure 5.10 shows that the gauge factor and temperature coefficients are inversely proportional to the level of doping.

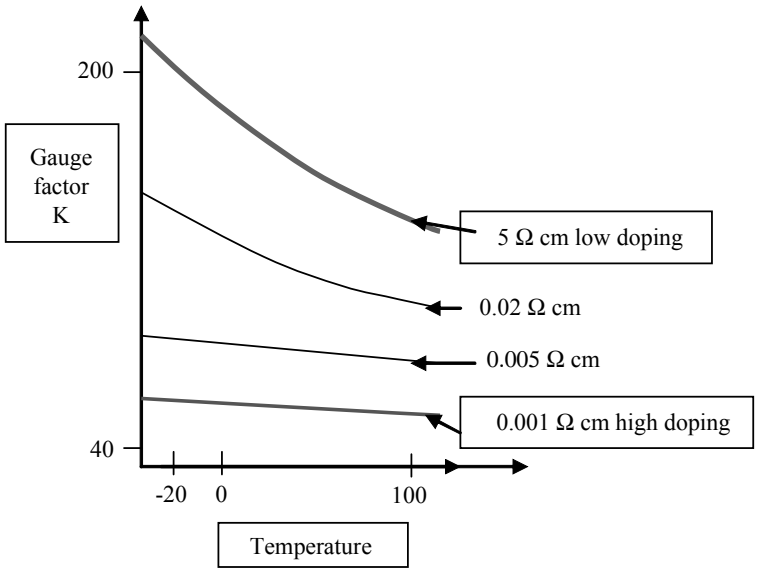


Figure 5.10. *Effects of the doping level and temperature on silicon type P (after [2])*

Assembly of the gauges

- Flat Gauges

These gauges have 2 relatively broad mounting plates, joined together by a narrow central element (Figure 5.11). In this configuration, strains are concentrated in a miniature element, the surface of which is polished, free from any potential unwanted strain.

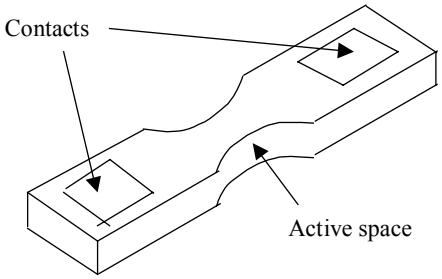


Figure 5.11. *Diagram of a flat gauge*

Thanks to broad contact surfaces at the ends, the strain induced by fixing will be kept to a small fraction of the useful strain at the throttling level. These strain gauges are made of a single silicon crystal with a high degree of purity.

The silicon doped with phosphorus gives a negative gauge factor, while doping with boron gives a positive gauge factor.

- Carved gauges

We can use the "notched gauge" principle by development of chemical etching. Figure 5.12 shows a carved monolithic element: the notches releasing the gauges and the seismic masses are the un-etched parts of silicon.

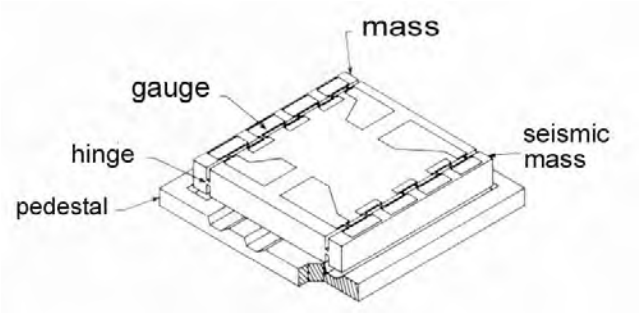


Figure 5.12. *Monolithic sensitive element for accelerometer 7270*

The linearity and the sensitivity are optimized because of:

– the monolithic structure;

– the extremely small size to ensure a very high force/weight ratio;

– the freedom of the gauges.

The resonance at several MHz and the linear range of more than 100,000 g exceeds the performance of former sensors. These sensors are particularly stable because there are no adhesive joints between the mass, the gauges and the substrate.

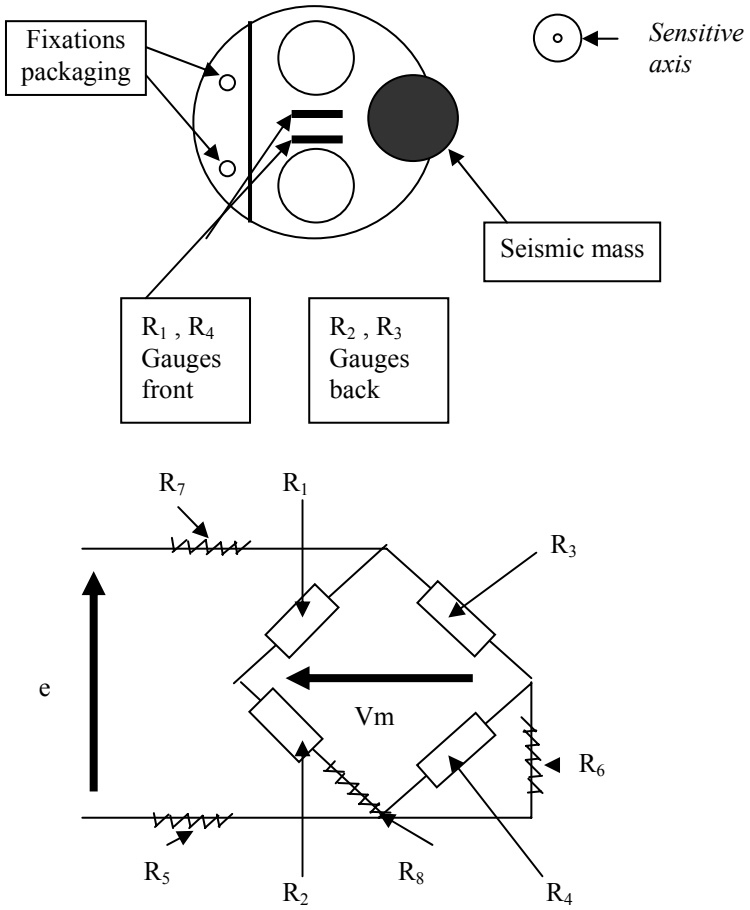


Figure 5.13. *Piezoresistive accelerometer for measurement of low frequencies*

Design examples (Figure 5.13)

a) Details of the creation of the sensor blade.

b) Wheatstone Bridge: G1 to G4: gauges; R5, R6, R7, R8 compensation of the thermal drift of sensitivity.

The decoupling cavities are created to direct inertias towards the exterior blade in order to decrease the influence of transverse parasitic accelerations. The blades are made of stainless steel or copper beryllium alloys. The seismic mass is made of steel or an alloy of sintered tungsten. The damping is carried out by blade displacement in a box filled with silicone oil.

5.4.2.3. *Features and limits of these accelerometers*

5.4.2.3.1. *Sensitivity, frequency response*

The sensitivity is defined by equation (5.6):

$$S = \frac{m}{a} = \frac{V_m}{\varepsilon}\frac{\varepsilon}{a} = S_1 * S_2 \tag{5.6}$$

V_m = output voltage of the Wheatstone Bridge

ε = deformation

a = acceleration

S_2 is the electric sensitivity of the Wheatstone Bridge formed by the 4 gauges.

$$S_2 = \frac{V_m}{\varepsilon} \tag{5.7}$$

S_1 characterizes the response of the mechanical part of the accelerometer:

$$S_1 = \frac{\varepsilon}{a} \tag{5.8}$$

The sensitivity varies from 1 to 25 mV/g according to the gauge.

5.4.2.3.2. *Bandwidth*

The accelerometers for measurement at continuous and low frequencies have critical damping $\tau = 1$. The useful bandwidth extends from 0 to ¼ of the resonance frequency.

Piezoresistive accelerometers have a factor of merit defined by equation (5.9)

$$\beta = S.f_0^2 \tag{5.9}$$

f_0 = natural frequency

S = sensitivity

β = factor of merit

For a given construction and technology, we cannot have, at the same time, an accelerometer having high sensitivity and wide bandwidth. Moreover, the higher the damping of the accelerometer is, the bigger the subsequent phase shift becomes. The typical frequency response range is from 0 to 3,000 Hz.

5.4.2.3.3. *Influence of temperature*

This can take 3 different forms:

– Influence on zero.

– Influence on sensitivity.

Due to thermal variations of the Young's modulus and the gauge coefficient, the sensitivity decreases as the temperature increases. The order of magnitude is 1 to 2.10^{-4} of nominal sensitivity/°C. The resulting error can reach a small percentage of the value at 20°C but this error is stable. It can thus be partially corrected by calculation.

– Variation of the damping coefficient: The piezoresistive accelerometers are damped using silicon oil. When temperature increases, the kinematic viscosity decreases and then the damping coefficient also decreases.

The achievable working temperature is –50 to 150°C. Above 150°C doping is no longer effective.

5.4.2.3.4. *Technological limitations: connecting cable*

Connecting cables bring a deterioration of the transmitted signal at the entry to the signal conditioner, which is related to its length and its frequency. In the case of a significant length of cable, the accelerometer with constant current instead of constant voltage can be supplied in order to eliminate the influence of the resistance of the cable.

The accelerometer requires a very stable source of power.

5.4.2.3.5. *Technological limitations: shocks and vibrations*

Accelerometers are sensitive to shocks and vibrations. For inertial navigation it is necessary to equip them with a mechanical filter eliminating high frequencies and to place them in an enclosure protected from shocks.

The main advantages and disadvantages of piezoresistive accelerometers are summarized in Table 5.12.

Piezoresistive Accelerometers	
Advantages	**Disadvantages**
– high sensitivity	– no significant linearity
– low cost	– high sensitivity to temperature
– quite high bandwidth	– generally average performance
– possibility of obtaining a very high natural frequency (> 30 KHz)	– the lower the sensitivity, the higher the bandwidth
– simple data processing	
– possible miniaturization	

Table 5.12. *Main advantages and disadvantages of piezoresistive accelerometers*

5.4.3. *Accelerometers with resonators*

5.4.3.1. *Principle*

A force-frequency transducer comprises a quartz beam vibrating at its natural frequency of flexion (see Figure 5.14). Under the influence of acceleration a, the seismic mass puts force on the beam and modifies its frequency of vibration ν_0 which becomes ν_1 (equation (5.10)):

$$\nu_1 = \nu_0 + ka + f(a) \tag{5.10}$$

a = acceleration

ν_0, ν_1 = frequencies of vibration

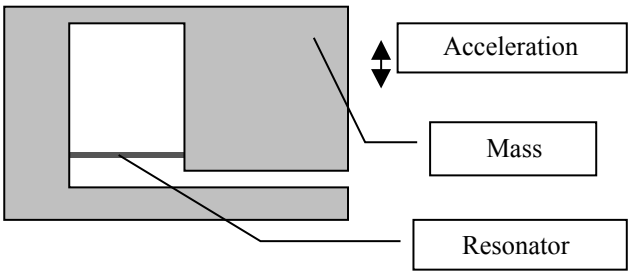


Figure 5.14. *Principle of Accelerometers with resonators*

This variation of frequency makes it possible to measure acceleration with a very high accuracy. Practically, the majority of the devices comprise two transducers in which acceleration imposes on the one hand a compressive stress and on the other hand a tensile stress. The difference between frequencies ν_1-ν_2 is approximately linear. To prevent that quartz beams embedding degrades the characteristics of the resonator, we sometimes use complicated beam shapes.

An example of a resonant accelerometer made of a monoblock of miniature quartz is shown in Figure 5.15.

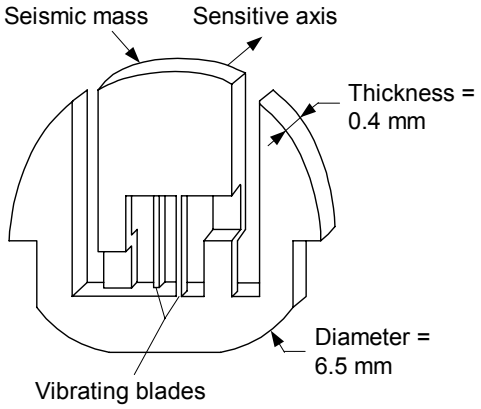


Figure 5.15. *Vibrating quartz accelerometer*

5.4.3.2. *Features and limits of these accelerometers*

The main advantages and disadvantages of such accelerometers are summarized in Table 5.13.

Accelerometers with Resonators	
Advantages	**Disadvantages**
– relatively simple construction	– low resistance to shocks
– very small energy consumption (approximately 1.5 10^{-8} W for the two oscillators associated with the transducers)	– generally non-linear output signal, hence the need for data processing
– excellent precision: they compete with the best accelerometers with servo controlled displacement	– bandwidth very dependent on the data processing
– excellent stability	– sensitivity to low frequency vibrations (< 2,000 Hz)
– high linearity and stability of the scale factor (up to 10^{-6})	
– the dynamic ranges of measurements extend according to the situation, from some tens of g to several thousands of g	

Table 5.13. *Main advantages and disadvantages of accelerometers with resonators*

5.4.4. *Capacitive accelerometers*

5.4.4.1. *Principle*

A mass, suspended at its ends by two membranes or springs, moves according to acceleration. One electrode of a capacitor is connected to this mass while the other is fixed. When the accelerometer is subjected to acceleration, the distance between the two electrodes changes and thus the capacitance changes. The capacitance is measured by embedded electronic circuits.

The general principle is explained by Figure 5.16.

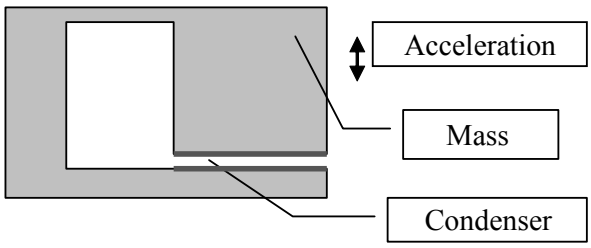


Figure 5.16. *Principle of a capacitive accelerometer*

There are several types of capacitive accelerometers:

– *the simple capacitor:* one plate of the capacitor is fixed, while the other is connected to the seismic mass. The acceleration of the system varies the distance between the two electrodes and thus the capacitance;

– *the differential double capacitor:* a plate is placed between two fixed plates creating two capacitors. The origin of the moving plate is placed symmetrically between the fixed plates in order to have two capacities of the same value at rest position. The acceleration of the system makes their capacitances variable in inverse direction according to the displacement of the moving plate.

– *others:* there are other capacitive accelerometers with reduced size, the most popular is the comb type. They work on the same principle but the condenser is a micro-machined comb.

EXAMPLES:

1. *Pendular capacitive detection accelerometer*

A seismic mass in the shape of plate or pendulum is fixed to a frame via hinges acting as a recall spring. Under the action of an acceleration imposed along the sensitive axis, the pendulum turns at an angle α (see Figure 5.17).

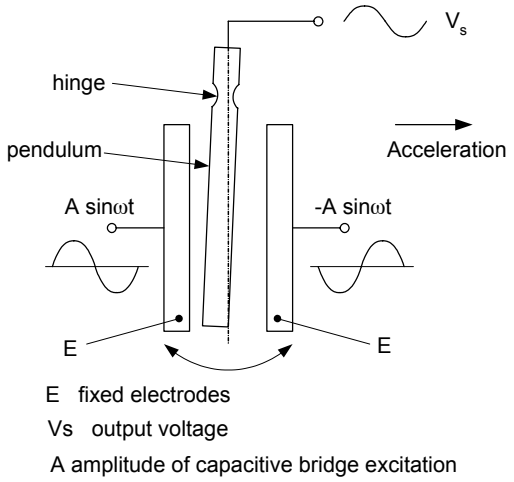


Figure 5.17. *Principle of pendular capacitive detection accelerometer*

Angle α follows the equation

$$\alpha = \frac{C}{\rho} \tag{5.11}$$

C = inertial origin torque around the axis hinge

ρ = angular stiffness of the pendulum.

α = angle

Angle α has a linear dependence on acceleration. It is measured by means of a detector with a capacitive bridge. Two equal alternating voltages with an opposite phase are applied to the fixed electrodes. The residual voltage, collected on the pendular electrode, quantifies the imbalance of the bridge.

2. *Ultra-sensitive Accelerometers*

The principle lies in the direct measurement of the electrostatic force necessary to maintain a mass (the sensing element) motionless in the center of the cage. This accelerometer can be used only in the open space, where the gravitational force is not present.

5.4.4.2. *Features and limits of these accelerometers*

The main advantages and disadvantages of such accelerometers are summarized in Table 5.14.

Capacitive accelerometers	
Advantages	**Disadvantages**
– high resolution	– quite low linearity
– high sensitivity	– high sensitivity to temperature ,hence the need for significant compensation
– mechanically simple detector	
– possible miniaturization	
– technological ruggedness	
– quite high bandwidth (several kHz)	

Table 5.14. *Main advantages and disadvantages of capacitive accelerometers*

5.4.5. *Potentiometric accelerometers*

5.4.5.1. *Principle*

The seismic mass suspended by two flat springs is connected directly, or through an amplifying system, to the slider of the potentiometer. A voltage proportional to acceleration can then be measured (see Figure 5.18).

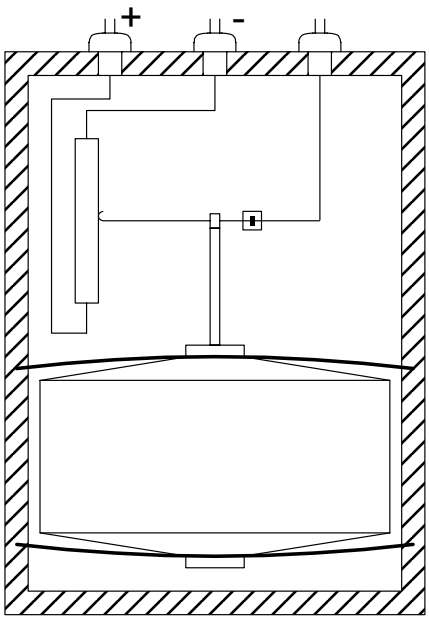


Figure 5.18. *Diagram of a potentiometric accelerometer*

5.4.5.2. *Features and limits of these accelerometers*

The friction between the slider and a potentiometer track causes important mobility errors and hysteresis errors. The choice of the potentiometer strongly influences the linearity of the accelerometer. This type of accelerometer is rather obsolete. Its main advantages and disadvantages are summarized in Table 5.15.

Potentiometric accelerometers	
Advantages	**Disadvantages**
– low cost	– high hysteresis error
– technological ruggedness	– linearity determined by the potentiometer
	– low reliability
	– relatively low performance

Table 5.15. *Advantages and disadvantages of potentiometric accelerometers*

5.4.6. *Optical detection accelerometers*

5.4.6.1. *Principle*

The development of microresonator fiber optic sensors requires the joint efforts of several branches of science and technology: microelectronics, radioelectronics and fiber optics.

Fiber optic sensors, in which the micromechanical resonator acts as a sensitive element, have been proposed for measurement of acceleration. The displacement of the seismic mass under the influence of the inertial force causes a variation of the coefficients of transmission of light generated by a laser diode and transported by an optical fiber. The quantity of transmitted light is measured by a photodetector and converted into voltage or current.

An all-optical approach using optical fibers for light transmission offers electrical passivity, while the output of the sensor, being a frequency, is transmission-line-independent. The use of metallic glasses as microresonator material opens new possibilities for detecting outer actions/forces through changes of the magnetic field.

The materials for microresonators can be:

– boron-doped silicon;

– silicon dioxide;

– silicon nitride;

– metallic glass.

Most frequently the microresonator is a silicon microbridge clamped at both ends.

Silicon microresonators can be bonded directly to the end of an optical fiber leading to a low-cost extrinsic sensor capable of providing the precision measurements based on frequency readout.

5.4.6.2. *Features and limits of these accelerometers*

Microresonator sensors being electrically passive can operate remotely in the presence of strong electromagnetic interference, hostile environments, explosives, while itself the passive sensors can not be detected electrically. Using the frequency as the output parameter presents two major advantages: it can be transmitted through extended systems and over large distance without any error and, secondly, it can be easily digitized by frequency counter.

The main advantages and disadvantages of such accelerometers are summarized in Table 5.16

Optical detection accelerometers	
Advantages	**Disadvantages**
– frequency coding of output signal – electrical passivity – light-weight – high shock durability – low mechanical hysteresis	– complex and expensive associated data processing – high sensitivity to temperature – average performance

Table 5.16. *Main advantages and disadvantages of optical detection accelerometers*

5.4.7. *Magnetic detection Accelerometers*

5.4.7.1. *Principle*

The displacement of the seismic mass, by the inertial force effect, causes a variation of the coupling coefficients of the magnetic circuit. Figure 5.19 shows the principle of an inductive accelerometer.

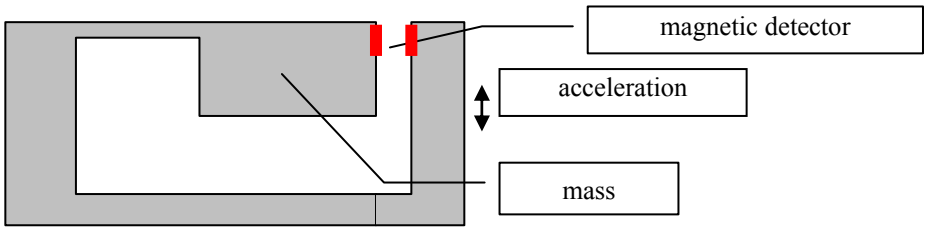


Figure 5.19. *Principle of an inductive accelerometer*

5.4.7.2. *Features and limits of these accelerometers*

Accelerometers using this type of detector are still very widespread. The variation of inductance is exploited to detect weak displacements of several hundredths of a millimeter.

The main advantages and disadvantages of such accelerometers are summarized in Table 5.17.

Magnetic Detection Accelerometers	
Advantages	**Disadvantages**
– possibility of measuring continuous and alternative accelerations	– need for AC excitation, thus imposing a need to compensate for parasitic capacitances
– excellent resolution	– low natural frequency limited by the relatively significant moving mass
– excellent sensitivity	– relatively low bandwidth (a few tens of Hz) limited by the excitation frequency
– allows a high level output signal	– obligatory demodulation for measurement
– low hysteresis if good quality sensing element	– high sensitivity to vibrations
– non-contact detection, and therefore no frictions	– sensitivity to ambient parasitic magnetic fields
	– need for an electronic linearity correction
	– sensitivity dependent on used materials
	– relatively complex measuring equipment
	– relatively expensive accelerometers

Table 5.17. *Main advantages and disadvantages of magnetic detection accelerometers*

5.4.8. *Servo accelerometers with controlled displacement*

5.4.8.1. *Principle*

In these devices, the inertial force applied to the seismic mass and causing its movement is compensated for by an equal and opposed force created by an electromechanical system. The driving current is then proportional to the measured force.

More precisely the inertial action on the seismic mass initially causes its micro displacement which at once starts a reaction of the torque motor, bringing it back to its initial position.

The gain from the feedback binding the detector to the actuator is such that the seismic mass displacements are extremely small, thus reducing the errors due to position variations like hysteresis.

The devices built are very varied. Indeed, the position detector can be an optical, capacitive or inductive sensor. In the same way the actuator system can be either electrostatic or electromagnetic with a permanent magnet. The control can be analog or digital. Accelerometers such as these have biaxial or triaxial versions.

5.4.8.2. *Servo accelerometers with balance of torque*

A very light square coil (frame) is suspended in a magnetic induction field between two pivots presenting the minimum of frictions (pivot or ribbon). The frame is provided on one side with a mass M (the sensing element) – see Figure 5.20. Under the influence of the inertial force due to acceleration, the mass tends to move, and it causes the rotation of the frame.

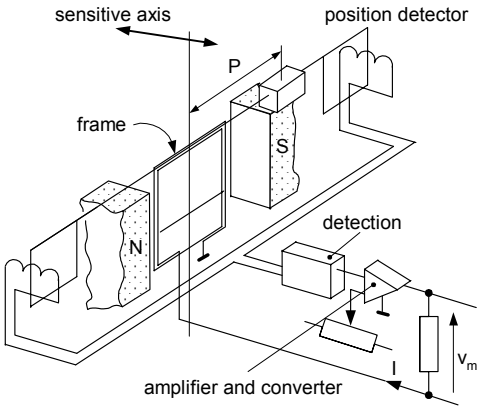


Figure 5.20. *Accelerometer with balance of torque*

Table 5.18 explains the principle of inductive and optical detection of servo controlled accelerometers with balance of torque.

<table>
<tr><th colspan="2">SERVO CONTROLLED ACCELEROMETER
WITH BALANCE OF TORQUE</th></tr>
<tr><td colspan="2">The frame is equipped with two perfectly balanced moving blades.</td></tr>
<tr><th>INDUCTIVE DETECTION</th><th>OPTICAL DETECTION</th></tr>
<tr><td>– The blades move in front of 2 detection coils supplied with an alternating voltage at high frequency</td><td>– The blades move in front of an optical system of detection</td></tr>
<tr><td>– The moving blades vary the inductance of the coils and the voltage on their terminals. The voltage is then rectified by a diode and compared with fixed reference voltage. The variation of the voltage is amplified and converted into a current I</td><td>– The blade changes the amount of light for the optical detector that generates a voltage. It is amplified and converted into a current I</td></tr>
<tr><td colspan="2">In both cases, the current traverses the frame coil and creates a reaction torque which restores the moving element to its initial position</td></tr>
</table>

Table 5.18. *Inductive and optical detection servo controlled accelerometer with balance of torque*

5.4.8.3. *Servo accelerometers with balance of force*

This consists of a pendular system in which the seismic mass is the sensing element.

During acceleration along the measurement axis, the position of the mass is detected by a sensor the output signal of which is amplified, in order to feed a system of recall of the mass to its initial position. This initial position corresponds in general to the zero of the accelerometer: the position without any acceleration or disturbing mechanical tension. This is a servo controlled system with almost no displacement, i.e. a system with great stiffness and high natural frequency. The position detector can be either inductive or capacitive.

In the case of capacitive detection, the seismic mass constitutes the moving plate of the capacitor.

5.4.8.4. *Features and limits of these accelerometers*

The main advantages and disadvantages of such accelerometers are summarized in Table 5.19.

SERVO ACCELEROMETERS WITH CONTROLLED DISPLACEMENT	
Advantages	**Disadvantages**
– excellent precision (+/–0.01% or better) generally much better than not servo controlled sensors	– sensitivity to shocks
– bandwidth: from DC to a few hundred Hz	– high cost
– threshold of detection: $< 10^{-5}$ m.s^{-2}	– need for significant data processing
	– relatively low bandwidth

Table 5.19. *Main advantages and disadvantages of servo accelerometers with controlled displacement*

They are very competitive to measure slowly variable phenomena.

5.5. The signal processing associated with accelerometers

At the input of the measuring equipment, the sensor subjected to the action of the acceleration produces an electrical signal:

– directly if it is active;

– by means of a signal conditioner if it is passive.

At the output of the measuring equipment, the electric signal is translated into a form which is directly readable. In order to optimize the acquisition and the processing of the signal we use building blocks which can be:

– a linearization circuit of the sensor output signal;

– an insulation or instrumentation amplifier intended to reduce the parasitic voltage;

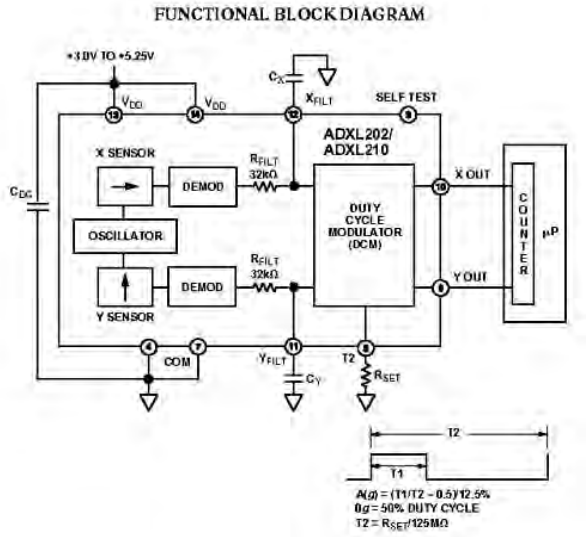
– a multiplexer;

– a digital-to-analog converter when the data must be digitally processed;

– a voltage-to-current converter or voltage-to-frequency converter when the signal must be transmitted remotely by cable;

– a frequency Modulator.

An example of electronics associated with a sensor is given in Figure 5.21.



*i*MEMS is a registered trademark of Analog Devices, Inc.

Figure 5.21. *Diagram of electronics associated with the ADXL 202/210*

5.6. Manufacturing process

The manufacturing processes of the electronic parts follow two types of process: monolithic or hybrid silicon.

5.6.1. *The monolithic processes*

The monolithic processes allow the simultaneous creation of the sensor and/or the actuator and the electronic control circuit on the same substrate of silicon.

They can be classified into four categories.

5.6.1.1. *CMOS (Complementary MOS) – BICMOS standard (Bipolar Technology and MOS)*

This approach uses CMOS and standard BICMOS used for the manufacturing of integrated circuits (IC). No modification is made in the nature, number or order of the process stages. Only some of the products can be made by standard process.

The economic advantages of this "fab-less" process are considerable.

5.6.1.2. *CMOS – BICMOS standard + back etching*

Like the previous one, this approach uses CMOS and standard BICMOS. A stage is added at the end of the microelectronics process consisting of an etching, which allows the release of moving structures like membranes.

Advantages of this process:

– development cost reduced;

– investment in "back end" limited by the etch stage.

However, the number and the thickness of the layers decrease the metrological performance of these acceleration sensors. In addition the circuit requires protection of the circuit before etching.

5.6.1.3. *Above IC*

This approach clearly separates the manufacturing process of the electronic circuits and the manufacturing process of the sensors and actuators. Initially the electronic circuit is created by a standard process until passivation is executed. Then this circuit is tested and transferred to the sensor and actuator production line. The sensor is physically built on the top of the integrated circuit, from where the name of the process "Above IC" comes.

This approach makes it possible to create a large variety of micro systems and in particular acceleration sensors. The current advantages of this technique are:

– separation of the functions;

– freedom of choice supplier;

– limited economic risks;

– economy of silicon surfaces.

5.6.1.4. *Specific process*

This approach integrates the manufacturing process of sensors, actuators and electronic circuits. It is a process resulting from the CMOS but the following are modified: nature of layers, thickness, number and order of stages. This process is completely specific to the manufactured acceleration sensor. Its advantage is adaptation to any specification.

Its limitations are cost and time of development.

5.6.2. *Hybrid process*

The hybrid process consists of creating the integrated circuit on a silicon substrate and the sensor or the actuator on another silicon substrate.

After testing and cutting, the respective chips of the electronic circuit and acceleration sensor are assembled, connected to each other on the one hand and connected to external contacts on the other hand, encapsulated and subjected to a final test.

This stage of assembly is increasingly often carried out on naked chips.

The advantages of this process are:
– separation of the functions;
– freedom of choice of the supplier;
– limited economic risks.

5.6.3. *Packaging*

An example of the packaging technique is shown in Figure 5.22.

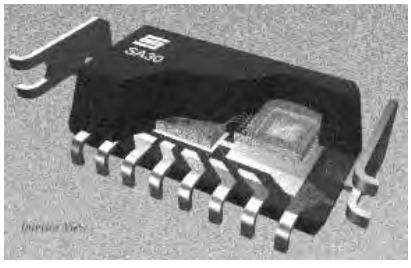


Figure 5.22. *Example of the hybrid packaging*

5.7. The calibrations:

Methods of calibration are classified into three groups according to the acceleration applied to the sensor:

– constant acceleration;

– sinusoidal acceleration;

– transient acceleration.

5.7.1. *Inclinometers and accelerometers with range lower than 1 g*

Here, we use an inclination plate to change the acceleration from –1 g to +1 g while varying the table angle from +180° to –180°. The precision is very high.

5.7.2. *Acceleration range higher than 1 g*

This calibration is carried out with:

– a rotary table (rate table);

– a centrifuge;

– a vibrating pot.

With a rotary table it is possible to generate an acceleration of a few g. The accelerometer is laid out in order to radially direct its sensitive axis compared to the vertical axis of rotation of the table. Acceleration follows the equation

$$a = \omega^2 r \tag{5.12}$$

ω = angular velocity

a = acceleration

r = distance between the axis and the seismic mass center of gravity G

It is thus possible to make variations while varying ω and r.

These tables are characterized by:

– a wobble (oscillation caused by an imbalance and expressed in seconds of arc);

– a small orthogonality error;

– a very good precision of position;

– a good speed stability.

Some of these tables are equipped with a thermostat to test the components at speed and temperature simultaneously. They are always associated with a processor which makes it possible to obtain the measurement values at the input of the sensors on the one hand (table drive) and to analyze output information of the sensors on the other.

The centrifuge

This comprises a rotary table based on the same principle as the tables presented above but with a much longer rotary support arm. These centrifuges can generate accelerations of several hundred g. In general for a calibration with centrifuges we proceed by comparison: the accelerometer to be calibrated is placed beside a known standard accelerometer as a reference and information from the two accelerometers is compared.

The vibrating pot

This is characterized by an electrodynamic generator imposing a perfectly sinusoidal movement on a table support rigorously guided in translation. The accelerometer to be calibrated is fixed on this table.

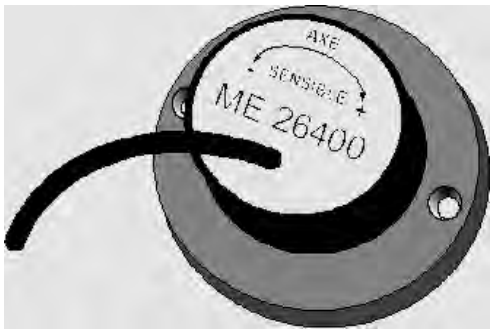
The vibrating pot is especially used to characterize the bandwidth of the accelerometers by observing their output signal when we apply a sinusoidal acceleration with variable frequency.

In addition the vibrating pot is essential for calibrating the piezoelectric accelerometers which provide only a dynamic signal.

5.8. Examples of accelerometers and inclinometers

The following illustrations are not commercial in nature but are merely intended as examples.

The range of products manufactured reflects the various technologies used for a great diversity of applications.

	
INCLINAISON **ECONOMIC INCLINOMETER MODEL *ME 26400***	**APPLICATIONS:** **angle measurement, deformation control, stabilization, regulation, safety.** **SPECIFICATIONS:** 3 ranges: +/- 30°, +/- 70°, +/- 80° Accuracy: 0.2% of the range Transversal error: < 1% Small size, low cost.

	MODELS		
	Case A		**Case B**
SPECIFICATIONS	**ME 26410**	**ME 26420**	**ME 26430**
Range (deg.)	+/- 30	+/- 70	+/- 80
Power supply (V/ma)	5/1	5/1	5/1
Sensitivity (mV/deg.)	5	3.2	4.3
Resolution (deg.)	0.01	0.01	0.01
Non-linearity (% range)	0.15	0.15	0.2
Offset (V)	2.5 +/- 0.1	2.5 +/- 0.1	2.5 +/- 0.1
Transverse sensitivity (% range)	< 1	< 1	< 1
Output impedance (KOhm)	10	10	10
Time rise (sec)	0.3	0.3	0.3
Thermal zero shift (deg./°C)	0.005	0.008	0.012
Thermal span shift (deg./°C)	–0.1	–0.1	–0.04
Operating temperature	–40 to +85°C	–40 to +85°C	–40 to +85°C
Maximum shock	100 g; 11 ms	100 g; 11 ms	100 g; 11 ms
Protection	IP 65	IP 65	IP 65
Weight (g)	18.5	18.5	72

Figure 5.23. *Economic inclinometer model ME 26400 [21] MEIRI*

QAT160/T185 Q-Flex® Accelerometers – HONEYWELL

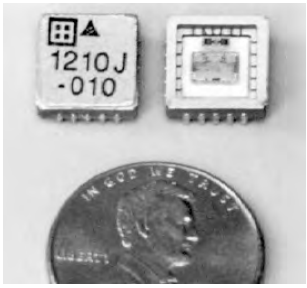
- **Two temperature ranges**
- **Field scaleable**
- **Square and round flanges available**
- **Form and fit compatibility**
- **Internal temp sensor**
- **Low power**

Performance	
• Input range	±20
• Bias	< 20 mg
- Residual modeling error	< 450 µg
• Scale factor	2.75 mA/g ±1.8%
- Residual modeling	< 450 ppm
• Axis misalignment	< 20 mrad
- One-year repeatability	< 400 mrad
• Vibration rectification (50-500 Hz)	< 100 $\mu g/g^2$
• Threshold and resolution	< 5 µg
• Bandwidth	< 200 Hz

Environmental	
• Vibration, operating & survival	
- Sine vibration	30g peak, 50 to 800 Hz
- Random vibration	20g rms
Electrical	
• Input voltage	±12.5 to ±15.5VDC
• Quiescent current	6 mA per supply
• Quiescent power	180 mW
Physical	
• Weight	55 grams
• Size	1.0 in. dia. × 0.73 in. high
• Core materials	Stainless steel

Performance by model	**QAT160**	**QAT185**
RSS bias & scale factor – one-year repeatability	1 mg	1.5 mg
Operating temperature	–40 to 160ºC	–40 to 185ºC
Survival temperature	175ºC	200ºC intermittent

Figure 5.24. *Model QAT160/T185 Q-Flex® [22] accelerometers Honeywell*

Model 1210 **CAPACITIVE one axis accelerometer**	Silicon Designs, Inc.
FEATURES • **± 4V Differential or 0.05 to 4.5 V Single Ended Output** • **Low Power Consumption** • **–55 to +125°C Operation** • **Built-in Nitrogen Damping** • **Calibrated to 1% Bias and Scale Factor (Typ)** • **+5 VDC Power** • **Responds to DC and AC Acceleration** • **Non-standard G Ranges Available** • **Hermetic LCC or J-Lead Surface Mount Package**	
DESCRIPTION The Model 1210 accelerometer is a low-cost, integrated accelerometer for use in zero to medium frequency instrumentation applications. It combines in a single, miniature, hermetically sealed package, a micro-machined capacitive sensor element and a custom integrated circuit that includes a sensor amplifier and differential output stages. It is relatively insensitive to temperature changes and gradients	

PERFORMANCE – all models: unless otherwise specified V_{DD} = 5.0 VDC, T_C = 25°C, differential mode

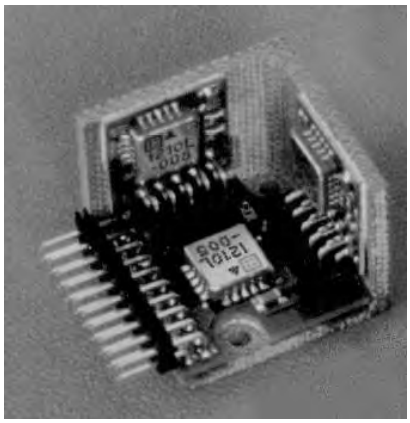
Parameter	Min	Type	Max	Units
Cross Axis Sensitivity		2	3	%
Bias Calibration Error [2]		1	2	% of Span
Bias Temperature Shift (T_C = –55 to +125 °C)		50	200	ppm of Span/°C
Scale Factor Calibration Error		1	2	%
Scale Factor Temp. Shift (T_C = –55 to +125 °C)		+300		ppm/°C
Non-Linearity (–90 to +90% of Full Scale)		0.5	1.0	% of Span
Output Impedance		90		Ohms

Figure 5.25. *Model 1210: capacitive one axis accelerometer [15] analog devices*

Model 2412 **Three Axis Analog Accelerometer Module**	Silicon Designs, Inc.

FEATURES

- **3 Axis Acceleration Sensing**
- **Contains 3 Model 1210 Accelerometers**
- **Built-in Nitrogen Damping**
- **Differential and Single Ended Outputs**
- **Low Cost Open Frame Design**
- **Low Power Consumption**
- **–40 to +85°C Operation**
- **+5 VDC Power**
- **Responds to AC/DC Acceleration**
- **Non-Standard Ranges Available**



OPERATION The Model 2412 accelerometer produces three individual analog outputs which can be used either in differential or single ended modes referenced to +2.5 volts. The voltage of each output is proportional to the orthogonal component of the applied acceleration. The Model 2412 operates with a single +5 volt power supply and generates its own +2.5 volt reference with an on-board resistive divider. The three sensitive axes are perpendicular to the bottom of each of the individual accelerometer packages, with positive acceleration defined as a force pushing on the bottom of the package.

PERFORMANCE Operating current is three times the current specified on the Model 1210 data sheet. The Bias Calibration Error is derated from the Model 1210 accelerometer to 1.5% typical and 3% max.

Figure 5.26. *Model 2412: three axis capacitive accelerometer [15] analog device*

PCB PIEZOTRONICS — DCS DIVISION capacitive accelerometers & static pressure

Series 3701 Single Axis Capacitive Accelerometers

Packaged in rugged titanium housings, the Series 3701 Single Axis Capacitive Accelerometers are able to measure acceleration from 30 mg to over 200 g within a frequency bandwidth from DC to 1,000 Hz. Built-in microelectronics, which operate from a 5 to 30 VDC power source, provide a conveniently standardized sensitivity and low-noise output signal capable of being driven over long cables through harsh environments with no loss in quality. Additionally, a self-test feature permits verification of sensor operation and allows for easy identification in multi-channel applications.

Voltage Sensitivity	Measurement Range	Frequency Range (± 5%)	Frequency Range (± 10%)	Damped Resonant Frequency	Broadband [rms] Resolution
10 mV/g (1.02 mV/ms^{-2})	200 g (1,961 ms^{-2})	0 to 800 Hz	0 to 1,000 Hz	≥ 2,500 Hz	1,000 µg
60 mV/g (6.12 mV/ms^{-2})	50 g (490 ms^{-2})	0 to 450 Hz	0 to 600 Hz	≥ 1,500 Hz	120 µg
100 mV/g (10.2 mV/ms^{-2})	20 g (196 ms^{-2})	0 to 300 Hz	0 to 500 Hz	≥ 900 Hz	70 µg
1,000 mV/g (102.0mV/ms^{-2})	3 g (29.4 ms^{-2})	0 to 100 Hz	0 to 150 Hz	≥ 400 Hz	30 µg

Performance	
Amplitude Linearity	≤ 1%
Transverse Sensitivity	≤ 3%
Environmental	
Maximum Shock – All Axes	3,000 g pk (29,400 ms^{-2} pk)
Operating Temperature Range	–40 to +185°F (–40 to +85°C)
Storage Temperature	–85 to +250°F (–65 to +121°C)
Temperature Coefficient	≤ 0.05%/°F (≤ 0.09%/°C)
Mechanical	
Housing	Titanium
Size (L × W × H)	0.85 × 0.85 × 0.45 inch (21.6 × 21.6 × 11.4 mm)
Weight	0.53 oz (15 g)

Figure 5.27. *Series 3701 single axis capacitive accelerometers (from [23])*

5.9. List of Manufacturers of Accelerometers

Manufacturers	URL
Analog devices	http://www.analog.com
BEI Sensors & Systems Co.	http://www.systron.com
Bosch Corp.	http://www.bosch.com
Breed Electronics	http://www.breedtech.com
CEC Vibration Products	http://www.cecvp.com
Columbia Research Laboratories Inc.	http://www.columbiaresearchlab.com
Condor Pacific Industries, Inc.	http://www.condorpacific.com
Crossbow Technology Inc.	http://www.xbow.com
Delphi-Delco Electronics Systems	http://www.delphiauto.com
Denso (JP)	http://www.denso.co.jp
Dytran (US)	http://www.dytran.com
Endevco Corp.	http://www.endevco.com
Entran Devices Inc.	http://www.entran.com
Fuji Electric	http://www.fujielectric.co.jp
GS Sensors (US)	http://www.gssensors.com
Honeywell (Allied Signal)	http://www.honeywell.com
Honeywell Sensing and Control	http://www.honeywell.com/sensing
KISTLER Instrumente AG	http://www.kistler.ch
Kulite Semiconductor Products Inc.	http://www.kulite.com
Measurement Specialties Inc.	http://www.msiusa.com/sensors.htm
MemSic	http://www.memsic.com
Motorola Sensor Products Div.	http://www.motorola.com
Murata (JP)	http://www.murata.co.uk
Raytheon	http://www.raytheon.com
SensoNor	http://sensonor.com
Sensotec Inc.	http://www.sensotec.com
Thales Avionic	http://www.thalesgroup.com/avionics
Silicon Designs Inc.	http://www.silicondesigns.com
Temic	http://www.temic.com
TRW/Schaevitz/Lucas	http://www.msiusa.com/schaevitz/index.htm

5.10. References

[1] Hecht Eugene, *Physique*, Translation from 1st edition by T. Becherrawy, revision by Joël Martin, ITP Deboeck University s.a. 1999.

[2] Wiley-VCH Verlag GmbH: Sensors A Comprehensive Survey, Vol. 8, Micro-and Nanosensor Technology/Trends in Sensor Markets, 1996.

[3] Wiley-VCH Verlag GmbH: Sensors A Comprehensive Survey, Vol. 1, Fundamentals and General Aspects, ed. by Göpel W., Shesse J., Zemel J.N., Lavoisier, 1996.

[4] Endevco, http://www.endevco.com

[5] ONERA, http://www.onera.fr

[6] SFIM SAGEM FRANCE, http://www.sfiminc.com

[7] Lide David R., *CRC Handbook of Chemistry and Physics*, 79th edition, CRC Press, Boca Raton, FL, 1998.

[8] Weast Robert C., *CRC Handbook of Chemistry and Physics*, 62nd edition, CRC Press, Boca Raton, FL, 1981.

[9] Wells A.F., *Structural Inorganic Chemistry*, 5th edition, Clarendon Press, Oxford, 1990.

[10] CSEM SA, http://www.csem.ch

[11] SensoNor, http://www.sensonor.com

[12] LETI (Laboratoire d'Electronique et de Technologie de l'Information) http://www.leti.cea.fr

[13] VAISALA SA, aix@vaisala.com

[14] Europractice MEMSOI TRONIC'S, Microsystems memsoi@tronics-mst.com

[15] Analog Device, http://www.analog.com/index_noflash.html

[16] MEMSIC, http://www.memsic.com

[17] Motorola, http://www.motorola.com

[18] Ultradex, http://www. agdavis-aagage.com/home.html

[19] LRBA, http://www.comelec.enst.fr/tpsp/visites/soc/lrba.html

[20] Summit Instruments, Inc, http://www.summitinstruments.com

[21] MEIRI, http://www.meiri.fr

[22] Honeywell Sensing and Control, http://www.honeywell.com/sensing

[23] PCB, http://www.pcb.com

5.11. Bibliography

1. *Proceedings of Eurosensors and Transducers conferences*

2. Campbell S.A. and Lewerenz H.J.: *Semiconductor Micromachining Vol. 1: Fundamental Electrochemistry and Physics*, Lavoisier, 1998.

3. Campbell S.A. and Lewerenz H.J.: *Semiconductor Micromachining Vol. 2: Techniques and Industrial Applications*, Lavoisier, 1998.

4. Chauffleur X.: Modélisation par la Méthode des Eléments finis du Comportement Thermomécanique de Capteurs de Pression Capacitifs et Piézorésistifs en Silicium, Thesis, 9th January 1998.

5. Esashi M.: Pressure Sensors, in: Sensors: a comprehensive Survey ed. by Bau H.H., de Rooij N.F., Kloeck B., Vol. 7, pp 331–358, 1994.

6. Mathieu J.P., Kastler A., Fleury P.: *Dictionnaire de physique*, Masson & Eyrolles, 1998.

7. Middelhoek S.: Celebration of the tenth transducers conference: The past, present and future of transducer research and development, *Sensors and Actuators*, A: Physical 2000, 82:1-3:2-23.

8. MST Benchmarking Mission to the USA – 13-25 November 1979. *Proceedings: Actes de la 1ère journée Nanotechnlogie et Industrie* – 13th April 1999.

10. Jornod R.A., Bergqvist J. and Leuthold H.: Precision Accelerometers with g Resolution, *Sensors and Actuators*, 1990, pp. 297–302.

11. *Second France-Japan Workshop*, ATRIA Hotel Toulouse, 8-10 November 1998, Book of Abstracts.

12. Van Drieënhuizen B.P., Maluf N.I., Opris I.E. and Kovacs G.T.A.: Force-Balanced Accelerometer with mG Resolution, Fabricated using Silicon Fusion Bonding and Deep Reactive Ion Etching, *International Conference on Solid-State, Sensors and Actuators*, pp.1229–30, 1997.

13. Wiley-VCH Verlag GmbH: Sensors A Comprehensive Survey, Vol. 6, Optical Sensors, Lavoisier, 1996.

Chapter 6

Chemical Sensors and Biosensors

6.1. Introduction

The world in which we live is rapidly becoming dominated by digital information. Initially, the digital revolution primarily involved stand-alone computers that gradually became networked. More recently, the merging of computing with wireless communications systems has led to an enormous growth in accessibility to, and hence demand for, this information. At present, this demand is dominated by a mixture of text, audio and image-based data driven mainly by almost ubiquitous accessibility to the internet. However, the "web" communications that has been assembled over the past decade will fuel demand for more sources of information and data about important aspects of our lives – our health, our environment, our food, our work. Sensors provide portals between the "real" or analogue world in which we live, and the digital world of computers and modern communications systems. They make it possible for us to obtain real time information about things we can see, touch, smell and hear, and about other things that we cannot detect – things that can be harmful or beneficial to us.

A chemical or biological sensor works by emitting a signal (typically a voltage or current) in response to an event such as binding between two molecules. This event typically involves a chemo- or bio-receptor (e.g. macrocyclic ligand, enzyme or antibody) binding with a specific target molecule in a sample, known as the "analyte". The electronic signal is passed to a circuit where it is digitized by an analog-to-digital converter or ADC (Figure 6.1). The digital information can then be

---

Chapter written by Gillian MCMAHON.

stored in memory, displayed visually on a monitor, or made accessible to the real world via a digital communications port.

As demonstrated by Table 6.1, there are many important markets for chemical and biological sensors, ranging from the continuous monitoring of chemical processes in industry to carbon monoxide sensing in homes.

Application	Example
Automotive	Fuel management systems, emissions monitoring
Defense	Biological and chemical warfare applications
Aerospace	Systems monitoring, air quality sensing within cabin
Agriculture	pH detection, controlled application of herbicides and pesticides
Chemical Industry	Materials testing, emission control, systems monitoring
Safety	Gas detection
Environmental	Detection of pollutants in air, water and soil, BOD
Medicine	Determination of the concentration of anesthetic gases, clinical diagnosis – *in vivo* and *in vitro*
Customs	Detection of illegal and dangerous substances, drugs and explosives
Food and Drink	Probing chemical composition, smell, freshness and flavor of food and wine, compliance checking of chemicals, fermentation checks in brewing industry

Table 6.1. *Typical applications of chemical sensors and biosensors*

A *chemical sensor* may be defined as "a device, consisting of a transducer and a chemically sensitive film/membrane, that generates a signal related to the

concentration of a particular species in a given sample". Schematically, it can be represented as shown in Figure 6.2. A *biological sensor* (a biosensor) may be defined as "a device, consisting of a transducer and a film/membrane that contains a biological material e.g. an enzyme or an antibody, that generates a signal related to the concentration of a particular species in a given sample" (see Figure 6.2).

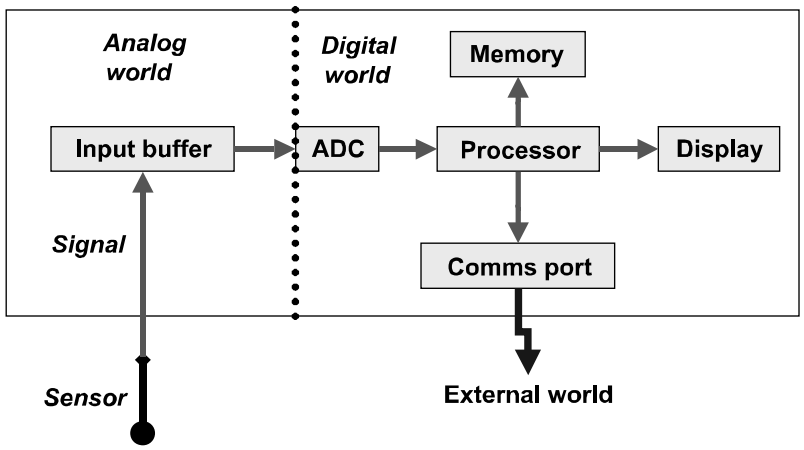


Figure 6.1. *Typical sensor instrument schematic*

The signal from the sensor passes to an input buffer (e.g. an operational amplifier) which provides high input impedance and protection for the circuit. From there, the signal is digitized by the ADC and passes from the "analog world" into the "digital world". In digital form it can be processed, stored, displayed and made electronically available to other locations through digital communications networks.

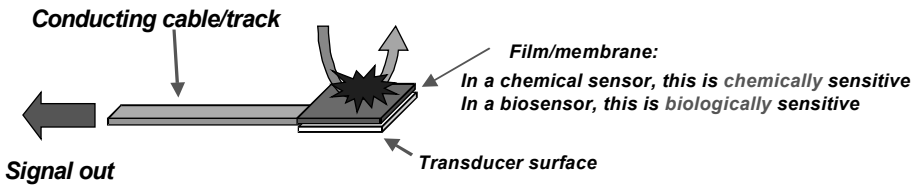


Figure 6.2. *Chemical sensor or biosensor*

In a chemical sensor, a chemically selective process occurring in or on a chemically sensitive film or membrane is coupled to signal generation at the

transducer. Examples of mechanisms commonly employed include host-guest binding, catalytic reactions or a redox process. In a biological sensor, a biologically selective process occurs in or on the film or membrane which is coupled to signal generation at the transducer.

One of the best known examples of a chemical sensor is the ion-selective electrode (ISE). These electrochemical sensors typically employ a silver-silver chloride (Ag/AgCl) wire as the transducer, a membrane at which the chemo-recognition signal generation occurs, and an internal electrolyte to electronically couple the membrane and the Ag/AgCl wire. For example, many cation-selective electrodes are based on highly plasticized poly(vinyl) chloride (PVC) membranes containing immobilized ion-receptors. When exposed to a sample containing the analyte, selective binding of the target cations occurs leading to a change in the membrane potential, which can be detected at the Ag/AgCl wire (Figure 6.3).

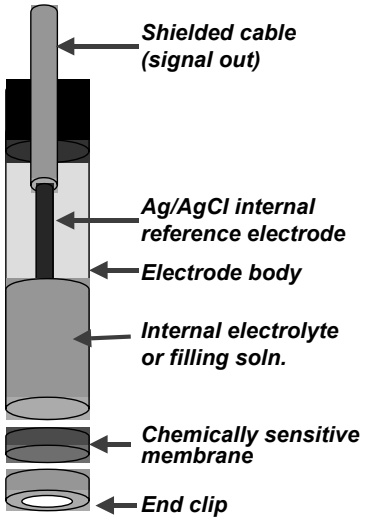


Figure 6.3.*Components of an ISE sensor*

In an ideal world, a perfect sensor would have the following characteristics:

– rapid (instantaneous) response;

– signal-specific rather than signal-selective response for the parameter/species of interest;

– be small, light, robust and easily linked to instruments and/or computers;

– able to be worn or used/left in remote locations (measurements at point of need);

– cost nothing to make and operate; never break or malfunction;

– zero energy consumption;

– compatible with automated fabrication techniques;

– no calibration required;

– no reagents needed (reagentless assays).

Obviously, this perfect sensor does not exist, and never will, but it is vital that chemical sensors and biosensors meet user demands for devices that are as close to the ideal reagentless, calibration free, reliable devices as possible.

The driving forces behind the recent surge in the development of sensor technology include general health monitoring, environmental awareness, food quality regulations, the momentum generated by major wireless communications initiatives such as "bluetooth", health and safety in the workplace and trends in consumer products, e.g. the "smart home" (ubiquitous computing).

Helping to speed these new sensors along are developments in planar fabrication techniques such as thick and thin film methods, the ability to make sensors smaller, the possibility of integrating more features (e.g. sample handling, processing and electronic communications) onto the same device and the availability of mass production techniques which mean lower unit cost and reproducibility.

6.2. What is involved in developing a sensor?

An ideal sensor is a device that will only detect a desired target analyte that is present within a given sample. Unfortunately, most samples usually contain many other analytes that may interfere with the sensor's ability to detect the target analyte. As a result it becomes necessary to design sensors which are as specific as possible for an analyte so that the sensor will discriminate against any interferents present. This is achieved using molecular recognition, where the sensor contains what is called a host molecule or chemoreceptor that will selectively bind the target analyte (which is quite often referred to as a guest). Once a suitable host-guest chemistry has been found, the host molecules need to be immobilized or incorporated in some way into the sensor. Finally, a means of signaling that the binding/recognition event has occurred has to be found (transduction).

6.2.1. *Molecular recognition*

One of the key requirements for molecular recognition is the existence of pre-organized groups within the host molecule that can selectively enclose or bind the guest ion (single atom) or molecule. Examples of this kind of host-guest recognition can be seen in everyday living processes. All living organisms use enzymes, which are proteins that contain "pockets" that are designed to recognize a specific analyte. This means that only one specific analyte is capable of entering into the enzyme pocket. Enzymes can be used in sensors (biosensors), but they are commonly unstable and may not be readily available. It is therefore necessary to synthesize new classes of molecules which are capable of acting as host molecules. In designing host molecules to be used in a sensor, the following criteria should be considered:

– The host molecule should be stable to the conditions in which it will be used.

– It must be able to selectively bind the target analyte in the sample.

– It must be capable of being immobilized in a film/membrane which is contacted with the sample.

– It must signal that a host-guest binding event has occurred.

– Ideally it should release the analyte after detection so the host is free to be used again.

The most commonly used synthetic hosts come from a class of molecules known as macrocycles. Various classes of macrocycles are outlined in Figure 6.4. The most common feature between the classes of macrocycles is that they contain cavities that behave as host pockets for guest molecules. Selectivity of hosts can be readily accomplished by varying the size of the preformed cavities. For example, 12-crown-4 has a small cavity ideal for binding small ions such as Li^+ whereas 18-crown-6 has a larger cavity, which is better suited for larger ions such as K^+. It is quite obvious that the size of cavities is important for host selectivity, but the question remains as to what attracts an ion or molecule into the pre-formed cavity, and what factors stabilize the host-guest complex.

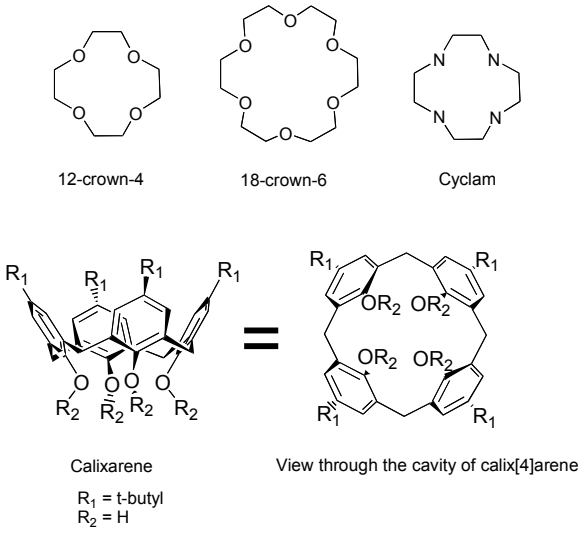


Figure 6.4. *An illustration of the various classes of macrocycles*

In enzymes, weak non-covalent interactions (H-bonding, electrostatic, dipole-dipole, Van der Waals, pi-pi) are used to bind the guest into the enzyme pocket. These interactions stabilize the host guest interaction. The macrocycles outlined in Figure 6.4 all contain polar functionalities which are capable of interacting with guests via H-bonding, electrostatic interaction and dipole-dipole interactions. It is also desirable that the binding in the cavity not be *too* strong, since it is important that the guest analyte be released from the host after it has been detected and measured. The oxygen-containing crown ethers and calixarenes are ideal for binding metal cations based on both their cavity size and the high electron density present on the oxygen atoms in the cavity.

In Figure 6.5, the structure of tetraethylestercalix[4]arene is shown. This compound was prepared from the parent macrocycle calix[4]arene shown in Figure 6.4. Although the parent compound selectively binds Li^+ over other metal cations [1], the modified version of the parent macrocycle has excellent selectivity for Na^+. Thus, with synthetic modification it is possible to increase the size of the host cavity and new functionalities can be introduced which will favor binding of certain ions and molecules. Another example of modified calixarenes, which further demonstrates this principle, is the group of novel calix[4]arene phosphine oxides (Figure 6.5). Changing the binding groups on the same calix[4]arene template from esters to phosphine oxides shifts the selectivity from Na^+ to Ca^{2+}. Increasing the number of repeat units in both esters and phosphine oxides to six increases the

cavity size, and the selectivity shifts in favor of longer cations such as Cs^+ and Pb^{2+} respectively.

Much work with sensors has involved the detection of metal cations of biomedical and environmental interest. Some host compounds have also been developed for the selective detection of uncharged small neutral molecules. The calixarenes have found widespread application in this area [1]. One example involves using a novel calixarene tetra-(S-propranolol) calix[4]arene amide containing four chiral moieties on the lower rim to selectively differentiate between enantiomers of phenylalanilol.

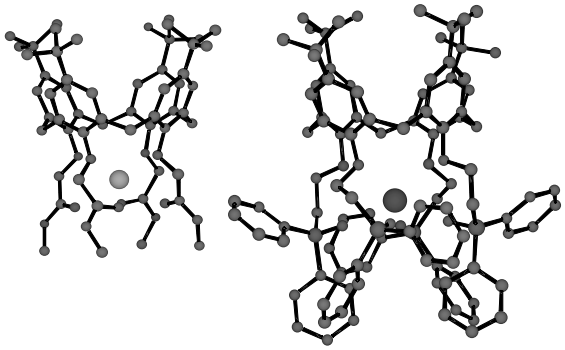


Figure 6.5. *Tetraethylestercalix[4]arene and tetra phosphine oxide calix[4]arene. The tetraethylester cavity is selective for sodium while the larger tetra phosphine oxide cavity is selective for calcium*

6.2.2. *Immobilization of host molecules*

Once hosts that are selective for specific analytes have been developed, it is then necessary to immobilize these compounds into the sensor device. The most widely used method is the incorporation of the host molecule into flexible polymer membranes, which are then fixed into the sensor. Most of these membranes are cast from solution and the most commonly used polymers and plasticizers are usually soluble in organic solvents. It is therefore essential that the host molecule be soluble in a variety of organic solvents. This is normally achieved by the introduction of lipophilic groups such as the t-butyl groups in calixarenes. Problems with this method of immobilization arise when the membranes come in contact with the sample (which is usually aqueous). It has been found that the host molecules leach from the membrane over time, thus lowering the lifetime of the sensor. One way of getting around this problem is to covalently link the host molecule into the polymer of the membrane. Unfortunately such an approach can be quite synthetically cumbersome since a reactive handle must be incorporated into the host molecule,

which may itself affect or inhibit selectivity. In the case of tetra-(S-propranolol) calix[4]arene, amide allylic groups have been introduced into the molecule. These groups can be used to covalently link the host into polymers and onto silica surfaces via platinum coupling.

6.2.3. *Transduction of signal*

As mentioned earlier, it is essential that the binding event between host and guest be detected. It is thus necessary for a mode of transduction to be available. Various modes of "transducting" the binding event are available such as via electrochemical signals (electrochemical based sensors), optical signals utilizing changes in fluorescence or absorbance (bulk optodes) and plasmon resonance. These methods of transduction will be discussed in more detail in the following chapters, but this section deals with the methods of incorporating a transduction element into the host itself.

With most sensors, transduction is accomplished either electrochemically or by optical detection. An electrochemical mode of detection usually requires that the membranes, which contain the host molecule, be placed onto an electrode surface and, upon binding of the guest, an electrochemical response is observed. This approach works very well when the target analytes are charged species such as the metal cations described above. Unfortunately neutral molecules are not as readily detected by this method. To overcome this problem, optical methods of detection have been successfully used. These optical methods are often more sensitive than electrochemical transduction techniques.

6.3. Electrochemical sensors

Chemical sensors are devices that provide information about the types, concentrations and chemical states of the species present within a sample. In this section, we focus on electrochemical sensors which represent a very important group of chemical sensors, some of them with properties approaching those of the ideal sensor. Broadly speaking, electrochemical sensors are based on one of three categories of transduction mechanism: amperometric and voltammetric, potentiometric or conductimetric. Amperometry senses current generated (at a fixed voltage) when an analyte is selectively oxidized or reduced resulting in the exchange of electrons. Two-electrode cells are common. In voltammetry, current is again measured, but as a function of the applied potential. The reference electrode's potential is constant, and the working electrode assumes the value of the applied potential. The working electrode is the site where electrolysis occurs and this generates the measured current. Three-electrode cells are usually used.

Potentiometry, where there is no current flow, measures the accumulation of charge density (voltage) at the surface. These reactions are spontaneous and two-electrode cells are used. Finally, conductometric methods measure conductivity through the sample between electrodes.

The variety of systems that can now be probed using electrochemical sensors is truly vast, ranging from the secretions of single cells to the inside of reaction vessels run at high temperatures and pressures. Electrochemical measurements can now be performed in the solid, liquid and even gas phases to accurately determine the concentrations of redox reactions active at concentrations as low as parts per trillion.

6.3.1. *Amperometric and voltammetric sensors*

An amperometric sensor depends on the maintenance of a fixed potential between two electrodes. More recently, pulsed techniques have come to the forefront (see Table 6.2). Usually amperometric detectors have a fixed voltage, while voltammetry implies a changing voltage. Enhancement of polarization may be achieved by using microelectrodes, since they consist of surface areas of a few square micrometers.

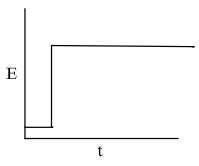
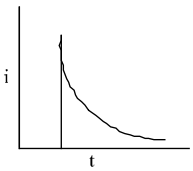
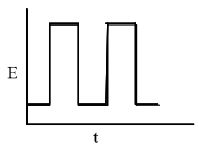
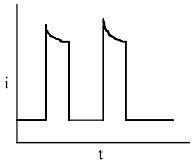
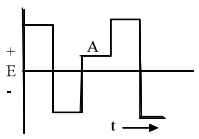
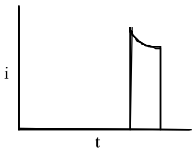
Method	Applied potential	Measured signal	Comments
Chrono-amperometry			The electrode potential is pulsed to a region in which the analyte is electroactive. The decaying current reflects the growth of the diffusion layer.
Pulsed amperometry			The electrode potential is pulsed for a short period of time to a region in which the analyte is electroactive. Between pulses the diffusion layer may be eliminated by forced or natural convention.
Pulsed amperometric detection			This allows for electrode conditioning, analyte sorption and catalytic electro-oxidation. The current is only measured in the last part of the cycle (point A).

Table 6.2. *Some amperometric methods*

In voltammetric systems the cell consists of three electrodes immersed in a solution containing the analyte and supporting (non-reactive) electrolyte. One of the three electrodes is the *working* electrode whose voltage changes linearly with time. The second electrode is the *reference* electrode whose potential remains constant through the procedure. The most generally available reference electrode for work in aqueous solutions is the saturated calomel electrode (SCE). Other useful reference electrodes are based on half reactions involving a silver electrode. The Ag/AgCl reference is quite common. The third electrode is the *auxiliary* or *counter* electrode, which is often a platinum wire, to serve as an electrical conductor from the source through the solution to the microelectrode. The basic requirement of the counter electrode is to provide an alternative route for the current to follow, so that only a small current flows through the reference electrode.

The typical voltammetric cell is of the order of 50 ml volume but there has been a drive towards smaller cells which would extend their range of applications. This has been provided by screen-printing techniques (see Figure 6.6).

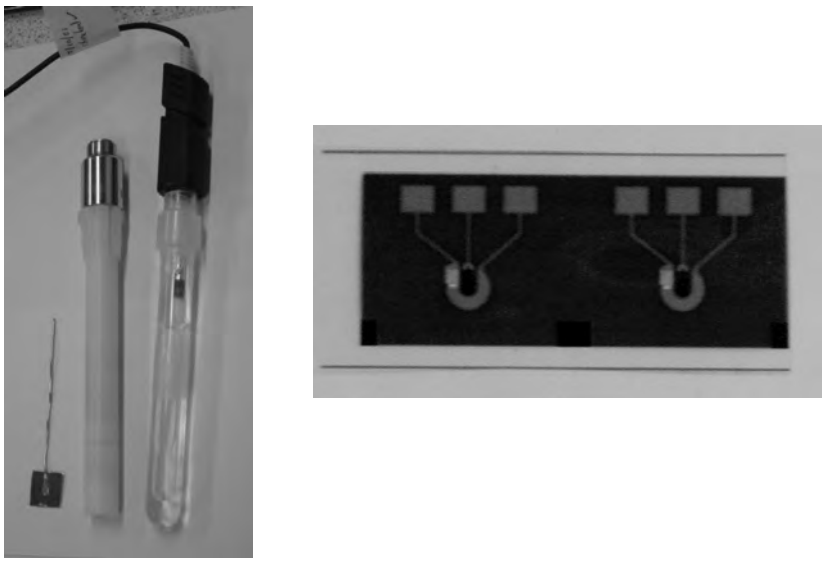


Figure 6.6. *The standard size components of a voltammetric cell and the new planar three electrode cell produced by screen-printing (approximately 2 cm × 2 cm)*

Amperometric sensors employ electrodes usually composed of solid materials, such as graphite fibers, glassy carbon, graphite paste, platinum, gold, copper, nickel or some other metals [2]. Electrodes employed in voltammetry are often small flat

disks of a certain conducting material that is pressed and fitted into a rod of an inert material such as Teflon that contains a wire contact. There are a variety of conductive materials which can be applied; these include an inert material such as platinum or gold, glassy carbon or pyrolytic graphite or a metal coated with a film of mercury. Mercury electrodes are used in voltammetry experiments as they tolerate the relatively large negative potentials since they have a high overvoltage of hydrogen. The dropping mercury electrode (DME) is usually used in polarography experiments since it benefits from a continuously renewed surface enabling stability and sensitivity to be achieved. Lastly the surface of a mercury electrode allows the formation of amalgams with many metal ions.

One disadvantage of amperometric sensors is that the sensor response is dependent on the hydrodynamic conditions and, accordingly, on the reagent flow rate in flow injection or liquid chromatography (LC) detectors. They also may suffer from interactions between the sensor and sample matrix, leading to poor stability and reproducibility of the electrode surface. Electrochemical analysis normally requires the presence of a suitable base electrolyte on which the background current is dependent. As such, incompatibility of gradient elution may occur in electrochemical detectors of this type.

In flow injection analysis (FIA), amperometric sensors have been used in direct current modes, which include the normal, reversed and differential pulse techniques. Electrochemical cells used for high performance liquid chromatography (HPLC) detection usually operate amperometrically. The detector cell must conciliate a number of factors when used with HPLC. These factors include a low-volume electrochemical cell, low dead volume connective tubing, low response time and low impedance between the electrodes. Therefore, it is desirable that the auxiliary and reference electrodes be placed as close as possible to the working electrode.

6.3.1.1. *Cyclic voltammetry*

In cyclic voltammetry, the working electrode potential is swept back and forth across the formal potential of the analyte. Repeated reduction and oxidation of the analyte causes alternating cathodic and anodic current flow at the electrode. The solution is not stirred. Experimental results are usually plotted as a graph of current versus potential, and a voltammogram as shown in Figure 6.7 is produced. The voltammogram displays two asymmetric peaks, one cathodic and the other anodic. The signal of primary interest to the analyst is the height of the peak, which is proportional to the analyte concentration.

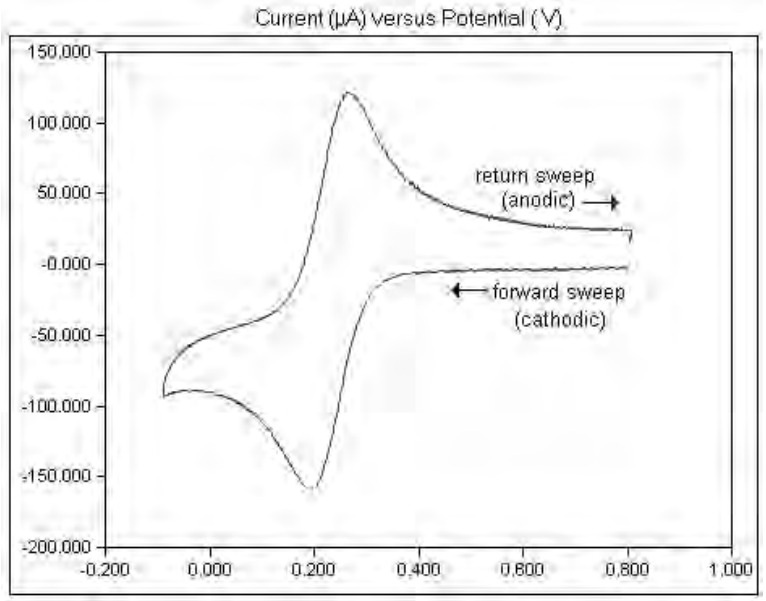


Figure 6.7. *A typical cyclic voltammogram*

6.3.1.2. *Hydrodynamic amperometry*

Methods involving forced convection (i.e. stirring) to aid the analyte to reach the electrode are called hydrodynamic methods. When the solution in stirred in a controlled mode, a non-turbulent flow of solution can be directed towards the working electrode. Convective diffusion may be obtained in two ways: firstly, by moving the solution relative to the electrode or, secondly, moving the electrode relative to the solution. Of the systems developed to move the electrode, the most common is the rotating disk electrode, which consists of a disk on the end of an insulated shaft that is rotated at a controlled angular velocity. Analyte is put across to the electrode surface by a combination of two types of transport. Firstly, the vortex flow in the bulk solution constantly brings fresh analyte to the outer edge of the stagnant layer, then the analyte moves across the stagnant layer via simple molecular diffusion. The thinner the stagnant layer, the faster the analyte can diffuse across it and reach the electrode surface. Faster electrode rotation makes the stagnant layer thinner. Thus, faster rotation rates allow the analyte to diffuse to the electrode faster, resulting in a higher current being measured at the electrode.

Experimental results are generally plotted as a graph of current versus potential, and a typical rotated disk voltammogram exhibits a sigmoidal shaped wave where the height of this wave provides the analytical signal.

6.3.2. *Potentiometric sensors*

Potentiometry is one of the oldest analytical methods which is still widely used for electrochemical analysis. The most common potentiometric sensors are ISEs, which yield information about the concentration of a compound in terms of the potential difference between two electrodes [3]. The method is popular because of its simplicity, selectivity and relatively low cost. Nevertheless, in some applications, problems may originate due to the slow sensor response, particularly at low analyte concentrations, and to the non-linear relationship between the potential and the analyte concentration.

The electromotive force (EMF) arises from a spontaneous process that has the potential or capacity to occur. Therefore, it is the aim of potentiometric measurements to operate the electrochemical cell in such a way that the spontaneous process does not occur, to obtain a value for the EMF and to correlate this in terms of the components in the cell solution. In order to do this the potential of the cell must be measured under equilibrium conditions, i.e. at zero current. A high impedance voltmeter is used to compensate the EMF by applying an external potential difference. Thus, by knowing the exact potential difference needed to compensate the EMF, such that no current flows, the determination of the EMF is possible. A potentiometric cell uses two electrodes, an indicator (working electrode) which is an ISE in this case, and a reference. The function of the reference is to maintain a constant potential in order to allow measurement of the potential at the indicator electrode. A typical cell, consisting of two electrodes (e.g. an ISE and reference electrode), a high impedance voltmeter and the sample solution, is shown in Figure 6.8. The ISE in this case is the indicator/working electrode and its purpose is to allow potentiometric determination of the activity of certain ions in the presence of other ions. The ISE thus constitutes one half of the Galvanic cell, consisting of an ion-selective membrane, an internal filling or contact solution (or solid contact in the case of solid state ISEs, e.g. a hydrogel) and an internal reference electrode. The external reference electrode gives the other half of the Galvanic cell. The external reference is typically an electrode such as Ag/AgCl, Cl^- or Hg/HgCl, Cl^- (calomel).

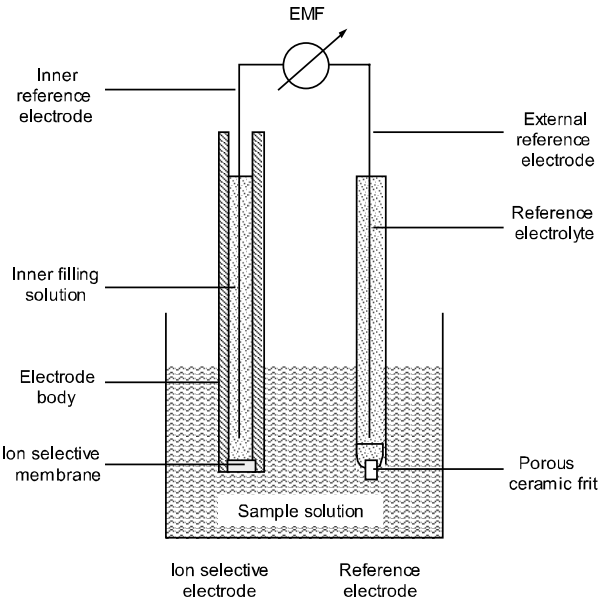


Figure 6.8. *Schematic of a typical potentiometric cell assembly incorporating an ISE and reference electrode as the galvanic half cells and a high impedence voltmeter for measurement of the cell EMF*

The ISE is very crucial in potentiometric analysis. It should only interact with the analyte of interest in a particular sample so that the sensor response reflects the concentration of the species in solution and not other components in the sample matrix. Such electrodes are classified by the mechanism by which the electrode potential is produced [4]. In analysis, solutions of constant ionic strength are commonly used and, therefore, this enables calibration to be described in terms of concentrations rather than activities. However, extraction of unknown concentrations from such calibration data assumes that the unknowns are also at the same ionic strength. This is usually achieved by adding a constant high background concentration of an electrolyte that does not affect the electrode signal – ionic strength adjuster (ISA) – to all standards and unknowns.

6.3.2.1. *Ion-selective electrodes*

ISEs are potentiometric devices selective for ionic species. The various types of electrodes used in potentiometric cells are summarized in Table 6.3.

Electrode type	Example
Electrode of the first kind	Metal electrode in reversible equilibrium with a solution of its own ions ($Ag/AgNO_3$)
Electrode of the second kind	Metal electrode in contact or coated with one of its own compounds: the interface is reversible (Ag/AgCl)
Electrode of the third kind	Metal electrode in reversible equilibrium with a solution of one of its chelates ($Hg/HgEDTA^{2-}$)
Membrane	Electrode in which a potential difference occurs across a membrane (pH glass electrode)

Table 6.3. *Types of potentiometric working electrodes*

ISEs based on membranes are by far the most useful for analytical sensor devices. A broad diversity of electrode materials has also been investigated such as carbon (impregnated graphite, glassy carbon, pyrolytic graphite, carbon paste), mercury, platinum, gold, and a variety of metallic wires. Also, coupling biological agents into electrode membranes leads to improved selectivity and response characteristics of ISEs. ISEs are ideally suited as detectors in flow analyzers and the development and application of ion-selective electrodes is an ongoing area of analytical research.

6.3.2.2. *Coated-wire electrodes and polymer-membrane electrodes*

Polymer based electrodes are used for the determination of various ions. Conventional ISEs and coated wire electrodes (CWEs) consist of a plasticized PVC membrane incorporating a carrier ionophore to provide ion discrimination. The lifetime of these electrodes is usually limited due to the leaching of the ionophore and plasticizer from the polymer or crystalline matrix. For applications in clinical and environmental chemistry, low cost and disposable or small sized sensors would be favorable. Many commercially available ISEs with polymer matrix membranes are very expensive due to them being hand-made whereas automation planar manufacturing of such devices would reduce this cost considerably.

CWEs are constructed with the deposition of a crystalline membrane on a metallic wire electrode either electrolytically deposited or in conjunction with a polymer material such as PVC. Electrodes which are coated with a layer of salt by a chemical reaction are known as electrodes of the second kind. These provide a source of the ionic species and ensure that the solution near the electrode is saturated

with it. Electrodes of this sort include the silver/silver chloride, silver/silver bromide, silver/silver iodide, silver/silver sulfide and mercury/mercurous chloride electrodes. Interest in these electrodes stems from the chance of developing electrodes selective for anions by incorporating a membrane containing selective anionic sites. The first of these electrodes was the fluoride ion electrode [5]. The membrane contains a single crystal of lanthanum fluoride, doped with europium (II) that generates the crystal defects required for its electrical conductivity. Many insoluble crystalline salts exist, which are selective toward both cation and anion exchange, but which, however, are not sufficiently conductive to be useful membranes in ISEs. In addition it has been found that electrodes based upon polycrystalline Ag_2S membranes become reliable electrical conductors owing to the mobility of the silver ions in the membrane matrix. Ag_2S mixed with PbS, CdS and CuS produces membranes that are selective to Pb^{2+}, Cd^{2+} and Cu^{2+} respectively, again owing to silver ions being electrically conductive in these solid membranes.

In polymer-membrane electrodes, electron transport occurs by ionic transport across a membrane from the sample solution to the metal electrode. The selectivity of a membrane electrode for a particular ion is determined by the mechanism of ionic transport across the membrane and not by the electron transfers process. The incorporation of membrane materials in ISEs is of great importance in the improvement of stability and reproducibility of the electrode response. Studies of membrane systems have promoted the selectivity features and fast response qualities that are desirable in sensors. The source of solvent polymeric membrane sensors for ISEs relies on the application of liquid ion exchangers trapped in the membrane matrix. The feasibility of PVC membranes containing selective ionophores for the detection of anions has been examined. The membranes were coated directly onto a metallic substrate by a simple method of dipping a wire, such as platinum, into a solution of PVC/ionophore/plasticizer components dissolved in tetrahydrofuran. PVC electrodes have been successfully applied to the fabrication of anion-selective ISEs.

The importance of ionophores in ISEs is associated with their greater selectivity compared with resinous ion exchangers, while the plasticizer/solvent in which this ionophore is dissolved plays additional roles for adjusting:

– mobility of the ion-exchanger sites; and

– the relative permittivity of the resultant organic phase.

The analytical characteristics of all PVC membrane-based ion-selective electrodes depend both on the types of sensor and the plasticizer utilized and also on their relative proportions in a membrane mixture after Solsky [6]. Plasticizers are used to lower the glass transition temperature of the polymer and make it soft and flexible. A typical membrane cocktail could be composed of (%w/w): 33% PVC,

66% NPOE (plasticizer) and 1% ionophore/exchanger (ratio at least 2:1 by mole) dissolved in a volatile solvent such as tetrahydrofuran.

Liquid membrane systems for ISEs have developed greatly by the use of naturally occurring ionophores such as valinomycin for potassium and by the synthesis of a range of materials and ionophores of selective complexing properties. The simple step of trapping an ionophore ensures sufficient robustness for the application of ISEs. The sensing components for say, nitrate, e.g., tridodecylhexadecyl ammonium chloride, in association with an appropriate plasticizing solvent mediator such as dibutylphthalate are readily setup in a PVC matrix membrane system. The membrane is readily fabricated into conventional style membrane electrodes as for glass pH electrodes or it is used in coated-wire electrodes and in specially designed systems for flow analysis.

The ion exchange processes that occur within the membrane itself describe the response mechanism of polymer membrane electrodes. The response has been found to be affected by the following factors: the addition of plasticizers and reduced temperature. The rate of ion-exchange is measured by the exchange current or exchange flux density. It is defined by the number of moles of ions that flow in opposite directions per second per square centimeter. The degree of ion exchange is measured by the equilibrium constant for the process. Two or more ions of equal charge or same sign of charge will exchange with a measurable ion exchange constant that is a ratio of the single ion partition constants. The determination of selectivity coefficients of a membrane electrode is important to the application of the sensor as well as for understanding the processes that occur during a measurement. Ionophoric reagents have widespread application in potentiometric ion-selective electrodes due to their ability to transport selectively targeted ions from an aqueous sample solution to a hydrophobic membrane phase.

6.3.2.3. *Potentiometric sensor arrays*

The use of sensor arrays allows multicomponent analysis to take place in a single sample. There are many analytical applications that require the simultaneous determination of mixtures of simple inorganic cations and anions. At present, these are typically achieved by splitting the samples and performing analysis of the cations and anions separately using instrumental methods such as ion chromatography, atomic absorbance for cations, electrophoresis or bench-type ISEs. In every case, more than one assay will be required, as cations and anions are determined separately. An exception to this is the widespread use of ISE arrays for blood electrolyte analysis, which typically requires sodium, calcium, potassium, pH and chloride to be determined, along with the blood gases CO_2 and O_2. Curiously, however, the array approach is not common outside this relatively focused area to solve other analytical problems with similar requirements. Furthermore, as the

analyzers are commonly flow-based instruments, the arrays are usually built into a flow-cell, which can easily be adapted for related approaches such as FIA. The manufacture of these arrays is relatively complex, and involves a combination of techniques such as screen-printing, spin coating, accurate drop-on-demand liquid handling, injection molding and, ideally, clean room conditions. An example of such a device, manufactured by SenDx Inc., Carlsbad, California, USA, is shown in Figure 6.9. This is used for profiling the general electrolyte balance of the blood e.g. K^+, Na^+, Ca^{2+}, etc.

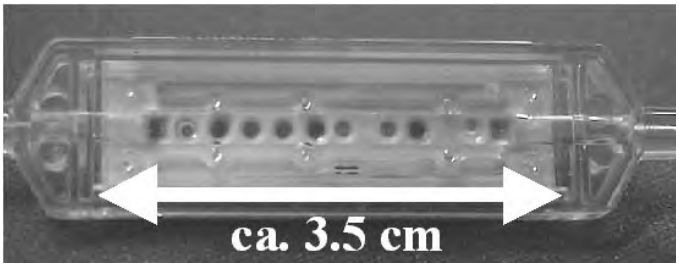


Figure 6.9. *Sendx sensor array and flowcell*

6.3.3. *Resistance, conductance and impedance sensors*

Some electrochemical sensors are based on measuring the change of electrical conducting properties of the sensing layer as a result of interactions between the analyte and (a certain component of) the sensing layer. A simple circuit is sufficient for most measurements where Ohm's Law is obeyed. The sensor acts as an additional resistor and the voltage change due to the change of the resistance of the sensor is recorded. The output may be in the form of a voltage drop (V) or a conductivity change (ohm). The materials used in these sensors are semiconductors such as metal oxides (e.g. tin oxides, zinc oxide, tungsten oxide, etc.), organic macromolecule-metal complexes (e.g. metal phthalocyanine), conducting polymers (e.g. polypyrrol, polyaniline and polythiophen, etc.) and carbon black-polymer blend. This class of sensors finds applications mainly in gas sensing.

A classical example is the tin oxide (SnOx) based gas sensor. The tin oxide layer is first activated by heating to >250°C to form a depletion layer where oxygen is chemisorbed on the surface, capturing two electrons from the conducting band of the tin oxide, thereby causing the electron mobility and carrier concentration to decrease. The conductivity of the activated sensor may be increased or decreased depending on the nature of the incoming gases. Reducing gases (e.g. alcohol) increase the conductivity and oxidizing gases (e.g. nitrogen oxides) further reduce the conductivity of the sensor.

The advantages of these semiconductor-based sensors are:

– easy fabrication (by sputtering);

– simple operation; and

– low production cost.

The main disadvantages of these sensors are the high-energy requirement for operation and low selectivity so recent trends include reducing their running cost and improving their selectivity. Hence, materials that operate at ambient temperature such as conducting polymer and carbon black-polymer blends have been extensively investigated. The array-sensing approach combined with statistical algorithms such as cluster analysis and principal component analysis (PCA) greatly improves the selectivity of the sensing technique. With arrays, it is very important that the sensors are very reproducible and have predictable behavior over time. Conducting polymer-based sensors work well and commercial sensing systems based on this technique are available. However, careful control of electrochemical deposition conditions is critical for reproducible sensors and the conducting polymer-based sensors are subject to poisoning. A more versatile sensing system is based on the carbon black-polymer blend where the carbon particles give the electrical conductivity and the polymer (any polymer) provides the sensor function (see Figure 6.10). The sensor response is a result of the swelling of the polymer [7], which causes the conductivity of the sensor to change. Weak Van der Waals forces between the polymer and the target gas molecules are responsible for the swelling of the sensing layer; therefore it is purely a physical change and is reversible, which makes the sensor reusable.

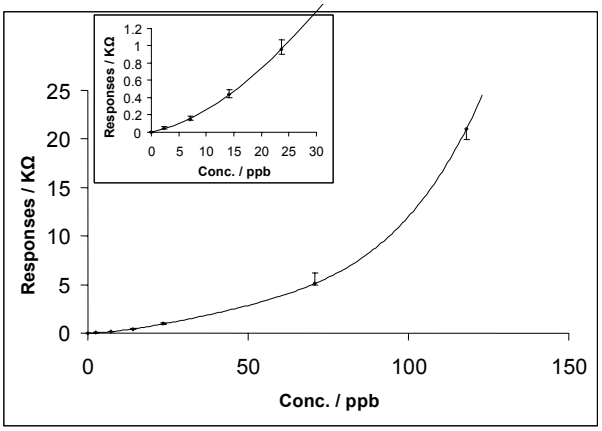


Figure 6.10. *A typical calibration plot obtained from a screen-printed electrode coated with poly(isobutadiene) (PIB)carbon black membrane. The analyte used was cyclohexane. The constructed sensor detects down to 3ppb of the vapor*

The advantages of using carbon black-polymer blend as a sensing layer are:

– the selectivity can be tailored by choosing polymers with desired functionalities;

– easy fabrication (screen-printable);

– robustness;

– reproducibility;

– low operation cost.

6.4. Optical sensors

Optical techniques for chemical and biological analysis are well established. In many cases, sensors based on these techniques use optical fiber technology, although planar waveguide configurations are increasingly favored. Optical sensor devices can be used for the detection and determination of physical or chemical parameters through the measurement of changes in some optical property. There is increasing flexibility in the measurement mode, e.g. evanescent wave, surface reflectance, and emerging technologies such as surface plasmon resonance (SPR). There is also a drive towards integration of optical components (source, waveguides, sampling region, detectors) on a planar platform. The cost, power consumption and size of such components is rapidly decreasing with the result that there is the potential for low cost, multichannel information on absorbance, color, fluorescence and turbidity etc. Portable UV-Vis, Raman and fluorescence instruments are now available for field measurements.

6.4.1. *Methods of detection*

Most optical sensors are based on a spectroscopic technique such as measurement of absorbance, reflectance or fluorescence, where the signal obtained is related to the concentration of the analyte. The two most popular methods are absorption and fluorescence. If both modes of detection are available for a particular compound, fluorescence would be preferable because of better sensitivity due to being measured against an almost zero background. It is also worth noting that a fluorescence signal is emitted by the molecules themselves (or reagent molecules) and therefore contains information about the molecule being measured, while absorption is based on the transmitted light left over after incident light has passed through a sample. However, absorptiometric detection is becoming more popular due to miniaturized inexpensive spectrometers. Figure 6.11 shows the principles of absorbance and fluorescence measurement. With absorbance, light is passed through a sample and the non-absorbed, transmitted light is monitored along the line of the

incident light. Fluorimetric detection is based on the absorption of a photon and the subsequent release of a photon that is generally of lower energy. The emitted photon has a fluorescence emission peak at a longer wavelength than the absorption peak and is usually measured at right angles to the incident radiation (see Figure 6.11).

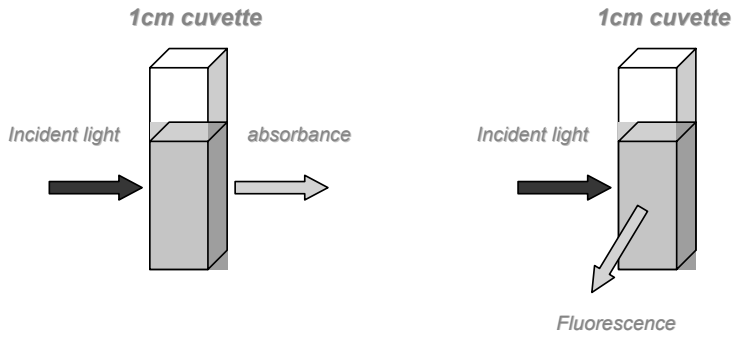


Figure 6.11. *The principles of absorbance and fluorescence measurement*

Raman spectroscopy is often used for "fingerprint" vibrational spectra. It can be used to identify the composition of a species or to quantify a species. Raman spectroscopy is based on a weak scattering phenomenon and so methods are emerging to enhance the signals. By locating the sample on metal film islands, the signal can be magnified and this technique is called surface-enhanced Raman spectroscopy (SERS).

Some chemical and biochemical reactions emit their own light – chemiluminescence or bioluminescence. This can be used to measure the concentration of some specific analytes under certain conditions. The major advantage of this phenomenon is that no light source is required.

6.4.1.1. *Evanescent wave sensors*

When light propagates in an optical fiber or waveguide (see Figure 6.12) under conditions of total internal reflection, a fraction of the radiation extends a short distance from the guiding region into the medium of lower refractive index that surrounds it (see Figure 6.13). This so-called evanescent field can interact with molecular species in close proximity to the fiber core e.g. in the cladding of the optical fiber itself. The motivation for embracing this evanescent wave (EW) approach is due to a number of advantages it offers:

– It is easy to miniaturize and no coupling optics are required.

– It is possible to discriminate between surface and bulk effects by controlling the launch optics.

– Sensitivity can be higher when compared to bulk optic approaches.

– Highly absorbing or highly scattering media are suited to this technique because the effective path length is so small and scattering does not interfere to the same extent.

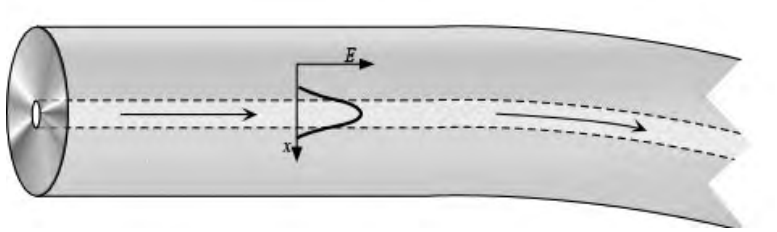


Figure 6.12. *Single mode of guided light in an optical fiber*

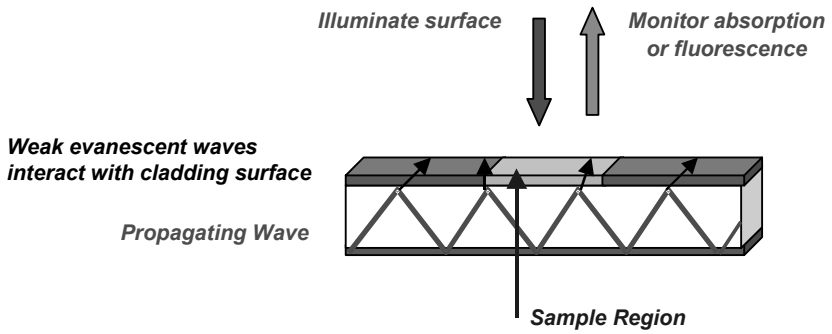


Figure 6.13. *Schematic diagram of evanescent waves and detection at the sample surface*

However, there are also disadvantages to the use of EW sensors. Surface contamination and build-up of sample can reduce sensitivity and require cleaning and recalibration on an ongoing basis. Commercial EW sensors may be best suited as single-use sensors or in situations where regular surface cleaning can be carried out. Also, the design of fluorimetric EW sensors is difficult since some of the fluorescent radiation is coupled back into the waveguide or the optical fiber and this is difficult to quantify. As such, the optimal design of absorptiometric EW sensors may prove to be less complicated.

Sensor systems based on SPR exploit the phenomenon of evanescent waves. A surface plasmon is a collective oscillation of free electrons in a metal film. The plasmon can be excited by evanescent waves under certain conditions which depend on the refractive index of the sensing material. Once excited, there is a sharp minimum in the reflected light at the precise angle of incidence. Essentially the system functions as a highly sensitive refractometer with selectivity imparted by the material coated on the sensing side of the metal film. Sensitivity is higher for larger molecules because their effect on refractive index will be proportionally greater. The BIAcore system from Pharmacia, Sweden is well established and works as a biosensor development system. This instrument and its applications are discussed further in section 6.6.1.1.

6.4.2. *Reagent-mediated sensors*

Reagent-mediated sensors (optrodes) employ an intermediate reagent which responds optically to the presence of the analyte of interest. These reactions are often adapted from well-established color chemistries and can be extremely sensitive. The optimal reagent should be very sensitive to and selective for the analyte, exhibit reversibility on removal of the analyte and respond quickly. If the sensor is not reversible, it is generally referred to as a *probe* as opposed to a sensor.

Whether the optrode is based on an optical fiber, waveguide or another platform, immobilization of the reagent molecules (e.g. electrostatic binding, adsorption, covalent binding, etc.) in an appropriate matrix is a major factor. Consideration must also be given to the ultimate transfer of the immobilization method to mass production. The example given below in Figure 6.14 shows a membrane-based optrode system using a ligand (L) to extract the analyte ion (M^+) into the membrane to form a ligand-analyte complex (ML^+). In the presence of this ML^+ complex, and in order to maintain charge balance, the dye loses a proton and changes color from red to blue. Membranes doped with compounds that are selective for an ion of interest are very popular. The two membranes used in Figure 6.15 were formulated using calix[4]arene compounds known to have cavities of the correct size for sodium and calcium, i.e. tetraethylestercalix[4]arene and tetraphosphineoxidecalix[4]arene respectively (the structures of these compounds are given in Figure 6.5). The sodium membrane exhibited an excellent spectral response to changing concentrations (over five orders of magnitude) in sodium and no response to calcium, and the calcium membrane exhibited the same phenomenon showing little interference from sodium.

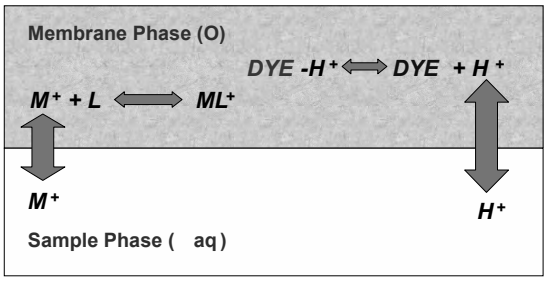


Figure 6.14. *Schematic diagram of a membrane-based optrode system*

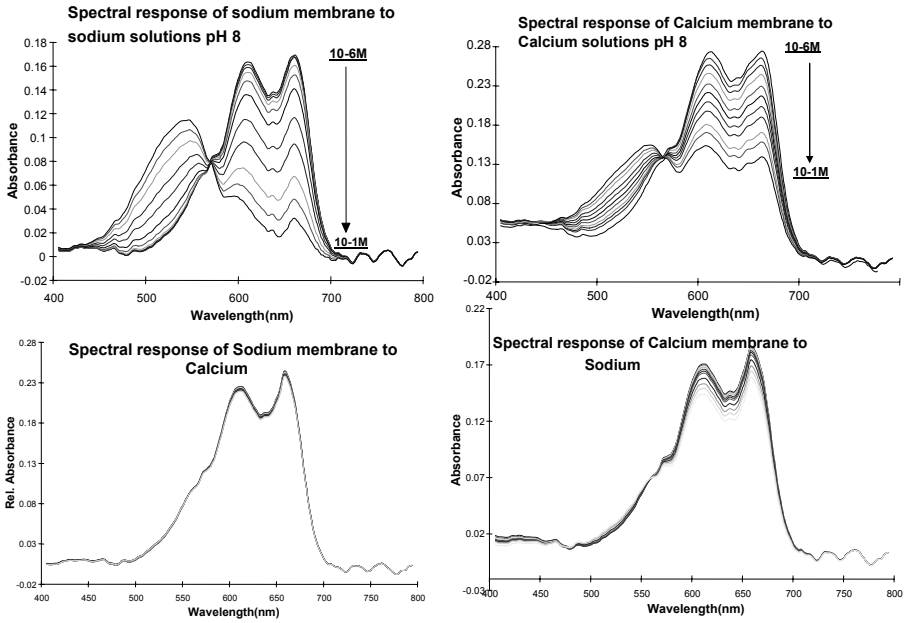


Figure 6.15. *Spectral responses of a sodium membrane and a calcium membrane to both sodium and calcium ions, demonstrating excellent selectivity in both cases. The sodium membrane contained tetraethylestercalix[4]arene and the calcium membrane contained tetra phosphine oxide calix[4]arene. The structures of these compounds are shown in Figure 6.5*

6.5. Acoustic (mass) sensors

In 1880, Jacques and Pierre Curie [8] showed that when anisotropic crystals were compressed in particular directions, a potential difference, or voltage, was produced between the deformed surfaces and this voltage was found to be proportional to the

force applied. The converse effect is also true. Therefore, when a voltage is applied across an anisotropic crystal, such as quartz, it will induce an acoustic wave, which will cause the crystal lattice to move. This acoustic wave will match that of the fundamental frequency, or harmonics, of the crystal and is dependent on its mass. These sensors can thus be called acoustic or mass sensors. In 1959, it was proposed that if a coating of uniform distribution and comparable density was attached to a quartz crystal, a change in the fundamental frequency would result and this change would be proportional to the coating mass. It was proposed that if the polymer film rigidly attached to the quartz surface was a gas chromatography (GC) stationary phase then the polymer could somewhat selectively absorb certain volatile organic solvents. The absorption of these solvents would change in the mass of the polymer, which would be then seen as a change in the fundamental frequency of the quartz to which it was bound. GC coatings were the first coatings used due to their excellent sorption properties and the pre-existing knowledge of their applications.

When the acoustic wave travels through the bulk of the crystal, the sensor is called a bulk acoustic wave (BAW) device and when the wave travels along the surface of the crystal, it is called a surface acoustic wave (SAW) device.

6.5.1. Quartz crystal microbalance sensors

Figure 6.16 shows a typical quartz crystal microbalance (QCM), which is a specially cut quartz wafer supported between two gold electrodes which generates a bulk acoustic wave through the crystal.

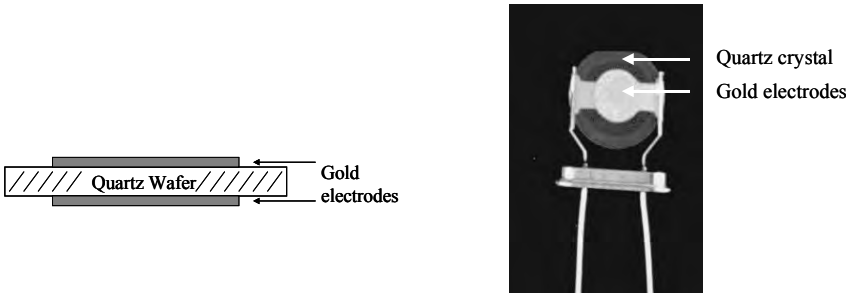


Figure 6.16. *(a) Schematic of a QCM (b) Photo of bare QCM*

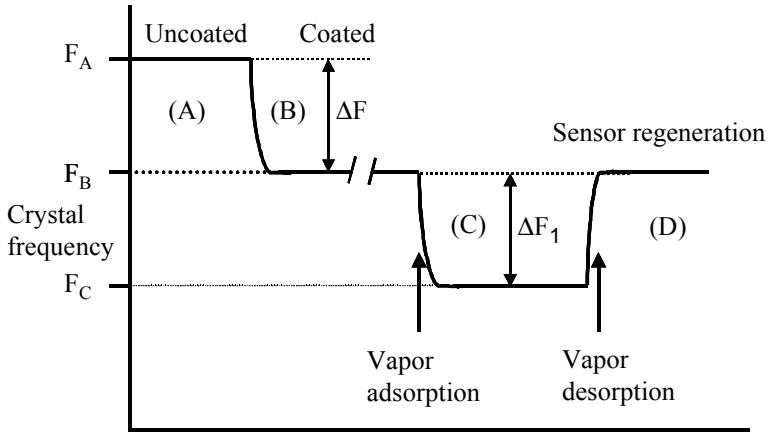


Figure 6.17. *Frequency profile of QCM during coating and sampling processes*

If an uncoated QCM (A) with a fundamental frequency of FA is coated with a polymer (B), the fundamental frequency of the QCM will drop to FB (Figure 6.17). When this coated QCM is then exposed to a vapor (C), some of the vapor will be absorbed into the polymer coating and will change the mass of the crystal, and hence its fundamental frequency**.**

Ideal coatings for QCM applications should be non-volatile so that it stays on the crystal surface and allows rapid and easy diffusion of vapors into and out of the material. The material should be stable over prolonged use and not undergo any hysteresis effects. Amorphous polymers satisfy these conditions. Elastomeric polymers are the best option as they will move on the vibrating crystal surface without any loss of energy (thus avoiding possible viscoelastic problems). Examples of such polymers would be poly(isobutylene) or poly(ethylene) glycol, both common GC stationary phases. The oscillating frequency of these sensors is based on the Sauerbrey equation [9], but there are limits on the coating material. The coating should be thin (preferably < 100nm) and of uniform thickness and density. This coating must also oscillate in perfect synchronization with the quartz surface.

QCM sensors are fast, cheap, reliable sensors, which can be used for various applications ranging from solvent detection to food quality analysis. They are inexpensive since they are mass produced for oscillator circuits, and their lifetime is approximately one year. They have good linearity and can be incorporated into an array and used to build up a library of characteristic responses for individual analytes, allowing for rapid and reliable identification of unknown compounds. They

can be used for the early warning detection of possible contaminants in many industrial processes such as beer and tobacco production.

6.5.2. *Sensor arrays*

There is no such thing as an ideal chemical sensor. However, there are sensors that are highly selective towards certain species, but such sensors tend to have reversibility problems. Nevertheless, incorporating different sensors possessing moderate selectivities into an array offers certain advantages. Such an array can be used to monitor different molecular interactions of a single sample. The success of such an approach is dependent on the type and quantity of information that each sensor gives. Therefore, to optimize such a system, it is necessary to include sensors that would be particularly selective to different classes of analytes. Therefore, using an array of sensors with different chemical properties will allow a specific pattern or fingerprint response for each solvent to be generated (see Figure 6.18). This fingerprint could be assigned to a particular analyte and a library of such responses could be built up.

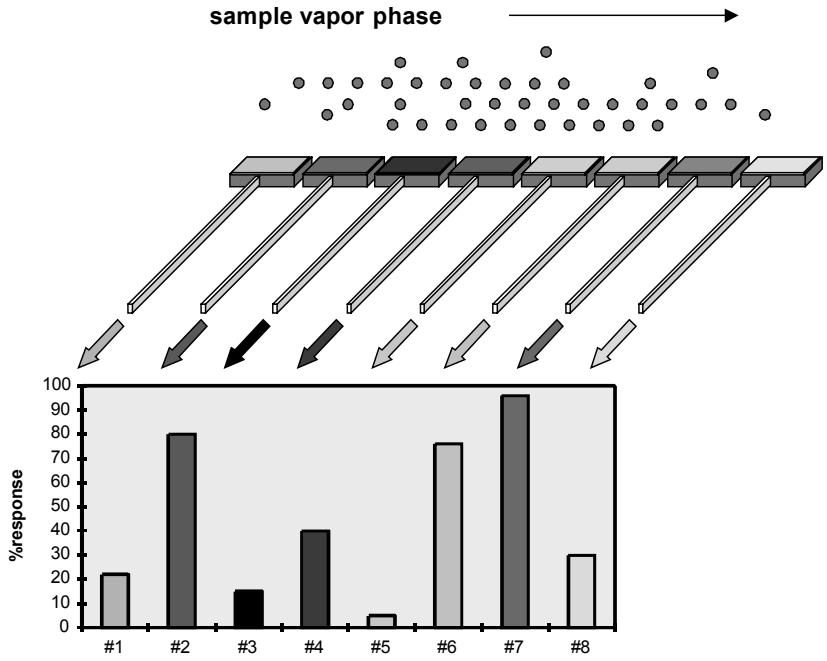


Figure 6.18. *Patterns obtained from a sensor array are characteristic of sample components*

Using these fingerprints, it will be possible to discriminate between these solvents using multivariate techniques such as principal component analysis (PCA) and discriminant function analysis (DFA). PCA is a method that is mainly used for data reduction as it can maintain the total variance within the data set while reducing the number of variables used. PCA is an unsupervised method that changes the original data set into a relatively smaller and simpler one that can still adequately give the same amount of information, by removing redundant variables. Of course some information is lost but the improvement in ease of use of the information compensates for this. From the original data set the first principal component (PC1) is calculated which contains most of the information of the variance of the data set. The second principal component (PC2), being orthogonal to the first, then contains information on the remaining variances and so on for subsequent principal components. To distinguish between known groups, DFA is used. DFA devises a *classification rule* that will allocate subsequent unknown individuals to these pre-assigned groups with the lowest error rate. DFA aims to maximize separation between groups of available individuals, while minimizing the misclassification rate over all possible future allocations. In a similar fashion to PCA, discriminant functions are generated, e.g. DF1-DF5, which contain information about the groups assigned. DF1 would contain most of the information relating to the grouping while the remaining information would be then sequentially distributed between the remaining factors in decreasing amounts. An example of an application would be the discrimination and identification of organic vapors, as can be seen in the following plot (Figure 6.19).

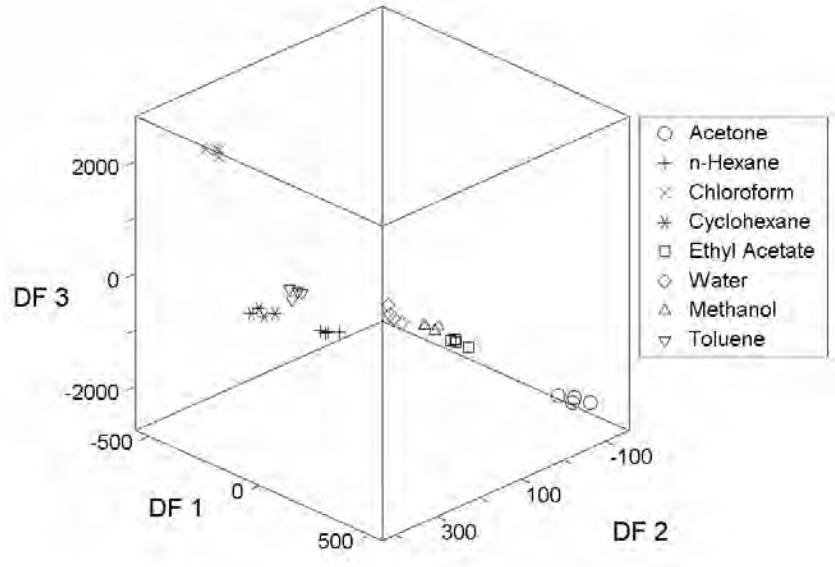


Figure 6.19. *A 3-D plot of DFA of selected solvents*

For a sensor to be reusable, the vapor-coating interactions must be reversible. Practically, the interactions between the vapor and polymer should be weak to promote fast adsorption and desorption. Weak interactions however tend to have rather poor selectivity, and so a range of sensors, each with a moderate selectivity, has to be incorporated into an array in order to generate analyte-selective patterns. For example, in such an array there could be a sensor that would be more responsive to alcohols, while another could be more responsive to aromatics.

6.6. Biosensors

Biosensors are devices incorporating a biological sensing element coupled to a variety of transducers such as electrochemical, optical, piezoelectric and colorimetric. The purpose of this is to convert a biological interaction into an easily measurable electrical signal. Research into biosensors has grown enormously over the past few decades. The specific and sometimes reversible interaction of biological components can produce changes in a variety of physical quantities that can be detected by these transduction methods. Such changes are the release or consumption of ions, electrons, changes in optical properties, mass, or the release or consumption of energy in the form of heat. Applications of biosensors have ranged from environmental monitoring to medical applications, for example, in the detection of coumarins and mycotoxins. The ideal characteristics of the biological component are that it is highly selective, stable, should not contaminate the sample being tested and should retain its biological activity when immobilized. The specificity of the biosensor is totally dependent on the properties of the biological component.

Biosensors may be characterized in a variety of ways. One such way is on the basis of the biological component used. In this way, there are two main categories: affinity biosensors and catalytic biosensors. Affinity biosensors are based on specific binding interactions. These use antibodies, cell receptors, nucleic acids and lectins, or their binding counterparts. The catalytic biosensor works on the principle that the analyte becomes altered in some way. All such biosensors are ultimately based on enzyme catalysis, but may employ purified enzymes, plant and animal tissues or whole cells. The use of biomolecules and biological systems in biosensors was reviewed by McCormack *et al.* [10]. Recently, commercial biosensors have been developed using optical transduction technologies. In particular, the BIAcore® (BIA is biochemical interaction analysis) instrument has become a very important tool for the detection of a number of analytes.

6.6.1. Affinity biosensors

Several biosensors have been developed based on antibody-antigen interactions. In principle, antibodies can be generated against any target (e.g. monoclonal antibodies) or against the recognition fragment only (receptobodies). Macromolecules can elicit an immune response alone while low molecular weight compounds must be bound to a macromolecule (usually a protein) before immunization. Antibodies are generally very selective, exhibiting little cross-reaction, and sensitive but many of them are unsuitable for continuous monitoring.

There are a variety of transduction modes for antibody-based sensors. Electrochemical, fluorimetric and photometric modes are very popular but require a labeled version of the analyte or the receptor for competitive assay. A more recent transduction mode is SPR which is an attractive option since no label is required. It is a direct method, currently expensive and lab-based but portable instruments are beginning to appear in other works.

6.6.1.1. *Electrochemical transduction*

It has been shown that immunoassays can be exploited by biosensors in much the same way as they are in immunoanalysis, opening up a very wide range of analytes that can be studied. Tests such as the ELISA use antibodies and enzymes in combination to bring about the optical recognition of the antibody-antigen interaction. The same principles can be used in electrochemical immunoassays by substituting a colorimetric substrate for an electrochemical redox mediator. However, immunosensors have been poorly exploited due to the complexities of reaction schemes required as reacted and unreacted species must be separated from one another and non-specific interactions excluded by washing, etc.

The novel application of biomolecular films has been used to bring about immunosensors with much greater simplicity. A gold-coated microporous nylon membrane has been used as an electrode on which to immobilize immunoreactants labeled with the enzyme horseradish peroxidase (HRP) (Figure 6.20). The mediator, hydrogen peroxide, was applied from the opposite side of the membrane to the immunoreactants. As it passed through the membrane, reaction with immobilized materials would occur first and little substrate would diffuse into the bulk solution to interact with unbound material.

Other researchers have exploited electrodes modified with conducting polymers to enhance similar separation-free principles. Conducting polymers such as polyaniline show some useful properties for immunosensors.

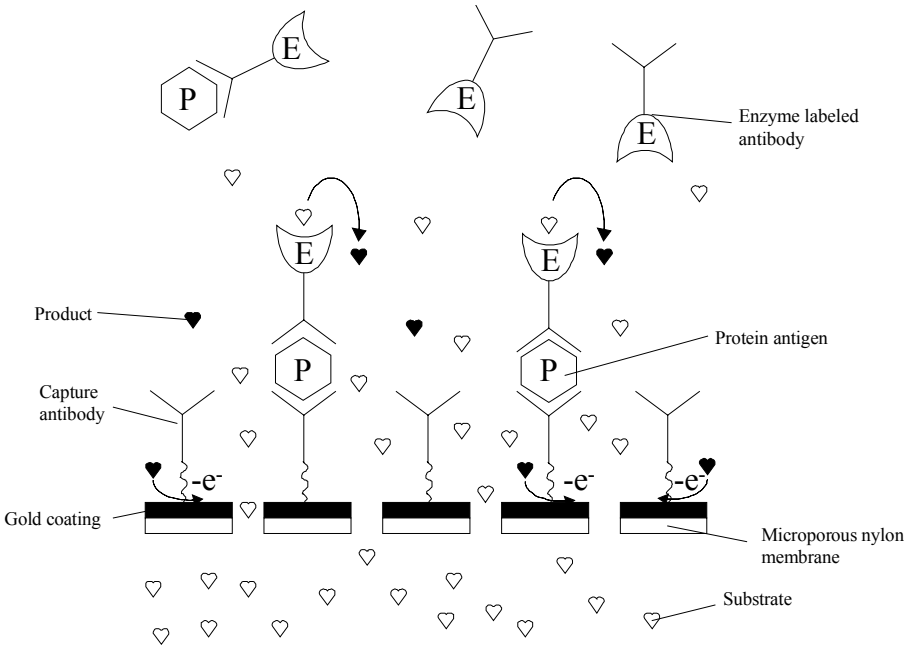


Figure 6.20. *A separation-free immunosensor principle based on microporous membranes. A gold layer was applied to microporous nylon for the immobilization of antibodies. An immunoassay was set up on one side of the membrane and substrate was introduced from the opposite side. The substrate was immediately consumed only by those enzyme labels taking part in the analytical measurement, and not by those remaining free in the bulk solution*

They can be used as the point of attachment of antibodies or antigens and can also be used to shuttle electrons from the electrode to the enzyme active site without the need for a soluble mediator. In addition, enzymes present in the bulk solution are not coupled catalytically to the electrode surface, so only immunological binding interactions are detected.

6.6.1.2. *Piezoelectric transduction*

The piezoelectric (PE) effect is based on the principle that certain crystalline materials – most notably quartz – oscillate at a certain frequency when stimulated with a voltage. This principle serves as the basis for two distinct types of biosensor. These are the resonant QCM and the SAW device. The type of device depends mostly on the configuration of the quartz crystal.

QCMs use a thinly cut (200 μm) quartz crystal disk which is sandwiched between two metal electrodes. When a voltage is applied, the crystal becomes deformed. Removal of the voltage allows the crystal to oscillate at its natural frequency. Crystals have resonant frequencies typically in the order of 10 MHz. However, the addition of any mass bound to the surface of the crystal when it is in air will result in a change in frequency according to the Sauerbrey equation [9]. Thus, decreasing the width of the crystal and so increasing its natural frequency will increase the sensitivity of the device. With SAW-based devices, vibrational waves travel across the planar surface of a thin crystal between a transmitter and receiver. These devices have typical frequencies of 100-200 MHz.

Whichever type of PE transducer is chosen, they all act principally as mass balances (i.e. they respond directly to the change in mass at their surface). For this reason, PE devices are not suited to metabolism biosensors but find great application as affinity biosensors, especially with antibody-based devices. Here, mass changes are brought about by the interaction of antibodies and antigens on the transducer surface. Either species can be immobilized, but it is typically the antibody as the free antigen that can then be measured. Most heterogenous immunoassay formats are applicable to this system. However, those techniques which bring about the greatest mass changes will prove most suitable. PE biosensors are thus capable of "direct" detection of biomolecular interactions. Various immobilization methods have been employed. The crystals are typically pre-coated with compounds such as polyethyleneimine and polyacrylamide to allow the cross-linking of the biological component. Other techniques have been to coat with protein A first and then attach the antibody.

The characteristics of PE devices make them predisposed to certain applications. The frequency of these crystals is affected by surface contact with liquids, being dependent on viscosity, density and temperature, and so work in the liquid phase has been limited. Much work has been performed with gas-phase devices. A dip-and-dry method is often the most widely employed technique where the sensor is immersed in the analyte, dried and measured. As the PE test is based on mass changes, it is particularly suited to the analysis of large antigens such as whole cells, and detectors for *Salmonella typhimurium* and *Candida albicans* have been established with this technique. The problems associated with liquid phase assays have not prevented researchers from working in this area. Some have developed techniques to compensate for the effect of liquids, successfully monitored antibody-antigen interactions in flow cells and have even followed the kinetics of the interaction in a liquid medium.

6.6.1.3. *SPR biosensors*

Several biosensors have been developed based on the phenomenon of SPR which allows the detection of biomolecular interactions in "real time". The principle behind SPR is described in the context of BIAcore for convenience. At an interface between two media of different refractive indices (e.g. glass and water), light coming from the side of the higher refractive index is partly reflected and refracted. Above a certain critical angle of incidence, the light is totally internally reflected and no light is refracted across the interface between the two surfaces of different refractive index – see Figure 6.13.

Under total internal reflection (TIR) conditions, an electromagnetic field component called the evanescent wave penetrates into the medium of lower refractive index a short distance in the order of one wavelength. As the evanescent wave moves further away from the interface into the lower dense medium, the wave decays exponentially. If the interface between the media is coated with a thin layer of metal (in the case of BIAcore, this metal layer is gold) containing electron clouds at the surface and the passage of the evanescent wave through this metal layer causes the plasmons to resonate, this results in a quantum mechanical wave known as a surface plasmon. Some of the energy of the reflected light (incident light) is taken up by the surface plasmon wave, resulting in a dip in the intensity of reflected light at a certain angle being observed. The incident light angle at which this dip is observed is known as the SPR angle. The SPR angle is dependent on a number of factors. These factors include the properties of the metal film (e.g. thickness, uniformity and composition), the wavelength and polarization of incident light and the refractive index of the media on either side of the metal film. In real-time analysis, the properties of the metal film, the wavelength and refractive index of the denser medium are kept constant. The SPR signal can be used to monitor the refractive index of the aqueous layer immediately adjacent to the gold metal layer. The light source in the BIAcore instrument is a high-efficient light emitting diode with a wavelength in the near IR region. This light is focused on an interface consisting of glass and gold on the sensor chip in a wedge shaped beam, producing a fixed range of incident angles. A two-dimensional diode array is used to monitor the reflected light. Changes in the refractive index are a direct result of changes in the mass or concentration at the surface of the chip and this characteristic of SPR has been used to monitor biological interactions. Figure 6.21 shows the operation of SPR in the BIAcore biosensor instrument.

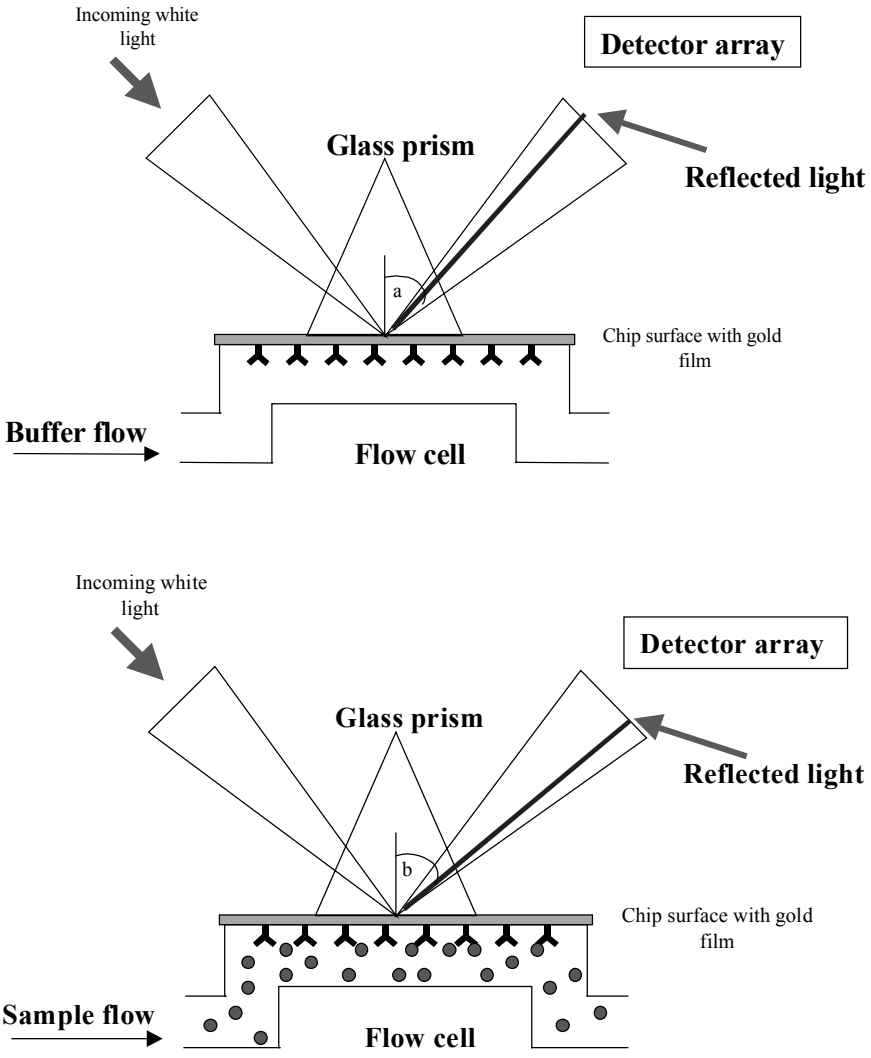


Figure 6.21. *These diagrams (i) and (ii) show the operation of SPR in the BIAcore biosensor instrument. See text for explanation*

Antibody (Y) is immobilized onto the surface of the chip using conventional carbodiimide coupling chemistry. Light from a high intensity light emitting diode (LED) is focused onto the gold chip surface by means of a prism under conditions of TIR. A two-dimensional photo-diode array was used to measure the reflected light.

Under conditions of TIR at a metal-coated interface, an evanescent wave propagates into the medium of lower refractive index on the non-illuminated side. This leads to a dip in the intensity of reflected light at a particular angle known as the SPR angle (SPR angle = angle a) and this is shown as the left line in Figure 6.21(i). This SPR angle is sensitive to a variety of factors such as the refractive index at the gold film side of the interface and any biospecific interactions. Figure 6.21(ii) shows the injection of antigen (which, from an analytical point of view, is the analyte) over a chip surface immobilized with antibody Y. The specific interaction of antigen binding to the antibody causes an increase in the mass bound at the sensor chip surface resulting in changes in the refractive index. This change in the refractive index is responsible for a shift in the resonant angle of the reflected light as the angle changes from a to b. As changes in the resonant angle are a direct result of changes in the mass or concentration on the surface of the chip, the mass of analyte binding to the chip surface may be determined. The change in reflected light is interpolated as a sensorgram (Figure 6.22).

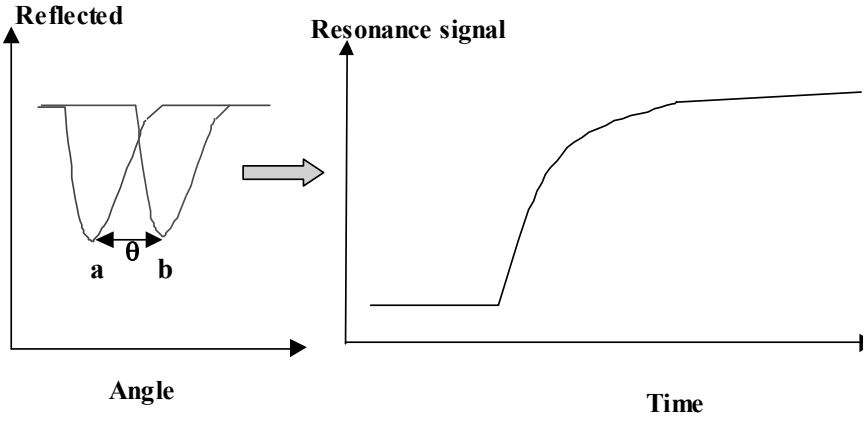


Figure 6.22. *The sensorgram displays the changes in resonant angle (due to antibody-antigen interaction) in terms of the graph shown*

The SPR angle may be seen as a dip in the intensity of the reflected light. As an antibody-antigen interaction occurs, changes in the resonant angle (θ) are monitored continuously and displayed as a sensorgram. Biomolecular interactions at the chip surface cause a mass change and an increase in the SPR angle which is seen as a gradual increase in the signal of the sensorgram. The signal is interpolated into response units (RU) by the instrument software. An increase of 1,000 RU as a result of a biomolecular interaction corresponds to approximately 1 ng/mm^2 of protein present on the chip surface.

At the heart of the BIAcore instrument, as with the majority of SPR-based commercially available instruments, is the sensor chip as shown in Figure 6.23.

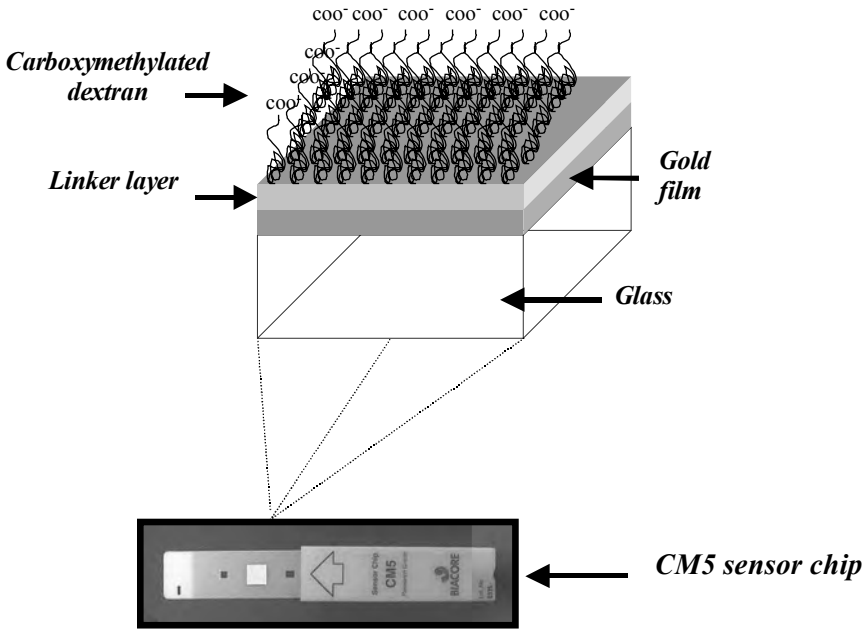


Figure 6.23. *Diagram representing the surface of the BIAcore carboxymethylated 5 (CM5) sensor chip. The surface of the chip consists of three layers: glass, a thin gold film and a carboxymethylated dextran layer. The carboxymethylated dextran matrix allows the covalent immobilization of analytes onto the surface of the chip*

The BIAcore sensor chip consists of a glass slide with a thin layer of gold deposited on one side. Gold was chosen, as it possesses the characteristics of chemical inertness and good SPR response. This gold layer was in turn covered with a covalently bound carboxymethylated dextran matrix attached by a hydroxyalkyl thiol linker layer. The matrix allows the covalent immobilization of analytes onto the surface of the chip and increases sensitivity by increasing the binding capacity of the surface. It also provides a hydrophilic environment with very low non-specific binding. The matrix forms one wall of a micro-flow cell where interactions are monitored. The carboxymethylated side of the chip comes into contact with the solution of interest, while the gold side of the chip is illuminated from the other side, through the glass. SPR is generated through the interaction of the light energy with the gold film and this is used to monitor concentrations of analyte on the surface of

the chip. There are currently a wide variety of different sensor chips available for the BIAcore instrument for varying purposes. However, all chips use the same optical principle.

A typical sensorgram for binding of analyte to immobilized ligand is shown in Figure 6.24. Running buffer flowing over the surface at a fixed volume per min produces the initial level baseline at the sensor surface. A sharp rise in signal is then observed at the precise moment of the injection of sample over the surface.

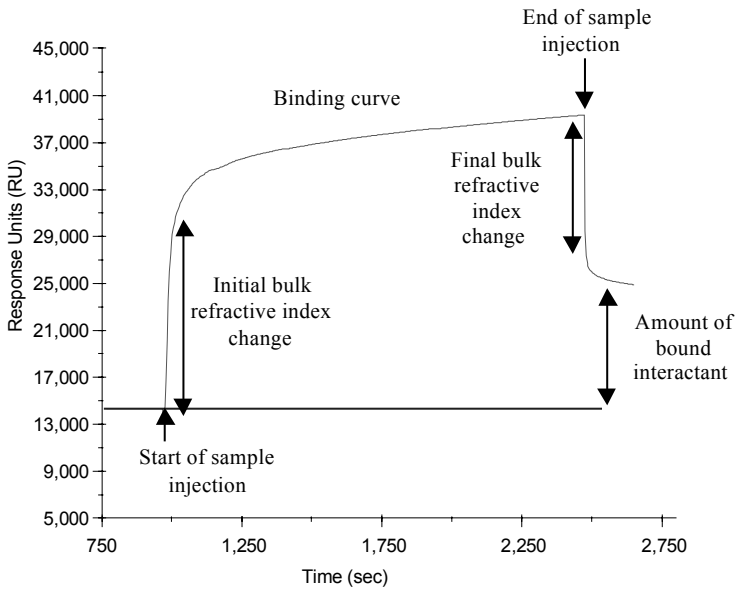


Figure 6.24. *Schematic sensorgram illustrating the binding of analyte to immobilized ligand*

This is due to a difference in the bulk refractive index between the running buffer and the sample buffer. The binding can be seen as an initial jump or initial bulk refractive index change in the signal followed by a steady increase in signal over the course of the injection. This is because large quantities of analyte bind followed by the sensor surface slowly becoming saturated, resulting in the steady signal. At the end of the sample injection, a sharp decrease in the signal is observed as a running buffer of lower refractive index is again allowed to pass over the surface. The difference in the recorded RU between the signal after the injection and the initial baseline RU corresponds to the amount of interactant that remains bound to the sensor chip surface. A change in response unit signal of 1,000 RU corresponds to a change in surface concentration on the sensor chip of about 1 ng/mm^2.

The BIAcore biosensor has a number of major advantages compared with conventional detection techniques. The absence of a need for labeling is of major significance as it minimizes interference with the binding interaction being studied and also eliminates expensive and time-consuming purifications in many situations. Real time analysis of interactants is also one of the main advantages of the BIAcore instrument as interactions may now be monitored as they occur providing valuable diagnostic information as well as kinetic data. Further advantages include the ability of the chip to be reused a significantly large (up to 1,000) number of times and also the ability of the instrument to allow rapid and automated analysis. The BIAcore instrument is provided as a complete system, which includes processing unit, sensor chip, controlling computer and the appropriate software for data collection and analysis.

6.6.1.4. *Proteomics*

Proteomics is an area of research devoted to determining the function of the encoded protein repertoire to find out how it regulates the behavior of individual cells and, ultimately, the whole animal throughout its development both in health and disease. In its infancy, proteomics was the word used to describe technology that allowed large-scale protein separation and mass identification. However, recently proteomics has divided into three areas, protein profiling, interaction analysis and structural genomics. Characteristics of the BIAcore instrument are real time monitoring, an auto injection system and the ability to analyze large numbers of samples while using small amounts of material. These provide a platform for analysis in the area of proteomics [11]. Generally 2D gel electrophoresis is used to separate and analyze 5,000 proteins simultaneously. Identification of the proteins is carried out by cutting the proteins from the gel followed by digestion with trypsin into fragments. The fragments are then identified by comparing them with databases of protein or DNA sequences with the use of mass spectrometry (MS). However, this process is very time-consuming and this is why new quicker methods of identification are being developed. Sonksen *et al.* [12] developed a method which combined BIA technologies with MS. Affinity-bound molecules were recovered from the surface of the BIAcore chip in a few microliters, ready for MS analysis. BIAcore provides information on the concentration of protein bound to the surface of the chip while MS reveals the identity of the compound by molecular mass determination. Calculation of the total surface molar concentration of affinity-bound molecules was possible by combining the information provided by the two instruments.

6.6.1.5. *IAsys biosensor*

The IAsys (IA means immunoaffinity) series of optical biosensors are analytical instruments that utilize advanced resonant mirror optical biosensor technology and apply it to the recognition of biomolecules. A dielectric layer of high refractive

index is used instead of the gold sensing layer in BIAcore. The sensor consists of a glass prism coated with a thin layer of silica and high refractive index resonant layer (e.g. titanium), which is in contact with the sample solution (see Figure 6.25). Polarized laser light is directed at the prism and under conditions of TIR illuminates the underside of the sensor surface at angles greater than the critical angle. At one angle, which is known as the resonant angle, a component of light can couple through the low refractive index layer and propagate along the high refractive index layer before being coupled back into the prism. During resonance a phase shift occurs between the reflected electric and magnetic modes, resulting in a phase shift in the measured response which is observed as constructive interference and can be measured in "real time". The resonant angle at which coupling occurs is essentially dependent on the refractive index at the surface of the sensor. As a result, changes in the refractive index or mass will change the resonant angle corresponding to signal increases as mass increases and signal decreases as mass decreases.

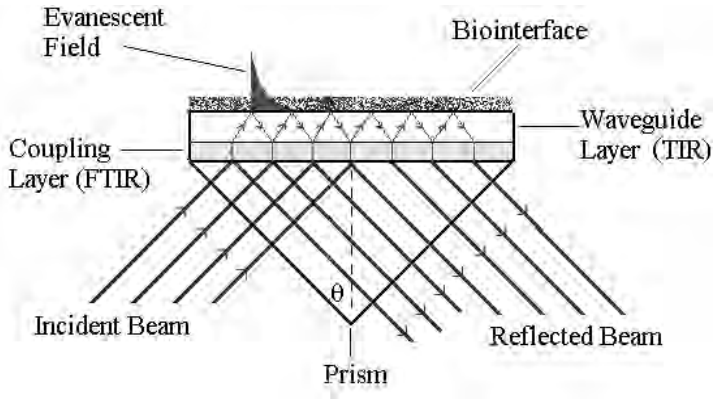


Figure 6.25. *Illustration of the resonant mirror configuration*

6.6.1.6. *Miniature TI-SPR sensor*

The miniature TI-SPR device was first released in 1996 by Texas Instruments, and consists of an LED, a polarizer, a thermistor allowing correction due to temperature changes and two 128 silicon photo diode arrays (Figure 6.26).

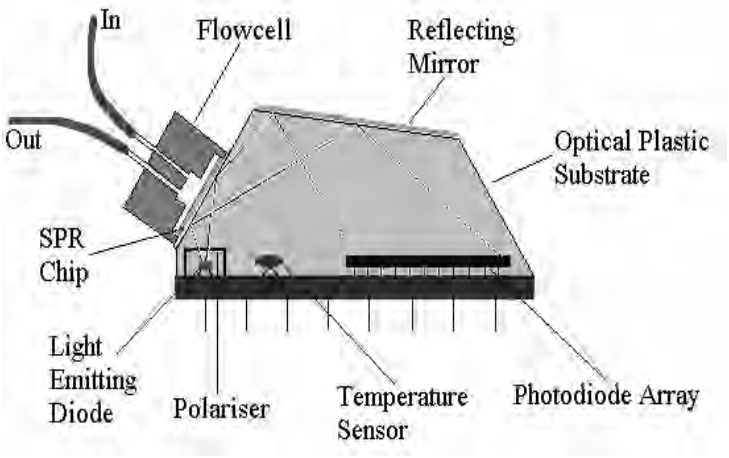


Figure 6.26. *Cross-section of miniature TI-SPR instrument*

These components are mounted on a single platform using conventional semi-conductor-based opto-electronic manufacturing techniques. The platform is encapsulated in an epoxy resin molding structure. Changes in mass are related to changes in the resonant angle and this change is measured by the photodiode array. The width of light produced by the light emitting diode is controlled by the polarizer and reduces the emission of transverse electric radiation. Under conditions of TIR the wedged shaped beam is directed onto a linear photo diode array by a mirror. An SPR-induced minimum is determined by processing the signal from the photo diode array in real time using dedicated signal software [13]. Temperature fluctuations can also be corrected during analysis as there is a built in temperature sensor in the device. This system is available in a hand-held format and is known as the Spreeta^{TM} device which possesses a control box housing the components of the instrument, which may be attached to a laptop to produce a portable device.

6.6.2. *Catalytic biosensors*

Enzymes are proteins that act as biological catalysts. They play very many essential roles in the functioning of all biological metabolism. As illustrated below, enzymes (E) transform a substrate (S) into a product (P) by forming a stable, short-lived enzyme-substrate intermediate (ES):

$$E + S \rightarrow ES \rightarrow E + P$$

Enzymes and their associated reactions have several properties that are extremely useful when used in conjunction with biosensors. The substrates consumed or the products produced can be detected in a variety of ways, e.g. colorimetrically, calorimetrically or electrochemically. Such changes can be coupled to biosensor transducers. In addition, enzymes are involved in the transformation of many clinically or industrially relevant substances such as glucose, lactate, creatinine etc. Using enzymes that transform such species is the basis of the largest group of biosensors in use. Enzymes can also be stably attached to transducer surfaces and can be stored for reasonable periods.

6.6.2.1. *Electrochemical transduction*

At present, the only biosensor systems to make it to widespread commercial application have been electrochemical enzyme biosensors, mainly in the guise of the glucose electrode. The principle was first illustrated by Clark and Lyons [14] in 1962. Their approach, which utilized an electrochemical transducer in combination with a bioselective layer containing an enzyme, is still the most widely studied and used configuration of biosensor. Their system was based on the oxygen electrode. This consisted of a platinum cathode, a silver anode and an oxygen-permeable membrane. To measure oxygen at the electrode passing through the membrane, the oxygen was reduced at a potential of –0.7 V to form hydrogen peroxide:

$$O_2 + 2\ e^- + 2\ H^+ \rightarrow H_2O_2$$

The current produced is proportional to the amount of oxygen reduced. Clark and Lyons then used an enzyme that required oxygen to bring about the oxidation of a biochemical molecule. Such enzymes are referred to as oxidases. They chose glucose oxidase, which oxidizes β-D-glucose to gluconic acid and hydrogen peroxide:

$$\beta\text{-D-Glucose} + O_2 + H_2O \xrightarrow{\text{Glucose oxidase}} \text{Gluconic acid} + H_2O_2$$

In this way, oxygen acts as the electron donor and is consumed in the reaction, itself being reduced to hydrogen peroxide. When the enzyme layer is held in intimate association with the platinum electrode surface (Figure 6.27), a decrease in oxygen concentration occurs at the electrode. The remaining oxygen diffuses through the Teflon membrane and is reduced at the electrode and measured. The

enzyme is held in place by a second membrane composed of cellophane, which allows the diffusion of glucose.

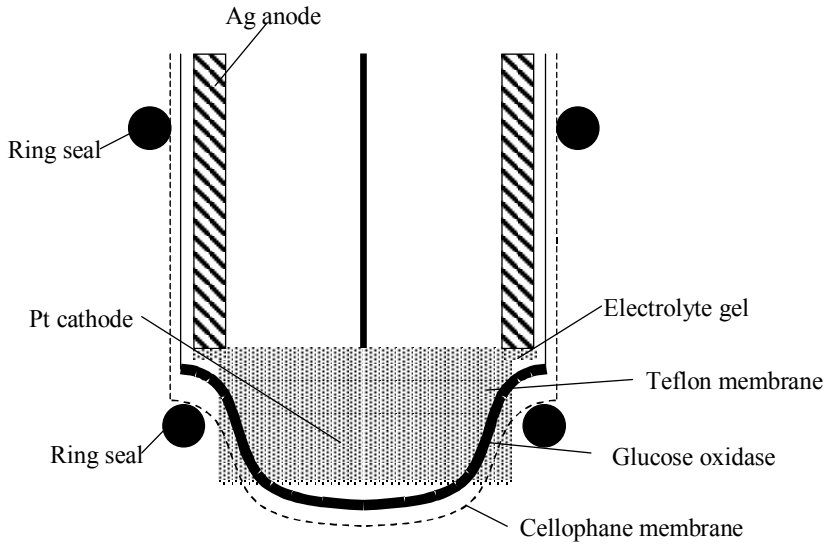


Figure 6.27. *The glucose biosensor of Clark and Lyons [14]. An oxygen electrode composed of a platinum cathode and a silver anode is coated with a gel electrolyte and a layer of glucose oxidase which is held in place by a cellophane membrane. Both β-D-glucose and oxygen can diffuse through the membrane to the enzyme layer where they undergo conversion to gluconolactone and hydrogen peroxide. The consumption of oxygen is measured at the platinum electrode*

Alternatively, the production of hydrogen peroxide can also be measured at the surface of an electrode. This configuration has been used in commercialized glucose biosensors such as that used by YSI Inc. (Figure 6.28). Here, two membrane layers act as barriers to the passage of substrate and oxygen. Only hydrogen peroxide comes into contact with the electrode and is measured amperometrically (monitoring a change in current). These methods are the basis of the measurement of blood glucose in many clinical test systems today.

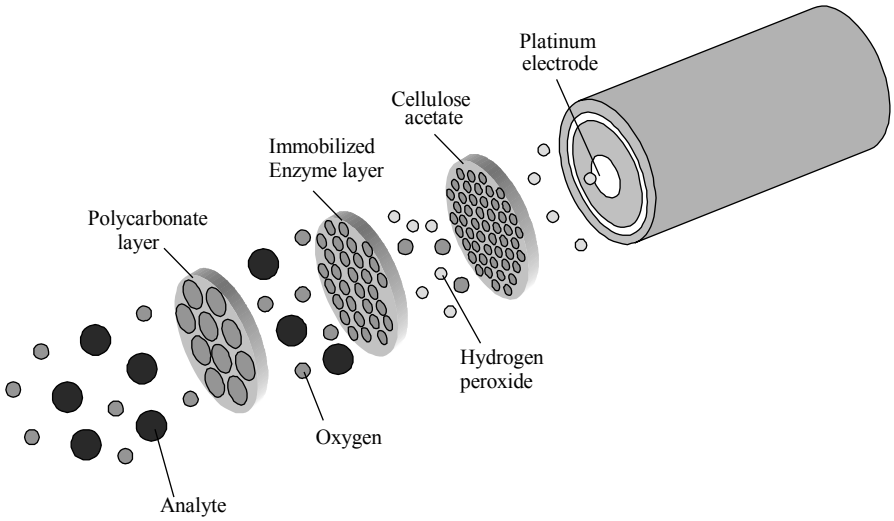


Figure 6.28. *The YSI electrode (from YSI Inc.). Initially, a polycarbonate membrane limits diffusion of the analyte (β-D-glucose) to the immobilized enzyme layer, where it undergoes oxidation. Hydrogen peroxide diffuses to the platinum electrode surface via a cellulose acetate membrane which excludes interferences such as ascorbate*

One way in which these biosensors have become so widely used is due to the low cost and large scale manufacture of these devices. Many enzyme biosensor electrodes are now manufactured using the technique of screen-printing. It involves the controlled deposition of inks onto solid planar surfaces. In the case of biosensors, the inks printed can be composed of many materials. Conductive inks based on carbon, silver, gold and others can be formed by combining fine particulates of the conductor with a suitable binder and solvent. This balance of active material and its various additives gives these inks specific properties such as conductivity, electrochemical activity, curing rate and temperature, as well as important viscometric and rheological properties, which are important during the printing process. Other materials can be incorporated into the inks such as enzymes, mediators, conducting and insulating polymers. Multiple ink layers can be patterned on top of one another to build up more complex electrode structures. Lifescan market a glucose biosensor called the One Touch® Ultra at-home blood glucose monitoring system based on this principle. Their system employs screen-printing processes to bring about immobilization of the active components at an electrode surface (Figure 6.29).

Many examples of screen-printed electrodes abound in the literature. Another example is a biosensor for the detection of breath alcohol based on the enzymatic

reaction of alcohol and nicotinamide adenine dinucleotide (NAD^+) with alcohol dehydrogenase (ADH) [15]:

$$CH_3CH_2OH + NAD^+ \overset{ADH}{\rightleftharpoons} CH_3CHO + NADH + H^+$$

NADH can be oxidized at an electrode at 400 mV vs. Ag/AgCl electrode:

$$NADH \rightleftharpoons NAD^+ + H^+ + 2e^-$$

The alcohol electrode is composed of several layers. A silver conductive layer is followed by carbon for the electrode surface. An insulative dielectric is then deposited which defines the carbon area.

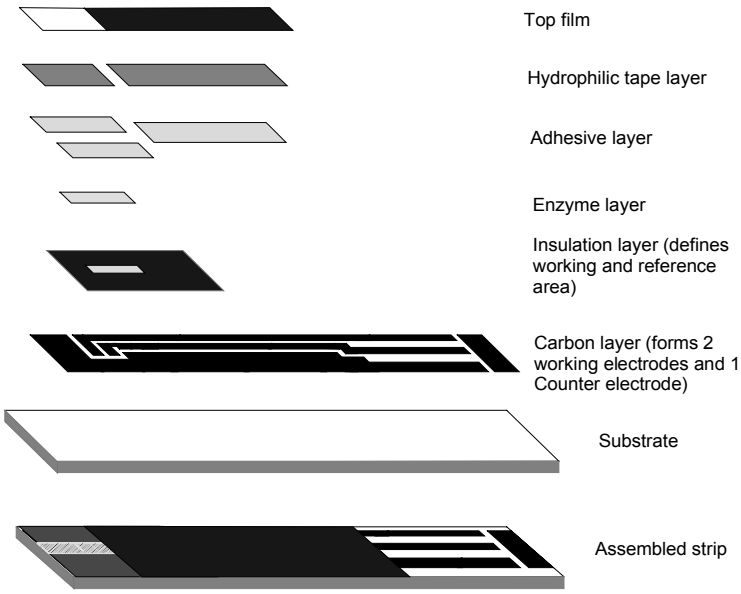


Figure 6.29. *Schematic of the Ultra™ glucose electrode (from Lifescan Inc). The electrode is assembled in several layers on a plastic substrate*

An Ag/AgCl reference is applied to an exposed silver conductive track and two working electrode inks are applied. An active electrode ink containing ADH and NAD+ in DEAE dextran, hydroxyethyl cellulose, ethylene glycol and carbon powder to form a printable paste was applied to working electrode 1 (WE1) and a

similar mixture containing BSA instead of ADH was applied to WE2 to act as a control. Finally, an outer membrane of hydroxyethyl cellulose was deposited. The sensor was capable of detecting ethanol in vapors from 20-800 ppm which is well within the range required for breath tests.

6.6.2.2. *Calorimetric transduction*

Probably the least used transduction method is based on thermometric devices. In these devices, the enthalpy changes that accompany all chemical and biochemical reactions can be measured and be related back to the quantity of analyte being measured. Most biochemical reactions are exothermic. Enzyme reactions typically have an enthalpy of around 80 $kJ.mol^{-1}$. The most widely used device is the enzyme thermistor. A thermistor is a resistor with a high negative temperature coefficient of resistance. They are often composed of ceramic semiconductors doped with various metal oxides. Such devices typically have the ability to resolve temperature differences as little as 10^{-5} K, with signal-to-noise ratios of 100. This gives such devices typical detection limits in the order of 1 μM.

Enzyme thermistors have the significant drawback of being very non-specific. To compensate, reference cells must be used. In addition, they must be carefully isolated from external temperature variations. The enzyme is normally immobilized in a small column. The choice of support in this application is controlled pore glass, to which the enzyme is attached via glutaraldehyde cross-linking. Columns have capacities of up to 1 ml, although miniaturized systems are being developed that require only microliter volumes of sample. Sample passing through the reaction column transfers heat to the thermistor.

6.7. Future trends

Miniaturization of flow-based analysis is today an actively pursued topic in analytical chemical analysis. The lab-on-a-chip concept promises to revolutionize chemical analysis systems. To reduce the cost of environmental analysis and *in vitro* diagnostics (IVDs), environmental and medical device designers are searching for new technologies that will enable them to develop high-quality instruments at a fraction of the cost of current laboratory systems. For many IVD and environmental manufacturers the best hope comes from recent successes with miniaturized sensor devices, which enable analysts to perform sophisticated diagnostic techniques in the field or at a patient's bedside, thereby removing routine testing from environmental laboratories and hospital settings and further reducing the costs of diagnostic testing.

The technologies associated with miniaturized or micro total (chemical) analysis systems (μTAS) are making a major contribution to the miniaturization of

instrumentation. μTAS technologies are already being applied for performing immunological analysis, polymerase chain reaction (PCR), *in vitro* fertilization, glucose sensing, and semen analysis. Continued development of μTAS technologies promises to produce inexpensive, self-contained microfluidic devices with performance and accuracy superior to their larger and expensive laboratory counterparts. In time, these instruments may replace many other time-consuming and less-sensitive laboratory instruments used to identify chemical compounds, biological species, and pathogens in forensic, environmental, clinical and industrial samples [16]. Miniaturized systems are a pathway to enhanced analytical performance and shorter analysis times.

6.7.1. *Microanalytical instruments as sensors*

The concept of micro total analysis systems (μTAS) or lab-on-a-chip has grown rapidly since its origins in the pioneering work of Manz *et al.* in the early 1990s [17]. μTAS involves using micromachining technologies to produce the flow conduits, mixing chambers, sample input and optical flow cells in an integrated planar package (see Figure 6.30). They are usually made from materials such as silicon, glass or plastic. Silicon is a versatile material with some excellent mechanical properties and it can be used to fabricate precise and reliable micro-mechanical microstructures that can be integrated with the microfluidic system.

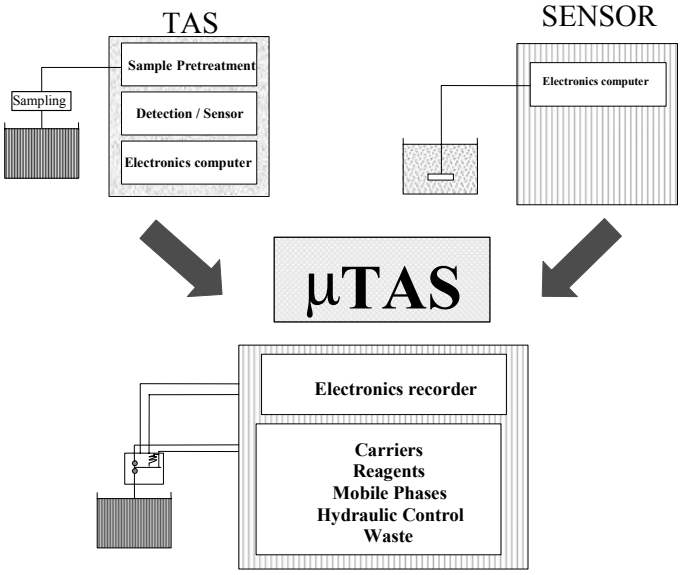


Figure 6.30. *Schematic of a μTAS system*
(courtesy of Dr. Elisabeth Verpoorte, SAMLAB, Neuchâtel, Switzerland)

Silicon is being employed in a variety of miniaturized systems because the active material silicon is inexpensive and abundant, and processing silicon itself is based on very thin deposited films, which are highly amenable to miniaturization. By employing micromachining techniques, it is possible to fabricate mechanical structures that have dimensions on the micron scale rather than millimeter scale, and therefore a large number of identical devices can be made in one batch.

In standard FIA, sample plugs are injected into a flowing carrier stream to which reagent(s) are added at various stages and reaction takes place in a coil before detection. In a coil of reduced cross-section, the flow is no longer turbulent. Laminar flow dominates in sub-10μl systems. Laminar flow can be described as a series of parallel layers, or lamina, moving at different velocities (see Figure 6.31). A flow is said to be laminar if the viscosity is high, the velocity is low and the length scale is small. In reducing the size of the coil in FIA, convective effects are minimized and transport distances shortened, thus making molecular diffusion a more significant transport mechanism. Poiseuille's law also shows the dependence of the volume flowrate on the radius and the length of the tube/channel. Diffusion is the process by which matter is transported from one part of a system to another as a result of random molecular motions. It is generally a product of intermolecular collisions rather than turbulence or bulk transport. Diffusion is the only available mixing option under laminar flow conditions, unless a mixing chamber is specifically introduced. Mass transport by diffusion occurs 100 times faster in a microfluidic device, which has dimensions 10 times smaller than a conventional bench FIA instrument. The mass transport by diffusion can be determined in a fluid system by calculating the Reynolds number which is a dimensionless quantity associated with how smooth the flow of a fluid is.

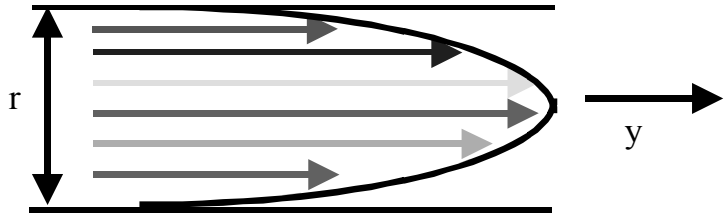


Figure 6.31. *Pressure-driven laminar flow with arrows indicating streamlines*

6.7.1.1. *Design considerations*

Miniaturization enhances a reduction in reagent consumption and waste generation, and increases sample throughput. The key goal of a microsystem is to manipulate and control sample and reagent delivery through miniaturization and

integration of the flow set-up. Three factors to contemplate when considering the design of a microfluidic chip are:

– the chemistry involved in the analysis (number of reagents, mixing channels, kinetics, materials compatibility, etc.);

– the mode of detection (electrochemical, direct optical, reagent-based optical, etc.);

– the physical and chemical characteristics of the material used to make the chip.

An example of a simple microfluidic chip design developed for reagent-based colorimetric detection is shown in Figure 6.32.

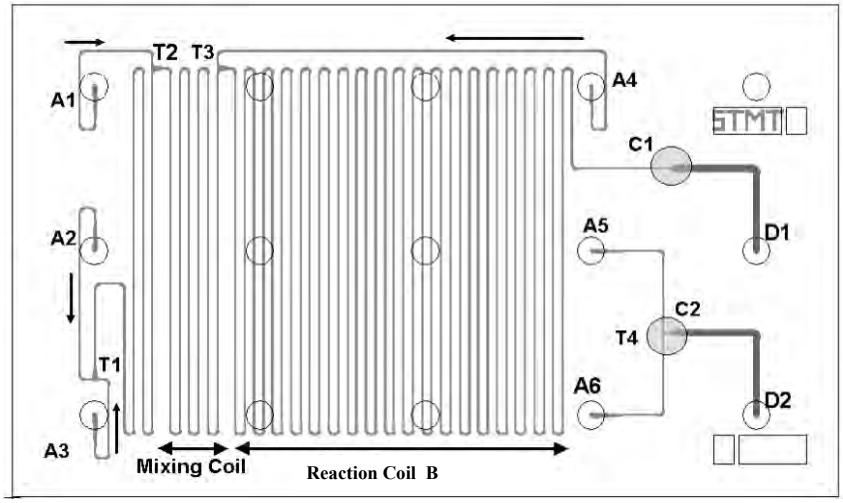


Figure 6.32. *Example of a Simple Chip Layout (courtesy of The Microchem Project, Danfoss A/C, Nordborg, Denmark, also GeSim GmbH, Dresden, Germany). The main system has 4 sample/reagent inlets (A1-A4), mixing T-junctions (T1-T3), reaction coil (B), optical cuvette (C1) and Waste Output (D1). An ancillary system has two inlets (A5, A6), a T-junction (T4), an optical cuvette (C2) and a waste outlet (D2)*

However, fundamental concepts in design and modeling defined on the macroscale may need to be re-assessed to accurately describe microsystem behavior. One such significant difference is dispersion and the role of diffusion as the dominant mixing mechanism in microfluidic systems compared with turbulent flow in macrosystems.

The dispersion of the sample plug increases with the square root of the distance traveled through a tube, decreases with decreasing flow-rate and decreases with increasing temperature rates. Variations in channel width introduce a skew to the flow profile and the deeper the channel, the greater the dispersion. Dispersive behavior in a channel is shown in Figure 6.33. Etch depth of channel plays a prominent role in behavior of dispersion in fluids. Generally it can be shown that decreasing the depth of the channel leads to a lower dispersion of the sample plug, which is a recognized prerequisite in a microsystem. Dispersion is a negative effect, therefore it is desirable to keep dispersion effects small because this in turn increases the number of plugs per unit time that can be generated. To reduce dispersion, the residence time in the tube has to be extended. This is accomplished by curtailing the flow-rate and by minimizing the tube dimensions.

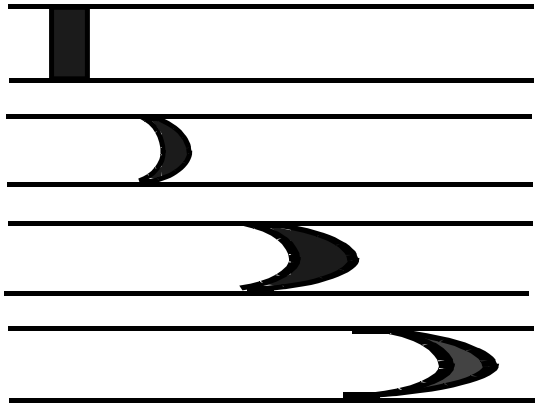


Figure 6.33. *Dispersion of sample plug in laminar flow (from [18])*

To comply with the various considerations that contribute to the optimum microfluidic chip design for a given application, ultimately a balance must be reached between low dispersion, low flow-resistance and short reaction time.

6.7.1.2. *On-chip chromatographic and electrophoretic separations*

In recent years there has been a growing interest in the miniaturization of analytical separation techniques to a scale at which separations can be performed "on-chip" as part of μTAS devices. The application of such devices as chemical and biological sensors has obvious potential. Fundamental benefits such as improved selectivity and the ability to determine analytes within complex matrices go hand in hand with the more practical advantages, such as reduced reagent consumption, shorter analyses times, less waste production and applicability to "*in situ*" and "online" analyses. Of course, as with any new technologies there exist the inevitable

technical hurdles to be overcome. With on-chip analytical separation devices these currently include problems with sample injection, limited range of suitable micro-scale pumps (for liquid chromatographic systems), inflexible detection options and rather poor concentration-based detection limits. For a more detailed evaluation of the "pros and cons" of miniaturized separation systems see Manz *et al.* [17].

Although miniaturized separation systems have experienced most interest in the past five or six years, the actual idea of chromatographic separations within a micro-chip is by no means new. In 1979, a paper was published on on-chip GC. However, compared to GC, the emergence of "on-chip" liquid chromatographic separations has been slow. This is due to several reasons, not least the lack of integratable high-pressure pumps and valves to provide a non-pulsating mobile phase flow at pressures of up to 400 bar. The design of micromachined pumps that fulfill these requirements is an engineering problem which has and is receiving much attention, although as yet few suitable systems are available. However, despite these technical problems, the principles and various modes of on-chip liquid chromatography have been extensively explored. Most research has focused on how to introduce a stationary phase into the narrow channels (typical dimensions: width 5-50 μm, depth 1-10 μm, length 5-15 cm) fabricated within the chip. Approaches taken include (a) slurry packing the channel with standard HPLC grade stationary phases, an approach which utilizes a frit to retain the packing material (see Figure 6.34), (b) coating the surface of the channel, an approach which is easy to implement but often results in a low phase ratio and therefore reduced analyte-stationary phase interaction, (c) *in situ* polymerization of a stationary phase, whereby a porous monolithic polymer is formed and functionalized within the channel itself, (d) *in situ* micromachining of monolithic support structures, an approach which results in a much increased stationary phase surface area with micro-channels of width ~1.5 μm and (e) agglomerated channels, a novel approach whereby functionalized latex particles (~75 nm) are coated onto the surface of the channel, again increasing the stationary phase surface area.

Once the mechanics of the stationary phase support structure have been mastered, it is possible to carry out many modes of on-chip LC depending upon the nature of the stationary phase itself. The mode of on-chip LC that has received most attention is of course reversed phase LC, although size exclusion LC, ion-exchange chromatography and chiral separations have all been achieved.

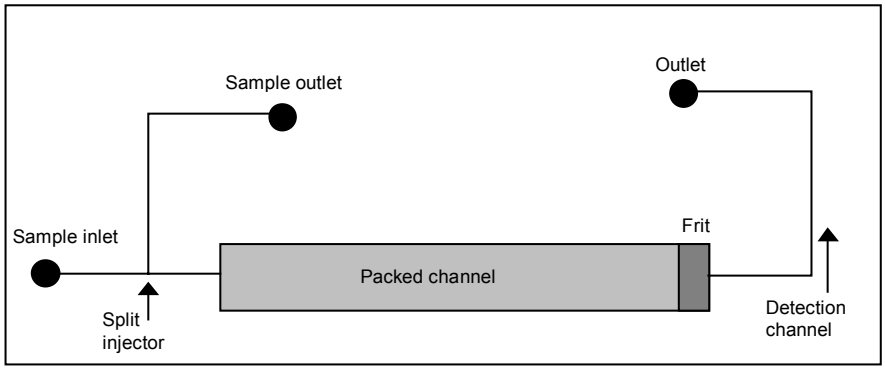


Figure 6.34. *Design for a packed column LC chip*

For many of the reasons mentioned above, on-chip electrophoretic separation techniques have dominated liquid chromatographic methods. The absence of moving parts and a pump and the ease of fabricating micro-electrodes into the chip itself (to provide a separation voltage) mean that electrophoretically-driven devices are, at least in theory, both mechanically and chemically, simpler than pressure driven LC devices. In addition to the above, the separation efficiencies obtainable when utilizing electrophoretically-driven systems are generally at least one order of magnitude greater than obtainable from using pressure-driven flow.

On-chip channel electrophoresis is simply based upon two phenomena. Firstly, the separation of charged species takes place under an applied electric field according to differences in size and charge. The rate of migration of these species under an applied electric field is given as their electrophoretic mobilities. Secondly, within a channel possessing a charged surface (i.e. glass chips) and containing an electrolyte solution, an applied voltage will result in the bulk flow of the electrolyte solution in one direction. This is termed electro-osmotic flow (EOF) and is the liquid driving force behind electrophoretic separations. The combination of these two phenomena results in the separation and migration of charged species along the channel.

Figure 6.35 shows the layout of a glass chip designed for carrying out an electrophoretic separation. An injection voltage is applied across the channels linking 1 and 4 which acts to fill the channel completely with the sample solution.

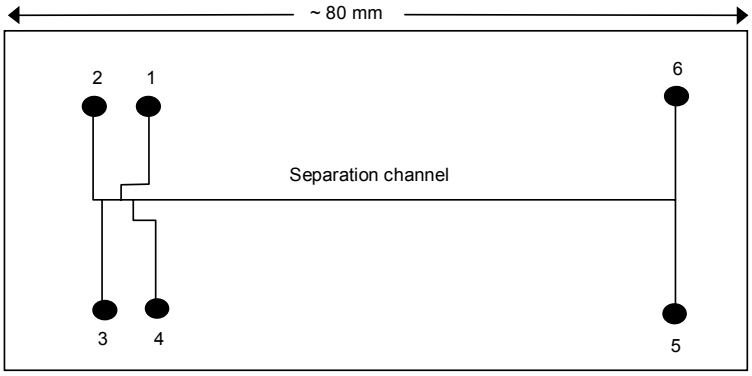


Figure 6.35. *Layout of a chip for capillary electrophoretic separations: 1 = sample inlet, 2 = carrier electrolyte inlet, 3 = modifier electrolyte inlet, 4 = sample outlet, 5 and 6 = outlets*

This results in a sample plug (typically < 100 pL) being introduced into the main separation channel, which has been previously flushed through with the carrier electrolyte. The separation voltage is then applied across the separation channel from electrodes at points 2 and 5, and the sample plug is carried along the separation channel, whereby the separation of the analytes takes place. On-chip detection, usually in the form of laser induced fluorescence or UV/Vis absorbance, takes place at the end of the separation channel, generally using the channel itself as the detector cell and employing a fiber optic cable to transmit the signal to a detector. As the actual volumes of sample are so small in on-chip separation devices, optimal detector conditions are vital if sensitive detection is to be possible. Special flow cells have been designed to maximize the optical pathlength without affecting separation efficiency, such as reflective channel surfaces to reflect the optical signal through the sample several times and designing the detection to probe longitudinally along the separation channel. In LC chip devices, electrodes can be built into the chip to allow sensitive electrochemical detection, although due to the high separation voltages required for electrophoretic separations (typically ~1,000 V/cm) electrochemical detection is less practical. The nature of on-chip separation devices allows sample handling and pre-treatment procedures to be carried out all within a single device (μTAS). Not only has this technique the advantage of reduced sample handling and thus less possible contamination, but it also allows the analyses to be carried out using much smaller sample volumes than previously possible.

6.7.2. *Autonomous sensing devices*

One of the most exciting developments in analytical science will be the appearance of "autonomous sensing devices" that incorporate wireless communications within a miniaturized instrument or sensing device. A key component in the development of these devices will be the ability to produce reliable analytical data over extended periods of time (ideally up to a year). This will require excellent reagent and instrument stability and very low power consumption and/or ability to search for power from the environment. The merging of miniaturized computer systems and wireless communications will increase a remarkable demand for small, autonomous analytical systems based on micro-dimensioned instruments (μTAS or lab-on-a-chip) and/or chemical sensors and biosensors, as these will become the primary information sources about the "real world" for these emerging network-communication technologies.

The combination of microfluidic systems and low power colorimetric detection constitutes a powerful tool for solving many analytical challenges in environmental monitoring and in medical applications. Potential integration of wireless communications systems will create enormous interest in the future as this will certainly lead to the emergence of extensive networked autonomous analytical "stations" that will offer high quality information about key chemical parameters that determine the quality of our environment. These devices will be the information gatherers of the future.

6.7.3. *Sub-micron dimensioned sensors*

Sub-micron dimensioned sensors are gaining in popularity, mainly due to the applications that are only open to sensors of this small size, in particular for biological and medical situations.

6.7.3.1. *Microamperometric sensors*

Microelectrodes, also commonly known as ultramicroelectrodes, may be defined as electrodes whose critical dimension is in the micrometer range, although electrodes with radii as small as 10 Å have been fabricated [19]. These small amperometric sensors have greatly extended the range of sample environments and experimental timescales that are useful for electroanalysis. In this section, we explore some of the exciting and innovative practical applications of these electrodes whose active surface areas are many times smaller than the cross-section of a human hair. Microelectrodes have several desirable attributes including small currents, steady state responses and short response times. The currents observed at these sensors typically lie in the picoamp to nanoamp range, which are several orders of

magnitude smaller than those observed at the conventional macroelectrodes where the radius is usually several millimeters. These reduced currents are a key element in the successful application of microamperometric sensors. In the past, the range of conditions under which electrochemical measurements could be made was restricted to highly conducting media, such as aqueous electrolyte solutions. This restriction arose because resistance between the working/sensing electrode and the reference electrode limited the precision with which the applied potential could be accurately controlled. The small electrolysis currents observed at microamperometric sensors often completely eliminate these ohmic effects. The immunity of microelectrodes to ohmic drop phenomena makes it possible to quantify the concentrations of electroactive analytes in previously inaccessible samples such as non-polar solvents, supercritical fluids, and even solids. The small size of these sensors makes diffusional mass transport extremely efficient. In fact, mass transport rates to a microelectrode are comparable to those of a conventional macroelectrode that is being rotated at several thousand rpm. This efficient mass transport makes it possible to observe steady state responses when the applied potential is slowly scanned in cyclic voltammetry. The sigmoidal shaped responses observed in these experiments are analogous to the polarograms obtained using a dropping mercury electrode, or a rotating disk electrode, but they are observed under entirely quiescent conditions. The steady state limiting current is directly proportional to the analyte concentration, making it extremely useful for determining the concentration of analytes in liquid, solid and even gas phases.

The low currents, high sensitivity and relative immunity of microamperometric sensors to ohmic effects greatly simplify electroanalysis. These attributes not only mean that simpler instrumentation can be used, e.g. two-electrode instead of three-electrode potentiostats, but also that microelectrodes can be used for electroanalysis in media of high electrical resistivity, such as soil, foodstuffs and solutions, without having to deliberately add a supporting electrolyte.

6.7.3.2. *Microelectrodes in biological systems*

The critical dimension of a microelectrode is typically in the 0.1 to 50 μm range. However, many fabrication methods give electrodes in which the sensing area is microscopic, but the complete electrode is macroscopic because the non-conducting body has a radius of several millimeters. These electrodes are not useful for performing electrochemistry in small volumes, or for obtaining information about redox activity with high spatial resolution. Therefore, other encapsulation methods have been developed to ensure that the non-conducting material is thin. One method involves insertion of carbon fibers into microscopic tapered glass pipettes that are subsequently sealed with epoxy resin. An active electrode surface is subsequently exposed by mechanical polishing. An alternative procedure involves electropolymerization of a passivating polymer film around the carbon fiber

electrode. Both of these methods can give electrodes with total diameters in the tens of micrometers range. These small electrodes are widely applied in studies of biological systems since their implantation causes little tissue damage, yet they still provide a sufficiently large area for sensitive extracellular measurements. The microprobes offer a relatively noninvasive means of *in vivo* monitoring, not only because they are physically small, but also because of the minute quantities of material electrolyzed.

Since the pioneering work of Davies and Brink in 1942 that measured the concentration of oxygen in animal muscle, microelectrodes have been instrumental in providing information about the concentration and temporal release of redox active biomolecules. This research is becoming more important in light of evidence that not only is the absolute concentration of a chemical messenger important in dictating a cellular response, but so too is the time profile (frequency) of the output. The chemical events of interest are often restricted to the interior or exterior surfaces of single cells. Therefore, to provide useful information about *in vivo* biochemistry, these measurements must be performed with a high degree of spatial and temporal resolution as well as a high degree of sensitivity and selectivity.

The mammalian brain has been the focus of a significant research effort over the last 25 years and represents an extraordinarily challenging environment in which to perform analytical chemistry. At every level of organization, the brain is temporally and spatially heterogenous with neuronal structures of differing sizes (from nanometer to meter) communicating with each other at timescales ranging from the microseconds to hours or days. When the objective is to elucidate the structure-function relationship of these assemblies, the ability to make spatially resolved measurements across a wide range of timescales is paramount. The first challenge is to ensure that measurements of the neurotransmitter concentration are sufficiently spatially localized to provide a meaningful insight into the brain's structure. Of particular importance is the volume of tissue that is sampled, since this dictates the size of structure that may be examined. For example, a microdialysis loop combining two 0.4 mm stainless steel cannulae cannot provide information on a micron length scale. In contrast, by using microelectrodes and experimental timescales less than 100 μs, the diffusion layer thickness will be less than 1 μm and a high degree of spatial resolution can be achieved. Wightman and co-workers have demonstrated that the distance between the microelectrode and the source cell can dramatically affect the nature of the microelectrode response. As illustrated in Figure 6.36, driving the electrode toward the cell causes the amplitude of the current spikes associated with catecholamines release to increase in amplitude and become narrower. This behavior arises because the neurotransmitter rapidly diffuses in the extra-cellular medium, giving rise to smaller peaks when the microelectrode is far from the release site.

A second key issue in bioelectrochemistry is sensitivity. In the case of localized release, e.g. neurotransmitter release through exocytosis, the local concentration may be high, but the total number of molecules released will be very small. In this regard, the high mass rather than the concentration sensitivity of electrochemical techniques, and the ability to routinely measure small currents, $\leq$ 1 pA, is important. We can confidently expect that both the range of materials and approaches to sensor design will continue to expand and increase in their sophistication. However, the complexity will be hidden from the end-user. For example, highly integrated sensor systems incorporating advanced signal processing and sensor arrays will provide real-time information simultaneously about several analytes in a portable format.

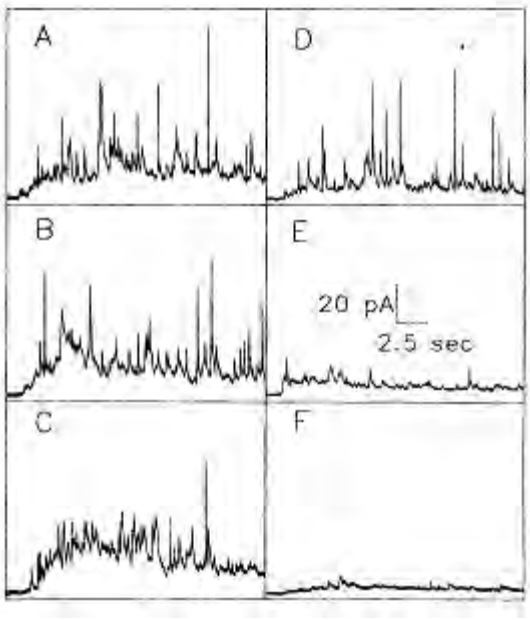


Figure 6.36. *Amperometric detection at isolated bovine adrenal medullary cells detected at glass encased (A-C) and etched (D-F) carbon fiber electrodes at 1 (A, D), 5 (B, E) and 10 μm (C, F). Measurements at each position were made simultaneously with the large and small electrode. Release of catecholamines was induced by a 3 second, 100 μM nicotine exposure applied at 1.25 seconds*

6.8. Conclusions

All of the types of sensors discussed in this book – electrochemical sensors, optical sensors, acoustic sensors and biosensors – are getting smaller, faster and more accurate as researchers continue to strive towards creating "ideal" sensors.

There are even now sensors that are genuinely autonomous and wireless being used in environmental monitoring applications. Medical devices with sensing technology are now on the market that are pocket-sized, with disposable test-strips and can be used "anytime, anywhere". DNA chips and other biochips already allow infectious diseases or genetic alterations associated with many cancers to be detected. It seems likely that electrochemical and optical detection of mutations in human DNA using microarrays will continue to develop so as to offer high speed biomedical diagnostic results. In the food and drink industry, there are sensors throughout the various processes to ensure that the products we buy are safe and fresh.

Many key problems in chemical sensing will be solved by the availability of new and exciting fabrication techniques, miniaturization and microfluidics and the surge of interest and developments in communications technology. Novel detection approaches will also appear; for example, a number of papers have appeared describing the detection of analytes based on principles traditionally associated with biological systems or semiconductors. The creation of highly selective electrochemical sensors relies on the ability to manipulate ensembles of molecules so as to achieve a selective response toward a target analyte. Molecular self-assembly will become an increasingly important approach to engineering sensor materials with useful electrochemical or photochemical properties.

All in all, chemical sensors and biosensors are experiencing some very exciting developments. Sensors are our way of finding out what is going on in the world around us – in our homes, at our work, inside our bodies, in the food we eat. They will never replace our own senses but by informing us about our environment, health and diet, they will certainly help to improve our quality of life.

I would like to acknowledge my colleagues in Dublin City University who contributed to this chapter - Dermot Diamond, Kieran Nolan, Paddy Kane, Robert Forster, Margaret Sequeiera, Kim Lau, Brendan Duffy, Richard O'Kennedy, Tony Killard, Paul Dillon, Brett Paull and Michaela Bowden.

6.9. References

[1] Diamond D. and Nolan K.: *Calixarenes, Designer Ligands for Chemical Sensors, Anal. Chem.*, Vol. 73, 2001, pp. 22A-29A.

[2] Dryhurst G. and McAllister D.L.: "Carbon Electrodes", in *Laboratory Techniques in Electroanalytical Chemistry*, pp. 289, P.T. Kissinger and W.R. Heineman (eds.), Marcel Dekker Inc., New York, USA, 1984.

[3] Weber S.G.: "Detection Based on Electrical and Electrochemical Measurements", in *Detectors for Liquid Chromatography*, pp. 229, E.S. Yeung (ed.), John Wiley & Sons, New York, USA, 1986.

[4] Hargis L.G.: *Analytical Chemistry, Principles and Techniques,* Prentice Hall, New Jersey, 1988, pp. 326.

[5] SKOOG D.A., WEST D.M. and HOLLER F.J.: *Fundamentals of Analytical Chemistry*, W.B. Saunders, New York, 5th ed., 1988, pp. 6-55.

[6] SOLSKY R.L.: Ion-selective electrodes, Anal. Chem., Vol. 62 (12), 1990, pp. 21-33R

[7] DIAMOND D., LU J., CHEN Q. and WANG J.: Multicomponent batch-injection analysis using an array of ion-selective Electrodes Anal. Chim. Acta, Vol. 281, September 1993, pp. 629-635.

[8] CURIE J. and CURIE P.: Bull. Soc. Min., Paris, Vol. 3, 1880, pp. 90.

[9] SAUERBREY G.Z.: Verwendung von Svhwingquarzen zur Wägung dünner Schichten und zur Mikrowägung, Z. Phys., Vol. 155, 1959, pp. 206-222.

[10] MCCORMACK T., KEATING G.J., KILLARD A.J., MANNING B. and O'KENNEDY R.: Biomaterials for Biosensors, in *Principles of Chemical and Biological Sensors*, D. Diamond (ed.), John Wiley and Sons, Chichester, England, 1998.

[11] NATSUME T.: Proteomics: Life's rich tapestry is now on show, *BIA Journal*, Vol. 7(1), 2000, pp. 5-7.

[12] SONKSEN C.P., NORDHOFF E., JANSSON O., MALMQVIST M. and ROEPSTORFF P.: Combining MALDI Mass Spectrometry and Biomolecular Interaction Analysis Using a Biomolecular Interaction Analysis Instrument, Anal. Chem., Vol. 70(13), 1998, pp. 2731-2736.

[13] KUKANSKIS K., ELKIND J., MELENDEZ J., MURPHY T., MILLER G. and GARNER H.: Detection of DNA Hybridization Using the TISPR-1 Surface Plasmon Resonance Biosensor, Anal. Biochem., Vol. 274, Issue 1, October 1999, pp. 7-17.

[14] CLARK L.C. Jr. and LYONS C.: Electrode systems for continuous monitoring in cardiovascular surgery, Ann. NY Acad. Sci, Vol. 102, (1962), pp. 29.

[15] PARK J.K., YEE H.J., LEE K.S., SHIN M.C., KIM T.H., KIM S.R.: Determination of breath alcohol using a differential-type amperometric biosensor based on alcohol dehydrogenase, Anal. Chim. Acta, 390 (1-3), May 1999, pp. 83-91.

[16] NORTHRUP M.A., CHANG M.T., WHITE R.W. *et al.*: DNA Amplification with a Micro-fabricated Reaction Chamber, in *Technical Proceedings of the 7th International Conference on Solid-State Sensors and Actuators*, Yokohama, Japan, Institute of Electrical Engineers of Japan, 1993, pp. 924-6.

[17] MANZ A., HARRISON D.J., VERPOORTE E. and WIDMER H.M.: in *Advances in Chromatography*, P.R. Brown and E. Grushka (eds.), Marcel Dekker, New York, 1993.

[18] RUZICKA J. and HANSEN E.L., *Flow Injection Analysis*, 2nd ed., Wiley, New York, 1988.

[19] FORSTER R.J., Microelectrodes: new dimensions in electrochemistry, Chem. Soc. Rev., Vol. 23(4), 1994, pp. 289-297.

Chapter 7

Level, Position and Distance

7.1. Introduction

7.1.1. *Classification of LPD sensors*

Measuring the level, position, distance and displacement of physical objects is essential for many applications: process feedback control, performance evaluation, transport, traffic control, robotics, security systems, to name just a few.

The sensors that can operate only when they are in direct contact with measured object belong to the class of contact sensors. By analogy, sensors which perform the measurement task without direct contact with a measured object form the class of non-contact sensors. Obviously the non-contact sensors offer many advantages as ideally they do not interfere with the measured object.

For measurement of time-varying quantities (e.g. vibrations) the dynamic properties of sensors are the key criteria for selection. The dynamic properties of sensors are determined by the frequency response of the sensor (the ratio of the amplitudes of output and input variables with sinusoidal waveform at the different frequencies).

Chapter written by Stanislav ĎAĎO and G. HARTUNG.

7.2. Resistive LPD sensors

7.2.1. *Potentiometer*

Potentiometers are resistive devices with a linear or rotary sliding contact whose position is affected by the position (movement) of the measured object. The resistance of the resistive material (winding or a layer of resistive material placed on insulated core) between the beginning of the winding and the wiper, is proportional to the position of the wiper. The potentiometer can operate as a variable resistor (rheostat) or resistive voltage divider. The operation in voltage divider (potentiometric) mode in which the voltage between the wiper and one end of the resistive material (output voltage U_2) is measured offers more advantages.

When operating in the rheostat mode the resistance R2 between the wiper at position x and one terminal of the resistive winding, x can be calculated from the simple equation:

$$\frac{x}{x_0} = \frac{R_2}{R}$$

where $R = R_1 + R_2$ is the total resistance of the resistance winding (or resistive track) with length of x_0.

The principle of operation in potentiometric mode is shown in Figure 7.1. In potentiometric mode the output voltage U_2 is given by the relation:

$$U_2 = \frac{R_2}{R_1 + R_2}$$

and it is proportional to the resistance R_2 which in turn is proportional to the position of the wiper *x*.

Proportionality is valid only if the loading resistance R_Z is much greater than *R* as it can be seen from the equivalent circuit (see Figure 7.1).

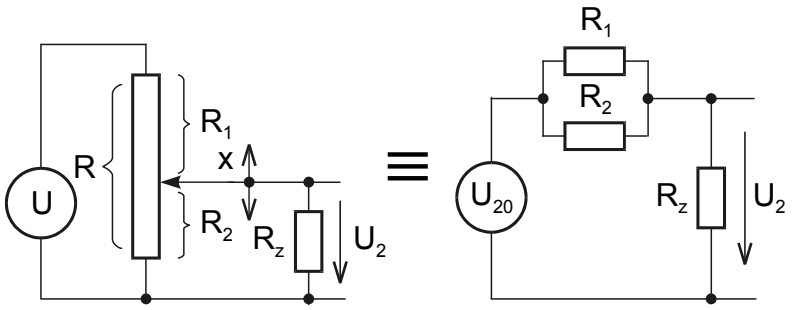


Figure 7.1. *Principle of operation of linear potentiometer with moving wiper and equivalent circuit diagram*

7.2.2. *Angular position measurement*

The rotational potentiometer is a common sensor for angular position measurement. Similarly as for the linear potentiometer, the output voltage U_α measured on the slider (Figure 7.2) is proportional to its position.

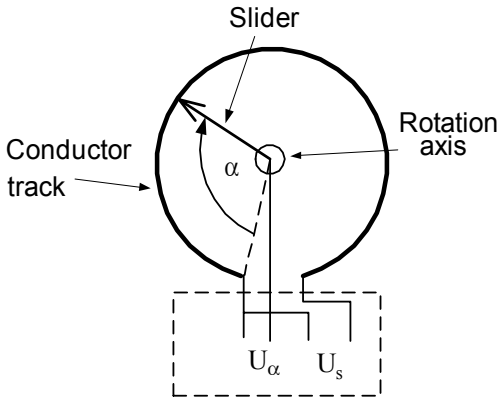


Figure 7.2. *The principle of rotational potentiometer*

The range of angular potentiometers is approximately 10^o to $3{,}000^o$ for the multi-turn potentiometers (helipots).

7.2.3. *Draw wire sensors*

In the draw wire sensors the transformation of translational movement to angular is performed by rolling up a cable made from metal or nylon on a spring-driven metal drum. The other end of the cable is fixed on the moving object, while the sensor case remains at a fixed position. The angular deflection is then measured by an angular potentiometer.

The measurement range of these sensors is from 0-50 mm up to 0-50 m.

Figure 7.3. *Draw wire sensors*

7.2.4. *Inclination detectors*

Inclination detectors measure the angle of direction of the Earth's center of gravity. The sensor (switch) is made of a glass tube having two electrical contacts and a drop of mercury. When the sensor is inclined the mercury "wiper" moves away from the contacts and the switch is opened.

For higher resolution, an *electrolytic tilt* sensor is used. A small slightly curved glass tube is filled with a semi-conductive electrolyte. Three electrodes are built into the tube: two small electrodes on the ends and an extended electrode along the length of the tube. An air bubble resides inside the tube and may move along its length as the tube tilts. Electrical resistance between the center electrode and each of the end electrodes depends on the position of the bubble. In some systems the "liquid wiper" made from electrolyte or mercury is used in order to decrease the effect of friction.

Accelerometers capable of measuring the DC component of acceleration (e.g. accelerometers produced in form of IC by Analog Devices) can be also used as inclinometers.

7.2.5. *Application of potentiometers*

Wire-wound potentiometers are fabricated with thin wires having a diameter in the order of 0.01 mm and have limited resolution. Resolution of resistive film potentiometers is restricted by the non-uniformity of the resistive material and the noise of the voltage measuring instrument. The resistive film (layer) is fabricated with conductive plastic, carbon film, metal film, or a ceramic-metal mixture which is known as *cermet.*

The life time of potentiometers is approximately 100 million cycles.

7.3. Inductive LPD sensors

In inductive sensors the measured quantity causes a change in magnetic flux in the coil or magnetic flux coupling between two or more coils (windings). A change of magnetic flux in the coil is done by inserting a ferromagnetic core into the coil which causes the change of magnetic reluctance (magnetic resistance). The variable reluctance R_m affects the inductance of the coil L defined by the following relationship:

$$L = \frac{N^2}{R_m}$$

where N is the number of turns of the coil winding.

The basic formula for magnetic reluctance is:

$$R_m = \frac{l}{\mu_r \mu_0 S}$$

where l is the length and S is the cross-section of the magnetic path (field line), μ_r is the relative permeability of the magnetic core and μ_0 is the permeability of the vacuum. From this formula it can be deduced that there are various alternatives to inductive sensors based on variable reluctance (*inductance sensors*).

In the typical example of a variable reluctance sensor shown in Figure 7.4 the change of core position causes the increase of the inductance of one coil $Z_1(j\omega)$ while simultaneously decreasing the inductance $Z_2(j\omega)$ of the second coil.

7.3.1. *Linear variable differential transformers*

In linear variable differential transformers (LVDTs), the movement of the ferromagnetic core (Figure 7.5) changes magnetic flux coupling, i.e. mutual inductances M_1, M_2 between primary coil P and secondary coils S_1 and S_2. The primary coil is connected to the AC voltage source $U(j\omega)$ and two secondary coils are connected in opposite phases. When the core is in the magnetic center of the transformer, the secondary output voltages cancel each other out and there is no output voltage. Moving the core away from the central position unbalances the induced magnetic flux ratio between the secondary coils producing an output. Consequently the amplitude of induced voltage is proportional (in linear operating region and steady state) to the core displacement.

Information about the direction of displacement can be determined from the phase angle between the primary voltage and secondary voltage.

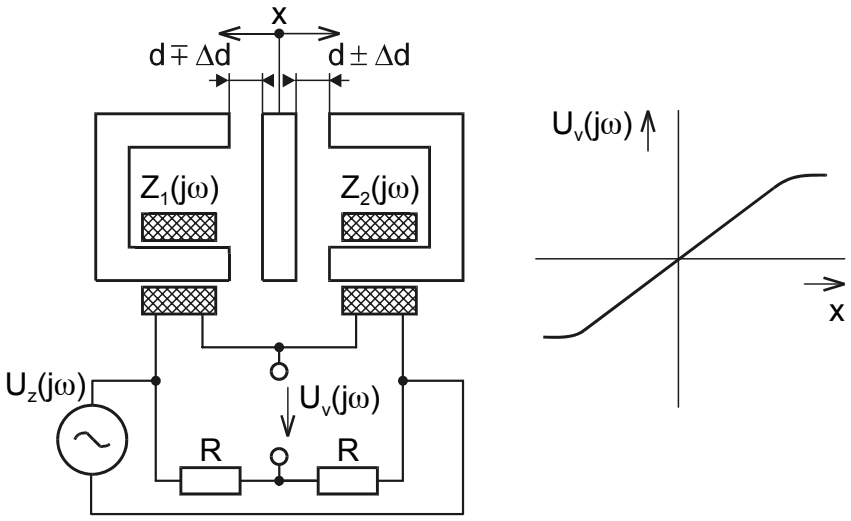


Figure 7.4. *Differential displacement sensor with variable reluctance*

A typical signal conditioning circuit for LVDT is a synchronous detector using primary voltage as a reference. The amplitude of the synchronous detector output

voltage is proportional to the magnitude of displacement while its polarity indicates the direction of movement.

The construction of variable differential transformers can be modified for angular position measurements. These types of sensors, which are commercially available and known as RVDTs (rotary variable differential transformers) have a core actuated by measured rotary movement.

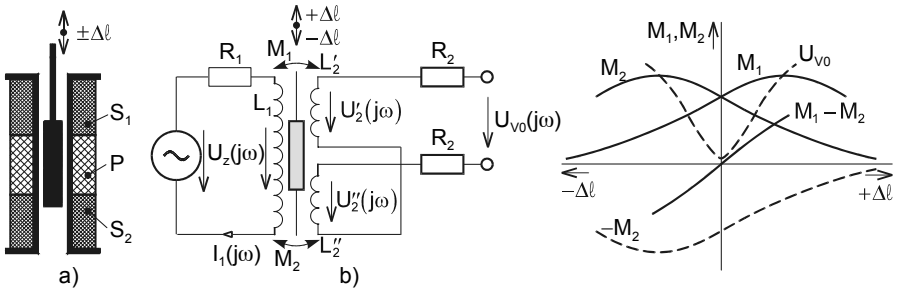


Figure 7.5. *The principle of LVDT*

7.3.2. *Inductosyns*

The primary and secondary coils of inductosyns are in the form of "printed circuits". The primary "stator" coil is a periodical pattern of a conductive layer placed on a strip of insulator *(*ruler, scale) material. The period (step) of the stator pattern is equal to p (usually p = 2 mm). On the slider two "rotor" coils with a pattern identical to that of the stator (Figure 7.6) are located and mutually geometrically shifted by an odd number of $p/4$.

The rotor windings are driven by two AC voltage sources with a phase angle mutually shifted by 90°. If the windings of the stator and rotor are located exactly in opposite positions, then the mutual inductance reaches a maximum value and a maximum voltage will be induced on the stator winding. If the patterns of the stator and one rotor coil are shifted geometrically by $p/4$, the voltage induced to the stator will ideally be zero. Thus, it can be deduced that the mutual inductance between the stator and the rotor coils depending on the position obeys the sinusoidal law and (for the shifted rotor) cosinusoidal function.

The total voltage induced in the stator for a mutual displacement of rotor and stator windings by x is equal to:

$$u_2(t) = KU(\cos\alpha\cos\omega t + \sin\alpha\sin\omega t) = KU\cos(\omega t - \alpha)$$

where $\alpha = 2\pi\frac{x}{p}$

The displacement x can be found from phase shift α between voltage $U_2(t)$ and voltage on the rotor windings.

Inductosyns can also be used with a driven stator and two outputs from slider (rotor) windings.

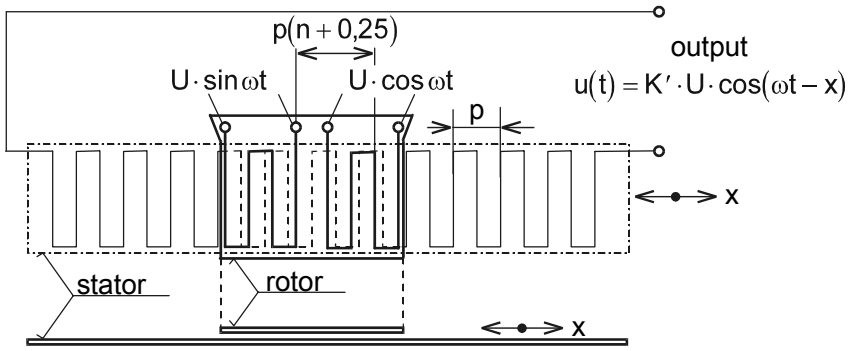


Figure 7.6. *Inductosyn with two rotor (slider) windings*

7.3.3. *Resolvers*

The resolver (Figure 7.7) consists of two stationary coils (stator) and a rotative coil (rotor). The rotor is driven by the current proportional to voltage U sinωt. As the magnetic coupling between stator and rotor depends on mutual angular position, the alternating current in the rotor induces alternating voltages with amplitudes dependent on the angular position of rotor α, i.e.

$$u_1(t) = K.U.\cos\alpha \sin\omega t,\ u_2(t) = KU\sin\alpha\sin\omega t$$

The amplitude ratio of the stator and rotor voltages (after demodulation) gives direct information on the angular position α.

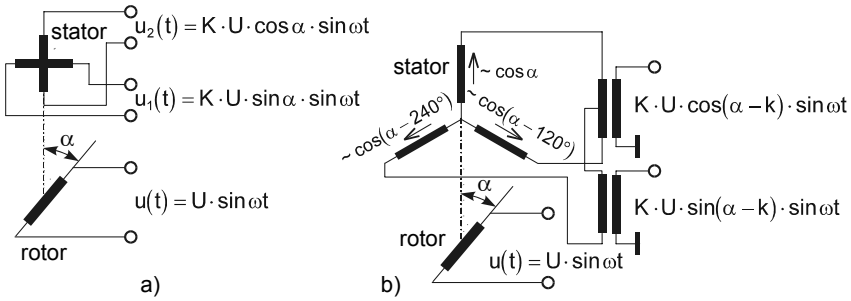


Figure 7.7. *The principles of resolver and selsyn*

7.3.4. *Selsyn*

Selsyn is conceptually similar to the resolver. It contains three windings – stator coils are 120° apart – and belongs to the oldest type of angle measurement sensors (radar antenna position measurements).

7.3.5. *Inductive sensors of angular velocity*

The principle of the inductive sensors is shown in Figure 7.8.

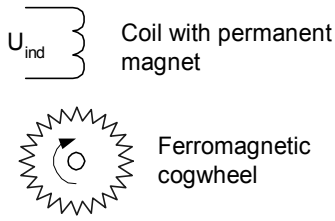


Figure 7.8. *Inductive sensor of angular velocity*

A cogwheel, made from ferromagnetic material with a number of teeth moves in the magnetic field of a permanent magnet. When a tooth tip is close to the coil, the magnetic flux is large. The rotation of the cogwheel causes periodical changes of magnetic flux. According to Faraday's law, the induced voltage in the coil is proportional to the first differentiation of the magnetic flux and is proportional to the angular speed ω.

A disadvantage of all sensors, which are based on induction of voltage caused by changes of magnetic field, is low amplitude of induced voltages for slowly changing variables (e.g. low rotational speed).

7.3.6. *Eddy current distance sensors*

AC excitation field H induces eddy currents in the conductive measured object (Figure 7.9 a). These eddy currents generate magnetic field H_v, which according to the Lenz rule opposes (counteracts) the excitation field, that of the driving (sensing) coil, and causes a decrease in the original field H. The effect of the eddy currents can be represented by the depth of their penetration given by the formula:

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}}$$

where f is the frequency, μ is the permeability and σ is the conductivity of the target.

For small penetration depth the distance between the sensor and measuring object can be measured independently on the thickness of the object.

The effect of eddy currents can be measured as the change in impedance of the coil. To measure the distance d (*proximity detector*) the change of impedance of active $\mathbf{Z}_m$ is measured by the AC bridge circuit (Figure 7.9 b).

Another possibility is to use the oscillator-based circuit as shown in Figure 7.10.

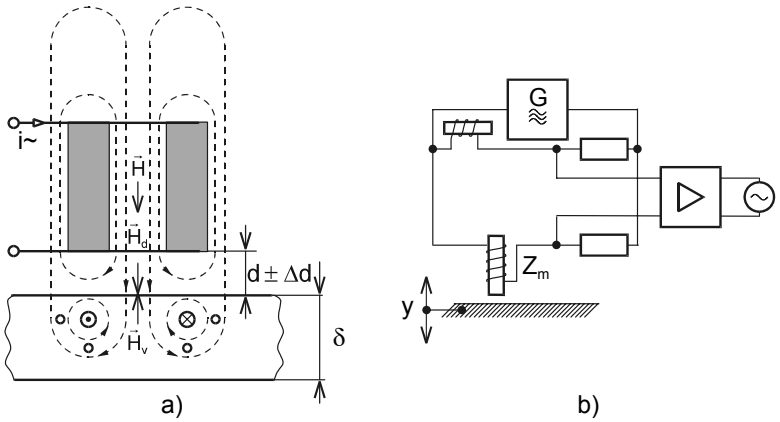


Figure 7.9. *a) The physical principle of eddy current based sensors*
b) Bridge type of the signal conditioning circuit

If a conducting object approaches the magnetic field of the coil, the intensity of eddy currents increases the losses of the tuning circuit of an oscillator causing the drop in the amplitude of oscillations. The value of demodulated amplitude of the oscillations is compared.

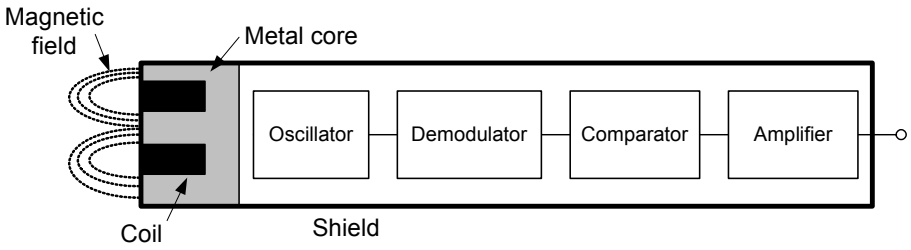


Figure 7.10. *The typical construction of eddy current proximity detector with the threshold*

The coil assembly which has a metal guard around the ferrite core is used to focus the electromagnetic field to the front plane of the sensor.

The construction of a shielded sensor with signal conditioning circuitry is shown in Figure 7.10.

Eddy current proximity detectors allow the detection of fast moving objects. Eddy current sensors could be used at high temperatures, for example for measuring the level of molten metals.

7.4. Magnetic LPD sensors

In magnetic sensors the position of the measured object (often carrying a permanent magnet) causes changes in the magnetic field, which are then measured by a magnetic field sensor.

7.4.1. *Magnetic field sensors*

The magnetic field sensors generally measure the vector of the magnetic field. The measurement may include magnitude of the vector, its direction or both. A detailed description of the magnetic field sensors is the subject of Chapter 10 and here will be only shortly reviewed.

Anisotropic magnetoresistive (AMR) sensors

These sensors are made of a Permalloy (NiFe) thin film deposited on a silicon wafer and are patterned as a resistive strip.

The properties of AMR films cause the resistance to change by 2-3% in the presence of a magnetic field. Typically four of these resistors are connected in a Wheatstone Bridge with a resistance around 1 kΩ. AMR sensors provide an excellent means of measuring both linear and angular position and displacement.

7.4.2. *Reed switches*

A reed switch consists of a pair of contacts hermetically sealed which are activated by an external magnetic flux (typically when a magnet approaches the switch by about 5 mm). The reed switches have a substantial amount of hysteresis which makes them immune to small fluctuations in the magnetic field. When a perpendicularly-oriented magnet moves close to a reed switch, several zones of ON and OFF states could occur, as shown in Figure 7.11.

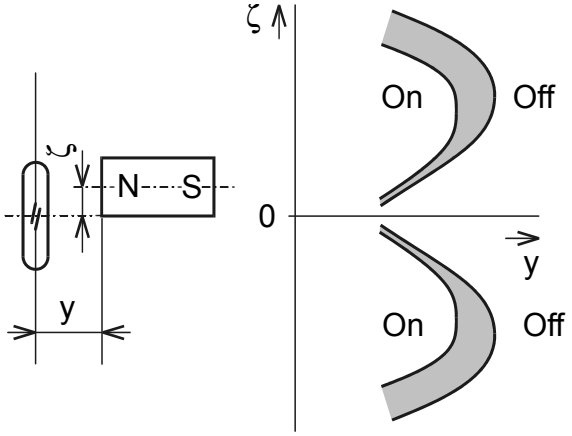


Figure 7.11. *Zones of ON and OFF state when permanent magnet moves close to the reed switch*

7.4.3. *Hall sensors*

In Hall sensors, the semiconductor strip with thickness *d* is exposed to the magnetic field with induction B according to Figure 7.12.

When the current I flows through the semiconductor strip, the magnetic field B (due to the Lorentz force) causes charge displacement towards the lateral electrodes resulting in voltage generation. This voltage is known as Hall voltage U_H given by the equation:

$$U_H = R_H. I. B/d$$

R_H is the Hall-constant depending on the semiconductor material.

For position and displacement measurement, Hall effect sensors must be equipped with a magnetic field source and signal conditioning circuit. Due to the popularity of Hall sensors in the automotive industry, the complex "smart" signal conditioning circuits are available in the form of integrated circuits.

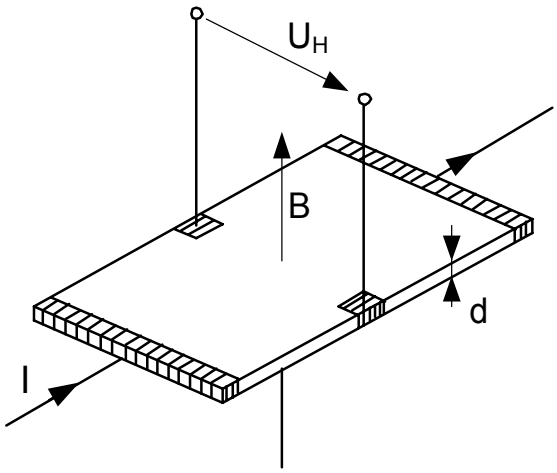


Figure 7.12. *The principle of the Hall effect*

7.4.4. *Semiconductor magnetoresistors*

If a voltage is applied along the length of a thin slab of semiconductor material, a current will flow and the resistance can be measured. When a magnetic field is applied perpendicular to the slab, the Lorentz force will deflect the charge carriers. The magnetic field increases the path length of the charge carriers and therefore the resistance. An increase in resistance of several hundred percent is possible in large fields. A permanent magnet is often incorporated to bias the magnetoresistor up to a

more sensitive part of their characteristic curve. The sensors are usually combined with external resistors to form a Wheatstone Bridge.

7.4.5. *Wiegand wire*

The Wiegand sensor is based on the generation of a voltage in a coil, which has a ferromagnetic core, when an external field causes a change in the magnetic field of the core. The main component of the Wiegand sensor is the Wiegand wire (or pulse wire), which has a magnetically hard surface with a high magnetic coercivity. The magnetic properties of the Wiegand wire are similar to those of a ferromagnetic material with only one domain. Changing the magnetic field in the wire from positive to negative saturation causes generation of large voltages and the increase of sensitivity.

The typical application of this sensor is the proximity switch with the advantage that there is no need for any external voltage source.

7.4.6. *Magnetostrictive sensor*

This sensor is based on the measurement of the time of travel of mechanical pulses in the magnetostrictive delay line. A waveguide (tube) contains a conductor (wire) which upon applying an electrical pulse sets up a magnetic field over its entire length (Figure 7.13). Another magnetic field is produced by a freely movable ring shaped permanent magnet that exists nearby. Superimposing the vectors of the two fields where the permanent magnet is located causes the minute torsional strain, or twist at the location of magnet (Wiedemann effect). This torsional pulse propagates as a mechanical wave in both directions of the material with the sound speed and it is absorbed at both ends. When the pulse arrives to the excitation head of the sensor, the moment of its arrival is precisely measured (e.g. by a variable magnetic reluctance sensor). The measured time of travel between the pulse of the electrical current in the wire and the detected torsional pulse is proportional to the magnet position. Applications of this sensor include hydraulic cylinders, detection of rock movements as small as 25 μm, elevators and other devices where fine resolution along a large dimension is a requirement.

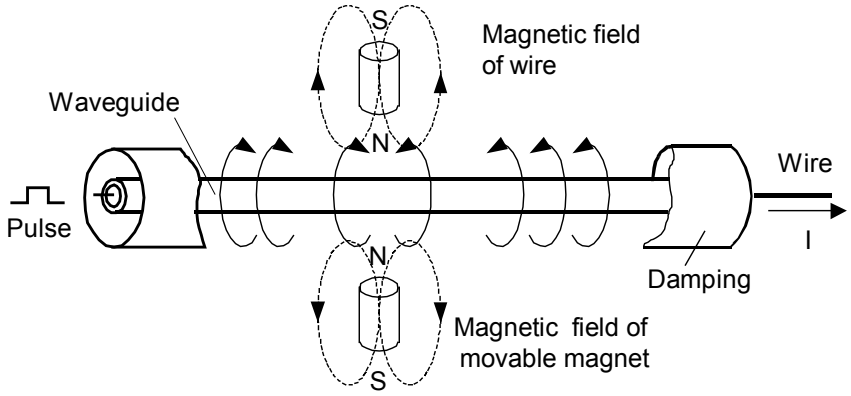


Figure 7.13. *The principle of magnetostrictive sensors*

7.5. Capacitive LPD sensors

7.5.1. *Introduction*

The capacitance between two parallel electrodes is given by the following relationship:

$$C = \varepsilon_r \varepsilon_0 \frac{A}{d}.$$

The principle of the capacitive measurement of displacement is based on the effects of measured displacement on the distance d, area of overlapping of electrodes A or relative permittivity ε_r of the medium between the electrodes. Permittivity of the vacuum is $\varepsilon_0 = 8.8542\ 10^{-12}$ As/Vm. According to this equation, different sensor types can be realized, as shown in Figures 7.18 and 7.19.

The influence of non-homogenous stray fields on the outer edges of electrodes can be eliminated by using the shielding techniques known as the Kelvin guard ring. When the width of the ring is at least five times larger than distance *d* between the electrodes and the gap between the ring and the main electrode is less than *d*/5, the capacity could be calculated according to the homogenous field formula with an accuracy of 1 ppm.

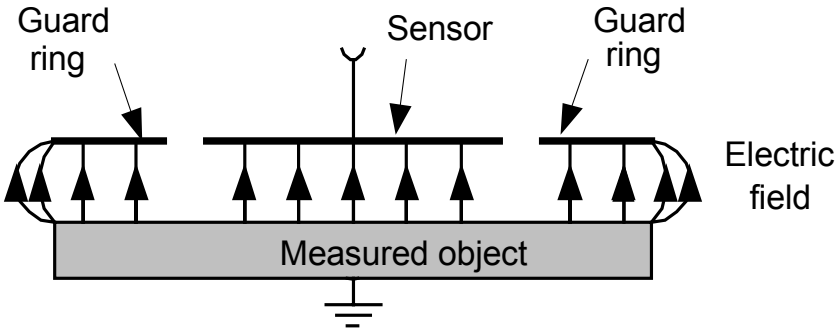


Figure 7.14. *Capacitive sensor with guard ring*

The principle of a capacitive sensor using a guard ring is shown in Figure 7.14. The main electrodes of the sensor and the guard ring have the same potential (connected via the voltage follower). The capacity is given by the distance between the electrodes and the overlapping area of the main electrode and the surface of the measured object where the field is nearly homogenous.

7.5.2. *Signal conditioning circuits for capacitive sensors*

The main problem is to avoid the capacitances of leads to the electrodes of the sensor. In order to avoid the influence of capacities with respect to shielding (ground) as with C_{13} and C_{23} the *current* flowing through sensor C_{12} should be measured (Figure 7.15). If for current measurement the operational amplifier is connected as a current-to-voltage converter, then the influence of capacity C_{23} is negligible as it is connected between the inputs of the amplifier (virtual ground). In order to avoid the capacity of the lead to the right hand electrode of sensor (C_{13}), the ideal voltage source should be used for driving current through the sensor.

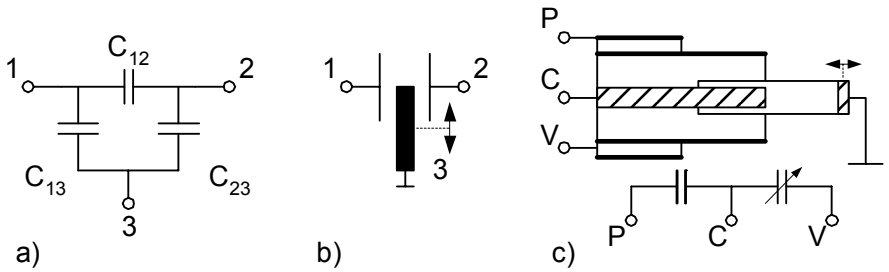


Figure 7.15. *Three terminal configurations of capacitive sensors a) equivalent circuit b) sensor with grounded movable electrode c) cylindrical sensor with a variable area of overlapping for large displacements. Insertion of movable electrode causes a decrease in the capacity when circuits for measurement of feedthrough capacity C_{12} (C_{CV}) are implemented*

This principle is fulfilled in circuit on Figure 7.16 where driving U_R and compensating U_E ideal voltage sources are used for measurement. As the phase of both sources is opposite, the difference of current I_o flows to the input of current-to-voltage converter. The capacity between leads and shielding is connected to nearly zero potential difference and thus has negligible influence. For automatic balancing using the feedback approach, the output voltage U_V controls the amplitude of compensating source U_E and in balanced state is valid:

$$\frac{C_{VC}}{C_{PC}} = \frac{U_E}{U_R}.$$

In the transformer bridge on Figure 7.17 the balance indicator is again a current-to-voltage converter formed by operational amplifier with a C_{zv} in the feedback path. Capacitive feedback eliminates the errors caused by changes of frequency of voltage sources. The influence of capacities between leads and shielding or ground is suppressed.

The principle operations of sensors having parasitic capacitances eliminating three terminal configurations are shown in Figure 7.15.

7.5.3. *Using capacitive sensors*

Besides the optical measurement methods the capacitive position measurement is one of the most accurate and rapid non-contact measurement methods. The sensors

have a relative resolution of more than 10^5 and allow high precision measurements of several hundred micrometers.

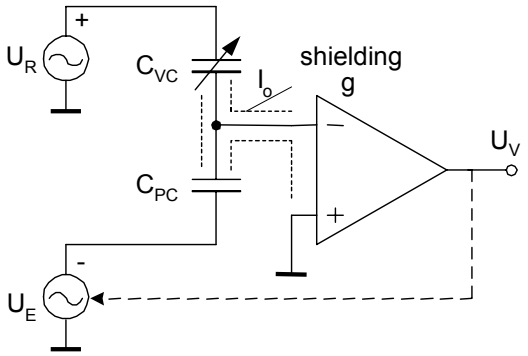


Figure 7.16. *The principle of signal conditioning circuit with automatic balance and suppression of parasitic capacitances of leads*

The application of capacitive sensors with variable air gaps for the measurement of small displacements can compete with the most sensitive optical sensors. For medium range displacement measurements, the capacitive sensor with variable area electrodes is used. During movement of electrodes the air gap distance should not change.

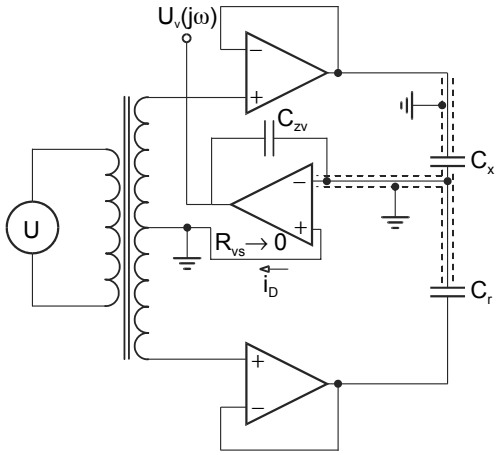


Figure 7.17. *Transformer bridge as signal conditioning circuit for capacitive sensors*

7.6. Optical LPD sensors

7.6.1. *Introduction*

The basic principles of optical sensors for position, movement, displacement and dimensions are described in Chapter 2.

From this reason the attention in this chapter will be devoted only to a short review and the extension of the basic types of optical sensors.

7.6.2. *Photo-electric switches (PES)*

These form a special class of optical sensors whose electric output signal has only two states assigned to the presence or absence of the certain feature in a measured object. Each PES contains at least two basic components: a system transmitter (light source and optical parts) and receiver (optical components and photoelectric detector).

The basic types of PES are determined by the mutual arrangement of the transmitter, the measured object and the receiver and may be divided into three groups:

– through beam;

– diffuse-reflective;

– retro-reflective.

The principle of the through-beam PES is based on an interruption of the beam between a light source and a detector (receiver) by the movement or presence of a measured object (Figure 7.20). Through-beam PES can be used for distances around 100 m.

The disadvantage of this method is the necessity for precise geometrical alignment of the source and the detector.

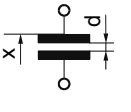
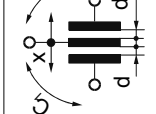
Type of sensor	Schematic symbol	Mathematical description	Transfer characteristic
single plate sensor with variable airgap		$C = \varepsilon \frac{S}{d(x)}; \quad \frac{\Delta C}{\Delta d} \doteq -\frac{C}{d}\left(1 - \frac{\Delta d}{d}\right)$	
differential plate sensor with variable airgap		$C_1 = \varepsilon \frac{S}{d(x)}; \quad C_2 = \varepsilon \frac{S}{d(x)}; \quad \frac{\Delta C}{\Delta d} \doteq -\frac{C}{d}\left[1 + 2\left(\frac{\Delta d}{d}\right)^2\right]$	
plate sensor with dialectric layer and variable airgap		$C = \frac{\varepsilon_1 S}{d_1(x) + \frac{d_2 \varepsilon_1}{\varepsilon_2}}; \quad \frac{\Delta C}{C} = -\frac{\Delta d_1}{d_1 + d_2} \cdot \frac{1}{\frac{1}{N} - \frac{\Delta d_1}{d_1 + d_2}}; \quad N = \frac{\varepsilon_2(d_1 + d_2)}{\varepsilon_2 d_1 + \varepsilon_1 d_2}$	
plate sensor with variable thickness of dialectric layer		$C = \frac{\varepsilon_1 S}{d_0 - d_1(x)\left(1 - \frac{\varepsilon_1}{\varepsilon_2}\right)}$	
plate sensor with variable area of overlapping		$C = \varepsilon \frac{S(x)}{d}; \quad \frac{\Delta C}{\Delta l} \doteq -\frac{C_{max}}{l_{max}}\left(1 + \frac{\Delta d}{d}\right)$	

Figure 7.18. *The principles of capacitive sensors for small displacement measurement*

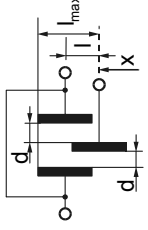
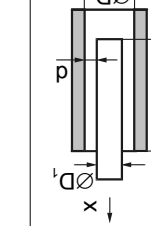
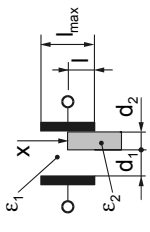
Type of sensor	Schematic symbol	Mathematical description	Transfer characteristic
plate differential sensor with variable area of overlapping		$C = \varepsilon \frac{S(x)}{d}$; $\frac{\Delta C}{\Delta \ell} \doteq -\frac{C_{max}}{\ell_{max}}\left[1+\left(\frac{\Delta d}{d}\right)^2\right]$	
plate differential sensor with variable area of dialectrical layer overlapping		$C = \frac{\varepsilon_1 S}{d_1 + d_2}\left[1+\frac{\ell(x)}{\ell_{max}} \cdot \frac{1-\frac{\varepsilon_1}{\varepsilon_2}}{\frac{d_1}{d_2}+\frac{\varepsilon_1}{\varepsilon_2}}\right]$	
coaxial sensor with variable area of overlapping		$C = \varepsilon \frac{2\pi \cdot \ell(x)}{\ln \frac{D_1}{D_2}}$; $\frac{\Delta C}{\Delta \ell} \doteq -\frac{C_{max}}{\ell_{max}}\left[1-\frac{1}{2}\left(\frac{\Delta d}{d}\right)^2\right]$	
differential sensor with variable area of overlapping		$C = \varepsilon \frac{S(\alpha)}{d}$	

Figure 7.19. *The principles of capacitive sensors for medium and large displacement measurement*

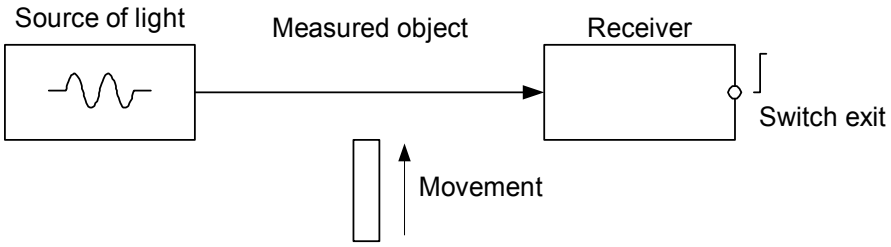


Figure 7.20. *Measuring principle of through-beam PES*

Diffuse reflective PES

The sensing distance depends on the reflectivity and the size of the object. There is a difference between ON and OFF points when moving the object from or to the sensor. This difference (hysteresis) also depends on the reflectivity of the target's surface.

The advantage of diffuse reflective PES is simple installation, the sensor is mounted on only one side and adjusting the sensor is relatively easy (Figure 7.21). On the other hand, a proper function strongly depends on the optical properties of the object, the switching point depends on the position of the object and the sensing distance is relatively short.

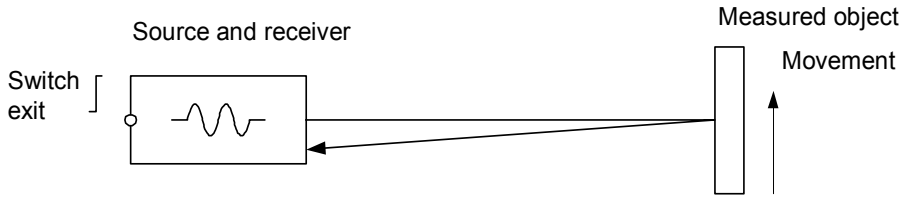


Figure 7.21. *Principle of PES with diffusive reflection from the measured object*

Retro-reflective PES

Their principle is similar to through-beam PES; however, the light source and receiver are located in one housing and a reflector is used on the end of the active zone (Figure 7.22).

The performance of a sensor of this type depends to a large extent on the properties of the reflector; the triple reflector is the most popular in applications. The active beam diameter depends on the size of the reflector and the distance from the sensor.

The presence of the same object may not be detected in a certain position with respect to the reflector or the housing of the sensor. Another disadvantage is that the detection of glossy objects requires polarized light. In case of interruption of the light waves by the measured object, the switch exit will be activated.

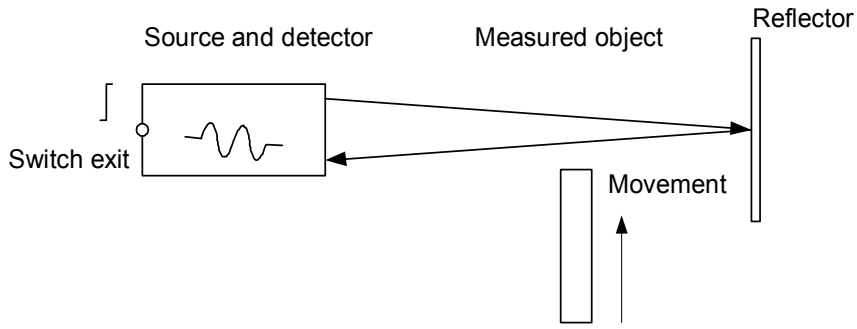


Figure 7.22. *The principle of retro-reflective PES*

7.6.3. *LPD Sensors based on triangulation*

The typical configuration of the sensor depicted in Figure 7.23 is designed for the measurement of the range, i.e. distance d of an object (target) from the light source, rather than the co-ordinates of its position. A light source with a narrow beam (laser diode or LED with the beam collimated to an angle < 2°) strikes the target (point P) and the beam is reflected back toward the PSD sensor (point X_1). The received low intensity light is focused on the sensitive surface of the PSD. The intensity of a received beam greatly depends on the reflective properties of the target. Nevertheless, the accuracy of the measurement depends very little on the intensity of the received light. As the surface moves within the measuring range on either side of the stand off distance, the image of the spot moves laterally (e.g. to point X_o) on the one-dimensional (single-axis) PSD. The output voltage of PSD depends linearly on the target displacement. The laser emitting the focusing laser beam is reflected on to the reference object at point P_0. The change from P_0 to P_1 is thus transformed into a lateral shift of X_0 to X_1 on the receiver level.

For the accurate operation and the satisfactory linearity, the exact orientation of the detector plane and the measured object surface should be chosen.

7.6.4. *Optical encoders*

Optical displacement transducers are based on light intensity modulation by two overlapping gratings. A grating principle of light modulation is used in rotating or linear encoders, where a moving mask has transparent and opaque sections.

There are two basic types of optical encoders:

– incremental encoders (sensors);

– absolute encoders.

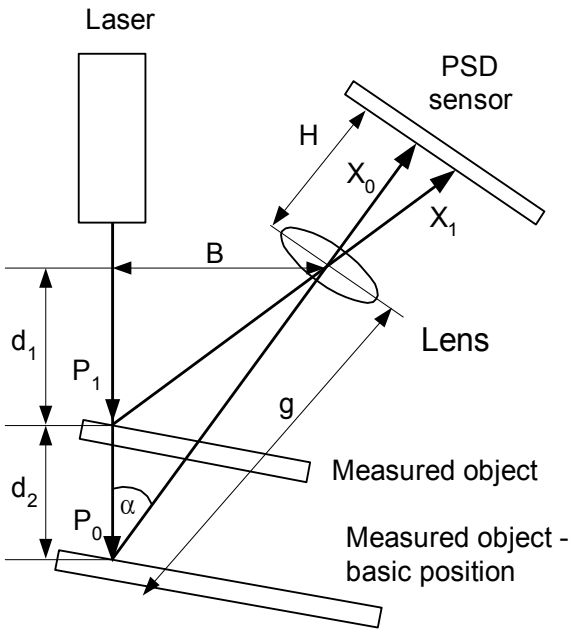


Figure 7.23. *Principle of distance measurement by triangulation*

7.6.4.1. *Incremental sensors*

Incremental optical encoders with grating patterned on a glass disk are often used for this purpose. Their principle of operation and signal conditioning circuits do not differ substantially from encoders of linear (translatory) position. One of the more sophisticated sensor constructions for angular displacement is shown in Figure 7.24. The transient (pulse) is produced whenever the disk is rotated for a pitch angle (the angle difference between two successive marks in a grating).

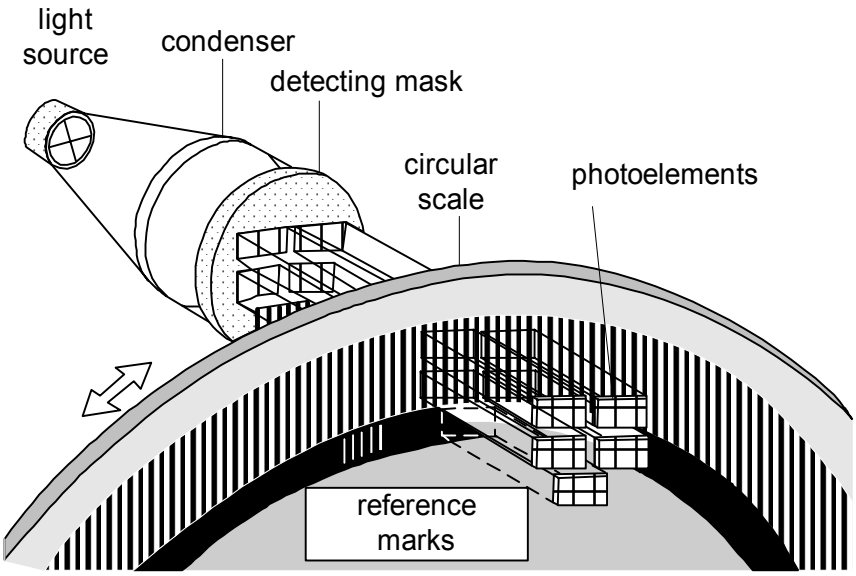


Figure 7.24. *Typical construction of rotary incremental encoder*

7.6.4.2. *Absolute encoders*

Absolute position encoders yield a unique digital output to each resolvable position of a movable element, rule or disk, with respect to an internal reference. The movable element is formed by regions having a distinguishing property, and designated by values 0 or 1. However, unlike incremental encoders, their tracks are so arranged that the reading system directly reads the coded number corresponding to each position. Each track corresponds to an output bit, with the innermost track yielding the most significant bit.

In contrast to incremental encoders, absolute encoders do not accumulate errors. The price to be paid is a more complex reading head as a separate head is needed for each track. In addition, they must be perfectly aligned, otherwise the output code may be ambiguous when changing from one position to a neighboring one.

Binary codes with unit distance in all positions including the first one and the last one are unambiguous. Using coded measurement procedures, the angular position is included as digital information.

The Gray code is commonly used because only one piece of coding information is changed per measurement step, which makes it easy to control transmitting errors.

7.6.4.3. *Gray Code*

Gray code belongs to the class of unit distance codes. In these codes the digital output word will change only in one bit (Figure 7.25) when the position changes by one step (smallest resolvable change of position).

The Gray code does not behave as a weighted code (such as binary code, for example), so for further logical signal processing, the detected signals have to be transformed into "weighted" binary codes.

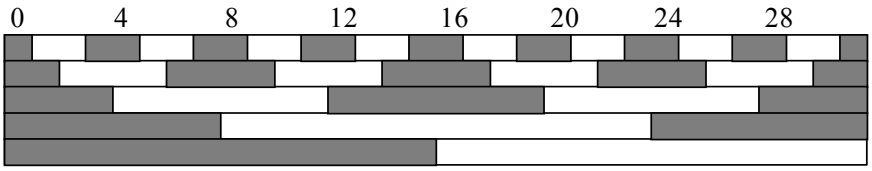


Figure 7.25. *A rule with Gray code*

7.6.5. *Interferometry*

The measurement principle of interferometry is based on the superposition (vector addition) of two coherent light waves (having equal wavelength) in the space. The vector addition results in amplification or attenuation of waves depending on their mutual phase shift which is then observed as a pattern of light and dark fields (interference pattern).

The distance of maximum to minimum light intensity of the interference is $\lambda/2$. Using typical wavelengths of nearly 1 μm, these systems reach a very good resolution.

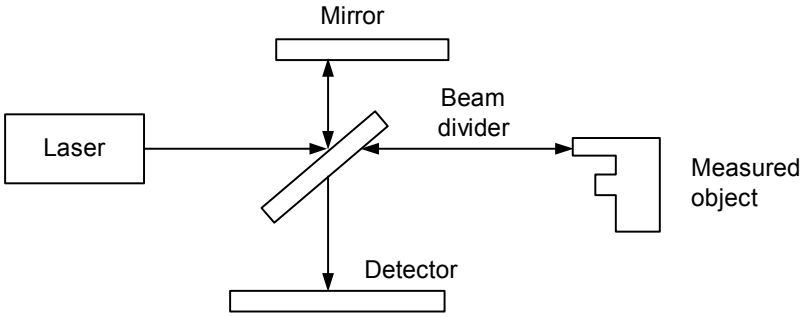


Figure 7.26. *Michelson interferometer*

The simplest case of interference measurement occurs in the "Michelson Interferometer" (Figure 7.26). This one consists mainly of a beam divider, a fixed reflector (reference wave), a moving reflector (measuring wave), a coherent light source (laser) and a detector. Using the Michelson interferometer a light beam is divided in a beam divider (a semi-transparent mirror) into two parts. One part of the light moves vertically to the mirror on which it is reflected. The other part of the light penetrates the beam divider and is reflected at the measuring object. Interference patterns occurring on the beam divider are observed by a light detector. By the analysis of these intensity changes, the displacement of the moving mirror (object) can then be calculated with a high degree of precision.

The equations describing quantitatively the principle of light interferometry can be found in Chapter 2 which is devoted to optical sensors.

The sensors operating on the interferometry principle are suitable for applications where a small number of measuring points with very high accuracy have to be measured. The main application fields of interferometry are surface quality checking and calibration.

7.6.6. *Optical LPD sensors based on travel time (time-of-fly) measurement*

A light wave is transmitted by the sensor, reflected by the measuring object and received by the detector.

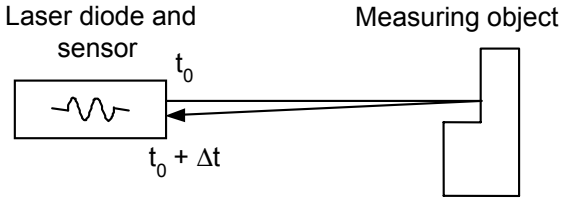


Figure 7.27. *Distance sensors using travel time measurement*

The simplest way to calculate distances with optical sensors is based on the determination of the elapsed time interval between emission and reception of a pulse (signal burst) or series of pulses.

Using laser diodes it is possible to produce pulse widths in the range of sub-pico-seconds, and the resolution of the procedure within the mm-range can be reached. The precision is determined by the resolution of time measurement procedures. Due

to the very high propagation velocity (light velocity) limits, very short time interval measurements are required. A light wave needs 1 ns to travel the distance of 300 mm. This explains the challenges for the time measuring systems intended for short distances.

Typical parameters

Measurement range: up to km range.

Resolution: 1mm

7.6.7. *Image-based measurement-machine vision, videometry*

7.6.7.1. *Introduction*

The optical sensors belonging to this class use the video signal from a camera or an array of optoelectronic sensors (most often 2D-CCD type) as the source of information about geometrical properties of the target. The methods used for the video signal processing could be very complicated as they try to imitate capabilities of the human eye and brain system for interpretation of images and for this reason they are known as *machine vision* or *computer vision*.

The basic principles of image based measurements and some practical examples are described in MM 2 "Optical sensors". Some additional image processing procedures will be introduced in this section.

7.6.7.2. *Light sheet method*

The light sheet method represents a generalization of the principle of the optical triangulation used for distance measurement. In the simplest case a sheet of laser light created by a rotating or oscillating mirror or by cylindrical lenses is projected onto the target object. The reflected light then follows the contours of the object and the recorded picture shows a profile cut of the object for the whole width of the picture. The measuring equipment based on the light sheet approach consists of a laser diode source projecting optical stripes (lines) and the CCD-matrix-camera with high resolution recording the image of the illuminated object. The matrix-video-camera will be located at a defined angle to the laser. The position where the light sheet "cuts" the object is a function of the shape of the object at this place. The principle is described below in Figure 7.28.

The known mathematical procedure used in triangulation is valid for each point on the line recorded by the camera. The light source, the sensor and the measured object represent the cornerstones for the calculation of distances based on light sheet

triangulation. The distance between the light source and the sensor is the basic width. If the basic width, the angle of the projected light beam and the basic width as well as the angle between the reflected light beams are known, then the distance between the basic width and the point of the object can be calculated.

For calculation of the shape along the surface (wall) of the measured object the determination of the relative difference in position will be sufficient. When an angle between the projection level and the optical angle of the camera is 45°, in a vertical observation of the scene, the recorded direction of the light sheet corresponds exactly to the height of the object.

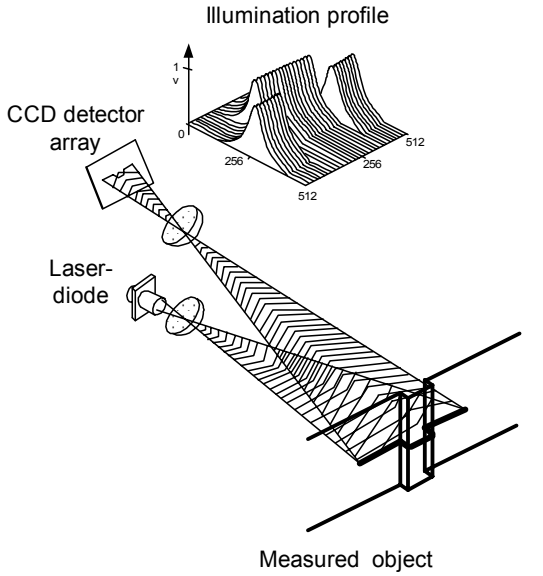


Figure 7.28. *Dimension measurement based on the light sheet projection*

7.7. Ultrasonic sensors

7.7.1. *Introduction*

Ultrasonic waves are mechanical oscillations at a frequency which is beyond the audibility of humans (more than 20 kHz). By a mutual interference of ultrasonic waves from the individual sensors, the high density of emitted energy in a certain direction can be reached. Using the proper phasing of the transducers configured in a

matrix, the concentration of energy in a small area is feasible. The same phasing principle can be implemented on the matrix of receivers for a reflected wave. By the computer control of phases of transducers operating in a transmitting or receiving mode, the scanning regime of operation is feasible (e.g. in ultrasonic tomography).

7.7.2. *Travel time principle*

The pulses containing a definite number of periodic waves propagate with sound velocity towards a measured object. The pulses are then reflected from the measured object and picked up by the receiver (Figure 7.29) with a lag time equal to the elapsed time between the emission and reception of pulses. The received pulses of ultrasonic waves are transformed to the electrical signal by means of the piezoelectric effect.

7.7.3. *Doppler effect*

The Doppler effect, (discovered by C. Doppler in 1843) is the change in frequency undergone by radiation (be it mechanical or electromagnetic) when it is reflected by an object that is moving with respect to the radiation transmitter. If the reflector moves with velocity *v* the shift of frequency is approximately given by relation

$$f_e - f_r = 2f_e \frac{v}{c} \cos\alpha$$

where f_e is the emitted frequency, f_r is the received frequency, and α is the relative angle between reflector velocity and propagation direction.

The relative velocity v can be calculated by measuring the frequency shift.

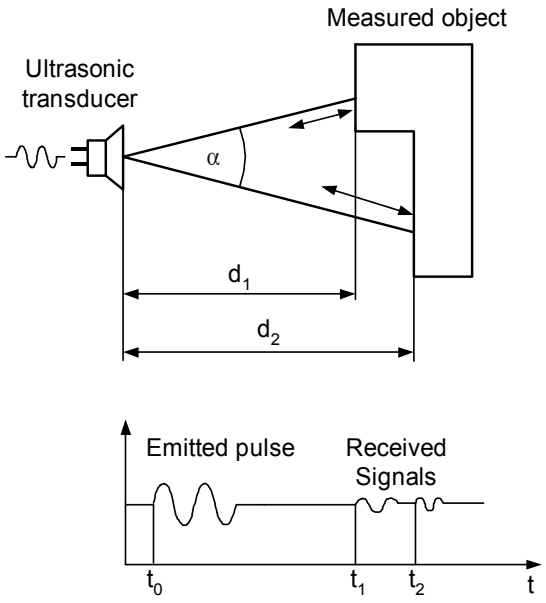


Figure 7.29. *Principle of ultrasonic distance sensor employing echo-travel time procedure*

7.8. Microwave distance sensors (radar)

7.8.1. *Introduction*

Microwaves are electromagnetic waves with high frequency (GHz range). The procedures used in microwave sensors for distance and position measurement are analogical to those of ultrasonic sensors. The measurements are based on:

– the travel time measurement;

– Doppler effect;

– Frequency Modulated Continuous Wave (FMCW).

Travel time measurement is the original procedure used in radar systems for a distance measurement.

As the electromagnetic waves propagate with high velocity the travel times are extremely short and this can put high demands on the electronic time interval measuring circuitry. The *Doppler effect* principle known from section 7.7.3 is especially used for the measurement of velocity and for detection of moving objects.

In many cases radar sensors could replace the detectors for object presence and movement based on infrared radiation (pyroelectric sensors) in door openers and traffic control. A typical example is the microwave sensor KMY 24 (Siemens) used for the detection of moving objects.

7.8.2. *Microwave sensors based on FMCW*

The FMCW-based sensor continuously sends signals with a periodical frequency modulated carrier wave (Figure 7.30). The frequency of the emitted wave changes linearly with time.

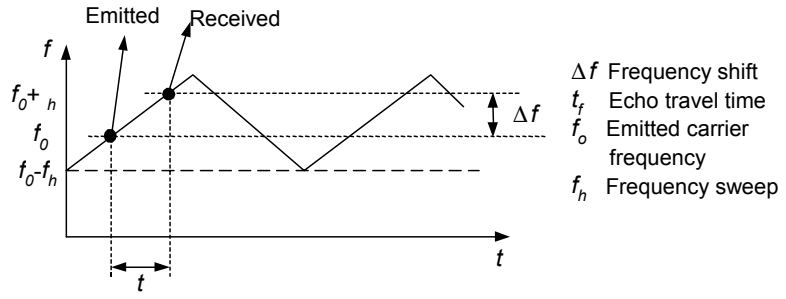


Figure 7.30. *The principle of the FMCW procedure*

The reflected signal has a frequency corresponding to the time instant of its emission, but when it returns the source frequency has already been shifted proportionally to the time interval corresponding to the double distance of the sender/receiver. Evaluation of the frequency difference between emitted and received signals (reflected) is performed in a mixer.

The frequency difference is proportional to the distance which can be calculated from the following equation:

$$d = c\frac{\Delta f}{f_h}\frac{1}{2f_m}$$

where f_m is the frequency of the triangular modulating signal and f_h is the frequency sweep.

The precision of a FMCW radar system depends mainly on the linearity of the triangular waveform of the frequency modulating signal.

In order to enhance the accuracy, sophisticated signal conditioning circuits utilizing digital signal processing procedures such as the Fast Fourier Transform are used.

7.8.3. *Properties of microwave sensors*

The reflectivity of a target is very important for the magnitude of the received signal. Generally, conductive materials and objects with high permittivity are good reflectors, while many dielectrics absorb energy and reflect very little. The best target for a microwave detector is a smooth, flat conductive plate positioned normally towards the detector.

Common FMCW systems work within the range of frequencies 9 GHz and 10 GHz or 24 GHz and 26 GHz. The minimum attenuation in the air is under 10 GHz; at higher frequencies, the attenuation increases and passes through the maxima and the minima (at 35 GHz and at 90 GHz).

Materials with low permittivity, e.g. synthetic materials, result in a lower attenuation than materials with high permittivity. Plastics and ceramics are quite transmissive and can be used as windows in the microwave detectors. The influences of dirt and dust sediments on the measurement are low.

7.9 Level measurement

7.9.1. *Introduction*

In principle there are two types of level measurements:

– continuous level measurement;

– detection of threshold values (limits).

Continuous level measurement provides information on the actual level. The *threshold detection* provides information whether given critical level(s) have been reached.

7.9.2. *Detection limits*

The sensors for level detection limits act as switches outputting a logical signal when a certain level has been reached.

7.9.2.1. *Capacitive level switch*

Whenever the bulk material has reached the position where the sensor had been installed the value of the sensor capacitance is higher than a preset value and switch is activated. The resolution of 0.01 mm can be reached and therefore they can be used for leakage control in oil tanks.

7.9.2.2. *Ultrasonic switch*

In ultrasonic sensors the time of travel t_f is evaluated and when the value of t_f reaches the set limit the output logical signal is generated. The sensor can also identify the presence of an object within the switch area.

7.9.2.3. *Vibrational switch*

The vibrational sensor uses a vibrator (rod, tuning fork) driven by a piezostrictive (or magnetostrictive) force and oscillates at its mechanical resonant frequency. If the material is in contact with the oscillating rod, its presence will dampen the oscillation amplitude which is sensed mostly by a piezoelectric sensor. When the vibration amplitude drops below a certain level, the switch is activated.

The sediments of the bulk material on the vibrating rod are removed by forced pulses of vibrations.

7.9.2.4. *Conductive sensors*

Conductive sensors consist of electrodes inserted into the container which measures the resistance of the bulk material. A conductive container can serve as a common electrode and in this case the sensor consists of only one electrode.

7.9.2.5. *Floating switch*

Due to the higher fluid density, the float (buoy) floats on the surface of the liquid. The float may carry a permanent magnet, which produces a magnetic field strong enough to activate a reed switch located at the position of the level being detected. Using two floats with a different buoyancy, it is possible to measure the respective level of immiscible liquids such as water and oil in a storage tank.

7.9.2.6. *Fiber optics level switches*

The function of these switches is based on the change of the optical fiber properties when the level of a liquid with appropriate refraction index reaches the end of the fiber immersed in the fluid. A more detailed description can be found in Chapter 2.

7.9.3. *Continuous level measurement*

7.9.3.1. *Principles of measurement*

A continuous level measurement sensor provides a signal proportional to the level of the material.

The following sensor principles can be used:

– capacitive;

– ultrasonic;

– microwave/radar;

– pressure difference (hydrostatic) methods.

7.9.3.2. *Capacitive sensors*

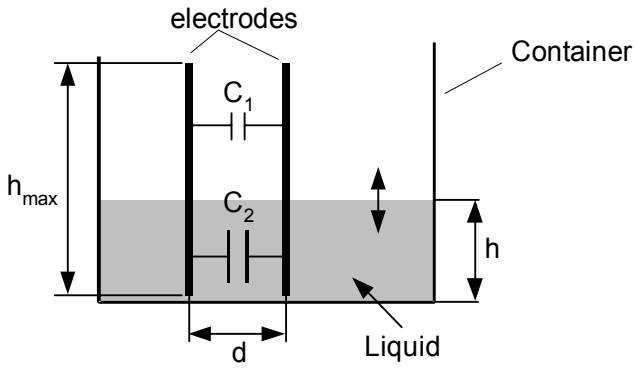


Figure 7.31. *Capacitive sensor with planar electrodes for level measurement*

Filling the space between electrodes with a material having a relative permittivity $\varepsilon_r > 1$ increases capacitance of the sensor proportionally to the relative permittivity and the level.

The capacitive sensors have planar or cylindrical electrodes. In Figure 7.31 the capacitive sensor with planar electrodes is depicted.

The total capacitance C is a result of a parallel connection of two capacitances C_1 and C_2 (Figure 7.31):

$$C = C_1 + C_2$$

Using C_0 for the capacity of the empty container and $\varepsilon_r = 1$ for air leads to a capacitance dependent on the level:

$$C = C_o + \varepsilon_o(\varepsilon_r - 1)\frac{bh}{d}$$

Figure 7.32 shows the capacitive sensor with cylindrical coaxial electrodes.

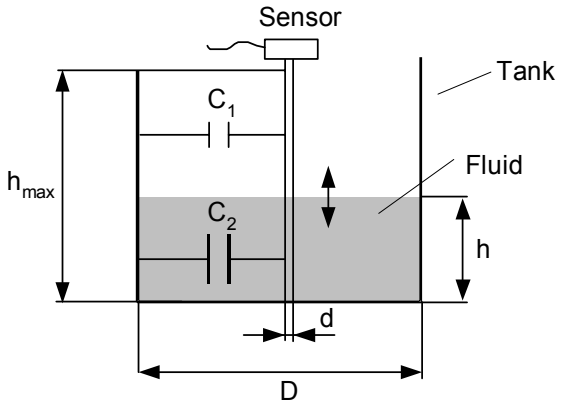


Figure 7.32. *Capacitive level sensor with cylindrical coaxial electrodes*

The capacitance of the cylindrical coaxial capacitor is given by the relation:

$$C = 2\pi\varepsilon_r\varepsilon_o \frac{h}{\ln(D/d)}.$$

The two capacitors C1 and C2 act as a parallel connection with the total capacitance:

$$C = C_1 + C_2 = 2\pi\varepsilon_o \frac{h_{max} - h}{\ln(D/d)} + 2\pi\varepsilon_r\varepsilon_o \frac{h}{\ln(D/d)}.$$

The first expression corresponds to the capacity of the empty container C_0.

Variation of the permittivity value causes systematic errors which could be compensated for by using an additional capacitive sensor which measures the permittivity.

Capacitance of level sensor also depends on density, concentration, temperature and humidity.

7.9.3.3. *Ultrasonic sensors*

The principle of operation of ultrasonic sensors is shown in Figure 7.33.

The transducer emits pulses of ultrasonic waves which are then reflected from interfaces between materials with different mechanical properties (discontinuities of acoustical impedance). The piezoelectric transducer can operate alternatively either as transmitter (piezostriction) or a receiver.

The distance d between the transducer and the level is found from the sound velocity and the travel time of the ultrasonic pulses. Compensation of temperature effect on sound velocity is carried out using data from an outside temperature sensor.

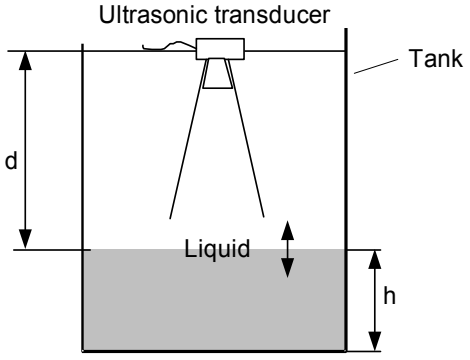


Figure 7.33. *The principle of ultrasonic sensors for level measurement*

The sensor axis must be perpendicular to the reflecting surface and reflections from tubes and bracing have to be avoided in order to produce an echo without disturbances (false echoes). Installation of the sensor in the focus of a parabolic tank cap increases the level of disturbing signals. The optimal position would generally be at half of the container radius in the center.

7.9.3.4. *Microwave sensors (radar)*

Regarding pulse-radar, the travel time of microwave signals is measured. Due to the high value of light velocity c_0, the duration of the pulses is only 1 ns, otherwise the transmitted and received pulse would overlap. Short pulses pose high requirements on time interval Δt measurement. Therefore, most of the level sensors use FMCW principle (section 7.8.2).

Microwave sensors are used in cases of high temperature, mist or dust, for rapidly moving objects or for long distance measurement.

Their disadvantage is the relative high cost of sensors compared to other types of level measurement principles.

The microwaves can penetrate non-conducting materials such as glass or nylon with low reflection, thus measurement can be made through the nylon container side without contact with the inner side of the container system. It is possible to choose the thickness of the penetrated material, so that the reflection of waves entering and leaving the material will cancel each other out. The resulting reflecting disturbance will then be zero. This is called *wave cancellation* and is based on the same principle as antireflection coating used for optical components and described in Chapter 2.

7.9.3.5. *Pressure difference (hydrostatic) sensors*

The pressure of a liquid or solid is proportional to the level h according to:

$$h = \frac{\Delta p}{\rho g}$$

where ρ is the density, g is the acceleration of gravity and Δp is the difference between the hydrostatic pressure in the liquid on the bottom of the container and in the space above the liquid level.

The principles and properties of pressure sensors are explained in Chapter 1.

7.10. Conclusions and trends

Current developments are directed to miniature sensors (micromechanical or MEMS) or to the use of new sensor materials for the improvement of measurement precision and the elimination of the environmental influences.

The development of sensor techniques in the field of position and distance measurement methods shows the growing importance of optical sensors. This trend was introduced with the development of image acquiring sensors (e.g. CCD camera) and digital image processing (machine vision). Thus, efficient solutions at acceptable prices can be offered.

Current trends are directed towards optical 3D measurement devices, which would allow full control of production processes.

The trend towards "intelligent" sensors through PC-based signal processing systems increases the possibilities of using "common" sensors. Thus, complex mathematical procedures are used to improve measurement results (filter, fuzzy-logic or complex analysis of single signals in neural networks), which opens new application fields.

7.11. References

[1] Profos, Pfeifer: Handbuch der industriellen Messtechnik, Oldenbourg 1994.

[2] J. Hoffmann: Messen nichtelektrischer Größen, VDI-Verlag.

[3] J. Niebuhr: Physikalische Messtechnik mit Sensoren, Oldenbourg 2002.

[4] K. Bonfig: Sensoren und Mikroelektronik, expert Ehningen 1993.

[5] P. Hauptmann: Sensoren, Hanser München 1990.

[6] E. Schoppnies: Lexikon der Sensortechnik, VDE-Verlag Berlin 1992.

[7] D. Bimberg: Messtechniken mit Lazern, expert Berlin 1993.

[8] Product information of MICRO-EPSILON MESSTECHNIK GmbH & Co. KG Königbacher Straße 15 D-94496 Ortenburg http://www.wiresensor.de.

[9] Product information of Newall Measurement Systems Ltd. http://www.newall.co.uk.

7.12. Online references

[1] Resistive LPD sensors

http://www.asm-sensor.de
http://www.altmann-gmbh.de
http://www.pewatron.com
http://www.baumerelectric.com
http://www.hengstler.de

[2] Inductive LPD sensors

http://www.orbitcontrols.ch
http://www.pewatron.com
http://www.twk.de
http://www.heidenhain.de/d0.htm
http://www.ruhle.com
http://www.unidor.de
http://www.micro-epsilon.de

[3] Magnetic LPD sensors

http://www.pewatron.com
http://www.heidenhain.de/d0.htm
http://www.newall.co.uk

[4] Capacitive sensors

http://www.micro-epsilon.de
http://www.capacitance-sensors.com
http://www.balluff.de
http://www.sie-sensors.com
http://www.pcb.com

[5] Optical sensors

http://www.pwb-technologies.com
http://www.safedoors.net/fraba_posital.html
http://www.renishaw.com
http://www.zeiss.de
http://www.abw-3d.de
http://www.lazer-zentrum-hannover.de
http://www.wenglor.de
http://www.lap-lazer.com
http://www.sitek.se
http://www.baumerelectric.com
http://www.thalheim.de
http://www.ivo.de
http://www.ifm-electronic.de

[6] Ultrasonic sensors

http://www.baumerelectric.com
http://www.novotech.co.at
http://www.sntag.ch
http://www.ad.siemens.de

[7] Microwave sensors

http://www.micas.de
http://www.milltronics.com
http://www.innosent.de

[8] Level measurement

http://www.milltronics.com
http://www.weka-ag.ch

Chapter 8

Temperature Sensors

8.1. Introduction

Drawing a good glass of beer, pasteurizing milk or producing electricity are all processes that require accurate temperature measurement. Various methods of performing the measurement, each with its own characteristics and possibilities, are described in this module. In the first part, concepts that are commonly used in thermal measuring techniques are explained. We start with a definition of heat and temperature and then give different methods to measure a temperature. The necessity of thermal equilibrium is demonstrated with an example.

The second part deals with ways to measure temperature: glass thermometer, liquid filled thermometer, liquid filled expansion thermometer, pressure temperature detector, vapor-pressure temperature detector and bimetallic thermometer.

These five possible ways to measure temperature are followed by measuring principles, the construction and application of sensors such as thermocouples, resistance temperature detectors (RTDs) and monolithic temperature sensors. In the final part the practical possibilities of pyrometry are described.

Chapter written by F. PEETERS, M. PEETERMANS and Luk INDESTEEGE.

8.2. Thermal measuring techniques

First some concepts and thermal units are explained: heat and temperature, static and dynamic readings, time constant and response time, thermal units, thermal equilibrium, temperature reading options and the quality of a measurement.

8.2.1. *Heat and temperature*

Heat is the total amount of kinetic energy of the molecules and atoms in a specific object. The molecules of each substance are constantly moving: in solids they move around a certain equilibrium point, while in liquids and gasses they basically move freely.

Because the molecules keep colliding with each other, their velocity changes all the time, both in size and direction. If there are no external influences, the average velocity remains unchanged. This average velocity changes when there is a change in temperature.

The absolute temperature of a body is proportional to the average kinetic energy of the molecules (so it is proportional to the mass and to the average velocity squared).

Adding or withdrawing heat does not necessarily result in an increase or decrease in temperature. When for instance heat (i.e. energy) is added to a piece of ice at 0°C the ice will melt and change into a liquid (water at 0°C). This heat is called *latent heat*.

8.2.2. *Static and dynamic readings*

To measure the water temperature in a pot of warm water, we place a thermometer in the water and wait a few minutes to let the water warm up the thermometer. Then we read the temperature; this is quite accurate because the temperature of the water has not changed much. When the pot is placed on a heater, the temperature will be hard to read because it rises continuously. It increases until the water is boiling. While the water temperature changes, the thermometer reacts with a time delay.

Reading the temperature when the pot is not heated or when the temperature remains stable is called a static reading (“steady-state”). When you heat the pot or when the temperature changes it is called a dynamic reading (“unsteady-state”). At the moment that the water temperature and the thermometer temperature change

there is a difference between them. This is called the dynamic error. Some thermometers react quickly to changes, having a fast response and a short response time. The dynamic error is small if we have fast responses and short response times or if we slowly heat up the water.

8.2.3. *Time constant and response time*

The response time is the period of time that renders how quickly or slowly a thermometer reacts to a change. Let us assume that there is an abrupt change at the entry (Figure 8.1).

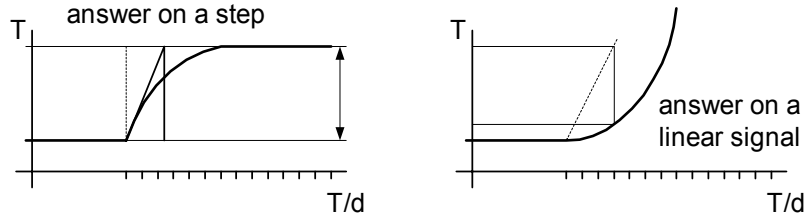


Figure 8.1. *Response time*

The output of the reading device shows a gradual change. This abrupt change that in theory occurs immediately is called step disturbance. The response time is expressed in time constants. With a first order process, τ is the time that the reading instrument needs to reach 63.2% of the step. The total response time is usually considered to be 5τ. A second method to measure response time is to use a periodically alternating signal, usually sinusoidal. This method is called frequency analysis. The result of this analysis is a frequency response. After a mathematical operation, the results of a frequency response yield the same information as the step response.

8.2.4. *Thermal units*

According to the International Unit system the unit of energy is expressed in *Joule* (J), while the older system uses calories (cal).

$$1\text{cal} = 4.18\text{J}$$

Temperature is expressed in Kelvin (K) or, more commonly, degrees Celsius (centigrades), degrees Fahrenheit and Rankine.

To draw up a temperature scale two invariable and easy to reproduce thermal conditions are chosen. Arbitrarily temperature values are attributed to them to obtain two reference temperatures T_1 and T_2. At the same time this determines the size of the graduation. (Figure 8.2)

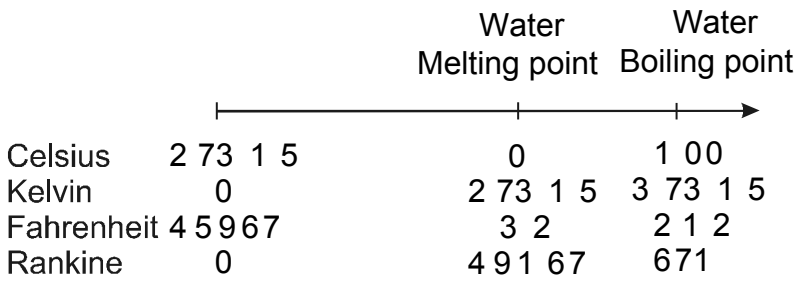


Figure 8.2. *Reference temperatures*

The same applies for centigrades:

– T_1: the melting point of ice – the freezing point of water (p_{atm} = 1.1013 bar) = 0°C.

– T_2: the boiling point of water – the condensation point of vapor (p_{atm} = 1.1013 bar) = 100°C.

Conversions can be made using the following formulae:

$$T_{°C} = \frac{5}{9}(T_{°F} - 32)$$

$$T_K = \frac{5}{9}(T_{°F} + 459.67)$$

8.2.5. *Thermal equilibrium*

When two bodies with a different temperature are connected to each other, heat will flow from the warmer body to the colder body. After some time they will both

reach the same temperature. This is called *thermal equilibrium*. This process has important consequences that can strongly influence the reading, both positively and negatively.

Close contact between the measuring probe and the substance to be measured accelerates the heat transfer. The thermal resistance of the surface layer is very important.

The *thermal capacity* of the reading instrument needs to be small enough compared to the body to be measured, on the one hand not to influence the temperature of the body, on the other hand to be able to faithfully follow changes in temperature.

The unit of the measuring instrument and the body to be measured needs to be sufficiently separated from the external world. *Heat transfer* from and to the surroundings can influence the reading.

When larger volumes are used, a *temperature gradient* can occur (e.g. temperature variations in a large room). The quality and quantity of this gradient, in combination with other factors, determine the way of reading. Sometimes *protective covers* or *thermowells* are used. Two different types exist: threaded thermowells and flanged thermowells (Figure 8.3).

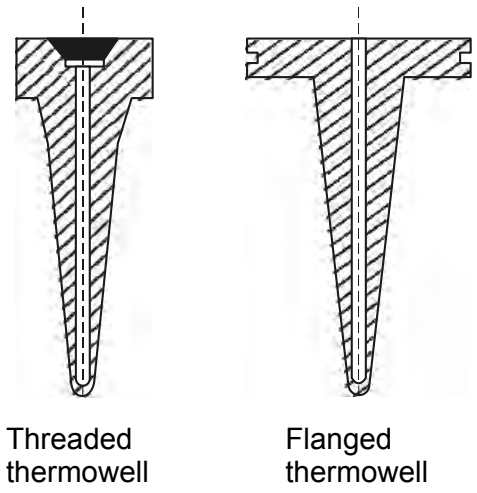


Figure 8.3. *Thermowells*

Threaded thermowells are usually made of stainless steel. Flanged thermowells can be made of stainless steel, monel metal, nickel, hastelloy or of clad metal for special uses. The plating can be silver, tantalum, lead, glass, etc.

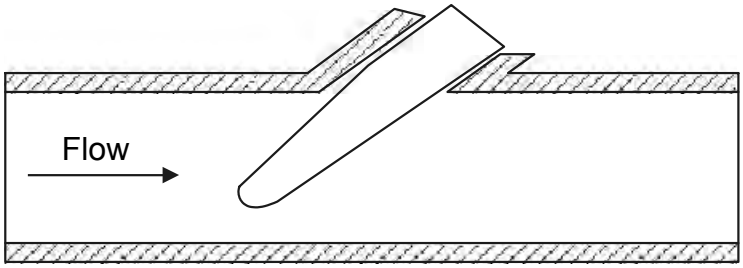


Figure 8.4. *Thermowells in a tube*

In general a threaded thermowell is used in those places where no corrosion occurs, while flanged thermowells are used in all circumstances.

When a thermowell is placed into a pipe or tube, it is necessary that the heat transfer is optimal. Usually, thermowells are placed at an angle in the opposite direction of the flow (Figure 8.4). An incorrect installation of thermowells can cause large reading errors. Thermowells (with reading device) can be placed at several positions in a tube (Figure 8.5).

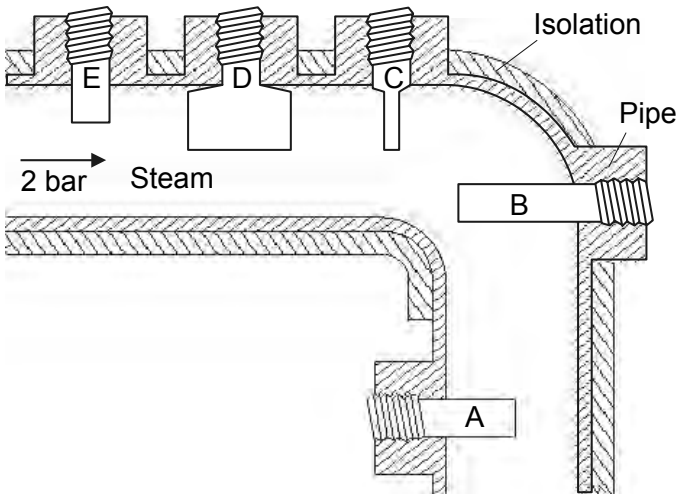


Figure 8.5. *Reading positions for thermowells*

The table below gives the reading errors for the different positions.

Position	Reading [°C]	Reading error [°C]
A	341	45
B	386	0
C	385	1
D	384	2
E	371	15

Table 8.1. *Reading errors for different positions*

– *Position A*: the tube has not been insulated. This causes losses in heat transfer because a heat flow runs through the thermowell into the exterior. In addition, the placement of the thermowell in the direction of the flow is not optimal. For a good temperature measurement a position on the outside (near A) would give a slightly better result, because the turbulence caused by the centrifugal force would be larger and that results in a better heat transfer. At an actual temperature of 386°C, it would not be strange to read a temperature of 341°C in position A.

– *Position B*: the tube has been insulated and the direction of the flow is ideal to measure the temperature. Thus, no reading error is expected.

– *Position C*: the heat transfer is slightly lower than in position B, because the direction of the flow is slightly less favorable. However, a thin-walled thermowell results in few losses in heat transfer so that only a small reading error of ± 1°C can be expected.

– *Position D*: these are the same conditions as in position C, but with a thick-walled thermowell. This results in more heat transfer losses compared to position C.

– *Position E*: the conditions are exactly the same, but the thermowell is shorter. Thus, the heat transfer will be worse and a larger reading error is to be expected.

The following factors determine a good temperature measurement:

– The position of the thermowell in the flow.

– The depth of insertion.

– The thickness of the wall of the thermowell.

– Insulation of the pipe.

Thermowells increase the response time. Suppose that a measuring instrument in a pipe reads a temperature of 25°C; at time T_1 the liquid flow suddenly acquires a temperature of 50°C.

Figure 8.6 shows the delay with and without a thermowell.

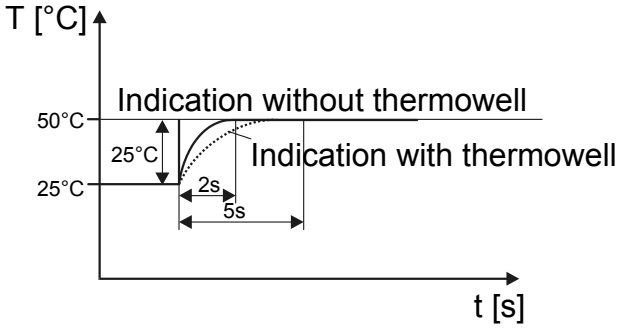


Figure 8.6. *Time constant in thermowells*

8.2.6. *Temperature measuring options*

Characteristics for choosing the type of temperature measurement are as follows:

– desired temperature span;

– desired accuracy, linearity, sensitivity, hysteresis, etc.;

– desired time constant;

– desired mechanical strength and built-in possibilities;

– desired readability (reading on the spot or at a distance);

– desired output unit;

– usefulness of the reading in a specific medium;

– relation of price to quality and budget.

For an accurate temperature measurement, several physical and electrical properties can be used:

– The expansion or contraction of a substance when it is heated or cooled down (bimetallic thermometer, liquid filled thermometer).

– The occurrence of a thermoelectric voltage at the connection of two metals (thermocouples).

– The changing of electrical resistance of a substance when the temperature is changed (resistance temperature detectors, thermistors).

– The changing of the properties of radiated electromagnetic waves at temperature changes (infrared photography, pyrometry).

8.2.7. *Quality of a measurement*

Every measurement has accuracy limits. The best approach uses the most accurate measuring instrument, applies the best measuring technique and takes the average of several measurements. Backlash, friction and other defects cause repeatability errors and if we measure the same value again, we do not get the same reading. There will also be a restriction in the resolution with which the reading is registered or rendered. The measuring system will fall within a certain span and have a specific, preferably linear, transfer within this span. These factors are generally referred to as the accuracy of a measurement.

Every measuring instrument has a minimal starting point and a maximum end point. The starting point of a reading is called zero; the difference between the end point and the starting point is called a span or range. The end point of a reading is also referred to as full scale.

Because the range of a process variable differs in each application, industrial reading devices are designed in such a way that zero and span can be adjusted.

The changing of zero and span depends on *repeatability, resolution, accuracy* and other sources of errors. Reducing the span for instance can result in increasing accuracy. When choosing a measuring instrument for a certain application, it is therefore advisable to choose an instrument with a range close to the range of the application.

8.3. Physical or direct temperature measurement

8.3.1. *Glass thermometer*

Principle: liquids that are warmed in a small reservoir and then pressed into a narrow tube along a graduation give useful thermometers.

Of all *liquid filled thermometers*, the *mercury thermometer* is the best known. Mercury, however, is not fit for lower temperatures (solidification point –39°C). However, it is suitable for higher temperatures, even above its boiling point (360°C or 633 K). In that case, precautions need to be taken, for instance by fitting a high-

pressured (80 bar or 8 Mpa) nitrogen filling above the mercury. For lower temperatures, alcohol (–110 to +50°C), pentane (–200 to +20°C) or toluene (–70 to +100°C) can be used.

Application: glass thermometers are very suitable for calibration. Because of their simplicity and accuracy they are extremely suitable for use in laboratories. In practice, however, glass liquid filled thermometers are not applied in the processing industry. They are too fragile for the operating conditions of the processing industry. If for instance a mercury thermometer broke during a food production process, the toxic mercury would come into contact with the food. As the glass thermometer cannot generate a signal (standard electrical signal), it is also not well-adapted for control engineering. The accuracy of glass thermometers depends on the quality, the span and the extent of immersion (complete or partial immersion).

8.3.2. *Liquid filled expansion thermometers*

Principle: liquid filled thermometers consist of a bulb connected with a capillary and a Bourdon element, a bellow or a diaphragm (Figure 8.7).

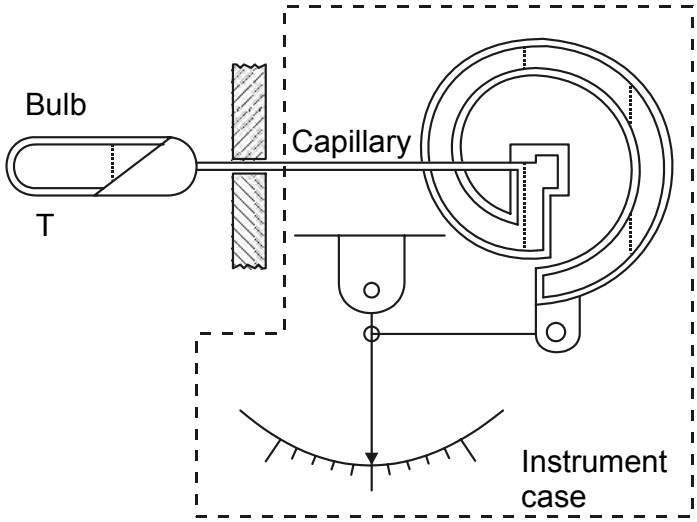


Figure 8.7. *Liquid filled thermometer*

The entity is closed and filled with a liquid. The filling occurs under high pressure (up to ca. 70 bar) to avoid every influence of the vapor pressure (highly raised boiling point).

	Minimum temperature	Maximum temperature
Mercury	–39°C	528°C
Xylene	–40°C	400°C
Alcohol	–46°C	150°C

Table 8.2. *Application temperatures of liquid-filled thermometers*

The volume of the bulb is warmed up by the substance to be measured. The volume of the liquid in the bulb will expand and unroll the free end of the Bourdon element through the capillary.

Practical realization: because the capillary can also expand or contract under the influence of a temperature change, it is important to keep the capillary as short and as thin as possible.

This is not always possible and therefore a compensation method has been developed which allows a maximum distance of 60 meters between the bulb and the indication. Using the second capillary in parallel with the first it is connected to a second Bourbon element. The second element works in the opposite direction and compensates for the errors of the first (Figure 8.8).

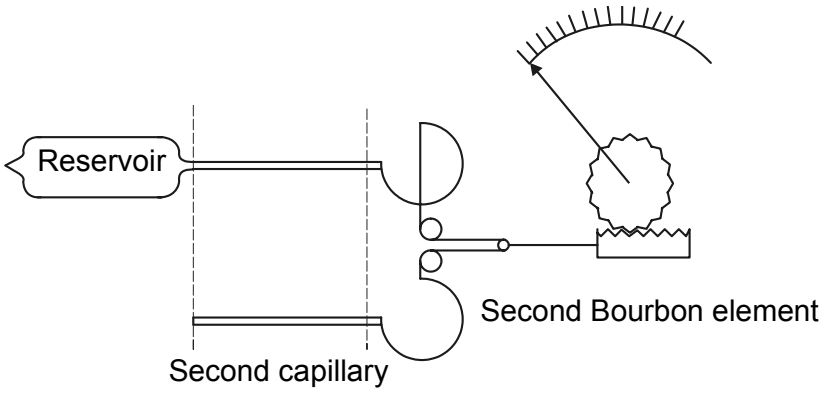


Figure 8.8. *Compensation with liquid filled thermometers*

The liquid filled expansion thermometer is an old-fashioned but reliable device, which is particularly interesting because no auxiliary energy is needed. The accuracy is about 0.5% FS.

8.3.3. *Gas filled expansion thermometer or pressure thermometer detector*

Principle: gas filled expansion thermometers are built in the same way as liquid filled thermometers, but the whole volume is filled with high-pressured gas. They operate in a slightly different manner than liquid filled expansion thermometers. At a first approach we can assume that the expansion of the bulb is negligibly small compared to the total volume, which means that we are dealing with a change of state at a constant volume.

The universal gas law applies to this:

$$p.V = n.R.T$$

The volume remains constant, so the pressure will change proportionally to the temperature. The Bourdon tube will purely expand following the pressure principle:

$$p_1 = \frac{T_1}{T_2 - T_1} . y$$

where:

– y: pressure span of the Bourbon tube

– p_1: pressure at temperature T1 [K]

– p_2: pressure at temperature T2 [K]

The formula above shows that the transfer between pressure and temperature is linear.

Practical realization: as with liquid filled thermometers temperature fluctuations of the capillary and the Bourbon tube cause false readings. A commonly used compensation is the double system and the bimetallic strip (Figure 8.9).

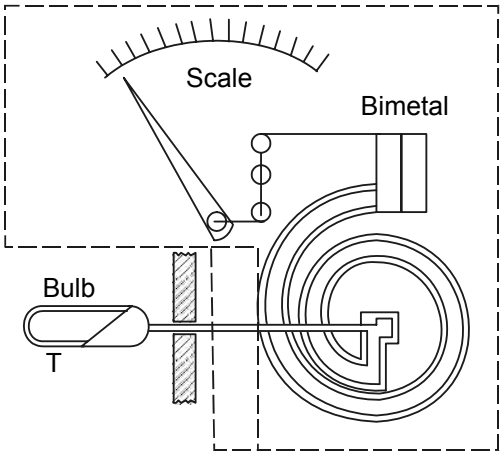


Figure 8.9. *Compensation with bimetallic strip*

In contrast with the liquid filled expansion thermometer, the level difference between the bulb and the Bourbon tube is negligibly small because of the low density of nitrogen.

Barometric pressure changes can cause small errors depending on the pressure of the filling.

The most frequently used gas is nitrogen, due to its almost ideal gas properties and its large coefficient of expansion. Lower temperatures can be attained with a helium filling.

Applications: this thermometer can be used almost everywhere because of its non-toxic nitrogen filling. Examples are the food industry, mechanical engineering, the pharmaceutical and chemical industry. Maximum spans from –250°C to +800°C are feasible with an extreme accuracy rate of ±0.6% of the span. The distance between the bulb and the indication is 100 m maximum.

8.3.4. Vapor-pressure systems

Principle: the vapor filled thermometer looks essentially the same as the liquid filled thermometer. The filling, however, consists of liquid and vapor (of the same liquid).

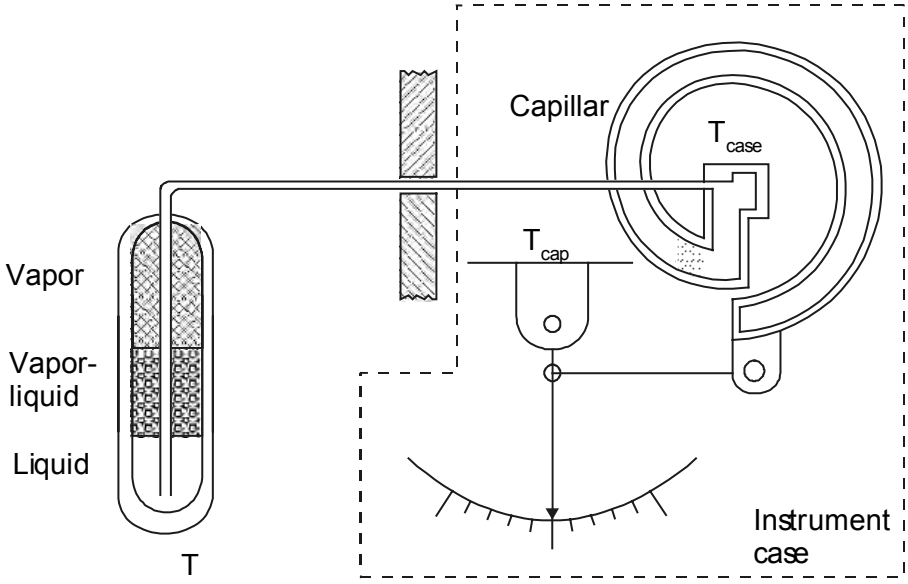


Figure 8.10. *Vapor-pressure thermometer*

The boundary between liquid and vapor is always situated inside the bulb, above the exit of the capillary. Liquid and vapor are in equilibrium. When the temperature rises, a small amount of liquid will evaporate, the pressure in the system rises (Figure 8.10). This causes the creation of a new equilibrium. The vapor pressure curve (Figure 8.11) of these liquids is non-linear, which results in a non-linear scale.

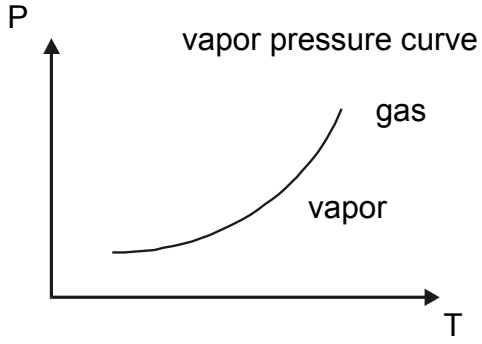


Figure 8.11. *Vapor pressure curve*

Practical realization and applications: liquids that qualify for this kind of measurement are butane, propane, hexane, toluene, etc. The span depends on the liquid used and can range from approximately 50°C to 260°C. This measurement is cheap, but suffers from inaccuracy, non-linearity and possible gross errors.

8.3.5. *Bimetallic thermometer*

Principle: when two metallic strips, with different coefficients of expansion α_A and α_B, are attached to each other at a specific temperature, changes in temperature will cause the strips to expand differently, and the compressed strip will experience circular bending (Figure 8.12).

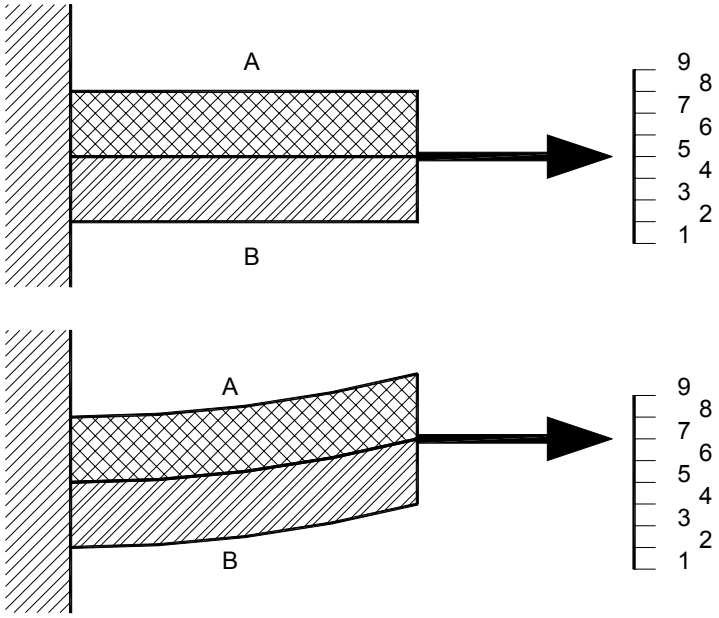


Figure 8.12. *Bimetallic thermometer*

The radius of the curve ρ is given by:

$$\rho = \frac{2.t}{3.(\alpha_A - \alpha_B).(T_2 - T_1)}$$

where

– t: total thickness of the strip (practically 12 μm < t < 3.5mm);

– $T_2 - T_1$: change in temperature.

Practical realization: because there are no metals with a negative coefficient of expansion that can be used in practice, Invar is usually chosen as the A-element (64% Fe, 36% Ni, $\alpha_A = 0.2 \cdot 10^5$ [K^{-1}]). Originally bronze was used for the B-strip (90% Cu, 10% Sn, $\alpha_B = 1.9 \cdot 10^{-5}$ [K^{-1}]), but depending on the requirements a wide spectrum of alloys is available nowadays.

To obtain a large span, the bimetallic strip can be rolled up into a spiral or a helix (Figure 8.13).

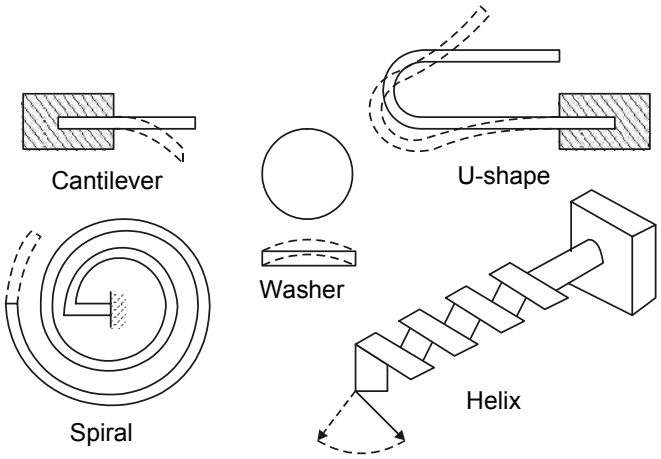


Figure 8.13. *Realizations of bimetallic thermometers*

A span ranging from –50°C to +500°C is feasible. The accuracy is about 1% of the span.

Application: bimetals have a very wide range of applications, from simple switches to industrial measurements. Due to the low price this method is frequently used for applications in which accuracy is less important. Measurements at a distance are possible but discouraging.

8.4. Thermoelectric measurements (thermocouples)

8.4.1. *Measuring principle: thermoelectricity*

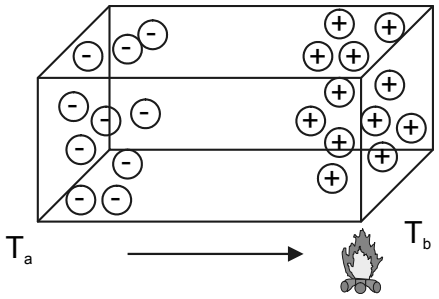


Figure 8.14. *Thermal diffusion of electrons*

When a bar or a ring of a homogenous conductor is heated locally, the concentration of free electrons will no longer be constant at every place of the material. The free electrons seek out the lowest energy point and diffuse towards the colder part. The warmer part becomes positively charged compared to the colder part (Figure 8.14). At a specific temperature difference a dynamic equilibrium arises; the generated *thermal voltage* will create an electric field that counteracts the diffusion of the electrons. The voltage between two points a and b is proportional to the difference in temperature and the Seebeck coefficient.

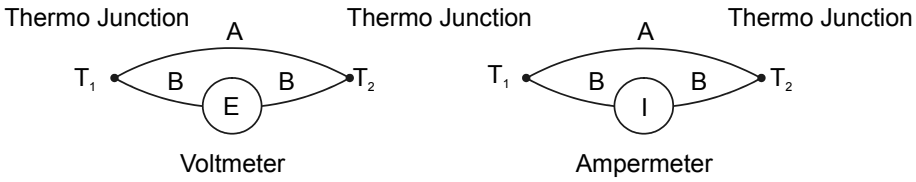


Figure 8.15. *Basic thermocouple*

When two conductors made of different materials A and B are connected as in Figure 8.15, with one junction at temperature T_1 and the other at temperature T_2, a voltmeter (with infinite internal resistance) is able to read an electromotive force (EMF) E. This is the Seebeck effect. The voltage E depends on the materials used (the difference of their Seebeck coefficients) and on the difference in temperature between T_1 and T_2.

When we replace the voltmeter by an ammeter, we see that a current flows in the circle. Because of this flow electrical energy can be generated, but only to a limited extent.

The effect is reversible, so when we send a current from an external source through the thermoelectric circuit, one junction will warm up and the other will cool down (Peltier effect).

Whereas we are interested in the thermoelectric effect as a way of measuring temperature, contemporary material research is developing the application of thermoelectricity for generating power, heating and cooling.

Thermocouples are based on the Seebeck effect, which is caused by thermal diffusion of electrons. Thermoelectric voltage is not caused by a contact potential (a common error in textbooks). Thermocouples are also not based on Peltier and Thomson effects, as these two exist only when the current is flowing in the circuit.

The set-up in the previous figure is called a *thermocouple*. For this thermocouple we can fairly accurately write:

$$E = C_1.(T_1 - T_2) - C_2.(T_1^2 - T_2^2)$$

where

– E: total voltage, expressed in [V] or [μV]

– T_1 and T_2: absolute temperature of joints A and B [K]

– C_1 and C_2: thermoelectric material constants

For example, for a copper/constantane thermocouple:

$$E = 37.5.(T_1 - T_2) - 0.045.(T_1^2 - T_2^2) \ [\mu V]$$

8.4.2. Thermoelectric laws

Practical temperature readings with thermocouples are based on some thermoelectric laws. These laws are formulated as follows:

– The resulting generated voltage E in a thermocouple with the joints at temperature T_1 and T_2 does not depend on the temperature elsewhere in the circuit (Figure 8.16).

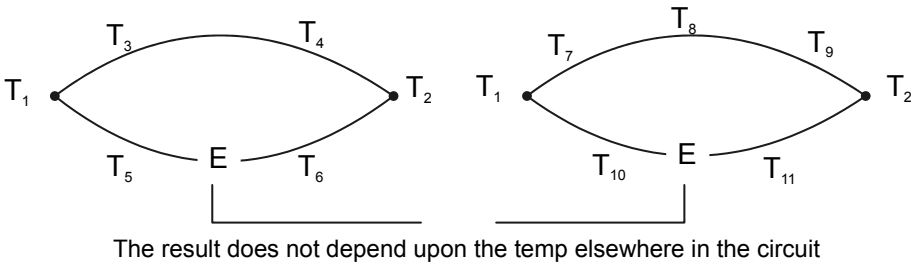


Figure 8.16. *First thermocouple law*

– When a third homogenous metal C is placed within either A or B, and the new joints remain at constant temperature T3, the resulting voltage E will be the same as without metal C (Figure 8.17).

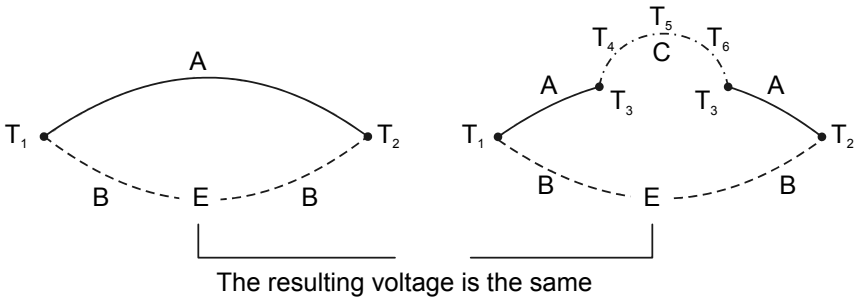


Figure 8.17. *Second thermocouple law*

– When metal C is inserted between A and B and the joints BC and CA remain at a constant temperature T1, the resulting EMF E will be the same as without metal C (Figure 8.18).

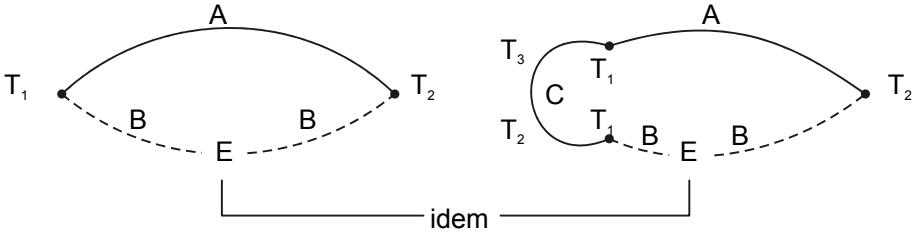


Figure 8.18. *Third thermocouple law*

– When the EMF of metals A and C is EAC and the EMF of B and C is ECB, the EMF of A and B is EAC + ECB (Figure 8.19).

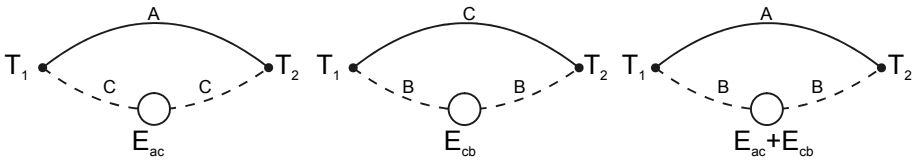


Figure 8.19. *Fourth thermocouple law*

– When a thermocouple gives EMF E1 at joint temperatures T1 and T2 and EMF E2 at temperatures T2 and T3, it will give voltages E1 + E2 at temperatures T1 and T3 (Figure 8.20).

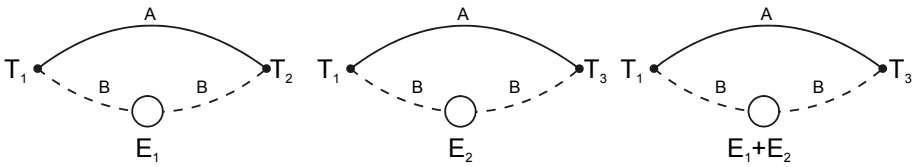


Figure 8.20. *Fifth thermocouple law*

These five laws are very important for the practical application of thermocouples.

The first law demonstrates that the connecting wires of the two joints can be exposed to unknown changing temperatures (of the surroundings) without a change in the resulting EMF.

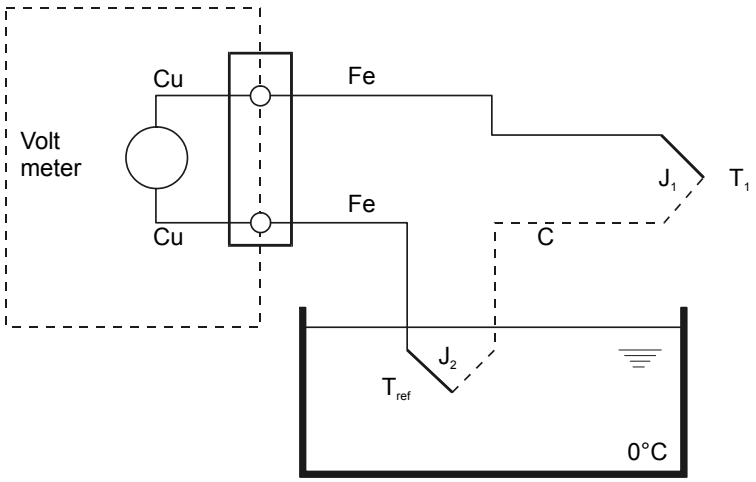


Figure 8.21. *Realization of a thermocouple*

Laws 2 and 3 allow the insertion of a voltmeter in the circuit in order to measure the resulting voltage. The metal C represents the internal circuit (usually copper) of the measuring instrument (Figure 8.21). The instrument can be connected in two different ways, between A and B or within A (or B) (Figures 8.17 and 8.18).

The third law shows that the thermocouple axle can be soldered without a changing EMF.

The fourth law proves that every metal can be calibrated against a standard metal (usually platinum) and that starting from this the EMF can be calculated between each metal.

When we look at *the fifth law*, the comment has to be made that the temperature of one of the joints needs to be known in order to measure an unknown temperature with a thermocouple. This joint will be kept at a reference temperature and is called a *reference joint*. Similarly the temperature of the measuring joint can be read as long as the voltage is known.

8.4.3. *Practical temperature measurement with thermocouples*

In a nutshell Seebeck voltage is a measure for the temperature T_{wj} of the warmer joint, if we can keep the *colder joint* at a reference temperature (e.g. 0°C). We cannot calculate the temperature T_{wj} accurately enough because the thermoelectric constants C_1 and C_2 do not remain constant at changing temperatures (Figure 8.22).

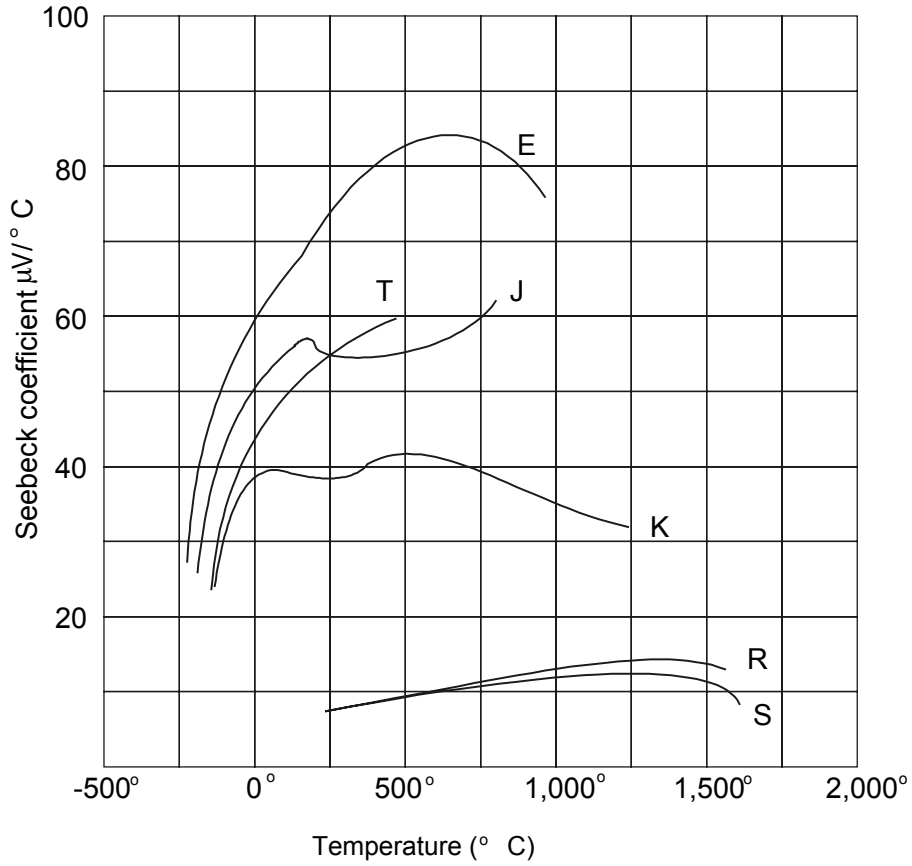


Figure 8.22. *Seebeck coefficient*

The diagram also shows that the transformation of temperature to voltage is non-linear (Figure 8.23).

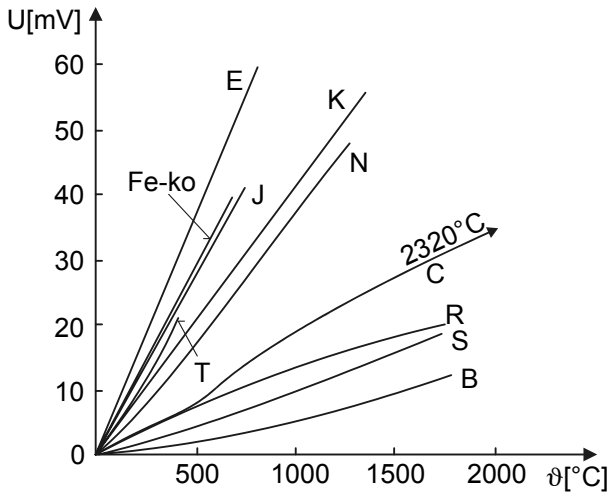


Figure 8.23. *Temperature versus voltage*

The thermocouple tables of the NBS (National Bureau of Standards) are used for the exact figures. For a specific type of thermocouple and at a reference temperature of 0°C, these tables show the temperature of the warmer joint that corresponds to the measured voltage.

The colder joint has to be kept at a constant temperature. The best solution is to keep the colder joint at a reference temperature of 0°C or to reduce the generated voltage to the voltage that corresponds with 0°C. This is called *cold joint compensation.*

The cold joint compensation can be done with software or with hardware.

Software compensation: by means of a temperature sensitive sensor we measure the temperature of the colder joint at the isothermal connecting block (isothermal: with the same temperature), and we use that information to calculate the unknown temperature T_1. A multimeter operated by a microprocessor that is especially programmed to measure temperatures (Figures 8.24 and 8.25) performs this calculation.

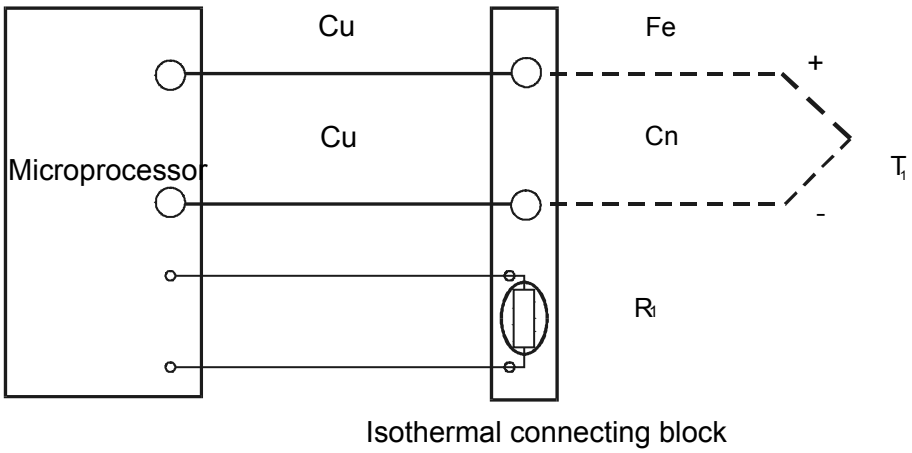


Figure 8.24. *Software compensation*

The sensor Rt can be any instrument that is capable of reading linear absolute temperatures, a Pt-100, a thermistor or a temperature IC. Conversion of temperature to voltage and the other way around can be obtained by storing the NBS tables in the memory.

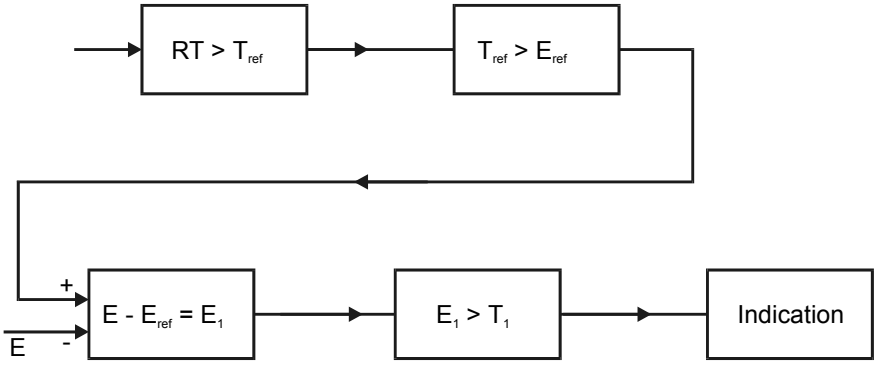


Figure 8.25. *Block diagram for software compensation*

Software compensation is the most versatile technique for measuring temperatures with thermocouples. Many thermo couples can be connected to a computer, whatever the type of thermocouple. A multiplex system takes care of the measuring data which is subsequently input into the computer. In this way the

advantages of *data acquisition* by computer can be combined with temperature measurement.

Hardware compensation: instead of measuring the colder joint temperature and calculating the corresponding voltage as illustrated in the software method, we can also include a voltage source in the circuit in order to eliminate the offset voltage in the colder joint. The combination of this hardware compensation voltage and the cold joint voltage is the voltage at a joint at 0°C (Figure 8.26).

Voltage E now refers to 0°C and the temperature can be read directly from the NBS tables. The largest disadvantage is that another switch is needed for every type of thermocouple.

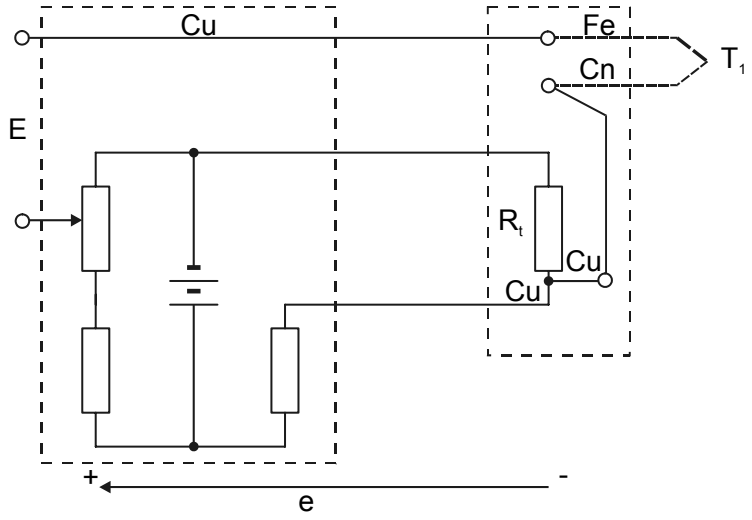


Figure 8.26. *Hardware compensation circuit*

8.4.4. Technological realizations of thermocouples

One of the most important steps in the use of thermocouples is the selection of the metal combination. Only a few of the enormous amount of possible combinations are actually used.

This choice is determined by:

– thermoelectric power;

– stability;

– reproducibility;

– electric resistance (as small as possible);

– melting point (as high as possible, depending on the application);

– mechanical qualities;

– price.

Some of these metal combinations are standardized. The most frequently used standards are the DIN standards and the ISA standards and an overview will be given of both.

The practical realization of thermocouples is more demanding than one would expect at first. The choice of materials, the way of connecting and control over this connection during the complete production process are all highly important to respect the imposed tolerances.

Different *realizations* are:

– *Ceramic beads insulation*: ceramic beads, which are available in various shapes and sizes, assure the electric insulation between both wires. Advantages are: low price, simplicity, and smaller time constant. Disadvantages are: mechanical sturdiness and bare thermocouple wires (Figure 8.27).

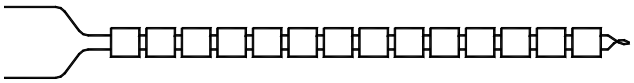


Figure 8.27. *Ceramic beads*

Casing thermocouples: in this type of thermocouple, both thermocouple wires are covered by metal and they are separated by, e.g., Al_2O_3 or MgO (Figure 8.28). The advantages of these thermocouples are: sturdiness, flexibility and weldable RVS casing.

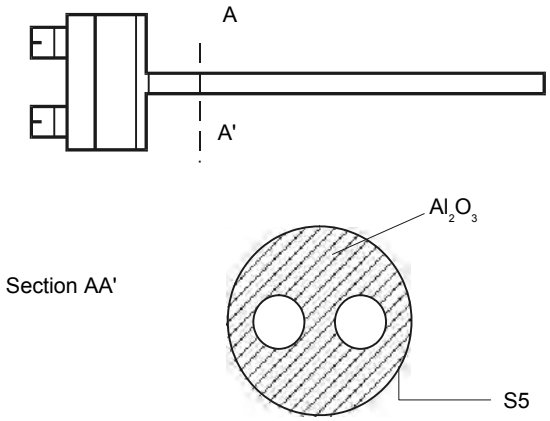


Figure 8.28. *Casing thermocouple*

Film thermocouple: film thermocouples are used for temperature readings at the surface, and consist of flattened thermal wires that are affixed to a polymer film. Advantages are: cheapness, small time constant, simple assembly and small size. Disadvantages are: fragility, small temperature range and inaccuracy. A film thermocouple is shown in Figure 8.29.

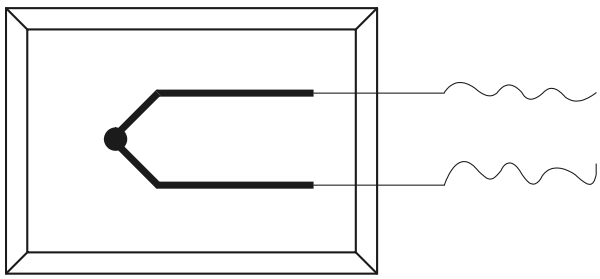


Figure 8.29. *Film thermocouple*

The distance between the reading instrument and the warmer joint can be quite large. As mentioned earlier the colder joint should be affected as little as possible by changes in temperature. It is thus advisable to move the colder joint to a place where the ambient temperature is as stable as possible. A *compensation wire* bridges the distance between the measuring instrument and the warmer joint (Figure 8.30). This wire possesses the same thermoelectric qualities as the thermocouple used for temperatures between 0 and 200°C and it does not influence the Seebeck voltage. Compensation wire is much cheaper because it can be made of materials different

from the thermocouple itself. The compensation wires have specific color codes, which possibly can be confusing.

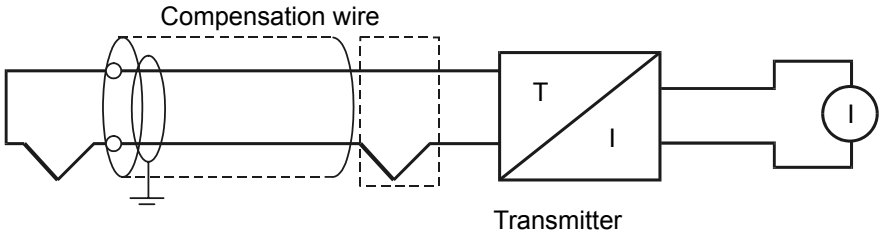


Figure 8.30. *Compensation wire*

8.4.5. *Applications*

Thermocouples are often used because of their large span. Thermocouples can also be used in different surroundings. Their construction is sturdier, so that they are often welded on a metal body or pressed under a screw. The use of thermocouples is as simple as connecting wires. A converter can take care of the conversion of the millivolt signal into a standard signal and of the colder joint compensation. Every thermocouple can be subjected to interference, especially as the generated voltage is quite low and approaches voltages that are generated by *electromagnetic noise* or *ground circuits*.

The following precautions can contribute to avoid interference:

– Shielding the wires;

– Using twisted pairs;

– Grounding in only one place;

– Proper, moistureproof insulation;

– Differential input of amplifier;

Just as for every other type of temperature measurement, a thermocouple has a time constant. The following table provides some typical response times of insulated and grounded thermocouples for a step from 25 to 100°C.

External diameter [mm]	Response time [s]	
	grounded	isolated
1	0.07	0.11
1.6	0.09	0.28
3	0.34	1.6
4.8	0.7	2.6
6	1.7	4.5

Table 8.3. *Response time variation with size*

8.4.6. *Parallel and series connections of thermocouples*

When applying heating and cooling in relatively large spaces it can be necessary to perform *average readings*.

The average temperature is defined as the mathematical average of the separate temperatures:

$$T_G = \frac{T_1 + T_2 + T_3 + ...T_n}{n}$$

Thermocouples can be placed in series or in parallel, each with its own advantages and disadvantages. In each case the thermocouples have to be of the same type to evaluate their voltage behavior. Circulation currents could lead to substantial measuring errors.

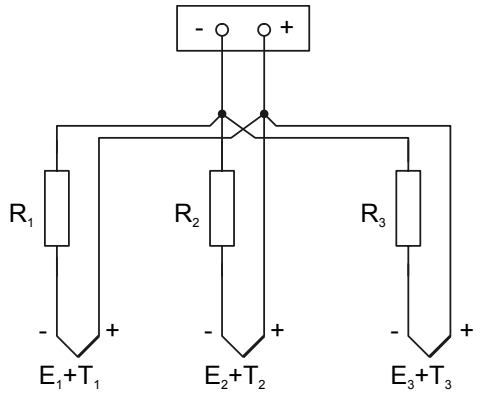


Figure 8.31. *Thermocouples, connected in parallel*

This circuit can be considered as a parallel connection of three voltage sources (possibly more) with an internal resistance R_1, R_2 and R_3, measured at the common joint of every thermocouple.

The accuracy of the average reading is larger as R_1, R_2 and R_3 approach each other. The resistance can always be adjusted until $R_1 = R_2 = R_3$. Of course limitations as far as the total resistance is concerned do apply.

Advantage: the connection can be calibrated as if it were only one thermocouple.

Disadvantage: there is no error indication when one of the thermocouples drops out.

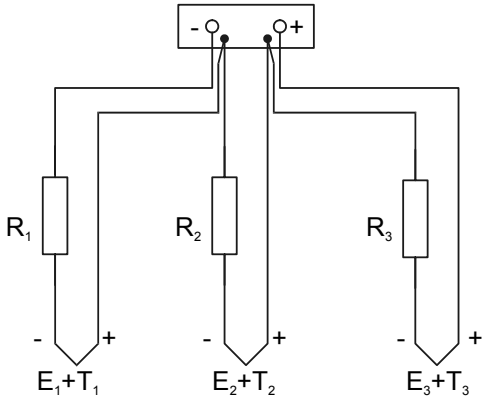


Figure 8.32. *Thermocouples, connected in series*

The output signal is $E_r = E_1 + E_2 + ... + E_n$.

This connection can be replaced by a source E_r in series with a resistance $R = R_1 + R_2 + ... + R_n$. Thus, every thermocouple functions like a voltage source with output E_r and resistance R. When you divide the reading by N, you find the average voltage or the average temperature T_g. The average temperature does not depend on the different resistances of the wiring.

Advantage: good sensitivity and any drop out of a thermocouple would be noticed immediately.

8.5. Resistance temperature detectors (RTDs)

8.5.1. *Principle*

For most metals the change of resistance R with temperature T can be expressed in an equation of the form:

$$R = R_0.(1 + \alpha_1.T + \alpha_2.T^2 + ... + \alpha_n.T^n)$$

in which resistance R_0 is the resistance at 0°C.

The number of terms of the equation depends on the material, the required accuracy and the temperature span. Platinum, nickel (and less often copper) are the most commonly used metals for RTDs and usually need 2 to 3 α-values for highly accurate measurements (Figure 8.33)

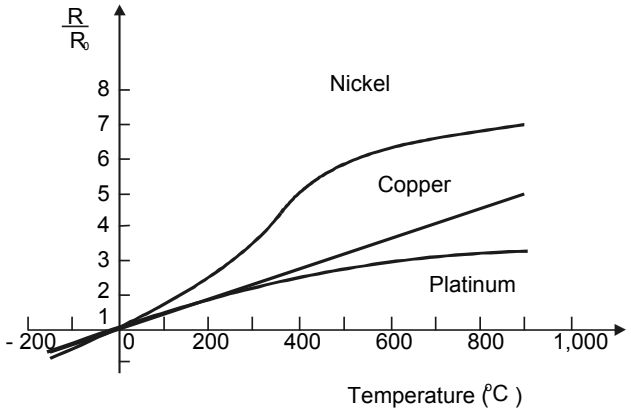


Figure 8.33. *Temperature versus resistance*

When normal accuracy and linearity are aimed for, and the span is not too wide, and the equation reduces to:

$$R = R_0 \cdot (1 + \alpha_1 \cdot (T - T_0))$$

For platinum $\alpha = 3.850.10^{-3}$

For a resistance thermometer, W is defined as: $\frac{R_{100}}{R_0}$

where

– R_{100}: the resistance at 100°C

– R_0: the resistance at 0°C

E.g. for a Pt-100 element W = 1.385

The resistance temperature coefficient, depending on the used materials and the purity of these materials, is defined by:

$$\alpha_0^{100} = \frac{R_{100} - R_0}{100 \cdot R} \quad [\frac{\Omega}{\Omega.°C}] = [°C^{-1}]$$

For a Pt-100 element (DIN 43.760) this is:

$$\alpha_0^{100} = \frac{138.5 - 100}{100 \cdot 100} = 3.850 \cdot 10^{-3} \ °C^{-1}$$

The name Pt-100 refers to a platinum resistance element with a resistance of 100 Ω at 0°C.

So a Pt-500 element has a resistance of 500 Ω at 0°C.

The difference between both is the change in resistance per °C:

– Pt-100: 0.385 Ω/°C;

– Pt-500: 1.960 Ω/°C.

Thus a larger sensitivity is attainable with the Pt-500.

A RTD is a resistance device and it needs measuring current to generate a useful signal. Because this current heats the element above the ambient temperature ($P = I^2.R$), errors can occur, unless the extra heat is dispersed. This forces us to choose a small-sized resistance device with a quick response or a larger resistance device and better heat release.

A second solution is to keep the measuring current low (usually between 1 mA and 5 mA).

8.5.2. *Used materials and construction*

Platinum has a temperature range of –260°C to 750°C, good linearity and stability. That is why it is one of the more frequently used metals.

Copper is almost perfectly linear, but possesses a low relative resistance. To obtain a reasonable resistance, and thus variation in resistance, very long elements have to be used. Copper can easily oxidize and therefore it is not suitable for sensors: however the temperature of copper winding of electric motors and transformers can easily be measured by measuring its DC electric resistance.

Nickel is the cheapest and most sensitive metal, but its span is limited.

MATERIAL	Span [°C]	R_{100}/R_0	Deviation from 0-100°C [°C]
Platinum	260-750	1.385	0.38
Nickel	80-300	1.672	3.2
Ni-Fe	200-230	1.518	2.84
Copper	200-260	1.427	0.00

Table 8.4. *Characteristics of material used*

The time constant depends on the construction of the element and varies from 0.1 to 10 seconds.

The resistance wire always needs to be wrapped and insulated.

We distinguish the following types:

– RTDs with glass casing: advised for corrosive surroundings, span from –220 to 500°C.

– RTDs with ceramic insulation and steel casing: span from –220 to 850°C, good mechanical qualities.

– RTDs with a metal film: a platinum film is deposited or sputtered on a flat, narrow ceramic sheet, etched with a laser and cut off from the surroundings by a polymer film. It is very compact and has a high specific resistance; it is cheap but less stable.

8.5.3. *Applications*

Just like thermocouples, resistance thermometers can be placed in series or in parallel to perform average readings. However, this requires more caution than using thermocouples: it is important to keep the connection resistance low. Usually golden contacts are used. The change in resistance is measured with converters available in retail. The measuring circuit of these converters is often based on the Wheatstone Bridge (Figure 8.34).

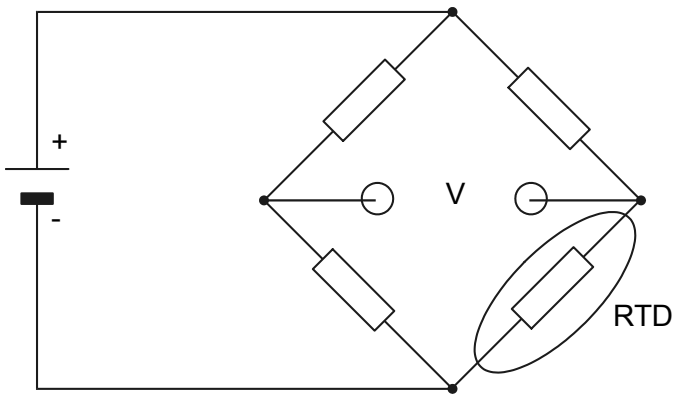


Figure 8.34. *Wheatstone Bridge*

The output voltage of the bridge is an indication of the RTD resistance. The method of connecting to the bridge and the effect of the wiring can lead to errors. The following three connecting methods are used: connection to two wires, three wires or four wires.

– Two wire connection

The bridge has four connecting wires, an external source and three resistance devices with a temperature coefficient equaling zero. To avoid the three bridge resistance devices reaching the same temperature as the RTD, connecting wires separate the RTD from the bridge. These connecting wires again cause the problem mentioned previously as the wire resistance influences the result (Figure 8.35).

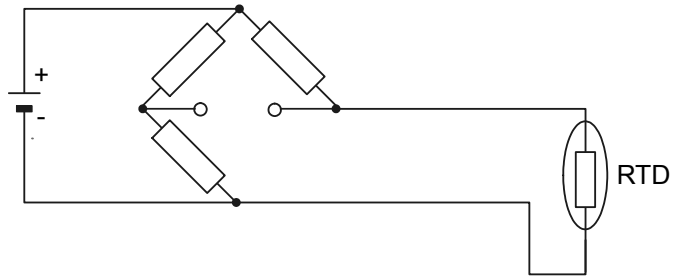


Figure 8.35. *RTD with long connecting wires*

– Connection with three wires

When wires A and B have the same length, and thus the same resistance, the effect of the added resistance is nullified, because every wire belongs to one half of the bridge. The third connection C does not influence the reading (Figure 8.36).

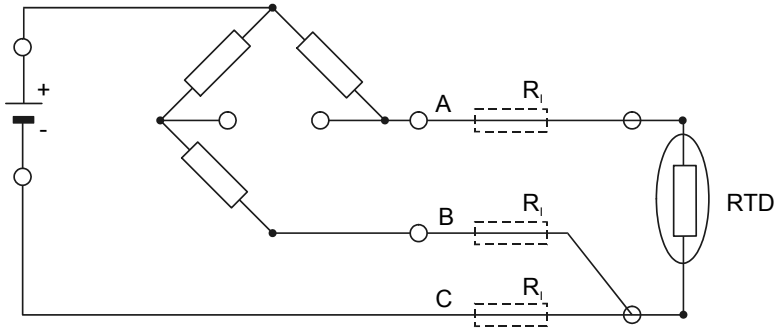


Figure 8.36. *Connection with three wires*

– Connection with four wires

The previous Wheatstone Bridge creates non-linearity between change in resistance and output voltage of the bridge and provides an accuracy rate of 5% FS. Current supply is recommended in order to reduce this error.

The four-wired connection considerably improves the accuracy (Figure 8.37).

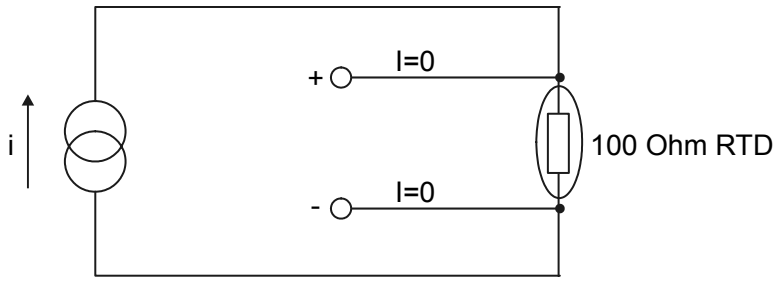


Figure 8.37. *Connection with four wires*

The set-up uses a power supply and a digital voltmeter. This voltmeter measures the difference in voltage of the RTD and is insensitive to the length of the wire (I = 0). The only disadvantage of this method is that four connections are needed.

The practical guidelines concerning interference that are mentioned in the section on thermocouples are also valid for RTDs.

8.6. Thermistors

8.6.1. *Principle*

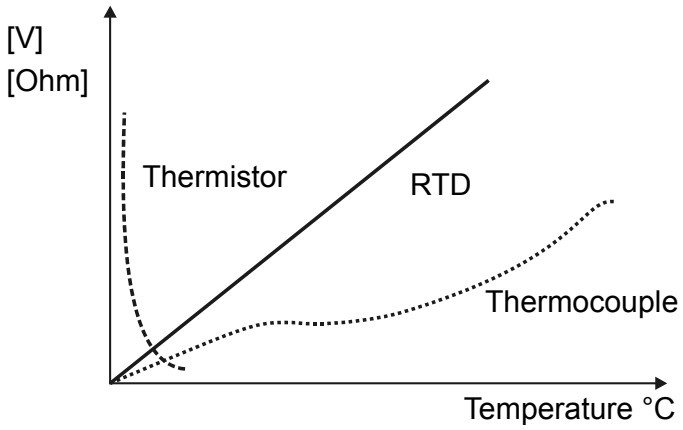


Figure 8.38. ***NTC thermistor compared with RTD and thermocouple***

Thermistor is a temperature sensitive resistance device made of semiconductor material. The thermocouple is the most versatile thermometer, the RTD is the most accurate and the thermistor is the most sensitive. Even though thermistors with a positive temperature coefficient (PTC) are available, most thermistors have a negative temperature coefficient (NTC) as the resistance decreases when the temperature rises (Figure 8.38). The temperature coefficient can change to several % per centigrade. This allows the thermistor to detect small changes in temperature, which cannot be done with RTDs or thermocouples. The main disadvantage of the thermistor is its strong non-linearity. Cheap thermistors have large spread of parameters ("tolerance") and calibration is usually necessary. Interchangeable precise thermistors are more expensive.

The transfer characteristic of a thermistor is given by the Stein-Hart equation:

$$\frac{1}{T} = A + B.\ln R + C.(\ln R)^3$$

where

– T: temperature [K]

– R: resistance of the thermistor [Ω]

– A, B, C: diagram constants

A, B and C can be found by inserting three known values of the thermistor in three equations and then solving the system of these three equations. If the three known values fall within a range of 100°C, an accuracy rate of 0.02°C is feasible.

8.6.2. *Thermistor technology*

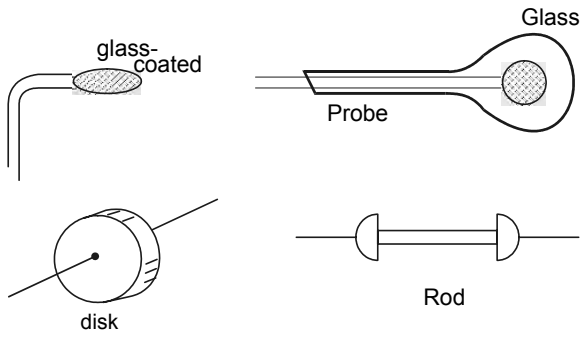


Figure 8.39. *Several thermistors*

Thermistors are usually made out of oxides of materials like nickel, cobalt and manganese. Other materials like iron, aluminum and copper in the form of silicates and sulfides are also used. A thermistor is made of a paste of solid-state material and then sintered to form the final element. You can find them in the form of pellets, disks, rings and rods (Figure 8.39). Because of the simple manufacturing method, thermistors can be made very small. The time constant of miniature thermistors can be units of milliseconds.

8.6.3. *Application*

In contrast with RTDs, thermistors have a large resistance value at room temperature (from KΩ to MΩ). Because of this the effect of the connecting wires is small to cause measuring errors. The measuring current may cause self-heating of the sensor. Just like resistance thermometers, a thermistor has to be insulated from media that can cause a short-circuit of the system. The insulation is usually made of glass or ceramics. Thermistors are preferentially used in a small temperature range, because that is where they are linear and very sensitive. An example of such an application is reading of the temperature of a reference junction of the thermocouple. Thermistors are commonly used in temperature monitoring systems (often as switches). Common applications include:

– safety switch in a coffee maker;

– flow switch: a heating element is situated inside the switch. When liquid or gas flows through the reading device, the liquid carries off the heat. Without flow the heat transfer becomes smaller and the temperature rises. The thermistor reads the temperature and operates the lock by means of a relay contact.

8.7. Monolithic temperature sensors (IC sensor)

A modern development in thermometry is the *monolithic temperature IC* (monolithic: all the elements of a full circuit are situated in only one chip or IC). These sensors are available in voltage and current output configurations (Figure 8.40). Both provide a linear output as a function of temperature. Typical values are 1 μA/K and 10 mV/K. With the exception of having a linear transfer these ICs have the same disadvantages as the thermistor. They are semiconductors and thus have a limited span (max. 0 – 100°C). They are very fragile and require an external power supply. These ICs are a good solution for measuring ambient temperatures or for hobby applications.

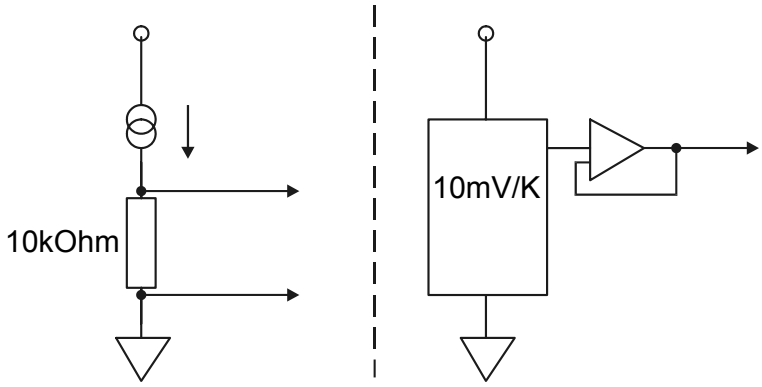


Figure 8.40. *Monolithic temperature sensors*

8.8. Pyrometers

8.8.1. *Introduction*

All temperature-measuring methods previously discussed require part of the thermometer to be in touch with the medium that needs to be measured. This is not possible for high temperatures, for instance in the case of melting furnaces or welding machines. Temperature measurement without contact is also a solution for moving substances. To measure temperature variations of the surface, the surface can be scanned with a contactless reader.

To cope with these problems we can use a wide range of instruments, all of them more or less based on the observation of radiation. In general we call them pyrometers, but depending on the manufacturer and depending on the instrument, we also speak of infrared thermometers, optical pyrometers, temperature radiation meters and so on.

An infrared image is not always accurate. Nitrogen filled cameras are still in use, but modern cameras are electrically cooled or uncooled in color and have the possibility of PC processing and analysis. Powerful software programs make it possible to process images. Histograms, three-dimensional representations, spot measurements and so on are all possible. The best-known application domains are medical science, the building trade and the industry. Classic applications are insulation measurement, energy management, climate control, process control, material research, preventive maintenance, etc.

Thermography makes it possible to judge the physical conditions during the production process in order to achieve maximal exploitation until nearly the end of life of each component. Because thermography measures without contact, this method can be used quickly at all times. This results in lowering the power failure, an increase in productivity, the prevention of fire, a decrease of maintenance costs, the prolongation of the life span, a cutback in material costs, safety, etc. The latest novelty is an uncooled camera.

8.8.2. *Basic principles of pyrometry*

Pyrometers receive electromagnetic rays in the visible and the infrared part of the spectrum. The spectrum of the visible part is quite narrow: wavelengths from 0.3 to 0.72 μm (Figure 8.41). To obtain a correct measurement "from a distance" (without a negative influence of the intermediate air on the reading), the measurement is performed in an infrared spectral band. In this band the layer of air absorbs little heat energy transmitted by the object to be measured and the air itself transmits little heat.

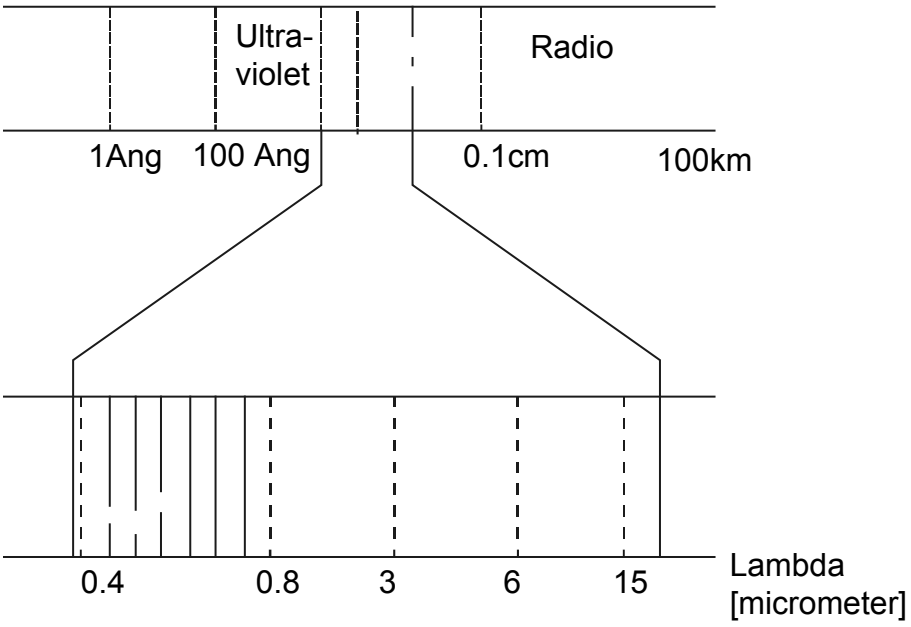


Figure 8.41. *Frequency spectrum of pyrometers*

Pyrometers usually take up a span from 2 to 14 µm, the so-called "atmospheric bands". This span is divided depending on the temperature range you want to measure. Window 1 (2 to 2.5 µm) and window 2 (3.5 to 4.2 µm) are used to measure higher temperatures (higher than 1,000°C). Atmospheric window 3 (8 to 14 µm) is used for temperatures between –50° and +600°C.

Every body warmer than 0 K emits electromagnetic radiation depending on its temperature. Figure 8.42 shows that the sensor measures also reflected radiation from other sources (ambient radiation).

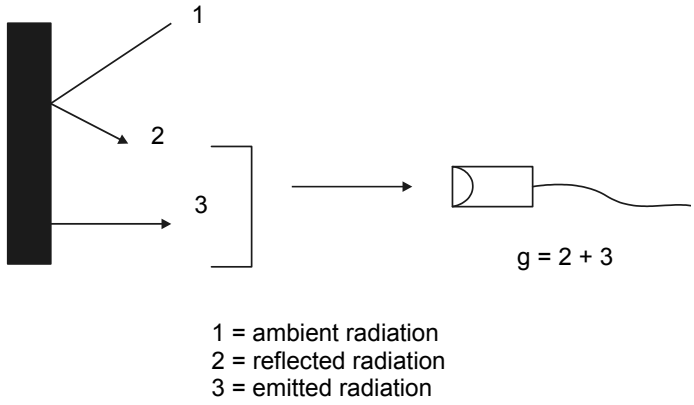


Figure 8.42. *Different kinds of radiation*

The ideal thermal radiator is called a black body. A black body absorbs all entering radiation (it has zero reflection coefficient) and emits a maximum amount of radiation, depending on its temperature (emission coefficient $\varepsilon = 1$). It also has zero transmission. In practice black bodies do not exist. Real materials have some degree of reflection or transmission. Every measuring instrument has a "correction factor" in which the (estimated or known) emission coefficient can be entered. A "black body model" is used as a source of radiation to calibrate pyrometers. It is black and cone-shaped, with a 15° cone angle.

8.8.3. *Measurement possibilities for pyrometers*

Pyrometers receive thermal radiation that the body to be measured has emitted from a distance, and they convert this radiation into an electrical signal from which we can calculate the temperature. When a pyrometer is calibrated against a black body, we can preset the necessary correction factor if the emissivity of the substance is known. Unfortunately emissivity does not only depend on the substance of a

body, but also on its size, its shape, the roughness of the surface, the angle of the measurement, etc. This leads to uncertainty in the values of the emissivity and can result in a largely inaccurate pyrometry. Another source of error pyrometry is the loss of energy during the transfer to the reading instrument. In atmospheric air the weakening in radiation is caused by the absorption in vapor, carbon dioxide and ozone or by the dispersion of radiation in particles or water drops. Figure 8.43 shows the absorption effect of H_2O, CO_2 and O_3.

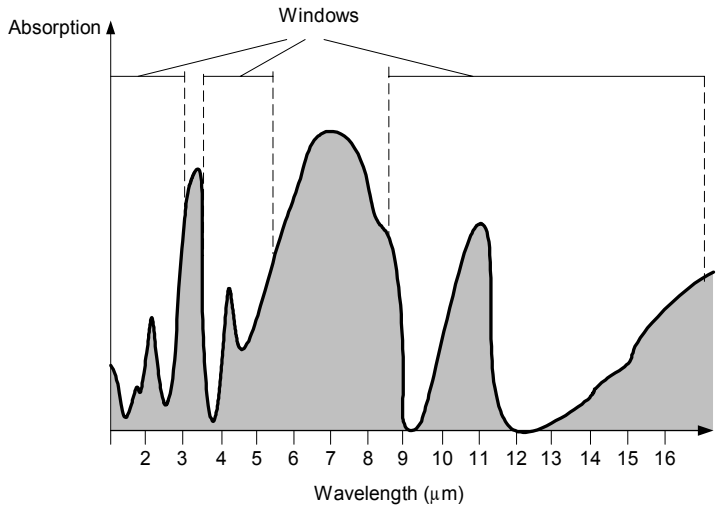


Figure 8.43. *Absorption of infrared radiation in atmosphere*

A good IR reading falls within a 1% accuracy rate. Sometimes the food industry demands an accuracy rate of 0.5%. When the reading needs to be traceable for legal reasons, the IR temperature reading has no juridical value. An extra method is an insertion sensor (usually based on NTC thermistor). So when the outcome of an IR reading approaches the marginal limit, it is better to perform an extra reading (core measurement in the food industry). Furthermore, in the food industry, it is important to be aware of the forming of condensation. The layer of condensation is a reflection mirror, so the reflected (and higher) temperature of the surroundings is measured instead of the temperature of the product.

The measuring method differs for each pyrometer. The most important measuring principles are as follows:

– *Measuring the total radiation*: these pyrometers measure the radiation in a band-span as wide as possible. This is the cheapest solution for a reasonable

sensitivity. A disadvantage of these instruments is that they are sensitive to sunlight and electric light, it is important to be careful about potential reflections.

– *Measuring within a standard band-span*: we measure within a narrow band-span, which usually measures between 500 and 1,000 nm. We speak of a standard span because the NBS-issued charts provide the emission coefficients of various substances for these wavelengths. These instruments are used to read higher temperatures.

– *Band-pass pyrometers*: we measure within a random band-span that is selected by two filters of which one passes larger wavelengths and the other smaller wavelengths. The instrument is usually optimized for the measurement of only one type of material.

– *Proportionate measurement*: the radiation intensity is measured in two different narrow bands. It can be demonstrated that the proportion of these two signals does not depend on the emissivity as long as this displays a constant proportion for the two bands. Practically it is proven that this type of measurement is very suitable for lower temperatures, but that it is inaccurate for higher temperatures.

8.8.4. *Implementation and construction of pyrometers*

The spectrum of pyrometers is fairly comprehensive as far as working principles and implementations are concerned.

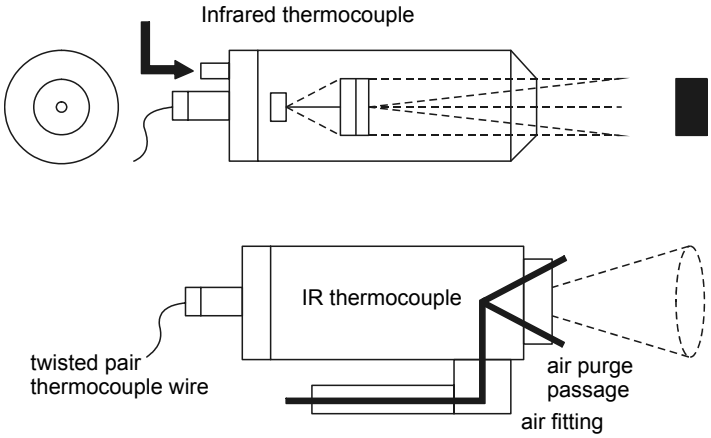


Figure 8.44. *Infrared thermocouple measurements*

Figure 8.44 shows an infrared thermometer that detects the infrared energy of the surface we want to measure. This energy is converted into an electrical signal that is sent to a processor.

This processor converts the electrical signal into an output signal. This can be an indicator, a recorder, a computer or a process controller. The data can also be sent to a data logger.

Specific instruments assure an optimal reading exists for specific applications such as the cement industry, glass industry, melting furnace industry, etc.

The last pyrometer is cooled with air to avoid disturbing influences. That is the aim of the air purge passage. The operating principle of the pyrometer is that the infrared energy is captured by the lenses and focused on the receiver (a thermocouple). The voltage on the thermocouple is electronically amplified into a useful signal that reflects the actual temperature of the object that is measured.

Above instruments can be carried out "stand alone". This means that the instrument can be integrated into a control circuit for temperatures between 0°C and 1,300°C. They have an output signal between 4 and 20 mA. This set-up is simple as far as the connection is concerned and cooling can be provided (either with compressed air or with cooling-water). The whole sensor is made robust in order to be used in industrial surroundings.

By using an optical fiber another option is created as an extension of the previous method (fiber optic thermometer). The use of "optic fiber" is an excellent solution for two problems. The lens and the cable can easily resist temperatures exceeding 200°C so that extra cooling is not necessary. The detector and the belonging electronics can remain in a cooler place. Furthermore it is often easier to reach the surface with the flexible optical cable (accessibility of the target).

Often a complete kit is offered:

– an accurate thermometer;

– a complete spectrum of protection jackets;

– a signal processing unit that emits a signal from 0 to 20 mA or from 4 to 20 mA;

– different lenses to provide a choice of measurement spot size at a given working distance.

Another possibility how to eliminate the errors caused by unknown emissivity is to place a "reflective cone" at the surface of the measured object. This causes the

effect of a black body at the exact spot of the reading. In addition, surrounding radiation does not affect the reading.

8.9. References

http://www.temperatures.com/links.html

http://www.enfm.nl/index.htm

http://www.op-ieder-potje-past-een-deksel.nl/

http://www.blwvisser.nl/

http://www.rosemount.com/products/temperature/accessories.html

http://content.honeywell.com/sensing/prodinfo/temperature/

http://www.minco.com/sensors.htm

http://www.mtisensors.com/rtds.html

http://www.rdfcorp.com/

http://www.sensycon.com/txt/txt_21_e.htm

http://www.heitronics.com/

http://www.ircon.com/

http://www.landinst.com/infr/index.html

http://www.mikroninst.com/

http://www.raytek.com/

http://www.endress.com/

http://www.pyrometer.com/

http://www.omega.com/

http://www.luxtron.com/

http://www.jumo.net/

http://www.spiraxsarco-usa.com/

8.10. Bibliography

MEYLAERS R., PEETERMANS M., PEETERS F., Meettechnieken cursus, 2000-2001, KHK Geel Departement TW.

List of symbols, acronyms and abbreviations

RTD	Resistance temperature detector
Pt 100	Platinum resistance (100 Ohm)
J	Joule
K	Kelvin
T	temperature
$T_{\circ F}$	temperature °Fahrenheit
$T_{\circ K}$	temperature °Kelvin
$T_{\circ C}$	temperature °Celsius
bar	bar
Mpa	Mpascal
γ	coefficient of cubic expansion
K^{-1}	Kelvin
V	volume
p	pressure
m	mass, proportion of thickness
y	pressure span of the Bourbon tube
n	proportion of the elastic modules
t	total thickness of the strip
C	thermoelectric material constant
σ_A	Thomson coefficient
ρ	radius of the curve
α	expansion coefficient
E	voltage
EMK	electromotoric force
I	current
Fe	iron
μV	microvolt
NBS	National Bureau of Standards
mV	millivolt
IC	integrated circuit
ISA	International Standard Association

DIN	Deutsche Industrie Norm (German industry standard)
NTC	negative temperature coefficient
IR	infrared
OPC	OLE [Object Linking and Embedding] for Process Control
τ	time constant

Chapter 9

Solid State Gyroscopes and Navigation

9.1. Introduction

The first practical experiment with a gyroscope was carried out in 1865, thanks to the provision of movement by an electric motor. The first gyrocompass was patented in 1904. Since acceleration and angular rate are measurable physical quantities without an external reference, they can be used for navigation of autonomous systems. These sensors are constantly improving because of their strategic importance. In the last 15 years in particular, thanks to optical and to micro-technologies, there has been enormous progress made in precision, linearity and stability as well as in the size and electric consumption of these sensors.

Inertial navigation has evolved continuously with the first combined accelerometer and gyroscope being produced in 1923, the first platform with three axes in 1924 and the first operational equipment being launched in 1940s on V2 rockets.

The gyrometers (= angular rate sensors) intended for inertial navigation must be able to detect revolution speeds varying from zero to 100°/s. In the standard case, a gyroscope has drifts of 10^{-2} degree per hour where errors are incurred at 1 mile per hour. For a ship, whose navigation can last several weeks, errors can accumulate with time, and a periodic correction is required. This is not necessary for a plane whose flight lasts only a few hours, or a missile where the duration of flight is measured in minutes.

Chapter written by André MIGEON and Anne-Elisabeth LENEL.

9.2. *The angular rate*

Let us consider a massive body in rotation with high initial angular rate ω_i and an inertia (inertial moment) I. According to the Newton's second law, the angular momentum, $I\omega_i$ of a body remains unchanged unless it is acted by a torque. A moment of force τ produces a term $\tau\Delta t$, where Δt is an interval of time. Let us suppose that this contribution is small, either because the moment of force is weak or because the interval of time Δt is short. By adding it vectorially to the great initial value $I\omega_i$ an end value $I\omega_f$, is found, which is not very different from its initial value. Thus, a body in rotation has a kind of gyroscopic stability. Gyroscopic stability explains why a spinning top amazingly remains vertical on its pointed end, defying gravity.

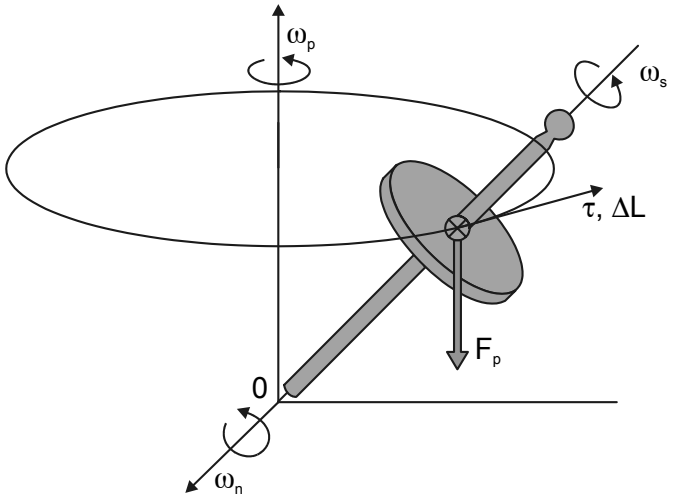


Figure 9.1. *A spinning top*

The spinning top in Figure 9.1 has the following movement:

– It turns around its axis of symmetry with an angular rate ω_s.

– At the same time, the axis of rotation (which is also the axis of symmetry of the spinning top) can undergo a slow vertical precession of angular rate ω_p.

– Its slope (inclination) ϕ can undergo a periodic variation, of frequency ω_n called nutation, on both sides of a certain average value.

The tilted spinning top of Figure 9.1 (that is turning quickly) is subjected to one constant moment of force, resulting from its own weight being applied on its center of gravity.

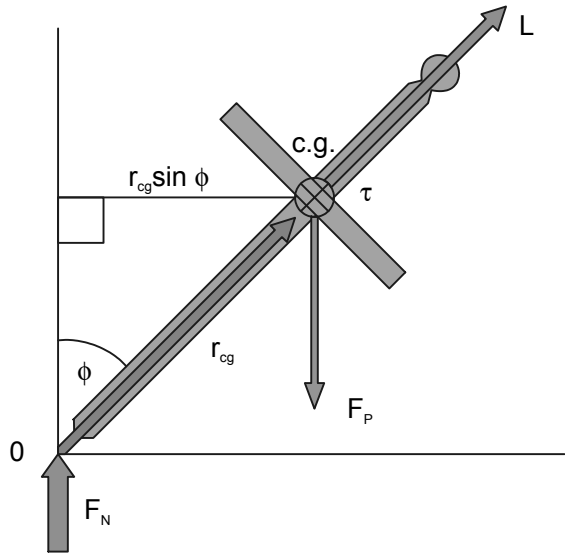


Figure 9.2. *Moment*

As shown in Figure 9.2, the moment τ is defined by vectorial multiple:

$$\vec{\tau} = \vec{r}_{cg} \times \vec{F} \tag{9.1}$$

– τ: moment of force of the force of gravity;

– r_{cg}: distance from center of gravity to axis of symmetry;

– F_p: force of gravity;

– and its size is $\tau = r_{cg}F_p\sin\phi$.

τ is perpendicular to the axis of rotation, which is also the direction of L, the angular momentum (kinetic moment). As shown in Figure 9.3, during a small interval of time Δt, τ produces a variation of the kinetic moment L

$$\Delta L = \tau \Delta t \tag{9.2}$$

– ΔL: a variation of the angular momentum (kinetic moment)

– τ: moment of force of the force of gravity

– Δt: a small interval of time

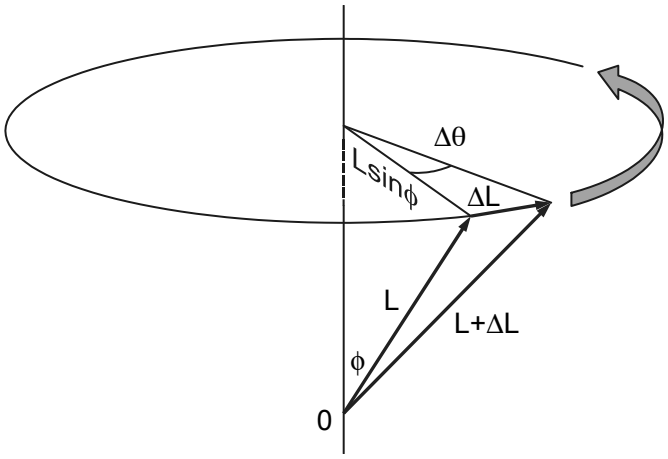


Figure 9.3. *Kinetic moment*

The kinetic moment varies from a certain initial value L to an end value L+ΔL and continues to change under the effect of the moment of force. As τ has no component in the direction of L, L is constant. The axis of rotation of the spinning top, tilted with a certain constant ϕ angle, sweeps a conical surface around the vertical without falling. This is *precession.* It is only when the angular rate becomes slow that the spinning top starts to fall. The real angular velocity of precession ω_p can be determined as

$$\omega p = \frac{\Delta\theta}{\Delta t} \quad \text{when } \Delta t \rightarrow 0 \tag{9.3}$$

– ω_p: real angular velocity of precession

– Δθ: variation of the angle between the position at t and t + Δt

We have:

$$\Delta\theta = \frac{\Delta L}{L \sin \Phi} \tag{9.4}$$

$$\Delta L = \tau \Delta t \tag{9.5}$$

$$\tau = mgr_{cg} \sin\Phi \tag{9.6}$$

then:

$$\Delta\theta = \frac{(mgr_{cg} \sin\Phi \Delta t}{L \sin \Phi} \tag{9.7}$$

– ϕ : angle of deviation of the axis of rotation of the spinning top

– g : gravity

Therefore,

$$\frac{\Delta\theta}{\Delta t} = \frac{mgr_{cg}}{L}. \tag{9.8}$$

when $\Delta t \to 0$, and supposing that $\omega_s >> \omega_p$.

We obtain:

$$\omega_p = \frac{mgr_{cg}}{I\omega_s} \tag{9.9}$$

Of course, a spinning top which is not in rotation must be supported by an additional vertical force appropriate to be maintained to an angle ϕ. For a gyroscope of guidance, it is desirable to minimize ω_p. It is thus necessary to make I and ω_s large and r_{cg} small.

9.2.1. *Definition of rate gyro*

9.2.1.1. Comparison between a gyroscope and angular rate meter (gyrometer)

The gyroscope is a fast rotating body suspended by a Cardan suspension allowing movements in all directions. The gyroscope generally describes a circular cone (movement of precession around an axis). The direction of this axis remains fixed in space (in a Galilean reference frame).

The gyroscopic effect is concerned with the fact that the gyroscope moves in a direction perpendicular to a force which is exerted on it. This effect is used to detect forces, angular movement, particularly in inertial navigation systems or in inertial platforms (which make it possible to define the position of a mobile unit per integration of its acceleration). In inertial navigation the gyroscope is a sensor commonly used to measure a rotary angle. The output of a gyroscope is an angle.

The gyrometer is a sensor commonly used to measure an angular rate.

The traditional gyrometer with an axis is primarily made up of:

– a spinning top turning around an axis Δ carried by a ring itself connected to the case of the gyrometer by the axis of exit S perpendicular to the axis Δ;

– a torque motor;

– a detector of variation acting on the moving element around S.

Because of the gyroscopic phenomena, an angular movement of housing with an angular velocity ω according to the axis of entry will cause the appearance of a couple being exerted around the axis of exit and proportional to ω. The ring is brought under control to remain in its position of balance by means of a torque motor, so that the driving current it being proportional to the couple applied. The output is a voltage across the sensing resistor, which is proportional to that current and thus a function of the measured angular velocity. The general diagram of the traditional gyrometer with one axis is shown in Figure 9.4.

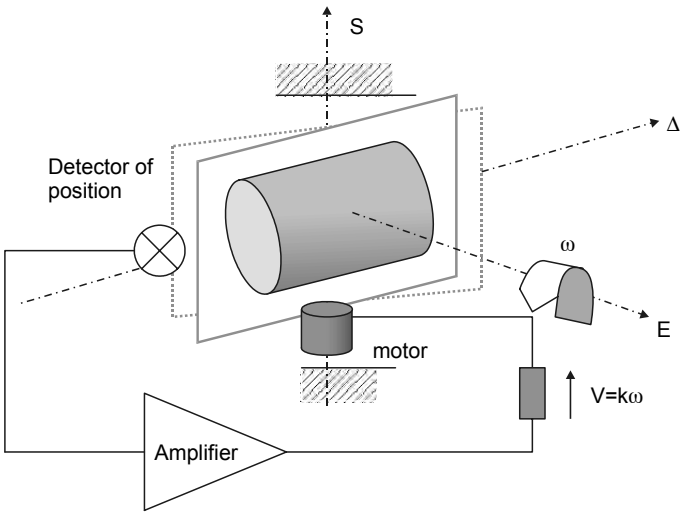


Figure 9.4. *Diagram of the traditional gyrometer*

9.2.2. *Use of rate sensors*

The gyrometers constitute the basic elements of the inertial frames of reference used to provide the functions of guidance, piloting, navigation and stabilization of various vehicles or particular platforms. In the systems with bound components, called "strapdown", only the gyrometers are used.

The gyrometers account for 30 to 40% of the end value of the system with an added value of high technology. In addition, the majority of the powerful gyrometers are not available on the "open" market of the sensors. Consequently the gyrometers are regarded as strategic products, explaining the fact that it is extremely difficult to obtain "serious" documentation about the top-of-the-range products. Also, all the major equipment suppliers in civil aeronautics and military development produce their own gyrometers.

Gyro for automotive applications

The first use of rotational sensors in automobiles was for traction control. In this application, the function of the gyro was to provide the vehicle rate of rotation while moving along a curved road. By comparing the angle of turn from the steering wheel and the actual vehicle's rotation obtained from the gyro, the control system can determine if the vehicle is experiencing loss of road traction. When unsafe conditions are detected, the system initiates commands to the Anti-Lock Breaking System to provide counter torque to prevent the vehicle from spinning out of control. Other automotive applications in development include using gyros to detect vehicle rollover conditions, vehicle dynamics control and navigation. For each application, the specifications may differ significantly. Each car in the future could potentially require three gyros.

From now on, many of the concerns regarding research and development are with economic considerations. All in all, the operational constraints are increasingly difficult to manage, especially when there is a compromise between cost and performance. The dominating parameters governing the level of their design and use are:

– duration of the mission;

– required precision;

– the dynamics of the mobile unit (°/s/s).

9.3. Different ranges of rate gyro

In this section, the sensors used for angular velocity measurement higher than a few tens of turns per minute have been eliminated. Indeed, these sensors belong

rather to the category of tachometers that are defined by other principles and technologies.

9.3.1. *Control of trajectory*

In order to carry out the control of trajectories of various vehicles, like cars, robots or projectiles, it is necessary to have gyrometers capable of measuring effective ranges of about 100°/s. The precision required would lie between 1°/s and 10°/s.

For these types of applications the quantities used are significant and the costs must be extremely low.

9.3.2. *Piloting and Stabilization*

For the functions of piloting and stabilization of planes, more particularly in the aeronautics and military fields, helicopters or machines would be required to have gyrometers having effective ranges of about 400°/s to 600°/s. The precision necessary in this case would be between 0.5°/s and 10°/h. In these applications, autonomy is relatively limited, and the quantities used would be fairly significant. The size and weight would be paramount.

9.3.3. *Guidance*

In the military field, the function of guidance requires gyrometers to have effective ranges of about 400°/s to 600°/s. The precision sought in this case is close to 1°/h. In these applications autonomy is significant.

9.3.4. *Navigation*

In the aeronautical field, the function of navigation for planes primarily requires gyrometers to have effective ranges of about 100°/s to 200°/s. The precision sought in this case lies between 0.1°/h and 0.01°/h. In these applications large-scale autonomy of and excellent precision are required.

We can distinguish between two types of navigation: navigation with periodic retiming and autonomous navigation which is most constraining from the point of view of performance. Figure 9.5 summarizes the various ranges of performances of gyrometers.

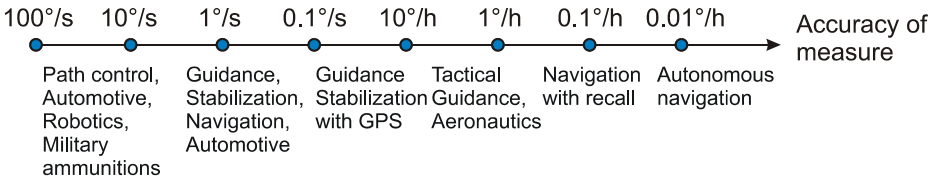


Figure 9.5. *Various ranges of performances of gyrometers*

Table 9.1 presents some examples of ranges of gyrometers that satisfy certain applications.

Market Major requirements	Uses	Range of (°/s)	Class of Performance (°/s)
Automotive Low cost, reliability, severe environment, lifespan	Security	200	10
	Active suspension	50	2
	Navigation	100	0.1
	ABS for automotive	50	0.5
Domestic/medical Low cost, low fuel consumption, small size, lifespan	Games	100	0.1
	Sports	50	0.1
	Cameras	50	0.5
	3D mice	100	2
Industrial Small size, reliability, severe environment	Robotic	10	0.1
	Machines monitoring	10	0.1
	Attitude control	20	0.2
	Stabilization	10	0.01
Aerospace Performance, reliability, lifespan, low fuel consumption	Guidance	600	10^{-5} to 10^{-3}
	Piloting	400	10^{-5} to 10^{-4}
	Navigation	200	10^{-5} to 10^{-7}
	Stabilization	600	10^{-5} to 10^{-3}
Military	Projectiles and rockets	100 to 800	10^{-7} to 10^{-3}

Table 9.1. *Some examples of ranges of gyrometers*

9.4. Main models of rate gyro

For the angular rate measurement one can arrange the various principles of gyrometry into three families:

– rotary gyrometers**,** which are the oldest;

– vibrating gyrometers, which are currently the subject of much development;

– optical gyrometers, whose particular characteristics place it between the two preceding families.

9.4.1. *Rotary gyrometers*

Mechanical gyrometers (gyros) have been used in aircraft for many years. Their main problem is long-term reliability and limited accuracy and resolution. The recommended operating life of most mechanical gyros is only several hundred hours.

9.4.2. *Vibrating gyrometers*

These gyrometers are based on the Coriolis effect. The Coriolis force F exerted on a body is

$$F = 2m\vec{\omega}\times\vec{v}_r \tag{9.10}$$

where

– m is a body mass

– ω is angular velocity

– v is linear velocity.

Vibrating gyrometers use a proof-mass mounted on a spring suspension. Rather than spinning as a conventional gyroscope rotor, the proof-mass vibrates back and forth in translational motion.

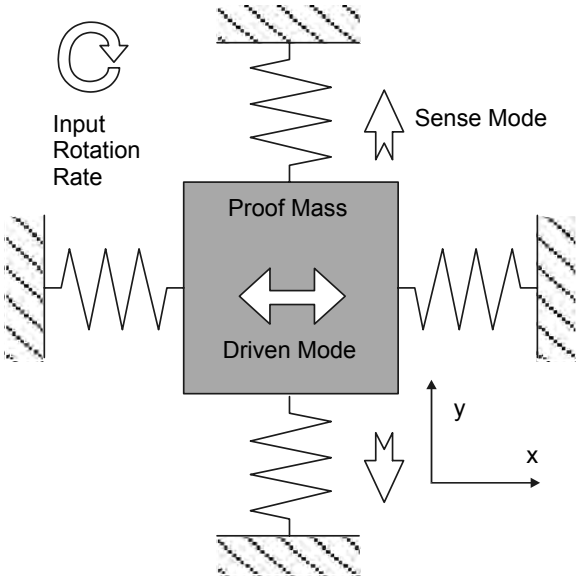


Figure 9.6. *The composition of the vibrating rate gyroscope*

The basic operating principle of all vibratory gyroscopes relies on the generation and detection of Coriolis acceleration. A very basic vibratory gyroscope is comprised of a proof-mass mounted on a suspension that allows the proof-mass to move in two orthogonal directions. In gyroscopic dynamics, there is a distinct motion about all three orthogonal axes as shown in the figure above. First, the proof-mass is put into oscillatory motion in the x-axis (called the drive axis) parallel to the substrate. Once in motion, the proof-mass is sensitive to angular rates induced by the substrate being rotated about the z-axis perpendicular to the substrate. The input rate induces Coriolis acceleration in the y-axis (called the sense axis) which is perpendicular to both the x-axis drive and the rate input z-axis. This Coriolis acceleration induces a Coriolis motion with amplitude proportional to the angular rate of the substrate.

The size of the Coriolis force is:

$$F = 2mv\,\Omega \sin \omega t \tag{9.11}$$

– Ω: angular rotation velocity

– ω : pulsation corresponding to the vibrating frequency of particle m

– t: time

The resulting vibration has components X and Y:

– the vibration of excitation in the direction of x;

– the vibration corresponding to the Coriolis force in the direction of y.

The measurement of the secondary vibration (along the y-axis) is used to determine the number of revolutions (Figure 9.7).

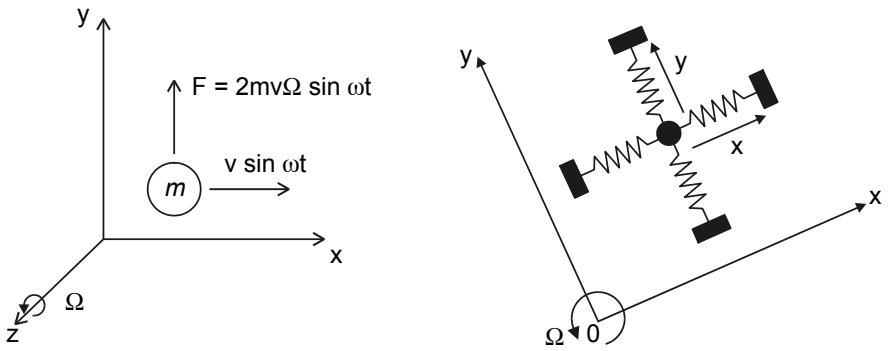


Figure 9.7. *The Coriolis force F*

The current research goals focus on the geometry of the sensing element and the means of exploitation of vibrations. Hereafter, various principles of exploitation of this physical law are presented, each of them derived from the geometry of the sensing element.

9.4.2.1. *Gyrometers with Elementary or coupled bars*

The sensing element of this type of gyrometer consists of a metal bar elinvar (elastic metal with invariable size with temperature). A piece of piezoelectric ceramics is stuck on the face of the bar and set in vibration. This is the excitation element. Other piezoelectric ceramics are stuck on faces at 90° and receive the vibrations. These are the detection elements. When an angular velocity, Ω, is applied to the system, it appears as a Coriolis force along the y-axis. The ceramic detection elements are subjected to a vibrating force, the resultant (vector sum) of X and Y (see Figure 9.8).

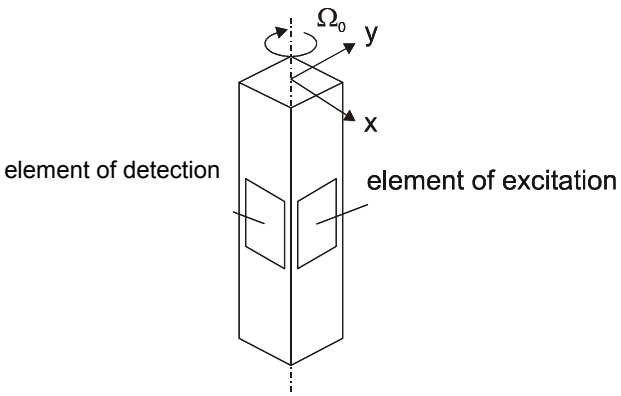


Figure 9.8. *The element of excitation/detection of the gyroscope with elementary bar*

Industrial example: the GYROSTAR from MURATA

In order to minimize the signal-to-noise problem, MURATA assembled three piezoelectric elements on the faces of the equilateral prism (Figure 9.9).

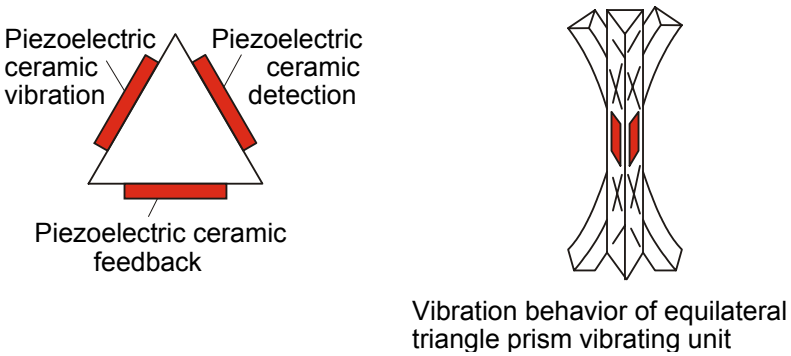


Figure 9.9. *The sensing element GYROSTAR (from MURATA [3])*

Principle

One of piezoelectric ceramic pieces sets the unit vibrating; this is termed the excitation element. Two other piezoelectric ceramics receive the vibrations: they are the detection elements. The inherent geometry of the equilateral prism and the position of the excitation elements direct the resulting force, F (product of the excitation force and Coriolis force) to be perpendicular to the detection elements.

This greatly improves the detection by several hundreds of mVs. The effects of the random vibrations will be thus negligible. Moreover, it is possible to overcome the signals generated by the two piezoelectric ceramic detection elements.

In the absence of rotation, the noise is cancelled because the vibrations due to vertical or different movements other than rotations will vary amplitude, but the difference will always remain zero.

In the presence of rotation the amplitude of one of two detection ceramics elements in any direction will increase (the force F will be almost perpendicular to the plane of ceramics), while the amplitude of the other elements of detection will decrease (the force F will be almost parallel to the plan of ceramics). This will give (A + a) – (A – a) = 2a (Figure 9.10).

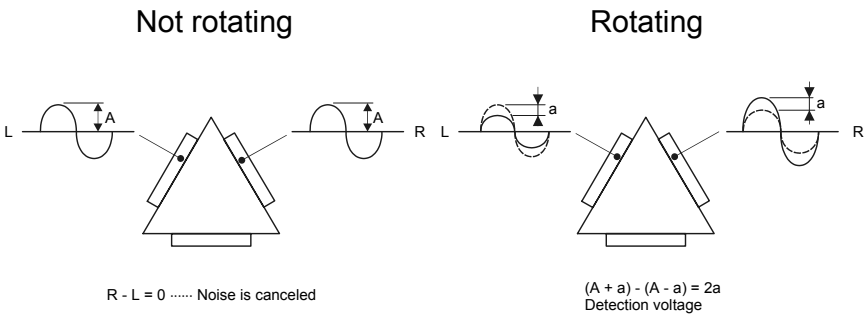


Figure 9.10. *Principle of the gyrostar (from [3])*

Thus, the signal of rotation is detected at the output with an amplitude twice that of a single detection element. These two combined factors give a very high signal-to-noise ratio.

The excitation and detection ceramic pieces are stuck by epoxy adhesive onto the nodal point of the prism (where no deformation is present). The ends of the prism are then free to move, giving the setup name freebar. An oscillator excites the ceramics at 8 kHz. Two detection ceramics vibrate electronically in phase and resonate in the three directions of the prism because of their identical resonance frequencies. The output signals of two ceramic detection elements are fed into a differential amplifier, and then into a synchronous detector (Figure 9.11).

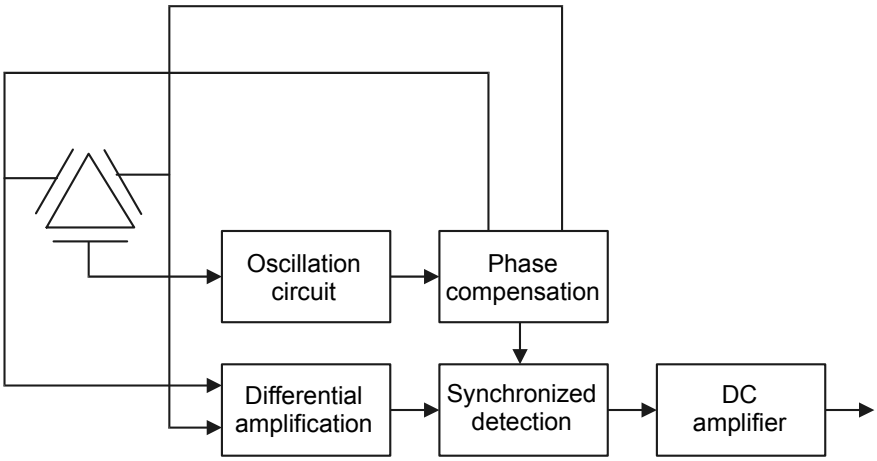


Figure 9.11. *Associated electronics*

The gyrostar vibrating gyrometer is able to yield measurable signals of good linearity and with a good signal-to-noise ratio. It has fast response time, an excellent behavior in temperature and low fuel consumption.

9.4.2.2. *Gyrometers with a tuning fork*

The principle of this gyrometer is similar to the previous example except that the piezoelectric excitation ceramics are assembled on the legs of a tuning fork. The detection ceramics are assembled perpendicular to excitation ceramics.

Therefore, if the excitation ceramics vibrate along the x-axis (X), when (Ω) is applied, Coriolis force is generated along the y-axis (Y) and detection ceramics are subjected to a vibrating force, which is the resultant force of X and Y.

In certain cases the material of the tuning fork itself is piezoelectric (Figures 9.12 and 9.13).

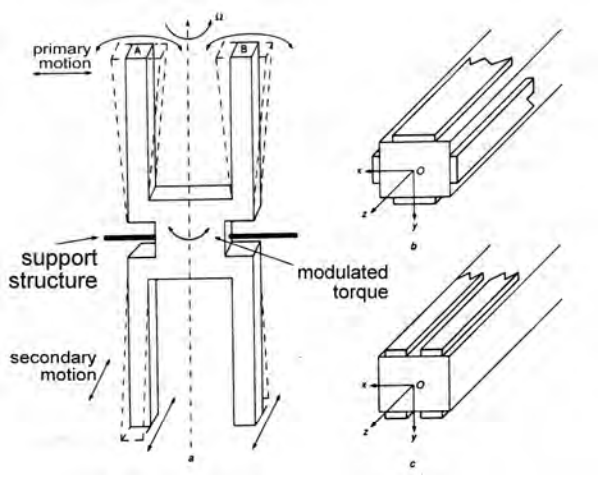


Figure 9.12. *The tuning fork*

Limits

As for any gyroscope, the detection voltage must be zero in the absence of rotation. This implies that ceramics must be assembled accurately and positioned perpendicular to the excitation ceramics very precisely. Good reproducibility of this standard is difficult to achieve in mass production. If the vibrations on the detection ceramics are not perpendicular to the sensor plane, then a movement of limited amplitude is generated, thus diminishing the output voltage to a few millivolts. Noise signals, due to external vibrations, can appear and disturb the detection signal. Complex and expensive electronic circuits are required to filter the signal. Therefore, the signal-to-noise ratio is unfavorable.

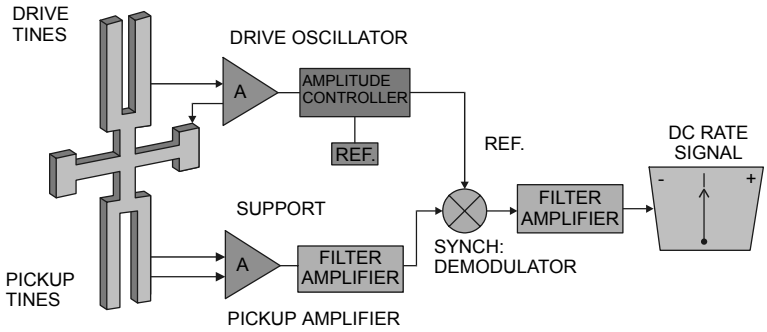


Figure 9.13. *A vibrating dual-ended quartz tuning fork and signal processing electronics constitute a complete gyroscope (after [5])*

9.4.2.3. *Gyrometers with coplanar interdigitated comb fingers*

This gyrometer uses a sensing element with coupled vibrating blades produced in silicon by surface micromachining. The detection of the various modes of vibrations is done by means of interdigitated capacitors realized simultaneously in silicon. To avoid some of the problems associated with single-mass sense elements, these gyroscopes employ two proof-masses as well as proprietary suspension designs. The gyroscope systems are less likely to confuse linear accelerations along the sense mode for Coriolis accelerations. Furthermore, the designs provide differential measurement capability of improved power supply noise rejection and measurement stability.

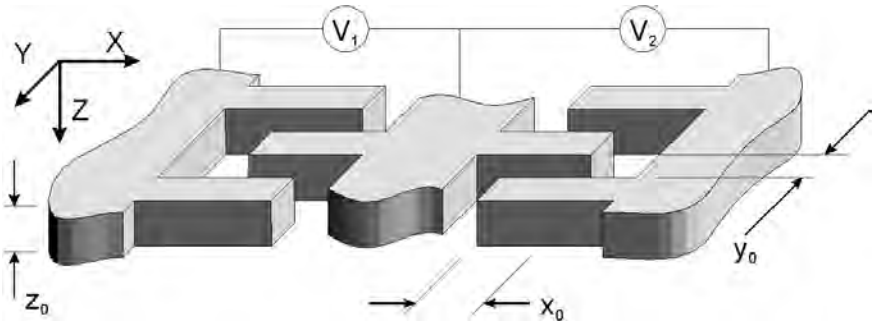


Figure 9.14. *Variable capacitor for applying X-axis force (from HSG IMIT [7])*

In order to generate the Coriolis acceleration necessary for rotation rate sensing, the gyroscope proof-mass must be driven into oscillation. The forces applied for sustaining the oscillation are electrostatic. A voltage is applied between the interdigitated comb fingers mounted on the proof-mass in the center and comb fingers mounted on the substrate. A voltage applied between the proof-mass and substrate combs invokes an electrostatic force that moves the proof-mass.

The interdigitated comb finger configuration has a number of advantages including force linearity and ample room for large displacements along the x-axis without collisions between comb fingers. The scale factor is proportional to drive oscillation amplitude, so keeping the oscillation amplitude constant is crucial.

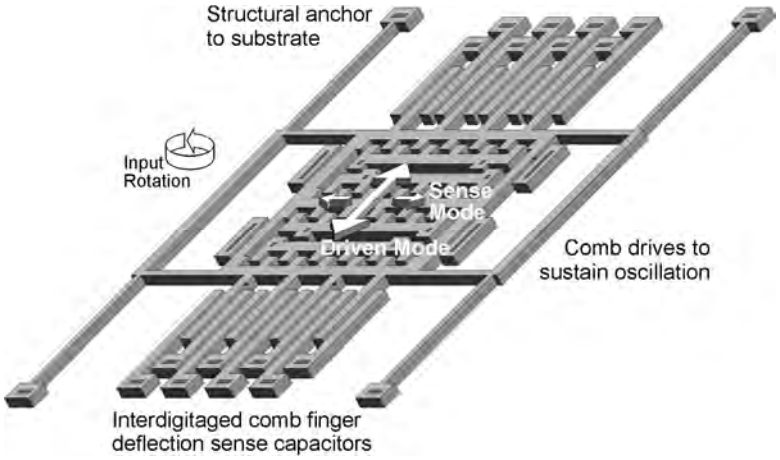


Figure 9.15. *Vibratory gyroscope design (from HSG IMIT [7])*

Given that the gyroscope proof-mass is designed to oscillate along the x-axis, deflections along the y-axis must be measured to infer the Coriolis acceleration. Parallel plate electrodes attached to the proof-mass and the substrate form air gap capacitors. The perspective view above (Figure 9.15) shows a single proof-mass gyroscope with all the electrodes needed for operation. Capacitive detection circuitry employing a full bridge layout is used to sense imbalances in differential pairs of air-gap capacitors. Capacitor imbalance is an indication of proof-mass displacement.

The acceleration of the proof-mass detected along the sense axis or y-axis must be processed to estimate the input rotation rate. The Coriolis acceleration is an oscillation along the sense axis, the magnitude of which is proportional to the input rotation rate. Because the frequency and phase of the Coriolis acceleration are known, the oscillation magnitude can be determined. This is a simple process provided that there are no interfering signals at the same frequency as the Coriolis acceleration. The final result is an output voltage proportional to the angular rate input.

9.4.2.3.1. *Micromachined dual input axis angular rate sensor [8]*

The inherent symmetry of the circular design allows the simultaneous angular rate measurement about two axes. Minimum detectable signals as small as 1.2°/sec over a 20 Hz bandwidth were achieved with this sensor (Figure 9.16).

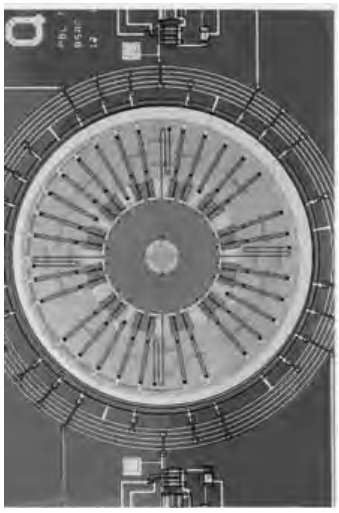


Figure 9.16. *Dual axis angular rate sensor (from Berkeley [8])*

Industrial example: "Butterfly-Gyro" from SensoNor

This silicon bulk gyroscope is good for automotive applications. The gyroscope structure, called the "Butterfly-Gyro", has a gyroscopic scale factor comparable to that of tuning fork gyros. The gyro is simple to manufacture, with single sided electrostatic excitation and capacitive detection. As the two masses vibrate in opposite phase, the offset is smaller and the gyro is less sensitive to linear and angular vibrations. The best samples have a resolution of approximately 0.1°/sec at 50 Hz bandwidth (Figure 9.17).

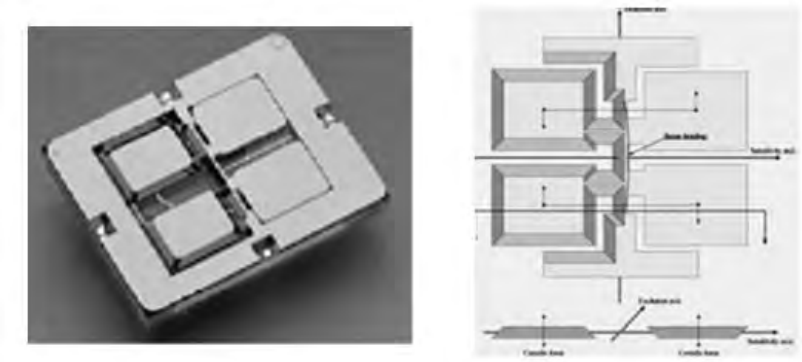


Figure 9.17. *The Butterfly-Gyro (from SensoNor [9])*

Manufacturing process of the Butterfly-Gyro

The silicon part of the sensor structure is formed in one self-stopping anisotropic etch step preceded only by a double-sided patterning of the front and the back of an oxidized silicon wafer. For the electrostatic excitation and the capacitive detection, the silicon itself constitutes one electrode while the counter electrodes are placed on an etched glass wafer anodically bonded to the silicon. The noise from the best sensors is equivalent to 0.07°/s at 50 Hz bandwidth [10].

9.4.2.4. *Gyrometers with vibrating shell and cylinder*

The rotation sensing principles of the vibrating shell gyroscope were first analyzed by G.H. Bryan in 1890 and are best illustrated by the ringing wine glass, vibrating in its fundamental flexural mode shown in Figure 9.18.

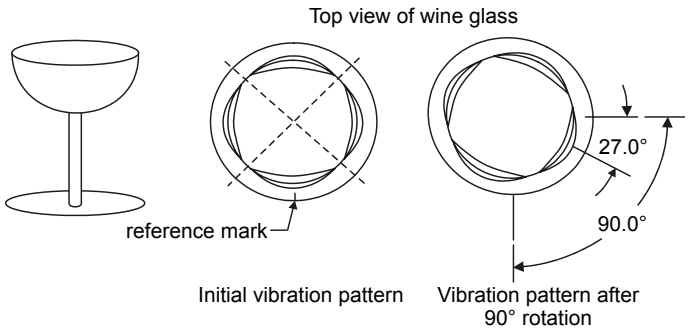


Figure 9.18. *Vibrating shell gyroscope*

In this mode of vibration, the lip of the wine glass vibrates in an elliptical shape mode that has two nodal diameters. When the wine glass is rotated, the node lines lag behind the rotation or precession, much like the precession of swing of the Foucault pendulum. During a 90° rotation, the node lines are observed to precess by about 27°. The precession rate is a geometric constant for each type of vibrating shell gyroscope. It is called the angular gain and is approximately 0.3 for a wine glass.

Due to their shape, vibrating shell gyroscopes are inherently rugged. They are less sensitive to spurious vibration than other vibratory gyroscopes. Only when the shells have mass or stiffness asymmetries can environmental vibrations induce a spurious response. Vibrating shell gyroscopes can operate in either whole angle open-loop or force-to-rebalance mode depending on the demands of the application.

The vibrating cylinders gyrometer is composed of a cylindrical sensing element, excited to vibrate in two principal directions. Four nodes and four "antinodes" are perpendicular to the excitation vectors. Displacements at the "antinodes" of vibrations occur in opposite phase for the two principal directions of excitation. When the sensing element is subjected to a rotation about an axis perpendicular to the two principal directions of excitation, the nodes of vibrations do not turn with the sensing element. They do not remain fixed in space either, but turn at an angular velocity that is a fraction of the angular velocity of the sensing element. The relationship between the angular velocities of the sensing element and the nodes of vibration depends on the geometry of the sensing element.

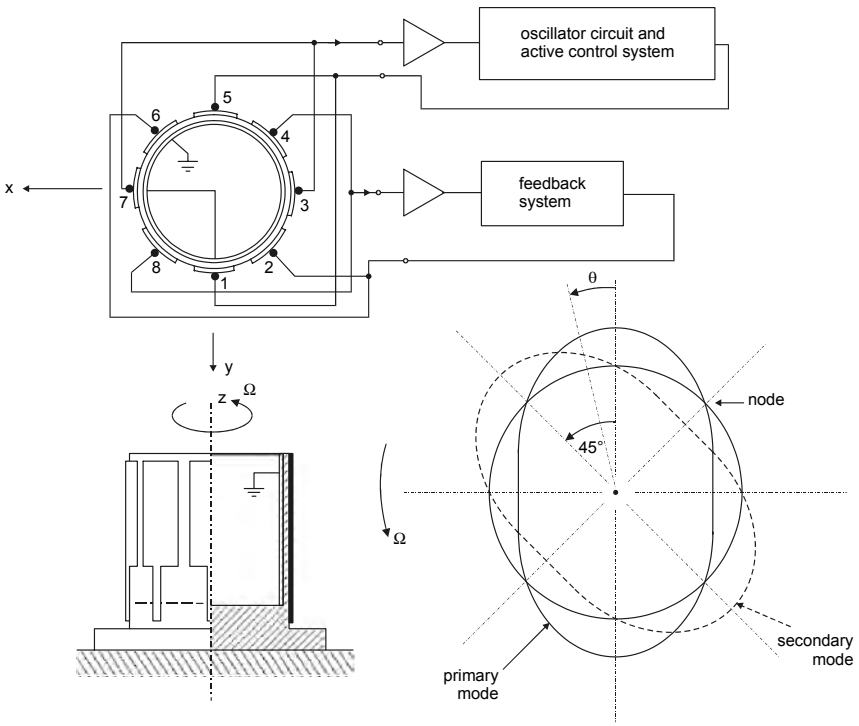


Figure 9.19. *Sensing element in a gyrometer with vibrating cylinder (after [2])*

The geometry of the sensing element is cylindrical. It has thin walls, open at one end and closed at the other, with a bottom wall thicker than the walls, forming a solid base for the element.

9.4.2.4.1. *Example of the Gyrometer with vibrating cylinder of Condor Pacific Industries, with magnetic excitation and detection*

The Condor Pacific Resonant Rate Sensor is a single axis device, which uses a magnetically controlled vibrating cylinder as the sensing element. The closed loop produces accurate scale factors and linearity. Because there are no rotating parts to wear out, the expected operating lifespan is greater than 20 years. This device is very rugged and can withstand higher shock than most rotating devices.

The rate sensor consists of a cone supported cylinder, with eight forcer coils, and eight pickoff coils. The cylinder is excited at its fundamental frequency mode (4,000 Hz). In this mode, the cylinder opening vibrates radially with maximum vibration amplitude every 90° around the circumference of the cylinder. At midpoints between these maxima nodes of zero amplitude of vibration exist. If the cylinder is rotated about its central axis, the nodes will not rotate with the cylinder but will lag behind (Figure 9.20).

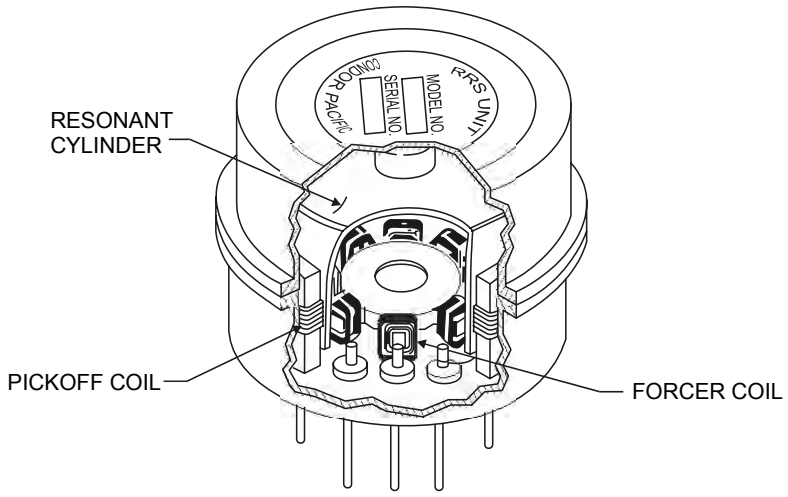


Figure 9.20. *Design of the resonant solid state rate sensor, gyrometer with vibrating cylinder from Condor Pacific Industries [10]*

Pickoff coils are arranged around the outside of the cylinder to sense the vibration of the drive axis and the nodes between the standing waves. The pickoff coils are excited with a 250 kHz sine wave. The drive and sense pickoff signals are both amplified and demodulated. The two demodulated signals are called “drive

motion" and "sense motion". The forcer coils are arranged around the inside of the cylinder and are located directly opposite the pickoff coils. Four of the forcer coils are connected to excite the vibrating axis of the cylinder and the other four forcer coils are connected to drive the node axis of the cylinder. Positive feedback from the drive motion signal to the drive axis forcer coils causes the resonator to vibrate at its natural frequency. The amplitude of vibration is measured and adjusted to maintain constant amplitude. In this state where there is no input rate, the sense motion signal is zero. If an angular rate is applied about the central axis of the cylinder, the sense motion signal will increase with increasing rate. If the rate is reversed, the phase of the sense motion signal will reverse.

To improve bandwidth, linearity and scale factor accuracy, another amplifier and circuit are connected from the sense motion signal to the sense axis forcer, thus making the rate sensor a closed loop device. The current through the sense axis forcer is proportional to the input rate about the central axis of the cylinder. This current is sensed, demodulated and filtered to provide the output for the rate sensor.

Gyrometer with an optical detector of the position of the nodes of vibration

In this gyrometer, the LED light source is placed against the stator using an adapted support, inside the vibrating roller. The emitted beam of light is coaxial with the vibrating roller and is turned towards the center of the reflectors. Reflectors, plated with two reflective faces, are used to orient and concentrate the light. The reflective surface can be slightly concave facets or plane mirrors placed 45° to the axis of the vibrating roller.

The detector is composed at least of two, though preferably eight, photoelectric sensors placed on a circle whose diameter is equal to the internal diameter of the vibrating roller. This circle is positioned outside the cylinder so that the surfaces of the photoelectric sensors are comparable with the air-gap formed between the vibrating roller and the diameter external of the stator.

9.4.2.5. *Gyrometers with vibrating disk*

The sensing element, in this case, is a planar disk from piezoelectric material (Figure 9.21). The excitation and detection electrodes are distributed on the surface of the disk according to the selected mode of vibration.

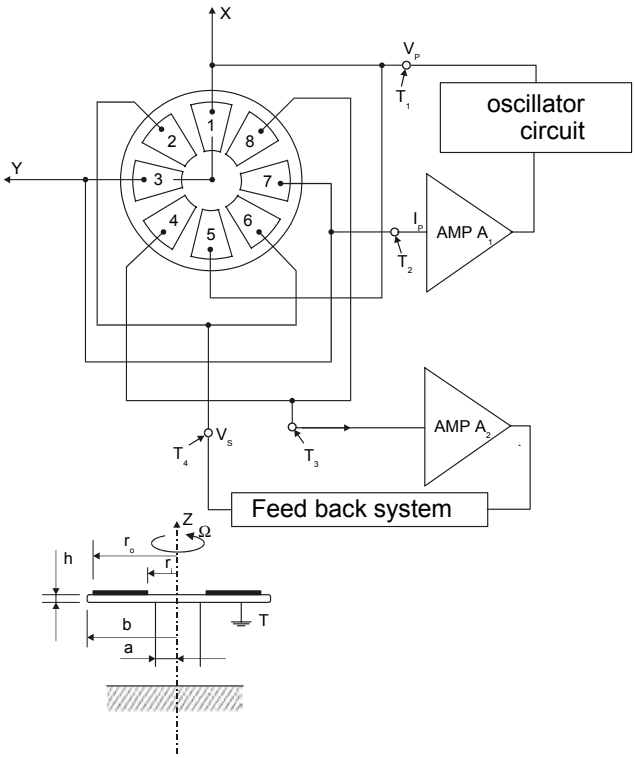


Figure 9.21. *Principle of gyrometer with vibrating disk (after [2])*

9.4.2.6. *Gyroscopes with vibrating ring*

A vibrating ring gyroscope is schematically shown in Figure 9.22. This device consists of a ring, semicircular support springs, and drive, sense and balance electrodes, which are located around the structure. Symmetry considerations require at least eight springs to result in a balanced device, with two identical flexural modes that have equal natural frequencies.

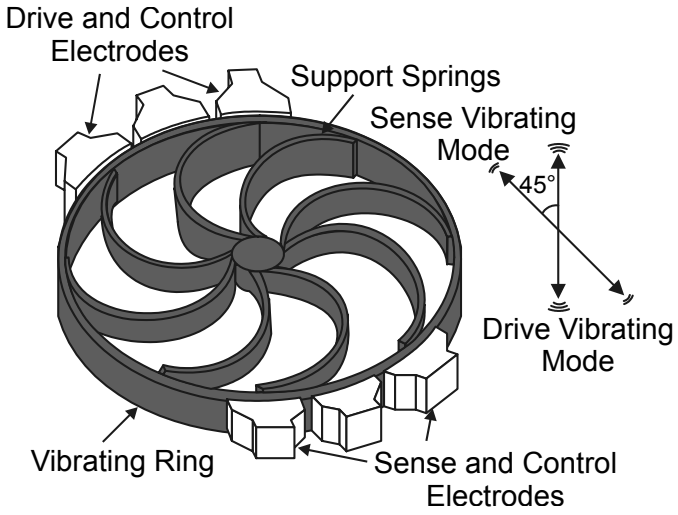


Figure 9.22. *A vibrating ring gyroscope (from [11])*

The ring is electrostatically vibrated into an in-plane elliptically shaped primary flexural mode with fixed amplitude.

When it is subjected to rotation around its normal axis, the Coriolis force causes energy to be transferred from the primary mode to the secondary flexural mode, which is located 45° to the primary mode. This causes the amplitude to build up proportionally in the latter mode, which is capacitively monitored.

Features

The inherent symmetry of the structure makes it less sensitive to spurious vibrations. Only when the ring has mass or stiffness asymmetries can environmental vibrations induce a spurious response. Since two identical flexural modes of the structure "with nominally equal resonant frequencies" are used to sense rotation, the sensitivity of the sensor is amplified by the quality factor of the structure, resulting in higher sensitivity. The vibrating ring is less prone to temperature changes as the vibration modes are affected equally by temperature. Any frequency mismatch due to mass or stiffness asymmetries that occurs during the fabrication process can be electronically compensated by use of the balancing electrodes that are located around the structure.

9.4.3. *Optical gyrometers*

9.4.3.1. *Ring laser gyrometers*

The gyrolaser exploits the Sagnac effect. Two coherent light waves traveling in opposite directions on the same closed loop surface (S) interfere. A rotation of the whole loop with an angular rate Ω around an axis normal to S produces a displacement of the interference strips. This corresponds to a path difference of 2 ΩS/c between the two directions (c being the speed of the light). Two coherent beams coming from the same source are separated by a semi-transparent mirror and reflected by mirrors. Using laser resonance modes, the sensitivity is increased (Figure 9.23).

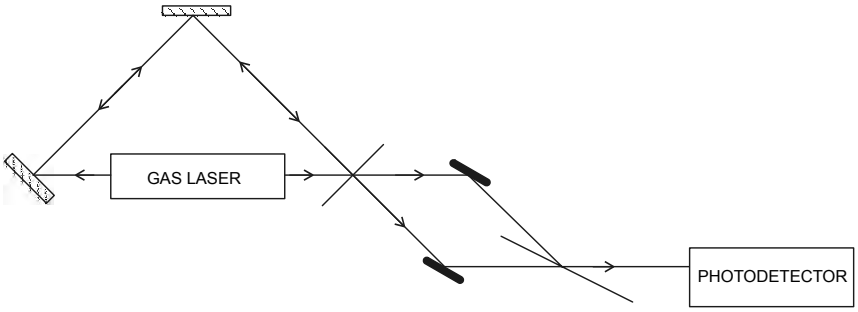


Figure 9.23. *Principle of the laser gyrometer*

Theoretical difficulties

1) Linearity error

In practical applications, e.g. civil avionics, linearity is desirable in the region of 0 to 100°/s. However, this depends on the stability in terms of the dimensions of the cavity, the evolutions of the laser, and the diffusion of the mirrors.

2) Blind zone

Below a certain angular rate, retrodiffusion by the mirrors of a weak part of the incidental waves involves a coupling of the two oscillators: the two contra-rotating waves are in phase and $\Omega_{measured} = 0$. The cause of this phenomenon is the index variation and the mirror surface irregularities.

To eliminate this defect, a wheel of activation is used, which mechanically adds a sinusoidal rotation movement of zero average effect to the block rotation. This maintains most of the measurements in the linear range.

3) Drift (false zero)

When motionless, the gyrolaser measures a non-zero angular rate. The reason for this phenomenon is rooted in the dissymmetrical movement of the gas of the amplifying medium, consecutive with the excitation discharge. The solution is to align the symmetry of the optical system (dimension, mirrors), with the discharges (position of the electrodes, supply power).

New evolutions of laser gyrometers have appeared to improving the characteristics of the traditional laser gyrometers. These include:

– triaxial gyrometers (PIXYZ of Thales, Monolithic Triad basic, Laser Monolithic Boxing ring Gyro of Kearfott);

– gyrometers without mechanical activation (ZLG or FLAG from LITTON, ZLG from POLYUS);

– miniature gyrometers made out of fusible glass (GG 1308 from HONEYWELL, GLC 8 from SAGEM).

9.4.3.2. *Fiber optic gyrometers (FOG)*

This gyrometer uses fiber optics as an interferometer of the Sagnac effect, which presents the following characteristic: two beams follow the exact same course, the only difference being the direction of propagation of the light (Figure 9.24).

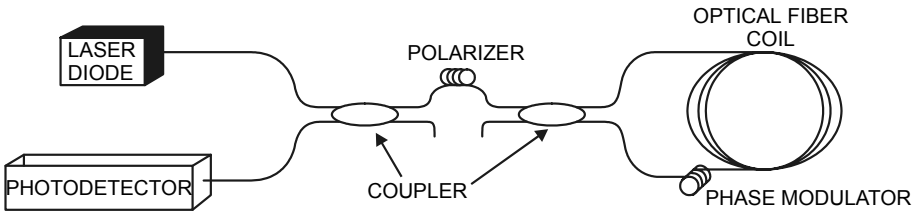


Figure 9.24. *Principle of the gyrofiber (from [14])*

A monochromatic wave delivered by a coherent source is divided into two waves a and b. These are simultaneously sent by two optical fibers and then recombined at their exit of fibers. When the interferometer is at rest, the optical ways of the two waves a and b are both equal to $2\pi R$ (assuming n = 1). If the interferometer is animated by an angular rate Ω the optical path of wave a is lengthened, and that of wave b shortened causing temporal dephasing:

$$\Delta\Phi = \omega\Delta t = \omega\frac{2\Delta l}{c} \tag{9.12}$$

With the difference of propagation time:

$$\Delta t = 4\pi R^2 \Omega / c^2 \tag{9.13}$$

$$\Delta \Phi = 4\pi R^2 \omega / c^2 \tag{9.14}$$

In a medium of unspecified index, this reasoning remains true. Similarly, if the fiber optics are rolled up N times around a support of ray R, dephasing is given by:

$$\Delta \Phi = \frac{2\omega R}{c^2} L\Omega \tag{9.15}$$

Dephasing is thus proportional to

– the angular rate Ω;

– the length of the fiber $L = 2\pi RN$.

Traditional fiber optic gyrometer

Sensitivity

The sensitivity of the gyrofiber is limited by the photonic noise similar to any optical interferometer. The signal-to-noise ratio is thus proportional to the square root of the return power on the detector. Current technology enables the detection of return powers in the range of 1-10 μW. This causes white noise in the measurement and dephasing of 10^{-6} to 3.10^{-7} rad/Hz$^{(1/2)}$, for λ = 840 nm, which is the wavelength most usually used in this application. This corresponds to a noise equivalent of rotation from

– 1.5 to 0.5 (°/h)/Hz$^{(1/2)}$ in the case of an average sensitivity (L = 200 m, D = 30 mm);

– to 0.03 (°/h)/Hz$^{(1/2)}$ in the case of an increased sensitivity (L = 1 km, D = 90 mm).

For a given diameter, the sensitivity of dephasing increases proportionally to the length of fiber used. This also increases the losses and thus the relative noise of detection. For a limitation by the theoretical photonic noise, there is an optimal length L_{op} which, expressed in kilometers, is $L_{op} = 8.7/\alpha$; α is the attenuation in decibels per kilometers. This length corresponds to an attenuation of 8.7 dB, that is, 4.35 km for 2 dB/Km silica fibers with 840 nm. In practice, however, other considerations come into play, and the lengths used are generally of a few hundred meters.

Noise and drift

The signal of a gyrofiber at rest is a random function, which is the sum of white noise and a function varying slowly to cover the long-term drift of the average value. The white noise is generally expressed in terms of standard deviation per square root of bandwidth, i.e. in $(°/h)/Hz^{(1/2)}$. The spectral density of noise is also used by taking the square of the preceding value and is therefore expressed in $(°/h)^2/Hz$. This noise is limited by the photonic noise, whereas the drift corresponds to the residue of non reciprocity of the interferometer in ring. For a coil of great sensitivity, a noise of 0.03 $(°/h)/Hz^{(1/2)}$ and a drift of ±0.003°/h are obtained.

For inertial navigation, the drift is a fundamental parameter. Indeed, the gyrometer measures the angular rate, which is then integrated numerically to deduce the variations in angular orientation. This process of integration produces an average of the white noise. After 100 seconds of navigation, the filtering of the noise reduces the standard deviation to 0.003°/h, and the system then becomes limited by the drift of 0.003°/h.

Retroreflexion and retrodiffusion

It is very difficult to completely eliminate parasitic noise waves. There are six waves which can interfere at the output of the interferometer:

– two reciprocal primary waves, which, after having been propagated in opposed directions, have perfectly identical amplitudes;

– two reflected waves at the input with different amplitudes;

– two reflected waves at the output with different amplitudes.

If the six waves are coherent, there are additional interferences. The error term appears as a phase difference, with an amplitude (not intensity) dependency between the primary waves and the reflected waves. To limit the effect of this term to 10^{-7} rad a suppression extinction of 140 dB is required.

To reduce the phenomena of retrodiffusion:

– by using a pulse source: in this case only the light retrodiffused in the medium of the coil can interfere with the transmitted primary impulses;

– the use of a source with broad spectrum and length of short coherence is another possibility, which is equivalent in theory, but is much more powerful in practice. Moreover, it is easier to obtain a broad source with great extent of spectrum (several percent of the central frequency) what leads to very low widths of correlation (20 to 50 μm). An impulse of equivalent correlation should last only about a 100 FS!

For a given average power (determining the quality of the signal-to-noise ratio), the peak power is greatly reduced and the non-linearities of the source, the medium of transmission and the detector are avoided.

Source

The first experiments of gyrometry with fiber optics used a standard monomode fiber coil, a He-Ne laser and traditional optical components. Nowadays, gas lasers have been replaced with semiconducting diodes. In particular, superluminescent diodes (SLD) give weak temporal coherence as well as good space coherence, ideal for the applications of high performances. The standard multimode laser diodes are not suitable because of their greater temporal coherence. They can be useful for the average performances. Electroluminescent diodes (LED) have an adequate temporal coherence, but their very weak space coherence diminishes their effectiveness in coupling the monomode fibers, limiting their use to low performance applications.

The majority of successful experiments used the 820 to 850 nm window of transparency of silica fibers, with GaAs diodes. This allows the use of standard silicon detectors with very good performances. There are other wavelength windows of 1,300 and 1,550 nm, but they are used for telecommunications or to obtain the best held with the nuclear radiation.

Coil

Long-term stability is an important characteristic when dealing with optic fibers. This is achieved by using fibers which conserve polarization within the fiber; therefore, there is no need to control polarization. To carry out compact coils, the fibers must have good resistance to the effect of a small radius of curvature in terms of:

– loss;

– conservation of polarization;

– and also of static fatigue.

Possible solutions to those problems:

– cores with strong doping;

– fibers of small section. This makes it also possible to limit the size of the coil.

Detector

The choice of the detector is significant not to degrade the performances of the set which should be limited by the photonic noise. In practice, the PIN diodes in

silicon have the best answer (0.55 A/W), equivalent to a quantum effectiveness of 80%.

Conclusion

Many solutions exist for eliminating the various sources of noise and drift (see Table 9.2).

Treatment of the sources of noise	
Sources of noise	**Solutions**
Rayleigh retrodiffusion	Use of a broad spectrum
Acoustic vibrations and disturbances	Good conditions of winding
Detection noise	Good technologies make it possible to reach the photonic noise
Limitation of the space rejection of the monomode filter ensuring the reciprocity	In practice the fiber ensures perfect filtering
Polarization	Limitation of the rejection in polarization
Heat gradient	Symmetrical winding and precaution of insulation
Magneto-optical Faraday effect	Use of fiber which conserves polarization
Non-linear Kerr effect	Source with broad spectrum
Demodulation error	Numerical solution

Table 9.2. *Various solutions to minimize the sources of noise*

It is possible to obtain high quality fiber gyros. However, they will not reach the sensitivity of the most sophisticated mechanical gyroscopes, just like they will not reach the precision of scale factor of the best gyrolasers. However, they are suitable and often used in military applications.

According to constructions the characteristics change:

– speeds from 40 to 500°/s;

– a skew (bias) from 60 to lower than 1°/h;

– accuracies from 1 to 20 (°/h)/$Hz^{(1/2)}$.

The benefit of using these gyrofibers of great sensitivity is their low cost.

9.4.4. *Other original principles*

Other principles have been studied, but are not commonly used by manufacturers because they are not industrially viable. These include magneto-hydro-dynamic gyrometers, resonant ring gyroscopes (RRG) and gyroscopes with nuclear magnetic resonance.

9.5. Calibration of rate sensors

The gyroscopic components are characterized and calibrated by means of rotary tables with 1, 2 or 3 axes. This allows a very good definition of the behavior of the components to be tested, in terms of position and speed.

These tables are characterized by:

– a small wobble (oscillation caused by an unbalance and generally expressed in second of arc);

– a small defect of orthogonality;

– a very good precision of position;

– a great stability of speed.

Some of these tables are equipped with thermal boxes to simultaneously test speed and temperature.

Calculations are used to simulate the values of the input measurements and to exploit information of exit of the sensors.

These standards are essential for the characterization of gyrometers. They can also be used to characterize accelerometers. Generally attitude or inertial measuring control units (IMU) are comprised of 3 accelerometers and 3 gyrometers.

9.6. General features of the gyrometers

Physical Principle	Advantages	Limits	Uses	Class of price A < B < C
Rotary	precise, well-known technology	parts in movement sensitive to vibrations and shocks, large size, power consumption	indicators of turn on planes, detectors for autopilot	B
Integrator	information specifies for localization, tested technology	old technology, expensive, significant size	inertial central of attitude or localization for missiles, planes, submarines	B
Matched suspension	good reliability, long lifespan	sensitive to temperature, sensitive to linear acceleration, significant size	navigation	B
Vibrating	high accuracy, digital output, large range, no moving parts, possible integration, great diversity of realization adapted to very diverse technologies, can be realized at low cost and in great quantity	high degree of precision of realization necessary to obtain a good quality of resonator, sensitivity relating to linear accelerations, complexes electronic treatment	control of stability of the vehicles, navigation and localization of the vehicles, recording accidents, stabilization of platforms, instrumentation, stabilization of cameras, robotics, positioning of antennae, guidance of missiles	A to C

Mechanical resonator	precision, digital output, no moving parts	complex implementation, sensitive to the vibratory environments	navigation, guidance	B
Magneto-hydro-dynamic	great dynamics, high degree of accuracy, broad bandwidth, low sensitivity to linear accelerations, electric low fuel consumption	semi-automatic technology	navigation, guidance	B
Gyrolaser	no moving parts (solid state), good linearity, great dynamics, high degree of accuracy (up to 10^{-2} °/h), digital exit, very short time of startup, not very sensitive to linear accelerations and the environment, good stability in time	plugs zone which complicates and increases its realization, relatively large size (the performance is proportional to the length of the optical path), delicate technology, high cost	inertial navigation for planes, launchers	C
Fiber optics	nuclear radiation resistance material, resistant to extreme environmental conditions, external vibrations and impacts, digital output, no moving parts	sensitivity limited by the photonic noise and the parasitic waves, limited volume, relatively complex electronics of exploitation	guidance, piloting in severe environments with precise details of the order of °/h.	B
Nuclear magnetic resonance	good resilience to the mechanical constraints of environments, no moving parts	low sensitivity, sensitivity to the local magnetic fields, complex electronics of treatment, significant volume	no industrial development identified to date	C

9.7. The main manufacturers

Manufacturers	Website
ATA Sensors	http://www.atasensors.com
BAE Systems	http://www.baesystems.com
BEI Sensors & Systems	http://www.systron.com
BEI Sensors & Systems Co.	http://www.beiied.com
Bosch (D)	http://www.bosch.com
Bosch Corp (US)	http://www.bosch.com
Charles Stark Draper Laboratory Inc.	http://www.draper.com
Condor Pacific Industries Inc.	http://www.condorpacific.com
Denso (JP)	http://www.denso.co.jp
Delphi Automotive Systems	http://www.delphi.com
Endevco Corp.	http://www.endevco.com
Fuji Electric	http://www.fujielectric.co.jp
Honeywell Sensing and Control	http://www.honeywell.com/sensing
Honeywell (Allied Signal)	http://www.honeywell.com
Kearfott Guidance & Navigation Corp.	http://www.kearfott.com
Litton Systems Inc.	http://www.littoncorp.com
Litef Gmbh *(μ-CORS)*	http://www.litef.de
Matsushita electronics	http://panasonic.co.jp
Microsensors Inc.	http://www.microsensors.com
Motorola Inc.	http://www.motorola.com

Murata (JP)	http://www.murata.co.uk
MPC Products Corp.	http://www.mpcproducts.com
Oregon Micro Systems Inc.	http://www.omsmotion.com
SensoNor asa	http://sensonor.com
Silicon Sensing Systems	http://www.siliconsensing.com
SpaceAge Control Inc.	http://www.spaceagecontrol.com
Thales Avionic	http://www.thalesgroup.com
Temic	http://www.temic.com
VI Technology Inc.	http://www.vi-tech.com
Xensor Corp.	http://www.xensor.com

9.8. References

[1] HECHT E., *Physique*, translation from 1st ed. by Becherrawy T., revision by Joël Martin, ITP Deboeck University S.A. 1999.

[2] BURDESS J.S., HARRIS A.J., CRUICKSHANK J., WOOD D. and COOPER G., "A review of vibratory gyroscopes", *Engineering Science and Education Journal*, December 1994.

[3] MURATA MANUFACTURING CO. LTD, http://www.murata.co.jp/products.com.

[4] WATSON INDUSTRIES Inc., http://watsongyro.com/company_info/technology.html.

[5] MADNI ASAD M., GEDDES ROBERT D., "A Micromachined Quartz Angular Rate Sensor for Automotive and Advanced Inertial Applications", BEI Technologies, Inc., 1999.

[6] BEI Sensors & Systems Co., http://www.beiied.com.

[7] INSTITUTE OF MICROMACHINING AND INFORMATION TECHNOLOGY (HSG-IMIT).

[8] JUNEAU T., PISANO A.P. and. SMITH J.H., "Dual axis operation on a micromachined rate gyroscope", in *Tech. Dig. 9th Int. Conference Solid State Sensors and Actuators (Transducers 97)*, Chicago, IL, June 1997, pp. 883–886.

[9] ANDERSSON G.I., HEDENSTIERNA N., SVENSSON P., PETTERSSON H., "A Novel silicon bulk gyroscope", *Extended abstract Transducers '99*, Paper no. 3D1.1 1(4), The IMEGO Institute, SensoNor, Monolitsystem AB, Autoliv AB, Research.

[10] CONDOR PACIFIC INDUSTRIES INC.

[11] YAZDI N., AYAZI F., NAJAFI K., "Micromachined Inertial Sensors", *Proceedings of the IEEE*, vol. 86, no. 8, 1998.

[12] SPARKS D., CHIA M. and ZARABADI S., "Reliability of Resonant Micromachined", paper series 2001-01-0618, Delphi Automotive Systems, Reprinted from Sensors and Actuators 2001, (SP–1609), SAE 2001 World Congress, Detroit, Michigan, March 5-8, 2001.

[13] LITTON SYSTEMS INC., http://www.littonapd.com/html/main.html.

[14] QUIST S., MARTINELLI V., IKEDA R., "Fiber-Optic Gyroscopes in Automotive and Industrial Applications", *Sensors*, April 1996, pp 42–45.

[15] BRITISH AEROSPACE DYNAMIC, http://www.baesystems.com/

[16] LRBA, Laboratoire de Recherches Balistiques et Aérodynamiques de Vernon, Eure, France Forêt de Vernon, 27200 Vernon.

[17] IDEAL AEROSMITH INC., http://www.ideal-aerosmith.com/idealcat.htm.

[18] CSEM Center Suisse d'Electromécanique, Neuchâtel.

[19] DELPHI-DELCO ELECTRONICS SYSTEMS http://www.delphiauto.com.

9.9. Bibliography

1. MARK J., TAZARTES D., FRIDIC B., CORDOVA A., "A Rate Integrating Fiber Optic Gyro", *Navigation*, Vol. 38, no. 4.Winter 1991-92.

2. LEFEVRE H., *The Fiber–Optic Gyroscope*, Artech House 1993.

3. EICHNER R., HANSEN R., OUELLETTE R., "A Ring Laser Gyro Inertial Measurement Unit Designed For System Integration Flexibility", Northrop Electronics Systems Division, Norwood.

4. "Senseurs inertiels", Direction des Recherches, Etudes et Techniques, Journée Thématique, 10 November 1994.

5. PUTTY M.W., NAJAFI K., *A Micromachined Vibrating Ring Gyroscope*, 1994, General Motors Research and Development Canter, 30500 Mound Rd., Warren, MI., 48090-9055, USA.

6. SYSTRON DONNER INERTIAL DIV., "Global Positioning System (GPS) and Inertial Measurement Units (IMU) for Combat ID", April 1995, Internal Report.

7. MADNI A.M., *et al.*, 12-16 October 1998, "Solid-State Six Degree of Freedom Motion Sensor For Field Robotic Applications", *Proc. 1998 IEEE/RSF Intl. Conf. on Intelligent Robots and Systems (IROS)*, Canada, Vol. 3:1389-1398.

8. MADNI A.M., *et al.*, 3-10 February 1996, "A Microelectromechanical Quartz Rotational Rate Sensor for Inertial Applications", *Proc. 1996 IEEE Aerospace Applications Conf.*, Aspen, CO, Vol. 3:315-332.

9. MADNI A.M., *et al.*, 9-12 November 1997, "A Miniature Yaw Rate Sensor for Intelligent Chassis Control", *Proc ITSC'97, IEEE Conf on Intelligent Transportation Systems*, Boston, MA, Paper No. 0002.

10. JUNEAU T., LEMKIN M., CLARK W.A., ROESSIG T.A., "Commercialization of precision inertial sensors with integrated signal conditioning", *Sensors Expo 1998*, San Jose, CA, May 1998.

11. JUNEAU T., CLARK W.A., PISANO A.P., HOWE R.T., "Micromachined Rate Gyroscopes", in *Microengineering for Aerospace Systems*, The Aerospace Press, El Segundo, California, 1999.

12. HSU Y., DEROO D., MURRAY J., "Low Cost Rate Sensor for Automotive Applications", The Fabless Strategy, MicroSensors Inc., 3001 Redhill Avenue, Bldg. 3 Costa Mesa, California 92626, USA.

13. SHMUEL MERHAV, "Aerospace Sensor Systems and Applications", Springer Verlag.

Chapter 10

Magnetic Sensors

The term "magnetic sensor" is broadly used for sensors using principles of magnetism. The main part of this chapter is devoted to magnetic field sensors and electric current sensors; magnetic position sensors are also covered in Chapter 7.

Magnetic sensors are usually contactless and robust and therefore they have reached a dominant position in the industrial and automotive sector. Basic information on magnetic sensors and their applications can be found in [1] and [2], more detailed information on some sensor types can be found in [3-5]. Basics on magnetism and magnetic materials are covered in [6-8].

10.1. Introduction

When a conductor carrying a current is placed in a magnetic field, the voltage distribution within the conductor is modified. The changes depend on the relative orientation of the current, the magnetic field and the measured changes. These effects have been classified into three groups:

1. *Hall effect:* the applied magnetic field is along the z-axis perpendicular to the current along the x-axis and the Hall voltage is measured along the y-axis perpendicular to both the current and the magnetic field.

2. *Longitudinal magnetoresistive effect:* the magnetic field is applied along the x-axis, parallel to the applied current also along the x-axis. The voltage change is

---

Chapter written by S. RIPKA and Pavel RIPKA.

measured along the x-axis. This is effectively a small change in conventional resistance.

3. *Transverse magnetoresistive effect:* the magnetic field is applied along the y-axis transverse to the current flow along the x-axis and the voltage changes are measured along the x-axis. Again, this corresponds to a change in the conventional resistance, but there are effects due to the thickness or cross-section of the conductor.

This classification is adequate for conductors which are isotropic such as most metals and many semiconductor materials. If the conductor is a magnetic material such as iron, cobalt, or nickel, then the direction of any prior magnetization of the metal has a profound effect on the magnetoresistive properties as we will see when we discuss anisotropic magnetoresistors.

10.2. *Hall sensors*

The most popular magnetic sensors are Hall sensors which are used for measuring magnetic flux densities greater than 1 mT and operate well in the temperature range from –100°C to +100°C and in the frequency range from DC to 30 kHz. They are used for measuring linear position, angular position, velocity and rotational speed. Hall magnetic sensors are also commonly incorporated into the brushless DC motors used in VCRs, CD-ROM drive ventilators and disk drives. Hall sensors are used in sensing electrical current where they have the advantage of maintaining galvanic isolation between the measured and the measuring circuits. Modern automotive solid state ignition and ABS braking systems make use of geartooth sensors and in these applications the ambient temperature can be as high as 180°C with junction temperatures as high as 200°C.

Integrated Hall sensors are robust, unaffected by dirty environments and are low-cost. In contrast to other magnetic sensors, the manufacture of Hall magnetic sensors does not require special fabrication techniques as they are compatible with microelectronics technology. Most of the sensors are low-cost discrete devices but an increasing proportion now come in the form of integrated circuits. The integrated Hall magnetic sensors usually incorporate circuits for biasing, offset reduction, temperature compensation, signal amplification and signal level discrimination. The most advanced Hall sensors incorporate digital signal processing and are programmable.

10.2.1. *The Hall effect*

We present a brief summary of the physics of the Hall effect here with more detailed descriptions being available elsewhere [4].

When a long current-carrying conducting strip is placed in a uniform magnetic field, the motions of all of the charge carriers in the strip are affected by the Lorentz force, F:

$$F = qE + q(v \times B)$$

where q is the electrical charge of the carrier, v is the velocity of the charge carrier and B is the vector representing the magnetic field.

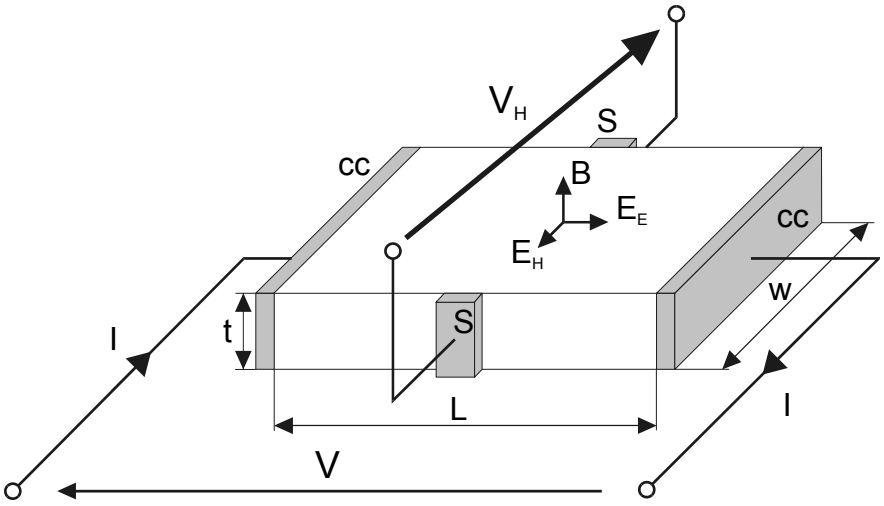


Figure 10.1. *The Hall sensor (adapted from [2])*

Assume that the strip material is a heavily doped n-type semiconductor so that we can neglect the presence of holes. A constant voltage is applied along the length of the strip in the x direction giving an electric field E_e which causes an average drift velocity v_{dn} of the charge carriers. Then:

$$v_{dn} = \mu_n E_e$$

where μ_n is the drift mobility of the n-type carriers (electrons).

The corresponding current density is then given by:

$$J_n = q_n \mu_n E_e$$

where q_n is the elementary charge.

If the thermal motion is neglected, the magnetic part of the Lorentz force is given by:

$$F_{mn} = q_n (v_{dn} \times B) = q_n \mu_n (E_e \times B)$$

The electrons are pushed towards the upper edge of the strip. Due to the increase in the electron concentration at the upper edge of the strip and the compensating decrease at the lower edge a transverse electric field, called the Hall field, appears in the region between the edges of the strip. This hall electric field, E_H, acts on the electrons with a force:

$$F_{en} = -q_n E_H$$

For an electron moving along the strip in the x direction these two forces balance each other. By combining the two previous equations we obtain:

$$E_H = -\mu_{Hn}(E_e \times B) = r_H \mu_n (E_e \times B)$$

where μ_{Hn} is the Hall mobility of n-type carriers and r_H is the Hall scattering factor which reflects the influence of the thermal motion of carriers and of their scattering on the Hall effect. It is found that $r_H \approx 0.8$.

The Hall electric field is perpendicular to both the applied electric field and to the magnetic field. The magnitude of the Hall field is also proportional to the carrier mobility. Since the mobility of p-type carriers (holes) is always lower than the mobility of electrons, it is better to use an n-type semiconductor than a p-type semiconductor as a field sensor.

The external electric field may be expressed in terms of the current density by substituting for J_n as follows:

$$E_H = -\mu_{Hn}(E_e \times B) = -\frac{r_H \mu_n}{q_n \mu_n}(J_n \times B) = -R_H (J \times B)$$

where we have introduced a Hall coefficient, R_H.

The sensitivity S_v of silicon, at room temperature, is $S_{Vmax} \approx 0.126$ V/VT, in GaAs 0.67 V/VT, and in InGaAs 0.78 V/VT. In the development of the Hall devices we are searching for other high-mobility materials, such as InSb.

There are many causes of electrical asymmetry such as small geometrical errors, variations in doping density, contact resistance, mechanical stresses and piezo-resistive effects in the Hall device.

We typically find $B_{off} \approx 10$ mT, 1 mT, 0.1 mT for Si, InGaAs, and InSb Hall devices, respectively, when microelectronics technology is used.

Offset voltage varies with temperature and time. Even if we eliminate all other influences, there remain long-term fluctuations of the output voltage due to $1/f$ noise. These fluctuations correspond to $B_{off} \approx 10$ μT in high-quality silicon Hall devices.

The temperature coefficient of the magnetic sensitivity Si is about 0.1%/K. If we want to extend the operating range to higher temperatures, wide band-gap semiconductors are used instead of Silicon (i.e. GaAs up to 175°C).

An example of a high accuracy Hall sensor is described in [8].

10.2.2. *New types of Hall sensors*

The main directions in the development of modern Hall sensors are application of new materials and geometries, and integration with analog and digital electronics. Also important are new types of packaging minimizing the stress on the chip while keeping the cost low.

10.2.3.1. *High mobility InSb Hall elements*

Early InSb Hall elements were mainly fabricated from thin bulk single crystal InSb, making them expensive and not suitable for mass production. The newly developed thin film InSb elements have a high input resistance of approximately 350 Ω so they are stable when driven by a drive voltage of 1 – 2 V and they can be driven by a constant voltage. Use of a constant voltage drive instead of a constant current drive reduces the temperature coefficient of the Hall output voltage, V_H, from –2.0%/deg to ±0.1 – 0.2%/deg near room temperature [9].

10.2.3.2. *Integrated Hall sensors*

Fabrication of the Hall element into linear bipolar silicon does not require any additional process steps, it is simple and cheap, and both quality amplification and

temperature compensation are easily achievable and reliable. An integrated Hall sensor is shown in Figure 10.2.

In order to cancel several of the components of Hall offset, dual or quad Hall elements are used. The magnetic sensitivity is also increased. Integration technology is used in many magnetic position-sensing switches.

Bipolar circuits are still preferred for industrial and automotive applications because of their high reliability at high temperatures and the ability to withstand repeated high voltage transients, but the application of MOS (metal-oxide-semiconductor technology)-based Hall effect circuits is increasing. Integrated Hall sensors, which can withstand voltage transients in excess of 100 V on the supply and output pins, are produced in high volume using standard linear bipolar silicon technology.

Their resolution is 0.5 mT over the temperature range –40°C to 200°C. CMOS (Complementary MOS) switches may be used to commutate the bias and sense contacts of symmetrical Hall effect elements so as to achieve maximum offset reduction [10].

Because the Hall voltage is quite small, high amplification is required in Hall effect sensors. For instance, 1.0 mV is a typical maximum Hall output voltage. DC offset of the amplifier will fundamentally limit the usefulness of the output. The DC offset can be eliminated if the application allows for AC coupling and other circuit limitations will set the overall achievable accuracy [11].

Constant current bias has a poorer signal-to-offset ratio than constant voltage bias, and biasing the Hall element with constant current makes it difficult to trim the offset. The nominal offset voltage of biased Hall element is independent of temperature; the temperature coefficient of the offset voltage is the result of packaging stress. Laser trimming can remove the basic offset.

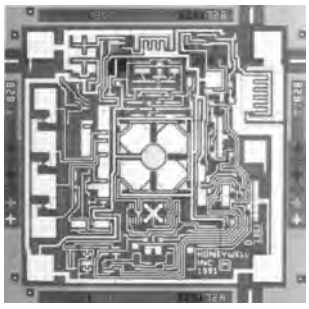


Figure 10.2. *Integrated Hall sensor (courtesy of Honeywell)*

10.3. AMR sensors

The anisotropic magnetoresistors (AMR) are, in general, suitable for use in the measurement of magnetic fields in the range up to 200 μT. The AMR sensors have high sensitivity, wide operating temperature range, offsets which are more stable than those of the Hall sensors and a wide operating frequency range approaching 10 MHz. Another advantage is their low sensitivity to mechanical stress when suitable packaging technology is utilized. The high sensitivity of AMR sensors has led to their application in traffic counting, contactless measurement of electrical currents, measurement of movements and rotational speed in machinery, earth field sensing and in electronic compasses and navigation systems.

At present, AMRs have increasing importance in the automotive industry with applications such as pedal position measurement, wheel speed sensors for ABS (anti-block system) and engine management systems where they are used to measure position to tenths of a millimeter and crankshaft angle for electronic ignition timing.

Details about the latest types of AMR sensors can be found in other works [12-13] and on the websites maintained by manufacturers such as Philips, Honeywell and others [14-15].

10.3.1. O*perating principles of AMR effect*

The magnetoresistance effect was discovered in 1857 but it is only in the last 30 years that it has become of any practical importance. In the anisotropic magnetoresistance (AMR) effect the specific resistance of ferromagnetic alloys (usually Permalloy which is an alloy of Fe and Ni) measured in a direction parallel to the direction of magnetization of the Permalloy, ρ_r, is slightly higher than the resistivity measured perpendicular to the direction of magnetization, ρ_k. We then have for the average value of the nominal resistivity:

$$\rho_p = \frac{(\rho_r + \rho_k)}{2} \quad \Omega m$$

and the absolute difference, $\Delta\rho$, of the resistivities in the parallel and perpendicular directions with respect to the magnetization vector, M, is:

$$\Delta\rho = \rho_k - \rho_r \quad \Omega m$$

Then the relative change of the nominal resistance value is given by:

$$\frac{\Delta\rho}{\rho_p} = 2\frac{\rho_k - \rho_r}{\rho_k + \rho_r}$$

and is a few percent (4.2% for Permalloy 82% Fe / 18% Ni, 3% for Permalloy 81% Fe / 19% Ni).

The AMR effect is due to the quantum mechanical scattering of electrons by the magnetized states of the material and a detailed explanation is beyond the scope of this text.

The basic principle of operation of the AMR sensor is shown in Figure 10.3. A thin strip of Permalloy is deposited on a substrate using a sputtering process. During the sputtering process a high magnetic field is applied in a direction along the strip which causes a characteristic uniaxial anisotropy with the easy axis aligned with the direction of the applied magnetic field. Let H_x represent the magnetization field after the deposition process has been completed. A typical value of this field is about 300 A/m.

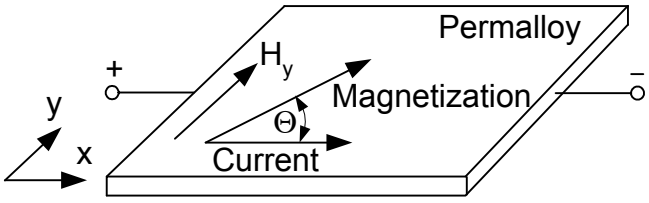


Figure 10.3. *AMR sensor*

The magnetic field which is to be measured, $H = H_y$, is applied along the y-axis and in the plane of the strip so that it is perpendicular to H_x. This field causes the internal magnetization of the strip to rotate away from the x-axis by an angle θ approximately given by:

$$\sin\theta = \frac{H_y}{H_x} \quad \text{for } H_y < H_x$$

This equation is valid only for ideal thin films for which the spontaneous magnetization changes by so-called coherent rotation. The angle of rotation is calculated from the principle of minimum energy.

For the field dependence of the resistance it can be derived:

$$R(H_Y) = R_0 + \Delta R\left[1 - \left(\frac{H_Y}{H_0}\right)^2\right] = R_0 + \Delta R \cos^2 \Theta$$

10.3.1.1. *Geometrical linearization of the AMR*

We can easily observe that the resistance of AMR is quadratically dependent on the measured field H (H_x is constant) and this quadratic factor makes it difficult to use the effect in a linear sensor. The change in resistivity as a function of the measured field is shown in Figure 10.4.

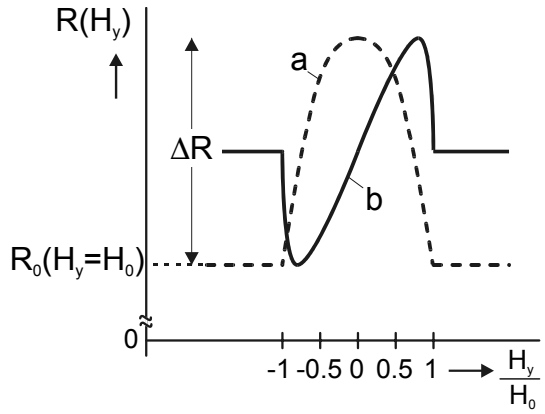


Figure 10.4. *Resistance of a) without linearization; b) with barber poles*

In this linear region where the current is flowing at 45° to the magnetization we find that any change in the magnetization gives a proportional change in the resistance. This configuration is realized by using the so-called “barber pole”

structure in which the Permalloy strip is covered with aluminum stripes oriented at 45° to the long axis of the strip. Aluminum has a much higher conductivity than Permalloy and the effect of the barber pole structure is to rotate the current direction through 45° as shown in Figure 10.5. This results in the change of the rotation angle of the magnetization M relative to the current from θ to θ – 45° between the aluminum stripes in the structure. Then resistance equation becomes:

$$R = R_0 + \Delta R \cos^2(\Theta + 45°)$$

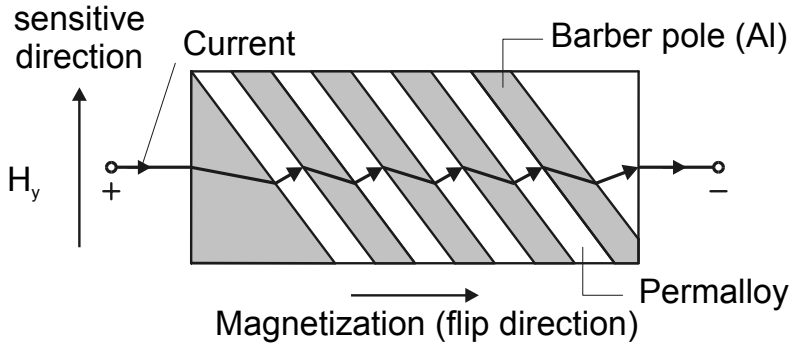


Figure 10.5. *Barber poles*

Applying the goniometric relation:

$$\cos(45° + \Theta) = \frac{\sqrt{2}}{2}(\cos\Theta - \sin\Theta),$$

we obtain the result:

$$R = R_0 \pm \Delta R \frac{H_Y}{H_0}\sqrt{1 - \left(\frac{H_y}{H_0}\right)^2}$$

where "±" represents two different possible orientations of the aluminum stripes on top of the P_y strip. In the case that $H_y << H_0$, we obtain:

$$R = R_0 \pm \Delta R \frac{H_Y}{H_0}$$

It is obvious that this equation represents the linear dependence on measured field.

10.3.2. *Measuring configuration of AMR*

AMR sensors are usually used in a Wheatstone Bridge configuration with diagonally opposite sensors having barber pole orientations of ±45°. This arrangement helps to:

– reduce the temperature drift of the sensor;

– increase the sensitivity of the sensor.

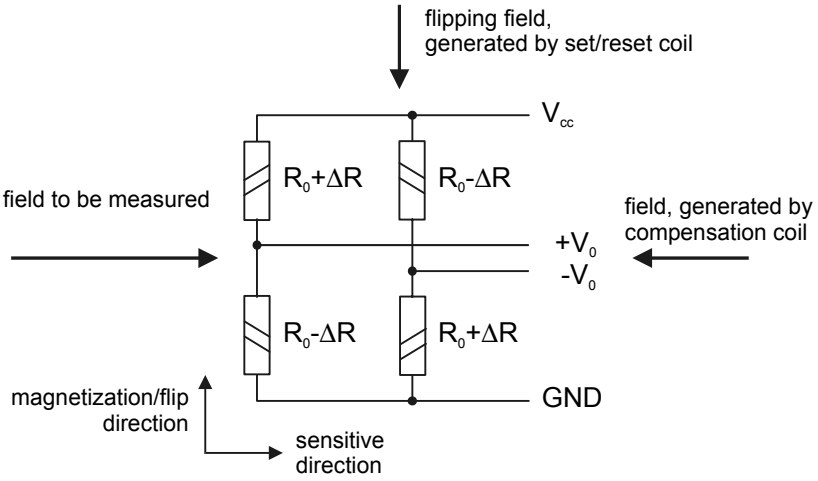


Figure 10.6. *AMR Wheatstone Bridge*

The best results are obtained from a Wheatstone Bridge configuration when a current source rather than a voltage source drives the bridge. Use of a current drive doubles the bridge linearity and, in the ideal case, the temperature dependence is reduced.

The typical layout of the AMR sensor is shown in Figure 10.7.

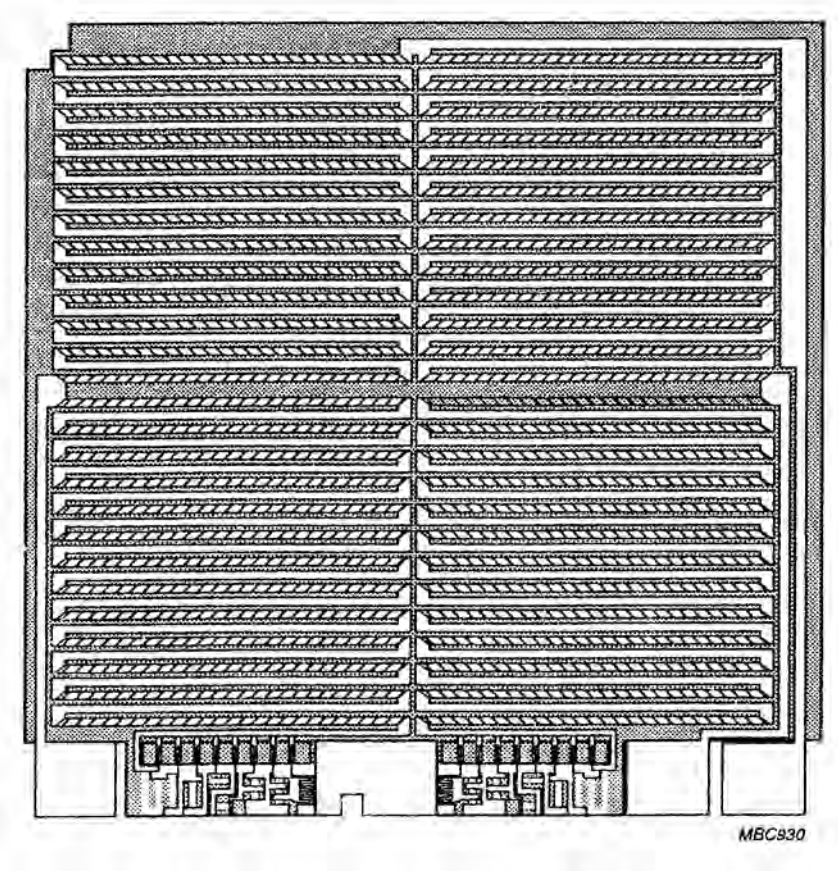


Figure 10.7. *AMR sensor layout (courtesy of Phillips)*

10.3.3. *Flipping*

Flipping is a technique which has been patented by Phillips which involves a periodic change or reversal of the AMR sensor magnetization. Flipping improves sensor precision at the expense of power consumption and achievable bandwidth.

When the AMR sensor output is plotted as a function of the magnetic field, H_y, one of the responses shown in Figure 10.8 is obtained. It can be seen that the voltage output is not zero for zero H_y. If a pulse of current, the flipping pulse, is passed

through a separate magnetization coil surrounding the sensor and if this current pulse is large enough to reverse the magnetization of the Permalloy of the sensor, then the other response curve in Figure 10.8 is obtained. If, at any value of the magnetic field H_y, the amplitude of the sensor voltage change due to flipping is measured, it will be found that this voltage change will represent twice the signal due to field H_y and that the offset signal will cancel out because it is present in both flipping states.

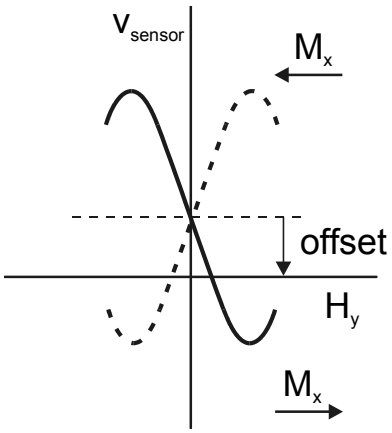


Figure 10.8. *AMR sensor output after flipping of both polarities*

Typical values of the magnetic fields which are applied in the x direction in order to cause this magnetization flip are of the order of 300 A/m. The flipping field is only applied for about 10 μs so the power consumption is not excessive.

A synchronous detector is used which adds the signal from one polarity flip state and subtracts the signal obtained in the other polarity of flip state and then takes the average of this composite signal. The flipping frequency is typically about 200 Hz and if the synchronous detector averages over a number of flip cycles, it can be seen that the system has a response time of 0.1 seconds or more. This is adequate for measurement of the weak magnetic field of the Earth. It is not necessary to use the flipping techniques when only changes of the magnetic field are to be measured such as in road traffic detection and counting applications.

10.3.4. *Magnetic feedback*

Sensor temperature drift of sensitivity

Although flipping eliminates the offsets that are important when low fields are being measured, it does not eliminate the effects of temperature drift which causes a reduction in sensitivity with increasing temperature. Compensation for this temperature sensitivity drift is achieved by using a current source to drive the AMR bridge. As the Permalloy in the AMR warms up, the resistance increases but the constant current drive causes the drive voltage across the bridge to increase and also causes the bridge output voltage to increase and compensate for the reduction in sensitivity.

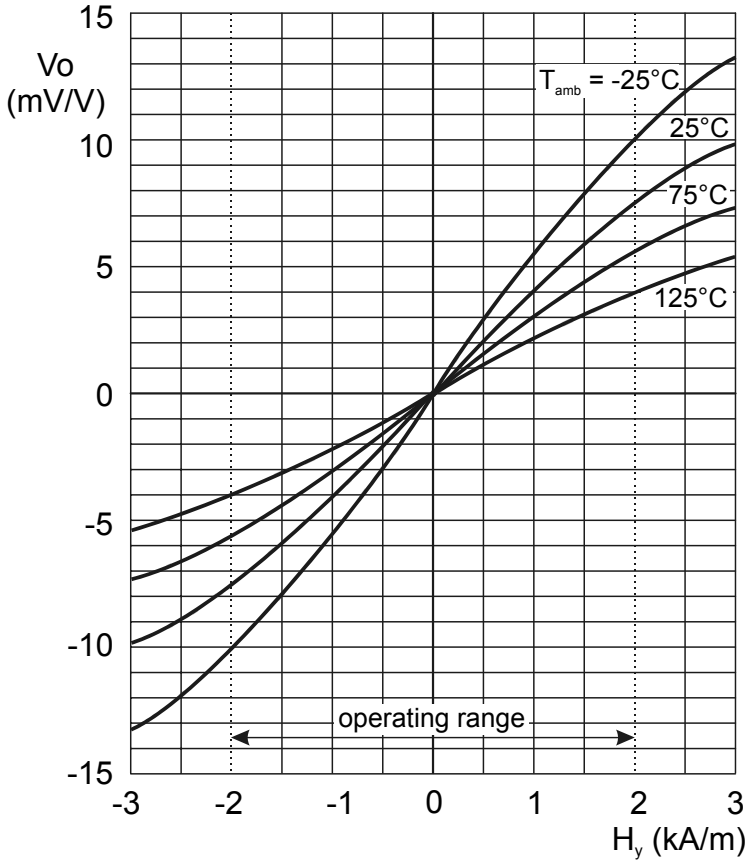


Figure 10.9. *Temperature dependence of AMR characteristics*

Another method to compensate for the temperature dependence of the sensitivity is to use magnetic feedback. We can see from the sensor characteristics in Figure 10.9 that the sensor output is independent of changes of temperature for zero values of the magnetic field, H_y. We now wind a second "compensation" coil around the sensor, perpendicular to the flipping coil, which gives a magnetic field in the same direction as the field which is to be measured, H_y. The output from the synchronous detector is amplified and used to drive a current through this compensation coil so as to cancel out the external field H_y and bring the output of the synchronous detector to zero. Use of this negative feedback nulling system means that the sensor is always operating at the zero field position where the temperature effects are zero. The final output signal is therefore the current in the compensation coil that is required to cancel the magnetic field that is being measured. This configuration thus has a highly linear response and has an increased sensor range. The disadvantage is that the frequency response is reduced.

Some modern magnetoresistors such as the Honeywell HMR 1000 and the Phillips KMZ 41 have feedback coils incorporated into the chip. It should be mentioned that magnetic feedback cannot reduce temperature drift of the sensor offset.

The basic structure of feedback compensated AMR magnetometer is shown in Figure 10.10.

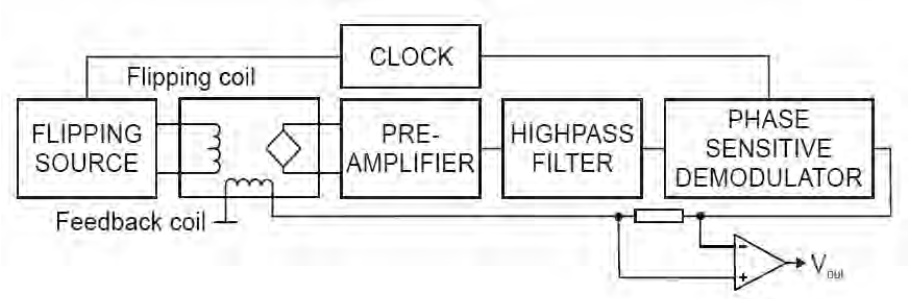


Figure 10.10. *AMR sensor with magnetic feedback*

10.4. GMR sensors

In the late 1980s, several researchers discovered the GMR (Giant Magneto Resistance) and it was soon used in disk drives [16]. As a young technology, it is relatively undeveloped compared to other magnetic sensing technologies. The development is very fast for applications with the large market (i.e. read heads), but

it is slower for applications with low volumes and very specific sensing requirements (e.g. low field magnetometers) [17].

In a giant magnetoresistive sensor, the resistance of two thin ferromagnetic layers separated by a thin non-magnetic conducting layer is changed if the magnetic moments of the ferromagnetic layers are changed from anti-parallel to parallel [18] (see Figure 10.11).

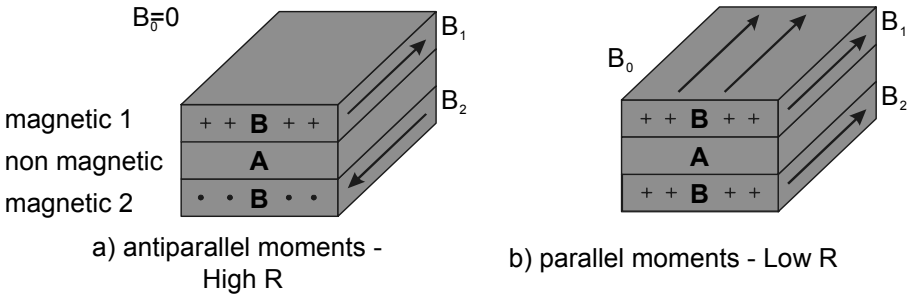


Figure 10.11. *Giant magnetoresistive effect: a) anti-parallel layer moments, b) parallel moments; measuring electric current is flowing in the laser plane (in the same direction as B_1) (after [18])*

Layers with parallel magnetic moments will display less scattering at the interfaces, longer mean free paths, and lower resistance than the layers with anti-parallel magnetic moments. The layers must be thinner than the mean free path of electrons (typically < 10 nm); if not, the spin-dependent scattering cannot be a significant part of the total resistance.

There are several methods of obtaining anti-parallel magnetic alignment in thin ferromagnet-conductor multilayers. The structures used in GMR sensors are unpinned sandwiches, anti-ferromagnetic multilayers, and spin valves.

"Unpinned" sandwich GMR structures consist of two soft magnetic layers (iron, nickel or cobalt) separated by a layer of non-magnetic conductor such as copper. There is a relatively small magnetic coupling between the layers which are 4-6 nm thick, separated by a conductor layer typically 3-5 nm thick. The magnetic field caused by a measuring current along the stripe is sufficient to rotate the magnetic layers into anti-parallel or high-resistance alignment. To rotate the magnetic moments of both layers into parallel alignment, an external field of 3-5 mT should be applied along the length of the stripe. The applied field perpendicular to the stripe has little effect on the resistance.

Anti-ferromagnetic multilayers consist of multiple repetitions of alternating conducting magnetic and nonmagnetic layers. They have more interfaces than the sandwiches and the nonmagnetic layers are thinner (1.5–2.0 nm), so the size of GMR effect is larger. For certain thicknesses of the non-magnetic layer, the polarized conduction electrons cause anti-ferromagnetic coupling between the magnetic layers. The condition needed for maximum spin-dependent scattering, i.e. each magnetic layer has its magnetic moment anti-parallel to the moments of the magnetic layers on each side, is then fulfilled. If a large external field is applied, it overcomes the coupling that causes this alignment, and can cause the low-resistance state by aligning the moments of all the layers so that they are parallel. If the wrong thickness of the conducting layer is chosen, the same coupling mechanism can cause ferromagnetic coupling between the layers resulting in no GMR effect.

It should be noted that the magnitude and sign (anti- or ferromagnetic) of the magnetic interaction depends on the thickness of the (non-magnetic) spacer. For ferromagnetically coupled layers the application of the magnetic field does not produce any large change in resistance.

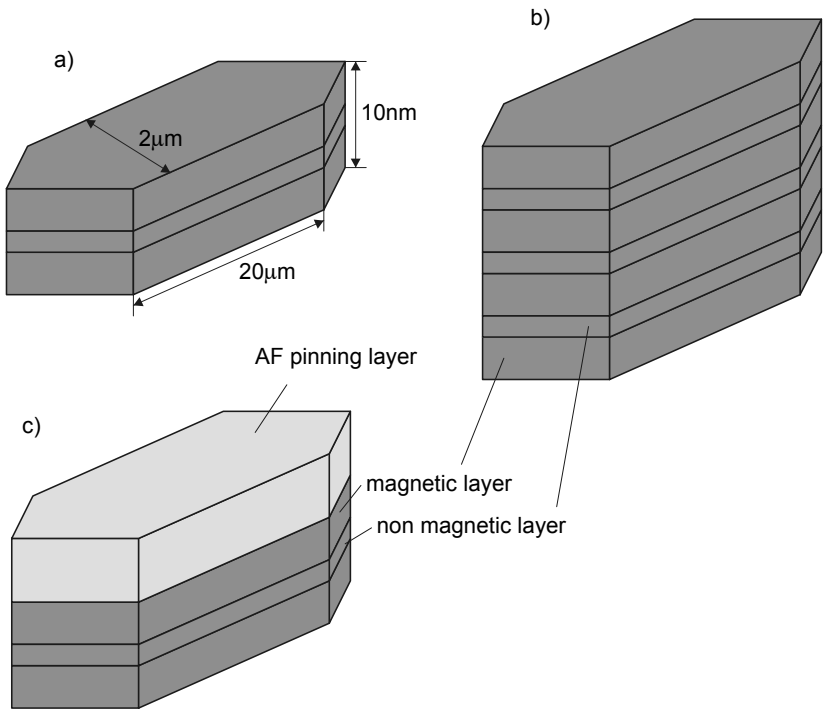


Figure 10.12. *Basic GMR structures: a) unpinned sandwich, b) multilayer, c) spin valve [2]*

10.4.1. *Physical mechanism*

The basic physical mechanism of GMR is "spin dependent scattering". The GMR effect is based on differences in the conduction properties of spin-up and spin-down conduction electrons. Certain ferromagnetic metals have a significant difference in the density of states for spin-up and spin-down electrons. We can visualize that spin-up electrons are those electrons whose spin is aligned with the magnetization and that this configuration is the preferred configuration with a longer lifetime and that there are therefore more spin-up than spin-down electrons which are aligned anti-parallel to the magnetization vector. When the non-magnetic film between two magnetic layers is thin enough to allow the electrons to pass from one magnetic layer to another, those electrons experience the state of the other magnetic layer. When the magnetizations of the two magnetic layers are parallel, their densities of states are matched and no unusual scattering occurs. However, in the case of two anti-parallel magnetized magnetic layers, the electrons traversing the non-magnetic layer experience an "unfriendly" environment. The electrons that were spin-up become spin-down and vice versa. These electrons are more scattered and therefore the conductivity is lower.

10.4.2. *Spin valves*

Spin valves consist of two soft ferromagnetic layers separated by a non-magnetic layer [19]. One of the ferromagnetic layers is "pinned" by an adjacent anti-ferromagnetic layer so that only one of the two magnetic layers is free to respond easily to an externally applied field. The layers are just a few nm thick and the total stack thickness is about 10 to 20 nm. The anti-ferromagnetic layer is usually made of an "artificial ferromagnet" structure consisting of several atomic layers with opposite magnetization. It is the layer closest to the magnetic layer which is responsible for pinning.

The response of the spin valve to a magnetic field parallel to the easy axis is rather complex and is shown in Figure 10.13. The applied field of about 1 mT abruptly reverses the magnetization of the soft layer. When the structure is exposed to a strong field (about 20 mT), the pinning of the hard layer is overcome and the "pinned" layer is also reversed. This effect is usually unwanted.

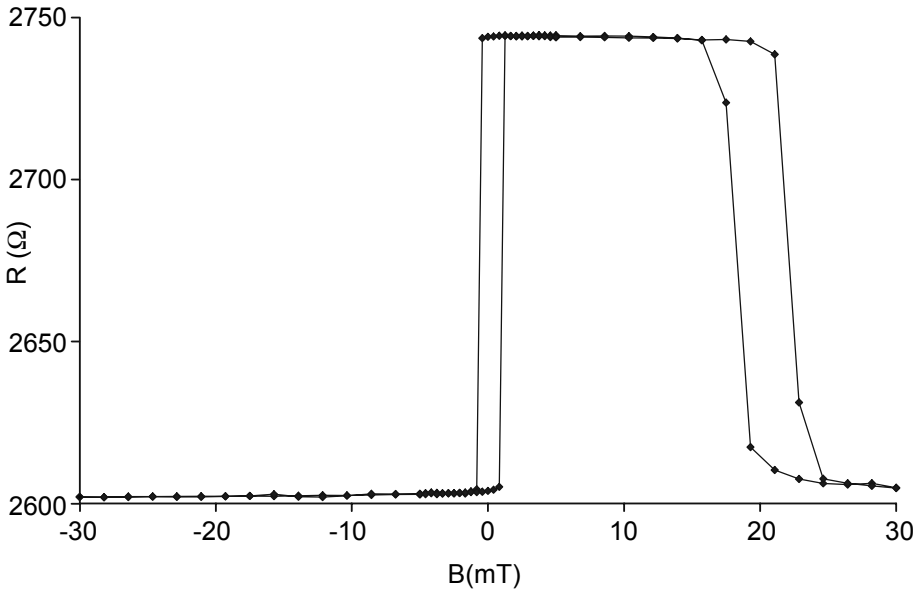


Figure 10.13. *The response of the Spin valve to the field in the easy direction (from [2])*

If the magnetization of the soft layer is rotated in the film plane rather than irreversibly switched, the resistance varies smoothly as a function of the angle between magnetizations of adjacent layers (this is also used in angular sensors, which will be mentioned later). The dependence of the resistance on the angle between magnetizations is given by:

$$R(\theta) = R_{par} + (\Delta R/2)[1 - \cos(\theta)].$$

or also,

$$\Delta R = R_{anti} - R_{par} = R_{max} - R_{min},$$

where ΔR is the magnitude of the maximum change of resistance as the angle θ between the two FM layers goes from parallel to anti-parallel, and R_{par} and R_{anti} are the resistances when the layers are parallel and anti-parallel, respectively.

Differentiating this equation shows that the maximum slope for R(θ) is at θ = 90º. Because of this, the element's R(θ) is at its most sensitive point for very small external fields, when the soft layer of a spin valve element is magnetically biased orthogonally to the hard layer. The total GMR is defined as the ratio (usually as a percentage):

$$GMR = \Delta R / R_{min} = (R_{anti} - R_{par})/R_{par}.$$

Figure 10.14 shows the measured characteristics which closely follows a theoretically predicted curve.

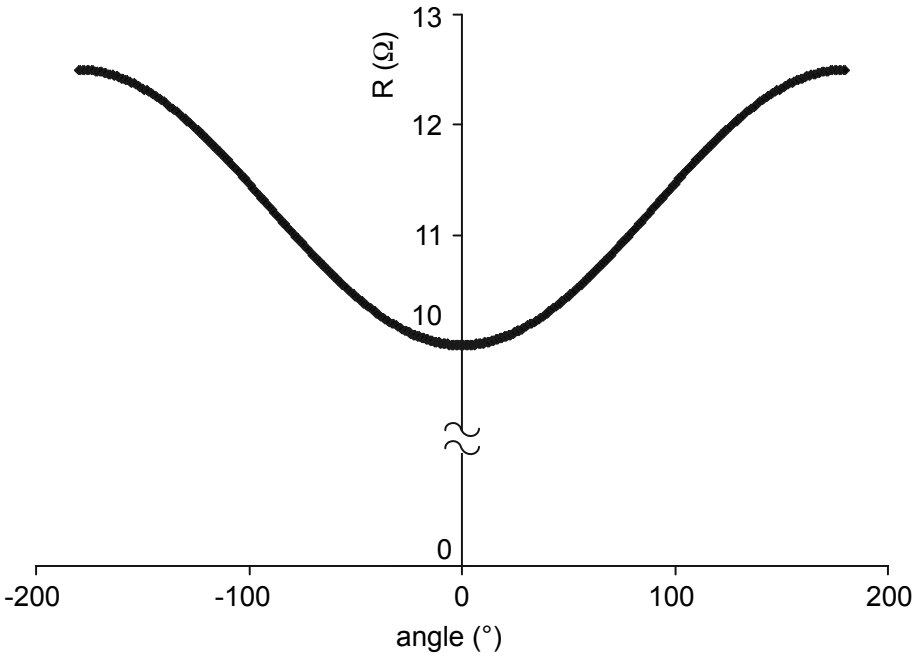


Figure 10.14. *Resistance of the spin valve as a function of angle between the free layer and the pinned layer (from [1])*

10.4.3. *Sandwiches and multilayers*

The magnetic properties of layered structures are strongly dependent on the layer thickness. For larger thickness of the non-magnetic layer the magnetic coupling is weak. This is the case of the sandwich: the layer magnetization is free to rotate, if one layer is pinned, the other is still free. For some smaller thicknesses the magnetic coupling can be ferromagnetic, i.e. individual layers have the same direction of magnetization. In this case we observe no GMR effect. For another value of (still small) thickness of the separation non-magnetic layer the coupling is anti-ferromagnetic, individual layers have magnetization opposite to their neighbors Structures such as this are called artificial anti-ferromagnets. They are used as pinning layers for spin valves. Artificial anti-ferromagnetic multilayers also exhibit GMR effect at high fields, which are able to align the magnetization vectors into their direction. A very large challenge in the construction of GMR films is to make very thin non-magnetic layers of high quality, so that they are not short-circuited by conductive particles. The resistance change (also called GMR ratio) can routinely be made to be 8% in simple sandwich structures and 15% in multilayer structures in practical commercial films, but it can be increased to over 100% by reducing the thickness of the non-magnetic layer to about 1 nm. However, these structures have extremely large saturation fields (> 1 Tesla).

10.4.3.1. *Temperature characteristics*

Some GMR materials are able to operate in environments above 225ºC. The output of all magnetoresistive material has some sensitivity to temperature: these effects come from two sources: 1) the usual increase in resistance with temperature (about 0.1%/K), which is typical for all metals, and 2) the decrease of the magnetic moment and subsequent decrease of the GMR effect with temperature. The GMR decrease with temperature differs according to the type of structure, but it is again around the order of 0.1%/K at room temperature.

10.4.3.2. *Cross-field error*

GMR sensors have very little cross-field sensitivity to z-axis fields due to their thin film (x-y plane) nature (demagnetizing factor is very strong in the z direction). However, some designs of magnetoresistive sensors have significant off-axis sensitivity to y-axis fields (10%). Flux concentrators further diminish the off-axis sensitivity to less than 1%.

10.4.3.3. *Unpinned sandwich*

While the sandwich structure has a simple layer composition, it is harder to control its magnetoresistive response. This is due to the freedom of rotation of the

both magnetic layers. Sensors using unpinned sandwich material have usually a unipolar response and hysteresis.

10.4.3.4. *GMR multilayer*

A GMR "multilayer" consists of repeated layers of ferromagnet/non-ferromagnet/ferromagnet. As already mentioned, the layers have anti-ferromagnetic coupling at zero external field (this comes from minimum energy condition). The advantages of multilayers are higher GMR and the opportunity to make them have linear output vs. field. However, because of the high saturation fields, multilayers are not suitable for sensing weak magnetic fields.

10.4.4. *SDT sensors*

The spin dependent tunneling (SDT) device is a very new structure. Using an insulating material (usually Al_2O_3), instead of the non-magnetic metal, in a sandwich or spin valve makes the measured resistance inversely proportional to the tunneling probability across the thin insulating barrier. The resistance vs. field response of SDT devices has the same shape as that of a sandwich or spin valve, but the details of the conduction mechanism are different [20].

SDT materials have significantly higher sensitivity than GMR spin valves. They are sensitive to the field in the film plane, like all GMR sensors, but the measuring current is perpendicular to this plane. Because of the insulation layer, the resistance of the SDT sensor bridge is high even for small dimensions – this is an advantage for battery-powered devices. SDT sensors are still in the development phase.

10.4.5. *Linear GMR sensors*

Typically, a complete sensor is made of several GMR resistors in a Wheatstone Bridge configuration. Two of the resistors must behave differently from the other two in an external field, in order for the bridge to have a non-zero response. In linear GMR sensors, this is accomplished with flux concentrators, which are two relatively thick (15 μm) layers of Permalloy plated on the GMR sensor chip. Two resistors are positioned in the "gap" between the two concentrators while the other two resistors are under the concentrators, so that the "gap" resistors experience a multiplication of the external field by a factor of about the length to gap ratio D2/D1. The resistors under the flux concentrators are effectively "shielded" from the external field, so their resistance does not change for moderate fields.

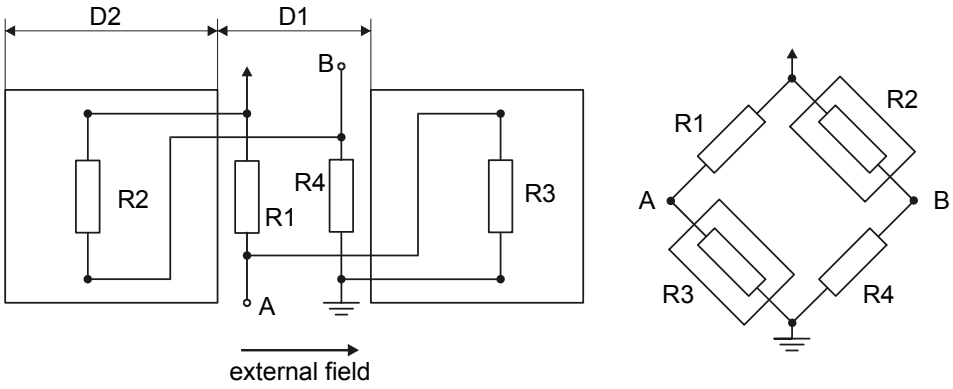


Figure 10.15. *Structure of a GMR sensor bridge with flux concentrators (courtesy of NVE)*

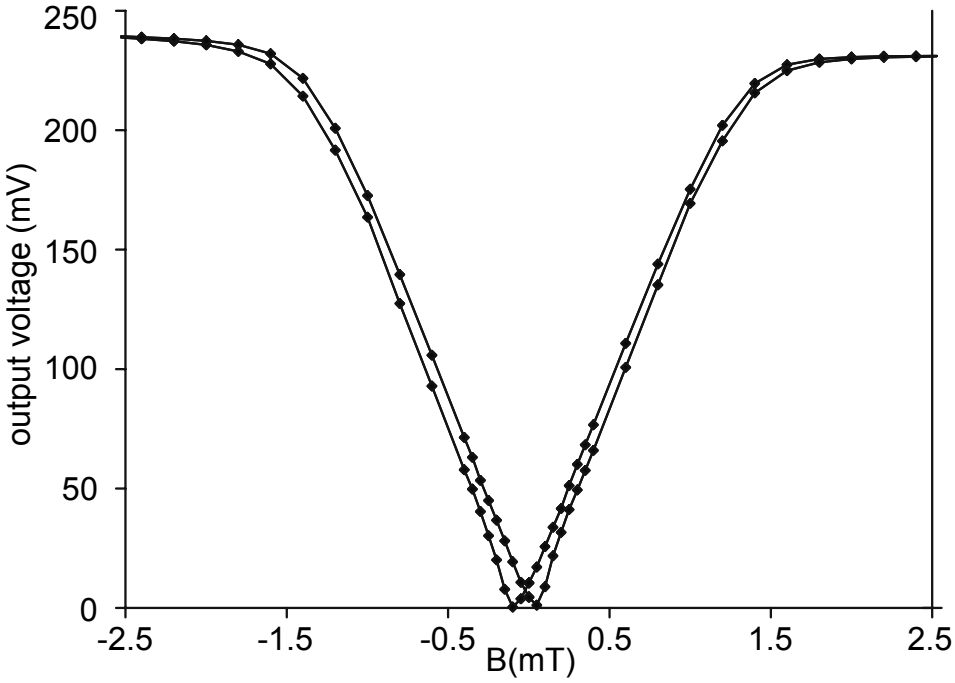


Figure 10.16. *Characteristics of the NVE AA00-02 multilayer bridge sensor*

10.4.5.1. *Bipolar response using biasing coils*

On-chip planar biasing coils offer another technique for generating a non-zero output from a bridge. These coils can be used to create a localized field in different directions for different GMR resistors on the chip.

Care should be taken in achieving sufficient stability of this bias field.

10.4.5.2. *GMR gradiometer*

Another useful configuration is to create a bridge whose opposite resistor arms are identical, but which are separated in space by a relatively large distance. The bridge output is then only zero when the field at the two sensor ends is equal. This type of sensor measures the field gradient rather than the absolute magnitude of the field and is very useful for detecting small nearby objects having relatively low magnetization.

10.4.6. *Rotational GMR sensors*

An Infineon rotational GMR sensor is made as a complex sandwich [21]. The cobalt layers couple to form an artificial anti-ferromagnet. The covering layers are made from soft magnetic iron and line up with an external magnetic field, while the cobalt layers retain their ("hard") magnetization.

The electrons undergo fewer scattering processes if the soft and nearest hard magnetic layers are aligned ("in line"), and the resistance reaches its minimum. The maximum is reached at the opposite orientation of soft and the nearest hard magnetic layers. The GMR effect is independent of the direction of the measuring current: only the angle between the hard and soft magnetic layers determines the total resistance of the system.

Within a wide range of magnetic fields (magnetic window) – where the soft magnetic layers turn with an external field while the hard magnetic layers remain unchanged – the resistance depends only on the direction of the magnetic field and not on its size (saturation mode). If the field is created by permanent magnet, this means that the sensor is tolerant to large changes in the magnet distance and its strength.

Bosch and Philips [22-23] have developed similar sensors.

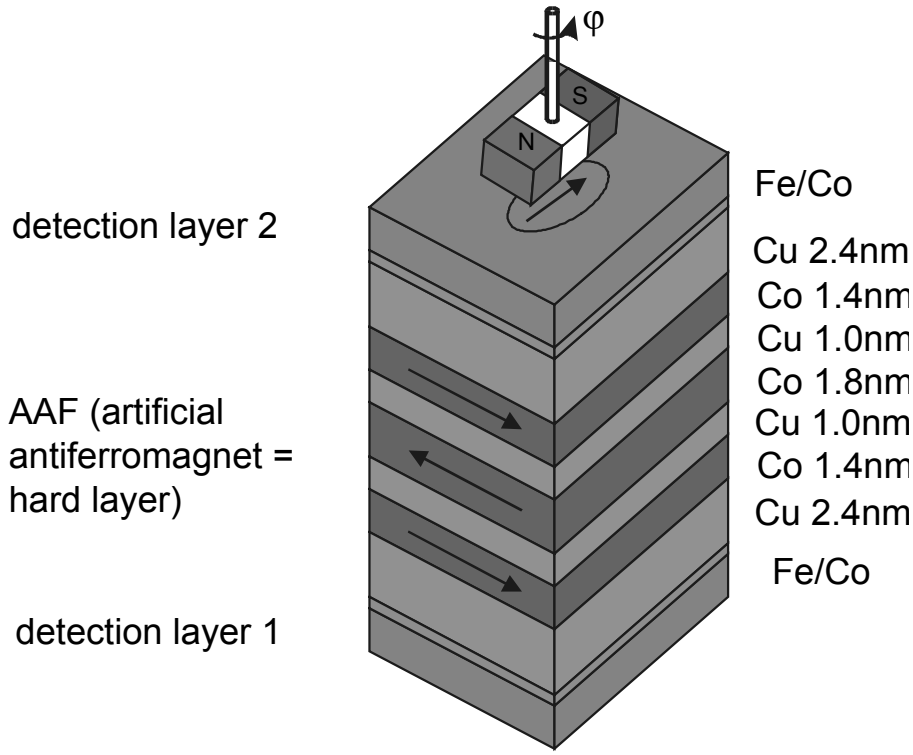


Figure 10.17. *Angular GMR sensor (from [22])*

10.5. Induction and fluxgate sensors

The magnetic flux, Φ, through a single turn coil which is orientated perpendicular to a uniform magnetic field, B Tesla, is:

$$\Phi = AB = A\mu_0\mu_r H \qquad \text{Webers}$$

where A is the area of the coil in m^2 and measurement is carried out in air ($\mu_r = 1$).

If the coil has N turns and has a core made of magnetic material with relative permeability $\mu_r(t)$ which may vary and if the coil is moved in the magnetic field so

that the flux through the coil changes, we then find that the voltage which appears across the terminals of the coil is given by Faraday's law of magnetic induction:

$$\begin{aligned} V_i &= -\frac{d\Phi}{dt} \\ &= -\frac{d}{dt}\{NA\mu_0\mu_r(t)H(t)\} \\ &= -NA\mu_0\mu_r\frac{dH(t)}{dt} - N\mu_0\mu_r\frac{dA(t)}{dt} - NA\mu_0 H\frac{d\mu_r(t)}{dt} \end{aligned}$$

In the final form, the three terms describe respectively:

– Static induction or search coil sensors.

– Rotating coil sensors.

– Basic fluxgate sensors.

10.5.1. *Induction coil sensors*

Traditional induction coil magnetometers consist of a multilayer solenoid. Air coils are very stable and linear, but have limited sensitivity. To optimize air coils they should have a large diameter and be relatively short [24-25]. Coils with a ferromagnetic core have higher sensitivity, but are less stable and markedly non-linear. In order to reach low demagnetization, they should be long and thin. In the design of induction coils for high frequencies the coil self-capacitance should be considered.

Induction sensors are passive, and they should be distinguished from inductance sensors, which are based on the change of the sensor inductance and which need excitation.

Basic applications of induction sensors are in geophysics, where they measure micropulsations of the Earth's magnetic field (1 MHz-1 Hz frequency range), in audiofrequency applications, and in magnetic recording techniques. The measurement of secondary magnetic fields, caused by the Earth's currents after artificial excitations at frequencies up to the audio range, is called magnetotelluric exploration. Geophysical exploration may also use natural electromagnetic field variations in the 1 Hz to 20 kHz band. Induction coils are also used in plasma experiments, in space research, in submarines and trains; magnetic antennae are used for navigation and communication.

Velocity and position detectors using induction coils are very important, but they fail at low or zero speeds. One of the most important applications is the electromagnetic compatibility (EMC) measurement [26]. The total field, measured by three orthogonal coils, is usually calculated by three-axial magnetometers of this type. That can cause large errors for rotating fields. Air-cored coils together with an integrator (fluxmeter) are used to map the dc induction B by an extraction method and also to measure field intensity H. Each induction coil has to be properly calibrated, as the coil constant can be calculated only for very simple air coils.

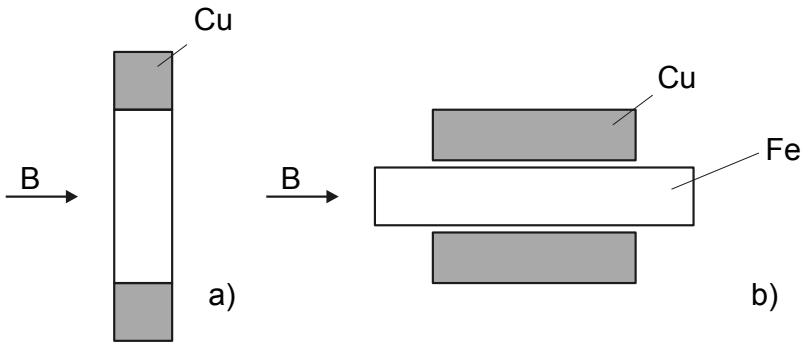


Figure 10.18. *Optimum geometry of induction coils: a) air coil, b) coil with ferromagnetic core*

10.5.2. *Fluxgate sensors*

Fluxgate sensors measure the static or low frequency magnetic field. They are vector devices and are sensitive to both the field direction and field magnitude up to 1 mT with achievable resolution of 100 pT. The principle of operation is that an excitation current I_{exc} through an excitation coil produces an alternating excitation field, which periodically causes saturation of the soft magnetic material of the sensor core (Figure 10.19).

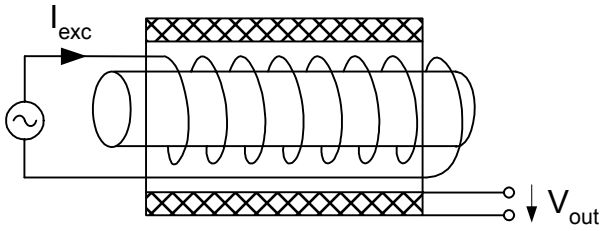


Figure 10.19. *Principle of fluxgate sensor*

Once saturated, the core permeability decreases and the flux associated with the magnetic field which is being measured B_0 decreases. The term "fluxgate sensor" comes from this "gating" or limiting of the flux that occurs when the core is saturated. In the presence of an external magnetic field B_0 the voltage V_i induced in the sensing or pick-up coil contains components even at the second and even higher harmonics of the excitation frequency. The voltage amplitudes of harmonic components are proportional to the external field B_0 and constitute the sensor output. However, some fluxgate magnetometers operate with current sensing in the pick-up coil rather than voltage sensing and in this case the pick-up coil is effectively shorted by the measuring electronics.

Fluxgate magnetometers are used in geophysical and space applications [1]. Fluxgate compasses are used in aircraft, land vehicle and submersible navigation systems. The fluxgate principle is also used in electrical current sensors and current comparators and for the remote measurement of direct currents. Simple fluxgate magnetometers are used for detecting metal objects and for reading magnetic marks and labels.

Fluxgate sensors are reliable solid-state devices and operate over a wide temperature range. Standard commercially produced devices have 100 pT resolution and 10 nT absolute precision but this can be extended to 10 pT resolution and 1 nT long-term stability.

Many fluxgate magnetometers have an upper cut-off frequency response of a few Hz but, when necessary, they can be operated up to kHz frequencies. Typical temperature stabilities are an offset drift of 0.1 nT/°C and a temperature coefficient of around 30 ppm/°C. Some fluxgates are compensated to 1 ppm/°C. If the fluxgate operates in feedback mode the linearity error may be as low as 10^{-5} [27 – 28].

If resolution in the nanotesla range is required, fluxgates are the best selection. Compared to a high-temperature SQUID (superconducting quantum interference device) they may have similar noise level, but the measurement range of a fluxgate is much larger. If pT or even smaller fields are measured, a low-temperature SQUID should be used. Magnetoresistors, mainly AMR (anisotropic magnetoresistance sensors), are the main competitors of fluxgate sensors. Commercially available AMR magnetoresistors have a resolution worse than 10 nT, but they are smaller and cheaper and may consume less energy.

The most commonly used modern low-noise fluxgate sensor is the "parallel" type with a ring-core. "Parallel" type means that the excitation and the measured field have the same direction. A phase-sensitive detector extracts the second harmonic in the induced voltage, and the pick-up coil also often serves for the

feedback. There exist other designs, used for special purposes, such as rod-type sensors for non-destructive testing or position sensing.

10.5.2.1. *Core shapes of fluxgates*

The main problem in using the basic single core design is the large signal at the excitation frequency which is present at the sensor output due to the sensor acting as a transformer. Therefore, the single core design is used only for simple devices. For precise fluxgates, double cores constructed in either the double-rod or the ring-core configuration are normally used as shown in Figure 10.20.

10.5.2.2. *Double-rod sensors*

In this configuration of the sensor, shown in Figure 10.20(a), the sensor core is made of two ferromagnetic wires or strips, which are excited by individual excitation coils in opposite directions. The sensing (pick-up) coil is wound around both rods.

10.5.2.3. *Ring-core sensors*

In this configuration of the sensor, shown in Figure 10.20(b), the excitation coil is wound toroidally around the ring-core which is oriented with a diameter of the ring-core pointing in the direction of the magnetic field which is being measured. In one half of the ring-core, the field due to current in the excitation coil is parallel to the external field B_0 and in the other half of the ring-core the excitation field is anti-parallel to the external field. The ring-core is usually constructed from several turns of thin tape of soft magnetic material [28]. The sensing coil is a simple solenoid with its axis parallel to the field which is being measured and the ring-core is placed at the centre of this sensing solenoid. The ring-core geometry is advantageous for low-noise sensors even though the ring-core sensors have low sensitivity due to large demagnetization.

10.5.2.4. *Race-track sensors*

The sensitivity of the race-track sensor shown in Figure 10.20(c) is higher than the other types and is also less sensitive to perpendicular fields due to the lower demagnetization factor. The advantages of closed-type sensors are still retained.

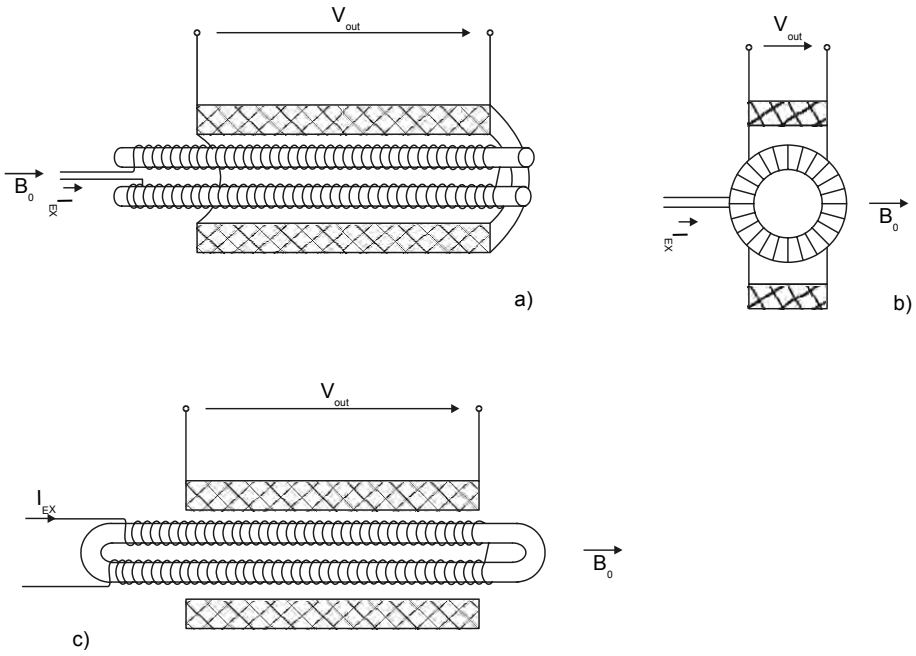


Figure 10.20. *Core shapes of fluxgate sensors: a) double-rod, b) ring-rod, c) race-track*

10.5.2.5. *Principles of fluxgate magnetometers*

The most frequently used principle of fluxgate magnetometers is second harmonic detection of the output voltage. The other principles have not brought any significant advantages except simplification of the circuitry. Fluxgate may also work in the short-circuited mode (with current output) [1].

A diagram of a common feedback-type fluxgate magnetometer is shown in Figure 10.21. The magnetic field which is being measured causes a second harmonic component in the voltage at the sensor output, that is, a component which is at twice the excitation frequency. The phase sensitive detector (PSD) demodulates this second harmonic to DC or near-zero frequency. Fields larger than 1 μT, which is the typical limit of the sensor's linear range, usually have to be compensated. Integrator, INT, gives a large feedback gain. The feedback current is sensed by the differential amplifier, DA, and serves as the magnetometer output. The generator is used to produce a sine wave or a square wave at a frequency typically between 400 Hz and 100 kHz. For crystalline cores a frequency of about 5 kHz is used.

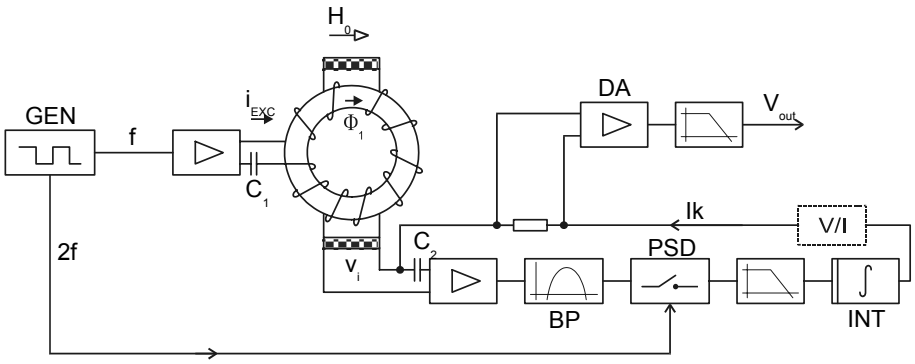


Figure 10.21. *Feedback fluxgate magnetometer [2]*

Increasing the frequency of the excitation current increases the sensitivity but can also lead to a situation where the eddy currents in the core material become important. Increasing the excitation frequency also improves the dynamical performance of the sensor. The excitation current should have a large amplitude and a low second harmonic distortion which would cause a spurious signal in the PSD. The sensor sensitivity also increases with the number of turns in the pick-up coil up to the limitation due to parasitic self-capacitance causing the circuit to become self-resonant. The feedback current source should have a large output impedance to prevent short circuiting of the sensor output leading to a lowering of the sensitivity. Any DC current flowing in the pick-up coil causes sensor offset so the amplifier input must be decoupled with a series capacitor.

In some simple magnetometers, the pick-up coil is also used for the feedback. However, the pick-up coil should be positioned close to the sensor core so as to keep the air flux low and, for homogeneity reasons, a large feedback coil is better, so two separated coils are usually used in precise magnetometers.

10.6 Other magnetic field sensors

10.6.1. *Resonance sensors*

Resonance magnetometers are scalar instruments and measure the magnitude of the field regardless of its direction. They are usually very precise but they often have a long measurement time and they fail in small fields and in large field gradients [1].

Nuclear magnetic resonance (NMR) is used for imaging diagnostics in medicine and for spectroscopic investigations of materials and also for absolute measurement of weak magnetic fields in proton magnetometers.

The proton precession magnetometer utilizes the precession of spinning protons or nuclei of the hydrogen atom in a sample of water or hydrocarbon liquid (e.g. petroleum). The protons behave as small, spinning magnetic dipoles. First their magnetic moments are perfectly aligned by a strong homogenous polarizing magnetic field created by a current in a coil. After the polarizing field is switched off, the magnetic moments are forced into the direction of the measured field. The spin of the protons together with this aligning force means that the trajectory of the proton magnetic moments is not straight: this phenomenon is called precession and it is similar to the precession of the mechanical gyroscope. The precessing protons generate a small signal at angular frequency $\omega = 2\pi f$, which is precisely proportional to the total magnetic field B and independent of the orientation of the sensor.

$$\omega = \gamma B$$

where $\omega = 2\pi f$ is the angular precession frequency, γ is gyromagnetic constant ($\gamma = 2.675\ 10^8$ rad/(s T) for a spherical sample of water) and B is the measured field.

The precession frequency for a field of 1 T is 42.5 MHz. This frequency, typically units of kHz for the Earth's field, is measured by counter. The instrument is capable of measuring the Earth's field of 50,000 nT with an absolute accuracy of 1 nT and resolution of 0.1 nT.

Other resonance magnetometers are based on Electron Spin Resonance (EPS) optically pumped magnetometers.

10.6.1.1. *Magnetic sensors based on electron spin resonance (ESR)*

Electron spin resonance (ESR) is the same phenomenon as NMR, but for electrons. The much higher spin frequency (about 600 times that of protons) offers the possibility of constructing highly sensitive and quick responding scalar magnetometers. The traditional proton magnetometer is excited by a DC-magnetic field followed by measurement of the frequency of the decaying nuclear spin signal. Contrary to that, the optically "pumped" electron spin magnetometers use light in resonance with an optical spectral line of the sample, and they produce a continuous electron spin resonance signal.

10.6.1.2. *Overhauser magnetometers*

The Overhauser effect proton magnetometer combines the nuclear magnetic resonance and the electron spin resonance phenomena [29]. A radio frequency magnetic signal is used for ESR excitation of the electrons in the sample, which then transfer their excitation energy to protons by collisions. Continuous excitation of the protons is thereby maintained and this produces a continuous proton resonance signal. Overhauser magnetometers work in larger field gradients and require a shorter time for the measurement than traditional proton magnetometers.

10.7. Magnetic position sensors

Magnetic position sensors sense either linear or angular position. They are produced with either linear or digital output. Digital output can be either bistable (proximity switches) or encoded (incremental and absolute position sensors). Some of the magnetic distance sensors are described in Chapter 7.

They can measure the position of a permanent magnet (induction sensors), soft magnetic material (LVDT, variable reluctance sensors), or just electrically conducting material (eddy current sensors). Special types of magnetic position sensors are magnetic compasses [30] which use the Earth's field to find the heading, and tracking systems which use an artificial field to find the heading and position.

10.7.1. *Sensors using permanent magnets*

This type of sensor measures the field of a permanent magnet which can be attached to the target. In an alternative arrangement, the magnet is attached to the field sensor and the target is made of ferromagnetic material. In both cases the moving target changes the sensed field. If the sensor is just a passive induction coil, only the movement is sensed and there is no sensing of static position. This type of sensor is commonly called an induction sensor, a speed sensor or a magnetic pick-up sensor. There are other types which use DC-magnetic sensors, most frequently Hall, AMR, GMR or semiconductor magnetoresistors.

10.7.1.1. *Induction position sensors*

Sensors of this type are based on induction effect. They are passive devices. If they contain a moving magnet, they are called "speed sensors". The target, a permanent magnet, is sensed by an induction coil, which is often in the magnetic circuit (Figure 10.22). Their advantage is that they consume no energy, because they are based on induction effect. However, their use is limited, because they fail at low

speeds. The failure is caused by the dependence of the amplitude of the output on the target speed.

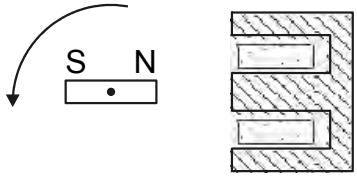


Figure 10.22. *Induction rotation sensor with moving magnet*

Induction sensors with variable reluctance and fixed permanent magnets are called "magnetic pick-ups". The DC field is generated by a fixed permanent magnet, and changing the position of the soft magnetic target, which is again often a part of the DC-magnetic circuit, produces the coil flux change (Figure 10.23). In other words the induction coil detects the perturbation of the DC-magnetic field caused by movement of the target. They are used in gear tooth sensing in shaft speed measurements and anti-lock brake systems (ABS). Because of their low sensitivity at low speeds they are not suitable for car ignition timing systems.

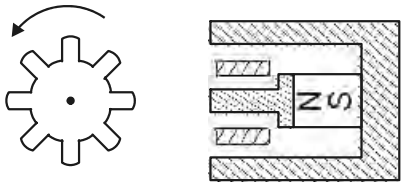


Figure 10.23. *Induction geartooth sensor (the "magnetic pick-up")*

10.7.2. *Eddy current sensors*

Eddy current-based instruments are used to measure displacement, alignment, dimensions, vibrations, and also to identify and sort metal parts in industrial applications. The target should be electrically conducting, but not necessarily ferromagnetic, and it has no lower speed limit [1].

The sensor coil, fed by an oscillator, creates an AC-magnetic field. The coil is tuned by a parallel capacitor. The LC circuit oscillates at the resonant frequency, and if the conducting target is present, the eddy currents (mostly on the target surface) create a secondary magnetic field decreasing the coil flux and thus the effective coil inductive reactance. The conductivity of the target influences the sensitivity. For

example, a good conductor such as aluminum is an excellent target, the best target sizes being lower than 0.3 mm of thickness and a diameter of 2.5 to 3 times the diameter of the sensor coil. The general rule is that the target thickness should be larger than the skin depth δ

$$\delta = \frac{1}{\sqrt{0.5\omega\mu\sigma}}$$

where σ is electrical conductivity.

If the target is ferromagnetic, the situation is complicated, as the coil inductance is increased by the target permeability $\mu > 1$.

10.7.3. *Linear and rotational transformers*

These devices are AC excited. Either the transformer winding or its core is moveable and the output signal depends on its position.

10.7.3.1. *Linear transformer sensors*

These AC powered sensors are based on the change of mutual inductance between two or more coils.

LVDT

The linear variable differential transformer (LVDT) is based on a variation of the transformer coupling factor between the primary and two secondary windings which is caused by movement of the ferromagnetic core. The configuration is shown in Chapter 7 (Figure 7.5). When the core is in the center position, the primary is coupled equally into the two secondary windings. If the core is displaced from the central position, there is more signal coupled into one secondary and less into the other secondary.

Standard measurement ranges are from 200 μm to 50 cm. Practical resolution may be better than 0.1% or below 1 μm. The excitation frequency is usually between 50 Hz and 20 kHz. The output signal is usually processed by PSD (phase sensitive detector = synchronous rectifier, lock-in amplifier) and sometimes ratiometric processing is used. The complete sensor electronics, including the excitation generator, can be integrated into the sensor housing.

The differential variable inductance transducer (DVRT ®, "half bridge LVDT") has only two windings.

Variable gap sensors

The principle of variable gap sensors is the change of the air gap in a magnetic circuit (between the core and armature) of inductor of transformer. Despite being less precise than LVDTs, they are often used in conjunction with mechanical transducers to measure pressure, strain, force, torque, and other mechanical variables that can be converted into mechanical displacement, due to design reasons. These sensors are also described in Chapter 7.

PLCD sensor

The core of PLCD (permanent magnetic linear contactless displacement sensor) is a long magnetic strip with two sectioned primary winding and homogenous secondary windings. The two sections of primary winding, which are supplied with ~4 kHz sine wave, are connected anti-serially. Localized core saturation, caused by a permanent magnet in the core vicinity, effectively divides the core into two halves whose lengths determine the signal induced into the secondary winding. The induced voltage is synchronously rectified in order to obtain linear output. The typical resolution is 0.2%, linearity 1% of the range (which is between 20 mm and 150 cm). The device is tolerant to changes of air gap between the magnet and core.

Inductosyn

This sensor consists of a scale and slider, two parallel flat meander coils. The two windings of the slider are shifted by 1/4 of mechanical period (pitch). The displacement is measured by inductive coupling between the scale and slider coils. Inductosyn has advantages of both incremental sensors (increment is one pitch) and analog sensors (sine wave dependence of the output voltage allows to interpolate the fine position with a resolution of up to pitch/65,000). Inductosyns are also made in a rotary fashion. Different patterns can be combined in one device in order to increase the incremental resolution by employing techniques known from optical encoders (such as N/N-1 method).

10.7.3.2. *Rotation transformer sensors*

Due to their applicability in extreme conditions, as they are more rugged than optical encoders, these sensors are still produced and used.

Synchros

Synchros are electromechanical devices, which replicate the rotor position at a distant location. They have three stator windings displaced by 120°. They combine the properties of sensor and actuator; typical application is the antenna rotator.

Resolvers

Resolvers have two windings displaced by 90°. Specialized resolver-to-digital converters often process the outputs of sine and cosine voltages. In brushless resolvers, another rotational transformer is used to supply the rotor. Resolvers can withstand temperatures from 20 K to 200°C, radiation of 10^9 rads, acceleration of 200 g (battleship cannons, punching devices), vacuum or extreme pressures.

"Linear resolver" is a linear position sensor which also has sin/cos outputs, but which is based on two AC supplied magnetoresistive elements.

10.7.4. *Magnetostrictive position sensors*

Magnetostrictive position sensors measure time of flight of a strain pulse to sense a position of a moving permanent magnet. The sensing element is a wire or pipe from magnetostrictive material (sonic waveguide). The devices are based on the Wiedeman effect: if the current passes through the waveguide and perpendicular DC-magnetic field is present, the torsional force is exerted on the waveguide (Figure 7.13).

The principle of the device is following: when the current pulse is applied, the torsional force is generated in the location of permanent magnet. This torsional strain pulse travels with ~3 km/s speed along the waveguide and is detected by a small induction coil at the sensor head. The hysteresis can be as low as 0.4 μm, uncorrected linearity is 0.02% FS, some devices have an internal linearization and temperature compensation. The maximum sensor length is about 4 m.

10.7.5. *Proximity switches*

A proximity switch can be made using any linear output magnetic sensor with an electronic (usually Schmidt) trigger. In fact most of the proximity switches are based on Hall sensors. Here we mention two very important types of sensors with naturally bipolar output.

10.7.5.1. *Reed contacts*

They are very simple, cheap and totally passive devices consisting of two magnetic strips of soft or semi-hard magnetic material sealed in a glass pipe filled with inert gas. There are two types of contacts: normally open contacts, which are connected to a certain field by an attractive magnetic force between the free ends, and normally closed contacts. Both of them have hysteresis and their switching zones have a complicated shape (Figure 7.11).

High security "balanced" switches use two reed contacts, one normally open and one normally closed, in the vicinity of the magnet. If the magnet moves, it causes transfer of one of the switches. The normally open contacts are usually crossed by a resistor, which makes it possible to monitor the continuity of the wires.

10.7.5.2. *Wiegand sensors*

A Wiegand sensor generates a high voltage pulse when the magnetic field reaches some threshold value. The voltage pulse is highly independent of the rate of the field change and the device is passive, having just two terminals (Figure 10.24). The sensors are made of 0.3 mm wire from Vicalloy (Co52Fe38V12) which is twisted to cause plastic deformation resulting in higher coercivity in the outer shell and elastic stress in the central part. The pulse is caused by one large Barkhausen jump when the single central domain reverses its magnetization. Eddy current damping determines the pulse width. Pulses of 2.5 V, under optimum driving conditions (asymmetric sensed field), which have a frequency independent amplitude between 1 MHz and 100 Hz, may be generated by 30 mm long wire with 1000-turn coil. Optimum working conditions are when the magnetization direction of the inner (magnetically soft) part reverses, while the magnetization of the (magnetically harder) outer shell is constant. Then the device characteristic is asymmetrical, and a large pulse is generated in only one direction of the field change. Due to the fact that the outer shell cannot be made really magnetically hard and therefore an external field can unintentionally remagnetize it, the main application field of Wiegand wires is marking and security application, not magnetic sensing. Wiegand wires are attached to access cards or anti-theft labels where the sensing coil is part of the stationary detection device. The detected pulse can be easily identified in noise, because it has a very characteristic shape. The presence of the switching field is necessary.

"Pulse wires" manufactured by Siemens consist of magnetically soft wire under stress and a parallel wire of magnetically hard material. While the magnetization direction of the magnetically soft wire is changing, the magnetization of the parallel wire has the same direction during the whole working cycle.

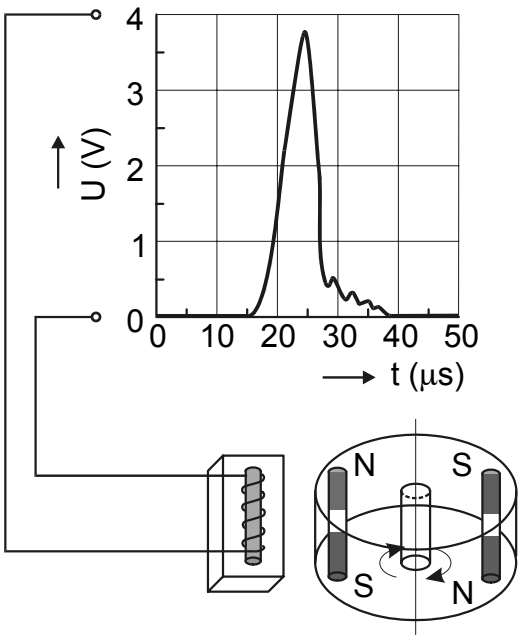


Figure 10.24. *Wiegand sensor (according to Siemens brochure)*

10.8. Contactless current sensors

It is sometimes impractical or impossible to measure electrical current by the usual method of measuring the voltage drop across a shunt resistor. For very large currents, the shunts are heavy and they cause voltage drops and dissipate heat. When it is necessary to measure the current in a conductor which is at high voltage, there is always the danger of electrocution of the operator. Optical current sensors for large currents are only in the development stage. Optical fiber devices are suitable for high-voltage applications, but the reported errors are large even after temperature compensation.

There are many requirements for contactless current sensors. They should be geometrically selective – i.e. sensitive to measured currents and resistant to interferences from other currents and external fields. This is achieved by using a closed magnetic circuit with a measured conductor inside. If this is not possible, magnetic sensor arrays may be used.

10.8.1. *Hall current sensors*

Most current sensors are based on use of a Hall element positioned in the air gap of a magnetic yoke. They have limited zero stability given by the Hall sensor offset: typical offset drift of a 50 A sensor is 600 μA in the (0ºC to 70ºC) range. It is twenty times worse than fluxgate-type current sensor modules. These devices are very sensitive to external magnetic fields and close currents, due to the magnetic leakage associated with the air gap. However, especially for larger currents (> ~10 A) the device precision is typically 1% for the uncompensated and 0.5% for the compensated type, and is sufficient for most industrial applications.

10.8.2. *Magnetoresistive current sensors*

Coreless magnetoresistive (MR) current sensors are based on an AMR bridge which is insensitive to an external field, but sensitive to measured current through the primary bus bar [31].

10.8.3. *AC and DC Transformers*

The primary winding of AC and DC transformers consists of few turns (or a single conductor through the core opening). The secondary winding should ideally be short-circuited.

Current comparators are usually very precise and expensive devices.

AC current comparators have three windings on the ring (toroidal) core. They have errors below 1 ppm in amplitude and $3 \cdot 10^{-6}$ deg in phase. DC current comparators are based on a fluxgate effect. Their core consists of two detection toroids excited in opposite directions and they are usually feedback compensated.

Much simpler are fluxgate-like DC current sensor modules. They are manufactured by VAC and others for measurement ranges between 40 and 200 A. The accuracy of a typical 40 A module is 0.5%, linearity 0.1%, current temperature drift < 30 μA (–25°C – 70°C).

10.8.4. *Current clamps*

These devices consist of a magnetic circuit which ensures that the reading is not dependent on the actual position of the clamped conductor and the device is insensitive to unclamped conductors.

AC current clamps are based on current transformers made on openable ferrite core. A measured conductor forms the primary winding and the secondary winding is terminated by a small resistor, or connected to a current-to voltage converter. Most of the available DC current clamps are based on the Hall sensor in the air gap. DC current clamps are usually made of Permalloy or Si-Fe laminated sheets. DC current transformers work on the fluxgate principle and they are much more precise.

Electric currents with unknown location can be localized and their magnitude can be calculated from the magnetic field measured at several points. This method is used to measure the currents in constructions such as bridges and pipelines.

10.9 References

[1] RIPKA, P.: Magnetic Sensors: Principles and Applications, in J. Buschow (ed.) *Encyclopedia of Materials: Science and Technology*, Elsevier, 2007.

[2] RIPKA, P. (ed.): *Magnetic Sensors and Magnetometers*, Artech, 2001.

[3] BOLL, R. and OVERSHOTT, K.J. (eds): *Magnetic Sensors, Sensors Vol. 2*, VCH Veiden, Germany, 1989.

[4] POPOVIC, R.S.: *Hall Effect Devices*, Bristol: Adam Hilger, 1991.

[5] TUMANSKI, S.: *Thin film Magnetoresistive Sensors*, IOP, 2001.

[6] JILES, D.: *Introduction to Magnetism and Magnetic Materials*, Chapman & Hall, London, 1998.

[7] KRAUS, J.D.: *Electromagnetics*, McGraw-Hill, 2nd ed., 1984.

[8] SCHOTT, C., BLANCHARD, H., POPOVIC, R.S., RACZ, R., HREJSA, J.: "High accuracy analog Hall probe", *IEEE-Transactions on Instrumentation and Measurement*, Vol. 46, no.2, pp. 613–616, 1997.

[9] SHIBASAKI, I.: "Mass production of InAs Hall elements by MBE", *J. of Cryst. Growth*, 1997, 175/176, p. 13.

[10] BELLEKOM, A.A. and MUNTER, P.J.A.: "Offset Reduction in Spinning-Current Hall Plates", S*ensors and Materials*, Vol. 5, no. 5, pp. 253–263.

[11] STEINER, R., MAIER, C., MAYER, M., BELLEKOM, S. and BALTES, H.: "Influence of Mechanical Stress on the Offset Voltage of Hall Devices Operated with Spinning Current Method", *Journal of Micromechanical Systems*, Vol. 28, no. 4, pp. 466–472, 1999.

[12] MAPPS, D.J.: "Magnetoresistive Sensors", *Sensors and Actuators A*, Vol. 59, no. 1, 1997, pp. 9–19.

[13] HAUSER, H., STANGL, G., HOCHREITER, J., CHABICOVSKY, R., FALLMANN, W. and RIEDLING, K.: "Field sensor by anisotropic magnetoresistance", *Journal of Magnetism and Magnetic Materials*, Vol. 216, 2000, pp. 788–791.

[14] DATASHEET, Magnetic field sensors, General, Philips Semiconductors http://www.semiconductors.philips.com.

[15] DATASHEET, Honeywell HMC 1001, www.honeywell.com. Also other datasheets and application notes at http://www.ssec.honeywell.com/products.

[16] TSANG, C., *et al.*: "Design, fabrication and testing of spin-valve read heads for high density recording", *IEEE Trans. Mag.*, Vol. 30, 1994, pp. 3801–3806.

[17] DAUGHTON, J.M., *et al.*: "Magnetic Field Sensors Using GMR Multilayer," *IEEE Trans. Mag.*, Vol. 30, 1994, pp. 4608–4610.

[18] SMITH, C.H. and SCHNEIDER, R.W.: Low Magnetic Field Sensing with GMR Sensors, *Sensors Magazine*, September-October 1999, also at http://www.sensorsmag.com.

[19] SPONG, J.K., *et al.*: "Giant Magnetoresistive Spin Valve Bridge Sensor," *IEEE Trans. Mag.*, Vol. 32, 1996, pp. 366–371.

[20] MOODERA, J., and KINDER, L.: "Ferromagnetic-insulator-ferromagnetic tunneling: Spin-dependent tunneling and large magnetoresistance in trilayer junctions", *J. Appl. Phys.*, Vol. 79, 1996, no. 8, pp. 4724-4729.

[21] INFINEON GMR DATASHEET at www.infineon.com, also Sensor-Applications for Your System Success, brochure, Siemens Electromechanical Components, at http://www.siemens.de/ec/eccs/sensors/magnetic.htm.

[22] RIEGER, G., LUDWIG, K., HAUCH, J., CLEMENS, W.: GMR sensors for contactless position detector, *Sensors and Actuators*, A 91 (2001), 7-11.

[23] LENSSEN, K.-M.H., *et al.*: Robust giant magnetoresistance sensors, *Sensors and Actuators*, A 85 (2000), 1-8.

[24] MACINTYRE, S.A.: "Magnetic field sensor design", *Sensor Review*, Vol. 11, 1991, pp. 7–11.

[25] MACINTYRE, S.A.: "A portable low noise low frequency three-axis search coil magnetometer", *IEEE Trans. Mag.*, Vol. 16, 1980, pp. 761–763.

[26] www.emi.com.

[27] ACUNA, M.H: "Space-based magnetometers", *Rev. Sci. Instrum.*, 2002, Vol. 73, pp. 3717–3736.

[28] NIELSEN, O.V., *et al.*: "Development, construction and analysis of the 'Orsted' fluxgate magnetometer", *Meas. Sci. Technol.*, Vol. 6, 1995, pp. 1099–1115.

[29] DURET, D.N., *et al.*: "Overhauser Magnetometer for the Danish Oersted Satellite", *IEEE Trans. Mag.*, Vol. 31, 1995, pp. 3197–3199.

[30] CARUSO, M.J., and WITHANAWASAM, L.S.: "Vehicle Detection and Compass Applications using AMR Magnetic Sensors", Honeywell, SSEC, 12001 State Highway 55, Plymouth, MN 55441, USA, http://www.ssec.honeywell.com.

[31] DRAFTS, B., "New Magnetoresistive Current Sensor Improves Power Electronics Performance", *Sensors*, September 1999, also on www.sensormag.com/articles/0999.

[31] SO, E., REN, S., and BENNET, D.A.: "High-Current High-Precision Openable-Core AC and AC/DC Current Transformers", *IEEE Trans. Instrum. and Meas.*, Vol. 42, 1993, pp. 571–576.

Chapter 11

New Technologies and Materials

11.1. Introduction: MEMS

MEMS is an abbreviation for "micro electro-mechanical system" [1]. These devices feature the integration of mechanical elements, sensors, actuators and operating electronics on a common silicon substrate with the use of microfabrication technology [2].

While the electronic circuits (either analog or digital) are fabricated using integrated circuit process sequences (CMOS, BiPolar or BiCMOS processes), the micromechanical part is made by micromachining. Micromachining techniques selectively etch away parts of the silicon wafer or add new structural layers to form the required mechanical and electromechanical devices [3].

Microelectronic integrated circuits (ICs) handle the signal processing, while micromachined parts serve as sensors and actuators which allows the microsystems to sense and control the environment.

Components of MEMS are:

– microsensors;

– microactuators;

– microelectronics;

– microstructures.

Chapter written by A. TIPEK, P. RIPKA and E. HULICIUS, with contributions from A. HOSPODKOVÁ and P. NEUŽIL.

Since MEMS devices are manufactured by batch fabrication techniques, similar to ICs, unprecedented levels of functionality, reliability and sophistication can be placed on a small silicon chip at a relatively low cost. With thin films, the photolithographic fabrication procedures make it possible to build extremely small, high precision mechanical structures using the same processes that have been developed for electronic circuits [4].

MEMS promises to revolutionize nearly every product category by bringing together silicon-based microelectronics with micromachining technology, thereby enabling the realization of a complete system-on-a-chip [5].

MEMS technology is enabling new discoveries in science and engineering such as the polymerase chain reaction (PCR) microsystems for DNA amplification and identification, introducing new technologies such as the micromachined atomic force microscopes (AFM), scanning processing microscopes (SPM) and scanning tunneling microscopes (STM), biochips for detection of hazardous chemical and biological agents, and microsystems for high-throughput drug screening and selection.

Examples of MEMS devices, which we encounter everyday, are inkjet-printer cartridges, accelerometers that deploy car airbags and miniature robots [6].

The successful production of MEMS needs the development of appropriate fabrication processes in four major areas:

– micromachining;

– microfabrication;

– micromechanics;

– microelectronics.

|Conventional silicon planar microelectronics technology has been adapted to the processing of both passive and active components. Passive material is one that does not play an essential role in the sensing mechanism (e.g. SiO_2 insulating layer in a pressure sensor) in contrast to an active material, which does (e.g. metal oxide layer in a chemical sensor).

The basic MEMS processes are:

– oxidation;

– diffusion;

– LPCVD (low-pressure chemical vapor deposition);

– photolithography;

– epitaxy;
– sputtering;
– micromachining processes;
– bulk micromachining;
– surface micromachining;
– wafer bonding;
– deep silicon RIE (reactive ion etching);
– LIGA (lithography, electroforming, molding);
– micromolding;
– etc.

MEMS devices are extremely small (e.g. electrically driven motors are smaller than the diameter of a human hair), but MEMS technology is not only characterized by the size [3]. Also, MEMS do not only include products based on silicon, even though silicon possesses excellent material properties (e.g. the strength-to-weight ratio for silicon is higher than for many other engineering materials). MEMS is a manufacturing technology; a new way of making complex electromechanical systems using batch fabrication techniques similar to the integrated circuits [7].

MEMS have several advantages: in traditional sensor-actuator electronic systems, sensors and actuators are the most costly and unreliable parts. In comparison, MEMS allows these complex electromechanical systems to be manufactured using batch fabrication techniques and therefore leading to a substantial decrease in the cost with increased reliability [8].

One example of the advantages of MEMS is the accelerometers for crash air-bag deployment systems in cars. The conventional system uses bulky accelerometers made of discrete components mounted in the front of the car with the separate electronics near the air bag and costs over $50. MEMS have made it possible to integrate the accelerometer and electronics onto a single silicon chip at a cost of $5 to $10. These MEMS accelerometers are much smaller, more functional, lighter and more reliable [3].

The microsensors produced using the silicon process have been developed since 1980, but for a long time they were just a laboratory curiosity. Circuits for preamplifier and logic elements have been integrated with transducers and used to make an intelligent chip element, called an intelligent sensor or smart sensor. Microsensors with moveable parts have been developed since 1985 and have become the first applications of micromechanical parts in the industrial field.

11.2. Materials

Materials used in electronics can play an active or passive role. Very often one material can play both roles.

11.2.1. *Passive materials*

Passive materials could be described as materials which are only used to provide either mechanical structure or electrical connection [9]. The following Tables 11.1 and 11.2 summarize some examples of the physical properties of several materials [10] that determine their use in applications [11]. Some of these materials can be used as active as well as passive materials, mainly silicon and gallium arsenide.

	Si	GaAs	SiO_2	Si_3N_4	GaN
Density [kg/m^3]	2,330	5,316	2,200	3,100	6,150
Melting point [°C]	1,414	1,238	1,600	–	2,500
Thermal conductivity [W/m/K]	168	47	6.5, 11	19	130
Dielectric constant	11.7	12	4.5, 4.3	7.5	4
Young's modulus [GPa]	190	–	380	380	–
Forbidden gap (300°C) [eV]	1.12	1.427			3.2

Table 11.1. *Physical properties of non-metallic materials*

	Al	Au	Cr	Ti
Density [kg/m^3]	2,699	19,320	7,194	4,508
Melting point [°C]	660	1,064	1,875	1,660
Thermal conductivity [W/m/K]	236	319	97	22
Work function [eV]	4.3	5.1	4.5	4.3
Young's modulus [GPa]	70	78	279	40

Table 11.2. *Physical properties of metallic materials (often used in the passive role)*

For more details see [12] or [13].

11.2.2. *Active materials*

These materials are essential to the sensing process used in various types of microsensors [9], such as photosensitive, piezoelectric, magnetoresistive and chemoresistive films.

Nowadays a wide range of functional materials are currently used in microsensors and these often take the form of thin or thick films and play an active role in the sensing system. Some of them can be deposited using IC-compatible deposition techniques (CVD or LPCVD) but others need special techniques such as electrochemical deposition as in the case of conducting polymers [11]. The properties of some active materials are given in Table 11.3.

	Density [kg/cm^3]	Melting point [°C]	Electrical conductivity [10^3 S/cm]	Thermal conductivity [W/m/K]
Thermal				
Pt	21,470	1,769	9×10^4	72
Radiation				
Ge	5,323	937	3×10^{-4}	67
Mechanical				
Quartz AT-cut	1,544	1,880	5.1	2.8
Magnetic				
Fe-pure	7,874	1,535	10^5	449
Chemical				
SnO_2	6,950	1,360	low	-

Table 11.3. *Some properties of active materials*

For more details see [12] or [13].

11.2.3. *Silicon*

Silicon makes up 26% of the Earth's crust by weight. Elemental silicon is not found in nature, but occurs in compounds like oxides and silicates. Silicon is

prepared by heating silica and carbon in an electric furnace, using carbon electrodes. Silicon is under normal conditions a relatively inert element, but it is attacked by halogens and dilute alkali [14].

Silicon is abundant, relatively inexpensive and exhibits a number of physical properties which are useful for sensor application [4].

However, a major problem with silicon is that many of its characteristics are temperature dependent. Silicon does not display the piezoelectric effect and it is not ferromagnetic. Silicon also has no efficient photo- or electro-luminescent properties with the exception of porous or some nanocrystalline forms of silicon, which have not yet found any industrial applications in this field. This is the reason for its limited role in the field of optoelectronic active sources. Although silicon does not display the desired effect, it is possible to deposit layers of materials with the desired properties on the silicon substrate [15].

Single-crystalline silicon

Single-crystalline silicon is the most widely used semiconducting material. It is a brittle material, yielding catastrophically rather than deforming plastically [4]. However, Young's modulus of silicon is similar to stainless steel and is above that of quartz and most types of glass. It is the basic material for the electronic industry. This material may be produced with high purity and quality (containing very few structural defects).

The silicon is cleaned by zonal melting (removing many impurities). Single crystals of silicon are then mostly prepared by cooling the melt using the Czochralski method.

Polysilicon

Polycrystalline layers may be formed by vacuum deposition onto an oxidized silicon wafer with an oxide thickness of about 0.1 μm. Polysilicon structures may be doped with boron or other elements by ion implantation or other techniques to reach the required conductivity. Even if the boron concentration is very high, the resistivity of the polysilicon layers is always higher than that of a single-crystalline material. The resistance change of the polysilicon with temperature is not linear. The temperature coefficient of the resistance may be changed over a wide range, positive or negative, through selective doping. The temperature sensitivity and the resistance of undoped polysilicon is substantially higher than that of single-crystalline silicon. For some specific doping concentrations, the resistance may become insensitive to temperature variation.

Polysilicon resistors are capable of reaching as high a level of long-term stability as can be expected from resistors in single-crystalline silicon, since surface effects play only a secondary role in the device characteristics [4].

11.2.4. *Other semiconductors*

There is a wide range of compound semiconductors that combine atoms from columns III and V, II and VI or IV and VI of the periodic table. The importance of compound semiconductors is the possibility of combining semiconductors from the same family (for instance III/V) to prepare heterostructures with unique properties. Only two compound semiconductors are presented in this chapter: GaAs, the most important and widely used compound semiconductor and InSb because of its use in magnetic sensors.

Gallium arsenide (GaAs)

GaAs is a compound semiconductor combining group III and V elements from the same row as the traditional group IV semiconductor, germanium. GaAs has a density of 5,317.4 $kg.cm^{-3}$ at room temperature and crystallizes into the zinc blended structure. A shift in valence charge from gallium to arsenic atoms produces a mixed ionic/covalent bond compared to the covalent bond of germanium and silicon, which increases the melting point (1,260^{o}C) but decreases hardness.

The most significant attribute is the electronic band structure of GaAs, which determines the major electrical and optical properties. Firstly, the optical absorption and luminescence across the band gap do not require the participation of momentum conserving phonons. This means that efficient luminescence is achievable for GaAs unlike Si and Ge. Secondly, the effective mass of electrons is substantially lower for GaAs than for Si (compare $0.3m_0$ for Si to $0.067m_0$ for GaAs), so faster electronic devices are achievable in GaAs. Thirdly, since the forbidden energy gap for GaAs (1.42 eV) is higher than that for Si (1.08 eV) superior device isolation is potentially available for GaAs.

GaAs is dominantly used in heterostructures combining other ternary compound semiconductors with wider band gaps such as AlGaAs, or lower band gaps such as InGaAs.

Indium antimonide (InSb)

InSb is useful for magnetic sensing devices such as Hall effect sensors and magnetic resistors. InSb magnetoresistors are used as magnetic position sensors in automotive applications such as crankshaft and camshaft sensors for engine control. The sensitivity of magnetoresistors is proportional to the square of the electron

mobility. Thus, the very large room temperature electron mobility of InSb is an advantage for these sensors. The narrow energy gap (0.18 eV) makes the intrinsic electron density high. Since the device operating temperature may be 200ºC in some applications, InSb is normally n-type doped to stabilize the electron density. InSb is also used for infrared imaging.

11.2.5. Plastics

Plastics are synthetic materials. They are made from monomers which consist of one chemical unit. The long chains of repeating units (ethylene) form polymers (polyethylene). In the same way, for example, polystyrene is formed from styrene monomers. Polymers consist of carbon atoms in combination with only seven elements -hydrogen (H), nitrogen (N), oxygen (O), fluorine (F), silicon (Si), sulfur (S) and chlorine (Cl). The combinations of these elements create thousands of various plastics.

The combination of the atoms must correspond to the rules of joining them with other atoms. Each atom has a limited capacity of chemical bonds if the compound should be stable. Polymers are also used as detectors of radiation, chemical sensors and other sensing applications.

Thermoplastics

Heavier molecules are created by adding more carbon and hydrogen to a chain, the step increase being 14 (one carbon + two hydrogens). For example: ethane gas (C_2H_6) is heavier than methane gas (it contains an additional carbon and two additional hydrogens, and its molecular weight is 30). Pentane C_5H_{12} is too heavy to be a gas and it is a liquid at room temperature. Further additions of CH_2 groups make progressively heavier liquids until $C_{18}H_{38}$ – this is not a liquid but is a solid – paraffin wax. If we reach a molecular weight of 1,402 ($C_{100}H_{202}$), the material is tough and is called a low molecular weight polyethylene – the simplest of the thermoplastics. Further addition of CH_2 groups increases the toughness of the material and we get medium and high-molecular weight polyethylene [4]. Polyethylene – the simplest polymer- is reasonably transparent in the mid- and far infrared spectral ranges and therefore is used for fabrication of infrared windows and lenses.

The long chains are formed by heat, pressure and by using catalysts. This process is called polymerization. The chain length (molecular weight) determines many properties of a plastic – toughness, creep resistance, stress-crack resistance, melt temperature, melt viscosity, difficulty of processing, etc. These polymers are called thermoplastic polymers (heat-moldable).

If we pack the chains closer to one another we get denser polyethylene. These plastics have crystal structures. Crystallized areas are stiffer and stronger. These polymers are more difficult to process because they have higher and sharper melting temperatures. The crystalline thermoplastics abruptly transform into low-viscosity liquids, while amorphous thermoplastics soften gradually.

Examples:

– Amorphous polymers include polystyrene, polycarbonate, polysulfone, etc.

– Crystalline plastics include polyethylene, polypropylene, nylon, acetal, etc.

Thermosets

Thermosets are another type of plastic. The polymerization – curing – is performed in two steps: 1 - material manufacturing; 2 - molding.

Example: phenolic compounds are liquefied under pressure during the molding process and a cross-linking reaction between molecular chains take place. After it has been molded, a thermoset plastic has all its molecules interconnected with strong chemical bonds, which are not reversible by heating.

Thermoset plastics resist higher temperatures and provide greater dimensional stability. Thermoplastics offer higher impact strength than thermosets. They are also easier to be processed and allow more complex designs.

The useful thermoplastics in sensor-related applications are: alkyd, alkyl, epoxy, phenolic and polyester.

A *Copolymer* is a polymer formed in a polymerization reaction with two different monomers.

Plastics are electrical insulators, but often we require them to behave as conductors. In order to make them conductive we may either use lamination of the metal foil, metallization (e.g. for shielding purposes) or we can mix plastics with conductive additives (graphite, metal fibers).

Piezoelectric plastics are made from PVF_2 and PVDF (crystalline materials). Initially, they do not have piezoelectric properties and they must be processed using high voltages or by corona discharge. These plastic films are used in some applications instead of ceramics, because they have better flexibility, stability against mechanical stress and they can be formed into any desirable shape [4].

11.2.6. *Metals*

Ferromagnetic metals (steel, iron, manganese, nickel and some alloys) are used in magnetic sensors, which are described in Chapter 10. Ferromagnetic metals are also used for magnetic shielding. Non-ferromagnetic metals such as copper, aluminum and certain alloys such as some stainless steels have relative permeability close to 1.

When selecting a metal for the sensor design, we must take into the account not only the physical properties but also its mechanical processing. For example, copper has excellent thermal and electrical properties, but it is difficult to machine. An alternative compromise in this case is very often aluminum.

11.2.7. *Ceramics*

Ceramics are crystalline materials which are very useful in sensor fabrication. The main common properties are structural strength, thermal stability, low weight, resistance to many chemicals, ability to bond with other materials and excellent electrical insulating properties. Another advantage of ceramics is that they mostly do not react with oxygen and thus do not create oxides.

Several metal carbides and nitrides belong to ceramics. Boron carbides and nitrides and aluminum nitrides (which have excellent heat transfer) are most commonly used. Silicon carbide has a high dielectric constant, so that it is ideal for designing capacitive sensors. Ceramics are usually hard, therefore they require special processing techniques. Various shapes of ceramic substrates are fabricated by scribing, machining, and drilling with the use of computer-controlled CO_2 lasers.

Ceramics for the sensor substrates are available from many manufacturers in thicknesses ranging from 0.1–10 mm [4].

11.2.8. *Glass*

Glass is an amorphous solid material made by fusing usually silica with basic oxide. Although its atoms do not form a crystalline structure, its atomic arrangement is rather dense. Glass is a transparent material available in many colors. It is hard and resistant to most chemicals (except hydrofluoric acid). A lot of glasses are based on silicate and are composed of three major components – silica (SiO_2), lime ($CaCO_3$) and sodium carbonate (Na_2CO_3).

Non-silicate glasses include phosphate glass (resistant to hydrofluoric acid), heat absorbing glass (made with FeO), glass based on oxides of aluminum, vanadium, germanium and other types of metal. For example, borosilicate glass is massively resistant to thermal shocks due to its low thermal expansion and is used for the fabrication of optical mirrors. Lead-alkali glass (lead glass) contains lead monoxide (PbO), which increases the index of reflection and it is a better electrical insulator. It is used for the construction of optical windows, prisms and as a shield against nuclear radiation.

Light-sensitive glass forms another group. Photochromatic glass darkens when exposed to ultraviolet radiation and clears when the UV is removed and/or the glass is heated. The photochromatic material may keep its color (at room temperature) from a few minutes to a week or longer [4].

11.3. Silicon planar IC technology

Microsensor processing has similar requirements as the current microelectronic technology. The basic processing steps in silicon planar IC (integrated circuit) technology often form the basis of microsensor technology. Conventional silicon planar IC technologies are subsequently modified to include some additional processing steps.

The monolithic fabrication processes can be divided into two basic types known as bipolar and CMOS. MOS is one of the most common IC technologies presently used in microsensors [16].

The silicon planar IC fabrication generally involves all of the following processes:

– crystal growth and epitaxy;

– oxidation and film deposition;

– diffusion or implantation of dopants;

– lithography and etching;

– metallization and wire bonding;

– testing and encapsulation.

The existence of an oxide is very important: SiO_2, the preparation of which is simple, and which is suitable for lithography and possesses high electrical resistivity. Therefore, most ICs and microsensors are produced whenever possible using silicon rather than gallium arsenide or other semiconductors. For the majority

of IC structures it is necessary to grow thin epitaxial layers, which have better crystallographic quality in comparison with the bulk material [17].

11.3.1. *The substrate: crystal growth*

There are two main techniques for bulk silicon crystal growth: Czochralski crystal pulling and floating zone process. Czochralski growth is used for the growth of larger diameter crystals (300 mm diameter crystals are already industrially used) and for doped Si crystals.

The advantage of the float zone crystals is purity not only with respect to dopants but also as far as non-doping impurities such as carbon, oxygen, heavy metals and others are concerned. By applying multi-pass zone melting, the purity can even be enhanced. This method is not suitable for the growth of doped crystals. The largest diameter, which can be grown by this method, is about 100 mm, because of the stability problems arising from the melted zone under gravity conditions.

11.3.2. *Diffusion and ion implantation*

MOS transistors are generally made from conducting or semi-insulating silicon wafers with layers that have been doped with n- or p-type materials [11]. Controlled amounts of dopants are inserted into the wafer by thermal diffusion, ion implantation or during epitaxial growth.

The procedure of thermal diffusion of n-type materials is following. The wafers are placed in a furnace and an inert gas containing the required dopant (e.g. AsH_3 or PH_3) is passed over them. The p-type diffusion can be achieved by passing an inert gas carrying, for instance, B_2H_6. (Note: AsH_3 and PH_3 are very toxic, in fact AsH_3 is the most toxic gas ever used in planar technology and they are typically replaced with less harmful ways of placing the same element into the silicon substrate. Arsenic is implanted from a solid source and phosphorus is doped in a furnace using $POCl_3$. B is also doped using a solid source.)

An alternative method to thermal doping is ion implantation. The charged ions of the desired dopant are accelerated to energies in the range of 10 to 1,000 keV and are fired at the surface. The technique is now commonly used by penetrating As, P and B to a depth of 0.5, 1 and 2 μm at 1,000 keV in silicon.

Ion implantation at 10 keV gives very little yield on the ion source so this energy is used only very rarely, minimum reasonable energy is about 40–50 keV. On the opposite side of the spectrum, typical implantation does not allow energy higher

than 200 keV. Besides the elements listed, BF^{2+} is also very common. After ion implantation the material should be annealed.

Doped layers can be prepared directly by epitaxial growth. By this way it is possible to dope layers homogenously or with a defined doping profile and at the exact doping levels.

11.3.3. *Oxidation*

The oxide layer is formed by placing the wafers into a furnace containing oxygen at 1,100°C. Oxygen reacts with silicon and diffuses through the growing SiO_2 layer [11].

The oxide films, which are used to prevent re-doping of areas with different types of doping materials, also form an electrically insulated region on the semiconductor device and are used for surface passivation.

11.3.4. *Lithography and etching*

Lithography is an image transfer process of a geometric pattern from a mask onto a thin layer of material, called a resist. It is used in traditional planar processes, but it also is the principal mechanism for pattern definition in micromachining. The name resist stands for a radiation-sensitive material. Firstly, a resist is usually either spin coated or sprayed onto the silicon wafer. The next procedure is to place the mask onto the resist correctly. Secondly, in optical lithography, ultraviolet (UV) radiation is used to change the solubility of the photo resist in a given solvent. The positive photo resist becomes more soluble after exposure to the UV light. The negative photo resists become less soluble due to a polymerization process.

Washing in an organic solvent (typically 3.9% solution of tetra methyl ammonium hydroxide) dissolves the uncured photo resists. The exposed SiO_2 is then etched away by HF solution or “buffered HF” (or BOE) which is a mixture of NH4H with HF. The remaining polymerized resist is then burned off. (The next procedure in processing the MOSFET transistor is the formation of the gate by thermal oxidation – the thickness of the layer is typically 0.1 μm. This procedure does not belong to lithography and etching.)

A few years ago the minimum line-width or resolution in optical lithography ($\lambda \approx 0.4$ μm) was determined by shadow printing or projection printing. This printing typically has a resolution of 1–5 μm. This resolution could be improved up to

0.3–0.5 μm through the use of stronger UV light and better optics. Sub-micron resolutions may be achieved by using of electrons, X-rays or ion beams [11]. The I-line stepper (UV at 365 nm) for 0.8 μm can print up to 0.4 μm. The high NA I-line stepper scan goes even below that.

Nowadays, ArF and F_2 sources emitting at 248 nm and 193 nm respectively are used. Off-axis optics can give a resolution under 200 nm.

A photoresist may also be used as a template for patterning material deposited after lithography (Figure 11.19b). The resist is subsequently etched away, and the material deposited on the resist is "lifted off".

11.3.5. *Deposition of materials*

The forming of the gate electrode is a typical example of deposition. The silane is pyrolyzed to produce the polysilicon layer. The polysilicon layer can be doped by the use of dopant gases during its formation or by diffusion or ion implantation. The higher reliability of the polysilicon as compared to the aluminum is the reason it is used in the construction of the gate. The polysilicon gate serves as a mask against source and drain implant.

The technology of deposition is described in more detail in section 11.4.

11.3.6. *Metallization and wire bonding*

At the end of processing the MOSFET, another oxide layer is formed and the window for metallization is exposed to lithography. The metal is then deposited by either physical vapor deposition – evaporation, chemical vapor deposition, or sputtering (≈1 μm). The metal, mostly Al or Au, forms the ohmic contacts to the source, drain and gate electrode.

The final wafer is then diced up. A saw or diamond scribing is used and the IC is mounted into the package. The ultrasonic welding of thin Al or Au wire or ribbons makes up the electrical connections between the pads (ohmic contacts) and package terminals. An interesting fact is that the crucial influence on the reliability of the whole IC depends on the wire-bonding [11].

11.3.7. *Passivation and encapsulation*

The final IC chip must be protected from the atmosphere. The sensing areas are often covered up by the photo resist or by a silicon nitride layer. Silicon nitrate can be deposited by LPCVD or CVD and acts as a firm barrier against water.

The film thickness for IC is usually limited to less than 0.2 μm, because thicker layers cause thermally-induced stresses.

The last step is the encapsulation of the IC in the following manner, e.g., sealed in a plastic resin or hermetically sealed in a metal case. This process is highly desirable, because it protects the silicon device from the surrounding environment. In the case of MEMS, isolating the silicon structure from the atmosphere it is not always required. The atmosphere could transmit the measured quantity [11].

11.4. Deposition technologies

11.4.1. *Introduction*

One of the basic steps in MEMS processing is to deposit thin and thick films of material, which provides the sensing surface with the required properties. For example, sensitivity to thermal radiation is given by coating with nichrome. Thick films are used to construct pressure sensors [18] or microphones, where the membranes have to be produced. The following processes allow the fabrication of films to have a thickness anywhere between a few nanometers and about 100 micrometers. The film can be locally etched using lithography and wet chemical etching processes. Dry physical etching and laser processing can also be used [19].

Depositions processes are:

1. CHEMICAL deposition
   - chemical vapor deposition (CVD)
   - epitaxy : liquid, vapor, molecular (rare in MEMS)
   - electrodeposition
   - thermal oxidation (see section 11.3.3)

These processes are based on the creation of solid materials directly by chemical reactions with gas and/or liquid compounds or from the substrate material.

2. PHYSICAL deposition
   - physical vapor deposition (PVD): evaporation or sputtering

- casting
- spray coating, screen printing
- laser ablation

In these processes, the deposited material is physically placed onto the substrate, with no traditional chemical reaction, which forms the material on the substrate [3], [4], [17].

11.4.2. *Chemical reactions*

Chemical vapor deposition (CVD)

The substrate is placed inside the reactor, into which a number of gases are introduced. The basic principle is that a chemical reaction takes place between the source gases. This reaction creates a solid material, which is deposited on free surfaces inside the reactor.

The CVD system is used to process thin films with good uniformity. This technology allows a variety of materials to be deposited, although some of them are less popular, because of hazardous by-products formed during the process. This is, however, the technology which is preferred by industry nowadays.

One of the simplified forms of CVD process is illustrated in Figure 11.1. The substrates or wafers are positioned on a stationary or rotating table whose temperature is elevated up to the required level by heating elements. There are three reasons for this: a) oxides are decomposed and evaporated from the wafer surface, b) the surface is smoothed, often at the atomic scale, and c) the source gases are thermally decomposed, which is necessary for the layer growth. The top cover of the chamber has an inlet for the carrier gas, which can be added with various precursors and dopants. These additives, while being carried over the heated surface of the substrate, form a layer. The gas mixture flows from the distribution cone over the top surface of the wafers and exits through the exhaust gas outlets.

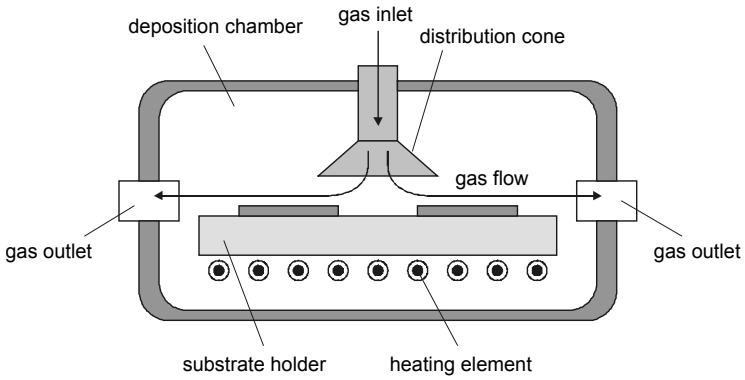


Figure 11.1. *Simplified structure of a reactor chamber (according to [4])*

CVD epitaxy

Epitaxial growth is used not only for layer deposition, but also because it is a way of preparing heterostructures of the highest quality from different materials. This technology is a special type of CVD process. It consists of depositing atoms of the desired material onto the substrate (e.g. semiconductor crystalsilicon, gallium arsenide) with the same crystallographic characteristic as the substrate. This concerns mainly crystallographic structures and orientation of the axes. In particular it is possible to produce almost perfect single-crystalline films on single-crystalline substrates, if the lattice constants of the two materials are very close to each other. Film grown on a polycrystalline or amorphous substrate will also be amorphous or polycrystalline.

The most important epitaxial growth is vapor phase epitaxy (VPE). In this process, a number of gases are introduced into an induction-heated reactor where only the substrate is heated. The temperature of the substrate typically must be high because of oxide decomposition and evaporation, smoothing of the surface at an atomic level as well as for the decomposition of precursors.

This technology is primarily used for deposition of silicon and is widely used for producing silicon-on-insulator (SOI) substrates. The advantage of epitaxy is the high growth rate of material – it allows the formation of films with a thickness ranging from $\approx 1\,\mu m$ to $> 100\,\mu m$. Some processes require high temperature exposure of the substrate, others do not demand significant heating of the substrate. Some processes can be used to perform selective deposition, depending on the surface of the substrate [17]. The typical VPE reactor is shown in the schema in Figure 11.2.

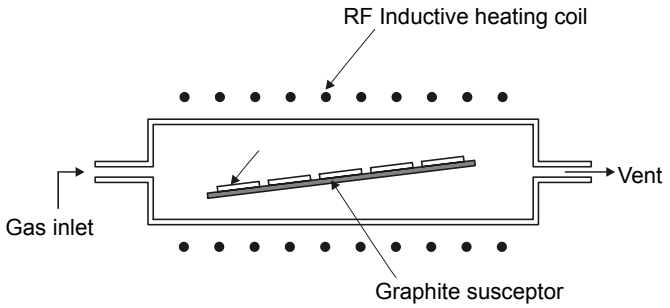


Figure 11.2. *Typical "cold-wall" vapor phase epitaxial reactor (after [3])*

Electrodeposition (electroplating)

This process is restricted to electrically conductive materials. The process is used to make films of metals such as copper, gold and nickel (the thickness ≈ 1 μm to 100 μm). The deposition is best controlled when used with an external electrical potential, but it requires electrical contact to the substrate. The typical set-up for electroplating is shown in Figure 11.3.

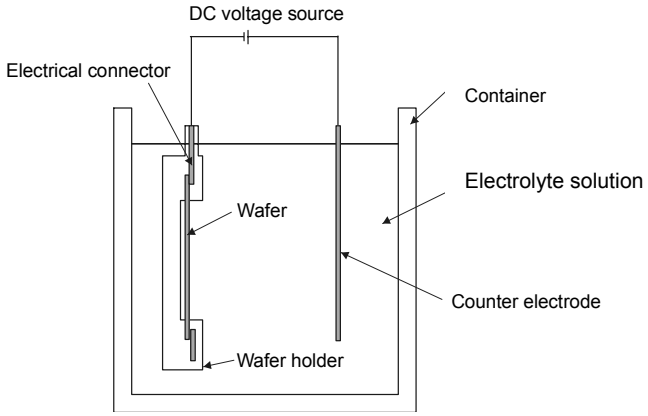


Figure 11.3. *Typical set-up for electrodeposition (after [3])*

The substrate is placed in a liquid solution (electrolyte) and an electrical potential is applied between a conducting area on the substrate and a counter electrode (usually platinum) in the liquid. The chemical process forms a layer of material on the substrate. Various types of gases are very often generated at the counter electrode [20].

Electroless plating does not require any external electrical potential and contact with the substrate during processing. These processes use chemical solutions in which deposition takes place spontaneously on any surface. The disadvantage of this fabrication is that it is more difficult to control the film thickness and uniformity [3].

11.4.3. *Physical reactions*

Physical vapor deposition (PVD)

PVD comprises technologies for deposition of metallic as well as dielectric films. It is more common than CVD for deposition of metals (lower process risk, cheaper), even if the quality of the films is inferior – higher resistivity for metals, more defects and traps for insulators. The choice of deposition method is in many cases arbitrary and depends on which technology is available for the specific material at a given moment. Two main techniques for PVD are evaporation and sputtering.

Evaporation

The main principle of this PVD technique is that metal can be converted into gaseous form and then deposited on the surface of the sample. The substrate is placed inside the vacuum chamber (usually 10^{-6} to 10^{-7} Torr). The block (source) of the material to be deposited is also located in the chamber. It is heated so that it evaporates. For some materials (Cr, Ti) sublimation temperatures are lower than the melting temperatures. The vacuum is required to allow the molecules to evaporate and move freely in the chamber. They subsequently condense on all surfaces. All evaporation technologies use this principle but the methods differ in the way the source material is heated and evaporated.

The two most popular evaporation technologies are e-beam and resistive evaporation. In e-beam evaporation, an electron beam is focused on the surface of the source material and causes local heating and subsequent evaporation. In resistive evaporation, a tungsten boat, containing the source material, is heated electrically.

The film thickness is determined by the evaporation time and the vapor pressure of the metal. The evaporation, just like all vacuum deposition processes, produces layers with large residual stresses and therefore these techniques are mostly used for depositing thin layers.

Very often it is important to keep the substrate at an elevated temperature, in order to evaporate moisture from its surface as well to remove some oxides or other surface impurities.

The resistive evaporation is shown in Figure 11.4.

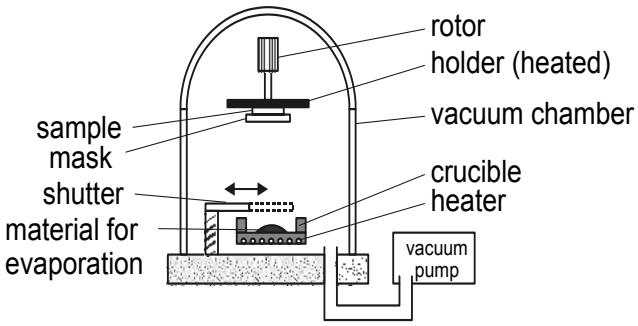


Figure 11.4. *The evaporation – deposition of thin metal film in a vacuum chamber (after [4])*

Sputtering

The source material in this PVD technology is subjected to a lower temperature compared to evaporation. The substrate is placed in a vacuum chamber (about 2.10^{-6} to 5.10^{-6} Torr) with the source material (denoted as the target or the cathode) and an inert gas (e.g. argon, helium) of low pressure. A gas plasma is ignited using an AC or DC high voltage power source. The gas becomes ionized. The target-cathode is connected to this voltage. The sample wafer is attached to the anode at some distance from the cathode. In some cases, when the non-conductive substrate is used, the wafer need not be connected to the electrode and it is sufficient to put the wafer between the anode and cathode. The ions are accelerated against the target. The kinetic energy of the bombarding ions is sufficiently high to free some atoms from the target surface. The source material, now in a vapor form, condenses on all surfaces including the substrate [3]. This principle of sputtering is common for all sputtering technologies. The differences are typically in the method of ion bombardment of the target [4].

The advantage of this technology is better uniformity, especially if a magnetic field is introduced into the chamber. The field allows improved flow of atoms towards the anode. Since this method does not require a high target temperature, theoretically any material, including an organic material, can be sputtered. The process of sputtering can be extended to the sputtering of more than one target at the same time (co-sputtering,), for example, sputtering nichrome (Ni and Cr) [17].

The sputtering process is shown in the schematic Figure 11.5.

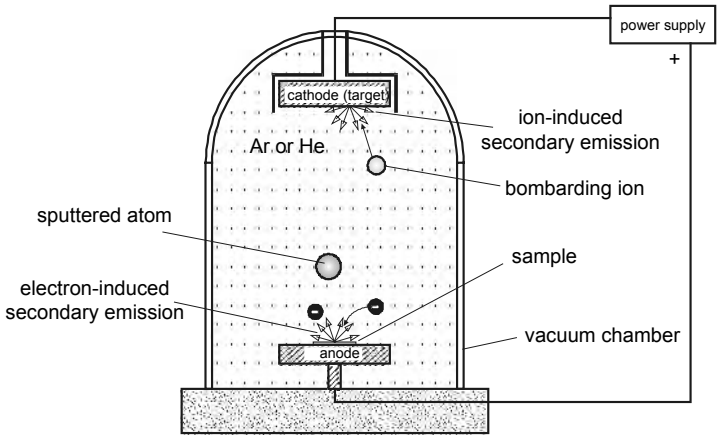


Figure 11.5. *The sputtering process in a vacuum chamber (after [4])*

Casting

In the casting process, the material to be deposited is dissolved in a volatile liquid solvent. When the solvent is evaporated, a thin layer of the material remains on the substrate [3]. The differences in this process are in the way the material is transported onto the substrate. The most widely used are transmission by spraying and spinning.

The thickness of the casting layer on the substrate depends on the solubility of the deposited material and can be in the range from a single monolayer of molecules/atoms (adhesion promotion) to tens of micrometers (0.1 to 50 μm). The control of the film thickness depends on exact conditions, but the thickness can be uniform within +/-10% in a wide range [4].

This process is often used for polymer, polyimide and other organic materials. The casting method is also used for transferring the photoresists to the substrate in the photolithography process and is an integral part of the photolithographic technique [17]. This technique is often used for fabrication of humidity and chemical sensors.

Spray coating

Thermal spray coating is used for metal deposition. The coating material is fed into the flame where it melts. The melted metal is atomized by a high-velocity stream of air or other gas. When the stream reaches the target, atoms are bonded to

the surface. This technology may replace traditional plating which often has serious pollution problems as the electrolytes contain toxic chemicals such as cyanide.

Low-temperature spray coating is used to deposit paints. One of the applications is in the production of thermal radiation sensors. The deposition layer (the surface of the sensor) is processed by covering it with a coating having a high infrared emissivity. The coating must be very thin to have a good thermal conductivity and a very small thermal capacity. However, the available organic materials have low thermal conductivity and cannot be effectively deposited with thicknesses less than 10 μm. This characteristic influences the sensor response [17].

Screen printing

This process has been used for many years as a cheap way to make hybrid circuits in electronics. The simplified explanation of the technique consists of the preparation of an ink paste using suitable organic solvents. The paste is then squeezed through a fine gauze mask and forms a 25 to 100 μm film in the desired areas. The film is then dried by heat treatment to form a conductive layer. The lateral resolution is only about 100 μm but the printing cost makes this technique commercially viable for low volume low-cost electronic circuits [11].

For example, platinum electrodes have been printed and used in electrochemical and bioelectrochemical sensors.

Laser ablation

Laser ablation in general is removing material from the surface using a laser beam. Localized heating causes the material to evaporate. This technology is also used for film deposition: in this case the target is heated by a pulsed laser and the substrate is positioned in the ablation plasma plume. This processed is made in vacuum or in the low pressure background gases. Laser ablation makes it possible to deposit complex materials (including organic) from the target to the substrate. High deposition rates are achievable, however the thickness is usually not perfectly even.

11.4.4. *Epitaxial techniques for semiconductor device preparation*

Crystallographic perfection (especially defect density) of *bulk single-crystals* makes them unsuitable for optoelectronics.

Epitaxy (from Greek epi-taxis – “arranged on”) is a coordinated crystalline growth on (usually) single-crystal wafer, up to crystal lattice mismatches of 15%.

Advantages of epitaxial growth are crystal quality, layer thickness, heterostructure preparation, composition and doping profile creation.

The main types of epitaxy are SPE (solid phase epitaxy); LPE (liquid phase epitaxy), LPEE (liquid phase electroepitaxy); VPE (vapor phase epitaxy) – CVD (chemical vapor deposition), PVD (physical vapor deposition), MBE (molecular beam epitaxy), solid-source MBE, CBE (chemical-beam-epitaxy), GS-MBE (gas-source MBE - hydride-source, halide-source), photo-enhanced MBE, plasma-enhanced MBE, MOMBE (metal-organic MBE), UHV ALE ultra-high-vacuum atomic-layer-epitaxy, MOVPE (metal-organic vapor phase epitaxy), MOCVD (metal-organic chemical vapor deposition), photo-enhanced MOVPE, plasma-enhanced MOVPE:

– *Principle of epitaxial growth:* atoms or molecules of the material which we would like to deposit on reasonable substrate are transported to its surface, which has to be "atomically" clean and flat (or with defined atomic steps). On the surface these atoms are "physically-sorbed" and consequently "chemically-sorbed" at places with minimum energy, and thus subsequent epitaxial layers and structures are grown.

– *Solid phase epitaxy* is an old method with newly found applications.

– *Liquid phase epitaxy* was the most important method during the 1970s and 1980s. It is still an important industrial method (simple cheap LEDs, solar cells, and for thick layer structures). LPE is good for the preparation of complex compounds at thermodynamic equilibrium and for doping by rare earth elements.

– *Vapor phase epitaxy* is fundamental for the preparation of optoelectronic devices. VPE is the basic technology today, not only for research but also in industry, and will remain so for at least for the next 10 years.

There are two principal branches of VPE depending on the type of agent used to transport growing material from source to substrate – PVD and CVD. PVD uses material evaporation (vacuum evaporation, sputtering, laser ablation, discharge, etc.) without chemical reactions. CVD uses gaseous chemical compounds (precursors) for material transport to the substrate where they are usually thermally decomposed. Layer growth is similar to PVD as well as structure parameters.

Epitaxial techniques for optoelectronics (and nanotechnology)

Basic methods for the preparation of optoelectronic devices and nanostructures are MBE and MOVPE. MBE is mainly used for research but also industrial technology, while MOVPE is the most important industrial technology but also very suitable for research.

Principle of MBE: the substrate (usually single-crystal semiconductor wafer, diameter is from 2 to 8 inches, 300–500 μm thick) is heated in ultrahigh vacuum

(10-(9-11) torr) chamber up to a temperature high enough that native oxides are desorbed and the surface is atomically clean and flat. Atoms (molecules) from preheated Knudsen cells propagate without collisions tens of centimeters through the growing chamber after opening the shutter, and shower impinge on the substrate surface (and evaporate its vicinity). Atoms of future epitaxial layer are captured on the surface (physi-sorption), migrate on it and after some time they are fixed at proper crystallographic positions (chemi-sorption). The principle is shown in Figure 11.4.

Principle of MOVPE: the substrate is heated in ultra-clean gas (hydrogen, nitrogen, purity better than ppb) up to a temperature high enough so that native oxides are desorbed and the surface is atomically clean and flat. Precursors (organometals, hydrides) with desired atoms are flushed over the preheated substrate, where they are thermally decomposed. Atoms of the future epitaxial layer are captured on the surface. The reactor scheme is shown in Figures 11.1 and 11.2.

The basic sum equation for growth of GaAs from trimethylgallium (TMGa) and arsine (AsH_3) is:

$$Ga(CH_3)_3 + AsH_3 \rightarrow GaAs + 3CH_4$$

11.5. Etching processes

Thin films (deposition methods have been described in the previous chapter) may be shaped by using masks during the deposition process. However, the achievable complexity of shapes and accuracy is limited. In order to form a functional MEMS structure it is necessary to selectively remove parts of the film. This is caused by surface micromachining. Compared to that, bulk micromachining only removes the material from the wafer. The websites are en excellent sources of information on micromachining technologies of the key manufacturers such as [21-28].

The most popular technique for surface and bulk micromachining is etching.

There are actually two main groups of etching processes:

– *Wet etching (micromachining)*, where the material not required is dissolved, when the wafer is put into a chemical solution.

– *Dry etching (dry micromachining)*, where the material is sputtered or dissolved by reactive ions or a vapor phase etchant.

11.5.1. *Wet etching/micromachining*

This simple etching method uses liquid etchant to dissolve the material. A mask should be made of material which will not dissolve during the process time.

The first etch solutions developed provided *isotropic etching*, the etch rate being independent of crystal orientation. Generally, isotropic etchants consist of a mixture of nitric, hydrofluoric and acetic acids.

Some single crystal materials (such as silicon) exhibit *anisotropic etching* in certain chemicals. The anisotropic etching means that the etching rates are different in different (crystallographic) directions. An example could be that the <111> crystal plane side walls appear when etching a hole into a silicon wafer with an <100> exposed plane in a chemical such as KOH [21]. The result is a pyramid-shaped hole in contrast to a hole with rounded side walls when using an isotropic etchant. This principle is illustrated in Figure 11.6.

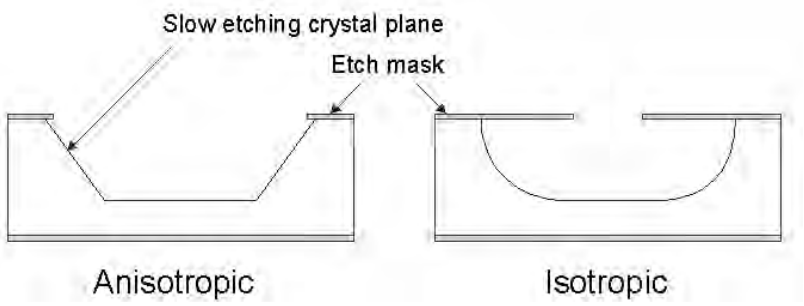


Figure 11.6. *Schematic results of anisotropic and isotropic wet etching (after [3])*

Anisotropic etch-stops

In the construction of microstructures it is often necessary to ensure that the etching stops at a predetermined position in order to control the shape of the structure. The main idea of the stops is that etch rates of alkaline anisotropic etchants strongly depend on doping of silicon by boron. The disadvantage is that a high level of boron doping is needed, which produces a residual tensile strain in silicon. This limitation can be overcome by using of electro-chemical etch stops. An anodic contact is made to the n-Si. The potential is held at a level higher than the passivation potential (0.6 V). The p-Si is then chemically etched until the n-Si layer is reached. At this point the cell current increases and the etching process is stopped. This process can be controlled to 1 μm accuracy [11].

Anisotropic wet etching is one of the main subtractive bulk microfabrication (micromachining) techniques. If the orientation of the substrate plane is the limitation of the application, more complicated and expensive dry etching should be used [17].

11.5.2. *Dry etching/micromachining*

Dry etching can be in general more precisely controlled than wet etching techniques. The disadvantage is the process requires more complex instrumentation including a vacuum chamber.

The methods of dry etching include:

– reactive ion etching (RIE)

– sputter etching

– vapor phase etching

In *RIE*, the substrate is placed inside a reactor chamber filled by several gases. The plasma is ignited in the gas mixture by an RF power source. In the plasma the gas molecules are dissociated into ions. The ions are accelerated towards and react at the surface of the material we want to etch. The output of the reaction is a formation of another gaseous substance. This is the chemical part of reactive ion etching and of course there is also a physical part. If the ions have sufficiently high energy, they can knock atoms out of the material to be etched without a chemical reaction. The etching by chemical reaction is very often isotropic and the physical reaction may be anisotropic. By changing the balance between these two types of etching, we can form side walls that have shapes from rounded to vertical. Figure 10.7 shows the typical RIE system [22].

The deep RIE (DRIE) is a special subclass of RIE. The side walls obtained with DRIE are almost vertical to a depth of hundreds of microns. In this process, two different gas compositions are alternated in the working reactor. The first gas composition creates a polymer on the surface of the substrate, and the second gas composition etches the substrate. The polymer is immediately sputtered away by the physical part of the etching. Only the horizontal surface is etched, not the side walls. The polymer is dissolved very slowly in the chemical part of the etching and therefore this chemical etching builds up the side walls. As a result, the etching aspect ratio of 50: 1 can be achieved. This process easily allows etching completely through a silicon substrate. The etching rates are 3–4 times higher than with wet etching.

An easier method of dry etching other than RIE is *sputter etching* (RIE without reactive ions). The substrate is subjected to the ion bombardment to remove material. Maximum etch rates are of the order of μm/min over a wafer-sized area [11]. The disadvantage of this technique is the lack of the etch stop layer.

The next easiest method of dry etching other than RIE is *Vapor phase etching*. The material to be etched is dissociated at the surface in a chemical reaction with the gas molecules. The two most popular processes are silicon dioxide etching using hydrogen fluoride and silicon etching using xenon difluoride. Both methods are isotropic dry etching processes.

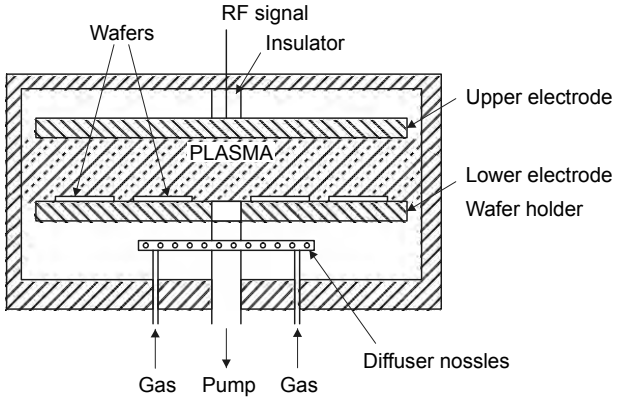


Figure 11.7. *Typical scheme of the parallel-plate reactive ion etching system (after [3])*

The technology of dry etching is more expensive compared to wet etching. If fine resolution in the thin film structures is required or in the construction of MEMS where deep etching in the substrate with vertical side walls is present, dry etching should be used.

11.6. 3-D microfabrication techniques

Conventional silicon processes allow two-dimensional structures to be built, as the deposition thickness is limited to about 10 μm. The 3-D processes extend this limitation to thicker structures, i.e. fabrication of MEMS to fully three-dimensional structures.

11.6.1. *LIGA*

The term LIGA originated in Germany and stands for Lithography, Galvanoformung, Abformung (lithography, electroforming, molding). LIGA allows high aspect ratio structures to by built, i.e. very high and narrow shapes.

The LIGA process begins by generating a photoresist pattern by short-wavelength X-ray lithography on a conductive substrate. A heavy metal is used for the mask and a type of polymethyl methacrylate is used for the resist.

After removing the exposed resist, spaces between the resist are electroplated. The created metal shape can be directly used, but in the LIGA process it is used as a tool for plastics molding. After curing, the mold is removed, leaving behind microreplicas of the original pattern. A lot of different materials are compatible with this process, e.g. a number of polymers, some metals and even some ceramic materials. This technology allows a large numbers of low-cost microstructures to be created. The disadvantage is the need for a short-wavelength, highly collimated X-ray source, typically a synchrotron orbital radiation (SOR) instrument.

The SLIGA (sacrificial layered LIGA) technique gains another degree of design freedom by combining LIGA with sacrificial layers. It is useful for making small gear linkages, or other released parts that can be assembled on a separate LIGA structure or used in more traditional products.

The LIGA process can be combined with the silicon process and this allows a novel application for the fabrication of micro turbines, gear trains and three-dimensional structures – filters, fluid-logic and conduits [29], [17].

11.6.2. Laser assisted etching (LAE)

This process belongs to the group of contactless subtracting methods of fabrication [30]. Electron beam or laser light could be used for the fabrication of the MEMS. If we scan these beams in three dimensions, we can process complicated three-dimensional microparts. The precision of the process is mostly decided by the wavelength of the beam. Photolithography or electron beam lithography use beams with a shorter wavelength. LAE is one of the methods using laser light.

LAE uses a photo-reactive or a thermal reactive process. The etching of the appropriate material (semi-conducting materials, metals, ceramics, high polymers etc.) in an etchant is obtained using a laser beam. LAE does not use patterning masks.

The disadvantage of the precision of traditional laser fabrication is the strong evaporation or melting caused by the high temperatures involved. LAE works at lower temperatures and is based on laser irradiation in a liquid with etchant (or in gas) which gives rise to reactive ions or plasma which change the material into a soluble or vaporized form. The output power of the laser used in this method is mostly very low. The laser is used only for activating an opto- or thermo-chemical reaction and not for abrasion, burning or evaporation off at a high temperature [29].

Proton beam micromachining (PBM) is a novel technique for the production of high aspect ratio three-dimensional microcomponents. PBM is a direct write process in which a focused beam of MeV protons is scanned in a pre-determined pattern over a suitable resist material and the latent image formed is subsequently chemically developed [31].

11.6.3. *Photo-forming and stereo lithography*

Photo-forming is a universal optical forming process. This technique includes plain lithography and stereo lithography.

The plain lithography technique uses a layer of photoresist. This process is often used in metal etching and silicon planar process. Even if the method is simple, it can be easily used to fabricate three-dimensional structures by multiple layers of photoresist. Exposition of the resist can be made with the use of a photo mask or a micro-optical fiber for irradiation [29].

Stereo lithography

This technique is based on scanning the ultraviolet laser beam on photo-polymerizing solution. The photopolymer quickly solidifies when the laser hits its surface. Once the complete layer is built, the support plate is slightly lowered and another layer is made. The result is a three-dimensional structure.

The process is illustrated in Figure 11.8.

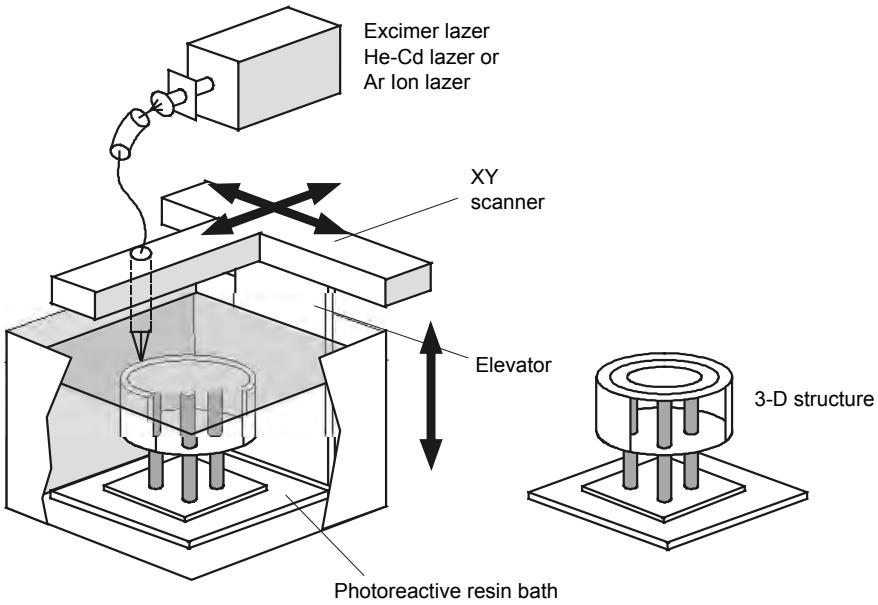


Figure 11.8. *The Macro-Photo Forming process (after [29])*

This technique is used for rapid prototyping of large parts: in this case the layer thickness is typically 0.2 mm and the production time is several hours.

In the microscale the model is fabricated from top to bottom or inversely. The microstereolithography process allows objects with a size of 5 μm to be built with an achievable resolution of 1 μm. The bending pipes, cords, micro-coil springs, combs, microturbine, etc., have been made using this fabrication method [29].

11.6.4. *Microelectrodischarging (MEDM and WEDG)*

Conventional precision machining is needed to construct fine units using metals, ceramics, alloys, and bulk silicon. The fine holes, gears and turbines of micro-size have been produced by microelectric discharge machining (MEDM) and wire electrodischarge grinding (WEDG). The principle of the three axis NC (numeric control) microelectrodischarge machining is described in Figure 11.9.

The discharge energy in a micro domain must be reduced to a low value 10-7 J. The resulting capacitance (C) should be reduced to less than 10 pF. The machine

has been constructed to minimize the use of metallic mechanical components to reduce the stray capacitance [31].

This method allows fabrication of electrode work pieces with complicated shapes, holes to a depth of about one tenth of the wafer. WEDG was used to make a fine electrode 4.3 μm in diameter and 50 μm in length. For example, a three-axis numeric control machining MEDM system was used to fabricate a micro-air turbine which was inserted into a metal catheter of external diameter 2.2 μm and rotated at 1,000 rpm (Matsushita Research Institute Inc., Tokyo).

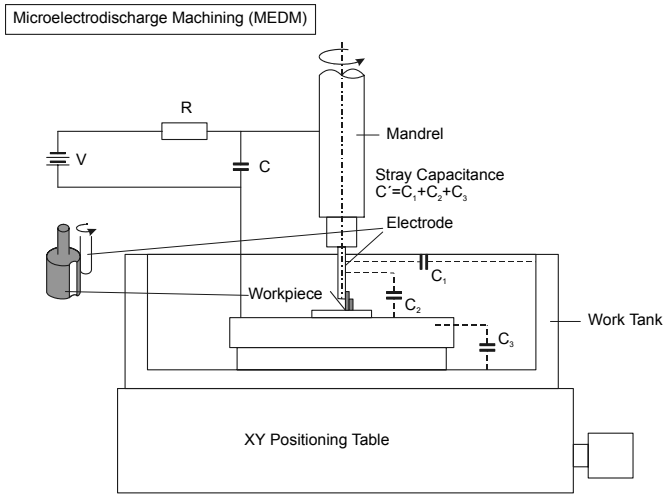


Figure 11.9. *The MEDM process (after [29])*

11.6.5. *Microdrip fabrication*

The main idea behind the process is to build micro objects from micro droplets. It is the same technology that has been used for thousands of years in nature by wasps, bees, and termites to construct their homes.

Wax (heated to around 90°C to melt it) was used at the beginning of the development of this technology. The wax was ejected through an adapted ink-jet head, and the droplets had a diameter of around 50 μm. The next step was to use the materials at higher melting temperatures. Some newly developed systems use photopolymers which are cured by UV light. This process is low-temperature and allows similar resolution and precision as stereo lithography.

11.6.6. *Manufacturing using scanning probe microscopes and electron microscopes*

This method allows fabrication of three-dimensional microstructures. The main instrument for this method is the scanning electron microscope (SEM) with the vacuum chamber used as a working table. A lithographic etching instrument called a multi-face fast atom beam (FAB), robot hands with four rotational and three translational degrees of freedom and many mechanical tools such as diggers, tweezers, blow pipes, scrapers, and sticking tools are placed on the table.

The University of Tokyo was one of the first laboratories to develop this complicated instrument. Figure 11.10 shows tools for microscope manufacturing. [29].

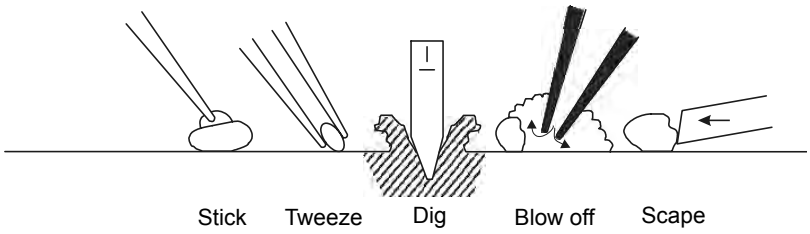


Figure 11.10. *Various tools of microscope manufacturing (after [29])*

11.6.7. *Handling of micro particles with laser tweezers*

The tiny forces are generated by the absorption, refraction or reflection of light by a dielectric material. Several milliwatts of power produced by a strong laser light generates a force of only a few piconewtons. This value of force is sufficient in the microscopic domain. We must be aware of the fact that radiation pressure could play a significant role in the handling of micro parts.

Engineers of AT&T Laboratories [28] proposed a laser-based optical trap for microscopic particles in 1980. The living material (viruses, bacteria, yeast, and protozoa) can be non-invasively manipulated with this device. The invention makes use of the so-called “gradient force” that appears in a light gradient when a transparent material with a refractive index different from that of the surrounding medium is placed in it. The same principle is used in optical tweezers, micromanipulators, tensiometers, etc.

Another principle: the fluctuating dipoles are induced, when the light passes through polarizable material such as dielectrics. These dipoles interact with the

electromagnetic field gradient and produce the force which directs microparticles towards brighter regions. When we direct parallel laser light onto a microsphere from above, the light is bent because the sphere acts as a lens. If the intensity profile of the incoming beam is uniform, the reaction forces on the left and right hand sides of the sphere cancel each other out and there is no net sideways component. When the light field is not uniform (gradient), an imbalance in the reaction forces is created and the object is pulled towards the bright side. These forces have been used in optical tweezers. Sharply focused light is required for these types of tweezers. Focusing a laser through a microscope objective can achieve sharp focus. A diode laser in the near-infrared region (780–950 nm) makes these devices practical for low-power use. Laser tweezers are relatively free of creep, backlash and hysteresis.

Laser tweezers are used in the assembly of micro-particles and to measure the dynamics of movement of micro mechanical particles [29].

11.6.8. *Atomic manipulation*

The finest possible tool for device assembly is provided by scanning probe microscopes. The first model, the scanning tunneling microscope (STM) invented by G. Binnig and H. Rohrer in 1981, gave us a tool capable of imaging and manipulating single atoms. Eye-catching images assembled by moving atoms by the STM tip appeared soon after and caught the imagination of the general public.

Since the seminal discovery of the STM, the field of nanotechnology has expanded into many directions:

– A large number of probe microscopes were developed which use various interactions between the sharp scanning tip and a specimen: atomic force microscopy (AFM), electric field microscopy (EFM), magnetic force (MFM), near-field scanning optical microscopy (NSOM), etc.

– Advanced probe designs use local heating, additional gate electrodes and sensitization by attached biological molecules.

– Multiple probes and/or probe arrays are investigated: arrays of probes in so called “millipede” or “nanodrive” project are used to store and read information with much higher density than the present hard drives, possibly reaching the atomic scale memory.

– Special molecules are designed and synthesized which could be used as building blocks for constructing functional nanostructures. One example is the “lander” molecule so named because of its similarity to the lander modules used in space exploration.

A probe with an additional electrode next to the tunneling tip of the STM can be used as a gate for the nanometer-sized field effect study of metallic nanoclusters.

More information on nanotechnologies can be found in [33–35].

11.7. References

[1] FUKUDA, T., MENZ, W.: Micro Mechanical Systems – Principle and Technology, Elsevier, 3rd ed., 2002.

[2] MALUF, N.: *An Introduction to Microelectromechanical Systems Engineering*, Artech House, 2000.

[3] MEMSnet, Corporation for National Research Initiatives, Reston, Virginia http://www.memsnet.org.

[4] FRADEN, JACOB: *Modern Sensors Handbook – Physics, Designs and Applications*, American Institute of Physics, NY, 1996.

[5] MEMSnet information service for the MEMS development community www.memsnet.org.

[6] European Union initiative to improve the competitiveness of European industry, http://www.europractice.com/technologies/microsystems/index.asp.

[7] BANKS, D.: "Introduction to Microengineering", "Microsystems, Microsensors & Microactuators", http://www.dbanks.demon.co.uk/ueng.

[8] Sze, S.M. (ed.): *VLSI technology*, McGraw-Hill, New York, 1988.

[9] CULSHAW, B.: *Smart Structures and Materials*, Artech House, 1996.

[10] Sze, S.M.: *Semiconductor Devices: Physics and Technology*, 1985, Wiley & Sons Inc., NY.

[11] GARDNER, W.J.: *Microsensors – Principles and Applications*, Wiley & Sons Inc., NY, 1994.

[12] WINTER, M.: WebElements™ Periodic table, 1993–2003, The University of Sheffield and WebElements Ltd, UK http://www.webelements.com.

[13] Professor Donald R. SADOWAY – Introduction to Solid State Chemistry, Massachusetts Institute of Technology http://web.mit.edu/3.091/www/index.html and http://web.mit.edu/3.091/www/pt/pert12.html.

[14] MOSELEY, P.T., CROCKER, A.J.: *Sensor Materials*, Institute of Physics Publishing Ltd, 1996.

[15] Okmetic Oyj, Finland (high-quality silicon wafers), http://www.okmetic.com.

[16] BALTES, H., BRAND, O.: "CMOS-based microsensors", *Sensors and Actuators*, A 2946, 1-9, 2001.

[17] MEMS Exchange® http://www.mems-exchange.org/catalog/ – Process Hierarchy http://www.mems-exchange.org.

[18] Micromachined Sensors for a Global Market, Norway, http://www.sensonor.com.

[19] Institute of Microelectronics (IME), National University of Singapore www.ime.org.sg.

[20] RASMUSSEN, F.E., RAVNKILDE, J.T., TANG, P.T., HANSEM, O., BOUWSTRA, S.: "Electroplating and characterization of cobalt-nickel-iron and nickel-iron for magnetic microsystems application", *Sensors and Actuators*, A 2982, 1-7, 2001.

[21] Colibrys SA (MEMS, MOEMS,MOC), US, http://www.colibrys.com.

[22] Protron Mikrotechnik GmbH in Bremen, Germany, (Deep Reactive Ion Etching), http://www.protron-mikrotechnik.de.

[23] Olivetti Group, IT http://www.olivetti.com.

[24] SUSS MicroTec http://www.suss.com/mems/st.

[25] Sandia National Laboratories, US mems.sandia.gov.

[26] microFAB – Bremen, Germany http://www.microfab.de.

[27] Texas Instruments Team, US http://www.us.st.com.

[28] AT&T, US http://www.att.com.

[29] Fujimasa, I: *Micromachines – A New Era in Mechanical Engineering*, Oxford University press, 1996.

[30] X-FAB, European-American group http://www.xfab.com.

[31] VAN KAN, J.A., BETTIOL, A.A., WEE, B.S., SUM, T.C., TANG, S.M., WATT, F.: "Proton Beam Micromachining: a new tool for precision three-dimensional microstructures", *Sensors and Actuators*, A 3000, 1-5, 2001.

[32] Aspen Technologies, US www.aspentechnologies.net.

[33] TIMP, G.L. (ed.): *Nanotechnology*, AIP, 1998.

[34] MADOU, M.J.: *Fundamentals of Microfabrication*, CRC Press, 2002.

[35] JACKSON, M.J.: *Microfabrication and Nanomanufacturing*, CRC Press, 2006.

List of Authors

Stanislav Ďado works as a Professor at the Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague.

Jan Fischer works as an Assistant Professor at the Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague.

G. Hartung was head of the Department of Process Automation at the Fraunhofer Institute for Factory Operation and Automation in Magdeburg, director of a software factory for BOS System House GmbH and managing director of ISIK GmbH. Since 2001 he has been a guest lecturer at the Magdeburg-Stendal University. Since 2002 he has been a member of executive committee of ISIK AG.

A. Hospodkova, *E. Hulicius* and *P. Neuzil* work at the Institute of Physics at the Czech Academy of Sciences in Prague.

L. Indesteege is the managing director of VIA, providing tailor made training to local industry in collaboration with schools and universities. He has previously worked as a researcher and teacher at various European universities on the subject of energy saving and measurements and control. He also works with many partners on European projects on lifelong learning.

Anne-Elisabeth Lenel graduated in 1989 from Ecole Ste Genevieve–Versailles and in 1991 from INA Paris–Grignon. She worked as a researcher for Pernod Ricard and Martini&Rossi. Since 1999 she has worked for M2A Technologies as a coordinator and manager of national and European projects.

Gillian McMahon works as a researcher and project manager for Dublin City University, Ireland.

R. Meylaers, F. Peeters and *M. Peetermans* work as lecturers for KHK University, Belgium.

André Migeon works as a project and development manager for Crouzet and Sextant Avionics. Since 1997 he has been a General Manager for M2A, a company which trains experts in measurement for companies and universities. He serves as an expert on sensor technologies for the French government and European Commissions.

J. Novak works at the Faculty of electrical Engineering at the Czech Technical University in Prague.

Pavel Ripka has worked at various European universities and in 2001 he held a Marie Curie Experienced Researcher's Fellowship at the National University of Ireland, Galway. He works at the Department of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague as a professor, lecturing in measurements, engineering, magnetism and sensors. He is a member of the IEEE, the Elektra society, the Czech Metrological Society, the Czech National IMEKO Committee and the Eurosensors Steering Committee. From 2001 to 2005 he served as an associate editor of the IEEE Sensors Journal. He was a General Chairman of the Eurosensors XVI conference held in Prague in 2002.

S. Ripka works at the Department of Social Sciences at Charles University in Prague.

Alois Tipek is currently a postdoctoral fellow and project manager at Tyndall (formerly NMRC), Cork, Ireland. He has participated in several Czech and international research projects. In 2002 he received the Siemens Award for PhD students.

Index

1-9

3-D microfabrication techniques 503, 505

A

acceleration 2, 28, 33, 193-197, 199, 201, 202, 204, 205, 208, 213, 217, 219-224, 226-236, 239-241, 309, 342, 395, 400, 405, 411, 412, 427, 428, 469,
accelerometer
- general 193-244, 309, 395, 426, 478, 479
- piezoelectric 144, 147, 194, 195, 205-212, 236
- piezoresistive 195, 205, 213-219
- with resonator 195, 206, 219-221
- capacitive 195, 201, 205, 206, 221-224, 239, 240, 241, 242
- potentiometric 195, 201, 206, 224-225, 226, 227, 229, 230
- magnetic 227-228
- servo 195, 205, 206, 221, 229-231

acoustic (mass) sensor 269, 270, 301
a/d conversion 142, 144-146, 148-151, 153-155, 180
AMR sensor 144, 316, 439-447, 460, 465
angular rate 395, 396, 398, 399, 400, 404, 405, 412, 413, 417, 420-423

B

balance of force 9, 33, 40, 43, 44, 230
Bernoulli equation 5, 86, 87, 88, 90, 94
biosensor 245-303
Bourdon tube 8, 11, 13, 356, 357, 358

C

capacitive sensor 9, 21, 22, 25, 114, 116, 117, 119, 130, 134, 144, 195, 201, 205, 206, 211, 221-224, 229-231, 239-241, 319-322, 338-341, 412-414, 419, 486
chemical sensor 79, 80, 245-303, 478, 484, 497
Coriolis
- mass-flow meter 127
- force 404-409, 411, 412, 419

current sensor 433, 434, 460, 465, 466, 471, 472

D

Dall tube 88, 99, 100
data processing 12, 142, 143, 144, 147, 151, 219, 221, 227, 231

deposition technology 16, 260, 264, 288, 440, 478, 481, 482, 487, 490-495, 497, 498, 500, 503
diaphragm 8, 11, 12, 13, 18, 21, 22, 23, 26, 27, 28, 30, 32, 35, 39, 73, 88, 356

E

eddy current sensor 144, 145, 314-315, 463, 465-467, 470
electret effect 9, 23
electrochemical sensor 248, 253, 254, 256, 258, 263, 264, 274, 293, 297, 299, 301, 302, 481, 498,
etching process 215, 233, 479, 487, 489, 491, 500-505, 507

F

flow meter
  float 101, 102
  target 103, 104
  turbine 104-107
  mechanical 108
  magnetic 111, 112, 116
  vortex 117-122
  ultrasonic 123, 127
flow nozzle 98, 100
fluxgate sensor 457-463, 472, 473

G

gauge
  with taut wire 9, 18
  with deposited film 9, 16
GMR sensor 447-457
gyrometer
  general 198, 199, 395-432
  rotary 404, 427
  vibrating 404, 405, 406, 407, 409, 414-419, 427
  optical 395, 404, 420, 421, 422, 424

H

Hall sensor 143, 149, 316-317, 434-438, 439, 465, 469, 473, 483,

I

IC technology 215, 232-234, 347, 384, 385, 370, 421, 477, 478, 481, 487, 490, 491
inductive sensor 9, 26, 144, 309, 313, 195, 205, 227, 229, 230, 309-315, 468,
interface
  analog 166, 167
  communication 147, 148, 149
  digital 166, 167, 168, 169, 170
  general 46, 164, 195, 448, 449,
  human machine (HMI) 143, 147, 148
  sensor 151, 155, 166, 171, 172
ionization 7, 8, 41, 42
inclinometer 200, 201, 235-237, 309
induction sensor 143, 314, 457-459, 465, 466
intelligent sensor 141, 142, 145, 146, 148-151, 153, 154, 155, 180, 343, 479

L

level measurement 201, 337, 339, 341, 342
light detector 54, 57, 331

M

magnetic sensor 142, 143, 144, 150, 227-228 315, 316, 317, 318, 416, 426, 428, 433-475, 483, 486
mass-flow 84, 86, 127, 137
MEMS 343, 477, 478, 479, 491, 500, 503, 504
microwave sensor 131, 135, 335-337, 339, 342
monolithic sensor 215, 232-234, 347, 384, 385, 370, 421, 487

O

optical fiber 9, 38, 39, 49, 50, 51, 70, 74-78, 226, 339, 390, 421, 471, 505,
optical sensor
  chemical 253, 265, 266, 267, 268, 282, 293, 301, 302
  general 9, 27, 38, 39, 49-82, 145
  level, position and distance 323, 326, 328, 331, 332, 343
orifice plate 88, 93-101
oscillator 9, 10, 30-32, 35, 105, 221, 314, 315, 408, 420, 466

P

photoelectric switch (PES) 66-70, 323-327,
piezoelectric sensor 10, 27, 28, 29, 35, 117, 119, 120, 194, 195, 205, 206-212, 236, 334, 338, 341, 407, 408, 409, 417, 481, 482, 485,
Piezoresistive sensor 8, 9, 15, 18-20, 144, 195, 205, 206, 213-219, 437
Pirani gauge 8, 43
Pitot tube 89, 90, 91, 93, 95, 100, 101
position Sensitive photo-Detectors (PSD) 57, 58, 59, 62, 63, 327,
position sensor 62, 63, 132, 167, 434, 459, 461, 465, 469, 483
potentiometer 9, 13, 14, 15, 57, 143, 205, 224, 225, 306-309,
Prandtl tube 88, 89
pressure difference 26, 84, 88, 90, 91, 98, 101, 119, 339, 342,
pressure sensor 1-48, 73, 74, 75, 76, 79, 119, 144, 147, 149, 150, 151, 342, 478, 491,
pressure standard 43, 44, 45
pyrometer 385-390

R

rate gyro 399, 401
resistance temperature detection (RTD) 143, 347, 355, 377-384
resistive sensor 146, 306, 309
Reynolds number 85

S

sensor networks 142, 148, 151, 154-160, 164, 171, 190, 192
signal conditioning 142, 143, 144, 147, 310, 314, 315, 317, 320-322, 328, 337,
static pressure 5, 9, 27, 28, 87, 88, 90, 95, 147, 151,

T

temperature sensor 75, 122, 143, 146, 147, 150, 166, 347-393, 426,
thermistor 117, 119, 143, 370, 382-384, 388
thermo electric measurement (thermocouple) 8, 143, 146, 347, 354, 363, 364-367, 369-377, 380, 382, 384, 389, 390

U

ultrasonic sensor 117, 121, 123, 127, 137, 333, 334, 335, 338, 339, 341

V

Venturi tube 88, 99, 100, 101
volume flow 84, 86, 137

W

waveguide 49, 51, 79, 318, 469,
weir 136, 137
Wheatstone bridge 15, 43, 213, 216, 217, 316, 318, 380, 381, 443, 454,
wireless networks 149, 156-160, 164, 171, 190, 192