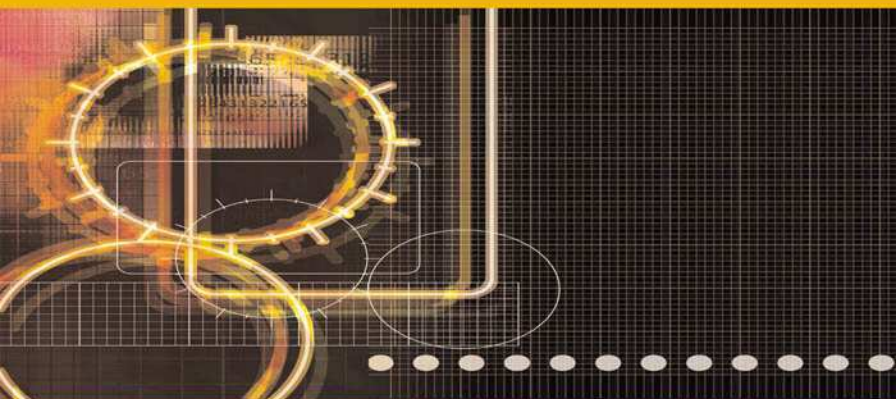


**INSTRUMENTATION AND MEASUREMENT SERIES**



# **Fundamentals of Instrumentation and Measurement**

**Edited by Dominique Placko**

**ISTE**

# Fundamentals of Instrumentation and Measurement

*This page intentionally left blank*

# **Fundamentals of Instrumentation and Measurement**

Edited by  
Dominique Placko

**ISTE**



First published in France in 2000 by Hermès Science Publications in two volumes entitled “Mesure et Instrumentation”

Published in Great Britain and the United States in 2007 by ISTE Ltd

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd  
6 Fitzroy Square  
London W1T 5DX  
UK

ISTE USA  
4308 Patrice Road  
Newport Beach, CA 92663  
USA

[www.iste.co.uk](http://www.iste.co.uk)

© ISTE Ltd, 2007

© HERMES Science Europe Ltd, 2000

The rights of Dominique Placko to be identified as the author of this work have been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

---

Library of Congress Cataloging-in-Publication Data

[Mesure et instrumentation. English]  
Fundamentals of instrumentation and measurement/edited by Dominique Placko.  
p. cm.  
Includes index.  
ISBN-13: 978-1-905209-39-2  
1. Mensuration. 2. Engineering instruments. 3. Scientific apparatus and instruments.  
4. Detectors. I. Placko, Dominique.  
T50.M394 2006  
620'.0044--dc22

2006020964

---

British Library Cataloguing-in-Publication Data  
A CIP record for this book is available from the British Library  
ISBN 13: 978-1-905209-39-2

---

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

# Table of Contents

<b>Introduction</b> . . . . .	xvii
<b>Chapter 1. Measurement Instrumentation</b> . . . . .	1
Mustapha NADI	
1.1. General introduction and definitions . . . . .	1
1.2. The historical aspects of measurement . . . . .	2
1.3. Terminology: measurement, instrumentation and metrology . . . . .	4
1.4. MIM interactions: measurement-instrumentation-metrology . . . . .	4
1.5. Instrumentation . . . . .	5
1.6. Is a classification of instruments possible? . . . . .	7
1.6.1. Classification of instruments used in cars . . . . .	9
1.7. Instrument modeling . . . . .	10
1.7.1. Model of a measurement instrument . . . . .	11
1.7.2. Load effects . . . . .	12
1.7.3. Estimating load effects . . . . .	12
1.7.4. Effort and flow variables . . . . .	13
1.7.5. Features and operating points of a system . . . . .	14
1.7.6. Generalized impedance . . . . .	16
1.7.7. Determining the load effect . . . . .	18
1.7.8. Measurement with a car battery . . . . .	19
1.7.9. Determining impedances . . . . .	20
1.7.10. Generalized admittance . . . . .	20
1.8. Characteristics of an instrument . . . . .	20
1.8.1. Components of static transfer functions . . . . .	21
1.8.2. Dynamic characteristics . . . . .	22
1.8.3. Instrument performance . . . . .	22
1.8.4. Combining transfer functions . . . . .	22
1.9. Implementing measurement acquisition . . . . .	23
1.9.1. Principles and methodology of measurement . . . . .	23

- 1.9.2. Field measurement constraints: instrumentation on the road . . . . . 26
- 1.10. Analyzing measurements obtained by an instrument. . . . . 26
  - 1.10.1. Error reduction. . . . . 27
  - 1.10.2. Base definitions . . . . . 27
- 1.11. Partial conclusion . . . . . 28
- 1.12. Electronic instrumentation . . . . . 28
- 1.13. Electronic instrumentation functionality . . . . . 30
  - 1.13.1. Programmable instrumentation . . . . . 32
  - 1.13.2. Example of an electronic instrument: how a piezoelectric sensor detects rattle in a combustion engine . . . . . 33
- 1.14. The role of instrumentation in quality control. . . . . 34
- 1.15. Conclusion . . . . . 35
- 1.16. Appendix . . . . . 36
- 1.17. Bibliography . . . . . 37

**Chapter 2. General Principles of Sensors . . . . . 41**

François LEPOUTRE

- 2.1. General points . . . . . 41
  - 2.1.1. Basic definitions . . . . . 41
  - 2.1.2. Secondary definitions . . . . . 43
- 2.2. Metrological characteristics of sensors. . . . . 43
  - 2.2.1. Systematic errors . . . . . 44
  - 2.2.2. Random uncertainties . . . . . 44
  - 2.2.3. Analyzing random errors and uncertainties. . . . . 45
    - 2.2.3.1. Evaluating random uncertainties. Standard deviations. Variances . . . . . 45
    - 2.2.3.2. Decisions about random uncertainties. . . . . 47
    - 2.2.3.3. Reliability, accuracy, precision. . . . . 48
- 2.3. Sensor calibration . . . . . 49
  - 2.3.1. Simple calibration . . . . . 49
  - 2.3.2. Multiple calibration. . . . . 50
  - 2.3.3. Linking international measurement systems . . . . . 50
- 2.4. Band pass and response time . . . . . 50
  - 2.4.1. Harmonic response . . . . . 50
  - 2.4.2. Response time . . . . . 56
- 2.5. Passive sensor conditioners . . . . . 59
  - 2.5.1. The effect of polarization instabilities. . . . . 59
  - 2.5.2. Effects of influence variables. . . . . 61
  - 2.5.3. Conditioners of complex impedance sensors. . . . . 63
- 2.6. Conditioners for active sensors . . . . . 64
  - 2.6.1. Direct reading . . . . . 64
  - 2.6.2. Using operational amplifiers . . . . . 66
- 2.7. Bibliography . . . . . 69

<b>Chapter 3. Physical Principles of Optical, Thermal and Mechanical Sensors</b> . . . . .	71
François LEPOUTRE	
3.1. Optical sensors . . . . .	71
3.1.1. Energetic flux . . . . .	72
3.1.2. Luminous flux . . . . .	73
3.1.3. The relative luminous efficiency curve $V(\lambda)$ of the human eye . . . . .	73
3.1.4. The black body: a reference for optical sensors . . . . .	76
3.1.4.1. Black body radiation . . . . .	77
3.1.4.2. Realization of black bodies . . . . .	78
3.1.5. Radiation exchanges between a source and a detector . . . . .	81
3.1.6. Definitions relating to optical sensors . . . . .	82
3.1.6.1. Darkness currents . . . . .	82
3.1.6.2. Spectral and total sensitivities . . . . .	82
3.1.6.3. Sources of fundamental noise sources in optical sensors . . . . .	82
3.1.6.4. Specific detectivity . . . . .	84
3.1.7. Semiconductors: the bases of optical sensors . . . . .	85
3.1.7.1. Molecular and crystalline bands . . . . .	85
3.1.7.2. Band structures in solids . . . . .	87
3.1.8. Current expression in a material containing free charges . . . . .	91
3.1.9. Photoconductor cells . . . . .	94
3.1.10. P-N junction and photodiodes . . . . .	99
3.1.10.1. Non-polarized junctions . . . . .	99
3.1.10.2. P-N junction with direct bias . . . . .	100
3.1.10.3. P-N junction in reverse bias . . . . .	101
3.1.10.4. Diode equation . . . . .	102
3.1.10.5. Illuminated P-N junctions . . . . .	103
3.1.10.6. Principle of photodiode fabrication . . . . .	103
3.1.10.7. Photodiode equation . . . . .	104
3.1.10.8. Electrical schema for a diode . . . . .	104
3.2. Force and deformation sensors . . . . .	109
3.2.1. Resistive gauges . . . . .	109
3.2.2. Piezoelectric effect . . . . .	110
3.2.2.1. Electrostriction, piezoelectricity and pyroelectricity . . . . .	111
3.2.2.2. The case of quartz . . . . .	111
3.2.2.3. Constraint tensors . . . . .	114
3.2.2.4. Other piezoelectric materials . . . . .	116
3.2.2.5. Construction of piezoelectric sensors . . . . .	117
3.2.2.6. Using piezoelectric sensors . . . . .	117
3.3. Thermal sensors . . . . .	119
3.3.1. Concepts related to temperature and thermometry . . . . .	119
3.3.2. Thermodynamic temperature . . . . .	120

- 3.3.3. Temperature scales currently in use and widely used measurements . . . . . 121
- 3.3.4. Heat transfers . . . . . 122
  - 3.3.4.1. Conduction. . . . . 122
  - 3.3.4.2. Convection. . . . . 125
  - 3.3.4.3. Radiation . . . . . 126
  - 3.3.4.4. Contact temperature measurement of solids . . . . . 127
- 3.3.5. Contact thermometers . . . . . 128
  - 3.3.5.1. Resistive thermometers . . . . . 128
  - 3.3.5.2. The Seebeck effect . . . . . 129
  - 3.3.5.3. The Peltier effect . . . . . 131
  - 3.3.5.4. The Thomson effect . . . . . 131
  - 3.3.5.5. The Seebeck electromotive force. . . . . 132
- 3.3.6. Features and uses of thermocouples . . . . . 134
- 3.4. Bibliography . . . . . 135

**Chapter 4. Analog Processing Associated with Sensors.** . . . . . 137  
Eduardo SANTANDER and Bernard JOURNET

- 4.1. Introduction. . . . . 137
- 4.2. The problem of electronic noise. . . . . 138
  - 4.2.1. The origin of electronic noise. . . . . 138
  - 4.2.2. Noise in an electronic chain. . . . . 143
  - 4.2.3. Signal-to-noise ratio . . . . . 145
- 4.3. Amplifiers. . . . . 147
  - 4.3.1. Operational amplifier . . . . . 147
    - 4.3.1.1. Feedback and counter-feedback in currents and tensions . . . . . 148
    - 4.3.1.2. Principle features of operational amplifiers . . . . . 153
  - 4.3.2. Instrumentation amplifiers . . . . . 160
  - 4.3.3. Isolation amplifiers . . . . . 162
  - 4.3.4. Logarithmic amplifiers. . . . . 163
  - 4.3.5. Multipliers . . . . . 164
- 4.4 Bibliography . . . . . 165

**Chapter 5. Analog Filters** . . . . . 167  
Paul BILDSTEIN

- 5.1. Introduction. . . . . 167
- 5.2. Technological constraints . . . . . 167
- 5.3. Methods of analog filter calculation . . . . . 169
  - 5.3.1. Attenuation functions of standard low pass prototype filters . . . . . 172
  - 5.3.2. Transfer functions of common prototype low pass filters . . . . . 174
  - 5.3.3 Transfer functions of derived filters . . . . . 174
  - 5.3.4. Filter synthesis carried out from the transfer function . . . . . 175

5.4. Passive filter using inductors and capacitors . . . . .	177
5.4.1. Sensitivity; Orchard's theorem and argument . . . . .	178
5.4.2. Low pass ladder filters . . . . .	179
5.4.2.1. Structures of basic low pass filters . . . . .	180
5.4.2.2. The Darlington analytic synthesis . . . . .	181
5.4.2.3. Examples of synthesis . . . . .	184
5.4.2.4. Direct digital synthesis . . . . .	187
5.4.3. L-C filters derived from a pass band . . . . .	189
5.4.4. Conversions of L-C filters; optimization . . . . .	190
5.5. Active filters . . . . .	191
5.5.1. Second order or biquadratic cells . . . . .	192
5.5.2. Biquadratic cells with one operational amplifier . . . . .	192
5.5.3. Universal biquadratic cells with three or four amplifiers . . . . .	195
5.5.4. Elevated order active filters (elevated by putting biquadratic cells in cascade) . . . . .	199
5.5.5. Simulating an L-C filter . . . . .	200
5.6. Switched capacitor filters . . . . .	202
5.6.1. Integrators without sensitivity to stray capacitances . . . . .	205
5.6.2. Analysis of switched capacitor integrators . . . . .	206
5.6.3. Synthesis of switched capacitor filters . . . . .	207
5.6.4. Operational simulation of an L-C filter (leapfrog simulation) . . . . .	208
5.6.5. Switched capacitor biquadratic cells . . . . .	211
5.7. Bibliography . . . . .	212

## **Chapter 6. Real-time Data Acquisition and Processing Systems . . . . . 215**

Dominique MILLER

6.1. Introduction . . . . .	215
6.2. Electronic devices for signal sampling and quantification . . . . .	216
6.2.1. Nyquist sampling . . . . .	216
6.2.2. Quantification noise . . . . .	217
6.2.3. Over-sampling . . . . .	219
6.2.3.1. Acquisition over-sampling . . . . .	219
6.2.3.2. Over-sampling and reconstruction . . . . .	222
6.2.4. Under-sampling . . . . .	224
6.3. Analog-to-digital converters . . . . .	229
6.3.1. Features of SINAD and ENOB converters . . . . .	230
6.3.2. $\Sigma - \Delta$ converters . . . . .	231
6.4. Real-time digital analysis by a specialized processor . . . . .	242
6.4.1. Fixed point and floating point analysis . . . . .	243
6.4.1.1. Fixed point notation . . . . .	243
6.4.1.2. Floating point notation . . . . .	243
6.4.1.3. Comparison between the two notations . . . . .	245

- 6.4.2. General structure of a DSP . . . . . 246
  - 6.4.2.1. Multiplication/accumulation structure. . . . . 247
  - 6.4.2.2. Time lag structures . . . . . 250
  - 6.4.2.3. Reframing structures . . . . . 252
  - 6.4.2.4. Resource parallelization . . . . . 254
- 6.4.3. Using standard filtering algorithms . . . . . 256
  - 6.4.3.1. General structure of a real-time filtering program. . . . . 256
  - 6.4.3.2. The FIR filter and simple convolutions . . . . . 258
  - 6.4.3.3. IIR filters. . . . . 260
- 6.5. Conclusion . . . . . 264
- 6.6. Bibliography . . . . . 265

**Chapter 7. The Contribution of Microtechnologies . . . . . 267**

François BAILLIEU and Olivier VANCAUWENBERGHE

- 7.1. Introduction. . . . . 267
  - 7.1.1. The vehicle: a system of complex, interdependent parts . . . . . 267
  - 7.1.2. Microtechnologies and microsystems . . . . . 268
  - 7.1.3. Appropriate architectures for electronic microsystems . . . . . 269
  - 7.1.4. Which examples should be chosen? . . . . . 270
- 7.2. Microtechnologies . . . . . 270
  - 7.2.1. Technologies derived from microelectronics. . . . . 275
    - 7.2.1.1. Si substrate. . . . . 275
    - 7.2.1.2. Si epitaxy. . . . . 275
    - 7.2.1.3. Si thermal oxidation . . . . . 276
    - 7.2.1.4. Photolithography . . . . . 277
    - 7.2.1.5. Polycrystalline silicon layer. . . . . 277
    - 7.2.1.6. Etching . . . . . 277
    - 7.2.1.7. Doping . . . . . 279
    - 7.2.1.8. Deposit of thin metallic and dielectric layers. . . . . 280
  - 7.2.2. Technologies specific to microstructures . . . . . 281
    - 7.2.2.1. Double face photolithography . . . . . 281
    - 7.2.2.2. Volume micromachining . . . . . 281
    - 7.2.2.3. Surface micromachining. . . . . 284
    - 7.2.2.4. Micromaching by deep anisotropic dry etching . . . . . 286
    - 7.2.2.5. Heterogenous assemblies . . . . . 287
  - 7.2.3. Beyond silicon . . . . . 288
- 7.3. Electronic architectures and the effects of miniaturization . . . . . 289
  - 7.3.1. Overall trends . . . . . 289
  - 7.3.2. Conditioning electronics for capacitive cells that are sensitive to absolute pressure . . . . . 291
    - 7.3.2.1. Measurement principle. . . . . 292
    - 7.3.2.2. The analog version . . . . . 293

7.3.2.3. Basic first order $\Sigma$ - $\Delta$ modulator with a one-bit quantifier . . . . .	297
7.3.3. Electronic conditioning for piezoresistive cells sensitive to differential pressure. . . . .	307
7.3.4. Electronic conditioning for cells sensitive to acceleration . . . . .	310
7.3.4.1. Direct applications of first-order $\Sigma$ - $\Delta$ modulators to 1 bit quantifiers . . . . .	310
7.3.4.2. Producing an accelerometer in true open loop by eliminating the effects of electrostatic forces . . . . .	312
7.3.4.3. Servo-control of an accelerometer using balanced mechanical forces through electrostatic forces. . . . .	316
7.3.5. Energy sources in microsystems . . . . .	322
7.4. Bibliography . . . . .	323

**Chapter 8. Instruments and Measurement Chains . . . . . 325**  
Bernard JOURNET and Stéphane POUJOLY

8.1. Measurement devices . . . . .	325
8.1.1. Multimeters . . . . .	326
8.1.1.1. Measurement principles . . . . .	326
8.1.1.2. Input resistance influence . . . . .	326
8.1.1.3. Intensity measurements . . . . .	327
8.1.1.4. Resistance measurements . . . . .	327
8.1.1.5. Two types of multimeters . . . . .	328
8.1.1.6. Measurement accuracy. . . . .	329
8.1.2. Frequency meters . . . . .	329
8.1.3. Oscilloscopes . . . . .	331
8.1.3.1. Introduction . . . . .	331
8.1.3.2. Input impedance and measurement . . . . .	332
8.1.3.3. Measurements done by an oscilloscope. . . . .	334
8.1.4. Spectrum analyzers. . . . .	334
8.1.4.1. Sweeping analyzers . . . . .	334
8.1.4.2. FFT analyzers . . . . .	336
8.1.4.3. Principles of possible measurements . . . . .	338
8.1.5. Network analyzers . . . . .	339
8.1.5.1. S parameters. . . . .	339
8.1.5.2. Measuring S parameters . . . . .	340
8.1.6. Impedance analyzers . . . . .	342
8.1.6.1. Method using a self-equilibrated bridge . . . . .	342
8.1.6.2. RF 1-V method . . . . .	343
8.1.6.3. Measurement with a network analyzer . . . . .	344
8.1.7. Synchronous detection. . . . .	345
8.2. Measurement chains. . . . .	347
8.2.1. Introduction . . . . .	347



- 8.2.2. Communication buses PC/instruments . . . . . 348
  - 8.2.2.1. The parallel bus IEEE488 . . . . . 348
  - 8.2.2.2. Serial buses . . . . . 351
- 8.2.3. Internal acquisition cards . . . . . 354
  - 8.2.3.1. Description of inputs/outputs and associated conditioning . . . . . 355
  - 8.2.3.2. Description of PC buses . . . . . 356
- 8.2.4. External acquisition cards: the VXI system . . . . . 357
  - 8.2.4.1. Functions of the VXI bus . . . . . 357
  - 8.2.4.2. Description of the VXI bus . . . . . 357
- 8.3. Bibliography . . . . . 359

**Chapter 9. Elaboration of Models for the Interaction Between the Sensor and its Environment.** . . . . . 361

Michel LECOLLINET

- 9.1. Modeling a sensor’s interactions with its environment . . . . . 361
  - 9.1.1. Physical description of the model . . . . . 361
  - 9.1.2. Phenomenological approach . . . . . 362
  - 9.1.3. Adjustment. . . . . 362
- 9.2. Researching the parameters of a given model. . . . . 363
  - 9.2.1. The least squares method . . . . . 363
  - 9.2.2. Application to estimate a central value . . . . . 364
  - 9.2.3. Introduction to weighting . . . . . 366
- 9.3. Determining regression line coefficients. . . . . 368
  - 9.3.1. A proportional relation. . . . . 368
  - 9.3.2. Affine relations . . . . . 370
  - 9.3.3. Weighting application . . . . . 378
    - 9.3.3.1. Calculation hypotheses . . . . . 378
    - 9.3.3.2. Weighting and proportional relations . . . . . 378
    - 9.3.3.3. Weighting and affine relations . . . . . 380
  - 9.3.4. The least measured-squares line: when two measured variables contain uncertainties . . . . . 384
- 9.4. Example of a polynomial relation. . . . . 390
  - 9.4.1. A simple example. . . . . 390
  - 9.4.2. An example using weighting . . . . . 394
  - 9.4.3. Examples with correlated variables . . . . . 395
- 9.5. A simple example . . . . . 398
  - 9.5.1. Linearizing the function . . . . . 398
  - 9.5.2. Numerical search for the minimum of the function of the sum of the squared gaps . . . . . 401
- 9.6. Examples of multivariable models . . . . . 402
- 9.7. Dealing with constraints . . . . . 405
  - 9.7.1. Presentation of the method . . . . . 405

9.7.2. Using Lagrange multipliers . . . . .	406
9.8. Optimizing the search for a polynomial model . . . . .	407
9.8.1. System resolution . . . . .	407
9.8.2. Constructing orthogonal polynomials using Forsythe's method . . . . .	410
9.8.3. Finding the optimum degree of a smoothing polynomial . . . . .	411
9.9. Bibliography . . . . .	413

**Chapter 10. Representation and Analysis of Signals . . . . .** 415  
Frédéric TRUCHETET, Cécile DURIEU and Denis PRÉMEL

10.1. Introduction . . . . .	415
10.2. Analog processing chain . . . . .	416
10.2.1. Introduction . . . . .	416
10.2.2. Some definitions and representations of analog signals. . . . .	416
10.2.2.1. Deterministic signals . . . . .	416
10.2.2.2. Random signals . . . . .	421
10.3. Digital processing chain. . . . .	422
10.3.1. Introduction . . . . .	422
10.3.2. Sampling and quantization of signals . . . . .	423
10.3.2.1. The Fourier transform and sampling . . . . .	423
10.3.2.2. Quantization . . . . .	427
10.4. Linear digital filtering . . . . .	429
10.4.1. The z transform . . . . .	429
10.4.2 Filtering applications . . . . .	430
10.4.3. Synthesis of IIR filters . . . . .	433
10.4.3.1. Methods using an analog reference filter . . . . .	433
10.4.3.2. Methods of synthesis by optimization . . . . .	434
10.5. Examples of digital processing . . . . .	436
10.5.1. Matched filtering . . . . .	436
10.5.2. Optimum filtering . . . . .	437
10.5.2.1. Wiener filtering . . . . .	437
10.5.2.2. Matched filtering . . . . .	439
10.5.2.3. Kalman filtering . . . . .	439
10.6. Frequency, time, time-frequency and wavelet analyses . . . . .	441
10.6.1. Frequency analysis . . . . .	443
10.6.1.1. Continuous transforms . . . . .	443
10.6.1.2. Discrete Fourier transform. . . . .	444
10.6.1.3. Algorithm of the fast Fourier transform . . . . .	446
10.6.2. Sliding window or short-term Fourier transform. . . . .	447
10.6.2.1. Continuous sliding window Fourier transform . . . . .	447
10.6.2.2. Discrete sliding window Fourier transform . . . . .	449
10.6.3. Wavelet transforms . . . . .	449
10.6.3.1. Continuous wavelet transforms . . . . .	450

- 10.6.3.2. Discrete wavelet transforms . . . . . 452
- 10.6.4. Bilinear transforms . . . . . 456
  - 10.6.4.1. The spectrogram . . . . . 456
  - 10.6.4.2. The scalogram . . . . . 457
  - 10.6.4.3. The Wigner-Ville transform. . . . . 457
  - 10.6.4.4. The pseudo-Wigner-Ville transform . . . . . 459
- 10.7. A specific instance of multidimensional signals . . . . . 459
- 10.8. Bibliography . . . . . 461

**Chapter 11. Multi-sensor Systems: Diagnostics and Fusion . . . . . 463**  
 Patrice AKNIN and Thierry MAURIN

- 11.1. Introduction . . . . . 463
- 11.2. Representation space: parametrization and selection. . . . . 465
  - 11.2.1. Introduction . . . . . 465
  - 11.2.2. Signal parametrization . . . . . 466
  - 11.2.3. Principle component analysis . . . . . 468
  - 11.2.4. Discriminate factorial analysis . . . . . 471
  - 11.2.5. Selection by orthogonalization . . . . . 474
- 11.3. Signal classification . . . . . 476
  - 11.3.1. Introduction . . . . . 476
  - 11.3.2. Bayesian classification . . . . . 477
    - 11.3.2.1. Optimum Bayes classifier . . . . . 477
    - 11.3.2.2. Parametric Bayesian classification . . . . . 480
    - 11.3.2.3. Method of the k-nearest neighbor . . . . . 480
    - 11.3.2.4. Parzen nuclei. . . . . 481
  - 11.3.3. Decision trees . . . . . 482
  - 11.3.4. Neural networks . . . . . 484
    - 11.3.4.1. Basic neurons . . . . . 484
    - 11.3.4.2. Mulilayered perceptrons . . . . . 486
    - 11.3.4.3. Radial base function networks . . . . . 488
    - 11.3.4.4. Neural networks and classification. . . . . 489
- 11.4. Data fusion . . . . . 490
  - 11.4.1. Introduction . . . . . 490
    - 11.4.1.1. Modelizing imperfections and performances . . . . . 490
    - 11.4.1.2. Different fusion techniques and levels. . . . . 491
  - 11.4.2. The standard probabilistic method . . . . . 492
    - 11.4.2.1. Modelization, decision and hypothesis choice . . . . . 492
    - 11.4.2.2. Multisensor Maysesian fusion . . . . . 494
  - 11.4.3. A non-standard probabilistic method: the theory of evidence . . . . . 495
    - 11.4.3.1. Mass sets of a source . . . . . 495
    - 11.4.3.2. Example of mass set generation . . . . . 497
    - 11.4.3.3. Credibility and plausibility . . . . . 498

11.4.3.4. Fusion of mass sets . . . . .	498
11.4.3.5. Decision rule . . . . .	499
11.4.3.6. Example . . . . .	499
11.4.4. Non-probabilistic method: the theory of possibilities . . . . .	501
11.4.4.1. Operations on ownership functions and possibility distributions . . . . .	502
11.4.4.2. Possibilistic multisensor fusion . . . . .	503
11.4.4.3. Diagnostics and fusion . . . . .	503
11.4.5. Conclusion . . . . .	505
11.5. General conclusion. . . . .	506
11.6. Bibliography . . . . .	506
<b>Chapter 12. Intelligent Sensors</b> . . . . .	509
Michel ROBERT	
12.1. Introduction . . . . .	509
12.2. Users' needs and technological benefits of sensors. . . . .	510
12.2.1. A short history of smart sensors . . . . .	514
12.2.2. Smart or intelligent? . . . . .	514
12.2.3. Architecture of an intelligent system. . . . .	515
12.3. Processing and performances . . . . .	516
12.3.1. Improving performances with sensors . . . . .	516
12.3.2. Reliability and availability of information . . . . .	517
12.4. Intelligent distance sensors in cars . . . . .	519
12.5. Fieldbus networks . . . . .	522
12.6. Towards a system approach . . . . .	523
12.7. Perspectives and conclusions. . . . .	524
12.8. Bibliography . . . . .	526
<b>List of Authors</b> . . . . .	529
<b>Index</b> . . . . .	531

*This page intentionally left blank*

## Introduction

# Instrumentation: Where Knowledge and Reality Meet

Instrumentation comprises scientific activities and technologies that are related to measurement. It is a link between physical, chemical and biological phenomena and their perception by humans. Constantly evolving, instrumentation changes how we live and plays a major role in industrial and life sciences; it is also indispensable to the fundamental sciences. In order to be credible, all new theories must undergo a series of experimental validations, of which instrumentation is the cornerstone.

Is curiosity a distinguishing human trait? Certainly, this characteristic leads us to question, to understand, to explain, and finally to “know”. The more we explore, the broader our range of investigation becomes. Since the 18<sup>th</sup> century, scientific and technical knowledge have undergone an exponential expansion, an explosive growth of combined learning, but this kind of growth leaves us with unanswered questions. In this context, instrumentation serves to stimulate scientific knowledge in the junction between theory and experimental practice.

Even before humanity developed a body of scientific knowledge, signs of technological progress had appeared in ancient civilizations. By 5,000 BC, humans had fashioned stone tools, and later began working in metal around 3,800 BC. Ancient Greeks, such as the philosopher Aristotle, who lived in the 4<sup>th</sup> century BC, were probably among the first thinkers to put forward logical explanations for observable natural phenomena. Democritus, a contemporary of Aristotle, already thought of matter as being formed of miniscule, indivisible particles. However, the

instrument of measurement most important to the Greeks was the gnomon, or needle of a sundial. The gnomon helped the Greek mathematician Euclid, living in the 3<sup>rd</sup> century BC, to measure the earth's radius by simultaneously observing the shadow cast by the instrument on two points of the same parallel. After this discovery, developments in mathematics, numerical theory and geometry followed, with Euclid's ideas dominating the world of science up until the Renaissance. From the 16<sup>th</sup> century onwards, Galileo, Newton, and Descartes brought forward new approaches that were truly objective, which meant that all new scientific theories had to be verified by observation and experiment. It was in this era that scientific instruments began to be widely developed and used.

The example we will discuss here will show, without forgetting Euclid's contribution as cited above, how instrumentation helped to join knowledge and reality. In the 18<sup>th</sup> century, both maritime navigation security and the possibility of complete world exploration were limited by current imprecision in measuring the coordinates of a ship traveling anywhere on Earth. The problem of calculating latitude already had been resolved some time before, thanks to fairly simple geometric measurements and calculations. Determining longitude presented more problems. As soon as a relation was established between the idea of time and space, scientists, especially astronomers, proposed using the movement of the stars as a cosmic clock: one example was the rotation of Saturn's satellites, discovered by the French astronomer Jean-Dominique Cassini in 1669. However, developing this idea further proved difficult and complicated. Determining longitude by relying on a measurement of time difference in relation to a given location required a precise measurement of time that was impossible to attain with the tools then available. To give an idea of the order of magnitude, let us recall that at the Equator, a nautical mile is defined as the length of a terrestrial curve intercepting an angle of a minute. The time zone being equivalent to 15 degrees, the lapse of time of a minute equals 15 minutes of curve or 15 nautical miles. Thus a nautical mile is equal to 4 seconds.

The problem was resolved in 1759 by the English clockmaker John Harrison, who invented a remarkable time-measuring instrument, a sea clock or chronometer that was only 5 seconds off after 6 weeks at sea, the equivalent of just 1.25 nautical miles. This revolutionary clock marked an important step in the search for precision begun in 1581 with Galileo's discovery of the properties of regularity in a swaying pendulum, a principle taken up and developed further in 1657 by the Dutch physician Christiaan Huygens, inventor of the pendulum clock. John Harrison's invention produced a number of other technological innovations such as ball bearings, which reduced friction that caused imprecision and errors. His chronometer stimulated progress in a number of other fields, among them cartography, leading to clearer, more geographically accurate maps. Today the Global Positioning System (GPS) stills depends on time measurement, but with a

margin of error of less than several centimeters, thanks to atomic clocks with a margin of error that never exceeds that of a second every 3 million years!

These kinds of remarkable discoveries became more frequent over time in all scientific and technological fields, often resulting in new units of measurement named after their inventors. Instead of the inexact and often anthropomorphic systems then in use, it became necessary to create a coherent system of measurement that could be verified by specific instruments and methods from which reproducible and universal results could be obtained. An example of one older unit of measurement was the “rope of 13 knots” used by European cathedral builders to specify angles of 30, 60 and 90 degrees. Other measurements long in use such as the foot and the inch obviously could not meet the criterion of reproducibility but did allow for the emergence of standards and the development of somewhat more regular measurements. The usage of these often varied from region to region, becoming more widespread over time. The ell, for example, differed not only according to place but also according to usage. The first tentative step toward a coherent system was clearly the British Imperial System, adopted in 1824 by Great Britain and its colonies. The SI, an abbreviation for the International System of Measurements today in use throughout much of the world, dates from 1960 and allows scientists to join all measurements in use to a group of specific and carefully chosen basic measurements, thus giving birth to a new field of science that could not exist without modern measurement: metrology.

As the development of the metrology shows, access to information, facts and measurements, all crucial to the interaction between knowledge and reality, also serve to stimulate technological innovation. Making use of the latest technology in the fields of sensors, measurement, communications, signal processing and information, modern instrumentation plays an unprecedented role in progress and science. An interdisciplinary field, instrumentation is itself present in almost all scientific disciplines, including the fundamental sciences, engineering science, medicine, economic and social sciences, promoting exchange of ideas and data between different scientific communities and researchers. The particle accelerator ring developed by CERN, the European Organization for Nuclear Research, is perhaps the newest instrument of measurement. With numerous subsets of specific measurements, this impressive instrument allows scientists to explore infinitely small things by studying and discovering new types of particles. As well, astrophysicists have attempted to validate certain elements of the big bang theory by more and more refined observations of the universe, making use of a vast array of extremely sophisticated technologies, among them the Hubble space telescope.

Resolving instrumentation issues frequently involves a very broad spectrum of theoretical abilities, as well as mastery of experimental techniques. This means that research teams in business and university laboratories, on the individual level, must



have scientists who can invest time in multi-disciplinary research; the teams themselves must also serve as conduits between research teams belonging to complimentary disciplines. This form of interdisciplinary activity, in which research teams are able to imagine and work out applications of their work beyond their own fields, is an extremely attractive challenge. But will this necessarily lead to innovative concepts – and if so, according to which scientific principles?

The reality is that of the vast range of solutions widely available to resolve any problem of measurement, very few are actually suitable. The emergence of an innovative and optimum system often appears as the result of an ingenious combination of a group of methods and technologies drawing on diverse disciplines. This approach does not necessarily mean a major development has occurred in each of the involved fields; it does, however, require in-depth knowledge of these fields. The innovation resulting from this mastery is not less rich, open and dynamic in terms of scientific, technological and economic terms, resulting as it does from interdisciplinary exchange.

The objective of this work on measurement and instrumentation is to present and analyze all the issues inherent in conceiving and developing measurement, from the source of a signal (sensor) to conveying quantitative or qualitative information to a user or a system. Le Colloque Interdisciplinaire en Instrumentation or Interdisciplinary Conference on Instrumentation held in November 1998 in Cachan, France gives a general survey of the range of this field (see C2I'98). This book cannot claim to be exhaustive. However, throughout the chapters, we give examples of our main theme – the idea of a system that brings together technologies, methods and complex components relating to theoretical, experimental, and scientific skills. All of these draw on the essence of instrumentation.

To give a well-known example of this theme, we look at the car, an object that has paradoxically retained the same function over decades even as it has never stopped changing and evolving. We are all aware of how new technologies, especially in the fields of micro-electronics and industrial computer science, have changed cars. We notice the continual appearance of new scientific concepts whose names and acronyms (such as the Antilock Braking System (ABS), the Enhanced Traction System (ETS) and controller area network (CAN) operating system) become familiar through widespread publicity and advertising of vehicles. In fact, the car as a symbol has become more interesting and inspiring than functions such as airbags or digital motor control which often make use of new, though hidden, technologies. These technologies usually develop within widely varying constraints such as safety, reliability, ease with which problems can be diagnosed and repairs can be made, and cost. Such technologies also are affected by marketing factors like style and comfort. The car is thus an illustration of an impressive technological

expansion that has taken place within the parameters of science and within the parameters of socio-economics.

This book has been written for technicians, industrial engineers, undergraduate students in the fields of electronics, electrical engineering, automation, and more generally those in disciplines related to engineering science who require in-depth knowledge of how systems of measurement are developed and applied. The chapters follow a fairly linear progression. However, our text falls into two complementary but somewhat different halves.

The first half of the book discusses fundamental ideas and issues of measurement and presents a range of physical phenomena that allow us to obtain measurable sizes and develop methods of pretreatment of signals. In these early chapters, our discussion of instrumentation focuses mainly on components. The second half of the book concentrates instead on the aspect of systems by looking at how data are processed and used. These two different emphases are linked in Chapter 6, which presents the carrying out of integrated functions, showing how microtechnologies have shown great promise in the fields of sensors and instrumentation.

Using the example of the car, the first chapter defines the links between instrumentation, measurement and metrology, explaining how units and tools of measurement are developed. Chapter 2 presents the general principles of sensors, while Chapter 3 gives a detailed description of the general principles of optical, thermal and mechanical sensors, and how these may be used in developing measuring tools and sensors. Chapters 4 to 6 discuss a range of methods and technologies that allow for a complete measuring process, from the conception of an electronic conditioning of signals, passage through discrete time, data conversion and quantification, filtering and numerical pretreatment.

Chapter 7 progresses from the idea of components to that of systems, concentrating on somewhat more technical aspects by discussing instrumentation in terms of microsystems, accelerometers, and pressure sensors. Chapters 8 to 11 present information on how systems and measurement networks are created, how models of interaction between sensors and their environment are developed, as well as ideas concerning representational space, diagnostic methods and merging of data. Chapter 12 summarizes the previous chapters and discusses the idea of intelligent systems and sensors, to which signal processing imparts valuable qualities of rapidity, reliability and self-diagnosis, available to us thanks only to the miniaturization of complex mechanisms that integrate a number of complex functions. We have chosen several examples from a specific field: the production of cars.

*This page intentionally left blank*

## Chapter 1

# Measurement Instrumentation

The purpose of this chapter is to review the essential definitions and characteristics of measurement. We discuss measurement systems and the roles and classifications of instruments in a comprehensive and descriptive way, with more detailed discussions to follow later in the book. Throughout this book, we use the example of the car to illustrate the importance and relevance of instrumentation.

### **1.1. General introduction and definitions**

Whether exploring Mars, measuring the brain's electrical signals for diagnostic purposes or setting up robots on an assembly line, measurement is everywhere. In all human activities, the idea of measurement establishes a relationship between a natural or artificial phenomenon and a group of symbols, usually numbers, in order to create the most reliable representation possible. This representation is classified according to an "orderly" scale of values.

Measurement is the basis of scientific and industrial research. It allows us to understand the phenomena we observe in our environment by means of experimental deduction and verification [ROM 89]; [HEW 90]; [PRI 95] and helps us keep records of the results of these observations. Established models and scientific laws are available for all of us, doing away with the need to begin each experiment with the most basic observations. This is why perpetuating knowledge is so important in the long term.

In the short term, this perpetuation guarantees the quality of products and commercial trade by connecting them to legal standards. Achieved through instrumentation, measurement is thus the basis of progress in many forms of knowledge, as well as being essential to production and trade. In the world of science, it allows us to make discoveries and confirm them. In terms of technology, instrumentation helps us control, improve and develop production, and in the world of economics, it makes commercial exchange possible, helping us assign value to objects and transactions.

Measurement therefore brings together knowledge and technological progress. Universal and essential to many disciplines [PRI 95], it is, in fact, fundamental to most human activity. This universality explains the recent interest among some researchers in improving the forms of knowledge related to instrumentation [FIN 82].

## 1.2. The historical aspects of measurement

We can look at the evolution of measurement by focusing on invented instruments or by using the instruments themselves. In this section, we will list the steps of progress in measurement, which we define somewhat arbitrarily, according to human needs as these emerged throughout history:

- the need to master the environment (dimensional and geographical aspects);
- the need to master means of production (mechanical and thermal aspects);
- the need to create an economy (money and trade);
- the need to master and control energy (electrical, thermal, mechanical, and hydraulic aspects);
- the need to master information (electronic and optoelectronic aspects).

In addition to these is the mastery of knowledge which has existed throughout history and is intimately connected:

- measurement of time;
- measurement of physical phenomena;
- measurement of chemical and biological phenomena.

Let us look at several examples from history regarding the measurement of time. The priest-astronomers of ancient Egypt were close observers of natural phenomena, especially the sky. Simply by observing the natural effects of solstices (including the floodings and harvests around the Nile coinciding with the rising of the star Sirius) they were able to invent a 365-day calendar. Their observations also enabled them to

develop a system of measurement based on a daily recording, made between summer solstices, of the shadows cast by a stick placed vertically in the ground. By about the year 2,500 BC, Egypt had three calendars: a civil calendar of 365 days, an equivalent lunar calendar, as well as one based on the earlier lunar year based on the heliacal rising of Sirius. Such progress was made by the Egyptian priest-astronomers that around the year 1,450 BC, during the reign of Thutmose III, they were able to measure days and hours, again only through observation. As can be seen on wall paintings of star clocks in tombs of that era, ancient Egyptians knew that the day consisted of 12 hours, compensating for the 12 dark hours of night. Their sundials – or, more accurately, “shadow clocks” – were very simple ancestors of the gnomons later used by the Greeks. These consisted of a rectilinear piece of wood in five sections, with a horizontal piece at one end. Through careful observations and corrections, the Egyptians of that era came very close to achieving our present level of knowledge of the duration and number of days in a year.

Throughout history, these kinds of advances in measurement have come about for specific motives. Economic motives drove the development of cartography and the growth of trade; militaristic motives spurred the creation of new armaments, with everything from cannon powder to the radiation levels emitted by nuclear weapons needing to be measured; strategic and expansionist motives prompted the need to control maritime routes and colonial territories; religious motives created a need to restrain and monopolize certain kinds of knowledge. Nowadays, these motives have developed with the disappearance of certain needs being replaced by new ones. An instance of this is how the need for sophisticated, three-dimensional maps of Earth that have become possible through the technology used by American space shuttles, has supplanted older colonial expansionist motives that gave birth to scientific bodies such as the *Bureau des longitudes* in France.

History is full of examples of the development of measurement to such an extent that no progress can be described or reported without a measurement being a result of completed experiences for the validation of theories [RON 82] [JAC 90], whether these are scientific, economic, technical, expansionist or even religious. Usually, the instrument used for such validation already exists but is used in a somewhat different way or is adapted for the new use. Instruments developed for a specific measurement are more rare. Religious motives have often brought about new ways and tools of measurement, especially in antiquity. As discussed above, ancient Egyptians used the sky to develop their calendar of 365 days and to measure days and hours. In our own time, some physicists confronting the mystery of particles and the Big Bang theory have turned to a spiritual explanation of these phenomena [HAW 89].

### 1.3. Terminology: measurement, instrumentation and metrology

The expression of measurement needs or tests are an everyday occurrence in science and industry [MAS 90]; [COM 92]. All existing tools that help us carry out measurement are part of instrumentation. Rules for using and guaranteeing measurement created metrology. It is important to point out that definitions<sup>1</sup> of these related terms are sometimes confused, as with “measure” and “metrology”.

The word *measurement* has many meanings. The International Vocabulary of Basic and General Terms in Metrology (VIM), using International Organization for Standardization (ISO) norms, has defined measurement as “a set of operations having the object of determining the value of a quantity”.

In other words, a measurement is the evaluation of a quantity made after comparing it to a quantity of the same type which we use as a unit. The concept of a *measurable quantity* goes beyond measurement. The VIM defines this as “an attribute of a phenomenon, body or substance, which can be distinguished qualitatively and determined quantitatively”.

*Metrology*, the science and “grammar” of measurement is defined as “the field of knowledge concerned with measurement”.

It guarantees the meaning and validity of measurement by strict accordance to established units [LAF 89]; [HIM 98]. These units are standardized on national and international levels [GIA 89]. Metrology plays a role in international agreements joining national systems of measurement to those used in other countries, making conversion between systems possible. Standardized measurement units mean that scientific and economic figures can be understood, reproduced, and converted with a high degree of certitude. The International Bureau of Weights and Measures based in France is one example of an international authority in charge of establishing international metrological rules.

### 1.4. MIM interactions: measurement-instrumentation-metrology

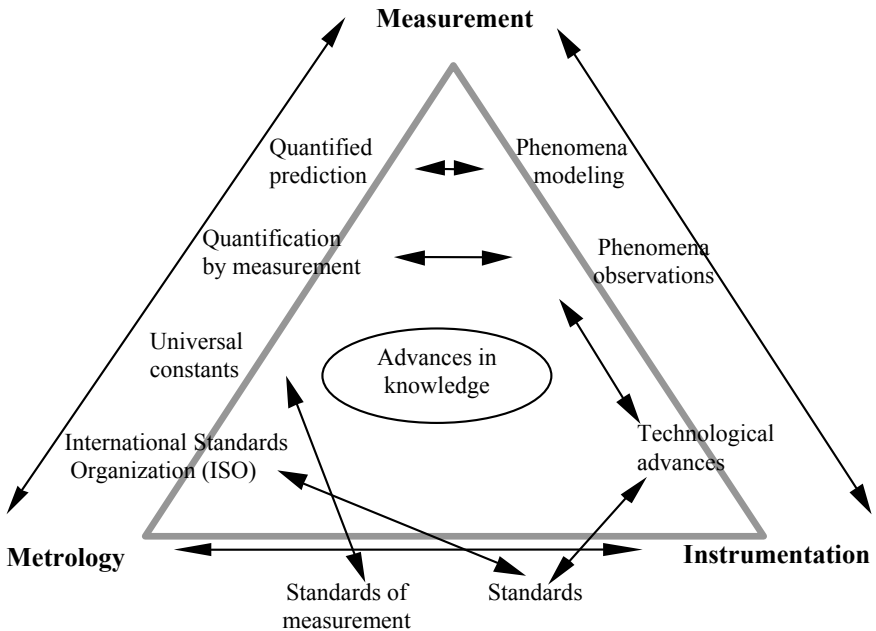
Knowledge fields have always grown according to measurement systems. “Experience” and “theory” interact and link together the “real world” and the “mathematical world” [DRA 83]. These interactions lead to overall progress in

---

<sup>1</sup> All definitions found in the text in italics come from the International Vocabulary of Basic and General Terms in Metrology.

scientific knowledge, with attendant technological advances that in turn benefit many disciplines (see Figure 1.1).

In scientific research, interactions between experiments and theories are permanent. Therefore, establishing a comparative relation between a quantity to be evaluated and a reference quantity or standard by means of an instrument of measurement is an interaction between instrumentation and metrology that guarantees the reliability of obtained results. Both the concept of measurement and the means used to obtain it, whether metrologic or instrumental, are part of interdependent evolutions. Technological advances develop and contribute to progress in the three fields defined above: measurement, instrumentation and metrology [RON 88]; [TUR 90].



**Figure 1.1.** *The MIM triangle: evolutions and permanent interactions of measurement, instrumentation, and metrology*

## 1.5. Instrumentation

The term instrumentation refers to a group of permanent systems which help us measure objects and maintain retroactive control of a process. In this sense,

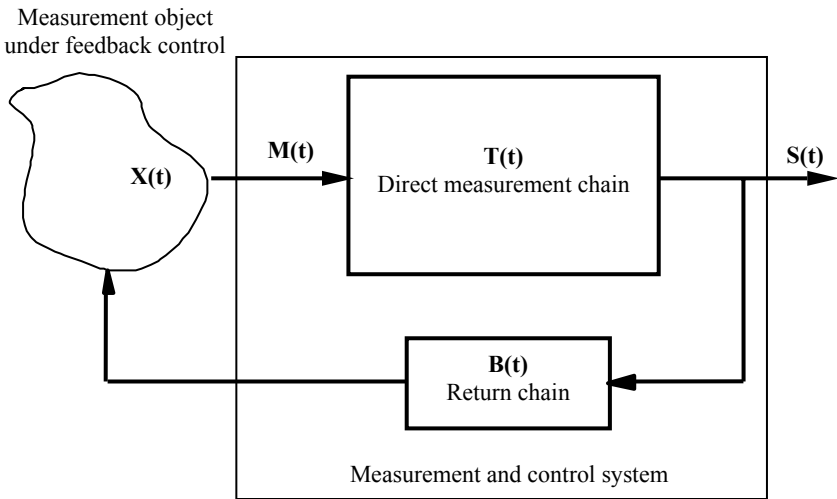


instruments and systems of measurement constitute the “tools” of measurement and metrology.

For our purposes, the following terms can be singled out:

- *measurement systems*: these are instruments used to establish the size of objects being scientifically tested. This kind of situation occurs in scientific experiments and industrial test trials to acquire information and data concerning the tested object. This data can be processed in real time or in batch mode (see Chapter 7);

- *control systems*: in addition to measuring objects, these instruments are also used to exert control over the feedback process. Figure 1.2 shows the conventional diagram of a measurement and control system.



**Figure 1.2.** Control and measurement system

The measurable quantity to be measured  $X(t)$  is transmitted by a signal  $M(t)$  at the input of the measurement chain. This, which is characterized by a transfer function  $T(t)$ , creates an exit signal  $S(t)$  in the form of  $X(t)$ . This can be completed by a feedback loop with a transfer function  $B$  that carries out the parameter control of the object being investigated according to preset or autoadaptive instructions. To simplify our explanation, we interchangeably use the terms measurement systems, instrumentation, and instruments. Physically, all measurement chains are based on a *measuring transducer*, which we define as “a measurement device which provides an output quantity having a given relationship to the input quantity”.

When an exit signal of a transducer is electric, we speak of a *sensor*, defined as “the element of a measuring instrument or a measuring chain to which a measurand is directly applied”.

The requirements of making correct and sophisticated measurements have meant that sensors have been increasingly used for this purpose. As instruments of management made electronically, sensors are capable of combining series of measurement results in a single indicator. This intelligence can be numerical; a data acquisition system connected to a computer directs the calculations and displays the measurements. As well, this intelligence can be integrated into a measuring sensor head in the form of microelectronic components that carry information to a compact and portable sensor with the capability of processing information in real-time. Research on sensors and their development is a rapidly expanding field; a fuller discussion follows in Chapter 6. In the next pages of this chapter, we will present other elements of the measurement chain, going into more detail later in the text.

### 1.6. Is a classification of instruments possible?

Does a taxonomy of instruments exist [WHI 87]? To the best of our knowledge, a universal classification of instruments has not yet been proposed.<sup>2</sup> The difficulties of even proposing such a classification are obvious, given that such an attempt would come up against problems of criteria choice. Would criteria have to be chosen according to the technologies being used or application fields?<sup>3</sup>

One approach would involve deciding on detailed utilization of a given approach, and thus criteria, allowing for a functional reading in terms of the research objectives. Starting from a given application field, we would index all required instruments in the most detailed way possible in terms of measuring function and nature, the dimensions of the instruments being used, and the sensors being used, to cite some elements of the process. Another approach would concentrate on different application fields, such as pressure measurement in agriculture, in medicine and in industry, to name several examples. Table 1.1 is an example of this kind of classification. It is far from exhaustive but shows some possible criteria.

---

<sup>2</sup> Two excellent books [ASC 87]; [FRA 96] and an article [WHI 87] all dealing with sensors are exceptions.

<sup>3</sup> For the definition of this term, see section 1.7.1.

Obviously, depending on the application field being used, it is generally difficult to carry out and validate reliable measurements. For example, the problems involved in measuring the pressure level of a submarine and of a machine tool are not the same. The constraints, consequences and precision demands are not comparable; neither are the environmental conditions.

However, other classification criteria are possible. Tables 1.4 and 1.5 (see also Appendix 1) give further examples of classification criteria in terms of the nature of the physical stimulus used and the physical quantity being measured.

Example of size to be measured	Application field
Fluid pressure	Industry
Sugar level in a fruit	Agriculture
Blood glucose level	Biology
Beam resistance	Civil engineering
Stock, currency exchange	Marketing, commerce, finance
Oilfield flow, power station output	Energy
Epidemiological monitoring, ECG signals, home health care monitoring	Health, medicine
Radar detection and surveillance	Military
Life span of an elementary particle	Scientific measurement
Flight speed, length of flight, altitude	Transportation
Battery fluid level	Automobile
Heavy metal level in wastewater	Environment
Atmospheric pressure, hygrometry level	Metrology
Presence detection	Home automation
Software performance, fiber optic flow, channel pass bands	Telecommunications
Undersea pressure and depth	Marine industry
Distance, speed, transmission time	Space

**Table 1.1.** *Examples of instrument classification criteria and related application fields*

### 1.6.1. Classification of instruments used in cars

The concept of the systems approach [ROS 75] is generally used in industrial design. Looking at the example of the car, it is possible to use this comprehensive approach to create a subset of instruments in this field. For purposes of brevity, we can say that the instruments necessary for a vehicle are centered around the driver and his or her needs [EST 95]. Driving a vehicle through traffic involves cooperation – and a certain amount of tactical skill. Planning the details of even a short car trip involves planning an itinerary, departure time and other details, all requiring strategic skill. Moreover, learning how to drive a car and ensuring its optimal and safe performance involves operational skills. A useful definition of instruments in this context would involve a classification by groups of functions, one flexible enough to accommodate technological changes and cost reduction.

A car or automotive vehicle is above all a group of interacting systems. The starting point of today's car designers is the idea that all components and modules must be planned, put together, and manufactured as integral parts of the same system. We can imagine the possible range of interactions within systems, a few being the smart sensor that activates an airbag and the data acquisition program that ensures a driver's safety. To further illustrate such interactions, we provide a list of some of the systems classes of a car in Table 1.2. This presentation follows the logic used in planning and production from an economic point of view.

<b>Temperature function</b>	<b>Chassis functions</b>	<b>String system functions</b>	<b>Passenger compartment and safety functions</b>
Temperature control functions	Chassis control systems	Steering systems	Inflatable airbags
Heating, ventilation, and air conditioning controls	Active suspension systems	Supporting columns and shafts	Passenger compartment amenities
Motor cooling systems	Chassis modules and systems	Steering shaft systems	Door control modules
Motor cooling controls	Brake systems.	Optimization, performance and fuel consumption systems	Electronic control systems
	Brake suspension components		

**Table 1.2a.** *Examples of classification in car instrumentation fields*

<b>Electronic functions</b>	<b>Transmission and wiring functions</b>	<b>Power and combustion functions</b>
Electronic antilock brakes (ABS) Electronic inflatable airbag unit Antenna systems Electronic passenger safety systems Audio components and systems Electronic engine monitoring Windshield projection system Collision protection systems Electronic dashboard (speed, gasoline levels, etc.) Integrated circuits Mechanical, electromagnetic, and electronic air conditioning regulators Energy sensors and controls Electronic steering and suspension VAN network	Electric and electronic generators Connection system (wiring) Electronic fittings Advanced data transmission Fiber optic lighting systems Lighting wiring Sensors Commutators and switches Modular side panels	Valve command Monitoring system Air and gasoline monitoring Exhaust system Sensors and thermostats Lighting Fuel supply and emission control Energy storage and conversion Advanced propulsion systems

**Table 1.2b.** *Examples of automotive instrumentation classifications*

### 1.7. Instrument modeling

From simple sensors and their conditioners to computer data acquisition systems, instruments must furnish reliable and accurate measurements. We can attempt to formalize data acquisition chains by using a global model to design an instrumental system to define different components and measures in use. Modeling an instrument of measurement depends on quantifiable and qualifiable knowledge of parameters – but these parameters cannot always be controlled. We can, however, attempt to

estimate these parameters quantitatively and qualitatively in order to evaluate their influence on acquisition and the representation of their real value.

### 1.7.1. *Model of a measurement instrument*

An instrument of measurement may be described in terms of input and output, according to the functional design of Figure 1.3 [NAC 90]; [PAR 87]. Input and output quantities allow for overall formalization in any measurement system. The sizes for which the system has been conceived are called the *measurands*, defined as “quantities subjected to measurement”.

The output phase of a measurement system delivers an “image” value  $S(t)$  of the characteristic being measured. Ideally, this value would be a faithful representation of the quantity to be determined, with the input and output linked by a characteristic transfer function of the measurement system or instrument.

In reality, however, we must add to the measurand  $M(t)$  additional quantities called *influence quantities*, defined as “quantities which are not the subject of measurement but which influence the value of the measurand or the indication of a measuring instrument”.

So, we can distinguish between:

- interfering quantities  $i(t)$  to which the system is unintentionally sensitive. The instrument takes their effects as disturbance that is taken into account as a supplementary transfer function that modifies output additively;
- modifying quantities  $m(t)$  that are all quantities capable of reacting on partial transfer functions when a temporary or permanent change in the structure of the instrument occurs.

These definitions identify the difference between real value  $M(t)$  and measured value  $S(t)$ . Metrology is a method used to rigorously analyze these differences. The role of the user is then to critically analyze the results using a thorough knowledge, by quantifying or qualifying influence quantities so as to estimate any possible errors they may cause [HOF 83]; [NEU 89].

In concrete terms, these errors manifest physically by an unwanted supplementary information transfer, which we will describe in the following sections.

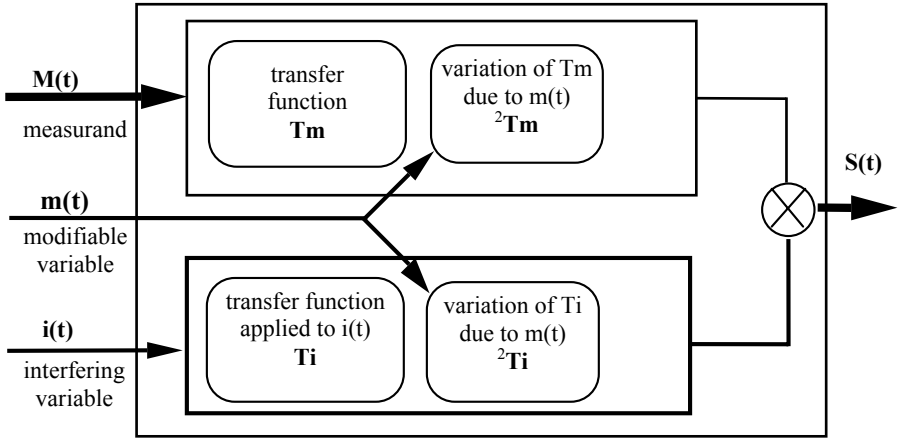


Figure 1.3. Modeling of a measurement system

**1.7.2. Load effects**

Any measurement operation requires connection (*in situ* invasive, semi-invasive or contact measurement) or measurement of an object using an instrument without contact. This linking of an instrument to an object or site of investigation means that a transfer of energy and/or information termed “a load effect” takes place [PAR 87, NACH 90]. This transfer directly affects the measured value.

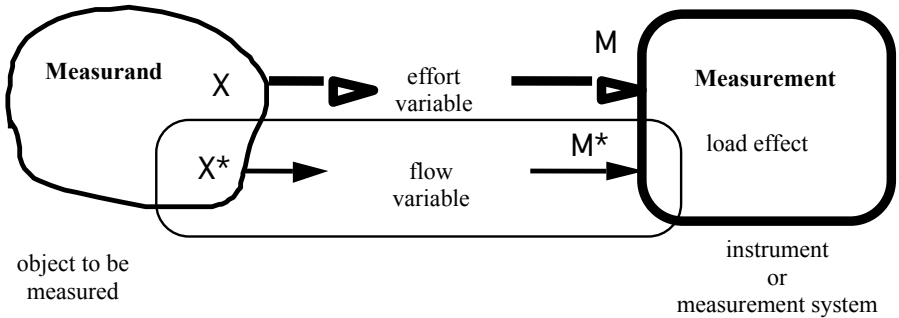
An example of this is shown by the insertion of a measuring probe into a solenoid which interferes with the electrical field, leading to a difference between the “true” value (the field by itself) and the value to be measured (the field disturbed by the probe).

We can, in certain cases, estimate and deduce errors that occur between measuring systems and the object to be measured. The measurand can then be achieved but may not be completely accurate; in such cases we must ensure that appropriate metrological procedures are followed. In other cases, measurement cannot be carried out, and being aware of this will help us find another solution to determining a quantity of interest.

**1.7.3. Estimating load effects**

If  $X(t)$  is the “true” value of the quantity to be measured when the object of measurement is not related to the measurement device, then  $M(t)$  stands for the

value of the quantity after measurement. The information conveyed from the object to be measured to the instrument of measurement represents an image of a measurand  $\mathbf{X}(t)$  upon which we superimpose information intrinsic to the energy of the connection, expressed as  $\mathbf{X}^*(t)$ . This energy transfer is a characteristic of measurement and means that the measured object delivers not only quantity  $\mathbf{M}(t)$  to the instrument of measurement but also a second quantity  $\mathbf{M}^*(t)$  (see Figure 1.4).



**Figure 1.4.** Load effect from contact of the object to be measured with a measurement system

This load effect can be described in terms of energy, a concept fundamental to all elements of all physical interactions, no matter what the quantity may be. In engineering sciences, we describe these interactions in terms of pairs of complementary variables whose product is equal to the power. We will further discuss the definition and role of these pairs.

#### 1.7.4. Effort and flow variables

The pair of variables associated with energy transfers is characteristic of all measurement operations. In a measurement system, one of its features is an “effort variable”  $\mathbf{M}(t)$  linked to a “flow variable”  $\mathbf{M}^*(t)$ . The result of these two variables to the dimension of a power:

$$P = M(t).M^*(t)$$

and its temporal integral:

$$W = \int M(t).M^*(t).dt$$

that of energy.



From the point of view of physics, one of these two variables is extensive: the *flow variable*, for example, current, speed, density flow or angular speed. The other is intensive and is a potential or *effort variable*: for example, tension, force or pressure. Sources of flow variables (current or speed) operate in constant flow and have infinite impedance. Information transfer follows the pair of complementary variables producing power or the “energy flow” that comes from interaction between the variables. In all pairs of variables found in classical physics such as electricity, mechanics, hydraulics and optics, we can define a size as equal to a power or form of energy:

$$P = \text{intensive variable} \times \text{extensive variable}$$

Certain pairs of variables are already familiar to us:

- *power = tension × current;*
- *energy = force × displacement;*
- *power = pressure × flow.*

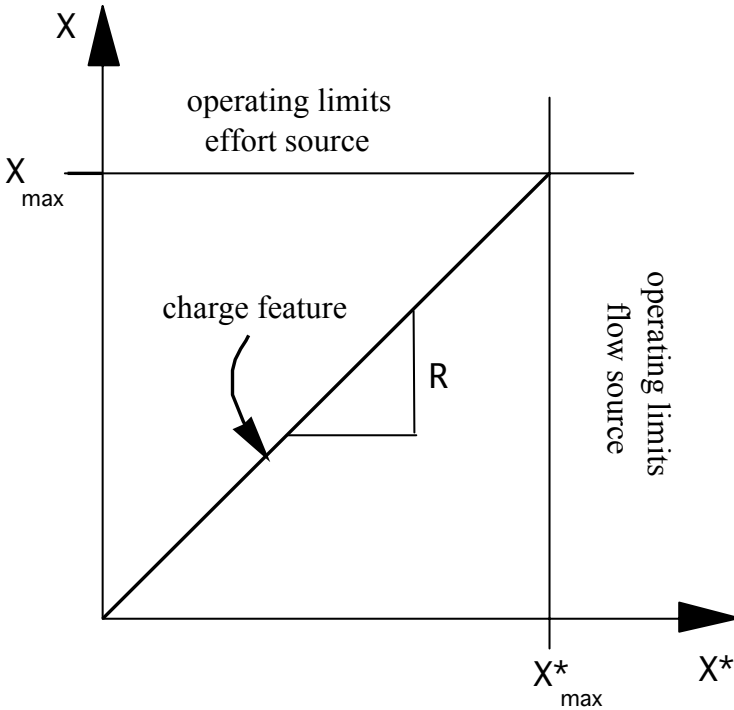
Less common examples are:

- *energy = load × tension;*
- *power = absolute temperature × entropy.*

The energy used by a measurement system may be finite and is therefore an energy transfer system (the balance carried by a mass) or it may be indeterminate and thus is a power transfer system; this is the case with voltmeters and wattmeters. The first is an energy transfer system; the second, an example of a power transfer system.

### **1.7.5. Features and operating points of a system**

In both linear and non-linear examples, the course taken by a flow variable or effort variable expresses the energetic limits of the system and determines an optimal operating point.

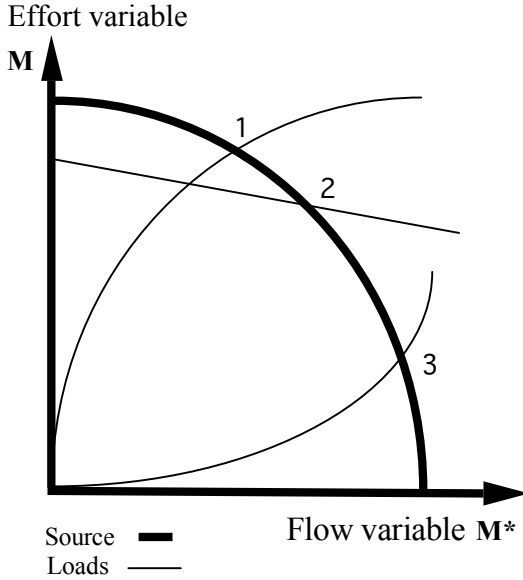


**Figure 1.5.** Operating limits in an example of a linear load

In general, a loaded source cannot operate when simultaneously emitting a flow variable and a maximum effort (see Figure 1.5). For example, a battery cannot both supply a maximum current and a different maximal potential to a charge.

Both the source feature and the load feature share one or several points of intersection (see Figures 1.6 and 1.7). These are operating points, determining the variable values and the load that permits their connection. From an energetic point of view, two conditions must be met:

- the continuity of shared flow variables;
- the continuity of shared effort variables.



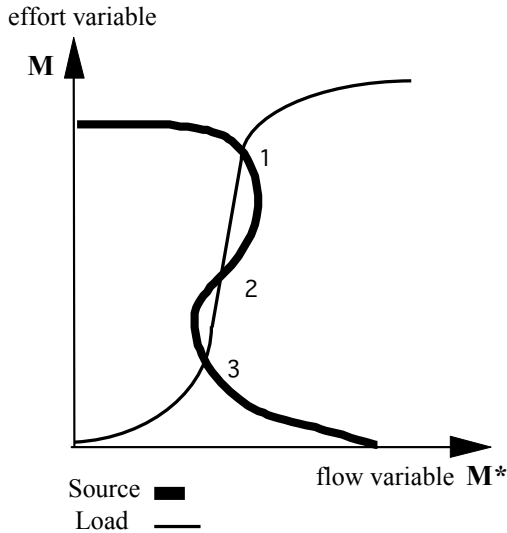
**Figure 1.6.** A source feature intersecting with different loads: all the operating points (1, 2 and 3) are stable

These conditions characterize operating points that are stable (see Figure 1.6) if the source and load features are simple (that is, linear, quadratic, logarithmic and so on) or unstable (see Figure 1.7) if the features are complex (curved, convex).

**1.7.6. Generalized impedance**

Determining the load effect (see Figure 1.4) makes use of the concepts of impedance and generalized admittance. In non-linear cases, the derivative in relation to the flow variable can be used but we will not discuss these cases here. We define the concept of impedance as a relation between intervening quantities in a power exchange. It is a specific transfer function of the system. The relation of the derivatives intersecting the associated variables is the determining factor:

$$Z = d(\text{Effort variable})/d(\text{Flow variable})$$



**Figure 1.7.** Intersections of a source feature with a load. Operating points (1, 2 and 3) are unstable

We use this concept in cases when the measurand is an effort variable. Going beyond equations that fit the model given in Figure 1.4, we here define two forms of generalized impedance that apply to cases of a power transfer system and an energy transfer. These are generalized resistance and generalized rigidity<sup>4</sup> shown in Table 1.3.

	Power transfer $X(t).X^*(t)$		Energy transfer $\int X(t).X^*(t)dt$	
Input variable	Associated variable	Generalized impedance $Z = \text{Var.Ext}/\text{Var.Int}$	Associated variable	Generalized impedance $Z = \text{Var.Ext}/\int \text{Var.Int} dt$
Stress variable $X(t)$	Flow variable $X^*(t)$	Generalized resistance $R = X/X^*(t)$	Flow variable $X^*(t)$	Generalized rigidity $S = X/\int X^*(t)dt$
Tension $U$ [V]	Electrical current $I$ [A]	$U/I$ [ $\Omega$ ]	Electrical charge $Q$ [C]	$U/\int Qdt$
Force $f$ [N]	Speed $v$ [m/s]	$f/v$ [ $N/ms^{-1}$ ]	Displacement $d$ [m]	$f/\int d dt$
Pressure $P$ [N/m <sup>2</sup> ]	Flow volume $D$ [m <sup>3</sup> /s]	$P/D$ [Nm/rad/s]	volume $V$ [m <sup>3</sup> ]	$P/\int D dt$

**Table 1.3.** Examples of interactions between effort variables, flow variables and corresponding generalized impedances in the case of the measuring object ( $X, X^*$ )

<sup>4</sup> The terms resistance and rigidity come from electronics and mechanics terminologies.

### 1.7.7. Determining the load effect

We can formulate two possibilities for a measurement system when the measurand is an effort variable:

– if the system is a power transfer  $\{P = M(t).M^*(t)\}$ , the generalized impedances reduced to a generalized resistance  $R = M/M^*(t)$  and the equation linking the measurement system and the object of measurement is:

$$M(t) = X(t) - R.X^*(t)$$

– if the system is an energy transfer  $\{W = \int M(t).M^*(t)dt\}$ , the generalized impedance is reduced to a generalized rigidity  $S = M / \int M^*(t)dt$  and the equation linking the measurement system to the measured object is:

$$M(t) = X(t) - S. \int X^*(t)dt$$

This gives us four possible cases according to which we consider that the quantity to be measured  $\mathbf{X(t)}$  can be an effort variable or a flow variable. When the measurement system is connected then  $\mathbf{X^*(t)} \neq \mathbf{0}$ ; if it is not, then  $\mathbf{X^*(t)} = \mathbf{0}$ . Generally,  $\mathbf{X(t)}$  depends on  $\mathbf{X^*(t)}$  and the relation between them differs according to whether they are viewed as the object to be measured or the measurement instrument. This is most simply and most often expressed as a linear relation.

To estimate the load effect, we write the equation linking the exact value  $X(t)$  and  $M(t)$  as:

$$X(t) = M(t) (1 + R/R_m)$$

$$\text{or } X(t) = M(t) (1 + S/S_m)$$

where  $R$  and  $R_m$  are the generalized resistances respectively of the measured object and of the measurement system, with  $S$  and  $S_m$  being both the generalized rigidity of the measured object and the measurement system, respectively.

If we want  $M(t)$  to tend towards  $X(t)$ , then we have  $(R_m \gg R)$  or  $(S_m \gg S)$ .

**1.7.8. Measurement with a car battery**

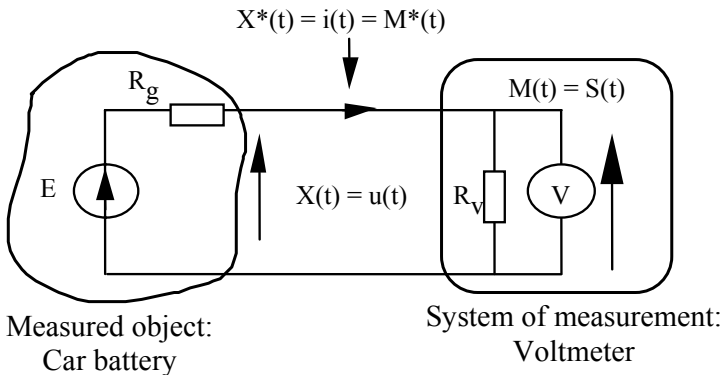
If we want to describe how the battery of a vehicle functions, we pay special attention to measuring its tension limits. Even in simple cases, this helps us describe certain aspects presented later in this chapter and, in particular, evaluate the load effect introduced by linking the measured object with the measurement system.

Modeling as a Thévenin source presents the measured object as made up of an electromotive force  $E$  (the measurand) and an internal resistance  $R_g$ . The voltmeter  $R_v$  is a measurement system or device allowing access to the measurand  $E$ . We model these according to Figure 1.7.

*Analysis of the load effect:* the pair of linked variables is made up of an effort variable  $X(t)$  that is the tension limit of the battery  $u(t)$  and an accompanying flow variable  $X^*(t)$  that is the current  $i(t)$ . The resulting measurement, made by a voltmeter, would ideally be  $M(t) = E$ .

If we express the given equations by the object of measurement, we derive:

$$u(t) = E - R_g \cdot i(t)$$



**Figure 1.8.** Contact of an instrument with a measured object and associated variables

The measurement system's tension limit (here the system is a voltmeter) is expressed by  $M(t) = S(t) = R_v i(t)$  where  $i(t)$  is the shared flow variable for the instrument and the measured object:  $M^*(t) = i(t) = M(t)$ . From these we derive:

$$E = u(t)(1 + R_g/R_v)$$

The load effect is thus represented by the term  $(R_g/R_v)$ . If we want  $u(t)$  to be equal to  $E$ , voltmeter resistance must show high resistance to the internal resistance of the battery; this is in fact the usual result.

The load effect is thus intrinsic to all measurement operations and as such is inevitable. We might want to think that the load effect is unimportant in view of the fact that it is responsible for easily tolerated errors. However, this is not always the case, and practical solutions (such as experimental precautions) and theoretical solutions (data treatment) are necessary to properly understand and analyze results.

### **1.7.9. Determining impedances**

We deduce impedances independently, then combine them according to the usual rules for combining functions (as for series and parallels). The problem is in determining the partial impedances of every source in the involved subsystem. In linear examples, we determine the transfer function between the effort variable and the flow variable at the point under study.

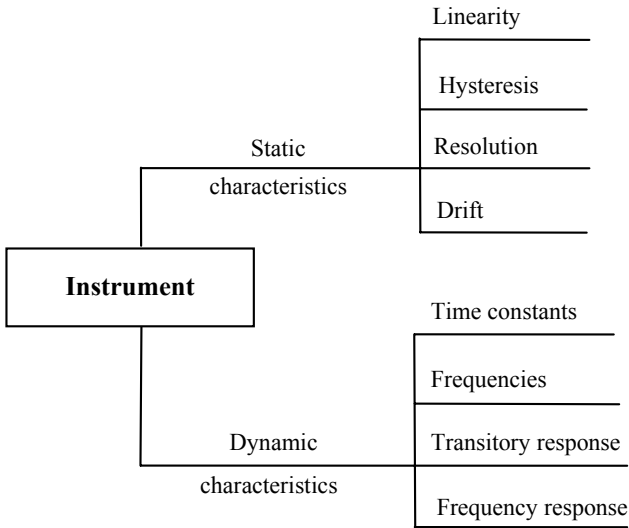
### **1.7.10. Generalized admittance**

Generalized admittance is, in electricity to cite one instance, the inverse of impedance. The definitions given above help us infer cases of generalized admittance. We will use this concept in cases of a measurand being a flow variable, just as in electronics where we usually apply the concept to parallel circuits.

## **1.8. Characteristics of an instrument**

To determine the design or choice of an instrument, we must consider the following three aspects:

- how we wish to use the instrument and for which purposes;
- whether it is an isolated system or connected to other systems;
- the features of the measurand and ease of accessibility.



**Figure 1.9.** *Principal characteristics of an instrument*

Responses are conditioned and numerous. We define the response of an isolated instrument by its static and dynamic characteristics (see Figure 1.9). These are a good starting point from which we may put into play a strategy for developing a measurement chain. These characteristics must be present in each part of the chain (especially in the transducer) in order to be certain of all characteristics of the system. The load effect produced by different connections along the chain can present problems. Certain provisional solutions, such as the use of impedances or adjusting of the linked parts can help, but usually we must compensate for the introduced errors through quantification or consider these errors in data analyses. The performance of each element is expressed as a transfer characteristic, which is often a complex data collection linking the output parameters of a system to its input parameters. This allows us to predict how a measurement system will perform by correctly combining the transfer characteristics of each of its elements. This operation will be efficient if all the transfer characteristics are expressed according to the same rules.

### **1.8.1. Components of static transfer functions**

Static transfer function characteristics are usually expressed in terms of parameter groups; we give some of these principles in Table 1.6. The relative



importance of these parameters depends on the metrological situation. Some of these parameters have several different names, according to the manufacturer. The units in which they are given tell us which parameters are in use.

Static calibration summarizes measurement system performances when all the input variables are maintained constant, excepting one which is varied by step by step. Output variables are collected according to steady-state functioning. These static transfer characteristics only make sense if static calibration conditions are established. In particular, variable values must remain constant, clear and exact.

### **1.8.2. *Dynamic characteristics***

Transfer characteristics do not easily combine when expressed in the form of transfer characteristics. One way of viewing an instrument is to see it as a black box with a known relation between the input (excitation) and the output (measured signal). This relationship is a transfer function  $S = f(s)$ . It can be linear ( $S = a + ks$ ) or non-linear (logarithmic, exponential, or polynomial). Often the gradient  $b$  is designated as sensitivity. In non-linear cases, sensitivity is not constant but is a variable that may be expressed at any point as  $x_0$  by  $k = d(S \text{ in } x_0)/ds$

### **1.8.3. *Instrument performance***

As we know the static and dynamic transfer characteristics of all the elements of a system, we can then combine them to obtain a description of the entire system, inferring these characteristics from the partial characteristics of various components described by their transfer functions.

There is no absolute rule for combining static parameters; each case requires different procedures. Often, contributions of most parameters may be negligible except those corresponding to a specific element. For dynamic characteristics, we use transfer functions.

### **1.8.4. *Combining transfer functions***

Combining transfer functions of constituent elements presupposes:

- an absence of initial conditions for all the elements;
- an absence of a load effect of one element on another.

We can deduce conditions that existed prior to the combining of transfer functions. Actually, the output signal of the transfer function of an isolated element has neither power nor energy. In view of this limitation, the following rules provide ways of verifying the transfer function of an entire system:

- the transfer function of  $n$  elements in a series must be equal to the product of the transfer functions of each element;
- the transfer function in Laplace notation of a feedback loop (see Figure 1.2) is expressed by:

$$T'(p) = T(p) / (1 + B T(p))$$

## 1.9. Implementing measurement acquisition

In a specific situation, choosing an appropriate measuring procedure requires adequate knowledge of measurement methods and implementing them with the means available. The ultimate functioning limit of a measurement system always depends on the level of noise present. Experimental research is based on the working-out of planning methods for experiments. However, most bench scientists avoid the preliminary planning stage. Rather than finding an appropriate method that would guarantee the best data with a minimum of measurement, they accumulate results as quickly as possible. Choosing the right strategy is essential for economic reasons; it saves time and money. In addition, it guarantees measurement reliability. The right strategy is important not only before beginning an experiment, but during each stage of it; the analyses carried out during experiments may lead to changes in strategies [BOI 89].

The planning of an instrumentation chain may differ according to the application field. Many books discuss the main principles for the working-out of an instrumentation plan, depending on research objectives or the desired outcome. Certain principles useful to planning research experiments are important to instrumentation: measured factors, positioning of sensors, measurement frequency and data analysis [CER 90]. Selecting from these principles depends on several types of constraints, including minimization of measurement error; sensor size, reliability of some sensors, minimization of measurement noise and field constraints.

### 1.9.1. Principles and methodology of measurement

In recent years, there has been much progress in improving techniques of measurement, instrumentation and data analysis [PRI 95]. Before going on to

describe the necessary elements for building the correct instrumental chain for a given situation, we will discuss ways to approach measurement and the methodology of measurement.

Many different situations may present themselves during a measurement operation. With the goal of measuring quantity, ascertaining this quantity begins with designing, choosing, defining and implementing:

- a *measurement principle* that serves as a scientific base for a measurement method;
- a *measurement method* or a group of theoretical and practical operations usually implemented during measurement according to a procedure;
- an *operating mode* of measurement or a precise series of theoretical and practical operations implemented during measurement according to a procedure;
- a *measurement process* or the sum of data, equipment and operations relating to a given measurement.

A *fundamental method of measurement*. We define this as “a method of measurement by which the value of a measurand is determined by measurement of the base quantities”. All implemented measurement system strategies begin with the accessibility or lack of accessibility of a variable, as well as from the physical principles determining variable acquisition. Concerning variable acquisition, three results are possible:

- *accessible real parameters*: these include the temperature of an oven, a current going through an element (resistance), or a person’s height;
- *inaccessible real parameters*: some of these are dielectric permittivity and conductivity of the brain, the pH factor of Jupiter’s subsoil, real time thermal cartography of a plane taking off, and a person’s age;
- *inaccessible unreal parameters*: some of these include negative time, ECG readings, temperatures below absolute zero and negative frequencies.

In practice, we can define two classes of measurement:

- *direct methods of measurement*: with these methods, we directly obtain the value of measured variable rather than measuring other, functionally related variables;
- *indirect methods of measurement*: with these methods, the value of a measured variable is obtained by measuring other, functionally related variables.

From these two classes, we deduce three measurement principles, one direct, two indirect:

- *the principle of direct measurement*: here, the measured object serves as an energy vector, carrying information to the measuring system;

- *comparison method of measurement*: in this indirect system, energy is carried by an external auxiliary system;

- *substitution method of measurement*: this system has conditions of the two above systems.

The description of all the stages and operations leading to quantitative measurement and its analysis make up the *methodology of measurement* [BOIS 89]. We base method choice on constraint definitions and objectives according to whether or not an appropriate model exists or the measurement principle (defined or set).

All strategies devised to implement an instrumental chain depend on the range of knowledge, instruments and operations used during the measurement process. How these strategies are defined and implemented depends on many factors, among them the “physical” conditions of the measurement, field constraints, economic constraints, measurement objectives (direct or feedback acquisition) and the operator’s level of competence.

This diversity of situations (including methods, means, models, operating modes) means that instrumentation requires a rigorous analysis at each step of the process in order to determine a strategy for the measurement, leading to the best model for the given situation.

When such a model exists, several situations are possible:

- a well-proven body of experimental data validated by mathematical analysis permits the development of a sufficiently representative model of the phenomenon being studied. If this model is robust (representative), we use it according to intermediary verifications; critical analyses then confirm the results;

- if the model is not robust and therefore is perfectible, it is the result of an analysis of prior results. In this case, however, we cannot be certain that it is a reliable representation of the measurand. Its use will be limited and we should not extrapolate beyond the reliability limits of this kind of model. In such cases, measurements must be carried out with rigorous care, leading to a confirmation, an improvement or an extension of the model being used.

If no model exists, the need for rigor remains, with the operator experimenting until finding, identifying and naming a model. Measurement results can lead to finding a preliminary model that can be improved later by successive approaches.

### **1.9.2. Field measurement constraints: instrumentation on the road**

This section will analyze a measurement situation, again using the example of the car. Suppose we want to measure the axle load of heavy trucks and similar vehicles and lighter cars. Using a measurement instrument involves the following field constraints that may effect measurement acquisition:

- road sensors that must set up and not be interrupted;
- climatic conditions such as ice or heat, among others;
- pressure resistance and other mechanical constraints;
- influence quantities that may be hard to control, such as variations in the main sensors due to the heat of the road, for example.

There are other constraints on transmitting and analyzing measurement not mentioned above. A Danish firm has developed a measurement tool that reliably measures different axle loads as well as the number of axles. Two rows of sensors are mounted in each lane across the road. Since the distance between the two rows of sensors and the time lapse between the obtained signals is known, it is easy to determine the speed and, therefore, the distance between the axles, as well as the type of vehicle. This system takes into account field constraints in order to choose which sensors (these are made of quartz) to use.

This study, carried out on a highway under the control of the Danish National Road Association (DNRA), has shown conclusively reliable results. With older systems, the sensitivity of sensors to lateral pressure resulted in a “phantom” axle count, resulting in an overestimation of the number of vehicles. This example shows the role of technology and its influence in defining and implementing a measurement system.

### **1.10. Analyzing measurements obtained by an instrument**

How we analyze data depends on the range of principles used in measurement. Implementing any measurement chain requires a quantified or estimated definition of any errors that may occur; in this way, precautions in using materials and software can reduce errors, leading to the closest probable value of the measurand [PRI 89].

### 1.10.1. *Error reduction*

There are two main approaches to error reduction. These are *experimental solutions*, such as the use of impedance matching and noise reduction, and *statistical solutions* or signal analyses such as error calculation, data analysis, spectral analysis and others.

Experimental solutions are physical in nature and are closely related to influence variables and load effects. They are implemented in design before measurement occurs. Statistical solutions are mathematical in nature. They are part of the analysis and correction of results and are carried out after measurement has taken place, since a measurement must be made before it can be corrected. In the section that follows, we discuss different bases for data analysis and their definitions.

### 1.10.2. *Base definitions*

Once a measurement has been carried out, we obtain a *result of the measurement*, which we define as “the value of a measurand obtained by measurement”.

This is the *uncorrected result*, or “the result of a measurement before correction for assumed systematic errors” which is necessary to obtain the *corrected result*. The corrected result is “the result of a measurement obtained after having made corrections to the uncorrected result in order to take account of assumed systematic errors”.

Correction of measurement results mostly depend on resolving errors. One kind of error is an *absolute error of measurement*, defined as “the result of a measurand minus the (conventional) true value of the measurand”.

A *systematic error* is “the component of the error of measurement which, in the course of a number of measurements of the same measurand, remains constant or varies in a predictable way”.

Understanding these different types of errors helps us make a *correction*. This is “the value which, added algebraically to the uncorrected result of a measurement, compensates for an assumed systematic error”. We deduce from this the *accuracy of measurement*, or “the closeness of agreement between the result of a measurement and the (conventional) true value of the measurand” and the *experimental standard deviation*, defined as for a series of  $n$  measurements of the same measurand, the parameter  $s$  characterizing the dispersion of the results given by the formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

$x_i$  being the result of the  $i$ th measurement and  $\bar{x}$  being the arithmetic mean of the  $n$  results considered.

### *Description of results*

Displaying results in an equation must take into account *measurement uncertainty*. This is “an assessment characterizing the range of values in which is found the true value of a measured variable”.

These definitions help us display the results:

$$\mathbf{X} = \mathbf{X}_0 \text{ [S.I. unit]} \pm \Delta \mathbf{X} \text{ [S.I.unit]} \text{ (probability \%)}$$

To be rigorous, a measurement result must contain the most probable value. The uncertainty range will include this probable value and the probability associated with both this uncertainty range and the unit being used.

### **1.11. Partial conclusion**

No matter which measurement method we use, there is always an energy transfer between the measured object, the measurement instrument, the influence variables, and physical disruptions such as connections and electromagnetism. This means that measurement systems must be validated by estimating disruptions that occur from contact of the measured object with the measurement instrument. This allows for corrections of introduced errors.

### **1.12. Electronic instrumentation**

The above definitions are applicable to all fields such as mechanics, hydraulics and biology. However, in this chapter, and in the rest of this book, we will discuss only measurement chains based on electronic technologies. In scientific measurement and industrial measurement, the observation, interpretation and control are increasingly carried out by electronic instrumentation. The very important development of microcomputers has meant that a range of separate devices dedicated to one or several functions became part of the instrumentation process. A few of these are the voltmeter, the frequency meter and the oscilloscope. This

development has also led to the creation of multi-component, interconnected systems that can be controlled and upgraded, with parameters that can be customized – all by a computer [NAD 98].

We measure a signal to gain information. A signal is a physical variable with characteristics which vary in level or in time. Computer-driven interpretation requires that a sensor convert the signal by means of electric signal tension or an amplified current. With signals, we must find out what information it contains and in which form it is being transmitted. This information is taken by an instrument whose function is determined by the characteristics of the signal and the type of information it carries.

The normalization of measurement systems is based on the definition of electric signal classes. Since the variable to be measured is converted into the form of an electric signal, we must identify the nature of the signal. For this reason, the analyzed signal becomes an essential criterion in the implementation of method and acquisition systems (see Figure 1.11). There are numerical signals (carried by pulses and pulse paths) and analog signals that may be continuous, temporal or frequential. We deduce information from one of the following parameters: logical state, cadence, level, form, or frequency.

Below we present a simple classification, based on the type of fundamental signal information.

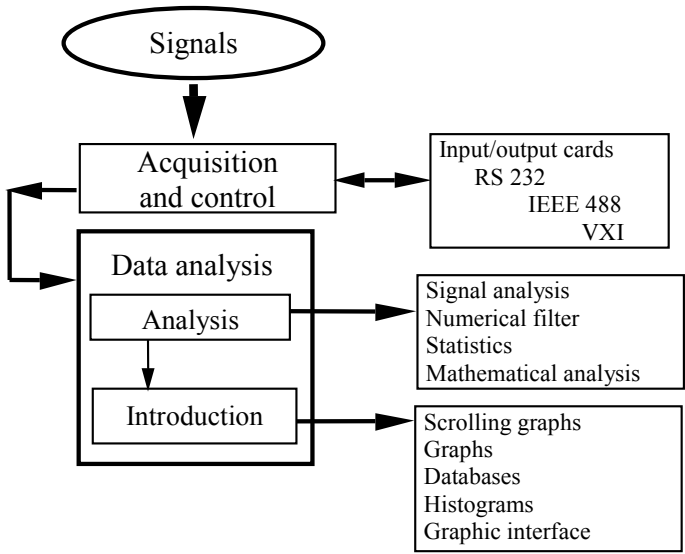
### ***Analog signals***

*Continuous signals* are static or vary slowly. With these signals, the level or amplification of the signal at any instant constitutes the information. While sensors (such as the thermocouple) may be used for measurement of these signals, an analog-to-digital converter (ADC) converts the signal into digital data. The precision, resolution, reliable pass band, and good synchronization of the ADC ensure parameters essential to continuous analog signal acquisition.

There are many kinds of *temporal signals*, such as ECG waves and temperature. Here, information is carried in the form of waves (amplitude and variation in time). Temporal signal acquisition means using the largest pass band possible and a precise time base (to avoid sampling problems), ensuring transfer speed, as well as beginning and ending the measurement sequence correctly.

*Frequential domain signals* contain information in signal frequency variations. Analyzing this type of signal involves converting the measurement into frequential data. Increasingly, specialized processors carry out this analysis, which includes Fourier's analysis functions.





**Figure 1.10.** *The signal is a definition criterion of an instrumentation procedure*

**Digital signals**

*Binary signals and pulse trains:* to carry out accurate measurements of numerical signals, an instrument must generate binary signals (an example is the start-stop mechanism of machines) and pulse trains (as with sequencing clocks and synchronization). These measurements are carried out by means of a counting function.

These classes of signals are not mutually exclusive. A signal can contain more than one type of information. This is especially true with transient and permanent states of second order oscillatory systems that transmit signals through microwave lines. Instruments to measure these signals range from the simple (logic-state detectors for TOR signals) to the complex (frequency analyzers for spectrum analyses).

**1.13. Electronic instrumentation functionality**

A simple classification of electronic instrumentation includes:

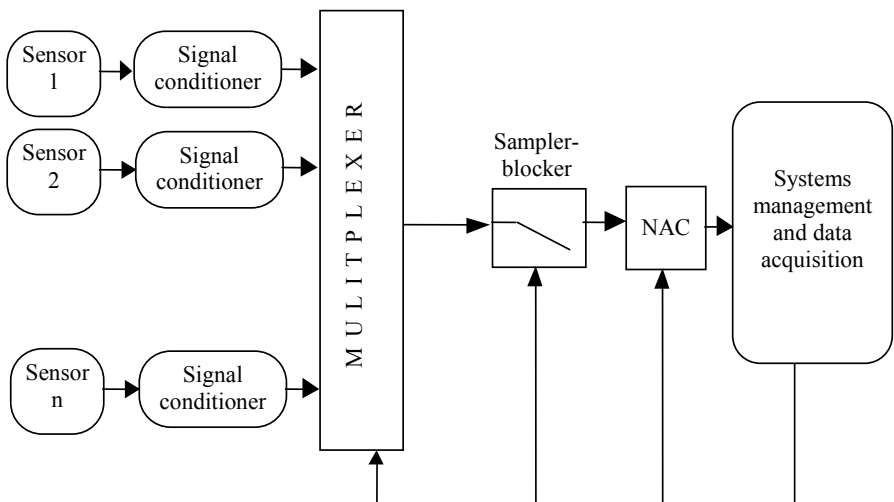
- sensors associated with electronic conditioning. Groups of these make up an autonomous instrument. This kind of instrument gives a directly usable measurement; they can also be combined with other groups containing several more

sensors. This means a pressure sensor can act as a surveillance device that gives instructions to an actuator; it can also be part of a network of several sensors supervised by a main microcomputer;

- instruments configured around a microcontroller. Another important field of electronic instrumentation includes measurement and control chains configured around an intelligent circuit. This pertains to many systems dedicated to a specific application characterized by reduced size, autonomy and reduced cost. In everyday living, this type of instrument plays a role in machines such as cars and household appliances, mostly because of their low cost, portability or small size;

- programmable electronic instruments. These are groups of instruments that have been configured to carry out customized functions according to the operator's needs. They are directed by computer instrumentation software.

The three types of instrumentation listed above depend on electronic measurement chains. There are two main classes of these: analog acquisition chains and digital acquisition chains. Usually, an analog chain can be represented by a functional block design which we show in Figure 1.11.



**Figure 1.11.** *An analog measurement chain*

Digital chains correspond to systems configured around a microprocessor or around programmable instruments [TRA 92]. We mention here the use of the

general functions of such chains in architecture; in later chapters, especially in Chapters 7 and 8, we will discuss digital chains in more detail.

### 1.13.1. Programmable instrumentation

Many measurement systems currently in use are based on programmable microelectronic instruments that are configured around a microcomputer. This type of modern instrumentation uses four combinable approaches, described in Figure 1.12 [STE 93].

The growth of these systems since the end of the 1980s has occurred because instrument manufacturers, using SCPI norms, have developed standards to ensure product quality, and because personal computers have improved and become more economical.

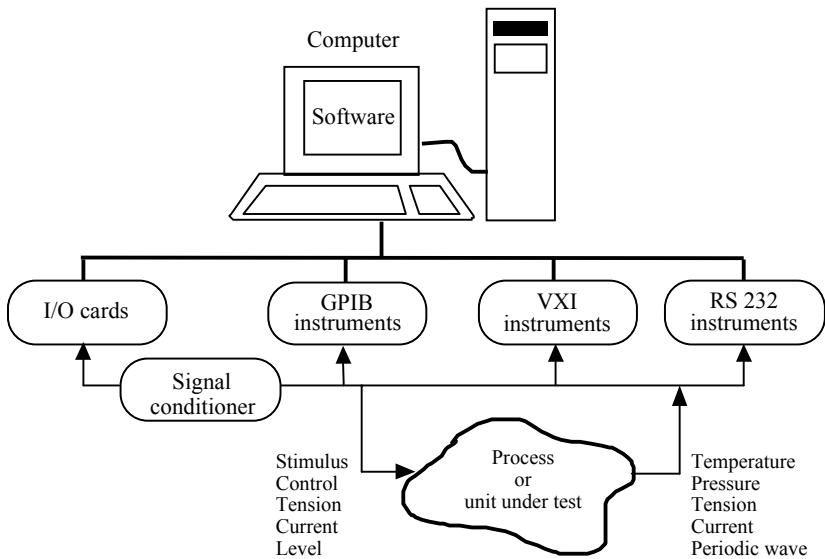


Figure 1.12. Programmable instrumentation

Modern instrumentation uses four methodologies to carry out signal acquisition. These are input/output cards, parallel type interfaces (IEEE488), series type interfaces (norm RS 232), interfaces using norm VXI (*Versa Module Eurocard Extended to Instruments*) and their derivatives. Depending on which methodology is chosen, signals are converted to data according to a normalized and programmable format of ASCII, binary or SCPI. Data are then analyzed and recorded. This data

processing treatment is important to correct instrumentation. “Virtual” instruments have become possible due to increasingly user-friendly software that aids in developing measurement chains; in addition, such software has meant there are more complex situations requiring analysis. The aesthetic aspect of screens and contours of computers is secondary to the need for careful analyses of results.

### ***1.13.2. Example of an electronic instrument: how a piezoelectric sensor detects rattle in an internal combustion engine***

The growing use of electronics in the automotive industry has given rise to the need for operational monotension amplifiers that can function with a supply tension of +5 V. The example given here has been taken from an application note published by Texas Instruments. In this note, a piezoelectric pressure sensor interfaces with a circuit integrated with operational amplifiers to detect rattling in an internal combustion engine.

The *piezoelectric sensor* operating in alternating current. Its electric model consists of an electromotive source in series with a capacitor. It operates in two modes: in one the sensor produces an alternating tension, and in the other, the sensor produces a charge.

The *conditioner* is an interface circuit made of two operational monotension amplifiers. The first amplifier delivers an amplification signal (gain 5) in broadband (530 Hz to 28 Hz); the second is configured in a pass band filter.

*Limiting load effects:* in order to limit the load effect between the sensor and the conditioner, a resistor with a high level of power is inserted between the two devices. The resistor ensures that polarization currents flow from the sensor. The operational amplifier must have a high driving point impedance to be adapted to the sensor and very weak polarization currents. These steps require a very precise signal conditioning.

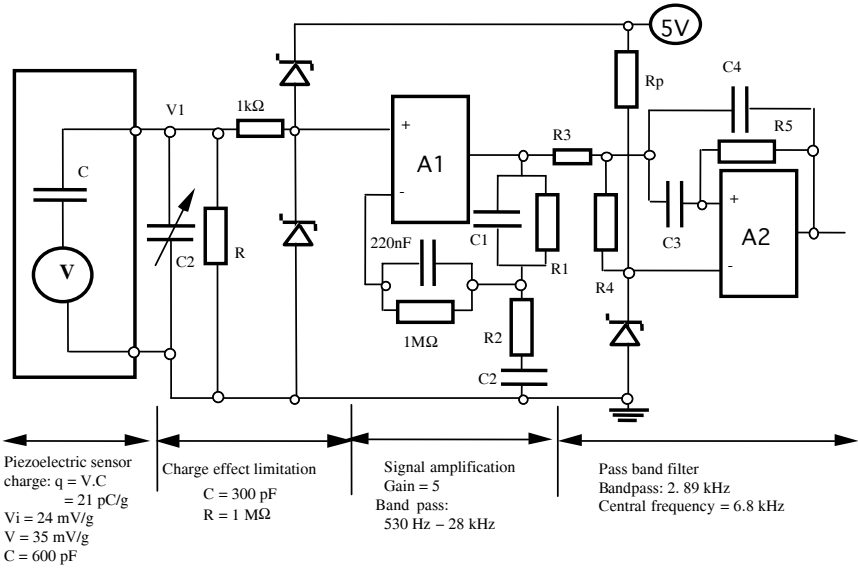


Figure 1.13. Example of an electronic instrument in a car

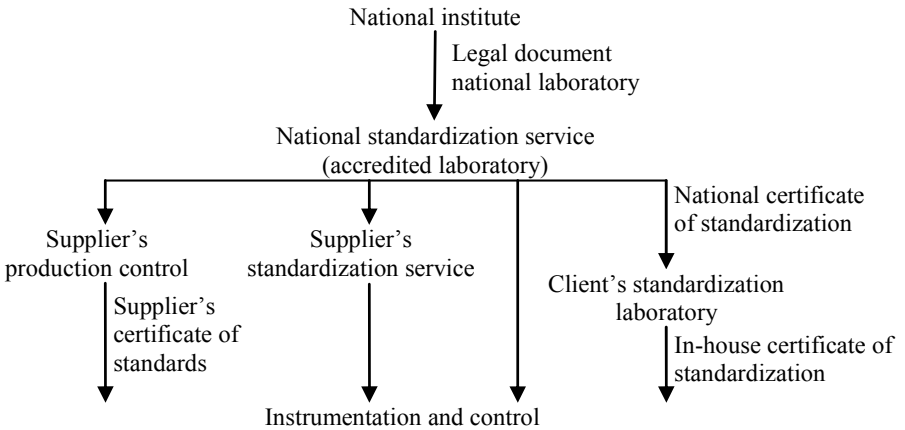
Functioning principles (see Figure 1.13): when rattling begins, the sensor produces a series of signals with a frequency that does not occur when the motor is functioning properly. An operational amplifier (OA) amplifies these signals. The frequency of the rattling varies according to the size of the cylinders and block cinder material. A broadband sensor and a very flexible operational amplifier must be used to be adaptable to all vehicles. Most sensors have band pass of several tens of kHz, meaning that one type of sensor is suitable for many applications.

**1.14. The role of instrumentation in quality control**

Throughout the world, the label “ISO 9000” has become a point of reference for companies wishing to maintain, guarantee and record the quality of their products through quality control [DEA 91]; [HOF 83]. These businesses need measurement instruments and standardized tests based on national norms.

Both the manufacturer of products to be standardized and certified and the client, who may have access to a laboratory with standardization capabilities, must conform to certain norms, depending on the quantities involved.

Overall, instrumentation and measurement play a crucial role in commercial quality control. There are instruments ensuring product *traceability* at every step of the production process. Traceability is “the property of a measurement whereby it can be related to appropriate standards, generally international or national standards, through an unbroken chain of comparisons”.



**Figure 1.14.** Role of instrumentation in quality control

To give an example, the sensors installed on a line can be calibrated with the supplier, but some must be calibrated *in situ*. Measurement is thus a key element in quality assurance. Figure 1.14 illustrates the positions occupied by the different levels of instruments of measurement and control in quality assessment processes in organizations.

### 1.15. Conclusion

The objective of this first chapter has been to introduce, through a comprehensive review, many of the issues involved in carrying out implementation. As well as the definitions given throughout, we have looked at how measurement, in all its diverse aspects, is a multidisciplinary science, drawing on mathematics, technology, and physics – all necessary to design and careful implementation.

1.16. Appendix

Transduction type		
Physical	Chemical	Biological
Elastomagnetic	Electrochemical process	Test effect on an organism
Electromagnetic	Spectroscopy	Spectroscopy
Magnetoelastic	Physicochemical transformation	Biophysical transformation
Photoelastic	Chemical transformation	Biochemical transformation
Photoelectric	Photochemical transformation	
Photomagnetic		
Thermoelastic		
Thermoelectric		
Thermomagnetic		
Thermooptic		

Table 1.4. Examples of possible instrument classification according to transduction type

Excitation	Operating physical variables
Acoustic	Amplitude, phase Spectrum Wave speed
Chemical	Components (concentration, states, etc.)
Biological	Biomass (concentration, states, etc.)
Electrical	Charge, current Potential, potential difference Electric field (amplitude, phase, polarization, spectrum) Conductivity, permittivity
Magnetic	Magnetic field (amplitude, phase, polarization, spectrum) Magnetic flow, permeability
Optic	Wave: amplitude, phase, polarization, spectrum Speed, wave length
Thermal	Temperature, thermal flow Specific heat Thermal conductivity
Mechanical	Position (linear, angular) Speed, acceleration Force, pressure Constraints, density mass Time Flow speed Form, hardness, orientation Viscosity
Radiation	Nature or type Energy Intensity

Table 1.5. Examples of possible classifications according to excitation type

Characteristics	Definitions
Zero offset	Zero offset is true relation of the zero output variable with the value of the measurand.
Drift	Temporal variations in system characteristics.
Dynamic	Admissible intervals of variation for input variables (in decibels).
Hysteresis	Maximum difference in output values, when the input variable is reached from minimum, then maximum admissible in algebraic value.
Linearity	Degree of concordance between the static state diagram and a straight line used as reference. (A straight line of the fewest squares calculated on calibration points, the line joining the farthest points throughout the measurement.)
Relaxation	Time lag between the cause and effect of a physical phenomenon, given in the form of a time constant.
Repeatability	Margin of fluctuation in output variable when the same input variable is applied several times under the same conditions.
Resolution	Smallest increase in the input variable leading to a change in the output variable.
Sensitivity	Ratio of change in output variables to the corresponding change in input variables.
Threshold	Threshold resolution is the smallest change of the input variable relative to zero value.
Response time	For a measurable excitation, this is the time required for an immediate value and a final value to be lower than a specified value (1%, for example).

**Table 1.6.** *Static characteristics*

## 1.17. Bibliography

- [ASC 87] ASCH G. *et al.*, *Les capteurs en instrumentation industrielle*, Dunod, 1991.
- [ATK 87] ATKINSON J.K., "Communication protocols in instrumentation", *J. Phys Sci. Instr.*, 20, p. 484-491, 1987.
- [BOI 89] BOISSEAU J.F., *Méthodologie de la mesure*, Techniques de l'Ingénieur, Mesures et contrôle, R140, 1989.
- [CER 90] CERR M. *et al.*, *Instrumentation Industrielle*, Techniques et Documentation, Lavoisier, 1990.



- [COM 92] COMMIOT D., “Les miracles de la mesure”, *Usine nouvelle*, 2373, p. 10-14, 1992.
- [DEA 91] DEAN, “Measurement, quality and trade”, *Meas. Sci. Technol.*, 2, 403-404, 1991.
- [DRA 83] DRAHEIM H., “Measurement as Science and Technology”, *Measurement*, 1, p. 68-74, 1983.
- [EST 95] ESTEVE D., COUSTRE A., GARAJEDAGUI M., *L'intégration des systèmes électroniques dans l'automobile du XXI siècle*, ouvrage collectif, CEPADUES Editions, 1995.
- [FIN 82] FINKELSTEIN L., “Theory and Philosophy of measurement”, in *Handbook of Measurement Science*, Wiley Interscience Publications, 1982.
- [FRA 96] FRADEN J., *Handbook of Modern Sensors*, 2<sup>nd</sup> Edition, Collection Engineering/electronics, Springer-Verlag, 1996.
- [GIA 89] GIACOMO P., Etalons métrologiques fondamentaux, Techniques de l'Ingénieur, Mesures et contrôle, R50, p. 1-16, 1989.
- [HEW 90] HELLWIG H., “The Importance of Measurement in Technology-Based Competition”, *IEEE Trans. on Instr. and Meas.*, 39/5, p. 685-688, 1990.
- [HIM 98] HIMBERT M., “La métrologie: un langage universel pour la science et la technologie”, in *Récents progrès en génie des procédés*, vol. 12, Lavoisier Tech & Doc, Paris, 1998.
- [HOF 83] HOFMANN D., “Modeling of errors in measurement”, *Measurement*, 1, 3, p. 125-128, 1983.
- [JAC 90] JACOMY B., *Une Histoire des Techniques*, Collection Points, Editions du Seuil, 1990.
- [LAF 89] LAFAYE P., Unités de mesure, Techniques de l'Ingénieur, Mesures et contrôle, R23, 1-13, 1989.
- [MAS 90] MASI C.G., “So, You Want to Measure Submicron Dimensions”, *Test & Meas. World*, p. 59-68, 1990.
- [NAC 90] NACHTIGAL C.L., *Instrumentation and Control*, Wiley Interscience Publications, 1990.
- [NAD 98] NADI M., “Rôle et évolution des systèmes de mesure électroniques”, in *Récents progrès en Génie des procédés*, vol. 12, Lavoisier Tech & Doc, Paris, 1998.
- [NAD 99] NADI M., “La mesure et l'instrumentation dans la recherche scientifique et dans l'industrie”, *Revue de l'électricité et de l'électronique*, no. 3, p. 38-42, March 1999.
- [NEU 89] NEUILLY M., Incertitudes de mesure et tolérances, Techniques de l'Ingénieur, Mesures et contrôle, R280, 1-21, 1989.
- [PAR 87] PARATTE P.A., ROBERT P., *Traité d'électronique et d'électricité, Systèmes de Mesure*, Dunod, 1987.

- [PRI 89] PRIEL M., Incertitudes de mesure et tolérances, Techniques de l'Ingénieur, Mesures et contrôle, R285, 1-11, 1989.
- [PRI 95] PRIEUR G., NADI M. *et al.*, *La mesure et l'instrumentation*, collection Mesures Physiques, Masson, Paris, 1995.
- [ROM 89] ROMANI L., *Structure des grandeurs physiques*, Blanchard Ed., Paris, 1989.
- [RON 88] RONAN C., *Histoire Mondiale des Sciences*, Collection Points, Editions du Seuil, 1988.
- [ROS 75] DE ROSNAY J., *Le Macroscopie*, Collection Points, Editions du Seuil, 1975.
- [STE 93] STENKON N., "Le standard VXI, un concept d'instrumentation électronique programmable", *Industronique*, 5, p. 54-56, 1993.
- [TRA 91] TRAN TIEN LANG, *Systèmes de mesures informatisés*, Masson, 1991.
- [TRA 92] TRAN TIEN LANG, *Electronique des systèmes de mesures*, collection Mesures Physiques, Masson, 1992.
- [TUR 90] TURGEL R.S., Instrumentation Everywhere, Feb. 1990, Vol. 39, no. 1, p. 1, 1990.
- [WHI 87] WHITE R.M., "A sensor classification scheme", *IEEE Trans. Ultrason. Ferroelec. Freq. Contr.* Vol. UFFC-34, no. 2, p. 124-126, March 1987.
- [WHI 93] WHITE R.M., "Competitive measures", *IEEE Spectrum*, 30, no. 4, p. 29-33, 1993.

*This page intentionally left blank*

## Chapter 2

# General Principles of Sensors

Sensors are the first components of a measurement chain. They convert the physical and chemical variables of a process or an installation into electrical signals that almost always begin as analogical signals. This conversion must also mirror as closely as possible the involved variables. Only a thorough knowledge of sensor responses guarantees success; sometimes sensors produce faulty signals due to interference, the conditions of use, or often because of the processes themselves.

We begin this chapter by discussing some of the basic principles of sensors and how they work [NOR 99]. These principles are based on calibration, evaluation of uncertainties, calculation of response time, and conditioning. Our aim is to provide the reader with a fairly general guide. Some relevant equations and formulae, as well as many issues relating to instrumentation and signal analysis, will be discussed in later chapters.

### **2.1. General points**

#### **2.1.1. *Basic definitions***

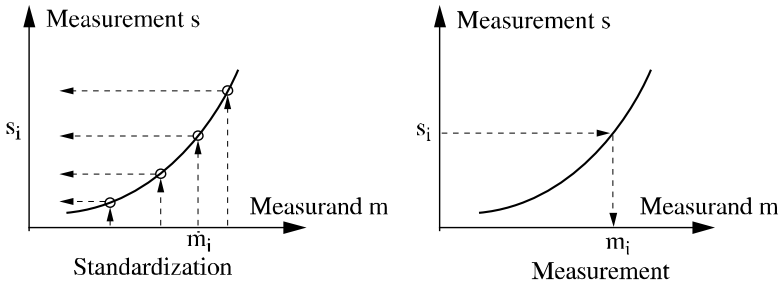
The quantity to be measured being the measurand, which we call  $m$ , the sensor must convert  $m$  into an electrical variable called  $s$ .

The measurement  $s$  can be an impedance, an electrical charge, a current, or a difference of potential. The relation that joins  $s$  to  $m$  can be called  $s = F(m)$  and depends on:

- the physical law determining the sensor;
- the structure and purpose of the sensor;
- the sensor's environment.

The expression  $F(m)$  is established by calibration. By using a standard or unit of measurement, we discover for these values of  $m$  ( $m_1, m_2 \dots m_i \dots$ ) electrical signals sent by the sensor ( $s_1, s_2 \dots s_i \dots$ ) and we trace the curve  $s(m)$ , called the sensor calibration curve (see Figure 2.1).

To use a sensor, we read the value of an electrical signal  $s$  when an unknown measurand  $m$  is applied. The calibration curve helps us to deduce  $m$ .



**Figure 2.1.** *Calibrating and reading a sensor*

We call sensitivity  $S$  the derivative  $ds/dm = F'(m)$ . In order to make sensitivity independent of the value  $m$ , the sensor must be linear:

$$s = S m + s_0 \tag{2.1}$$

where  $s_0$  is the value of the signal  $s$  when  $m = 0$ . Finally we have:

$$F'(m) = \text{constant} = S$$

Of course, we can always define a range of values in which  $S$  is constant: that is, when the sensor is linear.

### 2.1.2. *Secondary definitions*

It is important in what follows to remember some frequently used definitions given here that pertain to sensors in general:

- active and passive sensors and conditioners: the delivered electrical signals of passive sensors are impedance variations because these sensors require an electrical energy source in order to read s. Passive sensors are part of a circuit called the conditioner. All other sensors are active;

- measurement chains: in general, the signal cannot be used directly. We call a measurement chain the range of circuits or devices that amplify, adapt, convert, linearize or digitalize a signal before output readings;

- test specimen: especially in mechanics, converting m into s is not a direct process. For example, measuring a force means applying it to a deformable solid monitored by a deforming sensor. This deformable solid usually consists of all bodies found between the sensor and the measurand. These are called the test specimen;

- calibration: we distinguish between calibration carried out on the sensor itself and the global calibration carried out on all the test specimens, the sensor, the conditioner and the measurement chain;

- influence variables: the function  $F(m)$  often depends on physical variables in the environment, such as temperature or humidity. These variables are called influence variables;

- drift of sensor and response time: sometimes we find a specific instance of an influence variable that plays a role in measurement in one of two ways. It may cause long-term drifts that modify  $F(m)$ . This influences the drift of a sensor. Alternatively, the influence variable may modify the sensor's capacity to respond to measurement variations in time. This variable affects response time;

- band passes: when the sensor measures a measurand whose temporal dependence is sinusoidal, the sensor's sensitivity depends on the frequency of the measurand. The frequency range in which the sensor shows a constant sensitivity is called the band pass. Response time and the band pass are closely related.

## 2.2. **Metrological characteristics of sensors**

In this section we will not present methods measuring uncertainty. Here, we only discuss some general ideas to guide the reader through later chapters where uncertainty evaluation will be discussed on more detail, particularly in Chapter 10. Measurement uncertainty is the difference between the true value of the measurand and the measurement carried out by the sensor. The only known measurands are the standards with values that have been determined by convention. It is important to

distinguish between systematic errors and random uncertainties: they occur for different reasons and have very different consequences for measurement.

### **2.2.1. Systematic errors**

Systematic errors are always due to faulty sensor knowledge or utilization. This kind of error is detected by comparing the mean values of the same measurand, given by two different sensors. The most frequent causes of systematic errors are:

- incorrect or non-existent calibration due to an aging or altered sensor;
- incorrect usage: some examples include failure to reach steady state, a defect or error of one of the conditioner parts, or a modification of the measurand by the sensor itself;
- inadequate data processing: examples are error in linearization in the measurement chain or amplifier saturation in the measurement chain.

Obviously, detecting systematic errors leads to their elimination.

### **2.2.2. Random uncertainties**

We can know the cause of random uncertainties without being able to anticipate the measurement value. Their evaluation is always statistical. They are due to the presence of signals or interference; the amplitude of these is random and they are called, rather vaguely, “noise”. To cite a few examples:

- fluctuating supply sources in the measurement chain or in the conditioner (such as fluctuation of the electromotive force in a bridge);
- electromagnetic signals produced in an environment and picked up by a sensor element, conditioner or measurement chain;
- thermal fluctuation, including thermal turbulence of current carriers;
- fluctuation in influence variables, etc.

There are many other causes of random uncertainties, such as reading errors, defects in sensor mobility and hysteresis. Unlike systematic errors, random errors can never be completely avoided. We can, however, reduce them by using protection methods, such as electrical regulations, temperature stabilization, mechanical isolation and electromagnetic shields. In addition, filtering, synchronous detection and signal processing can reduce random uncertainties, which must always be evaluated carefully.

### 2.2.3. Analyzing random errors and uncertainties

Because, by definition, random errors cannot be anticipated, they become part of statistics and as such are labeled “uncertainties”. Because of this vagueness, statistics can prove a very useful tool. The British Prime Minister Benjamin Disraeli once said to Queen Victoria: “There are three kinds of lies: lies, damned lies, and statistics.” This witticism demonstrates why governments, for example, like using statistics. Statistics cannot anticipate a specific occurrence, such as a measurement result, but only give the probability of one value among many occurring. Statistics represent a crowd and not an individual; a forest, not a tree. That is why, Disraeli might have added, governments use statistics to direct and support as many actions as possible, without necessarily having to address or even be aware of specific problems. In the case of measurement, two statistical problems must be analyzed [DIE 92]:

- the evaluation of uncertainties;
- deciding how to analyze and treat these uncertainties.

#### 2.2.3.1. Evaluating random uncertainties. Standard deviations. Variances

Suppose we make  $n$  measures  $s_1 \dots s_i \dots s_n$  of the same mesurand  $m$ . We call the mean value of  $s$  the quantity  $\bar{s}$  so that:

$$\bar{s} = \frac{\sum_{i=1}^n s_i}{n} \quad [2.2]$$

The mean value of the difference  $s_i - \bar{s}$  is used to statistically analyze the random measurement of  $m$ . The variance  $v$  and the standard deviation  $\sigma$  must be introduced:

$$v = \frac{\sum_{i=1}^n (s_i - \bar{s})^2}{n - 1} \quad [2.3]$$

$$\sigma = \sqrt{\frac{\sum (s_i - \bar{s})^2}{n - 1}} = \sqrt{\frac{n \left[ \overline{(s^2)} - (\bar{s})^2 \right]}{n - 1}} \quad [2.4]$$

Suppose that the result of 10 temperature measurements are (in °C): 60.1; 60.6; 59.8; 58.7; 60.5; 59.9; 60.0; 61.2; 60.2; 60.2.

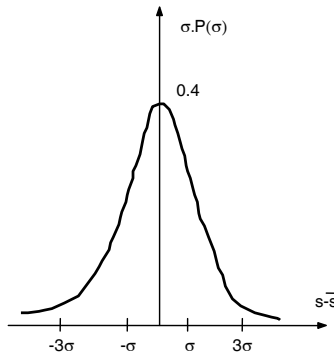


First of all, we present the results with one number after the decimal point. This shows that the uncertainty of these measurements is at best 0.1°C. We also find that the mean temperature is 60.11°C, and that the standard deviation  $\sigma$  equals 0.63°C. The mean value of 60.11°C does not mean that it is close to the true value of the measurand. A systematic error, perhaps in the form of defective measurement standards, could have produced it, possibly through a deviation of 1°C on all the values and therefore on the mean value. However, if only the systematic error is large in relation to the random uncertainty and the sensor is non-linear, the deviation type does not change after correcting the systematic error. The list of ten values will not be reproduced if we undertake another measurement, but this second list will still have something in common with the first. Looking again at the first list, we see that deviations between 0°C and 0.2°C occur seven times. There is one value that deviates more than 1°C from the mean value. Therefore, we can classify the obtained values by their occurrence probabilities. After many measurements have been carried out on the same measurand, if the uncertainty is truly random, we can demonstrate that this occurrence probability is a law called the Gaussian distribution. It expresses, according to the mean value and the deviation type, the probability density of finding the value  $s$  of a measurement. With the Gaussian distribution, the basic probability of finding the range of  $s$  through  $ds$  is given as  $dp = p(s) ds$  (where  $p(s)$  is called probability density) is given as:

$$dp = p(s) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(s-\bar{s})^2}{2\sigma^2}} ds \tag{2.5}$$

The probability of finding any value for  $s$  clearly is equal to 1:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(s-\bar{s})^2}{2\sigma^2}} ds = 1 \tag{2.6}$$



**Figure 2.2.** Representation of deviation type by the Gaussian probability density

For  $s - \bar{s} = \pm 3\sigma$ , the integral of  $\int_{\bar{s}-3\sigma}^{\bar{s}+3\sigma} P(s) ds \geq 0.99$  which means more than 99% of all measurements lead to a value of  $s$  expressed as  $(s - \bar{s}) \leq 3\sigma$ .

We return to our series of ten temperature measurements with an average reading of  $60.11^\circ$  and a deviation type of  $0.63^\circ\text{C}$ . With a second series of ten measurements, we have:  $59.5; 60.2; 60.6; 60.1; 59.6; 58.9; 60.9; 59.2; 60.1; 60.3$ . The mean value is  $59.94^\circ\text{C}$  and the deviation type is  $0.62^\circ\text{C}$ . These two values are slightly different from those of the first series; but if instead of taking ten measurements, we had taken an infinite series, we would have obtained the mean theoretical value of  $\mu$ . We can see that as  $n$  becomes larger the closer  $\bar{s}$  is to  $\mu$ . But the meaning of “close” is still not clear. To be more precise, we introduce  $\Delta\mu$  as a confidence interval. For example, we can demonstrate that the probability of finding  $\mu$  in the interval  $[\bar{s} - \Delta\mu, \bar{s} + \Delta\mu]$  with  $|\Delta\mu| = \frac{1.65 \sigma}{\sqrt{n}}$  ( $\bar{s}$  having been calculated for the  $n$  measurements) is 90%. With our first series, we had  $\Delta\mu = 0.33^\circ\text{C}$  and with our second series  $\Delta\mu = 0.32^\circ\text{C}$ . We obtain for the first series of measurements a confidence interval of 90%, close to:

$$59.78 \leq \bar{T} \leq 60.44$$

and for the second series

$$59.63 \leq \bar{T} \leq 60.26$$

Of course, we can define other confidence intervals in choosing probability values of finding  $\mu$  different from 90%.

### 2.2.3.2. Decisions about random uncertainties

Sometimes, when carrying out a series of measurements on the same measurand, a result may occur that deviates significantly from the mean value obtained through prior measurements. In such a case, it is hard to know if this measurement was produced by a rare but important increase in the random uncertainty or by an unforeseen phenomenon that has causally modified the measurand. We must then decide if, in fact, the measurand has been modified.

In this instance, we use statistical results to make our decision. For example, Chauvenet's criteria specifies that a non-random phenomenon has modified the measurand if the probability of obtaining the measurand, calculated with the help of the Gaussian distribution, is less than  $1/2n$ , with  $n$  being the number of

measurements made up to the point when the anomaly appeared. We can tabulate the results of this criterion by giving the deviation limit  $d_{max}$  to the mean value; beyond this the criterion applies (see Table 2.1).

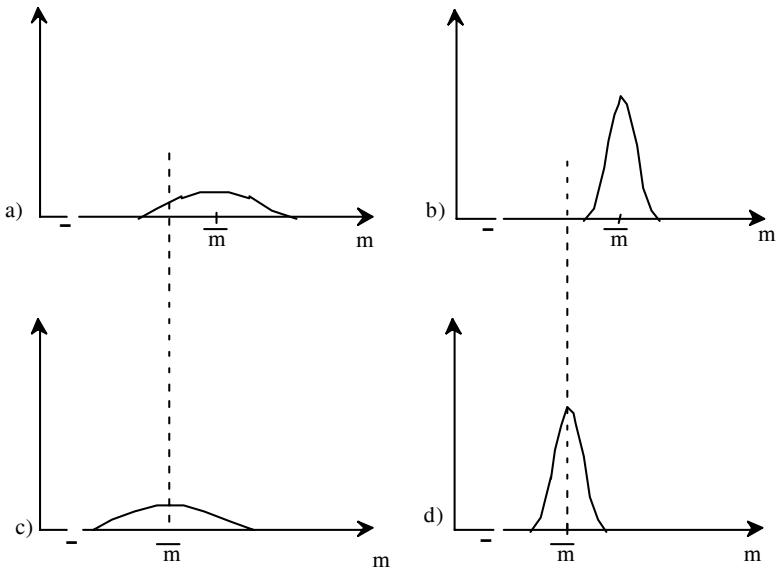
A way to use this criterion is to look at the threshold settings of an alarm system. Suppose that a presence detection has an analogical output of 0 V without any environmental interference. We can carry out 500 measurements and find a deviation type of 265 mV. If one measurement has a value superior to  $265 \times 3.29 = 872$  mV, something has modified the measurand – there has been some interference. We must then set the threshold of the system to the value of 872 mV, or set a rule:  $V < 872$  mV = no interference,  $V > 872$  mV = interference.

Number of measurements	$d_{max}/\sigma$
10	1.96
25	2.33
50	2.57
100	2.81
500	3.29
1,000	3.48

**Table 2.1.** Chauvenet’s criteria for a Gaussian distribution. Deviation from the mean value beyond which the engineer must consider if a measurand has been modified

2.2.3.3. Reliability, accuracy, precision

These three properties characterize sensor and sensor calibration. A sensor is reliable if its deviation type is weak, accurate if it has no systematic errors, and precise if it is both reliable and accurate. Figure 2.3 shows the derivative of the probability density in the four possible instances [ASC 91].



**Figure 2.3.** Reliability, accuracy, and precision. The dotted lines indicate the true value ( $\mu$ ). In a) the sensor is neither accurate nor reliable; in b) the sensor is reliable but not accurate; in c) the sensor is accurate but not reliable; in d) the sensor is accurate and reliable (from G. Asch [ASC 91])

## 2.3. Sensor calibration

Calibration is an operation that establishes the relation between the measurand and the electrical output variable. This relation depends not only on the measurand but also on influence variables. Where there are no influence variables, simple calibration is used. Multiple calibration is necessary with influence variables.

### 2.3.1. Simple calibration

There are two possible methods of simple calibration. These are:

- direct calibration: the measurand values come from standards or reference objects through which we know the measurand, with a given uncertainty;
- comparison calibration: with this method, we compare the measurements of the sensor to be calibrated with measurements made by another sensor that already has been calibrated and is being used as the reference. This means that its calibration is linked to standards and that the corresponding uncertainty is known.

### **2.3.2. Multiple calibration**

The existence of influence variables that may vary throughout measurement means we must set calibration parameters for different values of these variables. This is multiple calibration. Several specific situations that require multiple calibration should be mentioned:

- for sensors that show hysteresis, calibration must be carried out in a series of specific steps of measurand values;
- for sensors with dynamic variables, we determine the response as a function of frequency;
- in certain cases, especially for many mechanical and thermal sensors, when the manufacturer has not given instructions for usage, calibration is often carried out on-site and after installation. For instance, in this way, an accelerometer can be calibrated after being attached to the structure to be measured, even if the manufacturer has specified a different procedure in the calibration certificate.

### **2.3.3. Linking international measurement systems**

All industrialized countries have sets of standards. This means they have laboratories organized for specific purposes, established over time, that link measurements to basic standards [ASC 91]. Standards and transfer instruments assure traceability and successive stages from laboratories to industry. In France, for example, the National Bureau of Metrology is in charge of insuring traceability according to national standards. Throughout this linking process, successive operations are carried out within a standardization process that not only joins measurements to measurands but defines uncertainty levels in sensor measurements. As far as the legal aspect of some of these operations is concerned, it is important to remember that processes certified by ISO 9000 standards also require the traceability of all functioning sensors.

## **2.4. Band pass and response time**

### **2.4.1. Harmonic response**

The response of a sensor to a measurand that varies sinusoidally over time is particularly important: from it we deduce the response to all measurand variations in time. This is the sensor's transitory response. If we call  $S(\omega)$  the sensor sensitivity exposed to a sinusoidal measurand at pulsation  $\omega$ , the response of a temporal pulse is given by its Fourier transform:

$$h(t) = \frac{1}{\sqrt{2\pi}} \int_0^\infty S(\omega) e^{-j\omega t} d\omega \quad [2.7]$$

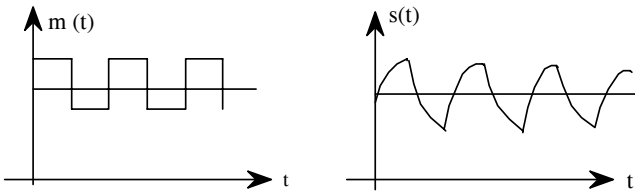
Lack of knowledge about this response can lead to systematic errors, even when carrying out stationary measurements.

When using sensors, the idea of band pass can be introduced through a discussion of distortion phenomena observed during measurement. If the measurand has a periodic temporal evolution described by Figure 2.4 that can be represented by:

$$m(t) = m_0 + \sum_i m_i \cos(\omega_i t + \theta_i) \quad [2.8]$$

By introducing the concept of static sensitivity  $S_0$  and dynamic  $S(\omega_i)$ , the delivered electrical signal  $s(t)$  can be expressed as:

$$s(t) = S_0 m_0 + \sum_i S(\omega_i) m_i \cos(\omega_i t + \psi_i) \quad [2.9]$$



**Figure 2.4.** Example of distortion with low band pass filtering

If the values  $S(\omega_i)$  are different, or if  $\psi_i$  is related in some way to  $\theta_i$ , a signal  $s(t)$  is obtained with a frequency content that changes in relation to the frequency content of the measurand. In such cases, we say the signal has undergone a distortion or that the system is dynamically non-linear.

The band pass is the frequency interval with a value of  $S(\omega)$  that is constant and in which  $\psi_i$  differs from  $\theta_i$  by a constant additive that can be written as  $\omega_i \tau$ , with  $\tau$  independent of  $\omega_i$ . Such systems are dynamically non-linear.

Generally, a sensor order is the order of the differential equation that governs its dynamic sensitivity. The simplest example is that in which an equation linking  $s$  to  $m$  in dynamic state is a first order differential equation:

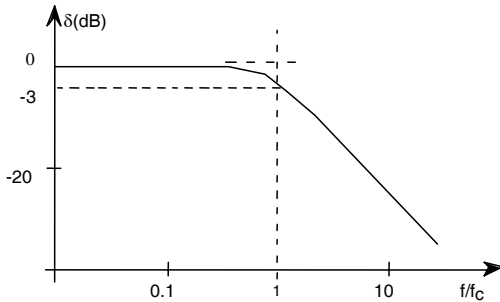
$$A \, ds/dt + Bs = m(t) \quad [2.10]$$

In this instance A and B are time-independent. When we call  $s_1$  and  $m_1$  the periodic parts of  $s$  and  $m$  and write the cut-off frequency  $f_c = B/2\pi A$ , the sensitivity  $|s|$  and the phase  $\psi$  of the sensor as a function of frequency  $f$  is written as:

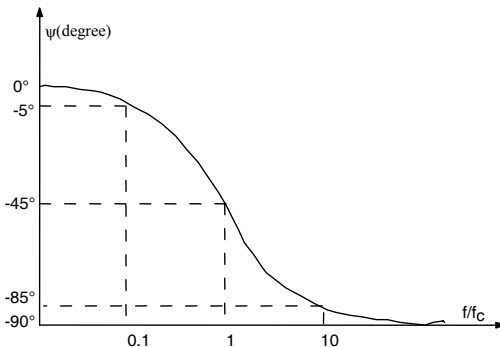
$$|s| = \frac{m_1}{B} \frac{1}{\sqrt{1 + \left(\frac{f}{f_c}\right)^2}} \tag{2.11}$$

$$\psi = -\arctg\left(\frac{f}{f_c}\right) \tag{2.12}$$

The amplitude response is often given in the form of attenuations in decibels (dB) as  $\delta = 20 \log_{10}(s(\omega)/s_0)$ . Figure 2.5 shows the general response slope of these first-order sensors with a cut-off frequency above  $f_c$  is 20 dB per decade.

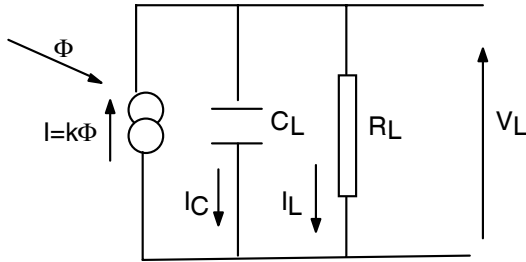


**Figure 2.5.** Dynamic sensitivity of a first-order sensor



**Figure 2.6.** Phase lag of a first-order sensor

An example of a first-order sensor, analyzed in more detail in section 3.1, is the photodiode. The photodiode works through the generation of electron-hole pairs “freed” from crystal-like bonding structures by the energy of absorbed photons. This sensor is a current source, often proportional to the absorbed luminous flux  $\Phi$ . No matter which utilization circuit is employed, the electrical signal delivered by the sensor can be given as the potential  $V_L$  at the limits of the charge resistance  $R_L$ . The differential equation of the sensor can be deduced from the equivalent diagram in Figure 2.7, where  $C_L$  is the diode capacity and  $R_L$  is a resistance, taking into account the charge resistance and the internal resistance of the photodiode.



**Figure 2.7.** *Equivalent design of a photodiode*

From this basic design, we can easily deduce:

$$I = I_C + I_L = C_L \frac{dV_L}{dt} + \frac{V_L}{R_L} = k\phi \quad [2.13]$$

and also:

$$\phi = \frac{C_L}{k} \frac{dV_L}{dt} + \frac{1}{kR_L} V_L \quad [2.14]$$

This is a first-order sensor equation. The cut-off frequency expressed by  $f_C = \frac{1}{2\pi R_L C_L}$  directly depends on charge resistance, and not only on the sensor.

This situation is actually very general. The response time of the sensors, whatever their order, is always influenced by the measurement chain (see Figure 2.8).



We can also establish that the sensitivity  $S_0$  is equal to  $kR_L$ , which means the product (gain  $\times$  band pass) is constant. Again, this result is quite general. We see that usually large band pass and strong gains do not exist together in most sensor measurement chains.

When the measurand is a mechanical variable [HAN 99], we often find second-order sensors. With these sensors, the equation joining the measure  $s$  to the measurand  $m$  is of this type:

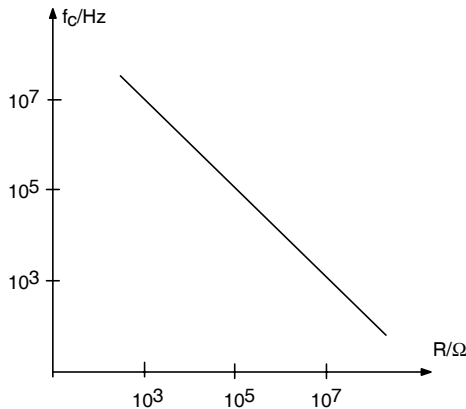
$$A \frac{d^2s}{dt^2} + B \frac{ds}{dt} + C = m \tag{2.15}$$

where  $A$ ,  $B$  and  $C$  are independent time constants. We introduce the cut-off frequency  $f_0$  and the damping factor  $\xi$  as:

$$f_0 = \frac{1}{2\pi} \sqrt{\frac{C}{A}} \tag{2.16}$$

and

$$\xi = \frac{B}{2\sqrt{CA}} \tag{2.17}$$



**Figure 2.8.** Example of variation in a photodiode band pass according to the charge resistance

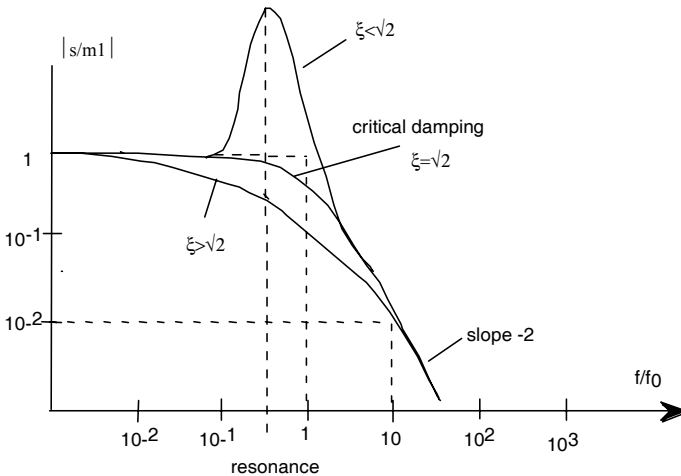
Resolving this equation leads to responses in amplitude and phase, equal respectively to:

$$\left| \frac{s}{m_1} \right| = \frac{1}{C \sqrt{\left(1 - \left(\frac{f}{f_o}\right)^2\right)^2 + 4\xi^2 \left(\frac{f}{f_o}\right)^2}} \quad [2.18]$$

and

$$\Psi = \arctg \left[ \frac{-2\xi}{\frac{f_o}{f} \left(1 - \left(\frac{f}{f_o}\right)^2\right)} \right] \quad [2.19]$$

Dynamic sensitivity can show a resonance with weak damping and damping beyond the frequency  $f_o$  equals 40 dB per decade. The phase lags are doubled compared to those of first-order sensors.



**Figure 2.9.** Amplitude response of a second-order sensor

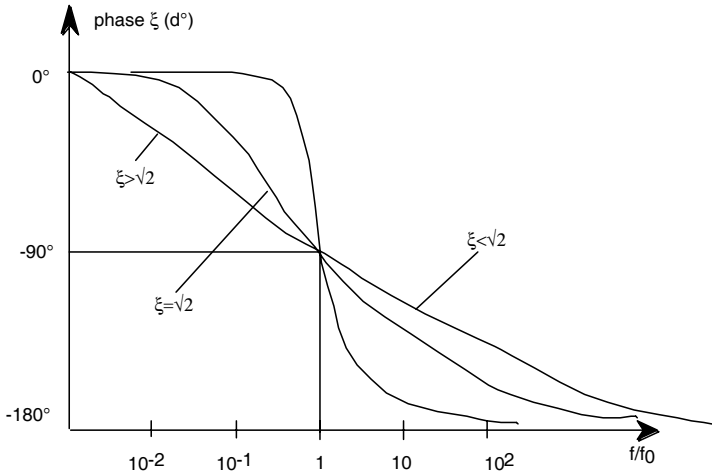


Figure 2.10. Phase of a second-order sensor

### 2.4.2. Response time

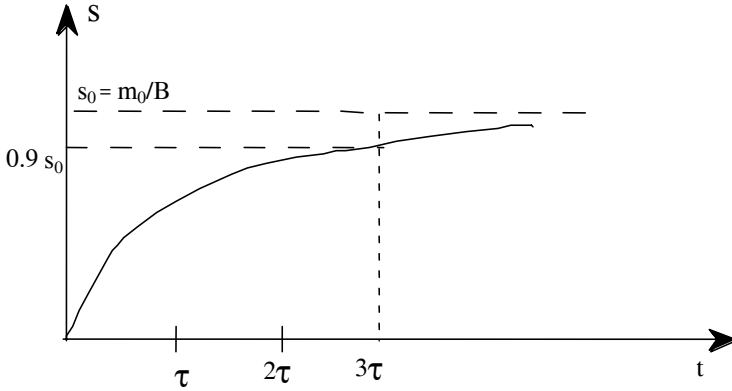
The response time of sensors can also be deduced from the differential equations presented above. For a first-order sensor, if  $m = 0$  for  $t < 0$  and  $m = m_0$  for  $t > 0$ , we get the solution:

$$s = s_0(1 - e^{-t/\tau}) \text{ with } \tau = \frac{A}{B} \tag{2.20}$$

and

$$s_0 = \frac{m_0}{B} \tag{2.21}$$

Here  $\frac{1}{B}$  is sometimes called static sensitivity.



**Figure 2.11.** First-order sensor response

In order to reach 90% of the sensitivity  $s_0$  of the steady state,  $t = 3\tau$  must be achieved. If the measurand is of the type  $m_0 = 0$  at  $t < 0$  and  $m = m_0$  (constant) for  $t > 0$ , a second-order sensor gives a response that is highly dependent on damping (see Figure 2.12). These transient solutions follow from this:

– if  $\xi > 1/\sqrt{2}$ ,

$$s = k_1 e^{-\eta_1 t} + k_2 e^{-\eta_2 t} \quad [2.22]$$

with:  $\eta_{1,2} = -\xi \omega_0 \pm \sqrt{\omega_0^2 (\xi^2 - 1)}$

– if  $\xi < 1/\sqrt{2}$ ,

$$s = k_1 e^{-\xi \omega_0 t} e^{j \omega_0 \sqrt{1-\xi^2} t} + k_2 e^{-\xi \omega_0 t} e^{-j \omega_0 \sqrt{1-\xi^2} t} \quad [2.23]$$

– if  $\xi = 1/\sqrt{2}$ ,

$$s = k_1 (1 + \omega_0 t) e^{-\omega_0 t} \quad [2.24]$$

If we take as initial conditions  $s = 0$  and  $\frac{ds}{dt} = 0$  with  $t = 0$ , we get the following complete solutions:

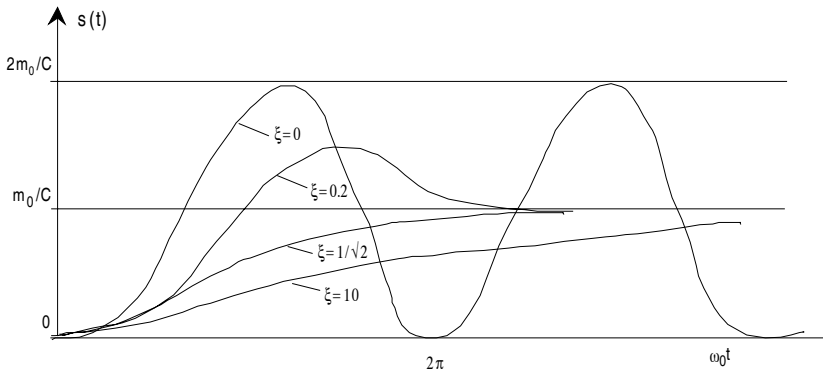
$$\xi < 1/\sqrt{2} \Rightarrow s(t) = \frac{m_o}{C} \left[ 1 - \frac{\exp(-\xi\omega_o t)}{\sqrt{1-\xi^2}} \sin \left[ (\sqrt{1-\xi^2})\omega_o t + \psi \right] \right]$$

with  $\Psi = \arcsin \sqrt{1-\xi^2}$

$$\xi = 1/\sqrt{2} \Rightarrow s(t) = \frac{m_o}{C} [1 - (1 + \omega_o t) \exp - \omega_o t]$$

$$\xi > 1/\sqrt{2} \Rightarrow s(t) = \frac{m_o}{C} \left[ -\frac{+\xi + \sqrt{\xi^2 - 1}}{2\sqrt{\xi^2 - 1}} \exp \left[ -\xi + \sqrt{\xi^2 - 1} \right] \omega_o t \right. \quad [2.25]$$

$$\left. + \frac{+\xi - \sqrt{\xi^2 - 1}}{2\sqrt{\xi^2 - 1}} \exp \left[ -\xi - \sqrt{\xi^2 - 1} \right] \omega_o t \right]$$



**Figure 2.12.** Temporal response of a second-order sensor

Second-order sensors are usually built with a damping  $\xi = 0.6$  (see Figure 2.12). With this the steady state  $m_0/C$  to almost 10% is reached for a period equal to

$2.4/\omega_0$ . This value must be taken into account to be certain that the sensor conforms to the calibrations set by the manufacturer.

## 2.5. Passive sensor conditioners

Passive sensors convert the measurand into an impedance variable. They must always be used with a circuit that has a source current or tension and, generally, several additional impedances. The circuit is called the conditioner. There are two main groups of impedance conditioners [BAU 61]. In the first group, the sensor's impedance variation is converted by a variation of potential difference. In the second group, the impedance variation is used to modify an oscillator frequency. In this case, the sensor reading is actually a frequency measurement. Here we will only discuss the first group.

### 2.5.1. The effect of polarization instabilities

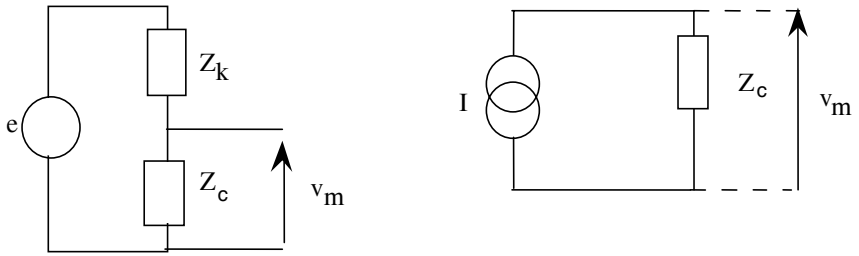
When a sensor's impedance variation is converted into a tension variation, the measured tension depends on the impedance of the sensor and the different conditioner elements that can be effected by influence variables and disturbances. Choosing the right conditioner can be critical for the signal to noise ratio [ANS 82]. With the simplest possible conditioner, which is often called potentiometric conditioner (see Figure 2.13), the signal  $V_m$ :

$$V_m = e \frac{Z_c}{Z_c + Z_K} \quad [2.26]$$

is proportional to the impedance of sensor  $Z_c$ , which makes it very sensitive to disturbances that could introduce random variations. If we take a source instability equal to  $\Delta e$ , the result is a variation of the measurement, expressed by:

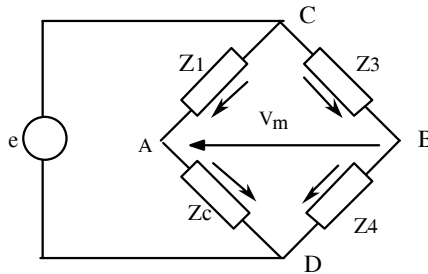
$$\Delta V_m = \Delta e \frac{Z_{CO} + \Delta Z_C}{Z_K + Z_{CO} + \Delta Z_C} \# \Delta e \frac{Z_{CO}}{Z_K + Z_{CO}} \quad [2.27]$$

This result can be found in the case of a simple polarization by a source current as well.



**Figure 2.13.** Potentiometric conditioner with voltage or current sources

On the other hand, bridge conditioners help eliminate the noise very efficiently (see Figure 2.14).



**Figure 2.14.** Impedance bridge in which  $Z_c = Z_{c0} + \Delta Z_c$  is the sensor

It is easy to show that the tension  $V_m$  appearing with the measurand is increased by  $\Delta V_m$  when  $e$  changes to  $e + \Delta e$  following:

$$\Delta V_m = \Delta e \left[ \frac{Z_{CO} + \Delta Z_C}{Z_1 + Z_{CO} + \Delta Z_C} - \frac{Z_{CO}}{Z_{CO} + Z_1} \right] = \Delta e \frac{Z_1 \Delta Z_C}{(Z_1 + Z_{CO} + \Delta Z_C)(Z_{CO} + Z_1)} \quad [2.28]$$

or:

$$\Delta V_m \# \Delta e \frac{Z_1 \Delta Z_C}{(Z_K + Z_1)^2} \quad [2.29]$$

For the same polarization instability, the noise generated by the measurement in the case of a bridge and a potentiometer are in the relation:

$$\frac{\Delta V_m \text{ bridge}}{\Delta V_m \text{ potentiometer}} = \frac{\Delta Z_C}{Z_{CO} + Z_1} \quad [2.30]$$

that is, in the variation order relative to the impedance. For measurements in which impedance varies from the order of % to around  $Z_{co}$ , the bridge is 100 times less sensitive to random variations of  $e$  than with potentiometric conditioners or the direct polarization by current source.

### 2.5.2. Effects of influence variables

It is important to remember that minimization influence variables follow a general rule that allows for optimization of conditioners. Suppose the measurement tension delivered by the conditioner depends on impedances, which we assume are all resistive, is written as:

$$V_m = f(R_K R_C) \quad [2.31]$$

We take  $g$  as the influence variable that modifies all the conditioner's resistances. A variation  $dg$  of the influence variable produces a variation  $dV_m$  of the measurement tension:

$$dV_m = \left[ \sum_K \frac{\partial V_m}{\partial R_k} \frac{\partial R_k}{\partial g} + \frac{\partial V_m}{\partial R_C} \frac{\partial R_C}{\partial g} \right] dg \quad [2.32]$$

If we get  $dV_m = 0$ , we have:

$$\sum_K \frac{\partial V_m}{\partial R_k} \cdot \frac{\partial R_k}{\partial g} + \frac{\partial V_m}{\partial R_C} \frac{\partial R_C}{\partial g} = 0 \quad [2.33]$$

For example, in the case of a potentiometric conditioner (or when one of the assembly resistances is sensitive to the influence variable  $g$ ) and when the



sensitivities to  $g$  of the two resistances  $R_c$  and  $R_k$  are the same, the condition  $dV_m = 0$  is equivalent to:

$$\frac{\partial V_m}{\partial R_k} = -\frac{\partial V_m}{\partial R_c} \tag{2.34}$$

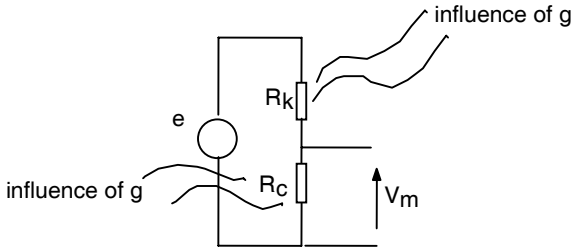
In the case of a potentiometric conditioner, we get:

$$\frac{\partial V_m}{\partial R_c} = \frac{R_k}{(R_c + R_k)^2} \tag{2.35}$$

and:

$$\frac{\partial V_m}{\partial R_k} = \frac{-R_c}{(R_c + R_k)^2} \tag{2.36}$$

Once again, this equals  $R_c$  and  $R_k$  (see Figure 2.15).



**Figure 2.15.** Elimination of the influence of the  $g$  variable in a potentiometric conditioner

With resistive bridges (see Figure 2.14), we see that sensitivity to influence variables is also minimal when:

$$R_1 = R_{CO} = R_3 = R_4 \tag{2.37}$$

These results also apply to complex impedance bridges.

**2.5.3. Conditioners of complex impedance sensors**

Conditioners of passive sensors with complex impedances are made with different bridges, for instance, Nernst’s bridge for capacitive sensors (see Figure 2.16) and Maxwell’s bridge (see Figure 2.17) for inductive sensors.

Nernst’s bridge is used for sensors with an impedance that can be represented by an impedance  $Z_c$ :

$$Z_c = \frac{R_c}{1 + jR_c C \omega} \tag{2.38}$$

For the value  $m_0$  of the measurand we adjust the following impedances of the bridge:

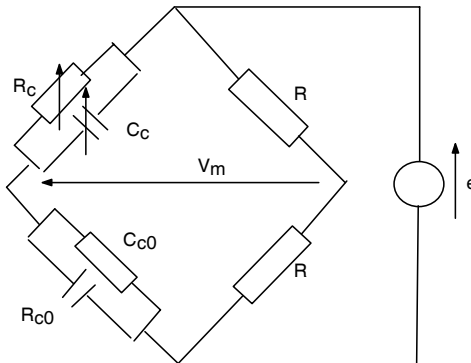
$$R_e = R_c = R_{c0} \tag{2.39}$$

and

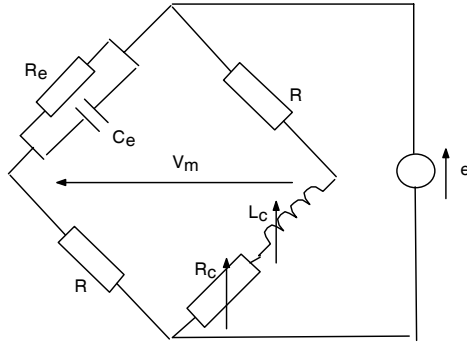
$$C_e = C_c = C_{c0} \tag{2.40}$$

So,  $V_{m0} = 0$ . If  $Z_c$  changes from  $Z_{c0}$  to  $Z_{c0} + \Delta Z_c$ , we get:

$$V_m \approx \frac{e}{4} \cdot \frac{\Delta Z_c}{Z_{c0}} \tag{2.41}$$



**Figure 2.16.** Bridge conditioner for a capacitive sensor



**Figure 2.17.** Bridge conditioner for an inductive sensor

For an inductive sensor, the impedance  $Z_c$  is given as:

$$Z_c = R_c + jL_c\omega \tag{2.42}$$

When the measurand equals  $m_0$ , the bridge changes to:

$$R_{c0} = \frac{R^2}{R_e} \tag{2.43}$$

and:

$$L_{c0} = R^2 \cdot C_e \tag{2.44}$$

so we then get:

$$V_m \approx e \cdot \frac{R \cdot \Delta Z_c}{(R + Z_{c0})^2} \tag{2.45}$$

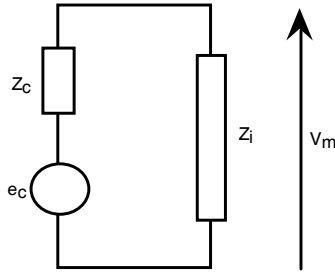
## 2.6. Conditioners for active sensors

### 2.6.1. Direct reading

Direct readings of active sensors are rarely satisfactory, whether or not these sensors are equivalent to tension, currents or charge sources. This is because this kind of reading presupposes a correction that is not always easy to evaluate.

When electrical information delivered by active sensors appears in the form of a tension source ( $e_c$ ) in series with an impedance  $Z_c$ , the electrical signal can be read to the impedance limits of  $Z_i$  and we get (see Figure 2.18):

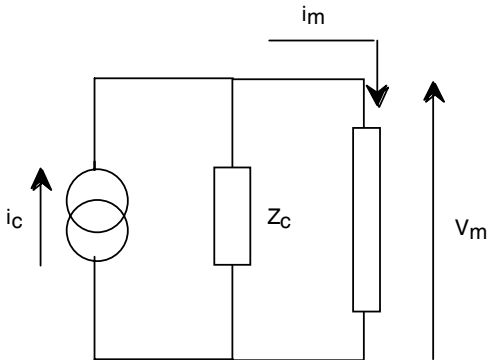
$$V_m = \frac{Z_i}{Z_i + Z_c} e_c \quad [2.46]$$



**Figure 2.18.** *Equivalence to a tension source*

For the measurement to be as close as possible to the tension delivered by the sensor, we must have  $v_m \approx e_c$ , and this implies that  $z_i \gg z_c$ , but this leads to a reduction of the band pass (see section 2.5.1).

The sensor can also appear in a form equivalent to a current source ( $i_c$ ) in parallel with an impedance  $Z_c$ . The electrical signal  $V_m$  is then given as in Figure 2.19.



**Figure 2.19.** *Equivalence to a current source*

$$v_m = Z_i i_m \text{ with } i_m = i_c \frac{Z_c}{Z_i + Z_c} \tag{2.47}$$

For  $i_m$  to approximate  $i_c$ , we need:

$$Z_i \ll Z_c \tag{2.48}$$

but in this instance, the signal will be very weak.

Lastly, with sensors that are also charge sources, it is clear that a simple measurement by the difference potential to the resistance limits affects the signal, since the measurement discharges the sensor.

### 2.6.2. Using operational amplifiers

Here we will not discuss in any detail the many schema used for signal processing of active sensors that resolve the problems presented in this section (see the following chapters for a more detailed presentation of these). Instead, we explain the three basic assemblies that correspond to the three types of equivalencies found in active sensors. These are basic to using operational amplifiers [FRA 93].

First of all, let us suppose that the sensor is equivalent to a tension source  $e_c$  in series with an impedance  $Z_c$ . With the assembly shown in Figure 2.20, we can easily see that, if we are close to the operational ideal for the amplifier, we get:

$$V_s = \left(1 + \frac{R_2}{R_1}\right) V_e = G V_e \tag{2.49}$$

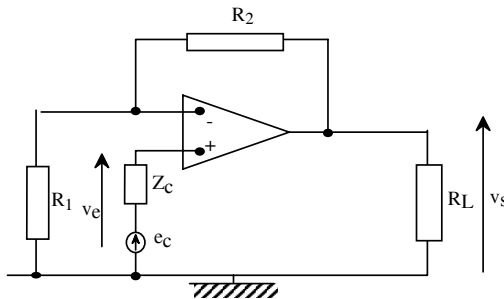


Figure 2.20. Assembly type for a sensor equivalent to a tension source

We can see that the sensor does not produce any current ( $i_+ = i_- = 0$  in an ideal amplifier) which means that it is connected to infinite impedance. The non-influence condition of the sensor's internal impedance  $Z_c$  is fulfilled when:

–  $V_s$ , as output, is independent of the current transmitted in the charge  $R_L$ . The tension  $V_s$  transmitted by the amplifier acts as a tension source of a zero internal impedance;

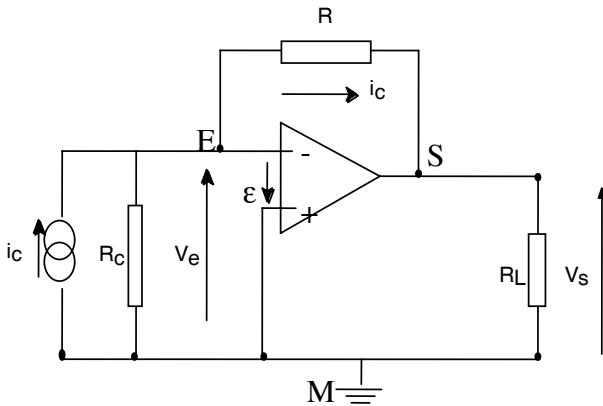
– the choice of  $R_1$  and  $R_2$  help regulate the desired gain  $G$ .

Remembering that choosing a gain is related to choosing a cut-off frequency, the product  $(G \cdot \omega_c)$  is a constant dependent on the type of operational amplifier chosen:

$$G\omega_c = \frac{\mu_o}{\tau_o} \tag{2.50}$$

where  $\mu_o$  is the open circuit gain and  $\tau_o$  is the response time of the operational amplifier.

Suppose now that the sensor is equivalent to a current source placed in parallel with a resistance  $R_c$ . We can then use the assembly seen in Figure 2.21.



**Figure 2.21.** Assembly type for a sensor equivalent to a current source

Since the input of an ideal amplifier would not transmit current, and since input differential tension is zero  $\epsilon \approx 0$ , the potential difference between E and M is zero

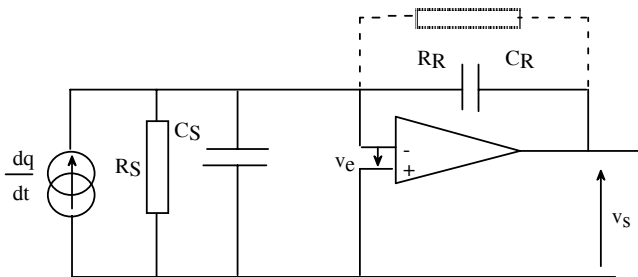
and no current circulates in the resistance  $R_c$  of the sensor. The output tension  $V_S$  is expressed as:

$$V_S = -Ri_c \tag{2.51}$$

As with amplifier tension, we will give a few basic descriptions of how basic assembly works:

- the value chosen for the feedback resistance  $R$  does not influence a sensor equivalent to a current source;
- input resistance is zero because source limits are maintained at the same potential to the input of an ideal amplifier;
- output provides a source of tension whose resistance is zero ( $V_S$  is independent of the charge resistance placed at output).

In the case of a sensor equal to a charge generator, it is often best to use a charge-tension convertor to short-circuit electrodes (see Chapter 3 for more detail). The most basic assembly is shown in Figure 2.22.



**Figure 2.22.** Assembly type for a sensor equal to a charge source

Because no current can come through amplifier inputs, all charge variations within the sensor’s limits are found within the limits of  $C_R$ . Here we get:

$$V_S = -\frac{Q}{C_R} \tag{2.52}$$

In reality we often have to take into account the resistance to leakage of capacity  $C_R$  ( $R_R$  in parallel with  $C_R$ ) especially always with low frequencies. It is easy to show that  $V_S$  becomes:

$$V_S = -\frac{Q}{C_R} \cdot \frac{j\omega R_R C_R}{1 + j\omega R_R C_R} \quad [2.53]$$

This is the expression of a high pass filter that shows a converter of this type does function properly at low frequencies and cannot transmit current.

## 2.7. Bibliography

- [AFN] AFNOR, NF X07001 norm.
- [ANS 82] Standard, Temperature measurement, ISA/ANSI standard MC 96, International Society for measurement and control, 1982.
- [ASC 91] ASCH G. *et al.*, *Capteurs en instrumentation industrielle*, Dunod, 1991.
- [BAU 61] BAURAND J., *Mesures électriques*, Masson, 1961.
- [DIE 92] DIECK R.H., *Measurements uncertainty*, ISA, Research Triangle Park, 1992.
- [FRA 93] FRADEN J., *AIP handbook of modern sensors*, AIP Press, 1993.
- [HAN 99] *Measurement, instrumentation and sensors handbook, CRC Handbook*, Springer, IEEE Press (1999).
- [NOR 99] NORTON H.N., *Handbook of transducers*, Prentice Hall, Englewood Cliffs, NJ, 1999.



*This page intentionally left blank*

## Chapter 3

# Physical Principles of Optical, Thermal and Mechanical Sensors

There are now so many sensors available [NOR 89] that it would be impossible to discuss the principles of all of them in a single chapter. We have therefore limited ourselves to three classes of measurands: optical, thermal and mechanical. However, even with this restriction, we still must limit our scope, and will only present the most frequently used laws for these types of physical sensors.

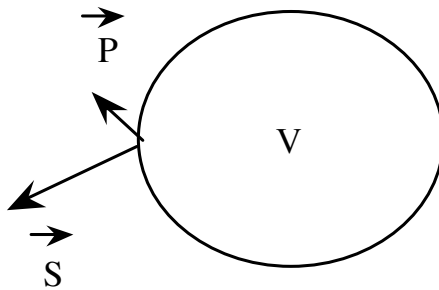
### 3.1. Optical sensors

One important class of sensors detects electromagnetic beams. Within this group, we will restrict our discussion to those optic sensors that are sensitive only to beams with wavelengths of 10 nm – 1 mm, that is, frequencies of between  $10^{16}$  and  $10^{11}$  Hz. In the specific case of light sensors, for reasons relating to the sensitivity of the human eye, it is necessary to introduce specific concepts when discussing visibility (0.4  $\mu\text{m}$  to 0.8  $\mu\text{m}$ ). After a brief recapitulation of the variables that act as measurands for optical sensors, we will define the reference light source used in making calibrations. We will then discuss the principles of sensors that are used in constructing semiconductors.

### 3.1.1. Energetic flux

An electromagnetic beam transmits energy. One way to see this is to place a thermometer with a darkened reservoir in an electromagnetic beam and check that the temperature increases. We know that the flux of Poynting's vector,  $\vec{P} = \vec{E} \wedge \vec{H}$  the vectorial product of electrical and magnetic fields, across a surface  $S$  surrounding a closed volume  $V$ , is equal to the quantity of electromagnetic energy  $W$  that comes from  $V$  by the surface  $S$  and by unity of time. This flux can be expressed by means of the divergence of  $\vec{P}$  and the local electromagnetic energy density, as we see in Figure 3.1:

$$\iint_S \vec{P} \cdot d\vec{S} = \iiint_V \text{div} \vec{P} \cdot dV = \iiint_V \left[ \frac{\partial}{\partial t} \left( \epsilon E^2 / 2 \right) + \frac{\partial}{\partial t} \left( \mu H^2 / 2 \right) \right] dV \quad [3.1]$$



**Figure 3.1.** Flux coming from a volume  $V$  across a surface  $S$

In the specific case of a plane wave, it is known that  $H = E/\mu v$  where  $v$  is the speed of the wave in the middle of the index  $n$  ( $v = c/n$  with  $c$  the light speed in traveling in a vacuum). The flux is reduced to:

$$\iint_S \vec{P} \cdot d\vec{S} = \iiint_V \text{div} \vec{P} \cdot dV = \iiint_V \frac{\partial}{\partial t} (\epsilon E^2) dV \quad [3.2]$$

From this we can deduce that the energetic flux is proportional to the square of the amplitude of the electric field. This relation is very commonly used by all sensors sensitive to energy and therefore to the square of the field, that is to say that any phase information is lost. The phase can only be retrieved by interference phenomena such as holography or speckle.

### 3.1.2. *Luminous flux*

In optical sensors with at least one capacity of visual sensitivity, measurands can be either energetic or luminous variables. Energetic variables are measurements that are completely independent of the sensitivity of the human eye. With luminous variables, on the other hand, the effect of the eye is taken into account as a variable according to the radiation wavelengths. It is important to remember that when choosing between these two measurands, the practical application is very important. For example, when studying photomultiplier sensors, within the framework of spectroscopy, it becomes clear that energetic variables are the only measurands we need to take into account. However, if we want to measure the flux coming into a camera producing images for the human eye, luminous variables are indispensable. As well, measuring the light in a room or road can only be expressed in units of luminous flux.

We have seen how energetic flux is directly related to the square of the electromagnetic field. Luminous flux is defined with respect to retina sensations. This kind of flux is a measurand, that is, a measurable variable, because we can define the equality of these two fluxes (the same sensation produced by two adjacent zones of the same shield) and the sum of several fluxes that superimpose their action on the eye. These measurements can also be made with a photoelectric cell with a spectral sensitivity set as close as possible to that of the human eye. No matter which methods are used, the measurements must be carried out with monochromatic waves, which means that the measurements must be taken in the brief interval  $d\lambda$  around the wavelength  $\lambda$  of the measurement. These are spectral variables.

### 3.1.3. *The relative luminous efficiency curve $V(\lambda)$ of the human eye*

“Normal” human eyesight is not the same for everyone across the spectrum of visual experience or for a single visual sensation. Even for one person, this sensation varies according to psychological and physical factors. Therefore, we define luminous efficiency of the eye by citing a large-scale statistical study carried out on people with “normal” eyesight. The results of this study have led to a definition of the average eye. The sensations of this standard eye, which we call the luminous flux  $F_\lambda$ , are at each wavelength  $\lambda$  proportional to the received spectral energetic flux  $\Phi_\lambda$ . The proportionality factor  $k_\lambda$  of course depends on  $\lambda$ . If this average eye receives the spectral energetic fluxes  $\Phi_\lambda$  and  $\Phi_{\lambda'}$  (to  $\lambda$  and  $\lambda'$ ) so that the luminous spectral fluxes  $F_\lambda$  and  $F_{\lambda'}$  are equal, we express them as:

$$\frac{F_\lambda}{F_{\lambda'}} = \frac{K_\lambda \phi_\lambda}{K_{\lambda'} \phi_{\lambda'}} = 1 \quad [3.3]$$

The experiment shows that  $K_\lambda$  is at its maximum at a wavelength  $\lambda = 0.555 \mu\text{m}$ . With  $K_m = K_{0.555}$  we can define the relative luminous efficiency  $V_\lambda$  given in Figure 3.2 with:

$$V_\lambda = \frac{K_\lambda}{K_m} \leq 1 \tag{3.4}$$

Using this definition, the luminous flux  $F_\lambda$  is given in the form:

$$F_\lambda = K_m V_\lambda \Phi_\lambda \tag{3.5}$$

The numeric value attributed to  $K_m$  leads to defining the relation between the unities of energetic flux and luminous flux. The unity of energetic flux, called the lumen (lm), has been set since 1979 as  $K_m = 680 \text{ lm} \cdot \text{W}^{-1}$ .

Three other variables are important for optical sensors. These are intensity, luminance and illumination. Of these, intensity is certainly the most well known because it is the most frequently used. Its origin resides in the fact that most luminous sources transmit fluxes that depend not only on surface points but also on the emission angle in a normal relation to the surface. For this reason, it is necessary to evaluate the elemental flux transmitted by an element  $dS$  of the surface around the point 0 in a small solid angle  $d\Omega$  around a given direction  $x$ . This leads to the definition of transmitted intensity in this direction (see Figure 3.3). The unity of energetic intensity  $d\Phi/d\Omega$  is clearly the  $\text{W} \cdot \text{sr}^{-1}$  and that of the luminous intensity is  $dF/d\Omega$  is the candela or  $\text{lm} \cdot \text{sr}^{-1}$ .

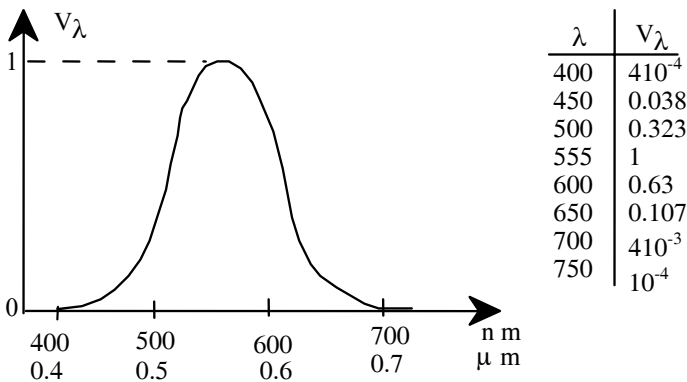
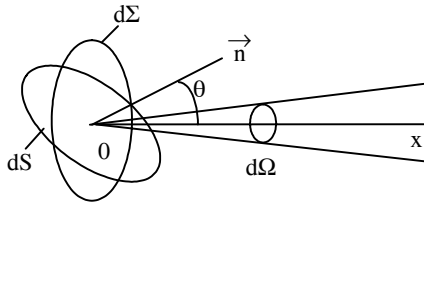


Figure 3.2. Luminous efficiency curve of a normal eye

Now, if we are within the range of a sensor transmitting waves in the direction  $x$  (see Figure 3.3), we see that the visible transmission surface is no longer  $dS$  but  $d\Sigma$ , which is the projection of  $dS$  on the plane usual to  $0x$ . As long as  $dS$  stays small, the luminous flux perceived in this direction is proportional to  $d\Sigma$ . We then introduce the luminance  $L$  as:

$$L = \frac{dI}{d\Sigma} = \frac{dI}{dS \cos \theta} \tag{3.6}$$



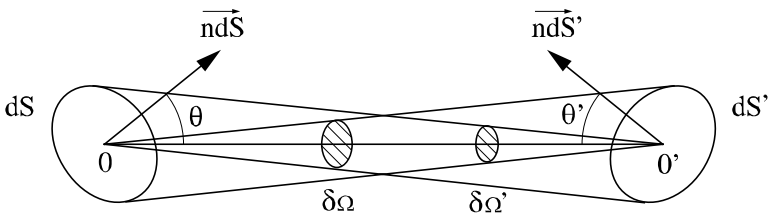
**Figure 3.3.** Definition of intensity

For many sources, luminance  $L$  is independent of the transmission angle  $\theta$ . These are called Lambertian sources. The flux  $d\Phi$  transmitted by the source of luminance  $L$  in the solid angle  $d\Omega$  and the direction  $\theta$  is written as:

$$d^2\Phi = dI d\Omega = L dS \cos \theta d\Omega$$

$L$  is expressed as  $\text{Wm}^{-2}\text{sr}^{-1}$  or for the luminous variable as  $\text{cd.m}^{-2}$ .

Let us look at the flux  $d^2\Phi$  transmitted by the surface  $dS$  in the direction of a sensor, delimiting the solid angle  $d\Omega$  by the surface  $dS'$  of the sensor.



**Figure 3.4.** Definition of the geometric area

The solid angle  $d\Omega$  can be easily expressed according to  $dS'$  and  $\theta'$ :

$$d\Omega = \frac{dS' \cos \theta'}{(00')^2} \quad [3.7]$$

and  $d^2\Phi$  can be written:

$$d^2\phi = \frac{LdS \cos \theta dS' \cos \theta'}{(00')^2} \quad [3.8]$$

where  $\frac{dS \cos \theta}{(00')^2}$  is the solid angle  $d\Omega'$  by which the sensor observes the source. We then get:

$$d^2\phi = LdS' \cos \theta' d\Omega' \quad [3.9]$$

The quantity  $dS \cos \theta d\Omega = dS' \cos \theta' d\Omega'$  is called the geometric area (see Figure 3.4). This quantity is conserved in stigmatic (that is, having to do with images) optical systems.

For most sensors, and especially optical sensors, the important measurand is the flux received by the unity surface called  $E$ :

$$E = \frac{d^2\phi}{dS'} \quad [3.10]$$

It is expressed as  $\text{W.m}^{-2}$  or  $\text{lm.m}^{-2}$ , also called lux.

### 3.1.4. *The black body: a reference for optical sensors*

Now that we have some knowledge of measurands, the next step is learning to calibrate optical sensors. To do this, we must be able to produce energetic and luminous fluxes whose properties are both completely known and reproducible. This can only be achieved through thermal radiation of black bodies. How these black bodies react depends on the temperature of the body and the universal constants.

3.1.4.1. *Black body radiation*

Maxwell's equations show that a continuous electrical current creates a magnetic field  $\vec{H}$ . If this current varies over time, this magnetic field  $\vec{H}$  itself creates an electrical induction  $\vec{D}$  that also varies over time. Thus, a current varying over time generates an electromagnetic wave ( $\vec{D}, \vec{H}$ ). In other words, for a current that varies over time, the electrical charges must vibrate around their position of equilibrium (a harmonic oscillator). These charges then accelerate, producing the couple ( $\vec{D}, \vec{H}$ ). In solid bodies, these vibrations can take many values. Because a harmonic oscillator only can take energies equal to  $p.h\nu$  with  $p$  integer and  $\nu$  the oscillator frequency [YAR 85], Planck shows that the spectral luminance  $L_\lambda$  of a solid, with the absolute temperature  $T$ , is expressed by:

$$L_\lambda = \varepsilon_\lambda \left[ \frac{C_1}{\lambda^5 (e^{c_2/\lambda T} - 1)} \right] = \varepsilon_\lambda L_\lambda^0(T) \quad [3.11]$$

where  $\varepsilon_\lambda$ , the transmissivity of the solid, depends on  $\lambda$  and is characteristic of the body (or sometimes only of its surface).  $L_\lambda^0(T)$  is the specific luminance of the black body that is only a function of the temperature of the body. This is Planck's law. The constants  $C_1$  and  $C_2$  are expressed by:

$$C_1 = 1.19 \cdot 10^{-16} \text{ W.m}^{-2}.\text{sr}^{-1} = 1.19 \cdot 10^{-8} \text{ W.}\mu\text{m}^{-2}.\text{sr}^{-1}$$

$$C_2 = \frac{hc}{k} = 1.438 \cdot 10^{-2} \text{ m.K} = 1.438 \cdot 10^4 \mu\text{m.K} \quad [3.12]$$

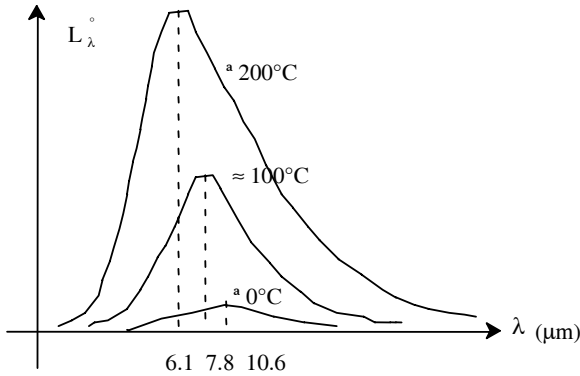
This function  $L_\lambda^0(T)$  is maximum for a wavelength  $\lambda_M$  obeying the relation ( $\lambda_M$  in  $\mu\text{m}$  and  $T$  in  $\text{K}$ ):

$$\lambda_M = \frac{2,898}{T} \quad [3.13]$$

The spectral luminance of the black body, shown in Figure 3.5 for three adjacent temperatures of the surrounding ambient, shows that the total luminance (integrated with the  $\lambda$  variant of 0 to  $+\infty$ ) is a rapidly growing function of the temperature. We get:

$$L^\circ = \int_0^\infty L_\lambda^\circ d\lambda = \frac{\sigma T^4}{\pi} \text{ with } \frac{\sigma}{\pi} = 1.8 \cdot 10^{-8} \text{ W/m}^2 \text{ K}^4 \quad [3.14]$$





**Figure 3.5.** Spectral luminance of black body

We note here that by explaining how energetic exchanges between matter and light occur in the form of photons, the discovery of the black body law marked the beginning of quantum physics.

3.1.4.2. Realization of black bodies

In the expression  $L_\lambda$ , the term  $\epsilon_\lambda$  is independent of temperature and, in fact, is solely dependent on the radiation properties of the surface. Kirchhoff has shown that transmissivity is equal to absorptivity ( $\epsilon_\lambda = \alpha_\lambda$ ). However, the conservation of energy at the interface between the two environments allows us to write a relation between  $\alpha_\lambda$  and the transmission coefficients  $\tau_\lambda$  and the reflection  $\rho_\lambda$ . We get:

$$\alpha_\lambda + \tau_\lambda + \rho_\lambda = 1 \tag{3.15}$$

To analyze the black body, the first idea is to look at opaque materials. An opaque object does not transmit energy at any wavelength  $\lambda$  and for all incidences ( $\tau_\lambda = 0$ ), which means it is a black body ( $\epsilon_\lambda = 1$ ) if it does not reflect energy for all wavelengths and incidences ( $\rho_\lambda = 0$ ). To see how this works, we will look at opaque materials.

Dielectrics are not helpful to realize black bodies because their behavior largely depends on the wavelength. For example, white paper has very weak remote infrared reflectance, which makes it close to being a black body beyond  $6 \mu\text{m}$  (an transmissivity of the order of 0.92). Unfortunately, it becomes a much stronger reflector in the visible, with almost no emissions at all.

In the case of metals,  $\tau_\lambda = 0$  at any  $\lambda$ , but we know that the reflection factor, independent of incidence except when it becomes low-angled, is expressed by:

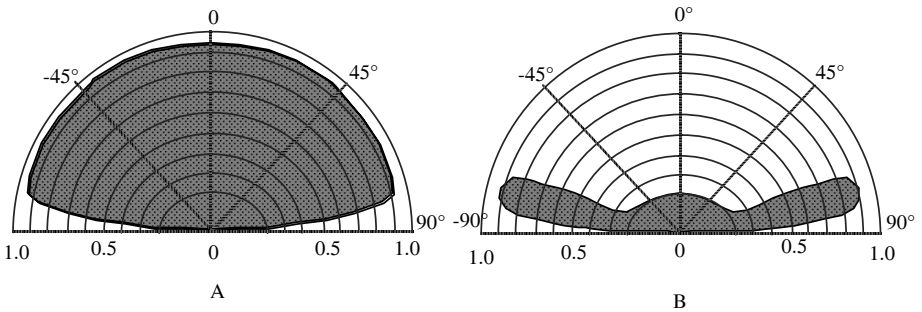
$$\rho_\lambda = \frac{4n}{(1+n)^2} \quad [3.16]$$

where the index is high for a metal. The transmissivity is written as:

$$\varepsilon_\lambda = 1 - \rho_\lambda = \frac{4n}{(1+n)^2} \quad [3.17]$$

and t is always weak.

Figure 3.6 shows that the use of opaque metallic bodies or dielectric bodies does not give the transmissivity of black bodies.



**Figure 3.6.** Transmissivity indicators of dielectric and metal bodies to a given wavelength

Given these findings, it becomes necessary to look at transparent bodies and treat them as opaque artefacts. The transmission coefficient  $\tau_\lambda$  is written as:

$$\tau_\lambda = \exp(-\beta_\lambda x) \quad [3.18]$$

where  $\beta_\lambda$  is the extinction coefficient and  $x$  is the distance light traveled in the material. If we choose a dielectric with the weak reflection coefficient, we get:

$$\varepsilon_\lambda = 1 - \tau_\lambda = 1 - \exp(-\beta_\lambda x) \quad [3.19]$$

This shows that augmenting  $x$  increases transmissivity. This can be done by creating an optical diffusion inside the material. The path traveled by the light becomes much longer than the simple thickness  $x$ ; this increases transmissivity [HOD 71].

Now suppose we create a cavity whose interior is covered with diffusing dielectric material, with an transmissivity  $\epsilon_\lambda$  necessarily inferior to 1. This cavity is closed everywhere except for a small opening. Each element of the internal surface  $dS$  transmits practically nothing in the direction of this opening and even this emission reaches other internal elements. The emission of the surface  $dS$  of the internal wall is written:

$$\phi_{emitted} = dS \epsilon \sigma T^4 \quad [3.20]$$

where  $T$  is the supposed uniform temperature of the cavity. This surface  $dS$  also receives, from another element of the surface  $dS$  of the cavity, a flux  $\epsilon \sigma T^4 dS$ , with one part shown as:

$$\phi_{reflected} = (1 - \epsilon) \epsilon dS \sigma T^4 \quad [3.21]$$

When we consider only  $dS$ , the radiative flux that comes from this term is:

$$\phi = \epsilon(1 - \epsilon) dS \sigma T^4 + \epsilon dS \sigma T^4 = \epsilon(2 - \epsilon) \sigma T^4 dS \quad [3.22]$$

Thus, we have a large transmissivity  $\epsilon(2 - \epsilon)$  that, for  $\epsilon = 0.9$ , already improves transmittance by 10%. Taking into account successive  $n$  reflections, we see that transmissivity becomes  $1 - (1 - \epsilon)^{n+1}$ . Beyond three reflections, noticeable transmissivity is nearly that of the black body, having become independent of the wall transmissivity; that is, of its nature.

In addition, a black body means the internal wall temperature will be uniform. This is realized with walls of excellent thermal conductivity (for instance, with copper), and in perfect thermal isolation from the exterior. The exact form of the internal cavity depends on the temperature at which the black body functions. For example, when a circulating liquid assures temperature uniformity, the cavity appears in the form of tubes that appear much longer than they do wide. Interior copper walls can be covered with a strongly diffusing dielectric (with layered painting, for example), to assure a transmissivity almost isotropic to the cavity's interior.

### 3.1.5. Radiation exchanges between a source and a detector

Radiation detectors are sensitive to the sum total of received radiative fluxes, that is, to the difference between entering fluxes and exiting fluxes. This leads to a measured variable which differs from the desired variable. Let us look at the example given in Figure 3.7, in which a light sensor is used to measure the transmitted flux by the object facing it. We call  $\varepsilon$  the object's transmissivity,  $\varepsilon_s$  the detector's transmissivity and  $\phi_{bo}$  the flux transmitted by the object to be measured. When the detector receives  $\phi_{bo}$ , part of  $\phi_{br}$ , the reflected flux is expressed as:

$$\phi_{br} = (1 - \varepsilon_s) \phi_{bo} \quad [3.23]$$

Furthermore, the detector itself transmits  $\phi_{so}$  and a part  $\phi_{sr}$  that comes back to it after reflection on the object. The state of received and measured flux  $\Phi$  is given as:

$$\Phi = \phi_{bo} - \phi_{br} - \phi_{so} - \phi_{sr} \quad [3.24]$$

We can express this, by calling S the facing surfaces, as:

$$\Phi = \left[ \varepsilon \sigma T^4 - \varepsilon (1 - \varepsilon_s) \sigma T^4 - \varepsilon_s \sigma T_s^4 + \varepsilon_s (1 - \varepsilon) \sigma T_s^4 \right] S = \sigma \varepsilon \varepsilon_s (T^4 - T_s^4) S \quad [3.25]$$

We see that the measured flux  $\Phi$  depends on the detector's temperature  $T_s$  and on its transmissivity  $\varepsilon_s$ . To find the flux  $\phi_{bo}$ , that is, the measurand, the transmissivity  $\varepsilon_s$  must be very close to 1 and that  $T_s \ll T$ . With light sensors, this condition is often not met. Not taking this fact into account can sometimes lead to significant systematic errors.

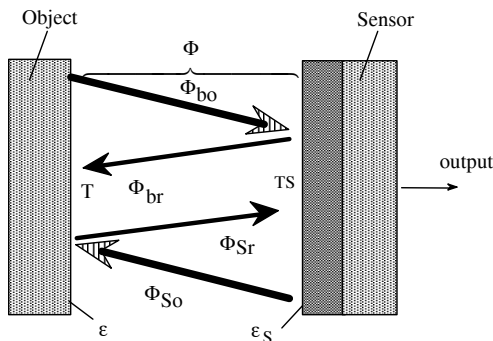


Figure 3.7. Measured flux and measurand

### 3.1.6. Definitions relating to optical sensors

The measurement (output variable) of optical sensors is usually a current. We define sensor performance by looking at variations in currents according to different parameters. Manufacturers of sensors give performance specifications through variables such as darkness currents, spectral sensitivity and specific detectivity, which we will discuss in the following sections. It is important to understand these variables because they are essential guides for anyone using a sensor, giving relevant criteria in the areas of sensitivity, detection limits in power at each wavelength, band pass and noise level, to name a few.

#### 3.1.6.1. Darkness currents

In the absence of any luminous flux, optical sensors almost always transmit a current called the darkness current  $I_0$ . This current is the result of the effects of noise related to influence variables, especially temperature, which create current carriers. The obscurity current fluctuates around its mean value, creating a fundamental noise that limits the detectable minimum amplitude of the luminous flux. For example, with photodiodes, most manufacturers indicate an ultimate detectivity whose value is related only to the value of the darkness current.

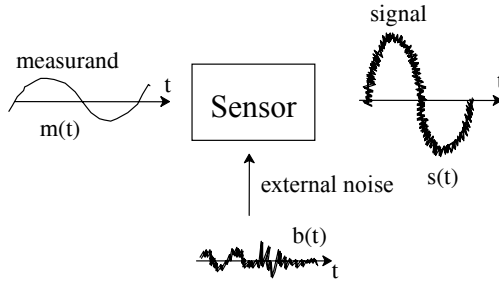
#### 3.1.6.2. Spectral and total sensitivities

When the sensor receives a flux  $\Phi$ , it delivers a current  $I$  that is the sum of the darkness current  $I_0$  and the light current  $I_p$ ,  $I = I_0 + I_p$ . The sensitivity  $S = \frac{\partial I}{\partial \Phi} = \frac{\partial I_p}{\partial \Phi}$  depends only on the light current. The spectral sensitivity is expressed with a monochromatic flux through the wavelength  $\lambda$  by  $S_\lambda = \frac{\partial I_{p\lambda}}{\partial \Phi_\lambda}$ . The total sensitivity  $S_t$  is defined for a flux whose distribution in wavelength is known. For instance, for a radiation whose limits in  $\lambda$  are  $\lambda_1$  and  $\lambda_2$ , we get:

$$S_t = \frac{\int_{\lambda_1}^{\lambda_2} S(\lambda) \left( \frac{d\phi(\lambda)}{d\lambda} \right) d\lambda}{\int_{\lambda_1}^{\lambda_2} \left( \frac{d\phi(\lambda)}{d\lambda} \right) d\lambda} \quad [3.26]$$

#### 3.1.6.3. Sources of fundamental noise in optical sensors

In addition to the sensor signal related to the measurement, there are always some random signals that come from sources internal or external to the sensor. These random signals perturb the measurement.



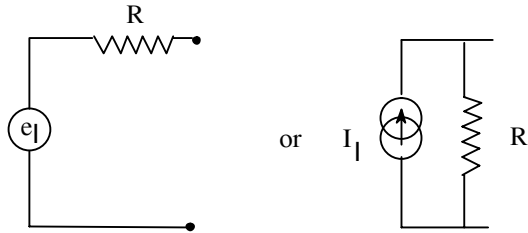
**Figure 3.8.** *External noise*

We say that  $b(t)$  is of external origin when it has an origin coming from the near environment (see Figure 3.8). It is then captured, either directly (an example is radiation interference in the detector band) or indirectly by the entire electric circuit, as with antennae. It is always possible to reduce noise of an external origin by placing the sensor in interference-free areas, by putting them in Faraday cages or shields, and by limiting the sensor's interference sensitivity. When these external noises are eliminated, we see that considerable noise levels still remain. These kinds of noise, called ultimate or fundamental, have their source in the corpuscular nature of electrical currents [BAU 61]. Every free charge is stimulated by random movements, resulting in an output current variation around a mean value. The following list gives two sources of these kinds of fundamental noises:

- thermal or Johnson's noise is the product of the collisions of the carriers with the lattice structure. The greater the number of these collisions, the more the mean quadratic current increases. It appears in the passive component and grows with their resistance. Related to thermal agitation, thermal noise increases with higher temperatures. The equivalent schema of this kind of noise is given in Figure 3.9 where  $\Delta f$  is the band pass of the sensor and of the sensor's electronic system where we get  $e_1 = 2\sqrt{RK T \Delta f}$  or  $I_1 = 2\sqrt{\frac{KT \Delta f}{R}}$ ;

- shot noise or Schottky noise characterizes the discrete nature of the current and obeys Poisson's law, which, unlike the case with Johnson's noise, is present when there are few carriers – that is, in charge-free zones. Shot noise appears in junctions or in a vacuum. The mean quadratic current of this kind of noise is expressed by:

$$\overline{I_S^2} = 2q I \Delta f \quad [3.27]$$



**Figure 3.9.** Equivalent schemata for Johnson's noise

When the current delivered by the sensor is the sum of light and obscurity currents, ( $I = I_0 + I_L$ ), the smallest Schottky quadratic current is:

$$\overline{I_S^2} = 2q I_0 \Delta f \tag{3.28}$$

These first noise sources that do not usually depend on frequency are called “white” noises. The other sources of noise are dependent on frequency. They decrease very quickly with lowered frequencies and are often described as 1/f noise (which is a very rough approximation) or “pink” noise. Pink noises are products of material defects and random recombinations of carriers on the irregularities of the crystalline lattice structures.

For all random uncertainties, the mean quadratic noise currents have to be added to give the total quadratic noise current  $\overline{I_b^2}$  and we call  $i_B$  the spectral noise current (i.e. in a band pass of 1 Hz):

$$i_B = \sqrt{\frac{\overline{I_b^2}}{\Delta f}} (A/\sqrt{Hz}) \tag{3.29}$$

#### 3.1.6.4. Specific detectivity

The Noise Equivalent Power (NEP) is the energetic flux that produces, as sensor output, a photocurrent equal to the spectral current of noise  $i_B$  to the wavelength  $\lambda$ :

$$NEP = \frac{i_B}{S_\lambda} \text{ (in W Hz}^{-1/2}\text{)} \tag{3.30}$$

where  $S_\lambda$  is the spectral sensitivity of the sensor (in  $A \cdot W^{-1}$ ). Some manufacturers sometimes use detectivity  $D$  (the reverse of NEP), but because  $i_b$  is generally proportional to the root of the sensitive surface  $A$  of the sensor, the specific detectivity  $D^*$  is more often used and is expressed as:

$$D^* = \frac{\sqrt{A}}{NEP} \text{ (in } W^{-1} cm \text{ Hz}^{1/2}) = D\sqrt{A} \quad [3.31]$$

For example, for a photodiode we find  $S_\lambda = 0.6 \mu A/\mu W$ ,  $A = 1 \text{ mm}^2$ ,  $I_o = 150 \text{ pA}$ ,  $D^* = 810^{12} \text{ cm Hz}^{1/2} W^{-1}$ . From this we can deduce the mean quadratic value of the total spectral noise current:

$$\overline{i_b^2} = \left( \frac{\sqrt{A} S_\lambda}{D^*} \right)^2 = 56 \cdot 10^{-30} A^2 Hz^{-1} \quad [3.32]$$

We can also obtain the mean quadratic value of the spectral current of Schottky's noise as  $\overline{I_S^2} = 2q I_o = 4,810^{-30} A^2 Hz^{-1/2}$ . From this formula we see that Schottky's noise is predominant in this photodiode. Apart from this kind of calculation, the user can verify that the assembly is not adding too many complementary noise sources to the already existing fundamental noise; in other words, he can check that the assembly is not compromising the performance of the light sensor.

### 3.1.7. Semiconductors: the bases of optical sensors

#### 3.1.7.1. Molecular and crystalline bands

When bringing together two atoms, crossing barriers of potential sometimes produces an aggregate that is the seed of the molecule or crystal. These aggregates, obeying the laws of electrons farthest from the core, are able to retain some of their energy. In addition, this new grouping tends to impart a greater electronic stability to the final state. There are five types of interatomic bands. These are ionic, covalent, metallic, Van der Waals and hydrogen. Some of these bands set electrons free to move in the lattice, producing electrical conductors. The covalent band fixes peripheral electrons to the crystalline mesh so that a covalent lattice does not conduct electricity. It produces insulators and semiconductors (valence states).

We know that the simple elements of the periodic table, peripheral electrons remain stable when the peripheral electronic strata called  $s$  (two electrons) and  $p$  (four electrons) are saturated. This means that the  $2 + 6$  electronic peripheral states

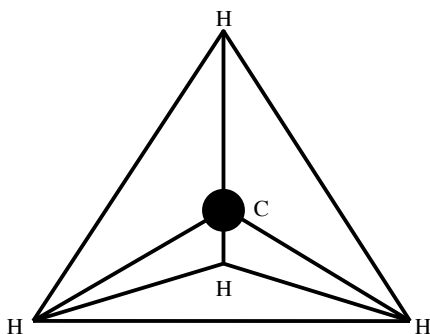


of the element must be occupied. This result is only obtained with rare gases (He, Ne, Ar, Kr, Xe and Ra). All other elements tend to group together, that is, they join with the peripheral electrons so that the strata  $s$  and  $p$  are complete, like those of the rare gases (eight electrons  $s$  and  $p$ ). For example, with carbon (C, atomic number  $Z = 6$ ), there are only four peripheral electrons  $s$  and  $p$ . Four electrons are lacking that are necessary to attain a chemical stability identical to neon ( $Z = 10$ ). Carbon thus tends to combine with the atoms which are capable of “lending” it four electrons. This can occur through covalent band.

Atom H, which has only one electron, therefore tends to gain a second and forms  $H_2$  to attain the stability of He. The molecule  $CH_4$  acquires the requisite structure in bringing the electron of each atom H to the central atom C. C itself takes each of its four peripheral electrons from H atoms placed around itself (see Figure 3.10).

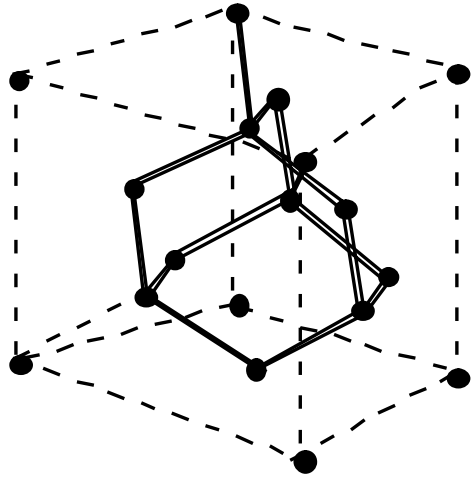
As a result of the formation of this kind of covalent molecule, the energy levels of the electrons reorganize themselves: we speak of orbital states or molecular states.

In other cases, the grouping reaching chemical stability will release many atoms that regroup in the form of a solid. This happens with silicon when a monocrystal is created in a solution. Silicon ( $Z = 14$ ) contains 14 electrons, of which four are valence electrons  $s$  and  $p$  (this situation is analogous to that of carbon). Silicon tends to group with atoms that give it the four electrons  $s$  and  $p$  it lacks to achieve valence saturation (argon stability). In preparing a monocrystal, the atoms of Si group together in a diamond configuration (the crystal of C of the face-centered cubic type) as shown in Figure 3.11. We notice that this structure has the same tetraedric arrangement of  $CH_4$ , but in the monocrystal, this basic figure is reproduced to infinity (in practice, just to the surface where, in fact, chemical stability is no longer a given).



**Figure 3.10.** Covalent bands of  $CH_4$

Cubic crystalization system, with centered faces allowing each atom in the space to be surrounded by four identical atoms

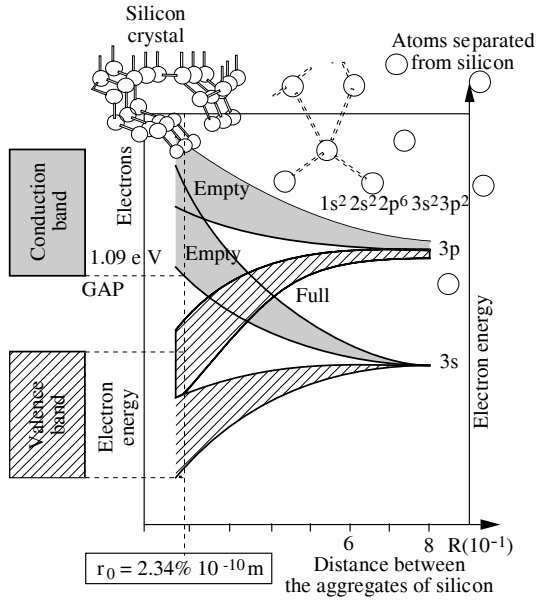


**Figure 3.11.** *Silicon crystal*

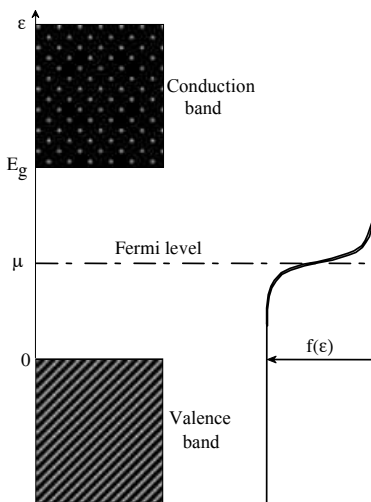
### 3.1.7.2. *Band structures in solids*

Each electron pair found in an orbital shares two neighboring atoms Si (called  $sp_3$  with reference to states  $s$  and  $p$  of the isolated atom it comes from). The crystal structure is fixed by these bands (called  $\sigma$ ) which correspond to the strongly negative energies (they are called strong because a great deal of energy is necessary to break them.) For this process to occur, the distance between the atoms of Si must be very weak ( $2.34 \text{ \AA}$ ).

The Schrodinger equation that regulates the states of the isolated atom Si is itself modified and the eigenstates of this equation become more numerous than for the isolated atom ( $Z$  times the number of atoms in the crystal). The energies regroup in bands in the solid. In a band, the energy levels are very close to each other [ZIM 72]. Figure 3.12 shows that in the case of silicon, two bands appear that are separated by a zone where there are no authorized levels.



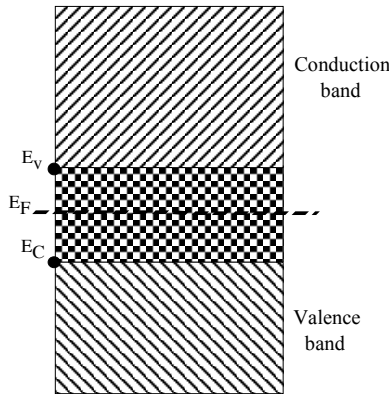
**Figure 3.12.** Modification of electron energy states occurring if the interatomic distance is diminished. A  $r_0 = 2.34 \text{ \AA}$  indicates a formed solid. The energy levels are grouped in two bands. The valence band corresponds to electrons bounded to a silicon lattice. The conduction band corresponds to free electrons (from J.-J. Bonnet [BON 84])



**Figure 3.13.** Organization of bands in a semiconductor or in an insulator

The valence band is the band of lowest energy (that is, greatest absolute value). The conduction band has the weakest absolute values and the band gap is the zone without allowable states [PHI 73]. The probability of the presence of an electron at a given energy  $\epsilon$  is given by the Fermi function  $f(\epsilon)$  (see Figure 3.13). This law is temperature  $T$  dependent. At  $T = 0\text{K}$ ,  $f(\epsilon) = 1$  when  $\epsilon \leq E_F$  and  $f(\epsilon) = 0$  when  $\epsilon > E_F$ . The Fermi level in pure silicon is exactly in the middle of the bandgap when the temperature reaches absolute zero ( $T = 0\text{K}$ ). The valence band is thus saturated and the conduction band is empty (see Figure 3.13).

However, other elements of the periodic table, called transition elements, in the isolated state have many peripheral electrons that occupy states other than  $s$  and  $p$ . These electrons cannot combine in interatomic bands when solids are created. The crystals that come from these elements have an energy state structure in which the bands are not separated by a gap (see Figure 3.14). The crystal has an enormous number of free electrons and nothing stops their displacement when a potential difference is applied to the crystal. In this case, we are describing a metal.

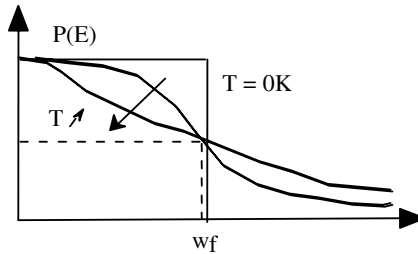


**Figure 3.14.** Organization of bands in a metal

In all solids, when the temperature increases electrons can occupy higher energetic states (less negative) than at absolute zero. Unoccupied places then appear in lower energy levels. The Fermi-Dirac statistics of electron's energetic states shows their probability  $P(E)$  occupies an energy state comprised of  $E$  and  $E + dE$ . We get:

$$P(E) = \frac{dn(E)}{dN(E)} = \frac{1}{1 + \exp\left(\frac{E - E_F}{KT}\right)} \quad [3.33]$$

when  $dn$  is the number of electrons whose energy is comprised of  $E$  and  $E + dE$ , and  $dN$  is the number of possible energetic states (states specific to the Schrodinger equation) between  $E$  and  $E + dE$ . The Fermi level  $E_F$  is therefore the energy for which the probability of an electron existing at this (possibly imaginary) level is  $1/2$  (see Figure 3.15). Of course, at absolute zero, the Fermi level gives the highest energy that an electron can attain (if it is authorized). In other words, all the energy states lower than  $E_F$  are occupied and the higher states are empty.



**Figure 3.15.** *Fermi probability*

In the case of metal, the Fermi level is an eigenstate and for this reason is the last state occupied at absolute zero. With semiconductors and pure (or intrinsic) insulators, the Fermi level is not a proper state but is found exactly in the middle of the restricted band. The Fermi energy is not reached at absolute zero and the electrons saturate the valence band. We see in Figure 3.15 that the probability of finding electrons in the conduction band becomes greater as the temperature increases. These electrons no longer take part in crystalline bands and become free carriers, as in metal. This increase of the number of electrons in the conduction band also frees an equal number of spaces in the valence band. These spaces act as positive mobile charge carriers.

The practical uses of semiconductors derive from the fact that the position of the Fermi level inside the gap can be moved by introducing carefully selected impurities into the crystal [SAP 92]. When these impurities have five peripheral electrons (donors), the Fermi level moves towards the top of the gap. The probability of finding electrons in the conduction band increases and very few holes remain in the valence band. This kind of semiconductor is called type N to remind us that most current carriers are electrons. On the other hand, when impurities are elements with only three peripheral electrons (acceptors), the Fermi level moves towards the bottom of the gap. The probability of finding electrons in the conduction band becomes very low, and there are many empty spaces in the valence band. In this case, the carriers are mostly holes (semiconductor type P).

### 3.1.8. Current expression in a material containing free charges

As soon as we apply a tension to the limits of a material that contains free charges, this tension forms an electric field  $\vec{E}$  and the Coulomb force  $\vec{F} = q\vec{E}$  sets these charges in motion. They take on a speed  $\vec{v}$  and the current  $I$ , formed by this movement, is the quantity of charges crossing the material per second.

When impurities are present, or because of thermal agitation or photon absorption (photoelectric effect), certain electrons go into the conduction band and holes appear [SMI 67]. These carriers are then free to create an electrical current when they are exposed to field  $\vec{E}$ . This current is the density flux of the current density  $\vec{J}$  across the section of the material. If we call  $\rho$  the number of mobile charges by unit volume and  $\vec{v}$  the carrier velocity, the current density  $\vec{J}$  in a conductor with electron type carriers ( $n$  by unit of volume  $e$ , charge of the electron) is given as:

$$\vec{J} = \rho\vec{v} = -ne\vec{v} \quad [3.34]$$

In addition, the relation between the speed  $\vec{v}$  of the electron mass  $m_e$  and the field  $\vec{E}$  is given by the Coulomb law and the fundamental relation of the dynamic, and is expressed by:

$$\vec{F} = -e\vec{E} = m_e \frac{d\vec{v}}{dt} \quad [3.35]$$

$$\frac{d\vec{v}}{dt} = \frac{-e}{m_e} \vec{E} \quad [3.36]$$

The resolution of this equation leads to a speed  $\vec{v}$  that linearly increases with time:

$$\vec{v} = \frac{-e}{m_e} \vec{E} t \quad (+ \text{constant} = 0 \text{ if } \vec{v} = 0 \text{ at } t = 0) \quad [3.37]$$

Actually, this law is completely unrealistic because the electron ( $e^-$ ) undergoes many collisions with the crystalline lattice. If we call  $\tau$  the mean time between two collisions, the average velocity is limited to the value it achieves at the end of this time  $\tau$ :

$$\vec{v} = \frac{-e}{m_e} \vec{E} \tau \quad [3.38]$$

The quantity  $\frac{e}{m_e} \tau$  is called mobility  $\mu_e$  of  $e^-$ :

$$\mu_e = \frac{e\tau}{m_e} \quad [3.39]$$

This relation is general for all charge types, and we get:

$$\mu = \frac{q\tau}{m} \quad [3.40]$$

where the quantity  $q$  is the absolute value of the mobile charge. The speed  $\bar{v}$  is thus an average velocity between two successive collisions and this velocity is equal to:

$$\bar{v} = \pm \mu \bar{E} \begin{pmatrix} + : \text{charges} > 0 \\ - : \text{charges} < 0 \end{pmatrix} \quad [3.41]$$

We see that  $\bar{v}$  is proportional to  $\bar{E}$  and independent of time because of the many collisions. The expression of current density  $\vec{J}$  can be given in the form:

$$\vec{J} = + ne \mu \bar{E} \quad [3.42]$$

The current density is proportional to  $\bar{E}$  and we get:

$$\vec{J} = \sigma \bar{E} \quad [3.43]$$

which is the local Ohm law with  $\sigma$  the given electric conductivity, expressed in the case of conduction by  $e^-$  by:

$$\sigma = ne\mu \quad [3.44]$$

We can generalize these relations by calling  $\mu_e$  and  $\mu_p$  the mobilities of  $e^-$  and of the holes ( $e^+$ ):

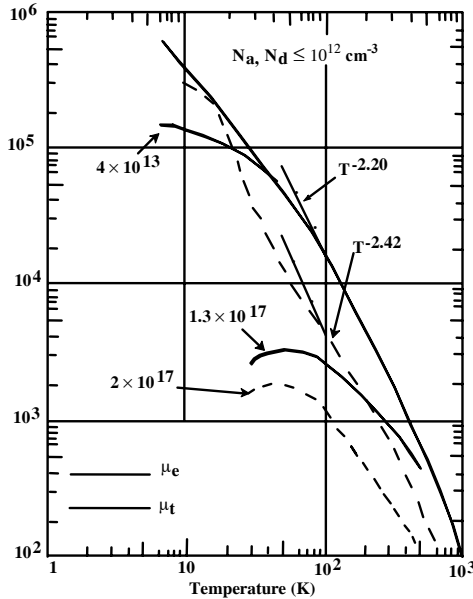
$$\vec{J} = \vec{J}_n + \vec{J}_p = e (n\mu_n + p\mu_p) \bar{E} \quad [3.45]$$

where  $n$  and  $p$  are the densities of  $e^-$  and of  $e^+$  and where  $\mu_n$  and  $\mu_p$  are given by:

$$\begin{aligned}\mu_n &= \frac{e \cdot \tau \cdot n}{m_n} \\ \mu_p &= \frac{e \cdot \tau \cdot p}{m_p}\end{aligned}\quad [3.46]$$

Then conductivity is written as:

$$\sigma = e [p \mu_p + n \mu_n] \quad [3.47]$$



**Figure 3.16.** Mobility variation ( $\text{cm}^2/\text{V s}$ ) of electrons  $\mu_e$  and of the holes  $\mu_p$  in silicon according to the temperature for materials with different dopings. The mobility of  $e^-$  is represented in continuous traits, with those of  $e^+$  in discontinuous traits. The indicated gradients correspond to the best linear adjustment of the experimental curves

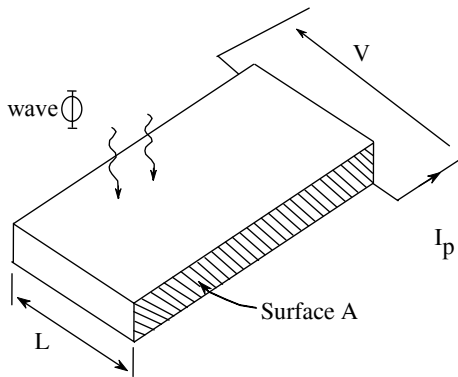
In reality, the problem is more complicated than it appears in this simple demonstration. For any given material, mobilities in particular depend both on doping and on temperature. Figure 3.16, as it relates to silicon, shows that  $e^-$  are more mobile than  $e^+$  and that mobility decreases rapidly with increased



temperatures; typically they are reduced from 1 to 2 orders of magnitude between 77 and 300 K. This complexity aids sensors because in selecting the doping and material type, we can produce components that are adapted to the desired measurand (see the following section for an example of this).

### 3.1.9. Photoconductor cells

Photoconductor cells are the simplest optical sensors to make use of semiconductors. Their basic operating principle depends on the photoelectric effect, in which a photon is absorbed by the semiconductor in giving its energy (by absorption) to an electron of the conduction band. In other words, the electron is freed from the crystalline bands and moves under the influence of an external electric field. Macroscopically, light absorption by the semiconductor means a rise in its conductivity, while its resistance diminishes. These sensors are made of a semiconductive plaquette with a large face that receives light and with two lateral faces of surface A that are metallized to be used as collection electrodes (see Figure 3.17). The sensor is passive, polarized by an exterior tension  $V$ . We carry out measurements according to the illumination received (the measurand) and the variation of the electrical resistance (the measurement). In the following paragraphs we will consider material of type N, which means the carriers are essentially of type  $e^-$ .



**Figure 3.17.** Schema of a photoconductive cell

Two phenomena occur in the semiconductor. These are: (i) carrier generation through the photoelectric effect and (ii) carrier recombination on the crystalline lattice.

(i) As we saw before, light must be represented in the form of photons that are energy grains  $h\nu$ . If  $h\nu > E_g$  (the gap energy)  $G$  electrons are created per second throughout the volume of the sensor or  $g$  electrons per second and by unit of volume:

$$g = \frac{G}{AL} = \frac{1}{AL} \frac{\eta(1-R)}{h\nu} \Phi = \frac{1}{AL} \frac{\eta(1-R)}{hc} \lambda \Phi \quad [3.48]$$

where A.L. is the sensor volume,  $\eta$  is the photon conversion into electrons,  $R$  is the coefficient of optical reflection of the receiving surface,  $\Phi$  is the flux incident of light on the sensor, and  $\lambda$  is the wavelength of the light flux.

(ii) The electrons freed by the photoelectric effect leave in the crystal many charged atoms that can trap them after their displacement. The variation by unit time of the number of free electrons occurring because of this recombination is proportional to the number of free electrons produced, as well as to the number of charged atoms. We get:

$$\frac{\partial n}{\partial t} = -r n^2 \quad [3.49]$$

where  $r$  is called the recombination rate. At equilibrium, there are as many electrons created as recombined:

$$\frac{\partial n}{\partial t} = 0 = -r n^2 + g \quad \text{from which we get } n = \sqrt{\frac{g}{r}} \quad [3.50]$$

We have seen that electrical conductivity is expressed by:

$$\sigma = e \mu_n n \quad [3.51]$$

since the majority carriers are electrons. In the following equation, we see that by replacing  $n$  with its expression as  $\Phi$  (see [3.49] and [3.51]), we get:

$$\sigma = e \mu_n \sqrt{\frac{1}{AL} \frac{\eta(1-R)}{hc} \lambda \Phi} \quad [3.52]$$

When the sensor is polarized by the tension  $V$ , the current going across the sensor is equal to:

$$I = \frac{V}{R} = \frac{V}{\rho \frac{L}{A}} = \frac{\sigma A V}{L} \tag{3.53}$$

Relationship [3.53] shows that the measurement  $I$  is not proportional to the measurand  $\Phi$  but to its root. The resistance  $R$  of the sensor, inversely proportional to  $\Phi^{-1/2}$ , tends toward infinity when  $\Phi$  tends toward zero. In practice, this value is obviously finite because there are always carriers even when light is absent. This darkness resistance  $R_0$  depends on the semiconductor being used and the geometric form of the sensor. This resistance can go from several dozen Ohms to several hundred M $\Omega$ .

In fact, the experimental dependence of resistance with the flux  $\Phi$  is not exactly the same as with the model shown above. This dependence, however, is not linear but is like the following:

$$R = k \phi^{-\gamma} \text{ with } 0.5 \leq \gamma \leq 1. \tag{3.54}$$

The number of carriers recombined by unit time can also be expressed with the help of the lifetime of the carriers  $\tau_n$  (which are the  $e^-$  in the examples given above):

$$\frac{\partial n}{\partial t} = -\frac{n}{\tau_n} \tag{3.55}$$

At equilibrium we get:

$$g = \frac{n}{\tau_n} \tag{3.56}$$

which lets us write the current delivered by the sensor as:

$$I = \frac{V}{R} = V \sigma \frac{A}{L} = V q \mu_n \tau_n \frac{G}{AL} \frac{A}{L} = \frac{V \mu_n \tau_n}{L^2} q G = F q G \tag{3.57}$$

where  $F$  is called the sensor's gain factor:

$$F = \frac{\tau_n \mu_n}{L^2} V \quad [3.58]$$

$F$  can attain several tens of thousands following the applied tension  $V$  and the geometric form of the sensor. In addition, this equation shows that it is necessary to use semiconductors with excellent mobility and lifetime.

The equivalent schema of photoconductive cells involves placing the darkness resistance  $R_0$  in parallel to a resistance that is sensitive to the flux  $R_{cp}$ . In practice, this schema can be reduced to one resistance  $R$ , expressed by:

$$R = \frac{R_0 R_{cp}}{R_0 + R_{cp}} \approx R_{cp} = a \varphi^\gamma \quad \text{if } R_0 \gg R_{cp} \quad [3.59]$$

The relation between the light current  $I_p$  and the flux  $\Phi$  is not linear:

$$I_p = \frac{V}{R} = \frac{V}{a} \varphi^\gamma \quad [3.60]$$

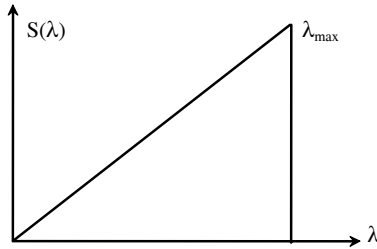
The total sensitivity depends on the value of the detected flux:

$$S = \frac{\partial I_p}{\partial \phi} = \gamma \frac{V}{a} \phi^{\gamma-1} \quad [3.61]$$

Furthermore, the spectral sensitivity can be deduced from the above formulae:

$$S(\lambda) = q \frac{\tau_n \mu_n V}{L^2} \eta \frac{(1-R)}{h_c} \lambda \quad [3.62]$$

This is applicable up to  $\lambda_{\max}$  of the gap. If there is an abrupt gap, the spectral sensitivity ends after  $\lambda \geq \lambda_{\max} = \frac{h_c}{E}$ , shown in Figure 3.18.



**Figure 3.18.** Theoretical spectral sensitivity of a photoconductive cell

Response time is directly related to the lifetime of the carriers  $\tau_n$ . Like the lifetime, response time is connected to parameters such as the temperature and doping of the semiconductor. If we know the nature and operating mode of the semiconductor, we can obtain values that range from 0.1s to  $10^{-7}$ s. Response time is noticeably reduced when the luminous flux is high; this is because the lifetime decreases with the number of free carriers.

The ultimate noise of these sensors is of the Johnson noise type and its minimum value depends on the value of the darkness current  $I_0$ . This obscurity current comes from the creation of carriers (for example, of electrons) by thermal agitation. In calling  $n_0$  the number by unit of volume of the thermal carriers, the electric conductivity in obscurity is  $\sigma_0$ :

$$\sigma_0 = e\mu n_0 \tag{3.63}$$

from which we get:

$$R_0 = \frac{1}{e\mu_n n_0} \frac{L}{A} \tag{3.64}$$

this helps us determine the current of Johnson's noise traversing  $R_0$ :

$$i_B = 2\sqrt{\frac{KT}{R_0}} \tag{3.65}$$

We remember that  $R_0$  largely depends on the kind of semiconductor used and on the temperature. Typically, the specific detectivities  $D^*$  are of the order of

$10^{10} \text{ W}^{-1} \text{ cmHz}^{1/2}$ , clearly inferior to what can be obtained with photodiodes, as we will see in the following sections.

### 3.1.10. P-N junction and photodiodes

Introducing impurities into a semiconductor displaces the Fermi levels, so that the current carriers become electrons (type N) or holes (type P). The P-N junction is basic to the creation of photodiodes. It is obtained by producing on the same semiconductor substrata two adjacent doping zones P and N. The junction has an advantage over a simple semiconductor because it creates a charge zone of space with a powerful electrical field. This field can efficiently separate charges created by photoelectrical effect, as well as significantly improving light sensor detectivity.

#### 3.1.10.1. Non-polarized junctions

Without bias, an electrostatic equilibrium appears between the two zones separating the junction. This equilibrium is converted by the equalization of the Fermi levels of the P and N regions. Both the energies  $W_{V_P}$  and  $W_{V_N}$  at the top of the valence bands, and  $W_{C_P}$  and  $W_{C_N}$  at the bottom of the band are displaced in the regions P and N and we get:

$$W_{C_P} > W_{C_N} \quad [3.66]$$

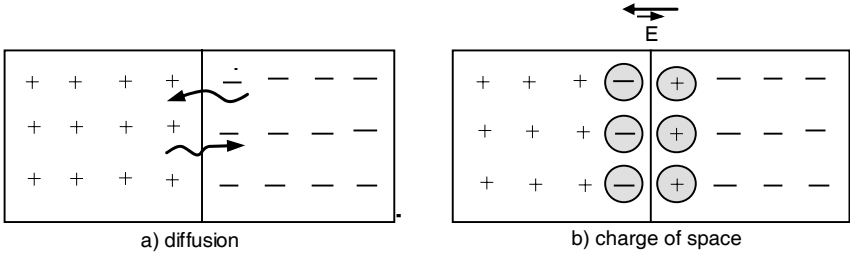
that is:

$$|W_{C_P}| < |W_{C_N}| \quad [3.67]$$

$$V_P < V_N \quad [3.68]$$

The electrostatic potential of region P has become inferior to that of region N. An electrostatic field  $\vec{E}$  has appeared, directed from N towards P. The majority carriers of each region are, for the most part, incapable of crossing this potential barrier. However, the minority carriers of each region do cross the barrier. They are launched by the field  $\vec{E}$  towards the adjacent region. Again, the equilibrium is converted by the equivalence of two currents (flowing in opposite directions) from majority carriers with enough kinetic energy to cross both the barrier and the minority carriers launched by  $\vec{E}$ .

It is important to remember that the barrier of potential  $V_N - V_P = V_B$  is an electrostatic tension which corresponds to an absence of current. This difference of potential converts an equilibrium and not an electromotive force.



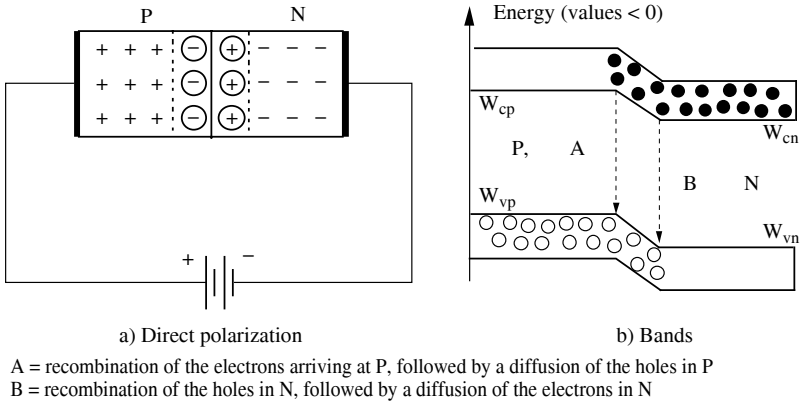
**Figure 3.19.** Diffusion and establishment of carriers after recombination of the static potential barrier in a non-polarized diode

3.1.10.2. P-N junction with direct bias

Applying a positive difference of potential between the limits P and N of a junction allows a large number of majority carriers to cross the junction. If the applied tension is superior to the electrostatic tension of the barrier (0.7 V in the Si),  $V_p$  becomes superior to  $V_N$  ( $W_{C_N} > W_{C_P}$ ). The electrons, very numerous in N, go into P, where they recombine in the many holes in that region. The current goes through P by diffusing through the holes towards the junction to fill the deficit produced by the electron recombinations near the junction. A direct current is essentially a current of majority carriers. In practice, this is the sole contribution of the total current as soon as  $V_A > V_B$ . It grows exponentially with  $V_A$ :

$$I \approx I_{majority} = I_0 \exp \frac{qV_A}{kT} \tag{3.69}$$

where  $q$  is the electron charge  $1.6 \cdot 10^{-19}$  C and  $k$  is the Boltzmann constant  $+ 1.36 \cdot 10^{-23}$  JK<sup>-1</sup>.

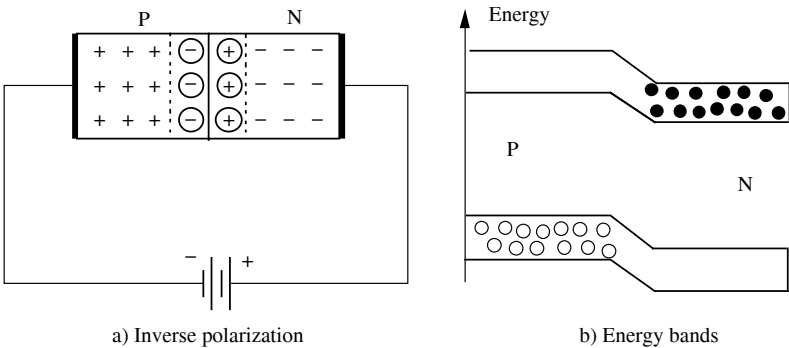


**Figure 3.20.** *P-N junction in direct bias*

3.1.10.3. *P-N junction in reverse bias*

When we apply a negative tension  $V_A$  between the extremities P and N of the junction, the electrostatic barrier is again amplified. The majority carriers capable of crossing the barrier become fewer. The current of the majority carriers which, before negative tension  $V_A$  was applied, exactly offset the current of the minority carriers, become very weak in comparison to the minority carriers current. The current crossing the inversely polarized junction is almost exclusively a current of minority carriers. This current is, by its very nature, opposed to a current obtained through direct bias, which fundamentally comes from majority carriers. This reverse current is independent of  $V_A$  because the movement of minority carrier across the barrier is produced by diffusion and not by a field effect. This is shown in Figure 3.21.

$$I_{\text{minority}} = -I_0 \tag{3.70}$$



**Figure 3.21.** *P-N junction in reverse bias*



## 3.1.10.4. Diode equation

No matter what  $V_A$  tension is applied to the limits of the diode the current is always the sum of the majority carriers current and of the majority carriers current:

$$I = I_{\text{majority}} + I_{\text{minority}} \quad [3.71]$$

We get:

$$I = I_0 \exp \frac{qV_A}{kT} - I'_0 \quad [3.72]$$

We know that for  $V_A = 0$ , the equilibrium imposes  $I = 0$ , so we get:

$$I'_0 = I_0 \quad [3.73]$$

Finally, the diode equation is written as:

$$I = I_0 \left[ \exp \frac{qV_A}{kT} - 1 \right] \quad [3.74]$$

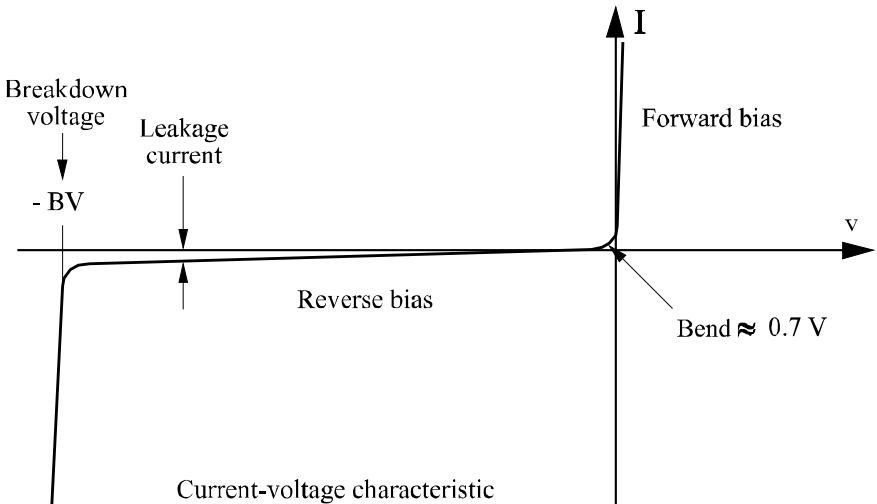


Figure 3.22. Typical current tension of a diode

Figure 3.22 shows us that a high negative current appears when reverse bias becomes strongly negative. This is called the breakdown voltage and is produced by a collision of minority carriers, which take on a high kinetic energy in crossing the barrier, ionizing the fixed centers of the crystalline lattice. This effect benefits some photodiodes by amplifying the photo current and reducing response time of light sensors.

### 3.1.10.5. Illuminated P-N junctions

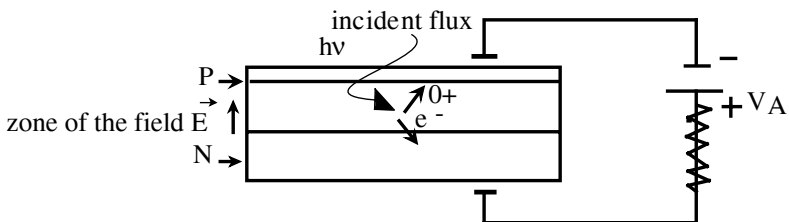
When diodes are exposed to a luminous flux with a wavelength  $\lambda$  that is inferior to the wavelength  $\lambda_s$ , corresponding to the gap energy (or the energy  $h\nu$  of the flux photons superior to the gap energy  $h\nu_s$ ), they produce electron hole pairs through the photoelectric effect, as we have seen with photoconductive cells. For these carriers to generate a current, the pair must be separated quickly. This only happens through the influence of electrical field, especially the one existing in the P-N junction. The photoelectrical effect must then be produced throughout the zone of the field  $\vec{E}$  of the P-N junction. Through the stimulating effect of the field  $\vec{E}$ , the generated photon holes go towards region P and the electrons go towards region N. When the holes are in P, they encounter negatively charged impurity sites which trap them. These sites are almost always taken to be an  $e^-$  from the atom's valence Si. The photon holes are trapped by these impurities; this produces a liberation of  $e^-$ . In region P, the electrons are diffused, moving toward the junction and filling in the deficit thus created. The same thing happens in region N.

We can see that the photoelectrical current acts as a current of minority carriers and therefore is negative.

### 3.1.10.6. Principle of photodiode fabrication

Several conditions must exist for the fabrication of photodiodes (see Figure 3.23):

- the photodiode must have a significant field of junction  $\vec{E}$  to efficiently separate the created photo carriers. Thus, it is clear that in this case, the diode must be polarized in reverse.



**Figure 3.23.** Schematic view of a photodiode

– in order for the photons to penetrate the field zone in large numbers, the incident flux must not be weakly absorbed by region P. Assume  $x$  is the thickness of region P and  $\alpha$  is its optical extinction coefficient at the frequency  $\nu$  of the photons of the monochromatic flux. The part  $\Phi_0$  that comes to the junction after reflection on the face in front and transmission by the doped zone P (see Figure 3.23) is equal to:

$$(1 - R_{opt})\phi_0 e^{-\alpha x} \tag{3.75}$$

where  $R_{opt}$  is the reflection coefficient of the photodiode surface. This transmitted flux will be accordingly greater than  $\alpha \cdot x$  and will be weaker if the entire flux coming into the junction is absorbed mainly in the field region;

– a field region must be designed that is sufficiently thick to allow the total absorption of the light. This becomes possible, for example, by creating the type of structure often called junction type PIN (P-intrinsic N).

### 3.1.10.7. Photodiode equation

The light current  $I_r$  is equal to the number of electrons (or number of holes) that have been created by the photoelectrical effect and that have come to the limit of the field zone per second:

$$|I_r| = \frac{q\eta(1-R_{opt})\lambda}{hc} \phi_0 e^{-\alpha x} \tag{3.76}$$

Because this current is created by minority carriers, it becomes entrenched in the diode current. The equation (or feature) of the photodiode is:

$$I = I_0 \left[ \exp\left(\frac{qV_A}{kT}\right) - 1 \right] - |I_r| \tag{3.77}$$

or:

$$I = I_0 \left[ \exp\left(\frac{qV_A}{kT}\right) - 1 \right] - \frac{q\eta(1-R_{opt})\lambda}{hc} \phi_0 e^{-\alpha x} \tag{3.78}$$

Equation [3.78] shows that if the dark current  $I_0$  is weak with respect to  $I_r$ ,  $I$  is proportional to  $\Phi_0$  when  $V_A$  is negative.

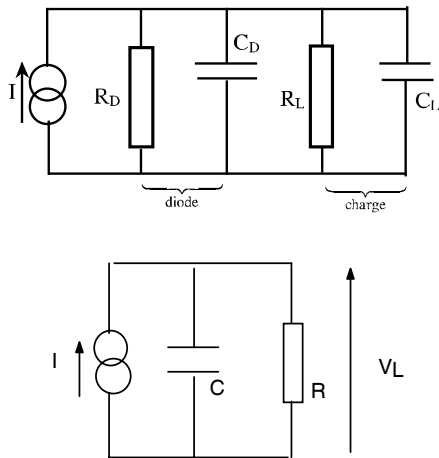
### 3.1.10.8. Electrical schema for a diode

Photodiodes are current sources, being the sum of the obscurity current and light current. They are in parallel with the resistance and the capacity of the reverse polarized junction. On the other hand, the charge impedance of the photodiode can

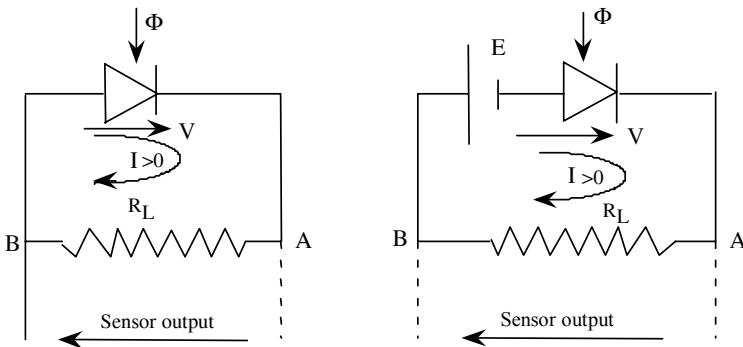
always be brought back to a resistance  $R_L$  in parallel with a capacity  $C_L$  (see Figure 3.24).

Photodiodes can be used as photoconductors with a reverse bias or as photovoltaic sources without bias (see Figure 3.25). These two functioning modes need some explanation in order to choose correctly between them.

Using photodiodes as photoconductors allows for a reduced response time because reducing the diode capacity results in a shrinking of the junction zone in reverse bias.

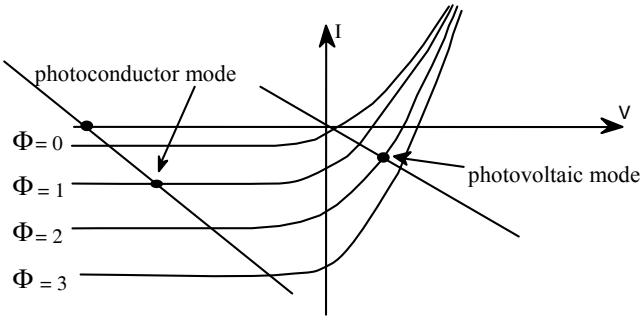


**Figure 3.24.** Equivalent schema for photodiodes



**Figure 3.25.** Photovoltaic and photoconductor modes

With the photoconductor mode, there is always a darkness current that produces an internal noise of the Schottky type, which means that this mode is not especially favorable for detecting very weak fluxes. On the contrary, with the photovoltaic mode, the straight line of the charge goes through the origin  $I = V = 0$ , and the darkness current no longer limits the very weak flux measurements. In the photoconductor mode, photodiodes are linear, but in the photovoltaic mode, photodiodes behave in a logarithmic fashion, except under very weak charges when a photovoltaic generator is used with a battery charge as for solar cells (see Figure 3.26).



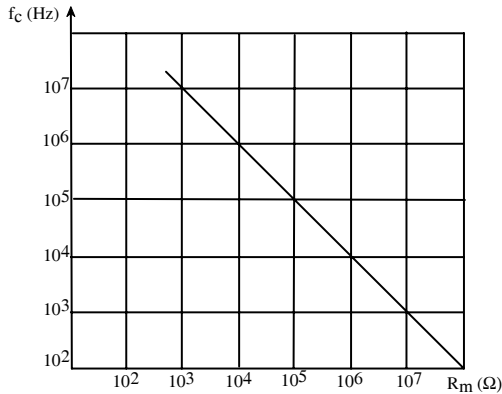
**Figure 3.26.** Functioning points of photoconductor and photovoltaic modes

Table 3.1 reviews some of the points just covered.

Implementation	Photoconductor mode	Photovoltaic mode
Bias	Reverse	No
Output signal	$I_T = I_0 + I_L$	$V_{c0} = \frac{kT}{q} \cdot \ln\left(\frac{I_L - I_0}{I_0}\right);$ $I_{cc} \approx I_L$
Advantages	Wide band pass, short response time	Low noise

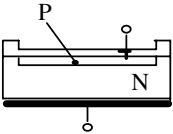
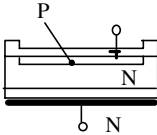
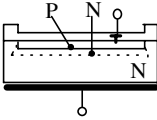
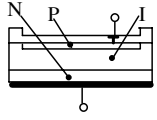
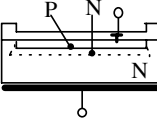
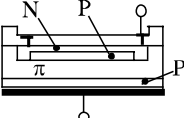
**Table 3.1.** Choice criteria for photoconductor and photovoltaic modes

Since the equivalent final schema can be reduced to the one shown in Figure 3.24, photodiodes are first order sensors (see Chapter 2). The cut-off frequency given by  $f_c = 1/2\pi RC$  depends on the charge resistance R (see Figure 3.27).



**Figure 3.27.** Example of a photodiode cut-off frequency versus charge resistance  $R_m$

Depending upon the structure of the different doped regions of a photodiode, the performances are different. Table 3.2 reviews some of these possibilities. In the schemata of Table 3.2, we see that the front face of the sensors, excepting the Schottky structures, is covered with a fine protective layer of transparent  $\text{SiO}_2$ . The first semiconductive layer is generally of type P. This doping is often preferred for its simplicity of preparation and transparency. The first schema (called planar type) corresponds to the junction already described. The interfaces are flat and establish a potential barrier between P and N. For those applications that concern the visible light and are in use, this technique, used in silicon, is sufficient. When response time needs improvement, the sensor's thickness must be augmented, which reduces its capacity. To effect this, a structured layer can be created by successive dopings (weak planar capacity). A progressive passage (type P-N-N) can also be created to realize the P-N junction. The spectral sensitivity is then modified by ensuring that certain wavelengths are not absorbed into the depletion zone. This allows better sensitivities in the infrared or ultraviolet. As mentioned above, the best way to reduce diode capacity and improve the separation of photogenerated couples is by creating an enlarged depletion zone. This is done by inserting an intrinsic stratum between P and N. This produces rapid photodiodes (type P-I-N). Flux detection in the ultraviolet field is difficult with semiconductive junctions because their fluxes are rapidly absorbed in the P region when the wavelength decreases. To avoid this problem, metal semiconductive Schottky junctions are realized. These junctions are similar to P-N junctions, especially relating to their potential barriers. Although this kind of junction has a narrower depletion zone, the metallic stratum can be made very transparent to ultraviolet rays. Often, the width of the charge zone of space can be augmented by creating a dopage gradient in the semiconductive area. In addition, an internal amplification can be produced by a correct and controlled breakdown stage.

Type	Construction	Properties	Material
Planar		Weak darkness current	Silicon or GaAsP
Planar, low capacity		Weak darkness current, rapid response, strong sensitivity in UV and IR rays	Silicon
PNN		Weak darkness current, strong sensitivity in UV rays, insensitivity to IR rays	Silicon
PIN		Very rapid response	Silicon
Schottky		Very high sensitivity in UV rays	GaAsP, GaP
Avalanche		Internal amplification, ultra rapid response	Silicon

**Table 3.2.** Examples of adaptability of photodiode structures to measurands

The breakdown stage also influences the noise current so that photodiode detectivity does not increase proportionally to the internal gain. In fact, the resulting strong bias from the breakdown stage augments the width of the charge zone, considerably reducing response time. Avalanche photodiode structures always thicken gradually (zone  $\pi$ ), so that the electrical field is moderated, even under strong tension, sometimes up to 1,000 V. Photodiode structures can also be created heterogeneously with different kinds of semiconductors. In these cases, we speak of heterojunctions as opposed to homojunctions created by dopings of a single semiconductor. Since we have been for the most part discussing visible effects, we have been describing photodiodes essentially as silicon homojunctions.

### 3.2. Force and deformation sensors

Among the range of mechanical sensors (position, speed, acceleration, shock, among others), we propose discussing two typical examples, one passive, the other active, whose measurands are forces and deformations.

Passive sensors are mostly used in mechanics. In Chapter 2, we saw that these sensors can be resistive, capacitive or inductive. Inductive sensors are often used for displacement measurements. On the other hand, resistive sensors are often used for deformation measurements and are sometimes called, somewhat incorrectly, constraint gauges.

The piezoelectric effect [CAD 64] is the most widely used basic principle in active mechanical sensors. It is used in its simplest form with force and deformation sensors. In the following sections, we will explain the principles of piezoelectricity and some methods used to analyze the signals it generates.

It is important to remember that very little difference exists between force and constraint measurements. In order to measure a force, a kind of dynamometer is generally used. This is a tool that helps us establish an equilibrium between the force we want to measure and the constraint produced by the deformations undergone by a solid that makes up a part of the sensor under the action of this force. Working in the elastic domain of deformations, we see that constraints and deformations are proportional. A simple calibration allows the same sensor to carry out both a constraint measurement (or force by surface unit) and a deformation measurement.

#### 3.2.1. Resistive gauges

Resistive gauges are simply resistive circuits that can be attached to a structure to determine its local deformations. These kinds of resistances represent an important percentage of deformation sensor sales. Their ability to function in many conditions and their low prices explain their widespread usage. In addition to these long-known assets, more recently gauges have been developed that aid in producing very small, high-resolution sensors. As well, these resistive gauges are associated with proof bodies and conditioners that improve sensitivities and signal to noise ratio. Following these processes, the measurable relative elongations go from  $10^{-7}$  to  $10^{-1}$ . The relative deformation error, seldom below to  $10^{-3}$ , is more often of the order of  $5 \cdot 10^{-3}$  to  $10^{-2}$ . These gauges are sometimes in the form of wire, sometimes thin layers of some material, or sometimes they are created by doping in the semiconductors. For a wire of section  $S$  and the width  $l$  made of a material of resistivity  $\rho$ , the

relative variation  $\frac{\Delta R}{R}$  of the resistance  $R$  given by  $R = \frac{\rho l}{S}$  is written:



$$\frac{\Delta R}{R} = \frac{\Delta \rho}{\rho} + \frac{\Delta l}{l} - \frac{\Delta S}{S} \quad [3.79]$$

If we call  $d$  the diameter of the section of surface  $\pi d^2/4$ , we get, by introducing the Poisson coefficient  $\frac{\Delta l}{l} = -\frac{1}{\nu} \frac{\Delta d}{d}$  and in writing  $\frac{\Delta \rho}{\rho} = C \frac{\Delta V}{V}$ :

$$\frac{\Delta R}{R} = \left[ (1 + 2\nu) + C(1 - 2\nu) \right] \frac{\Delta l}{l} = k \frac{\Delta l}{l} \quad [3.80]$$

where  $k$  is the gauge factor and  $C$  is Bridgman's constant. The resistance values are usually from several hundreds to several thousands of ohms. With metals, when  $\nu = 0.3$  and  $C$  is of the order of 1, we get  $k$  of the order of 2. With semiconductors,  $C$  can reach 200 and the gauge factor is high, of the order of  $C$ . This means that measuring very weak deformation must be done with semiconductive gauges, but in this case it is important to remember that  $R$  is very dependent on the temperature, which, in practice, limits the use of these gauges to temperatures below 200°C.

There are gauge assemblies able to determine deformation components following several axes. When the situation does not help us know the main deformation directions, we use gauges grouped in three resistances, each 120° from the other, that is, in rosettes. The term rosette extends to more complex deformation measurements known as shell deformations.

The main shortcoming of these gauges is that they must be attached to the structure. This limits their use to medium temperatures (up to 500-600°C for metallic wire gauges). At higher temperatures, resistive gauges are increasingly being replaced by optical methods using coherent light beams, among them speckle interferometry.

### 3.2.2. Piezoelectric effect

Piezoelectricity derives its name from the Greek word "piezo", meaning "to press". The piezoelectric effect is the conversion of pressure into electricity. To be more precise, the term describes the appearance, due to the action of microscopic deformations, of charges on the surface of a solid. These charges are produced by local displacements of centers linked to the crystalline mesh.

In fact, the piezoelectric effect only exists in crystals, ceramics and polymers that are anisotropes; that is, that have no symmetrical center in the elementary mesh.

Although the Curie brothers discovered piezoelectricity in 1880, it was only put to practical use during World War I, first by Paul Langevin, who developed sonar, then in 1918 by Walter Cady, who built the first quartz oscillator. Today, the popularity of high quality quartz oscillators makes piezoelectricity an integral part of electronics. The production of products using piezoelectricity employs hundreds of thousands of people throughout the world.

### 3.2.2.1. *Electrostriction, piezoelectricity and pyroelectricity*

Because the electric field is zero in metals, piezoelectricity cannot exist in it. However, when we apply an electric field to a dielectric, the equilibrium positions of electric charges bound to a solid undergo a slight displacement. In the absence of a permanent dipole moment (corresponding to the moment when barycenters of positive and negative charges join), a related dipole moment appears. In the presence of a permanent dipole moment, its value is changed. The charge displacements lead to a geometric deformation of the solid, with mechanical constraints compensating the electric forces produced by the action of the resulting dipole moment.

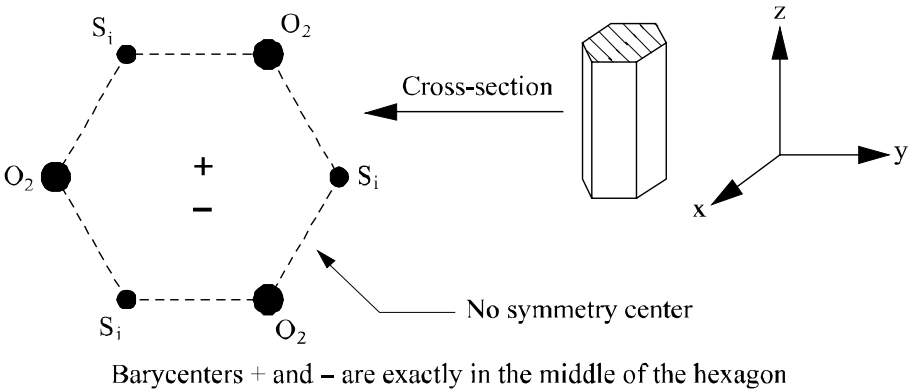
When the solid has a center of symmetry, the deformation is very weak and is proportional to the square of the applied electric field: this is electrostriction.

When the solid does not have a center of symmetry, the charge displacement, which is clearly more important, is proportional to the applied field. This is the inverse piezoelectric effect. Often, when there is no center of symmetry, the material already has a permanent piezoelectric bias. This dielectric bias varies not only according to the applied field but also according to temperature: this is the piezoelectric effect.

The following four points summarize these effects: in metals there is no effect; in dielectrics with a center of symmetry, the weak effect is called electrostriction; in dielectrics without centers of symmetry, their effect becomes stronger and is called piezoelectricity; and in anisotropic dielectrics with permanent bias, there is a piezoelectric disturbance of the piezoelectricity because the temperature is in this case an intruding influence variable.

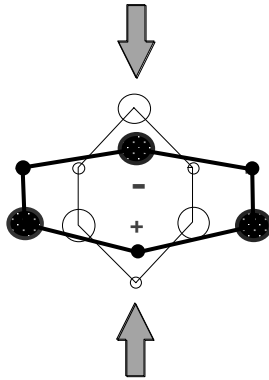
### 3.2.2.2. *The case of quartz*

Quartz is a silica crystal composed of  $\text{SiO}_2$  in which the peripheral electrons of Si atoms and  $\text{O}_2$  groups of chemicals combine and reach the stability level of rare gases (see section 3.1 for a discussion of chemical stability). The resulting solid structure has a rhombohedral symmetry. This means the  $\text{SiO}_2$  groups form hexagonal structures by projecting on the perpendicular plane to the optical axis  $z$  at the top, where we find either Si atoms or  $\text{O}_2$  groups (see Figure 3.28).



**Figure 3.28.** Quartz structure

Si atoms can be seen as positively charged centers and O<sub>2</sub> as negatively charged centers. Both are bound to the crystalline mesh. If a force is applied following the y direction shown in Figure 3.29, the hexagon becomes distorted and the barycenters of charges + and - stop commingling. Electric dipoles appear that are all directed toward the direction of the applied force  $\vec{F}$ .



**Figure 3.29.** Appearance of dipole moment  $\vec{P}$  under the action of a directed  $\vec{F}$  force following the y-axis of quartz

Suppose we metallize the two opposite faces of a piezoelectric crystal, exposing these to an  $\vec{F}$  force. In this case, a dielectric bias appears that will follow the vertical direction shown in Figure 3.29. The effect of the elementary dipoles is to produce electrostatic charges + and - respectively on the metallic electrodes M<sub>1</sub> and M<sub>2</sub> and

the total charge of the ensemble remains zero. Thus, the electric induction vector  $\vec{D}$  is also zero (Gaussian law), and we get:

$$\vec{D} = \epsilon_0 \vec{E} + \vec{P} = 0 \text{ from which we get } \vec{P} = -\epsilon \cdot \vec{E} \tag{3.81}$$

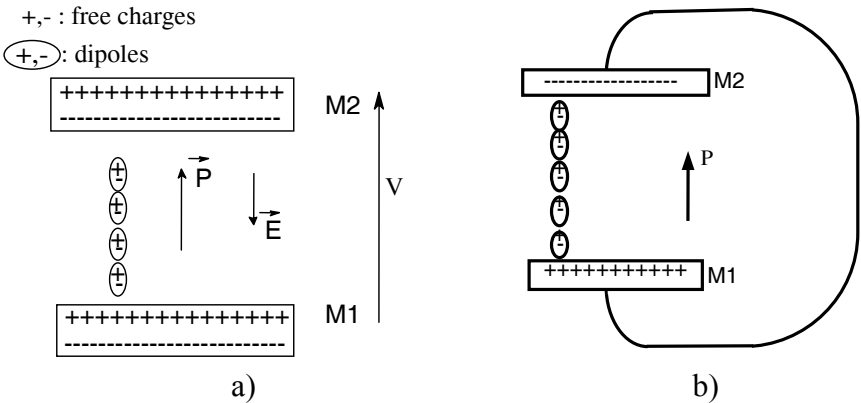
Finally, the potential  $V$  that appears at the limits of the piezoelectric crystal is deduced from:

$$\vec{E} = -grad \vec{V} = -\frac{\vec{P}}{\epsilon_0} \tag{3.82}$$

Let us now suppose that we have short-circuited the two electrodes by means of a metallic wire (see Figure 3.30b). The potential and the field between the electrodes become zero:

$$\vec{D} = \vec{P} \text{ and } div \vec{D} = \rho = div \vec{P} \tag{3.83}$$

which means that the flux of  $\vec{P}$  across the system shown in Figure 3.30b (the crystal, the electrodes and the metallic wire) is equal to the total sum of the contained charges. In the volume of the crystal  $\rho = 0$  and the specific charges that produce the flux of  $\vec{P}$  are the electrode charges  $M_1$  and  $M_2$  that we call the images of internal bias, that is, constraints.



**Figure 3.30.** Quartz with constraints in open circuit (a) and in short circuit (b)

3.2.2.3. Constraint tensors

Charges appearing on piezoelectric crystal faces depend on the direction of applied forces that break down into axial components and shearing. In Figure 3.31, the axial forces tend to stretch the cube in the direction of Oy and the shearing forces tend to make the planes xOy slide toward one another.

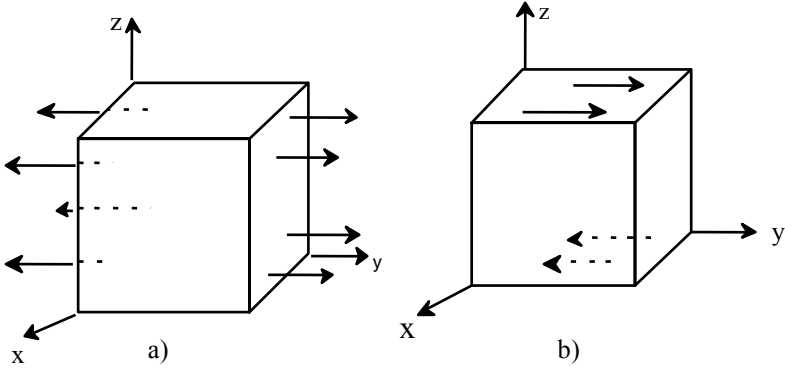


Figure 3.31. Axial constraints (a) and shearing (b)

In elasticity [GER 62] we note  $\sigma_{ij}$ , the constraint components in a solid, the index  $i$  giving the direction of the component, and the index  $j$  showing the normal of a facet to which this component is applied (see Figure 3.32).

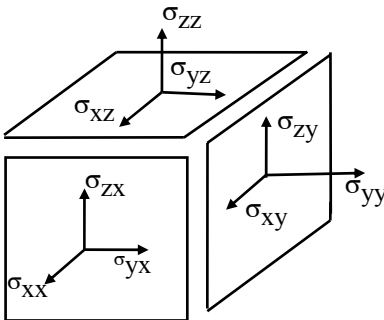


Figure 3.32. Definition of constraints

It has been shown [ROY 99] that constraint tensors ( $\sigma_{ij}$ ) are symmetrical ( $\sigma_{ij} = \sigma_{ji}$ ) and we note that  $\sigma_{ii} = \sigma_i$  and  $\sigma_{ij} = \sigma_k$ . In general, we describe the

piezoelectric effect through the linear relations which bind the constraints to charges by surface unity ( $q_i$ ) appearing on normal of facets in the direction  $i$  when the crystal is short-circuited:

$$\begin{aligned} q_1 &= d_{11}\sigma_1 + d_{12}\sigma_2 + d_{13}\sigma_3 + d_{14}\sigma_4 + d_{15}\sigma_5 + d_{16}\sigma_6 \\ q_2 &= d_{21}\sigma_1 + d_{22}\sigma_2 + d_{23}\sigma_3 + d_{24}\sigma_4 + d_{25}\sigma_5 + d_{26}\sigma_6 \\ q_3 &= d_{31}\sigma_1 + d_{32}\sigma_2 + d_{33}\sigma_3 + d_{34}\sigma_4 + d_{35}\sigma_5 + d_{36}\sigma_6 \end{aligned} \quad [3.84]$$

We write equations [3.84] in the form of  $q_i = (d_{ij})(\sigma_j)$ . Here  $(d_{ij})$  represents the piezoelectric tensor. Noting the degree of symmetry of the crystals, we can show that most of the  $d_{ij}$  are zero. For example, with quartz in the cross-section (with the  $z$  axis optical, the  $x$  axis mechanical, and the  $y$  axis electric; see Figure 3.28), the tensor is reduced to:

$$(d_{ij}) = \begin{pmatrix} d_{11}, -d_{11}, 0 & d_{14}, 0, 0 \\ 0, 0, 0, 0, -d_{14}, -2d_{11} \\ 0, 0, 0, 0, 0, 0 \end{pmatrix} \quad [3.85]$$

with  $d_{11} = 2.3 \cdot 10^{-12} \text{ CN}^{-1}$ ,  $d_{14} = 0.7 \cdot 10^{-12} \text{ CN}^{-1}$

For this Curie cut of crystal, we metallize the faces perpendicular to  $Ox$  (see Figure 3.33). If we apply a force of module  $F$  in the direction  $Ox$  so that  $\sigma_1 = \frac{F}{Ll}$  the charge by unity of surface  $q_1$  that appears on the electrodes equals:

$$q_1 = d_{11}\sigma_1 = d_{11} \frac{F}{lL} = \frac{Q_1}{Ll} \quad [3.86]$$

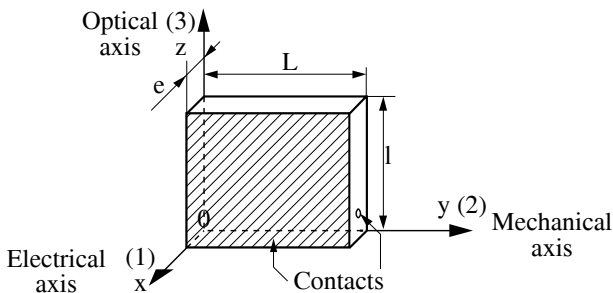


Figure 3.33. Quartz in Curie cut

If this same pressing force is applied following the Oy axis, the total charge  $Q_1$  increases in the relation  $\frac{L}{e}$  :

$$q'_1 = d_{12} \sigma_2 = -d_{11} \sigma_2 = -d_{11} \frac{F}{\ell e} = Q_1 \frac{L}{e} \quad [3.87]$$

Studying the piezoelectric matrix also makes clear that no charge can appear on the faces perpendicular to Oz. In addition, the sensor can be used for measuring shearing constraints of large surfaces (L, l), but not for hydrostatic measurements (pressure P). Indeed, such a constraint leads to:

$$q_1 = d_{11} \sigma_1 - d_{11} \sigma_2 = d_{11} P - d_{11} P = 0 \quad [3.88]$$

#### 3.2.2.4. Other piezoelectric materials

Aside from quartz, piezoelectric materials are usually ceramics made from piezoelectric polycrystals fritted in the presence of an electric field that directs the microscopic electric dipoles and finally produces a macroscopic bias. The only ceramic that causes significant bias is PZT. By calling Oz (3) the direction of the electric field, the matrix  $d_{ij}$  of this ceramic is written:

$$\begin{pmatrix} 0, & 0, & 0, & 0, & d_{15}, & 0 \\ 0, & 0, & 0, & d_{15}, & 0, & 0 \\ d_{31}, & d_{31}, & d_{33}, & 0, & 0, & 0 \end{pmatrix} \quad [3.89]$$

Ceramics cut perpendicularly to Oz are used in the manufacture of pressure sensors applied to faces xOy. The planes perpendicular to Ox or Oy are generally used for shearings. However, ceramics cannot be used for measuring hydrostatic pressure because the sum ( $2d_{31} + d_{33}$ ) is almost zero. After preparing PZT (composition, applied field, type of wiring), the coefficient  $d_{33}$  can be of the order of some  $10^2 \text{ pCN}^{-1}$ . Piezoelectric sensors can also be constructed with a polymer base. As with ceramics, the bias of these materials is obtained, during polymerization, through exposure to an electric field, this time by stretching them in heat, then cooling them. The material most often used today is PVDF whose piezoelectric tensor is written as:

$$\begin{pmatrix} 0, & 0, & 0, & 0, & d_{15}, & 0 \\ 0, & 0, & 0, & d_{24}, & 0, & 0 \\ d_{31}, & d_{32}, & d_{33}, & 0, & 0, & 0 \end{pmatrix} \quad [3.90]$$

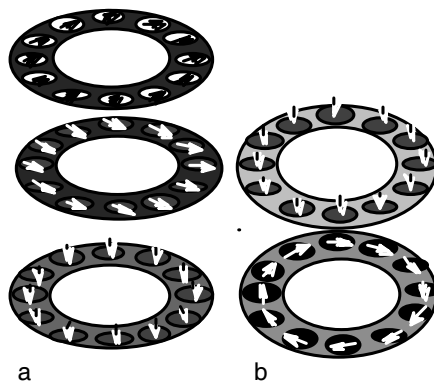
The values of  $d_{ij}$  are slightly above those of quartz and are sensitive to and dependent on preparation procedures.

### 3.2.2.5. Construction of piezoelectric sensors

These kinds of sensors are usually constructed in the form of “charge slices” (see Figure 3.34). The extent of the measurement varies from several kN to more than 100 kN. In order to detect traction, we constrain the charge slice or slices between two bolts. To increase sensitivity, assemblies are often used. For example, piezoelectric sensors are often in pile assemblies with electric connections so that either the tension or the piezoelectric charge is multiplied by the number of pile plates. Sometimes sensors are installed on a proof body adapted to the kind of constraint to be measured. Aside from compression sensors or simple shearings, which have principles very close to those we described above, elementary cell assemblies have different principles that allow for the construction of sensors that are sensitive to directional forces or structure (such as moment, pair and torque sensors).

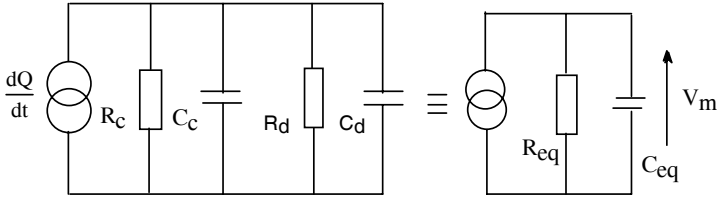
### 3.2.2.6. Using piezoelectric sensors

The equivalent electric schema of piezoelectric sensors can be deduced directly from its principle: a generator of variable charges in time that is represented by a current source  $dQ/dt$  in parallel with the capacity  $C_d$  of the dielectric between the two electrodes, and a flow resistance  $R_d$  that is also a characteristic of the dielectric. In practice, because of the values of  $C_d$  and  $R_d$ , we must always take into account the capacities and resistances  $R_c$  and  $C_c$  of the connecting cables, as shown in the schema in Figure 3.35.



**Figure 3.34.** Assemblies for measuring the three components of a force (a) or of the component following  $z$  and of the couple around  $z$  (b) (from G. Asch [ASC 91])



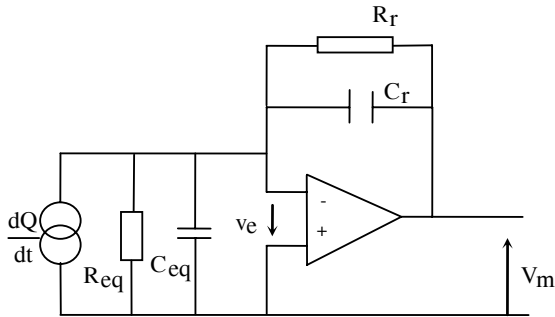


**Figure 3.35.** Equivalent schema of a piezoelectric sensor

If we measure the tension to the limits of the sensor or even the output of a tension amplifier (which is equivalent), we get:

$$V_m = \frac{Q}{C_r} \frac{R_r C_r j\omega}{1 + R_r C_r j\omega} \tag{3.91}$$

This is a first order low pass function with a cut-off frequency of  $\omega_c = 1/2\pi R_{eq} C_{eq}$  and the permanent value  $Q/C_{eq}$  depends on the impedance of the connecting cables and the input impedance of the tension amplifier. This situation is less promising than using a charge-tension convertor (see Figure 3.36) [BAU 61].



**Figure 3.36.** Piezoelectric sensor used with a charge-tension convertor

If we take the operational amplifier as an ideal with  $V_e$  zero, we carry out the short circuiting of electrodes and the transfer function depends only on the counter-reaction impedance:

$$V_m = \frac{Q}{C_r} \frac{R_r C_r j\omega}{1 + R_r C_r j\omega} = \frac{Q}{C_r} \cdot \frac{j\omega\tau}{1 + j\omega\tau} \tag{3.92}$$

The response time  $\tau = R_r C_r$  of this kind of assembly no longer depends on the sensor or on the converter connection. For the values of  $C_r$  of the order of several hundreds of pF, and of  $R_r$  of the order of several  $10^9 \Omega$ , we reach time constants of the order of several hundred ms and measurement tension of several mV/pC. For values of  $d_{ij}$  of the order of about 100 pC/N, we see that it is easy to reach sensitivity of the order of V/N.

### 3.3. Thermal sensors

Thermal measurands are the temperature and the variables related to accumulation or transfers of heat (specific heat, conductivity, thermal diffusion, heat flux, among others). In this book we have limited our discussion to certain sensors. With thermal sensors, the choice is simple, since in fact all these sensors are to some extent dependent on the temperature measurement. So, we are going to discuss temperature sensors in the following sections, beginning with the definition of temperature, modes of heat transfer [CAR 69] and, lastly, the principle of thermometry by contact. We will discuss the principle of thermoelectric sensors in some detail [GOL 60].

#### 3.3.1. Concepts related to temperature and thermometry

Our bodies can qualitatively evaluate the concept of hot and cold objects, but in this sense, though this concept is assimilated to the concept of touch in the normal five senses, it is both non-linear and residual: it depends on prior experience. Like the other human senses, hot and cold cannot be measured – we do not even know what it is we are trying to measure. The basic thermal sensor is, as we said above, a temperature sensor. The first question to be answered is how to define this odd measurand, which is not well understood in the physiological sense.

Temperature is a macroscopic concept which, even though it makes sense only in terms of a number of sufficiently large atoms, is dependent on a microscopic variable. This variable is the kinetic energy of each particle of the macroscopic system. The temperature of a system is an expression of the mean kinetic energy of all the particles contained in the system. The connection between the microscopic kinetic energy and temperature is part of the field of statistical mechanics, and we will not discuss it here, but will only note that this field proves that temperature sensors must be instruments capable of evaluating the mean kinetic energy of a system.

There are two ways a body can transfer its kinetic energy to another body. One is by contact that communicates the agitation of the first system to the second system; the second is electromagnetic radiation exchanged between the particles of the two

systems. The particles of the first body play the role of an electromagnetic source and those of the second body play the role of receiver. Temperature sensors are thus systems that can transform the kinetic energy of agitation communicated by contact or radiation into another form of energy, usually electrical.

We will only describe transfer by contact, since transfer by radiation really only occurs with optical sensors. Temperature sensors were called thermometers for the first time in 1624.

### 3.3.2. *Thermodynamic temperature*

Many physical properties of materials depend on thermal agitation of their elementary components (of their mean kinetic energy). It is always possible to take such a property, measure it and link it to the temperature (in terms of location and equality). However, because it would be dependent on the body used and the property measured, such a temperature scale would be completely arbitrary. Only thermodynamic temperature has a universal character.

Looking for this universality, Carnot stated that energy takes two forms: thermal agitation (heat) and organized energy (existence of privileged directions of speed) which is called work. This means there is a relation between the concept of temperature and that of heat conversion in work by means of a motor. Carnot introduced the idea of the ideal or reversible motor able to constantly operate the conversion of heat in work or the reverse (reversibility). He showed that the production  $\eta$  of such a motor, functioning between two heat sources of temperatures  $\theta_1$  and  $\theta_2$ , is independent of the technology used to construct it. This production depends only on  $\theta_1$  and  $\theta_2$ :

$$\eta = 1 - \frac{F(\theta_1)}{F(\theta_2)} \quad [3.93]$$

where  $F(\theta)$  is a function that depends on the temperature scale chosen. Thermodynamic temperature is defined by the choice of the scale as  $F(\theta) = T$ , so that:

$$\eta = 1 - \frac{T_1}{T_2} \quad [3.94]$$

To construct a reversible motor, we can follow a compression cycle and expansions that are, successively, isothermal and adiabatic to a mole of perfect gas. At this point we can see that the temperature appearing in the state equation of the mole of perfect gas,  $PV = RT$ , is Carnot's thermodynamic temperature.

If the value of  $R$  were known with a sufficiently weak uncertainty, temperature measurement would be reduced to a pressure measurement. This is not the case and we must eliminate  $R$  in carrying out a relative measurement. To do that, we chose a physical phenomenon that depends solely on the temperature, one that can be reproduced with a better uncertainty than we find with  $R$ . The triple point of water fulfills this condition. With  $T_0$  as the value of the temperature of this fixed point, we carry out two pressure measurements  $P_0$  and  $P$  by placing the thermometer containing perfect gas first in contact with the triple point of water, then with the body whose temperature  $T$  we want to measure. The thermodynamic temperature scale is fixed, just as its Kelvin unit (K), upon which we impose the value  $T_0$ . We take  $T_0 = 273.16$  K in order for  $R$  to have a value equal (to the closest uncertainty) to that produced by other measurement systems ( $R$  being the product of Boltzmann's constant and of Avogadro's number) in units of international measurement equal to 8.32 J/K.

The thermometer of this scale in principle must be a perfect gas thermometer. However, this kind of thermometer is too delicate for industrial use, and other devices have been developed which are capable of giving thermodynamic temperature measurements through the variations of other physical variables. The ITS 90 scale stipulates the physical methods of measurement and the formulae that help link diverse variables to temperature. These formulae are defined on the basis of reproducible temperatures (changing of state of pure bodies) measured once with a perfect gas thermometer. These fixed points, and interpolation formulae between the fixed points under exact conditions, constitute the temperature scale.

### ***3.3.3. Temperature scales currently in use and widely used measurements***

Two other temperature scales are in use today. Much of the English-speaking world uses the Fahrenheit scale for measuring thermodynamic temperature. This scale differs from the Celsius system used through most of the world.  $R$  values are different in the two scales. The  $R$  value of the Fahrenheit scale is related to a thermodynamic temperature scale called the Rankin scale.

The two non-thermodynamic temperature scales are:

– Celsius temperature that attributes the value  $0^\circ\text{C}$  to the freezing point of water saturated with water to the pressure of 101,325 Pa (273.15 K). The relation between the Celsius temperature and the thermodynamic temperature is expressed as:

$$\theta (^{\circ}\text{C}) = T(\text{K}) - 273.15 \quad [3.95]$$

– Fahrenheit temperature, used mostly in the USA and the UK, takes from the Rankin scale the value  $32^{\circ}\text{F}$  as the freezing point of water. The relation between  $^{\circ}\text{F}$  and  $^{\circ}\text{C}$  is expressed as:

$$\theta (^{\circ}\text{C}) = 5/9 \{ \theta(^{\circ}\text{F}) - 32 \} \quad [3.96]$$

Following the example of the perfect gas thermometer, thermal dilation is the physical variable most useful for making a thermometer. Thermal dilation is at the basis of many thermometers used today. In all the bodies, the increase of mean kinetic energy, that is, of temperature, is expressed by a modification of the mean distances separating the elementary particles (atoms or molecules). In solids, this modification is often different depending on the direction, and taking into account this anisotropy, we define the linear dilatation  $\alpha_l$  in the following way:

$$\alpha_e = \frac{1}{\ell} \frac{\partial \ell}{\partial T} \quad [3.97]$$

Strictly speaking,  $\alpha_l$  is weakly dependent on the temperature. However, in most industrial applications, small variations are disregarded. For isotropic materials, and therefore for fluids, we introduce the volumic dilatation coefficient  $\alpha_v$ :

$$\alpha_v = 3\alpha_l = \frac{1}{V} \frac{\partial V}{\partial T} \quad [3.98]$$

Many thermometers, among them perfect gas thermometers, make use of fluid dilatation for temperature measurements, but this does not mean the measurand is easily converted to an electric variable. Using the phenomenon of dilatation helps us understand that it is crucial that thermometers carry out an efficient transfer between the system we want to measure and the thermometer. We will discuss this further in the following section.

### 3.3.4. Heat transfers

#### 3.3.4.1. Conduction

The origin of heat (of thermal agitation) cannot really be known. We can say that the increased temperature of a system does not retain the memory of whatever produced the increase. This is due to the fact that it is impossible to contain heat within a system. The agitation inevitably spreads to the outside environment by contact or radiation. This process can be slowed to some extent by isolation procedures, but the “leakage” is, in fact, inevitable. These phenomena are explained

by heat transfers; and we must understand these transfers in order to establish the equilibrium temperature of a system, which is fundamental to procedures of taking and recording temperatures.

Figure 3.37 schematizes the group of thermal transfers that can occur between a temperature source  $T_0$  and the environment surrounding the temperature  $T_a$ . The source is maintained at a constant temperature  $T_1$ , partly to compensate for losses it undergoes, and partly to give it the highest possible calorific capacity  $C$ . Three successive plates of different materials have been attached to this source, and Figure 3.37 shows the temperature distribution from the source to the limit of the material stacking making contact with the exterior air.

The transfer that occurs in the plates  $P_1$ ,  $P_2$ , and  $P_3$  is transfer by conduction. It lowers the temperature of  $T_0$  to  $T_b$ . The last temperature decrease (from  $T_b$  to  $T_a$ ) is produced by two other possible types of transfers: convection, which is a specific form of conduction, and radiation. It is important to note that Figure 3.37 does not show the order of importance of different thermal transfers. Transfer by conduction is not systematically more efficient than other types of transfer.

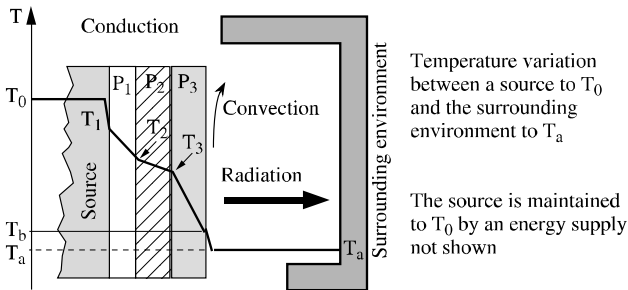


Figure 3.37. Three types of thermal transfers

Each of the gradient plates, which are made of different materials, show transfer by contact, that is, conduction. With the notations of Figure 3.37, we get:

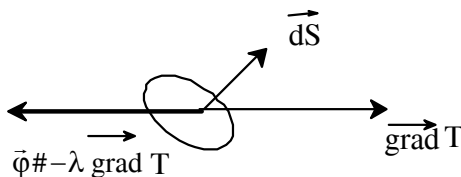


Figure 3.38. Transfer by conduction

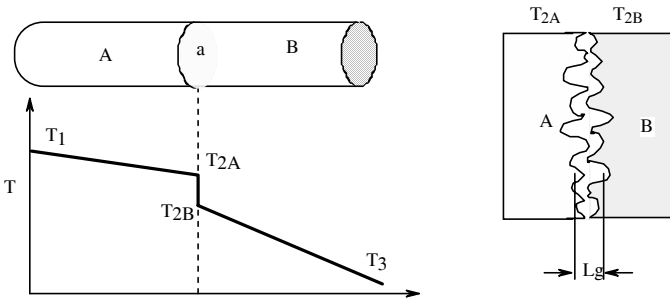
$$\frac{1}{S} \frac{dQ}{dt} = \lambda_1 \left( \frac{T_1 - T_2}{L_1} \right) = \lambda_2 \left( \frac{T_2 - T_3}{L_2} \right) = \lambda_3 \left( \frac{T_3 - T_6}{L_3} \right) \quad [3.99]$$

This equation shows that the quantity of heat leaving the source by time unit and the surface crosses without modifying the plates P<sub>1</sub>, P<sub>2</sub>, and P<sub>3</sub>. In other words, the heat flux has been conserved. The factor λ<sub>i</sub>, which defines the gradient  $\frac{\Delta T}{\Delta x}$  of the temperature distribution in each material, is the thermal conductivity (Wm<sup>-1</sup>K<sup>-1</sup>). More generally, thermal conductivity is defined through Fourier’s law. This vectorially expresses the heat flux transmitted by conduction:

$$\vec{\Phi}_{conductive} = -grad \vec{T} (\lambda T) \cong -\lambda grad \vec{T} \quad [3.100]$$

This expression only holds absolutely for the first equality because thermal conductivity is weakly dependent on temperature (almost always increasing).

In the schema shown in Figure 3.37, we supposed there was no temperature discontinuity at the level of contact surfaces between the different plates. In reality, this would be impossible. At the junction between solid materials, there is always a jump in temperature, as shown in Figure 3.39.



**Figure 3.39.** Contact resistance producing a temperature jump ( $T_{2A} - T_{2B}$ )

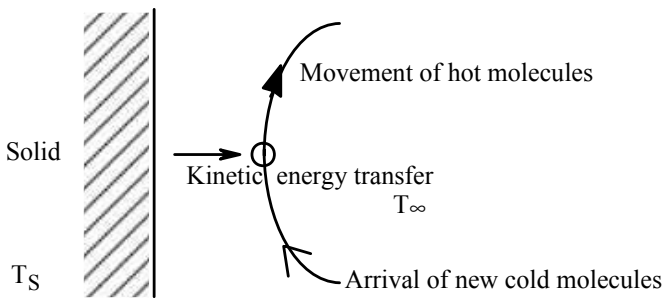
We characterize this temperature jump with a variable called contact resistance R<sub>a</sub>. Despite the presence of this resistance R<sub>a</sub>, which essentially depends on the quality of the surfaces in contact in the thickness zone L<sub>g</sub>, there is always continuity of heat flux φ because no energy can accumulate. We get:

$$|\vec{\phi}| = \frac{1}{S} \frac{dQ}{dt} = \frac{T_{2A} - T_{2B}}{R_a} \quad [3.101]$$

Taking into account the weak conductivity of air, the contact zone between the two surfaces is responsible for the value of the thermal resistance. This explains why the value  $R_a$  cannot really be calculated precisely. This value is deduced from the relation  $\frac{T_{2A} - T_{2B}}{|\vec{\phi}|}$ , which is measurable.

### 3.3.4.2. Convection

Convection takes as a starting point the fact of a mobile fluid taking part in a thermal transfer. In this transfer, the transmitted thermal flux increases considerably in relation to the flux produced by the conduction between a solid and a fluid. The fluid particles, their kinetic energy having been increased by contact (conduction) with the solid wall, displace and are then replaced by other molecules of weaker kinetic energy, capable of harnessing the heat of the wall. The movement of the fluid components permanently renews the fluid molecules in contact with the solid.



**Figure 3.40.** Principle of convective transfer

Convection appears of its own accord when different temperature zones coexist in a fluid. Actually, its volumic mass  $\rho = PM/RT$  decreases with the temperature, so that the hot fluid tends to rise and the cold fluid tends to drop with the effect of Archimedes' principle (or law of buoyancy). This type of convection is called natural convection. Of course, we can also force the movement of fluid by using a turbine. We then speak of a forced convection. The speed of the particle group becomes much higher and the flux exchanged by forced convection becomes higher by several orders of magnitude to that of natural convection. Calculating convection is difficult and often must be carried out by means of numerical calculation. In this text we will limit ourselves to phenomenological expressions of the transmitted flux by convection of a solid to a fluid. This occurs by means of a proportionality coefficient  $h_c$  that exists between the exchanged flux and the temperature



differences, between the surface  $S$  of the solid  $T_s$  and the fluid far from the surface  $T_\infty$  (in practice to several tens of thousands of  $\mu\text{m}$ ):

$$\frac{1}{S} \frac{dQ}{dt} = h_c(T_s - T_\infty) \tag{3.102}$$

In the air,  $h_c$  is clearly equal to  $5 \text{ W/m}^2\text{K}$  for natural convection and can reach some hundreds of  $\text{W/m}^2\text{K}$  for forced convection.

3.3.4.3. Radiation

We have seen in our discussion of optical sensors (see section 3.1) that the surface  $dS$  of a body carried to a temperature  $T$  produces a radiation. The energetic flux of the radiation transmitted in a solid angle  $d\Omega$  around a direction  $\vec{n}$  (see Figure 3.41) is expressed by:

$$d\phi = \int_0^\infty (\epsilon_\lambda L_\lambda^0(T) d\Omega dS \cos \theta) d\lambda \tag{3.103}$$

where  $L_\lambda^0(T)$  is the black body luminance at temperature  $T$  and  $\epsilon_\lambda$  is the transmissivity of the surface  $dS$  at the wavelength  $\lambda$ .

If the transmissivity  $\epsilon_\lambda$  does not depend on  $\lambda$  ( $\epsilon_\lambda = \epsilon$  is the gray body), the flux transmitted by radiation can be calculated in the half space above the surface  $dS$  by using the expression of the black body luminance  $L_\lambda^0(T)$  (see Figure 3.41):

$$d\phi = \epsilon \sigma T^4 dS \tag{3.104}$$

with  $\sigma \# 5.68 \cdot 10^{-8} \text{ Wm}^{-2} \text{ sr}^{-1} \text{ K}^{-4}$ .

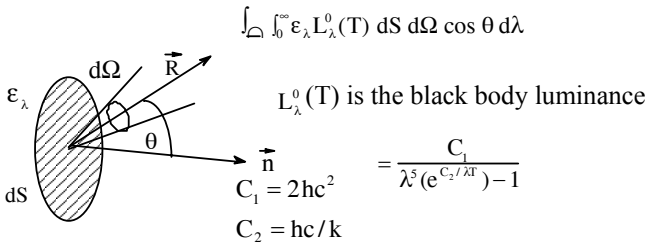


Figure 3.41. Transfer by radiation

We can see that the flux exchanged by radiation grows very quickly with the temperature. In section 3.1 we also saw that with two facing objects of transmissivity  $\varepsilon$  and  $\varepsilon_a$  that are at two temperatures  $T$  and  $T_a$ , the flux lost by the surface  $dS$  at the temperature  $T$  is:

$$d\phi = \varepsilon \varepsilon_a \sigma (T^4 - T_a^4) dS \quad [3.105]$$

If  $T$  is close to  $T_a$ ,  $d\phi$  can be developed to the first order:

$$d\phi \cong 4\varepsilon\varepsilon_a \sigma T^3 (T - T_a) dS \quad [3.106]$$

This is how we find a phenomenological law identical to that of convection with:

$$h_R = 4\varepsilon\varepsilon_a \sigma T^3 \quad [3.107]$$

For a body at ambient temperature, we find  $h_r$  of the order of five. So, in air without movement, at a temperature close to ambient, the sum of the convective and radiative transfers can be expressed as:

$$\phi = h(T - T_a) dS \text{ with } h = 10 \text{ Wm}^{-2} \text{ K}^{-1} \quad [3.108]$$

#### 3.3.4.4. Contact temperature measurement of solids

When a thermometer comes in contact with the solid body to be measured, several thermal transfers will assure the exchange of kinetic energy between the thermometer, the solid and the surrounding environment (which we assume to be fluid). In order to establish an equation that regulates, at any time, the relation between the temperatures of the thermometer, the body and the surrounding environment, we must make note of the instantaneous heat reading of the thermometer, the equality between the flux gained by unit of time and the amount of heat accumulated by unit of time. The factors which govern the equation are the thermal capacity of the thermometer  $C$ , its geometry and the contact resistance between the thermometer and the solid  $R$  and the phenomenological coefficient  $h$  of the equation (see equation [3.108]) expressing the transfer between the thermometer and the ambient fluid.

This differential equation is of the first order. The response time  $\tau$  of the thermometer is expressed by:

$$\tau = \frac{C}{hs + \frac{1}{R}} \quad [3.109]$$

with  $S$  being the surface of the thermometer in contact with the ambient fluid.

For the usual values of  $C$ ,  $S$ , and  $R$ , we get  $\tau$  of the order of the second. This order of magnitude shows that temperature sensors usually are very slow. To improve this response time, the most frequently used solution is reducing the caloric capacity of the thermometer by reducing to the minimum its geometric dimensions. However, it is important to keep in mind that response time is not intrinsic to a thermometer. It depends on how the thermometer is used.

Another consequence of the heat equation is that the equilibrium temperature of the thermometer  $T_\infty$  is not equal to the temperature of the solid  $T_s$ :

$$T_\infty - T_s = (T_a - T_s) \frac{hs}{hs + \frac{1}{R}} \quad [3.110]$$

The deviation between  $T_\infty$  and  $T_s$ , which is a systematic error of the thermometer, can be minimized by improving the thermal exchange between the thermometer and the solid and by reducing the exchange with the surrounding ambient fluid. This is done in several ways: by reducing the contact resistance (with thermal sticking or welding between the thermometer and the solid), by increasing the contact surface by burying the thermometer, and by reducing the surface exchange with the surrounding environment (reducing the length and diameter of the connection wires between the thermometer and the instrumentation).

### 3.3.5. Contact thermometers

Thermometry by contact is done by two types of thermometers. These are resistive thermometers that use the dependence of electric resistance on temperature and thermocouples that use the Seebeck effect.

#### 3.3.5.1. Resistive thermometers

Resistive thermometers are made of two materials: metals and semiconductors.

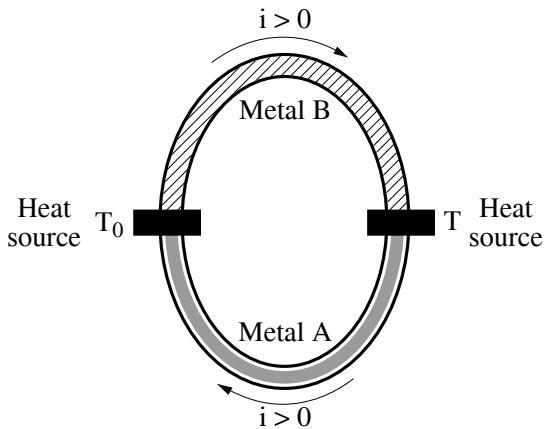
We saw in section 3.1 that electric conductivity is proportional to mobility and to the number of carriers by unit of volume.

With metal, electric conductivity decreases with the temperature because the number of current carriers in a metal does not, in practice, depend on the temperature. The only dependence on temperature comes from mobility, which decreases with the number of collisions per second and thus with the temperature. This decrease is approximately linear.

In the case of semiconductors, the temperature dependence is mostly controlled by the exponential growth of carriers with the temperature. Resistance decreases very rapidly with any drop in temperature. This variation is exponential, which explains why resistive thermometers are highly sensitive to semiconductors. However, their stability is clearly inferior to that of metallic thermometers. This factor, more than their linearity, explains why metallic thermometers are preferred as reference thermometers.

### 3.3.5.2. The Seebeck effect

Suppose we weld together the ends M and N of two different kinds of metallic wires (see Figure 3.42) maintained by an outside energy supply at different temperatures  $T_0$  and  $T$ . We can observe that in the closed circuit between the two sources of heat  $T_0$  and  $T$  circulate not only a heat flux but also an electric charge flux, which is another term for a current.



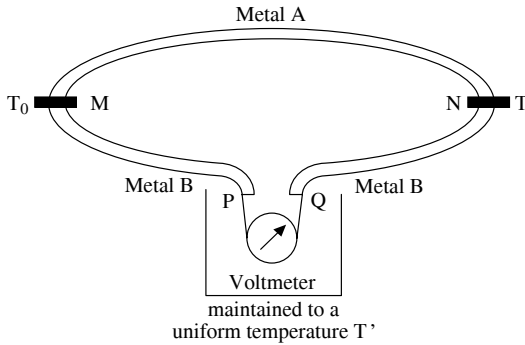
**Figure 3.42.** Schema showing principle of the Seebeck effect

We can express the creation of this electric current by saying that maintaining constant temperatures from two heat sources leads to the appearance of an electromotive force in the loop metal A + metal B. This A-B pair is called a thermocouple. We make  $E_{AB}(T, T_0)$  the algebraic value of this electromotive force in considering it to be positive because it makes the current circulate from A towards B in the junction M to  $T_0$ .

Of course, by reversing A and B or the temperatures T and  $T_0$ , we get:

$$E_{AB}(T, T_0) = -E_{BA}(T, T_0) = -E_{BA}(T_0, T) = E_{AB}(T_0, T) \quad [3.111]$$

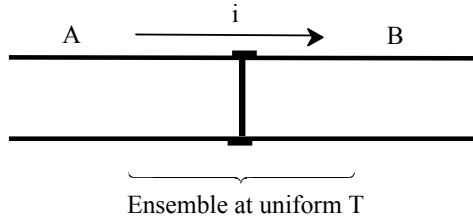
The electromotive force of the thermocouple can be proved experimentally by opening the circuit A-B and placing a voltmeter with all points at a single temperature between the two entry points P and Q (see Figure 3.43).



**Figure 3.43.** *Measuring the Seebeck effect*

If  $T = T_0$ , we observe that the deviation of the voltmeter is canceled and we can also verify the sign  $(V_p - V_q)$  by reversing T and  $T_0$  or the metals A and B described by equation [3.111].

The laws of thermodynamics (the first principle and the Onsager relation for irreversible processes) allow us to establish the fundamental laws governing thermocouples [MAC 62]. Thermodynamics does not give any explanation for the basic physics of the phenomenon (which can be found by studying, through solid state physics, the electron distribution in the different energies in the two metals [KIT 83]). However, the description physics gives us of the phenomenon is sufficient for studying temperature sensors. Thermodynamics shows us specifically that the Seebeck effect is the result of the Peltier and Thomson effects.

3.3.5.3. *The Peltier effect***Figure 3.44.** *The Peltier effect*

Suppose that two different metals A and B are welded and traversed by a current  $i$  (see Figure 3.44). When the current  $i$  traverses the welding in the direction  $A \rightarrow B$ , with the ensemble maintained at a temperature  $T$ , a certain power is freed in addition to the Joule effect. This power  $dQ/dt$  is called the Peltier effect. It is proportional to the current intensity  $i$  which traverses the welding and its sign depends on the direction of  $i$ , helping us to differentiate it from the Joule effect, which is always positive:

$$\frac{dQ}{dt} = \pi_{A-B}(T) i \quad [3.112]$$

The proportionality coefficient to  $i$ ,  $\pi_{A-B}(T)$ , depends on  $T$  and changes sign when we reverse  $i$ :

$$\pi_{A-B}(T) = -\pi_{B-A}(T) \quad [3.113]$$

where A-B means that the direction of the current is of A towards B.

3.3.5.4. *The Thomson effect*

Suppose a conductor is made of one metal with two extremities P and Q at different temperatures  $T + dT$  and  $T$  ( $dT > 0$ ). The conductor is traversed by a current  $i$  of P towards Q.

During a time interval  $dt$ , a certain quantity of heat is transmitted by the Joule effect. But we also observe an emission or a heat absorption  $dQ$  of a different physical nature. This quantity  $dQ$  is positive (heat emission towards the outside) when  $i$  circulates from P towards Q and is negative (the metal absorbs the heat)

when  $i$  circulates from  $Q$  towards  $P$ . Between two distant points of  $dx$  on the conductor with the different temperatures of  $dT$ , the exchange of  $d^2Q$  is given as:

$$d^2Q = \sigma_A(T) \overrightarrow{\text{grad}}(T) \cdot \vec{i} \cdot dx \cdot dt \quad \text{or again} \quad \frac{d^2Q}{dt} = \sigma_A(T) \cdot dT \cdot i \quad [3.114]$$

where  $dT$  and  $i$  are algebraic variables.

### 3.3.5.5. The Seebeck electromotive force

In applying the first principle of thermodynamics, we find that the variation of the electromotive force of the thermocouple ( $M - N$ ),  $dE_{AB}(T, T_0)$ , when the welding temperature  $N$  varies from  $dT$ , is equal to:

$$\frac{d E_{AB}(T, T_0)}{dT} = \frac{d\pi_{AB}}{dT} + (\sigma_A(T) - \sigma_B(T)) \quad [3.115]$$

The Onsager relation then helps us find a second relation:

$$\frac{d E_{AB}(T, T_0)}{dT} = \frac{\pi_{AB}(T)}{T} \quad [3.116]$$

after derivation in relation to  $T$ , the above equation is also written as:

$$\frac{d\pi_{AB}(T)}{dT} = \frac{d E_{AB}(T, T_0)}{dT} + \frac{d^2 E_{AB}(T, T_0)}{dT^2} \quad [3.117]$$

from which we deduce, by using the relation in [3.115]:

$$\frac{d^2 E_{AB}(T, T_0)}{dT^2} = \frac{\sigma_A - \sigma_B}{T} \quad [3.118]$$

These equations clearly show that  $E_{AB}$  is not a linear function of  $T$ . With the help of the relations just presented, we establish the Seebeck effect created in the circuit shown in Figure 3.43. For this, we maintain  $M$  to  $T_0$  and progressively raise  $N$  from  $T_0$  to  $T_1$  so that the integration of  $dE_{AB}(T, T_0)$  from  $T_0$  to  $T_1$  leads to:

$$E_{AB}(T_1, T_0) = \int_{T_0}^{T_1} d E_{AB}(T, T_0) = [\pi_{AB}(T_1) - \pi_{AB}(T_0)] + \int_{T_0}^{T_1} [(\sigma_A(T) - \sigma_B(T))] dT \quad [3.119]$$

This fundamental relation shows that we can deduce the electromotive forces of the two couples A-B and C-B from the couple A-C:

$$E_{AB}(T_1, T_0) - E_{CB}(T_1, T_0) = E_{AC}(T_1, T_0) \quad [3.120]$$

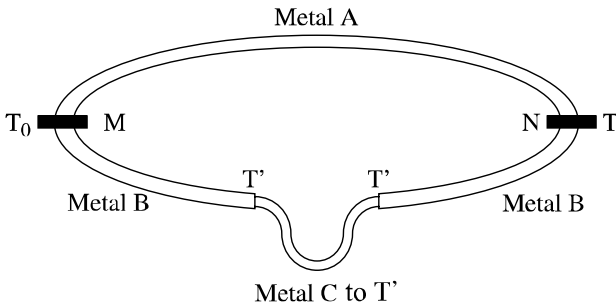
Equation [3.120] leads to establishing tables giving the electromotive force (EMF) of couples made by combining a metal B of reference (Pb or Pt) with different metals or alloys. From these tables we can deduce the EMF of all thermocouples using a variety of metals.

The measurement of a temperature  $T_1$  with a thermocouple A-B that has a welding reference at  $T_0$  can be deduced from the measurement with the same thermocouple having a welding reference at  $T'_0$  by:

$$E_{AB}(T_1, T_0) = E_{AB}(T_1, T'_0) + E_{AB}(T'_0, T_0) \quad [3.121]$$

This relation is used when  $T'_0$  is the ambient temperature and we want to deduce the EMF value in relation to  $T_0 = 0^\circ\text{C}$ , the latter being the reference temperature found in standardized tables. The electromotive force is often created by compensation housing that allows for thermocouple use without a reference source.

It is important to remember that all isothermal metals C introduced into the loop A-B do not modify the EMF of the thermocouple A-B. This allows us to construct couples by heterogeneous welding. This fact also helps explain measurement with the help of a voltmeter, as we have already described in section 3.3.5.2. Figure 3.43 can be schematized by Figure 3.45, in which we see that the branch PQ develops a zero EMF if the metal C is at a uniform temperature  $T'$ , so the only EMF is that of the thermocouple A-B. We can introduce as many metals as we wish between P and Q provided the temperature of the ensemble remains uniform.



**Figure 3.45.** Measurement principle of the EMF of a thermocouple



### 3.3.6. Features and uses of thermocouples

A normalized notation [ANS 82] of thermocouples (in capital letters such as E, J, K, N, T, S, etc.) helps us recognize their alloys. While value tables of electromotive forces of these couples [WIL 90] are cross-referenced to maintain standards, it is preferable to carry out calibrations for each thermocouple before its use, especially if the sensor is old and has already had several or more assemblies [ASH 81]. In fact, all chemical or physical variations of the components can cause variations of the EMF.

Even though many alloys can be used in making thermocouples, less than ten or so are currently used for this purpose. The couple platinum-platinum plated with 10% rhodium with sensitivity of the order of  $10 \mu\text{VK}^{-1}$  offers the lowest uncertainty ( $< 0.1^\circ\text{C}$ ) because of the attainable purity of its components and their chemical stability. It can be used from 300 to 1,800 K but more often is used above  $> 500$  K in which its stability and manageability often make it preferable to other thermometers, excepting platinum resistance thermometers, which are considered the reference.

Thermocouples made of metal alloys have very weak sensitivities. The most sensitive of these is the type E thermometer (chromel-constantan); this has a value of around  $80 \text{ mVK}^{-1}$ . We can replace the metal base couples with semiconductive junctions (for example tellurium, bismuth or germanium with different dopings). Although the transmitted electromotive forces are in this case clearly higher, the manageability of these semiconductive thermocouples is still today limited, in any case, for industrial usages. It is important to keep in mind that the calibration curve of a thermocouple, whatever its composition, is never linear – but only within a narrow temperature range. This means that sensitivity is strongly temperature-dependent.

The sources of disturbances in thermocouples are, at low frequencies, Johnson's noise in resistance wires and at high frequencies, electrostatic and magnetic couplings.

The manufacture and usage of thermocouples depends on two factors:

- the welding must be as small as possible in order to insure a weak response time;
- the assembly must be both mechanically solid and protective against disturbance.

The welding and connecting wires are usually placed within a protective metallic sheath. The welding is sometimes in electrical contact with this sheath. The wires are always isolated by a silica powder or a compacted alumina. When the structure to be tested is metallic and can be grounded with the instrumentation, the assemblies

with which the welding is connected to the sheath are preferred for suppressing usual disturbances.

In principle, two thermocouples are necessary for measurement (see Figure 3.45). One of them is placed in contact with the structure whose temperature  $T$  is to be measured; the other is placed in a protective shield whose temperature  $T_0$  is known and fixed. It is best to have  $T_0$  equal to  $0^\circ\text{C}$ , the reference temperature of tables.

However, this kind of assembly is seldom used in industrial settings. In such situations, the welding reference is replaced by an electromotive force that, for an ambient temperature  $T_a$ , constantly gives the value  $E_{AB}(T_a, T_0)$ . This force is placed in series with the welding A-B fixed on the structure. The total EMF of the ensemble is expressed by:

$$E_{AB}(T, T_0) = E_{AB}(T, T_a) + E_{AB}(T_a, T_0) \quad [3.122]$$

this is the value we would have by using a second couple at the reference temperature  $T_0$ . The EMF  $E_{AB}(T_a, T_0)$  (compensation casing) can be produced by the disequilibrium tension of a Wheatstone bridge containing a thermistor sensitive to  $T_a$ . Regulating different impedances and tensions of the bridge gives us  $E_{AB}(T_a, T_0)$  with a weak uncertainty if the ambient temperature does not vary more than  $50^\circ\text{C}$  during measurement. When the compensation casing cannot be directly connected to the thermocouple, intermediary cables, called compensation cables, must be used. This avoids systematic errors that can occur by creating parasite EMFs at A-B junction connections if the connection was made without proper precautions. Obviously, these compensation cables are dependent which A-B couples have been used. Their sheath is usually the standard color of the thermocouples currently in use.

### 3.4. Bibliography

- [AFN] AFNOR, norm NF X07001.
- [ANS 82] Standard, Temperature measurement, ISA/ANSI standard MC 96, International Society for measurement and control, 1982.
- [ASC 91] ASCH G. *et al.*, *Capteurs en instrumentation industrielle*, Dunod, 1991.
- [ASM 81] Manual on the use of thermocouples in temperature measurement, ASMT Publication 04-470020-40, 1981.
- [BAU 61] BAURAND J., *Mesures électriques*, Masson, 1961.
- [BON 84] BONNET J.J., *Physique générale*, Dunod Université, Paris, 1993.

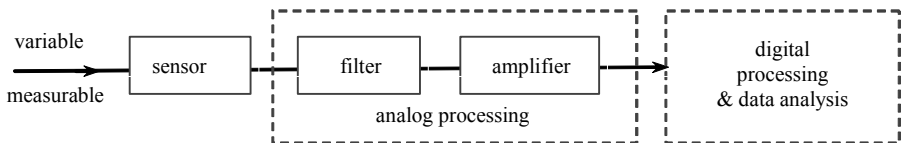
- [CAD 64] CADY W.G., *Piezoelectricity*, McGraw-Hill, 1964.
- [CAR 59] CARLSLAW and JAEGER, *Conduction of heat in solids*, Clarendon Press, Oxford, 1959.
- [GER 62] GERMAIN P., *Mécanique des milieux continus*, Masson, 1962.
- [GOL 60] GOLDSMID H.J., *Applications of thermoelectricity*, Wiley, 1960.
- [HOD 71] HODGSON J.N., *Optical absorption and dispersion in solids*, Chapman Hall, 1971.
- [KIT 83] KITTEL C., *Physique de l'état solide*, Dunod, 1983.
- [MAC 62] MACDONALD D.K.C., *Thermoelectricity: an introduction to the principles*, J. Wiley, NY, 1962.
- [NOR 89] NORTON H.N., *Handbook of transducers*, Prentice Hall, 1989.
- [PHI 73] PHILLIPS J.C., *Bonds and bands in semiconductors*, Academic Press, 1973.

## Chapter 4

# Analog Processing Associated with Sensors

### 4.1. Introduction

Correct measurements are crucial to the field of instrumentation. No matter what sensor is being used, many influence parameters or disturbances such as temperature, pressure, mechanical constraints and electromagnetic environment can contribute to measurement error. These kinds of problem are intrinsic to sensors. Furthermore, the acquisition chain must link an electronic device that can condition information and send it to a transducer. This information relates to the variable to be measured and provides the closest possible representation of the observed physical phenomenon.



**Figure 4.1.** *Simplified functional schema of a measurement chain*

Most analog processing of a signal sensor contains filtering and amplification functions (Figure 4.1). These functions help us retrieve relevant information from signal sensors and take it to a compatible and sufficient electric level so that the information can be then used by the system or equipment. This process assures

proper interface between user and displays, measurement devices and microcomputers.

Because measurement systems depend on reducing electronic noise in order to function correctly, in this chapter we will discuss in detail the aspects related to amplification functions, as well as the different sources of intrinsic noise in electronic devices. The concept of filtration will be dealt with in Chapter 5.

## 4.2. The problem of electronic noise

### 4.2.1. The origin of electronic noise

Electronic devices are subject to exterior noise sources (the field of electromagnetic compatibility studies these sources) and to internal noise sources caused by voltage variations and by circuit currents themselves.

There are two sources of exterior noise:

- one source of noise comes from electric disturbances transmitted by conduction. These can include: the influence of network distribution of electric energy of 230 V to 50 Hz; supply undulations (for example, alternating phase recovery at 100 Hz); power signals functioning at commutation frequencies from about 100 Hz to 100 kHz (breaks in energy supply);

- another source of noise can be radiated electric disturbances, including radiofrequency transmitters, electromagnetic fields created by high variations of voltage or current sometimes emitted by electric machines, such as motors or transformers or, most frequently, by static converters such as clippers or inverters.

Noise sources that come from inside components have many origins and fall into five categories. We discuss these below.

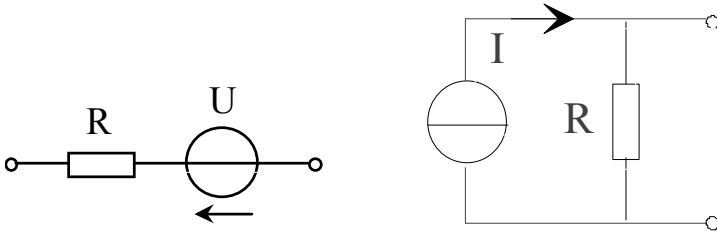
#### *Thermal or Johnson's noise*

This kind of noise corresponds to the electrons in resistive components. Its nature is random and does not depend on the value of the current traversing the resistive element. This is because the displacement speeds of the charges linked to thermal phenomena are much higher than the speeds of the group creating the current in the conductor.

Thermal noise is expressed by an effective noise voltage, given by the relation:

$$\overline{U^2} = 4.k_B.T.R.B_b \quad [4.1]$$

Here,  $k_B$  is Boltzmann's constant ( $k_B = 1.38 \cdot 10^{-23} \text{J/K}$ ),  $T$  is the absolute temperature in Kelvin (K),  $R$  is the ohmic resistance of the element, and  $B_b$  is the frequency range of usage (the bandwidth). The equivalent schema becomes that of a voltage generator (or Thevenin generator) as shown in Figure 4.2.



**Figure 4.2.** Equivalent sources of a resistance noise

It is possible to use another, equivalent schema for a noise source using a current or Norton generator. In this case, it is enough to divide the expression given above by  $R^2$  to get a clear expression of noise current.

$$I^2 = 4k \left( \frac{T}{R} \right) \Delta f \quad [4.2]$$

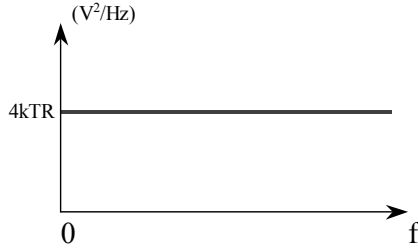
The spectral density of the voltage of thermal noise is written as:

$$\Phi_U^+(f) = 4k_B \cdot T \cdot R \quad [4.3]$$

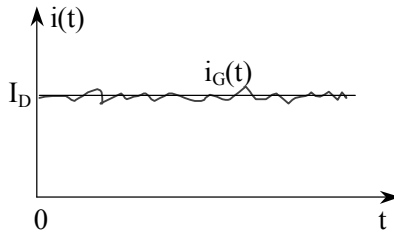
For a resistor of 1 k $\Omega$ , for example, at ambient temperature ( $T = 300 \text{ K}$ ), this spectral density of noise is close to  $16 \cdot 10^{-13} \text{ V}^2/\text{Hz}$ , or  $4 \text{ nV}/\text{Hz}^{1/2}$ . Consequently, the spectral density of thermal noise is constant (see Figure 4.3). That is why this kind of noise is also called white noise. Thermal noise is present in all the resistive linear elements, including microphones, loudspeakers, and antennae. With antennae, noise is the product of the thermal agitation of air molecules.

#### *Shottky's or shot noise*

Shot noise is present in all semiconductors (diodes and transistors), and is due to random instances of charge carriers crossing P-N junctions. The direct external current, though it seems constant, in fact fluctuates randomly around its mean value  $I_D$  (see Figure 4.4).



**Figure 4.3.** Spectral density of resistance noise



**Figure 4.4.** Fluctuation of the direct current  $i(t)$  of a P-N junction

An electric charge ( $q$ ) crossing a potential barrier is a Poisson process, so the current  $i(t)$  is written as:

$$i(t) = \sum_j q \cdot \delta(t - t_j) \tag{4.4}$$

The sum  $j$  corresponds to the crossings that occur per second. The probability density of the time interval  $T_1$ , separated by two successive crossings, is expressed by (this is a Poisson process):

$$p_{T_1} = \lambda \exp(-\lambda t) \tag{4.5}$$

where  $\lambda$  represents the average number of carriers crossing the barrier by unit of time. We then see that the intercorrelation function  $R_i(\tau)$  of  $i(t)$  can be given as:

$$R_i(\tau) = (q \cdot \lambda)^2 + q^2 \cdot \lambda \cdot \delta(\tau) = I_D^2 + q \cdot I_D \cdot \delta(\tau) \tag{4.6}$$

Here  $I_D$  represents the average value of the current  $i(t)$ . The spectral density of the power of the current  $i(t)$  is calculated with the help of the Wiener-Khintchine theorem applied to equation [4.6]. With a direct current we get:

$$\Phi_i(f) = \int_{-\infty}^{+\infty} R_i(\tau) \cdot \exp(-j2\pi f\tau) d\tau = I_D^2 + qI_D \quad [4.7]$$

The second term of this expression corresponds to the spectral density of the shot noise ( $\Phi_{iG}$ ). This means that shot noise, like thermal noise, is white noise. Because of this, shot noise cannot be differentiated from thermal noise as soon as they both become present in the same electrical circuit. On the other hand, shot noise only exists when a current  $I_D$  crosses the potential barrier.

In conclusion, an effective shot noise current ( $I_G$ ) depends on the direct current ( $I_D$ ) and the range of working frequency ( $B_b$ ) according to the relation:

$$I_G^2 = 2B_b \Phi_{iG}(f) = 2qI_D B_b \quad [4.8]$$

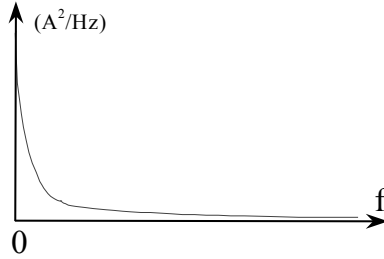
#### *Flicker or 1/f noise*

This noise has various origins. In bipolar transistors it is due to defects on the semiconductor surface in the depletion zone of the base-emitter, the carriers having been randomly trapped in the crystalline lattice (also called mesh). The energy of the resulting noise is mostly concentrated at low frequencies (between 0.1 Hz and 10 Hz) and the spectral amplitude density has, on a logarithmic scale, a linear decrease according to the frequency ( $f$ ). Flicker noise is also a function of the direct current  $I_D$  and can be written approximately as:

$$I_F^2 = K_1 \cdot \frac{I_D^\alpha}{f} \cdot B_b \quad [4.9]$$

where  $K_1$  is a constant belonging to the component and  $\alpha$  is a number included between 0.5 and 2. This type of noise is mostly caused by active components but also by carbon resistances. Flicker noise only exists in carbon resistances if the latter are traversed by a current  $I_D$ , with thermal noise always being present. That is why metallic resistances in low noise assemblies (replacing carbon resistances) never present flicker noise.





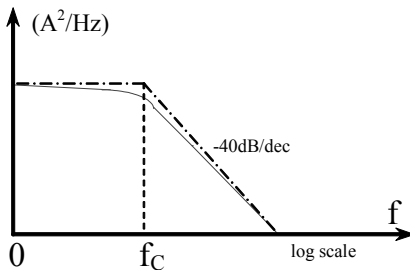
**Figure 4.5.** Spectral density of flicker noise

*Burst or popcorn noise*

This is another kind of low-frequency noise (lower than several kHz). It is found in integrated circuits and discrete transistors. The nature of popcorn noise is not completely known, but we can say that it is linked to the presence of contaminating heavy metal ions in the circuits. For example, components doped with gold present an especially high level of popcorn noise. The spectral density of noise can be expressed as:

$$\frac{I_P^2}{B_b} = K_2 \cdot \frac{I_D^\beta}{1 + \left(\frac{f}{f_c}\right)^2} \tag{4.10}$$

where  $K_2$  is a constant belonging to the experimentally established component,  $\beta$  is a number between 0.5 and 2, and  $f_c$  is a frequency of the given noise process. The spectral density of the jump has a speed similar to that of a low pass filter with a decrease at higher frequencies in  $1/f^2$ .



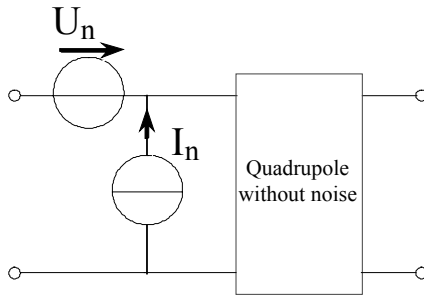
**Figure 4.6.** Spectral density of popcorn noise

### *Avalanche or Zener noise*

Avalanche noise is produced by creating an avalanche in the P-N junctions of Zener diodes. In the depletion zone, holes and electrons acquire enough energy to create electron hole couples by colliding with silicon atoms. This phenomenon occurs by means of random series of noise peaks. The resulting overall noise level is higher than that of shot noise for the same current amplitude. The voltage of useful noise, together with the Zener voltage, is strongly dependent on component structure and the homogeneity of the silicon crystal. In practice, we measure spectral densities of noise of the order of  $10^{-14}$  V<sup>2</sup>/Hz for a Zener current  $I_z = 0.5$  mA, which is the equivalent of a resistance of 600 k $\Omega$  to the ambient temperature. This means that it is not advisable to use Zener diodes in assemblies requiring low noise levels.

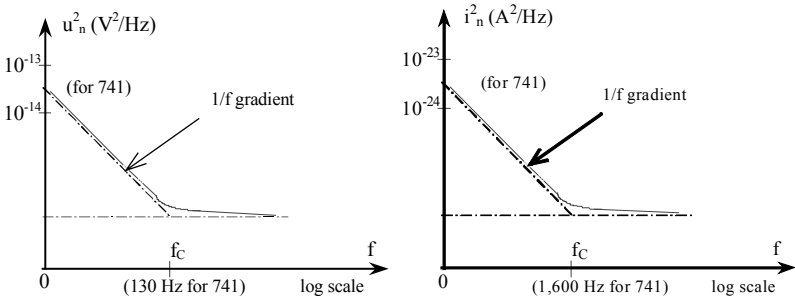
### **4.2.2. Noise in an electronic chain**

An electronic chain is made of many discrete components and integrated circuits that themselves have a large number of transistors and resistances integrated into the same substratum. This means, in practice, that studying noise with the help of an equivalent schema representing the noise of the ensemble is a complex, often irresolvable task. To simplify the problem, we represent the noise of each component, independently of the true schema, by considering the noise as a quadropole (without noise) having a source of voltage noise at input ( $U_n$ ) and a source of noise current ( $I_n$ ), as shown in Figure 4.7.



**Figure 4.7.** Noise generators at quadropole input

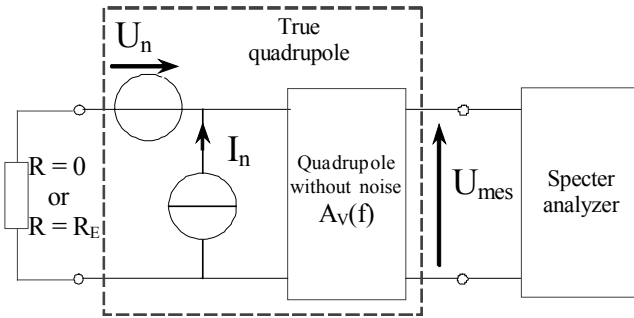
The two sources of noise are established experimentally by measuring their respective spectral densities. For integrated circuits or operational amplifiers, these spectral noise densities in voltage and in current have speeds shown in Figure 4.8.



**Figure 4.8.** Speed of spectral noise densities in an operational amplifier

These noise specters have both a white noise component and a noise component in  $1/f$ .

Measuring the noise sources of a quadropole is done with a spectrum analyzer. This device analyzes the output signal for several input configurations (see Figure 4.9):



**Figure 4.9.** Measurement of spectral noise densities of quadropole noise

– When the input is short-circuited ( $R = 0$ ), the spectral density of the noise voltage equivalent to the input is simply expressed as:

$$u_i^2 = u_n^2 \tag{4.11}$$

and the spectral density measured by the spectrum analyzer is:

$$u_{mes}^2 = |A_v(f)|^2 \cdot u_n^2 \quad [4.12]$$

Knowing  $A_v(f)$ , we can easily deduce  $u_n^2$ .

– When the input is charged by a source resistance  $R = R_E$ , the spectral density of the equivalent noise voltage at input becomes:

$$u_i^2 = u_n^2 + u_E^2 + R_E^2 \cdot i_n^2 + 2C \cdot R_E \cdot u_n \cdot i_n \quad [4.13]$$

where the term  $u_E$  corresponds to the thermal noise of the source resistance, we get:

$$u_E^2 = 4k \cdot T \cdot R_E \quad [4.14]$$

where  $C$  represents the correlation coefficient (a number between -1 and +1) of the noise sources  $u_n$  and  $I_n$ .

By taking a very high value of  $R_E$ , the third term of the equation is preponderant even if the measured spectral density is as follows:

$$u_{mes}^2 = |A_v(f)|^2 \cdot R_E^2 \cdot i_n^2 \quad [4.15]$$

This allows us to establish the source of the noise current  $I_n$ . An intermediary resistance value for  $R_E$  helps us find the correlation coefficient  $C$ .

### 4.2.3. Signal-to-noise ratio

Let  $P_s$  be the power of the useful signal at the output of an electronic chain and  $B_s$  is the power of the corresponding noise. We then define the signal-to-noise ratio  $(S/B)_s$  at the output of the chain with the relation:

$$\left( \frac{S}{B} \right)_s = 10 \cdot \log \frac{P_S}{B_S} \quad [4.16]$$

In the same way, the signal-to-noise ratio at the input of the chain can be written as:

$$\left(\frac{S}{B}\right)_E = 10 \cdot \log \frac{P_E}{B_E} \tag{4.17}$$

These two relations help us establish the quality (with regard to noise) of an electronic chain by defining the noise factor as:

$$F_{dB} = \frac{(S/B)_E}{(S/B)_S} = 10 \cdot \log \left( \frac{P_E \cdot B_S}{P_S \cdot B_E} \right) = 10 \cdot \log \left( \frac{B_S}{G \cdot B_E} \right) \tag{4.18}$$

where G is the power amplification of the electronic chain. If this amplification is ideal, it carries no supplementary noise to output, only amplifying noise at input. We get:

$$B_S = G \cdot B_E \tag{4.19}$$

In this case, the noise factor is equal to the unity (0 dB). Another definition of the noise factor is:

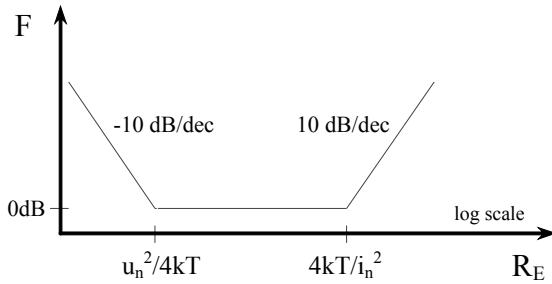
$$F = \frac{\text{total power of output noise}}{\text{power of output noise due to input generator}} \tag{4.20}$$

By using the spectral densities of equivalent noise at input of a quadrupole (equation [4.13]), the noise factor is written:

$$F = \frac{u_i^2 \cdot |A_v^2(f)|}{u_E^2 \cdot |A_v^2(f)|} = \frac{u_n^2 + u_E^2 + R_E^2 \cdot i_n^2 + 2C \cdot R_E \cdot u_n \cdot i_n}{u_E^2} \tag{4.21}$$

$$F = 1 + \frac{u_n^2 + R_E^2 \cdot i_n^2 + 2C \cdot R_E \cdot u_n \cdot i_n}{u_E^2} \tag{4.22}$$

When the noise sources  $u_n$  and  $i_n$  are weakly correlated, the speed of F according to the source resistance is shown in Figure 4.10.



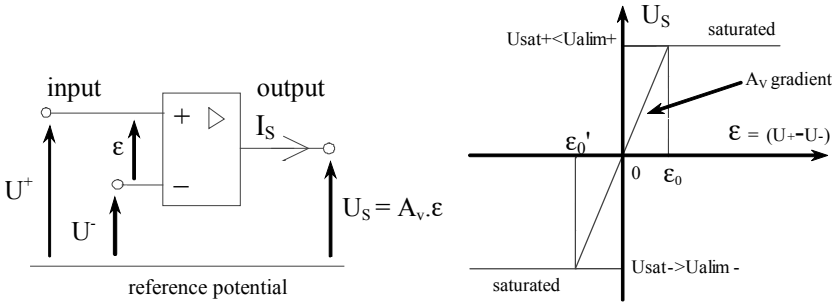
**Figure 4.10.** Evolution of noise factor  $F$  according to source resistance  $R_E$

The above expression shows that the noise factor of a quadrupole is always higher than the unit. Consequently, signal analysis always involves some degradation of the signal-to-noise ratio (see equation [4.18]). With reception systems, or with weak signal amplification coming from sensors, the additive noise at the first stage of reception or at preamplification plays an essential role in conditioning the signals to be analyzed.

## 4.3. Amplifiers

### 4.3.1. Operational amplifier

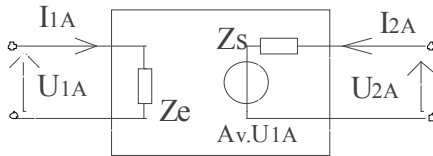
An operational amplifier or integrated differential amplifier is an integrated circuit which, in ideal conditions, provides an output voltage proportional to the *difference* of input tensions  $U_+$  and  $U_-$  of the observed limits “+” (not inverted) and “-” (inverted) of the component. The positive coefficient  $A$  is the voltage differential amplification of the amplifier. In practice, this is very high, of the order of  $10^4$  to  $10^6$ . The power supply of these circuits goes from several volts to several dozen volts maximum. The circuits cannot function by themselves (during open loop), since they need differential tensions lower than a few  $\mu\text{V}$  in order not to “saturate” (see the transfer feature in Figure 4.13).



**Figure 4.11.** *Schema for a differential amplifier and its features*

This is why, from the perspective of linear functioning, the operational amplifier is always used with a retroaction feature. This kind of amplifier can also be used in a non-linear functioning mode, with or without a positive reaction. In this instance, the feature is used for a broader input voltage dynamic. For these applications (comparator, trigger, astable, etc.), the output can only take two values:  $Usat^+$  or  $Usat^-$ .

In order to study all possible feedback types, the operational amplifier can be assimilated to a quadrupole of simplified design. This is shown in Figure 4.12.

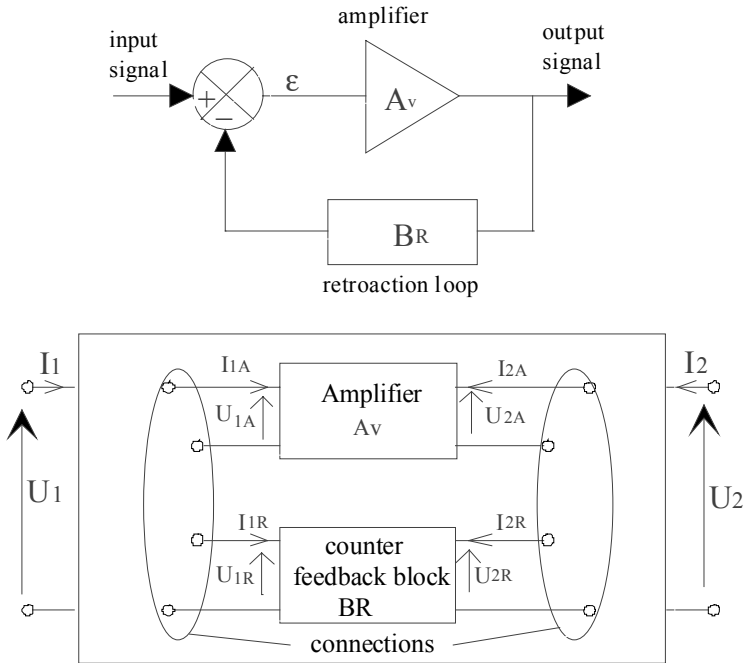


**Figure 4.12.** *Quadrupole design of a differential amplifier*

$Z_e$  represents the input impedance of the operational amplifier, with  $Z_s$  its output impedance, and  $A_v$  the amplification of the differential voltage.

4.3.1.1. *Feedback and counter-feedback in currents and tensions*

We can study an operational (or differential) amplifier that functions by means of amplification in linear regime, by viewing the amplifier feedback (or counter-feedback block) as the linking of two quadrupoles (Figure 4.13). One is part of the direct chain ( $A_v$ ) and the other part of the return chain or retroaction loop ( $B_R$ ).



**Figure 4.13.** *Schema of an amplifier with retroaction block*

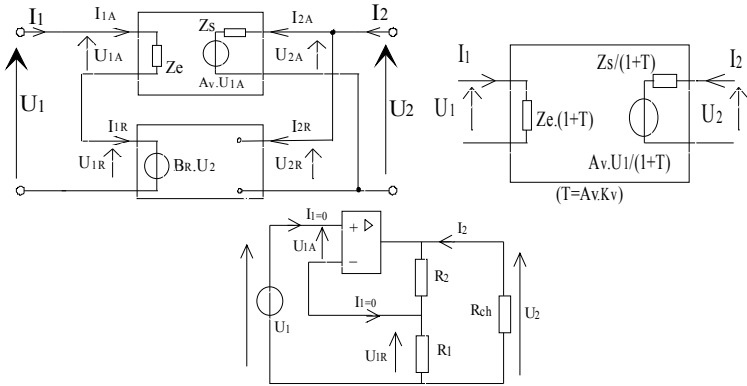
Following the serial or parallel linkage of input dipoles of the two quadrupoles, we can distinguish four possible configurations:

- serial input dipoles and parallel output dipoles;
- serial input dipoles and serial output dipoles;
- parallel input dipoles and parallel output dipoles;
- parallel input dipoles and serial output dipoles.

These four possibilities correspond to four types of feedback, either of voltage or of current. They can be created with an operational amplifier. The following examples illustrate, respectively, these four configurations.

- The counter-feedback of applied voltage applied in voltage (serial-parallel).

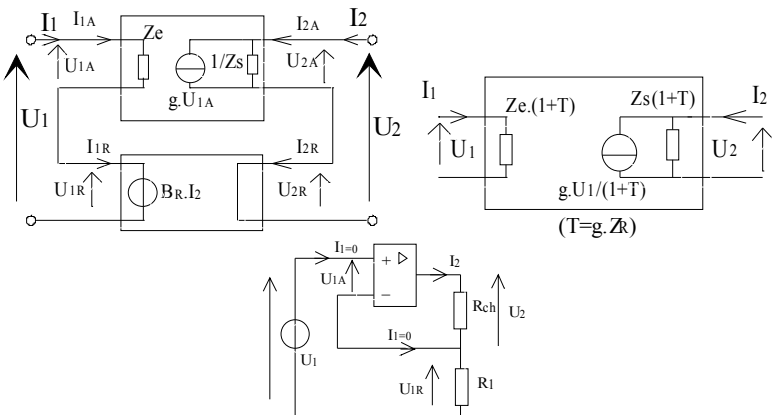




**Figure 4.14.** Counter-feedback assembly of applied voltage in voltage

In this assembly (known as a non-inverted amplifier), the output voltage  $U_2$  has servo-control to the input voltage  $U_1$  through the intermediary of  $R_1$  and  $R_2$ . The transfer function of the return block is a constant without dimensions  $B_R = K_v = R_1/(R_1 + R_2)$ . Since this is a retroactive effect of re-injection of voltage, input impedance is multiplied by  $(1 + T)$ , where  $T = A_v K_v$ . The output impedance is approximately divided by this term, on the condition that the generator impedance is, ideally, negligible in view of  $Z_e$ . If the amplification  $A_v$  of the direct chain is high enough, which is in practice true for operational amplifiers, the input impedance can be considered as infinity ( $I_1 = 0$ ), the output impedance as zero, and the voltage amplification as equal to  $1/K_v = 1 + (R_2/R_1)$ .

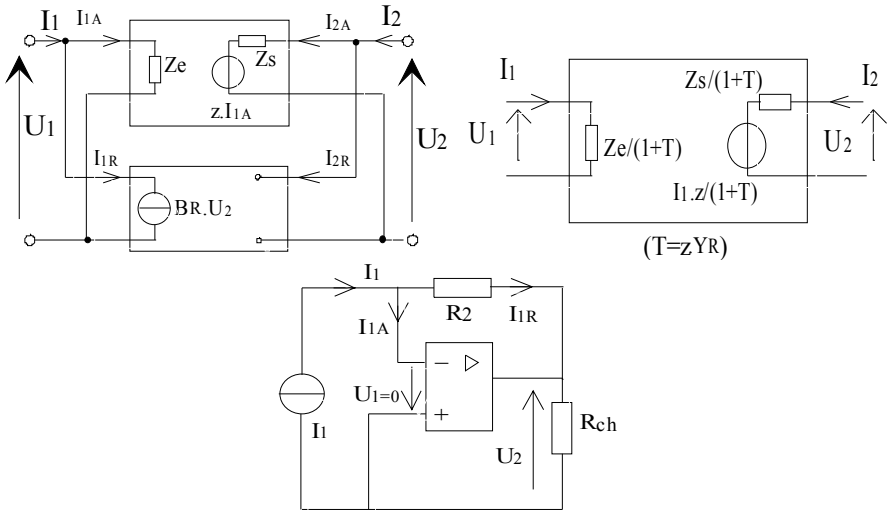
– The counter-feedback of a current applied in voltage (series-series).



**Figure 4.15.** Counter-feedback assembly for a current applied as voltage

As with the output current  $I_2 = U_{1R}/R_1 \approx U_1/R_1$ , this has servo-control of the input voltage  $U_1$ : this is how voltage current convertors or transconductance amplifiers are made. The transfer function of the return block is consistent with an impedance  $B_R = Z_R = U_{1R}/I_2$ , (equal to  $R_1$  if we disregard  $I_1$ ). This assembly is similar to the one shown before it, except that the charge resistance fluctuates because it replaces  $R_2$ . The input and output impedances are very high; they are both multiplied by  $(1 + T)$ . The term  $T$  corresponds to the transfer function in open loop of the assembly. It is written  $T = g \cdot Z_R$ . Here,  $g$  represents the assembly transconductance. If  $g$  is high enough, ideally infinite (or if  $I_1$  can be disregarded), the input impedances are infinite. This means the voltage-current convertor is perfect and the transconductance equals  $1/Z_R = 1/R_1$ .

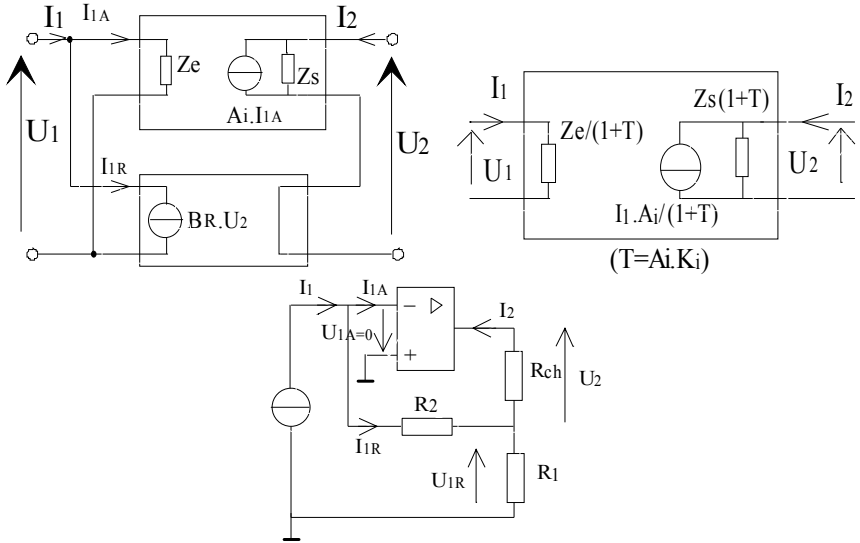
– The counter-feedback of voltage applied as current (parallel-parallel).



**Figure 4.16.** Counter-feedback assembly of voltage applied as current

This assembly is also called a current-voltage convertor or transresistance amplifier. The transfer function of the return block is an admittance  $B_R = Y_R = -1/R_2$  (it is negative for this assembly.) If we suppose that the operational amplifier imposes a nearly zero voltage ( $U_1 = 0$ ), or is negligible in relation to  $U_2$ , we obtain  $U_2/I_1 \approx U_2/I_{1R} = -R_2$ . This means the output voltage has servo-control of the input current. In such cases, the input impedance is zero. In reality, it is divided by  $(1 + T)$  where  $T=Y_R \cdot z$ ,  $z$  being the “transimpedance” of the direct chain. The output signal is only approximately divided by  $(1 + T)$  if the generator impedance of the current is not too small in comparison to the input impedance  $Z_e$  of the amplifier. If  $z$  tends

towards infinity, we get an ideal current-voltage convertor with an impedance transfer of  $1/Y_R$  ( $-R_2$  for the assembly shown in Figure 4.16).



**Figure 4.17.** Counter-feedback assembly of current applied as current

With this assembly, the output current  $I_2$  has servo-control of the input current  $I_1$ . This current crossing the charge resistance is then shared among the resistances  $R_1$  and  $R_2$ , which are both linked in parallel because  $U_1 = 0$ . Under these conditions, we can easily show that  $I_2 = I_{1R}(R_1 + R_2)/R_1 \approx I_1(R_1 + R_2)/R_1$ . The return chain is thus a coefficient of current transfer without dimension  $B_R = K_i$ . The loop transfer function is  $T = A_i \cdot K_i$  with  $A_i$  being the amplification in current of the direct chain. Since this is a retroactive effect in current, the input impedance is divided by  $(1 + T)$ , and the output impedance is multiplied by this term. When  $A_i$  is very high or if we suppose that  $U_1 = 0$ , the input impedance is zero and the output impedance is infinite, the current generator is then ideal and the current amplification equals  $1/K_i = (R_1 + R_2)/R_1$ .

These assemblies are important for conditioning signals before the measurement chain is introduced, especially with signals coming from sensors. According to the nature of the available signal, we can make use of one or the other of these counter-feedbacks.

### 4.3.1.2. Principle features of operational amplifiers

An operational amplifier is ideal if the differential amplification  $A_v$  is infinite, if the output voltage is zero when the input differential voltage is zero, and if it has an infinite input impedance as well as a zero output impedance. These conditions must exist throughout the entire range of frequency usage. Since in fact this kind of ideal amplifier does not exist, we will summarize the principle features of integrated differential amplifiers.

#### 4.3.1.2.1. Bandwidths

When the input signals are of high speed and amplitude, the linear functioning of the amplifier is limited by its maximum variation speed (also called the slew rate). This slew rate corresponds to the maximum gradient of the output signal, so that when the input signal exceeds this value, there is a reduction in amplitude and major distortion. In such cases, we speak of a “high signal” bandwidth. With these circuits, the term varies from 0.5 V/ $\mu$ s to 20 V/ $\mu$ s, but specific amplifiers can reach several hundred V/ $\mu$ s and 1,000 or 2,000 V/ $\mu$ s for hybrid components.

When the input signals are of low amplitude (“small signals”), the slew rate no longer appears. However, sometimes disturbances can create a bandwidth through the Miller effect. At the first order, we can see that the transfer function  $A_v$  in the open loop of an operational amplifier (the relation between the output voltage  $V_s$  and the differential voltage of input  $\epsilon$  in sinusoidal regime) is given as:

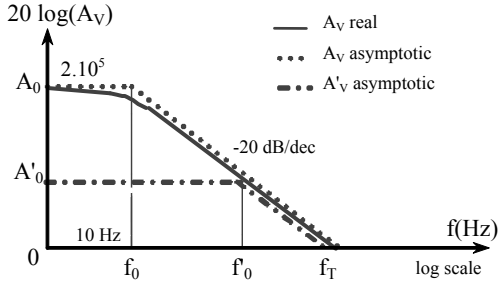
$$A_v = \frac{A_0}{1 + j \frac{f}{f_0}} \quad [4.23]$$

where  $A_0$  is the amplification in static open loop,  $f$  is the working or using frequency, and  $f_0$  is the cut-off frequency. For example, for an operational amplifier TL081,  $A_0$  is higher than  $2 \cdot 10^5$  and  $f_0$  is close to 10 Hz (see Figure 4.18).

The asymptomatic forms of the module (which is that of a first order bandwidth filter) of this function, in discontinuous features, are:

$$- A_v = A_0 \quad \text{for } f \ll f_0 \quad [4.24]$$

$$- |A_v| = \frac{A_0 \cdot f_0}{f} \quad \text{for } f \gg f_0 \quad [4.25]$$



**Figure 4.18.** Frequency response of an operational amplifier

Subsequently, the transfer function of an assembly using an operational amplifier in linear regime (with retroaction) keeps approximately the same expression. Here we suppose that  $A_0/A'_0$  is much higher than the unit in which  $A'_0$  indicates the static amplification of the assembly:

$$A'_v = \frac{A_v}{1 + \frac{A_v}{A'_0}} = \frac{\frac{A_0}{1 + j \frac{f}{f_0}}}{1 + \frac{A_0/A'_0}{1 + j \frac{f}{f_0}}} = \frac{A_0}{1 + j \frac{f}{f_0} + \frac{A_0}{A'_0}} \approx \frac{A'_0}{1 + j \frac{f}{f'_0}} \quad [4.26]$$

In this case, we get a new cut-off frequency  $f'_0$ :

$$f'_0 = \frac{A_0}{A'_0} \cdot f_0 \quad [4.27]$$

or a product called the bandwidth gain:

$$A'_0 \cdot f'_0 = A_0 \cdot f_0 = f_T \quad [4.28]$$

This bandwidth gain corresponds to the transition frequency  $f_T$ , for which the voltage amplification in open loop becomes equal to the unity (gain of 0 dB). This expression shows that the multiplication term between gain and cut-off frequency called “bandwidth” is conserved. We see that the bandwidth of an operational amplifier assembly is accordingly reduced when the static amplification of the closed loop is high.

4.3.1.2.2. Input impedances

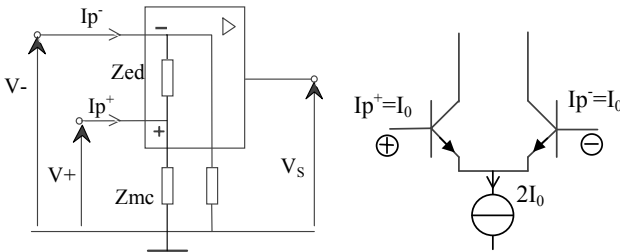
Two types of impedances exist at the input of an operational amplifier (see Figure 4.19). These are:

- differential impedance ( $Z_{ed}$ );
- common mode impedance ( $Z_{mc}$ ).

The differential input impedance of amplifiers with bipolar transistor bases rarely exceeds a few dozen  $M\Omega$ . At low frequencies, this impedance can be assimilated to a resistance with a “small signal” expression of:

$$R_{ed} = 2h_{11} = \frac{2\beta U_T}{I_C} = \frac{2}{40I_B} \tag{4.29}$$

where  $\beta$  is the current amplification of transistor currents,  $I_C$  is the polarization current, and  $U_T = kT/q = 25 \text{ mV}$  (at  $25^\circ\text{C}$ ), just as  $I_B = I_P^+ = I_P^- = I_0$ .



**Figure 4.19.** *Input impedances and polarization currents*

The input resistance in common mode (as seen on inverting and non-inverting inputs in relation to the ground), in the same conditions, is expressed by:

$$R_{mc} = 2\beta.R_E \tag{4.30}$$

where  $R_E$  indicates the polarization resistance that links the emitters to the power source.  $R_{mc}$  is in practice much higher than the differential input resistance ( $R_{ed} < R_{mc}$ ).

However, using JFET or MOSFET transistors in the differential stage of operational amplifiers helps us get, at low frequencies, very high input impedances (of the order of  $10^{12} \Omega$ ).

Nevertheless, for high frequencies it is important to bear in mind the differential capacities and the common mode of field-effect transistors, which are of the order of a picofarad. These can make certain assemblies oscillate or reduce their performances at high frequencies.

#### 4.3.1.2.3. Polarization currents and gap currents

Input impedances and input currents are obviously closely linked. The higher the input impedances are, the lower input currents become. For bipolar input stages, these currents vary from 1 to several dozen nA and from  $10^{-3}$  to  $10^2$  pA for input stages using field effect transistors or FETs. Also, because of the inevitable dissymmetry of the differential input stage, the input currents<sup>1</sup>  $I_p^+$  and  $I_p^-$  of the inverting and non-inverting limits of the amplifier are different.

The polarization current ( $I_p$ ) is then defined as the average value of the currents  $I_p^+$  and  $I_p^-$ . Their difference is called the gap current ( $I_d$ ):

$$I_p = \frac{I_p^+ + I_p^-}{2} \quad \text{and} \quad I_d = I_p^+ - I_p^- \quad [4.31]$$

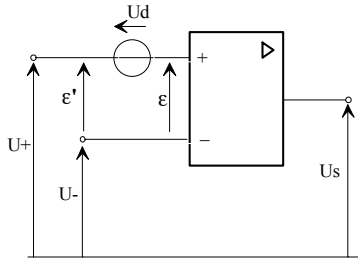
It is important to limit these currents. In practice, they bring about an output gap voltage whose value is dependent on the assembly being used. This is why amplifiers with FETs at the input stage are by far the best performing and most popular amplifiers currently used for instrumentation purposes.

#### 4.3.1.2.4. Input gap voltage

This residual voltage (or offset voltage  $U_d$ ) at the input of the amplifier (see Figure 4.20) is a product of the inevitable dissymmetry of the differential stage. To be more precise, it has to do with the different tensions that make up the stage (base emitter tensions for bipolar transistors and gate-source for field effect transistors). This voltage increases according to the gain of the amplifier assembly in closed loop. This can be a drawback for applications in which the amplifier signal is of low amplitude.

---

<sup>1</sup> Because the terms  $I_p^+$  and  $I_p^-$  correspond to base (or gate) currents of the differential stage, their input or output direction depends on the transistors being used (PNP, NPN, N-channel or P-channel).



**Figure 4.20.** Gap voltage  $U_d$

Advances in microtechnologies have helped to lower this voltage, which in previous generations of amplifiers sometimes ranged from several mV to as high as 100 mV. Today, thanks to current techniques of resistance adjustment by laser or Zener short circuiting (sometimes called *Zener zapping*), gap voltage is around 25  $\mu\text{V}$  for amplifiers using matched bipolar transistors and around 0.1 mV for amplifiers with FET transistors.

Gap voltage is a function of temperature, of the supply voltage of the amplifier, and of the amplifier's age. If the supply voltage is well-designed, correctly regulated and filtered, its influence is minimal in relation to that of the temperature.

Because of the evolution of the  $U_{be}$  and the  $U_{gs}$  with the temperature of transistors of the differential stage, the thermal drifts of the gap voltage are of the order of several  $\mu\text{V}/^\circ\text{C}$ . These drifts can be minimized by using the following techniques:

- we can try to create two differential stages to compensate for the drifts (LM121, for example);
- we can stabilize the substratum temperature, close to the differential stage, by means of a transistor that heats it (to a certain extent).

There are also self-switching circuits (one example is the ICL 7605 of Intersil) that use two differential amplifiers functioning alternately, switching to the rhythm of a clock. These are similar to the chopper-stabilized amplifiers like the circuit ICL 7650, also made by Intersil. The gap tensions we get are of the order of 2  $\mu\text{V}$  and the temperature drift of the first circuit is of the order of 0.1  $\mu\text{V}/^\circ\text{C}$ . However, the use of these components is limited to frequencies below that of their internal clocks.

In any case, most circuits have an interior adjustment that provides for an offset compensation.



4.3.1.2.5. Common mode rejection ratio

A gap voltage can come from common mode amplification  $A_{mc}$  of the operational amplifier. Actually, the output voltage depends not only on the input differential voltage but also on the common mode voltage  $U_{mc}$ , defined by:

$$U_{mc} = \frac{U^+ + U^-}{2} \tag{4.32}$$

The output voltage is then written as:

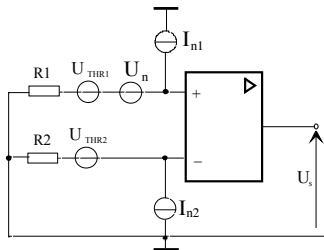
$$U_s = A_v \cdot U_d + A_{mc} \cdot U_{mc} = A_v \cdot (U^+ - U^-) + A_{mc} \cdot \left( \frac{U^+ + U^-}{2} \right) \tag{4.33}$$

It is clear that the higher the relation  $A_v/A_{mc}$ , the smaller the gap voltage is at the amplifier’s output. We call this common mode rejection ratio the Common Mode Rejection Ratio (CMRR). It is expressed in decibels in manufacturers’ specifications and can vary from 80 dB to 140 dB depending on the circuits.

We see that this ratio decreases with the utilization frequency and, in practice, is only a factor in assemblies with fairly high to high common mode voltage (for example, in differentiating assemblies and instrumentational amplifiers). This is because with assemblies having one input directly linked to the ground (as with reversed assemblies), we have:  $U^+ \approx U^- = 0V$ .

4.3.1.2.6. Noise

We model the noise sources appearing at the input of a differential amplifier assembly by three generators of basic noise ( $U_n$ ,  $I_{n1}$  and  $I_{n2}$ ), and by two other sources ( $U_{THR1}$  and  $U_{THR2}$ ). These are relative to the equivalent thermal noise resistances analyzed through inverting and non-inverting inputs, as shown in Figure 4.21.



**Figure 4.21.** Equivalent schema of noise of a differential amplifier assembly

The generators of noises specific to the amplifier ( $U_n$ ,  $I_{n1}$ , and  $I_{n2}$ ) are defined from the respective spectral densities ( $u_n$ ,  $i_{n1}$ , and  $i_{n2}$ ) for a frequency utilization band  $B_b = f_{\max} - f_{\min}$ :

$$U_n = \sqrt{\int_{f_{\min}}^{f_{\max}} u_n df}, \quad [4.34]$$

$$I_{n1,n2} = \sqrt{\int_{f_{\min}}^{f_{\max}} i_{n1,n2} df} \quad [4.35]$$

the sources of thermal or Johnson's noise being defined as:

$$U_{R1,R2} = \sqrt{4k TR_{1,2} B_b} \quad [4.36]$$

If we let  $I_n = I_{n1} = I_{n2}$ , the voltage of total noise at the input of a differential amplifier is written as:

$$U_{nT} = \sqrt{U_n^2 + (R_1 \cdot I_n)^2 + (R_2 \cdot I_n)^2 + U_{R1}^2 + U_{R2}^2} \quad [4.37]$$

Then the output noise voltage is:

$$U_s = A' U_{nT} \quad [4.38]$$

where  $A'$  is the assembly amplification (in closed loop) as defined in section 4.3.1.2.1.

Manufacturers indicate the spectral density values ( $u_n$ ,  $i_n$ ) of these components whose order varies from several  $\text{nV.Hz}^{-1/2}$  to about  $100 \text{ nV.Hz}^{-1/2}$  for  $u_n$ , and of  $0.01 \text{ pA.Hz}^{-1/2}$  to  $1 \text{ pA.Hz}^{-1/2}$  for  $i_n$ .

A well-known method for minimizing noise consists of using two amplifiers to get an amplification  $A'$ . The first one must be a low noise preamplifier ( $U_{n1} \ll U_{nT}$ ) with gain  $A_1$ , while the second can be an initial amplifier ( $U_{n2} = U_{nT}$ ) but used with a lower gain ( $A_2$ ), so that  $A' = A_1 \cdot A_2$ .

The total noise at assembly output is:

$$U'_s = \sqrt{(A_1 \cdot A_2 \cdot U_{n1})^2 + (A_2 \cdot U_{n2})^2} \quad [4.39]$$

$$U'_s = A' \cdot \sqrt{(U_{n1})^2 + (U_{nT} / A_1)^2} \quad [4.40]$$

If  $A_1 \gg 1$  then we get:

$$U'_s \approx A' U_{n1} \ll U_s \quad [4.41]$$

This demonstrates the importance of the first stage of amplification in the global noise level of the amplifier device.

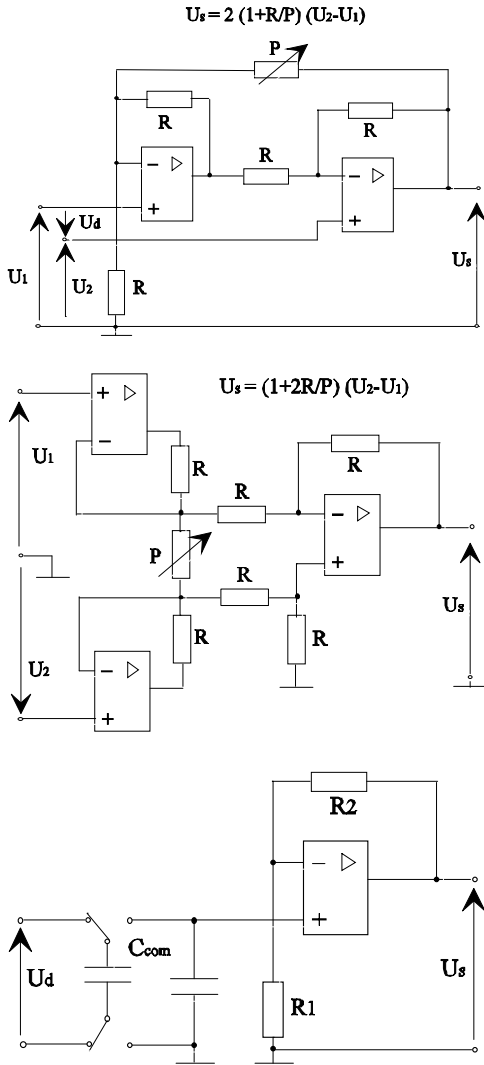
### 4.3.2. Instrumentation amplifiers

The name of these amplifiers comes from the fact that they are designed for the amplification of very low measurement signals (of the order of  $\mu\text{V}$  or of  $\text{mV}$ ) coming from sensors, transducers (constraint gauges, thermocouples) and measurement bridges such as Wheatstone's bridge, among others. They must perform well and must have the following features:

- a significant static amplification ( $> 10^6$ );
- a very low offset error ( $\approx 1 \mu\text{V}$ );
- a lower drift versus temperature and time ( $< 1 \mu\text{V}/^\circ\text{C}$  and  $< 1 \mu\text{V}/\text{month}$ );
- efficient values of noise, low in voltage and current, (of the order of  $1 \text{ nV}/\text{Hz}^{1/2}$  and of  $1 \text{ pA}/\text{Hz}^{1/2}$ );
- a high common-mode rejection ratio ( $> 100 \text{ dB}$ );
- a high input impedance in well-functioning high impedances;
- polarization currents below  $1 \text{ pA}$ ;
- a bandwidth and a high slew-rate according to the frequency and nature of the signal.

Several structures basic to operational amplifiers in order to get a differential amplifier are in use. The most frequently used are (see Figure 4.22):

- an instrumentational amplifier with two operational amplifiers;
- an instrumentational amplifier with three operational amplifiers;
- an instrumentational amplifier with switched-capacitor.

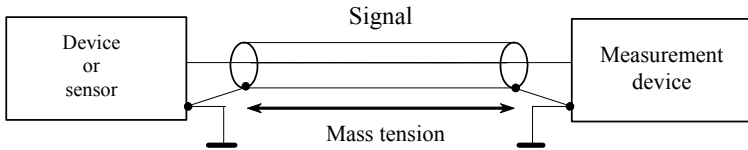


**Figure 4.22.** *Different assemblies of an instrumentation amplifier*

The first two assemblies offer a very high input impedance and regulation of the gain with one variable element (P). However, the common-mode error is, relatively speaking, higher for a structure with two amplifiers. The commuted capacity assembly can function up to frequencies of close to MHz and the common-mode rejection ratio reaches 120 dB. In addition, manufacturers have developed other schemata and specialized components.

### 4.3.3. Isolation amplifiers

This kind of amplifier is used when it is important to isolate an electronic monitoring sensor, such as those used in medical instrumentation (electrodes applied to the human body, to cite one example). These amplifiers help resolve issues arising from problems of ground that can occur when measurement signals are of low amplitude. For example, when we must connect two devices with a coaxial or sheathed cable, or connect a sensor with a measurement tool, each with its own ground, a fairly large “ground” voltage appears between the two ends of the cable.



**Figure 4.23.** Ground voltage resulting from use of a coaxial cable

There are three types of isolation amplifiers that we define according to the physical principle being used. “Galvanic” isolation is achieved by any of the three following ways:

- it can be achieved by electromagnetic coupling with the help of a transformer. In this case, a high frequency carrier is modulated in frequency or in impulse width by the signal that must be isolated. In particular, this principle is used in the isolation amplifiers made by Analog Devices. This company has demonstrated the advantages of using only one supply voltage that is shared between the “emission” and “reception” points of the transmission;

- it can be obtained by optical coupling (with a DEL emitter and a photodiode receiver). This technique does not require a high-frequency carrier and is well used by Burr-Brown. This company has succeeded in reducing linearity defects by using a retroaction mechanism with a second photodiode in the emission point;

- it can be obtained by capacitive coupling of a high frequency carrier modulated in frequency by the signal to be transmitted (the ISO 122 model made by Burr-Brown is an example).

The isolation tensions are of the order of 4 kV for isolation amplifiers with magnetic and capacitive coupling, and of several hundreds of volts for those using optical coupling. The bandwidths are lower by about 100 kHz.

#### 4.3.4. Logarithmic amplifiers

When a sensor's output dynamic is of a high amplitude (10 mV to 10 V, for example), it can be useful to compress the signal by using a logarithmic amplifier. After amplification and digitization, the signal can be easily transmitted across a transmission line. At reception, it is enough to carry out the reverse operation to restore the measurement signal. This principle allows us to lower noise sensitivity. Thanks to the compression, the output voltage to be digitized can be amplified (between 1 V and 5 V, for example) to get the most precise conversion. This is possible because the noise is independent of the level of the transmitted signal; in fact, it is easier to extract a signal of 1 V from noise than to do this with a signal of 1 mV.

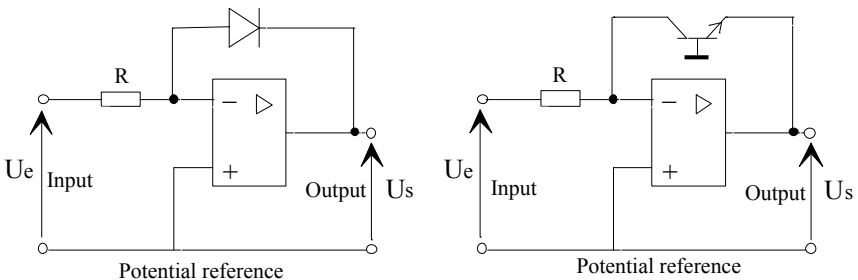
Logarithmic amplifiers also help us “linearize” sensors, carry out multiplications, divisions, elevations in the square, and extractions of the root squared.

To construct this type of amplifier (see Figure 4.24), we use the feature of a P-N junction with an equation (Ebres-Moll equation) in the following form:

$$i = i_o \left( \exp\left(\frac{qU}{kT}\right) - 1 \right) \quad [4.42]$$

$$i \approx i_o \cdot \exp\left(\frac{qU}{kT}\right) \quad [4.43]$$

where  $q$  is the electron charge,  $k$  the Boltzmann's constant,  $T$  is the absolute temperature, and  $U$  is the direct voltage and  $i_0$  is the flow of reverse current (extrapolated from  $U = 0$ ).



**Figure 4.24.** Schemata of logarithmic amplifier principle

We see that it is also possible to construct an exponential amplifier (or antilogarithmic) by using diodes or transistors, as shown in Figure 4.25.

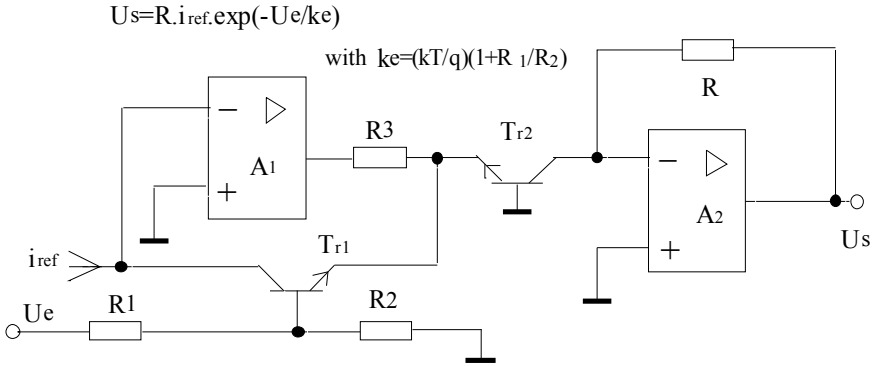


Figure 4.25. Schema of an exponential amplifier principle

4.3.5. Multipliers

Because of logarithmic and exponential amplifier circuits, it is possible to construct a multiplier assembly between two inputs, as shown in the functional schema in Figure 4.26.

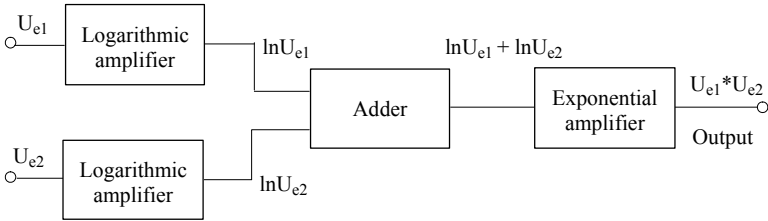


Figure 4.26. Functional schema for a multiplier

If we replace the functional block adder with a subtractor, we get a divider. Multipliers are usually sold in integrated forms; their price depends on how well they perform. They are mostly used to modulate amplitude, as well as in instrumentation for constructing synchronous detectors.

#### 4.4 Bibliography

- [BIC 92] BIQUARD M., BIQUARD F., *Signaux, systèmes linéaires et bruit en électronique*, Ellipses, 1992.
- [CHA 90] CHATELAIN J.D., DESSOULAVY R., *Electronique, tome 1 – Traité d'électricité d'électronique et d'électrotechnique*, Dunod, 1990.
- [COR 99] CORVISIER P., "Filtres à ondes de surface", *Mensuel Electronique*, no. 94, July 1999.
- [GRA 93] GRAY PAUL R., MEYER ROBERT G., *Analysis and Design of Analog Integrated Circuits*, Wiley International Edition, 1993.
- [HAS 89] HASLER M., NEIRYNCK J., *Filtres électriques – Traité d'électricité d'électronique et d'électrotechnique*, Dunod, 1989.
- [HOR 96] HOROWITZ P., HILL W., *Traité de l'électronique analogique et numérique*, Techniques Analogiques Elektor, 1996.
- [LET 87] LETOCHA J., *Circuits intégrés linéaires*, McGraw-Hill, 1987.
- [PAR 86] PARATTE P.A., ROBERT P., *Systèmes de mesure – Traité d'électricité d'électronique et d'électrotechnique*, Dunod, 1986.
- [PRI 97] PRICE T.E., *Analog Electronics: An Integrated PSpice Approach*, Prentice Hall, 1997.
- [TIE 78] TIETZE U., SCHENK CH., *Advanced Electronic Circuits*, Springer-Verlag, 1978.
- [TRA 96a] TRAN TIEN LANG, *Electronique analogique des circuits intégrés*, Groundon, 1996.
- [TRA 96b] TRAN TIEN LANG, *Circuits fondamentaux de l'électronique analogique*, 3<sup>rd</sup> ed., TEC & DOC, 1996.



*This page intentionally left blank*

# Chapter 5

## Analog Filters

### 5.1. Introduction

Sensors in instrumentation systems usually emit analog signals that must be conditioned before they are digitized (see Chapter 1). Analog filtering, which is indispensable in an electronic conditioning device, has two principal functions:

- improving the signal-to-noise ratio;
- eliminating any frequencies that might be aliased by sampling that precedes digital processing.

Analog filters are also used in the output stages of instrumentation systems when it is necessary to reconstruct an analog signal from the conversion of a digital signal.

Filters are complex and expensive mechanisms, with features that are often crucial to the overall performance of a system.

### 5.2. Technological constraints

The role of a filter is to separate the useful frequencies in a signal (those that carry information) from unwanted frequencies, such as noise or other signals.

The basic circuit of a filter is the resonator, a mechanism whose performance varies very selectively according to the frequency. A resonator is for the most part modeled by a transfer function of the second order, with parameters that are the resonance frequency  $f_0$  and the quality coefficient  $Q$ . For example:

$$H(p) = \frac{\omega_0^2}{p^2 + \omega_0^2 \frac{p}{Q} + \omega_0^2} \quad \text{with} \quad Q = \frac{|H(j\omega_0)|}{|H(0)|}$$

An efficient instrumentation filter must have the following qualities:

- it should have high quality coefficients  $Q$  to get a good frequency discrimination;
- it needs very precise and stable resonance pulsations, according to temperature and time. In particular, imprecise component values must have only a slight repercussion on the values of  $Q$  and  $\omega_0$ . If this condition is met, we can say that the filter presents a low sensitivity to component imperfections;
- it should have a low noise level and a high dynamic. Passive filters, which are constructed without electronic components, are especially important here;
- this kind of filter should be relatively small;
- it should entail reasonable production costs. Adjustments that are often necessary for producing precise filters are incompatible with this constraint.

Electronic resonators respond badly to the first two constraints, especially if the filter must be produced in the form of integrated circuits. That is why high quality filters are basically made with mechanical resonators with surface or volume waves. In this filter type, mechanical resonance is transformed into electrical resonance by piezoelectricity [DIE 74]. However, this technique does not work except with high frequencies (more than a few MHz). For lower frequencies that mostly exist in instrumentation systems, we get around this problem by using the following three techniques:

- we can use inductors and capacitors. If the L-C resonators individually have mediocre stability, the filters made with the help of this technique have, overall, an excellent stability (with a few restrictive conditions). This paradoxical property is directly turned to good account when technological constraints making adequate inductors are not prohibitory. Otherwise, we can make electronic copies of filters from L-C models;
- we can use active filters that have only resistors, capacitors and amplifiers. These kinds of filters are made by putting basic filters into cascade or by copying L-

C filters. The mediocre stability of these mechanisms can be partly compensated by settings that eventually have servo-control;

– we can use switched capacitor filters having only capacitors, amplifiers and switches. Because of an ingenious device invented around 1980 [ALL 78], these filters, which can be completely integrated into the system, have excellent stability and precision and do not require any adjustment. However, their application field is limited to several hundred kHz.

These three basic techniques are presented in this chapter, after a brief summary of general calculation methods. These techniques are still evolving, connected as they are to rapid, constant and unforeseeable progress in electronic component technologies.

### 5.3. Methods of analog filter calculation

An analog filter used in instrumentation must respond to fixed behavioral specifications in the frequential domain, with attenuation  $A$  (expressed in decibels dB) and phase difference  $\varphi$ . For a filter with inputs and outputs  $V_e$  and  $V_s$ , these variables are defined by the relations:

$$A = 20 \log \left| \frac{V_e(j\omega)}{V_s(j\omega)} \right| = 20 \log |H^{-1}(j\omega)|$$

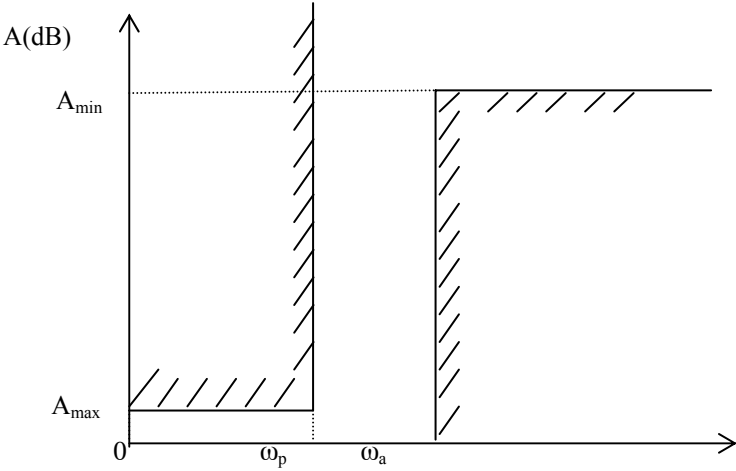
$$\varphi = \arg(H(j\omega))$$

In these expressions,  $H(j\omega) = V_s(j\omega)/V_e(j\omega)$  is the isochronous transfer function of the filter.

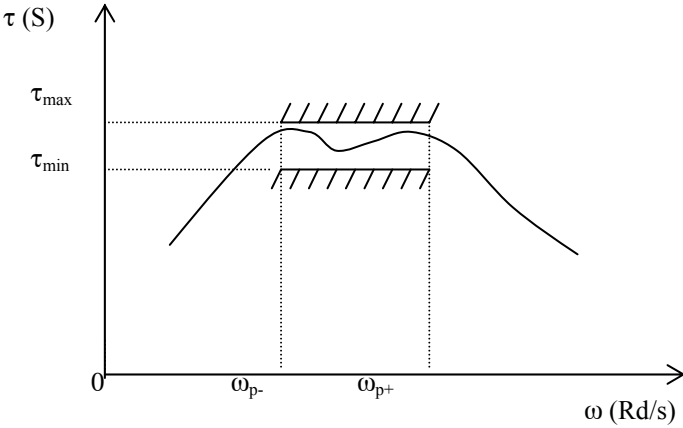
In order for a filter to transmit a signal without deformation, the phase difference must vary linearly according to the frequency. If this occurs, the derivative of the phase difference in relation to the frequency has a constant value. This variable is the group delay of group  $\tau$  defined by the relation:

$$\tau = - \frac{\partial \varphi}{\partial \omega}$$

In practice, the values of  $A$  and  $\tau$  become written within the attenuation gauge and the group delay gauge, as shown for the attenuation in Figure 5.1 and the group delay in Figure 5.2.

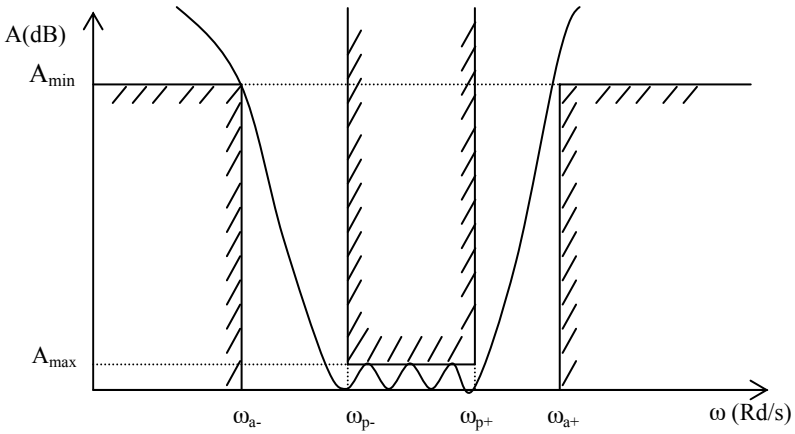


**Figure 5.1.** Gauge of a low pass attenuation filter



**Figure 5.2.** Gauge in group delay of a band pass filter

In most instrumentation applications, the attenuation feature is the most important. We calculate a filter from the attenuation gauge and, if the application requires it, we correct the inequality of the group delay with the help of an additional corrective filter [SED 79].



**Figure 5.3.** Attenuation gauge of a low pass filter

There are four types of gauges that measure attenuation, with corresponding parameters:

- low pass, with parameters  $\omega_p$ ,  $\omega_a$ ,  $A_{\max}$  and  $A_{\min}$  (see Figure 5.1);
- high pass, with parameters  $\omega_p$ ,  $\omega_a$ ,  $A_{\max}$  and  $A_{\min}$ ;
- band pass, with parameters  $\omega_{p+}$ ,  $\omega_{p-}$ ,  $\omega_{a+}$ ,  $\omega_{a-}$ ,  $A_{\max}$  and  $A_{\min}$  (see Figure 5.3);
- band reject, with parameters:  $\omega_{p+}$ ,  $\omega_{p-}$ ,  $\omega_{a+}$ ,  $\omega_{a-}$ ,  $A_{\max}$  and  $A_{\min}$ .

In instrumentation systems, band pass filters are most widely used (for anti-aliasing, improvement of signal-to-noise ratio, and for reconstruction). However, band pass and band reject gauges are also used when a major amplitude disturbance must be eliminated before conversion into digital signals.

In all cases, the usual method of calculating is to determine, from the attenuation gauge, the transfer function of a band pass filter which is called the prototype. From this we deduce, by conversion, the transfer function of the filter being planned. The synthesis is then carried out from this function. We see that the transfer function of the prototype is always normalized; that is, it is calculated by taking the cut-off pulsation  $\omega_p$  as unity.

### 5.3.1. Attenuation functions of standard low pass prototype filters

There are four types of important and well identified attenuation functions of prototype filters. These give rise to transfer functions that can be physically produced and verified:

– Butterworth filters:

$$\left|H^{-1}(j\omega)\right|^2 = 1 + \varepsilon^2 \omega^{2n}$$

– Direct Tchebycheff filters:

$$\left|H^{-1}(j\omega)\right|^2 = 1 + \varepsilon^2 T_n^2(\omega)$$

$$T_n(\omega) = \cos(n a \cos(\omega))$$

In these relations,  $n$  is the order of the filter, that is, the degree of its transfer function.  $\varepsilon$  is a parameter depending on attenuation tolerance in the pass band  $A_{\max}$ :

$$\varepsilon^2 = 10^{A_{\max}/10} - 1$$

For these two types of filters, the inverse of the transfer function is a polynomial of variable  $\omega$ . They are thus called polynomial filters. The attenuation curve of Butterworth filters is said to be maximally flat because all the attenuation derivations are zero at pulsation 0. However, the pass band response of Tchebycheff filters fluctuates  $n + 1$  times between the values 0 and  $A_{\max}$  (in dB). These are called equiripple filters.

In order to meet the requirements of a given filter, the necessary order  $n$  is much higher for a Butterworth filter than for a Tchebycheff filter. Polynomial filters are easy to make because their transfer function is simple.

The following are formulae for elliptical filters:

$$\left|H^{-1}(j\omega)\right|^2 = 1 + \tilde{\varepsilon}^2 \frac{\prod_{i=1}^{n/2} (\omega^2 - \omega_{0i}^2)}{\prod_{i=1}^{n/2} (\omega^2 - \omega_{\infty i}^2)} \text{ if } n \text{ is even}$$

$$\left|H^{-1}(j\omega)\right|^2 = 1 + \bar{\epsilon}^2 \frac{\omega^2 \prod_{i=1}^{n-1/2} (\omega^2 - \omega_{0i}^2)}{\prod_{i=1}^{n-1/2} (\omega^2 - \omega_{\infty i}^2)} \quad \text{if } n \text{ is odd}$$

$$\text{with: } \omega_{0i} \omega_{\infty i} = \frac{\omega_a}{\omega_p}$$

where  $\omega_{0i}$  and  $\omega_{\infty i}$  are respectively normalized pulsations for which the attenuation is zero and infinite and  $\bar{\epsilon}$  is a constant. The numerical values of these parameters can be calculated analytically, as Cauer has shown. However, nowadays it is simpler and more efficient to calculate these values by direct digital optimization; the existence and unicity of the solution have been shown, leading to a rapid and certain convergence of the algorithm for the calculation.

The attenuation curve of Cauer filters fluctuates  $n + 1$  times between the extreme values allowed by the gauge. This is true both for pass bands and attenuated pass bands. Of all filters, these meet the requirements of a given gauge with a transfer function of minimal order  $n$ . Unfortunately, their group propagation time is extremely irregular, which means they cannot be used when the temporal form of a signal must be preserved.

– Inverse Tchebycheff filters:

$$\left|H^{-1}(j\omega)\right|^2 = 1 + \frac{1}{\bar{\epsilon}^2 T_n^2\left(\frac{1}{\omega}\right)}$$

$$T_n(\omega) = \cos(n a \cos(\omega))$$

$\bar{\epsilon}$  is a scale constant.

These filters are among the best available in terms of stiffness of the attenuation curve and pass band group delay regularity. These properties make this kind of filters very useful in instrumentation systems.



### 5.3.2. Transfer functions of common prototype low pass filters

Attenuation functions define the square of the module of the transfer function. A standard, but sometimes difficult to realize, mathematical calculation helps us deduce the transfer function. It is carried out as follows, by noting the transfer function  $H(p) = P(p)/E(p)$ :

$$H^{-1}(p)H^{-1}(-p) = \frac{E(p)E(-p)}{P(p)P(-p)} = 1 + \frac{F(p)F(-p)}{P(p)P(-p)} = \left| H^{-1}(j\omega) \right|^2$$

By identifying this expression with the attenuation functions calculated above, we get polynomial values  $P(p)$  and  $F(p)$ .

We then get  $E(p)$  from:

$$E(p)E(-p) = P(p)P(-p) + F(p)F(-p)$$

This equation is known as the Feldtkeller equation. It has a unique solution because  $E(p)$  is a Hurwitz polynomial (its roots are real negative parts) so that the transfer function is that of a stable field.

Resolving the Feldtkeller equation requires a computer (there are many software programs available that can do this, such as Matlab and Mathcad), except for Butterworth filters, which allow for a simple analytic solution:

$$E(p) = H^{-1}(p) = \varepsilon \prod_{i=1}^{n/2} \left( p^2 + 2\delta p \cos \frac{(2i-1)\pi}{2n} + \delta^2 \right) \text{ if } n \text{ is even}$$

$$E(p) = H^{-1}(p) = \varepsilon (p + \delta) \prod_{i=1}^{(n-1)/2} \left( p^2 + 2\delta p \cos \frac{i\pi}{n} + \delta^2 \right) \text{ if } n \text{ is odd}$$

with  $\delta = \varepsilon^{(-1/n)}$

### 5.3.3 Transfer functions of derived filters

To obtain transfer functions of filters made from the normalized prototype, we apply a frequential transformation according to the transfer function of the low pass prototype. These transformations, which are always used, are as follows:

– low pass low pass transformation:  $p \rightarrow \frac{p}{\omega_0}$

- low pass high pass transformation:  $p \rightarrow \frac{\omega_0}{p}$
- low pass pass band transformation:  $p \rightarrow \frac{(p^2 + \omega_0^2)}{Bp}$
- low pass band rejector transformation:  $p \rightarrow \frac{Bp}{p^2 + \omega_0^2}$

In these formulae,  $\omega_0$  is the central pulsation of the band pass or band reject, expressed in radian/second, with B the width of the band pass or band reject, in the same unity. These transformations are complicated and difficult to carry out, necessitating the use of a computer. Many commercial software programs contain these transformations (Matlab, Mathcad and Matrixx are examples, as are most signal analysis programs).

### 5.3.4. Filter synthesis carried out from the transfer function

Two synthesis methods of an analog filter, done from transfer functions, are widely used.

#### *Cascade synthesis*

This is based on the decomposition that can always occur of  $H(p)$  in biquadratic terms (and of a first degree term when order n of the filter is odd):

$$H(p) = \frac{\alpha_m p^m + \alpha_{m-1} p^{m-1} + \dots + \alpha_0}{\beta_n p^n + \beta_{n-1} p^{n-1} + \dots + \beta_0} = G \prod_{i=1}^{n/2} \frac{a'_i p^2 + b'_i p + 1}{a_i p^2 + b_i p + 1} = G \prod_{i=1}^{n/2} B_i(p)$$

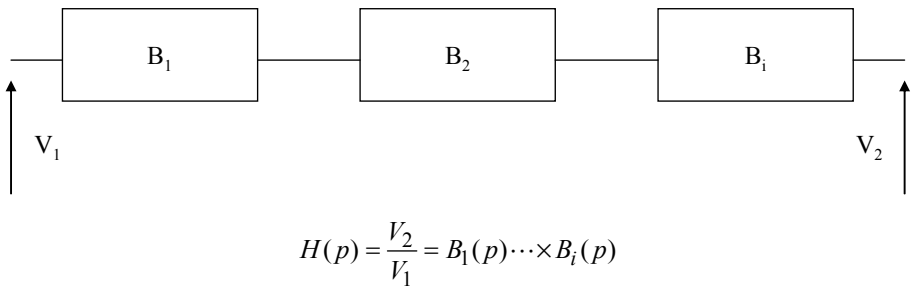
if n is even, and:

$$H(p) = G \frac{a'_0 p + 1}{a_0 p + 1} \prod_{i=1}^{\frac{n-1}{2}} \frac{a'_i p^2 + b'_i p + 1}{a_i p^2 + b_i p + 1} = G \frac{a'_0 p + 1}{a_0 p + 1} \prod_{i=1}^{\frac{n-1}{2}} B_i(p)$$

if n is odd.

To construct this kind of filter, we need to cascade basic biquadratic circuits with transfer function  $B_i(p)$  (and an additional first order circuit when n is odd). Here we

must ensure that no circuit interacts with another (see Figure 5.4). This condition is met if the output impedance of each basic circuit is zero, or very low in comparison to the input impedance of the next circuit.



**Figure 5.4.** Principle of cascade synthesis when circuits  $B_i$  are independent

However, it is impossible to meet this condition with passive elements, although it is easy to do with active circuits because of the very low output impedance of operational amplifiers.

Each biquadratic element depends on only four parameters. These are  $a_i$ ,  $b_i$ ,  $a'_i$  and  $b'_i$ , and can be created by standard circuits that are easy to adjust individually and available in the form of hybrid or integrated modules from many manufacturers. The best way to adapt these elements to specific needs is to construct them, with the help of resistors, capacitors and operational amplifiers, using methods which we will describe later. These methods are very easy to implement, both in terms of calculations and from the point of view of adjustments and maintenance. They are universally applicable and work no matter which transfer function needs to be synthesized. Unfortunately, this structure does not have a good sensitivity in relation to component value variations. In particular, it does not allow us to construct narrow band pass filters (< 5%) as soon as the frequency exceeds several tens of kHz.

### *Comprehensive synthesis*

This method involves synthesizing the filter with one network in order to minimize sensitivities. Calculations, adjustments, and optimizations are much more complex. Since L-C filters have an optimal sensitivity (see section 5.4), most of these methods directly or indirectly stimulate L-C structures.

#### 5.4. Passive filter using inductors and capacitors

Passive filters using inductors  $L$  and capacitors  $C$  were first developed in 1923. These are called Zobel filters [ZOB 23]. They are made of a quadripole that only contains this type of element, inserted between two resistors  $R_1$  and  $R_2$ , respectively the generator resistor and the charge (Figure 5.5).

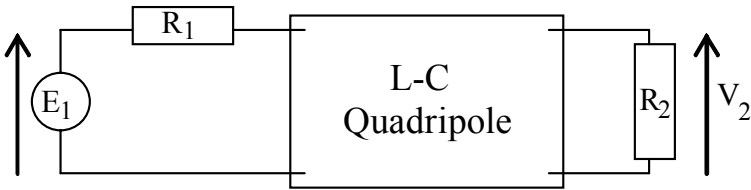


Figure 5.5. Schema of the principle for an L-C filter

These filters have the following advantages.

- they do not require a power supply;
- they have good dynamics;
- they have a very low noise level;
- they have low response sensitivity to the values of their own components.

However, these filters require the use of inductors, components that are costly, take up much space, and cannot be adjusted for industrial purposes. This is why, in mass production, L-C filters have been increasingly replaced by active filters, at least for in low and medium frequency applications. However, the technologies involved in manufacturing inductors have improved in recent years. Miniaturization, quality and costs have continued to improve. Laboratories have created higher quality inductors that can be integrated and electronically regulated.

Moreover, the very low sensitivity of L-C filters cannot be matched. Therefore, we use these filters as models to produce “electronic copies” with the same sensitivity features, but without inductors. These techniques will be discussed in sections 5.4.2 and 5.4.3. In practice, the only structures that are still used are ladder filters, as shown in Figures 5.6 and 5.7.

### 5.4.1. Sensitivity; Orchard's theorem and argument

The basic schema of an L-C filter is shown in Figure 5.5. The voltage generator  $E_1$  of internal impedance  $R_1$  can supply charge  $R_2$  with a maximum power  $P_{1m} = |E_1|^2 / 4R_1$ .

We define the transmission function in power by the relation of  $P_{1m}$  to power  $P_2$  effectively furnished to the charge by the relation:

$$\frac{P_{1m}}{P_2} = \frac{|E_1|^2}{4R_1} \frac{R_2}{|V_2|^2} = \frac{R_2}{4R_1} \left| \frac{E_1}{V_2} \right|^2 = \left| H^{-1}(j\omega) \right|^2$$

with:  $\frac{E_1}{V_2} \sqrt{\frac{R_2}{4R_1}} = H^{-1}(p)$

The attenuation  $A(\omega)$  of the filter expressed in decibels is given by the relation:

$$A(\omega) = 20 \log \left| H^{-1}(j\omega) \right|$$

This variable is always positive or zero.

Let  $X_i$  be the value of an element of the filter (inductor or capacitor), and  $\partial A / \partial X_i$  the partial derivation of  $A$  in relation to the value of this element. If  $X_i$  diverges from its nominal value by a quantity  $\Delta X_i$  (because of temperature variation or a time drift, for example), the corresponding variation of attenuation  $A$  moves near the first order by means of:

$$A(X_i + \Delta X_i) = A(X_i) + \Delta X_i \frac{\partial A}{\partial X_i}$$

At frequencies at which adaptation is carried out,  $P_2 = P_{1m}$  and  $A = 0$  (attenuation zeros), which gives us:

$$\Delta X_i \frac{\partial A}{\partial X_i} = A(X_i + \Delta X_i) \geq 0$$

Since variation  $\Delta X_i$  is a rather vague sign, the relation can only be verified with an equal sign. From this we get the following theorem: the partial derivation of the L-C filter attenuation inserted between resistors in relation to the value of each of the elements is annulled to the attenuation zeros.

This property is extended to all the non-dissipating quadrupoles whatever synthesis method is used. It is valid only for attenuation zeros. However, throughout the pass band of a filter, attenuation remains, by definition, close to zero; and we can see that  $\partial A/\partial X_i$  remains low, even lower than the maximum attenuation is low. This last proposition has not been proven by very rigorous methods, but it has been confirmed often in practical experience. It is known as Orchard's argument.

Orchard's argument is very important because it shows that a non-dissipating filter inserted between two resistors allows important tolerances on the value of its components without affecting its performance. This characteristic is especially important since the function of a filter is to accurately discriminate narrow frequency fields. This feature demands, *a priori*, very stable and precise components. Zobel filters could attain very high selectivity despite the use of the relatively mediocre components then available.

However, active filters do not have the advantages of this property. That is why, at this time, constructing a high performance filter using an L-C prototype must be done by electronic simulation. We must remember the restrictive conditions of Orchard's argument when carrying out this procedure:

- it is not valid except for the pass band (and not for the stopband where admissible tolerances for attenuation are notably higher);
- it only applies if attenuation zeros correspond to the transmission of maximum available power. Filters taking in a constant non-zero pass band attenuation are therefore excluded;
- it applies only to attenuation and not to propagation delay.

#### 5.4.2. Low pass ladder filters

In practice, most L-C filters have a ladder structure. This topology avoids the use of transformers and gives better results. From a given gauge, we must establish the schema and calculate the value of the elements of this schema.

Cauer and Darlington developed a systematic method of synthesizing ladder filters [CAU 41] from the transfer function and the characteristic function. This very powerful method makes use of very complex theoretical development. However,

implementing it is simple if we limit ourselves to common and straightforward ladder structures. This simplified procedure works for most filtering applications used in acquisitions systems. That is why we limit ourselves in this book to this presentation. The reader who is interested in more complete developments may refer to [SED 79] and [SAA 58]. The simplified method presented here consists of first calculating a low pass filter prototype with a simple repetitive structure (canonic). We can then deduce the planned filter from frequential transformations.

5.4.2.1. Structures of basic low pass filters

These filters have topologically simple schemata. They are made of branches that have, at most, two elements: an inductor and a capacitor. With the branches in series, these elements are in parallel. With the parallel branches, these elements are in series. The filter has  $n$  branches if its transfer function is of the order  $n$ .

The first branch can be a series branch (structure in T) or a parallel branch (structure in  $\Pi$ ). We get only two schemata type; these can be deduced from one another by duality (see Figure 5.6). The filters in  $\Pi$  of odd order terminate with a parallel branch of those of even order with a series branch. The opposite is true for T filters. We see that if a branch contains two elements (an inductor or a capacitor), this branch provokes a signal interruption transmitted to the charge to its resonance frequency (zero transmission to finite frequency). Polynomial low pass filters thus have only one element per branch and their structures are quite simple.

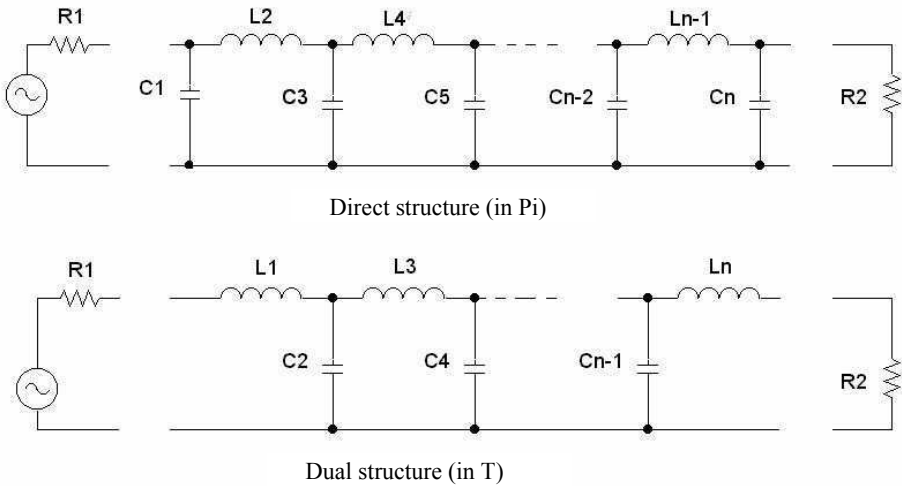
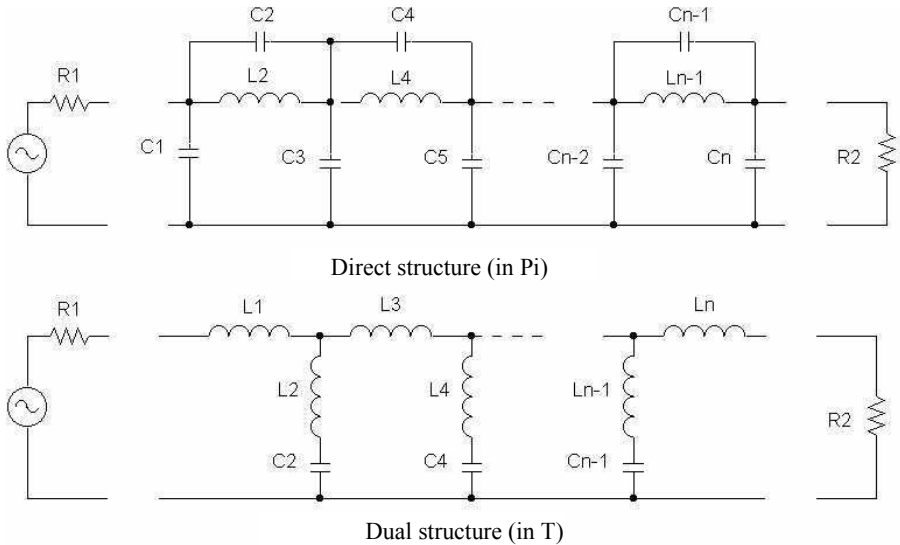


Figure 5.6. Schemata of polynomial low pass filters

The transmission zeros of Cauer and low pass inverse Tchebycheff filters are produced by trapped circuits in series in parallel branches. Only odd order filters can be produced by L-C technology (the explanation for this is given later). We thus have one schema in T (and its equivalent in  $\Pi$ ) as shown in Figure 5.7.

The schemata shown help produce standard low pass filters, such as the Butterworth filter, the direct and inverse Tchebycheff filter, and the elliptic filter. These give excellent results in most practical applications.

The schema being established from the transfer function, we need only calculate the value of the elements. This calculation can be carried out analytically following Darlington's decomposition method, or it can be done directly by modern methods of digital optimization, if the necessary software is available.



**Figure 5.7.** Schemata for elliptic and inverse Tchebycheff filters

#### 5.4.2.2. The Darlington analytic synthesis

Even though the theoretic developments of this method are complex, the method is simple to use in the following ways:

a) From the attenuation gauge, we choose the type of filter we want to complete. We determine the minimal order and calculate the transfer functions and polynomials  $E(p)$ ,  $P(p)$ , and  $F(p)$  using the method described in section 5.3.2.



b) By knowing polynomials  $E(p)$ ,  $P(p)$ , and  $F(p)$ , we can deduce these L-C quadrupole impedances:  $z_{11}$ ,  $z_{22}$ ,  $y_{11}$ , and  $y_{22}$ . The results of this calculation are given in Table 5.1. Indices  $_p$  and  $_i$  indicate the even and odd parts of polynomial  $E(p)$  and  $F(p)$ . This means  $E_p$  indicates the even part of polynomial  $E(p)$ . These results have been established from Darlington’s original calculation [HAS 81].

c) From any one of these four impedances, we calculate the successive values of the branch impedances by a procedure of iterative extraction. For polynomial filters, this procedure is very simple; it is only a development of continued fractions of the initial impedance. For example, for a structure in T shown in Figure 5.6, it is:

$$z_{11} = L_1 p + \frac{1}{C_2 p + \frac{1}{L_3 p + \frac{1}{C_4 p + \dots}}}$$

	$z_{11}$	$z_{22}$	$1/y_{11}$	$1/y_{22}$
Structure in $\Pi$	$R_1 \frac{E_i + F_i}{E_p - F_p}$	$R_2 \frac{E_i - F_i}{E_p - F_p}$	$R_1 \frac{E_p + F_p}{E_i - F_i}$	$R_2 \frac{E_p + F_p}{E_i + F_i}$
Structure in T	$R_1 \frac{E_i - F_i}{E_p + F_p}$	$R_2 \frac{E_i + F_i}{E_p + F_p}$	$R_1 \frac{E_p - F_p}{E_i + F_i}$	$R_2 \frac{E_p - F_p}{E_i - F_i}$

**Table 5.1.** *Quadrupole impedances*

This decomposition can occur element after element by making  $p$  tend towards infinity:

$$L_1 p = z_{11} \Big|_{p \rightarrow \infty}$$

$$C_2 p = \frac{1}{z_{11} - L_1 p} \Big|_{p \rightarrow \infty}$$

etc.

For elliptic and inverse Tchebycheff filters, the procedure is a bit more complicated. This is because the frequency of L-C branches must be positioned to their values, that is, to frequency pulsations  $\omega_{\infty i}$ , which annul  $P(p)$  and correspond to a transmission zero. For example, for the structure in T shown in Figure 5.7, we get:

$$z_{11} = L_1 p + \frac{1}{Y_2 + \frac{1}{L_3 p + \frac{1}{Y_4 + \dots}}}$$

with:  $Y_i = \frac{C_i p}{1 + L_i C_i p^2} = \frac{C_i p}{1 + \frac{p^2}{\omega_{\infty i}^2}}$

In this case, the fractional decomposition continues as follows:

$$L_1 p = z_{11} \Big|_{p^2} \rightarrow -1/L_2 C_2$$

$$Y_2 = \frac{1}{z_{11} - L_1 p} - Y_r$$

etc.

We see that in all cases, we must use a calculator to carry out these operations precisely in order to avoid cumulative errors at each stage. We illustrate this mechanism with two examples.

d) Terminal resistor calculation. For filters with zero attenuation at zero frequency, terminal resistors  $R_1$  and  $R_2$  are equal. Their value determines the impedance level of the filter and can be taken as unity of normalization:  $R_1 = R_2 = R_0$ .

For filters with attenuation equal to  $A_{max}$  at zero frequency, and infinite at infinite frequency (such as direct even order Tchebycheff filters), the resistor of the generator is taken as unity and the terminal resistor is calculated in the following way. By writing that the attenuation at  $\omega = 0$  is obtained by bridge divider  $R_1, R_2$ , the L-C filter acting as a short circuit between the two resistors is expressed as:

$$\frac{P_m}{P_2} = \frac{E_1^2}{4R_1} \frac{V_2^2}{R_2} = \frac{(R_1 + R_2)^2}{4R_1 R_2}$$

The attenuation at  $\omega = 0$  is:

$$A_{\max} = 10 \log \left[ \frac{(R_1 + R_2)^2}{4R_1 R_2} \right]$$

If  $R_1 = 1$ , we deduce from it  $R_2^2 + 2R_2(1 + 2\alpha) + 1 = 0$

with  $\alpha = 10^{A_{\max}/10}$  and so:

$$R_2 = (2\alpha - 1) \pm \sqrt{(1 - 2\alpha)^2 - 1}$$

In this relation, the sign + is used for structures in T and the sign – for those in  $\Pi$ .

It is important to be aware that elliptic and inverse Tchebycheff filters of even order can not be completed in L-C filter format. In fact, at pulsations  $\omega = 0$  and  $\omega = \infty$ , an L-C filter acts as a short circuit between the generator and the charge, or simply as a short circuit. For these types of filters, and at two frequencies, attenuations have distinct and non-zero finite values  $A_{\max}$  and  $A_{\min}$ : this means that synthesis is not possible. In practice, we avoid this difficulty by carrying out a prior transformation of the transfer function [SED 79].

#### 5.4.2.3. Examples of synthesis

##### Example 1

We can produce a low pass Tchebycheff filter of the order 6,  $A_{\max} = 1$  dB, and whose cut-off frequency is 1 MHz. The retained structure is in T. The calculation done according to the method discussed in section 5.3.2 gives the characteristic polynomials  $E(p)$ ,  $F(p)$  and  $P(p)$  in the normalized form (that is, that the cut-off frequency is taken as frequency unit):

$$E(p) = 16.28p^6 + 15.11p^5 + 31.44p^4 + 19.57p^3 + 15.3p^2 + 5p + 1.22$$

$$F(p) = 16.28p^6 + 24.425p^4 + 9.159p^2 + 0.5088$$

$$P(p) = 1$$

Since  $F(p)$  is even,  $F(p) = F_p$  and  $F_1 = 0$ . This gives us  $z_{11} = y_{22}$  and  $z_{22} = y_{11}$ .

$$E_p = 16.28p^6 + 31.44p^4 + 15.3p^2 + 1.22$$

$$E_i = 15.115p^5 + 19.57p^3 + 5p$$

By carrying over the values in Table 5.1, we get parameters  $z$  and  $y$  of the filter in normalized form ( $R_1 = 1$  for the calculation of  $z_{11}$  and  $y_{11}$  and  $R_2 = 1$  for the calculation of  $z_{22}$  and  $y_{22}$ ):

$$z_{22} = y_{11} = \frac{32.57p^6 + 55.86p^4 + 24.45p^2 + 1.631}{15.11p^5 + 19.57p^3 + 5p}$$

$$z_{11} = y_{22} = \frac{7.015p^4 + 6.136p^2 + 0.6132}{15.11p^5 + 19.57p^3 + 5p}$$

The schema obtained by the fractional decomposition of  $z_{11}$  is an L-C ladder in T whose normalized component values (and therefore without dimensions) are:

$$L_1 = 2.1546 \qquad C_2 = 1.1041 \qquad L_3 = 3.0634$$

$$C_4 = 1.1518 \qquad L_5 = 2.9367 \qquad C_6 = 0.8101$$

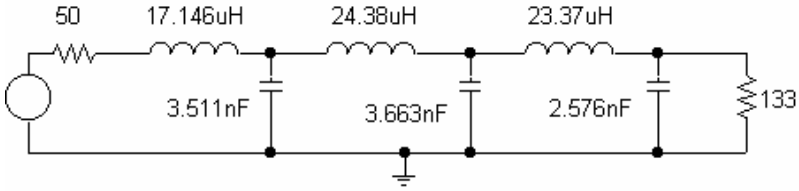
The terminal resistors are calculated as shown above:

$$\alpha = 10^{A_{\max}/10} \quad \text{which gives us:} \quad R_2 = (2\alpha - 1) + \sqrt{(1 - 2\alpha)^2 - 1} = 2.66$$

In order to obtain the real values of the components, we must determine the unitary values of the capacitor and the inductor, after having determined an impedance level by choosing resistor  $R_1$ , which we take as unity. For example, if we choose  $R_1 = R_{\text{unitary}} = 50 \, \Omega$ , we deduce  $R_2 = 133 \, \Omega$ , and from relations  $R_u C_u \omega_u = 1$ , and  $L_u \omega_u = R_u$ :

$$C_u = \frac{1}{2\pi f_u R_u} = 3.18 \text{ nF} \quad \text{and} \quad L_u = \frac{R_u}{2\pi f_u} = 7.958 \, \mu\text{H}$$

The definitive schema is shown in Figure 5.8.



**Figure 5.8.** Schema of a low pass filter calculated in Example 1. The resistors are in Ohms

*Example 2*

We can construct a low pass elliptic filter of order 5,  $A_{max} = 0.2$  dB and  $A_{min} = 40$  dB, with a cut-off frequency of 50 kHz. The structure is in T. The calculation follows the method presented in section 5.3.2 and gives characteristic polynomials  $E(p)$ ,  $F(p)$ , and  $P(p)$  in normalized form.

$$E(p) = 17.87p^5 + 26.18p^4 + 43.68p^3 + 36.07p^2 + 23.26p + 8.141$$

$$F(p) = 17.87p^5 + 24.53p^3 + 7.319p$$

$$P(p) = p^4 + 6.12p^2 + 8.14$$

In carrying over the values in Table 5.1, we get parameters  $z$  of the filter (if we take  $R_1 = 1$  and  $R_2 = 1$ ):

$$z_{11} = z_{22} = \frac{26.18p^4 + 36.06p^2 + 8.14}{35.74p^5 + 68.21p^3 + 30.58p}$$

The normalized value of the elements is obtained by the fractural decomposition according to the method shown above:

$$L_1p = z_{11} | p^2 \rightarrow -1/L_2C_2 = \omega_{\infty 1}^2 = 1.3978 = 0.8965p$$

or  $L_1 = 0.8965$

$$Y_2 = \frac{1}{z_{11} - L_1p} - Yr = \frac{0.8179p}{1 + 0.5115p^2}$$

or  $C_2 = 0.8179$  and  $L_2 = 0.6258$

Also:  $L_3 = 1.6857$ ;  $C_4 = 1.140$ ;  $L_4 = 0.2104$ ;  $L_5 = 1.1742$

Terminal resistor  $R_2$  is equal to  $R_1$ , because  $z_{11} = z_{22}$ .

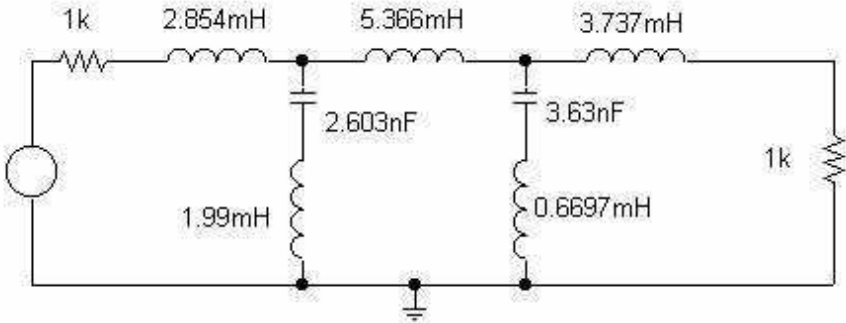
By taking a value of  $1 \text{ k}\Omega$  for these resistors, as unitary values:

$$C_u = \frac{1}{2\pi f_u R_u} = 3.183 \text{ nF} \quad \text{and} \quad L_u = \frac{R_u}{2\pi f_u} = 3.183 \text{ mH}$$

Eventually, the schema for creation shown in Figure 5.9, where we see that real component values have been obtained by multiplying the normalized values by the unitary values.

#### 5.4.2.4. Direct digital synthesis

When using numerical optimization software, calculating the digital value of elements can be directly carried out, with the condition that several precautions must be taken. Actually, the weak sensitivity to value variations of L-C filters brings about, in mathematical terms, very low partial derivation values. Numerical methods are almost always based on calculations of these derivations. The algorithm can easily converge on a local minimal more or less removed from ideal values; or these may not converge at all.



**Figure 5.9.** Schema of a band pass filter calculated by using Example 2.  
The resistors are in Ohms

For polynomial filters of order  $n$ , the variable is the vector of  $n$  values of inductors and capacitors. We take as initial values the unity and we carry out the calculation from a least mean square criterion by about 100 frequency values distributed between the band pass and the stopband. The attenuation is expressed in dB.

For elliptic and inverse Tchebycheff filters, the variable is also a vector of  $n$  values of capacitors and inductors; that is, a value by branch. For branches with a tuned circuit, the value of the second element is calculated based on knowing transmission zeros  $\omega_{\infty i}$  of the transfer function, which means of the zeros of  $P(p)$ :

$$L_j C_j \omega_{\infty i}^2 = 1$$

This procedure is shown by the direct numerical calculation of the two same filters that were previously calculated by the analytic method, using Matlab software and its “optimization toolbox”:

Step 1: we have the initial values of the components. The unity value is always a good choice for  $L$  and  $C$  elements. The generator resistor is taken as equal to unity. The terminal resistor is calculated as shown above.

Step 2: we choose  $m$  pulsation values, for which we calculate the filter attenuation. Between 100 and 1,000 values included between normalized pulsations 0 and 5 constitute a correct order of magnitude.

Step 3: for these  $m$  values, we calculate the filter attenuation with the help of a function written for this purpose.

Step 4: for these  $m$  frequency values, we calculate the ideal attenuation from the transfer function.

Step 5: we calculate the distance between these two series of values with a quadratic criterion.

Step 6: we minimize this distance by using a universal optimization function made by Matlab. Here, we use the “constr” function which allows us to introduce a positivity constraint on the element values. This helps us to avoid arriving at a solution that cannot be put into practical use.

The calculation takes only several dozen seconds working with a Pentium III and much less time at a workstation. We usually get a correct convergence. We see that if the calculation is not carried out very precisely, we can get a value set quite different from the values obtained by analytic calculation, even if the filter has a satisfactory response. This fact proves the weak sensitivity of  $L$ - $C$  filters, since equivalent results can be obtained with quite a different value set.

For an elliptic filter, the procedure is almost exactly the same, with one difference. It is necessary in this case to calculate the values of the second resonant branch elements from the initial values.

*Example 1*

The same Tchebycheff filter of order 6 calculated by this method gives, after 1,800 iterations, the following normalized values (without unities):

$$\begin{array}{lll} L_1 = 2.1467 & C_2 = 1.1113 & L_3 = 3.0182 \\ C_4 = 1.1690 & L_5 = 2.9180 & C_6 = 0.8130 \end{array}$$

We notice that the values are more or less different from those calculated by Darlington's exact method, even though the precision required by the algorithm of the calculation was very high. However, the difference between the response curves is undetectable.

*Example 2*

The elliptic filter of order 5 calculated by the direct numerical method gives the following normalized values (without unities) after 800 iterations:

$$\begin{array}{llll} L1 = 0.7407 & C2 = 0.8750 & L2 = 0.5295 & L3 = 1.5832 \\ C4 = 1.1855 & L4 = 0.1787 & L5 = 1.0338 & \end{array}$$

Even if the values are quite different from the exact values, the response curves differ by less than a thousandth of a dB.

**5.4.3. L-C filters derived from a pass band**

When the filter is not a band pass, we first calculate the transfer function of the corresponding low pass prototype. Then, we carry out the synthesis of this prototype. The last step is deducing the final schema by applying the frequential transformations shown in section 5.3.3 to the impedance values of the low pass schema. In this way we get the schema of the transformed filter, with element values expressed in real unities. This transformation is simple to effect.

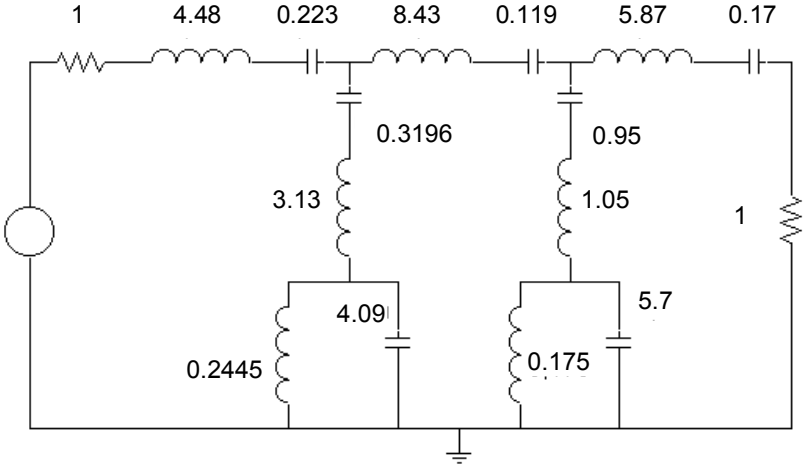
*Example*

Suppose we want to produce an elliptic band pass filter of order 10 with  $A_{\max} = 0.2$  dB and  $A_{\min} = 40$  dB. Its central frequency is of 100 kHz and its band width  $B = 20\%$ , that is, 20 kHz.

The low pass prototype of this filter is exactly the filter calculated in Example 2 of the above section (see Figure 5.9). Transformations of low pass impedances  $\rightarrow$  band pass are seen in Figure 5.10. Each impedance is converted into a capacitor in



parallel with an impedance. In this schema, the elements appear with normalized values.

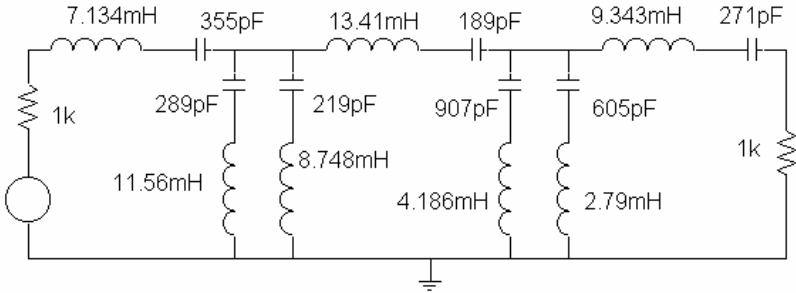


**Figure 5.10.** Schema of a band pass filter obtained by conversion of the low pass prototype shown in Figure 5.9. The elements are in normalized values (without unities)

**5.4.4. Conversions of L-C filters; optimization**

The schemata obtained by the method just described are not always compatible with the technological constraints presented by capacitors and, most importantly, inductors. We can then obtain schemata derived from network transformations.

Here, there are many possibilities, such as the Norton transformation and the Star/Triangle conversion. Implementing them is a delicate task and needs to be done by an engineer. These techniques are described in [HAS 81]. We present here one very common example of conversion, universally used in creating band pass filters. This consists of replacing the four elements of parallel branches with two resonant circuits in series. We then have a schema similar to that shown in Figure 5.11, but less sensitive to disturbances and with values that are less dispersed. In this schema, the values of elements are in real unities, by taking a value of 1 kΩ as the generator resistor.



**Figure 5.11.** Schema expressed in real values obtained after transformation of the schema in Figure 5.10. The resistors are in Ohms

### Optimization

The calculations presented above are carried out by supposing that the capacitors and resistors have no losses. In reality, even if this hypothesis is generally valid for capacitors, it is not as valid for inductors whose coefficients of quality  $Q$  are limited. The response curves diverge so much from ideal curves that these losses become significant.

It is possible to partly remedy these imperfections by modifying the element values by using a numerical optimization tool such as the “optimization toolbox” made by Matlab. However, losses can be effectively compensated only if component quality is sufficiently raised. There must be values of  $Q$  of the order of about 100 to produce high performance L-C filters.

## 5.5. Active filters

Active filters are made of capacitors, resistors and active elements (almost always operational amplifiers). Less bulky, easier to produce, and thus less costly, active filters are useful when frequencies are not too high (typically up to several MHz). Nevertheless, active components introduce noise, limiting the maximum voltage that can be filtered, and requiring a power supply.

Active filters are usually produced by putting basic second order cells in cascade, as shown above. Simple and very widely used, this method does have the limitation of producing filters that are very sensitive to imprecisions or variations in component values. To avoid this problem, we also make active filter copies of L-C filters, following several very good methods. The following sections will present these different approaches.

### 5.5.1. Second order or biquadratic cells

RC circuits of varying complexity can be linked successfully to one to four operational amplifiers. In this chapter, we present increasingly complex applications, beginning with cells linked to only one operational amplifier and a minimal number of capacities. Presenting more complex applications at each level, only the most current constructions and their main results will be presented and analyzed. The most complex configurations will be discussed only if they offer some practical advantage.

### 5.5.2. Biquadratic cells with one operational amplifier

Their general schema is given in Figure 5.12. Quadrupoles  $(RC)_1$  and  $(RC)_2$  are respectively inserted into a positive and negative reaction loop. In order for this to be cost-effective, quadrupoles  $(RC)_1$  and  $(RC)_2$  must not have more than a minimal number of capacities, at most two or sometimes three. From there, we have three circuit families, each one having approximately the same qualities.

#### Negative reaction biquadratic cells

The positive input of the amplifier is linked to the mass (if there is no  $RC_1$  quadrupole). The most widely used cell of this type is the Rauch band pass cell (see Figure 5.13). Its transfer function is given by the following relation:

$$\frac{V_2}{V_1} = \frac{-R_2 C p}{R_1 R_2 C^2 p^2 + 2R_1 C p + 1}$$

$$Q = \frac{1}{2} \sqrt{\frac{R_2}{R_1}} \quad \text{and} \quad \omega_o = \frac{1}{C \sqrt{R_1 R_2}}$$

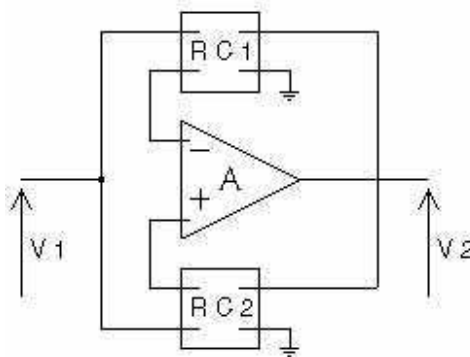
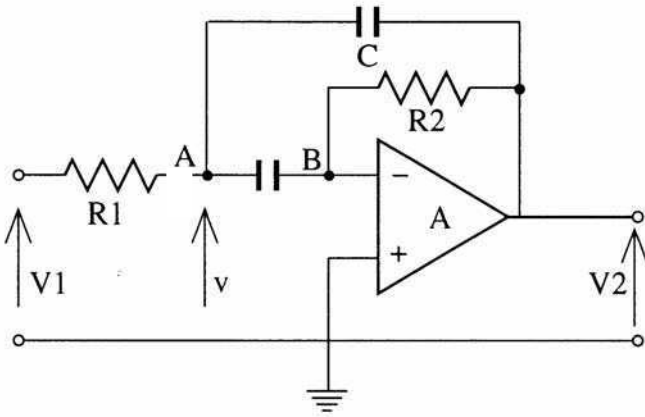


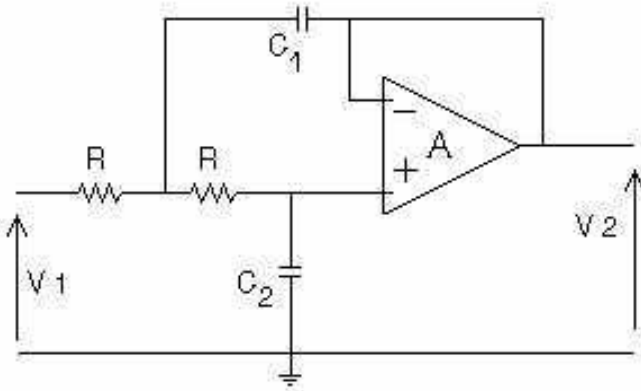
Figure 5.12. General biquadratic cell with one amplifier



**Figure 5.13.** Biquadratic band pass cell in negative reaction (Rauch cell)

#### *Biquadratic cells in positive reaction*

With these types of circuits, the operational amplifier is assembled as a controlled source; that is, as a constant gain amplifier, usually equal to or near unity. Low pass and high pass cells of this type, called Sallen-Key cells, are very widely used (Figure 5.14).



**Figure 5.14.** A Sallen-Key low pass cell

The transfer function of the low pass cell is:

$$\frac{V_2}{V_1} = \frac{K}{R^2 C_1 C_2 p^2 + R p [2C_2 + (1-K) C_1] + 1}$$

$$Q = \frac{\sqrt{C_1 C_2}}{2C_2 + (1-K) C_1}$$

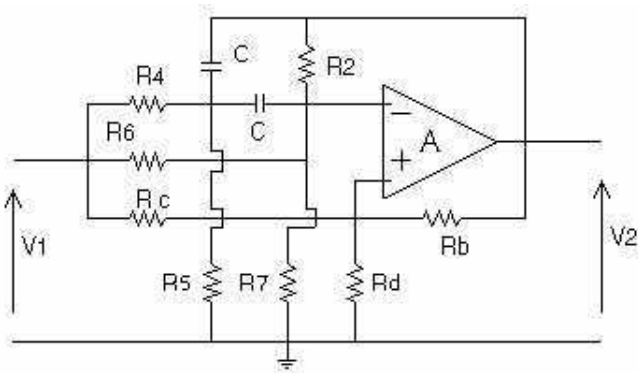
$$\omega_o = \frac{1}{R\sqrt{C_1 C_2}} \quad \text{with } K \approx 1$$

*Biquadratic cells in mixed reaction*

The two types of cells described above are simple but specific: they have no transmission zeros. By combining positive and negative reactions, we can obtain all-purpose biquadratic cells. However, these cells are much more complex. The most widely used cell of this kind is the Friend cell (see Figure 5.15). Its transfer function is very complex and the use of an appropriate software program is indispensable for calculating a cell of this type, otherwise the cells may not be viable. However, the high number of components does allow for a very flexible design.

The major advantages of these simple structures are their low number of components and reduced energy use. But there are many disadvantages. Some are:

- difficult adjustments and settings (not independent for Q and  $\omega_0$ );
- high dispersion of component values (proportional to  $Q^2$ );
- high active sensitivities (proportional to  $Q^2$ );
- structures specific to a response type.



**Figure 5.15.** General schema of a Friend cell

These disadvantages can be minimized by using more complex structures. These structures should have dispersions and active sensitivities proportional to  $Q$  and, especially, universal schemata. Because of this last characteristic, cells with three or even four operational amplifiers are mass produced and found in all manufacturers' catalogues. Their relatively high cost and required use of more amplifiers is compensated by their wide availability and by the advantages of their simple design and adjustments.

### 5.5.3. Universal biquadratic cells with three or four amplifiers

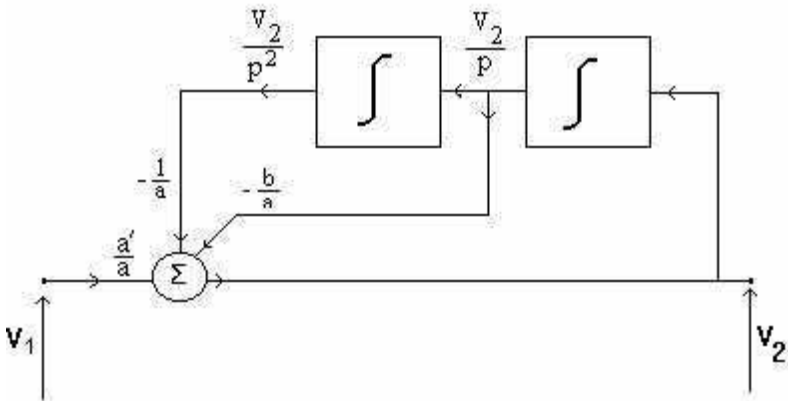
Manufacturers have presented several schemata for these cells. All are based on the theory of state variables. According to this theory, it is always possible to decompose a transfer function of the order  $n$  in an ensemble of  $n$  functions of the first order that have been simulated by integrators and combined with adders-subtractors. As an example, let us look at the high pass function of the second order with a transfer function as follows:

$$\frac{V_2}{V_1} = \frac{a'p^2}{ap^2 + bp + 1}$$

This equation can be written as:

$$V_2 = \frac{d}{a} V_1 - \frac{b}{a} \frac{V_2}{p} - \frac{V_2}{ap^2}$$

This equation can be carried out by the analog circuit shown in Figure 5.16, which has two integrators and an adder.



**Figure 5.16.** Principle of producing a biquadratic state variable high pass transfer function

Several schemata derived from this principle have been developed. We will present two of them, both of which are widely used and important.

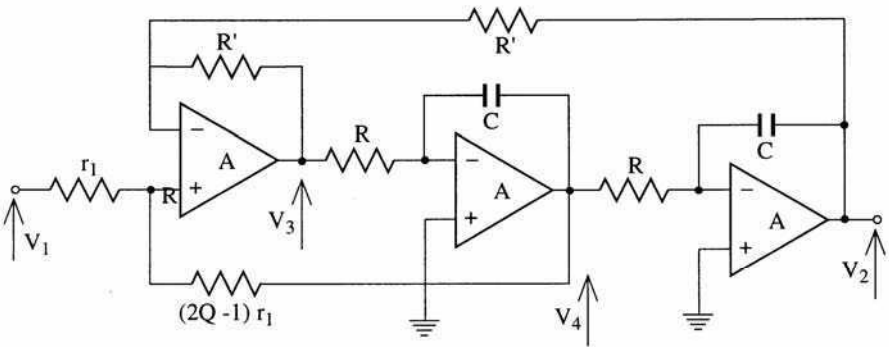
*The Kerwin, Huelsman and Newcombe cell (KHN)*

This cell was the first to be developed. It has been the basis for later cells. It is produced by rearranging the signs of Figure 5.16 by using only three amplifiers (see Figure 5.17). The calculation is carried out from the adder-subtractor forming the first operational amplifier. We get the following transfer function:

$$\frac{V_3}{V_e} = \frac{2Q-1}{Q} \frac{R^2 C^2 p^2}{R^2 C^2 p^2 + \frac{RCp}{Q} + 1} \quad \text{with} \quad RC = \frac{1}{\omega_0}$$

The KHN cell has the remarkable characteristic of simultaneously presenting, on the same circuit, a low pass transfer function (between  $V_1$  and  $V_2$ ) high pass (between  $V_1$  and  $V_3$ ) and pass band (between  $V_1$  and  $V_4$ ). This quality is important; it allows for easy production standardization. In addition, the gain and the quality coefficient  $Q$  can be independently adjusted by  $r_1$  and  $R$ .

These cells are sold by several manufacturers under the name “Universal Filters”. A fourth amplifier is also sold along with these cells, allowing the user to adjust transmission zeros using a technique we will discuss below.



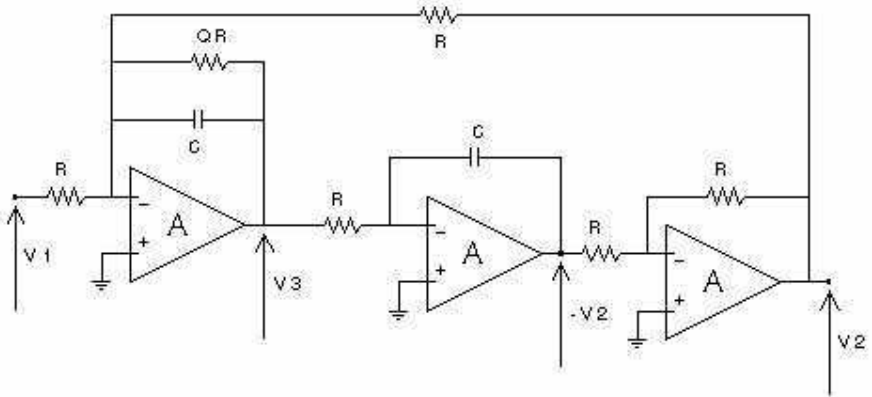
**Figure 5.17.** KHN biquadratic state variable cell

*Tow-Thomas cell*

Because of a small modification generalized in the Tow-Thomas cell (Figure 5.18), amplifiers have their positive input linked to the mass. By supposing three

amplifiers, whose gain in open loop is  $A \gg 1$ , to be identical, the transfer function of this kind of cell is written:

$$\frac{V_2}{V_1} = \frac{1}{R^2 C^2 p^2 + \frac{RC}{Q} p + 1} \quad \text{with} \quad \omega_o \approx \frac{1}{RC}$$



**Figure 5.18.** Tow-Thomas biquadratic state variable cell

We simultaneously get a band pass output at point  $V_3$ . This circuit has excellent qualities, low passive and active sensitivities, and gives the possibility of producing circuits with very high quality coefficients  $Q$ .

*Active KHN universal filter*

KHN cells simultaneously have three outputs: band pass, high pass and low pass. In adding these three outputs weighted by factors  $a'$ ,  $b'$ , and  $c'$ , we can obtain any biquadratic function (see Figure 5.19):

$$\frac{a'p^2}{ap^2 + bp + 1} + \frac{b'p}{ap^2 + bp + 1} + \frac{c'}{ap^2 + bp + 1} = \frac{a'p^2 + b'p + c'}{ap^2 + bp + 1}$$

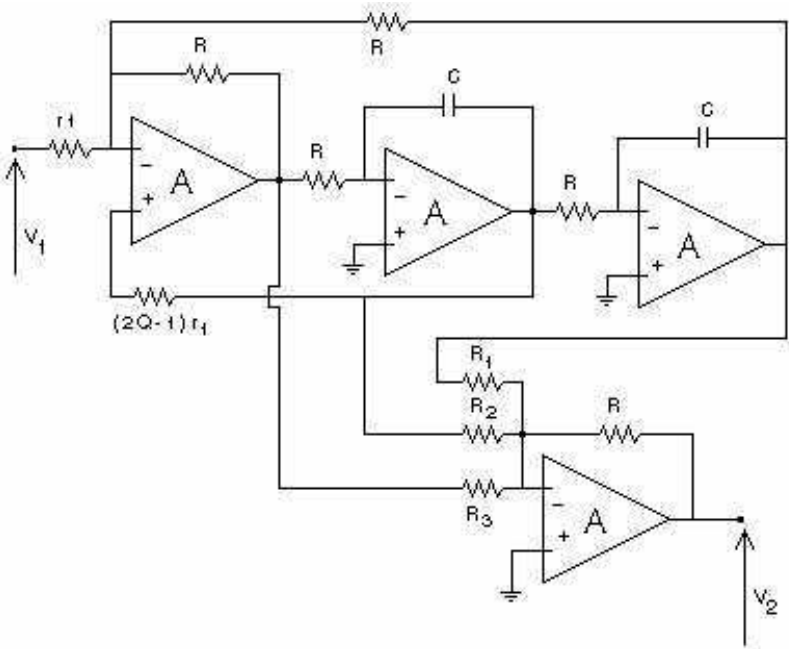
This is why component manufacturers sell KHN cells with an amplifier, which makes summation possible. These cells are very easy to use, in addition to being adaptable and affordable.



*Universal Fleisher-Tow filter*

Another technique, called feed forward, helps obtain universal biquadratic cells. It requires just three amplifiers. This cell is called the Fleischer-Tow cell. The schema of its principle is given in Figure 5.20, if the desired transfer function is:

$$\frac{V_2}{V_1} = \frac{a' p^2 + b' p + c}{ap^2 + bp + 1}$$



**Figure 5.19.** KHN biquadratic universal cell obtained by weighted summation of band pass, high pass and low pass outputs

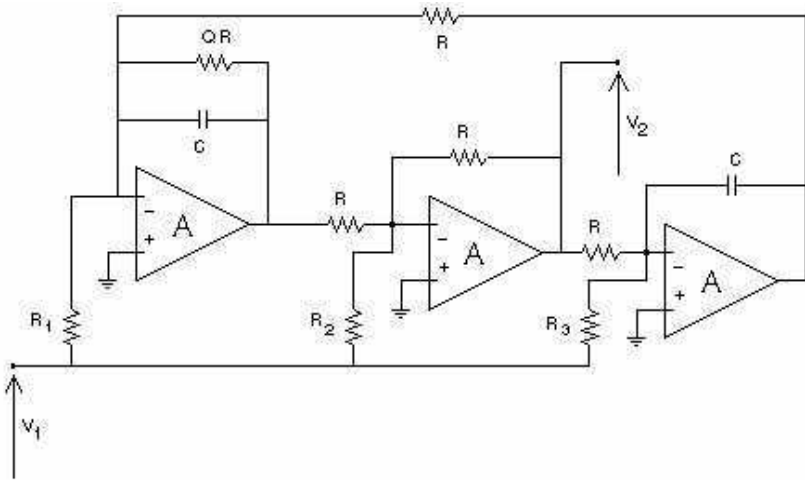
The normalized values of the schema elements are, if we suppose that  $C = 1$ :

$$R = \sqrt{a} \quad R_1 = \frac{a}{aa' - b'} \quad R_2 = \frac{\sqrt{a}}{a'b} \quad R_3 = \frac{\sqrt{a}}{c'} \quad Q = \frac{\sqrt{a}}{b}$$

### 5.5.4. Elevated order active filters (elevated by putting biquadratic cells in cascade)

The principle of these cells has been shown in section 5.3.4 (Figure 5.4). In order to minimize noise and maximize the dynamic, three precautions must be taken:

- we must ascertain that the order used to put the cells in cascade are such that the cells with high overvoltage coefficients are placed near output. This preserves the filter dynamic;
- for filters with transmission zeros, we should ensure that the poles and zeros of biquadratic functions are optimally matched. Practically speaking, we must often be satisfied with linking the closest pair of zeros to a pair of poles;
- we must ensure that the gain repartitions of the different biquadratic cells preserves the ensemble dynamic. A good way to do this is to equalize all cell responses.



**Figure 5.20.** Universal Fleischer-Tow cell with three amplifiers

It is important to note that this type of synthesis is limited to certain filters. These are:

- low pass and high pass filters with a degree not exceeding eight to ten;
- band pass filters with a band width not much below 20%.

If these limitations are not considered, the synthesis may produce excess sensitivity to component value variations.

### 5.5.5. *Simulating an L-C filter*

Orchard has shown that L-C filters inserted between two resistors have the best possible sensitivity. This means that an active filter with low sensitivity is produced by “copying” an L-C filter, which is used as a model. Among the solutions proposed for this, three methods yield a workable way of doing this, and are easy to implement. All three techniques produce a filter with a sensitivity considerably lower than their homologues using the cascade technique.

#### *Copying by simulation of inductors using gyrators and capacitors*

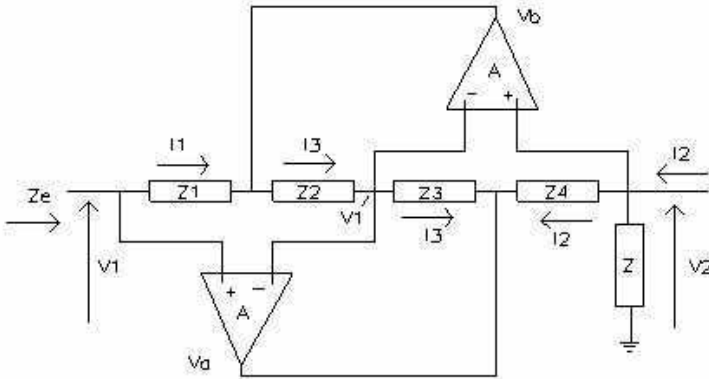
A gyrator is an electronic ensemble that converts a capacitor impedance to an inductor impedance. At the time of writing, the best gyrator is made with a generalized impedance converter, called a GIC (Figure 5.21). This mechanism is only useful for inductors that have a point linked to the mass. All high pass filters and even band pass filters satisfy this constraint, making possible adequate conversions. The synthesis method based on the use of gyrators (see Figure 5.22) is excellent for producing filters with very narrow bands, up to a few hundred kHz.

#### *FDNR-based simulation*

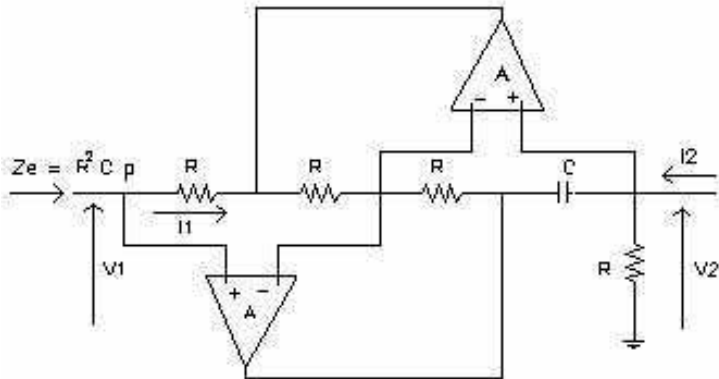
Another method for simulating L-C filters is based on an impedance conversion called the Bruton transformation. This transformation suppresses the inductors while activating the Frequency Dependent Negative Resistor (FDNR). This element is an electronic ensemble that converts the capacitor impedance into a negative resistor whose value is inversely proportional to the square of the frequency. The best FDNR schema is obtained with the help of the GIC shown in Figure 5.21, in which  $Z_1$  and  $Z$  are capacitors of value  $C$  and the other impedances are resistors of value  $R$ . The

input impedance of this mechanism is then  $Z_e = \frac{-1}{RC^2\omega^2}$ . Here, we are describing

a negative resistor dependent on the frequency.



**Figure 5.21.** Generalized impedance converter (GIC):  $Z_e = \frac{V_1}{I_1} = \frac{Z_1 Z_3}{Z_2 Z_4} Z$



**Figure 5.22.** Inductor simulation by a GIC:  $L = R^2 C$

The Bruton transformation involves multiplying all the impedances of an L-C filter by  $1/p$ . This leads to the following steps:

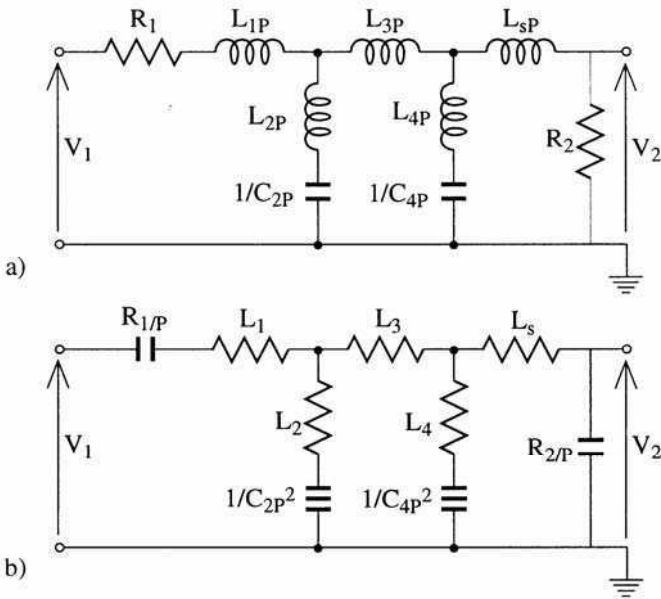
- inductors are converted into resistors;
- capacities are converted into FDNR;
- terminal resistors are changed into capacities.

The overall response is not affected by this transformation. This method works particularly well with low pass filters (see Figure 5.23), if we begin with a schema in T in which resonant circuits are serial circuits, in parallel branches. As with

simulating inductors by gyrators, it is not easy to produce floating FDNRs. We must also look for the L-C configuration that reduces their number. A combination of Bruton transformations and a gyrator can often help us avoid the use of floating FDNRs and gyrators.

*Operational simulation of an L-C filter*

A third method of copying L-C filters is to produce an electronic circuit that has no inductors but is responsive to the same differential equations. This method is widely used to make integrated filters with switched capacitors. This method will be presented in the next section.



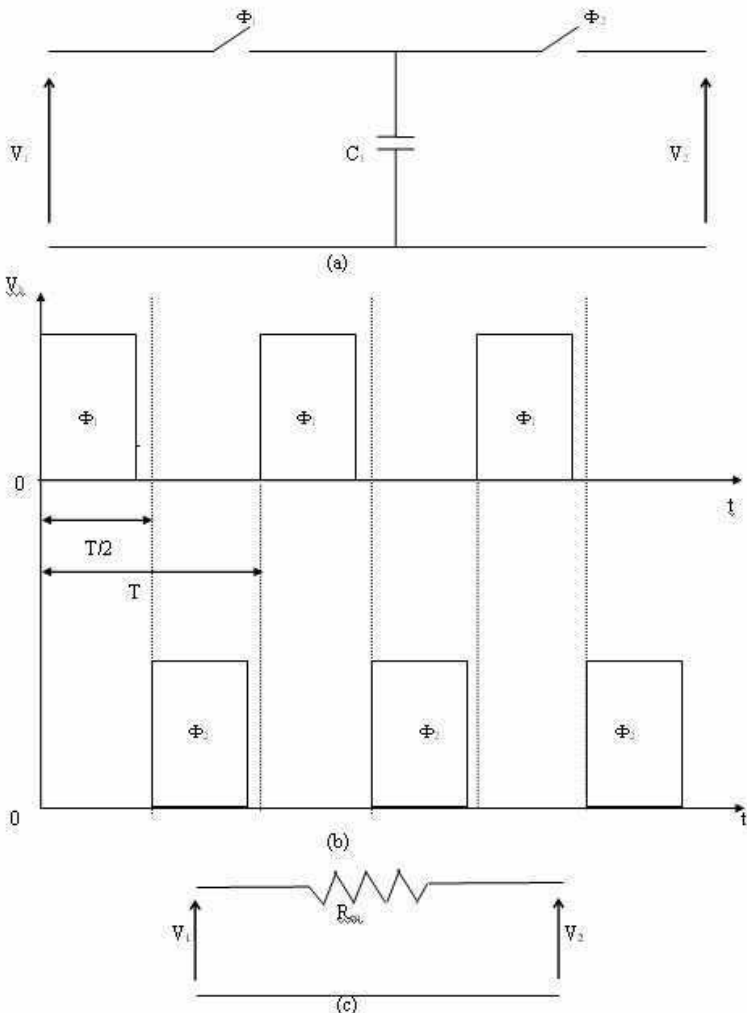
**Figure 5.23.** Low pass L-C filter (a) and a filter converted by Bruton's method (b) producing two FDNRs

**5.6. Switched capacitor filters**

Producing a filter requires very precise and stable components. This condition is incompatible with the technological constraints of integrated circuits. An ingenious and very useful device invented by Friend in 1972 [FRI 72] avoided this problem and produced excellent filters that were completely integrated and required no adjustments. Unfortunately, at the time of writing this type of filter can only be produced at frequencies lower than a few MHz.

The basic assembly, shown in Figure 5.42, consists of replacing the resistors of an active filter with an assembly that only contains capacitors and interruptors that alternately open and close to the rhythm of a clock of period  $T = 1/f_h$ . Each  $T$  period decomposes in two non-overlapping phases  $\Phi_1$  and  $\Phi_2$ . If this mechanism is connected to two voltage sources  $V_1$  and  $V_2$ , we can write that a charge  $\Delta Q$  is transferred from the output source at each period:  $\Delta Q = C_1(V_2 - V_1)$ .

During time  $t \gg T$ , the transferred charge and the mean current are:



**Figure 5.24.** Principle of switched capacitor circuits: a) basic schema; b) control voltages of interruptors 1 and 2; c) equivalent schema

$$Q = \frac{t}{T} C_1 (V_2 - V_1)$$

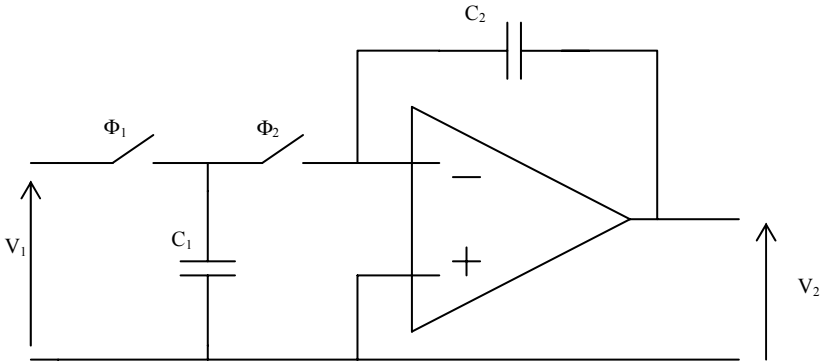
$$I_{\text{mean}} = \frac{Q}{t} = \frac{C_1}{T} (V_2 - V_1) = f_h C_1 (V_2 - V_1)$$

In terms of the transferred charge, the circuit becomes equivalent to a resistor:

$$R_{\text{equivalent}} = \frac{V_2 - V_1}{I_{\text{mean}}} = \frac{1}{C_1 \cdot f_h} = \frac{T}{C_1}$$

This resistor of value  $T/C_1$  can be easily integrated, since MOS transistors can be excellent analog interruptors. This mechanism becomes extremely useful when it is used as the resistor of an analog integrator (see Figure 5.25). The corresponding transfer function is:

$$\frac{V_2}{V_1} = \frac{-1}{R_{eq} C_2 P} = -\frac{C_1}{T C_2}$$



**Figure 5.25.** *Switched capacitor integrators*

The time constant of this integrator only depends on the relation of capacities  $C_1$  and  $C_2$ , and not on their individual value. If these elements are produced on the same substratum, this relation depends solely on the relation of their physical surfaces, which can be established by construction, with excellent precision (of the order of 0.1%).

Since it is possible to make active filters that use only integrators (the KHN cell is one example), we can create very precise integrated filters without carrying out adjustments. Another advantage is that the time constant depends on the clock period  $T$ , so we can modify cut-off frequencies by digitally programming the clock. In practice, this is a very helpful property.

### 5.6.1. Integrators without sensitivity to stray capacitances

As such, the basic assembly cannot be used in an integrated filter. Actually, the values of  $C_1$  and  $C_2$  must be very low (lower by several pF) in order to conserve the silicon surface. This means that it is of crucial importance that inevitable stray capacitances, which have values that are difficult to control, should not influence the charge transfer. We can achieve this by using an integrator with four interruptors (see Figure 5.26). This assembly is insensitive to the principles of the stray capacitances of MOS interruptors. It can be used with two phasings of separate clocks. The following steps describe two different phasings:

- interruptors 1 and 4, as well as 2 and 3, are simultaneously activated. The transfer function is that calculated previously. We will call this a “type 1” integrator;
- interruptors 1 and 3, as well as 2 and 4, are simultaneously activated. The sign of the transfer function is reversed, because capacitor  $C_1$  is re-set before ceding its charge to  $C_2$ :

$$\frac{V_2}{V_1} = \frac{1}{R_{eq} C_2 P} = \frac{C_1}{TC_2}$$

This is called a “type 2” integrator.

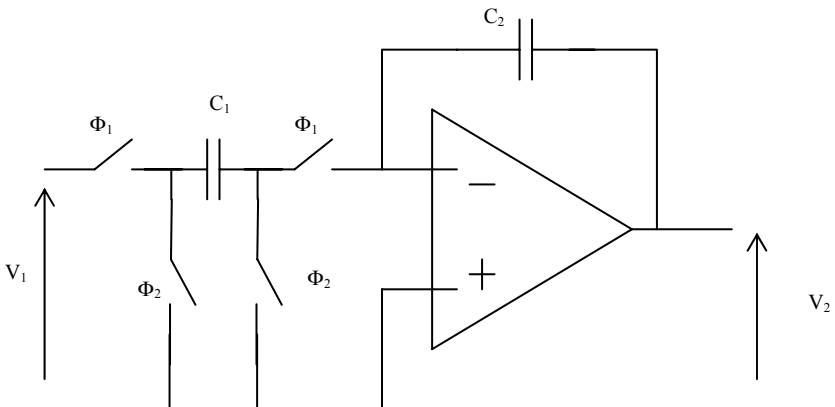
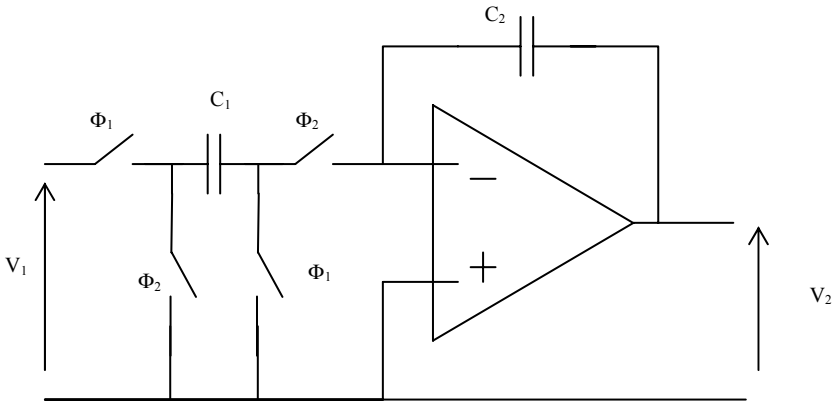


Figure 5.26a. Type 1 integrator (reverser)



This dual possibility facilitates filter synthesis by avoiding the use of inverters. We will see that, strictly speaking, these two circuits do not share the same transfer function in  $z$ , which is an added advantage.



**Figure 5.26b.** Type 2 integrator (non-reverser)

As a standard analog electronic mechanism, it is possible to place several branches with switched capacitors at the input of an integrator. In this way, we have an adder or subtractor with multiple inputs. This is why these two schemata are the bases for creating all integrated filters with switched capacitors.

### 5.6.2. Analysis of switched capacitor integrators

Equivalence between a switched capacitor resistor and an ohmic resistor is only approximate because this equivalence has been established by assuming a very high clock frequency in view of the filtered signal frequency. A more precise analysis comes up against a fundamental problem: a circuit with switched capacitors is not time invariant and therefore cannot be analyzed by using traditional methods. In particular, the concept of transfer function does not really apply to these circuits.

Attempting to overcome this problem, we see that, aside from input nodes, currents are zero and voltages are completely stationary, except for moments of interruptor switching. If we restrict our observation of circuits to instants that immediately follow these transitions, we can write equations from divergences that create the charge conservation between two consecutive clock periods. From this we can deduce a transfer function in  $z$ , allowing for a rigorous analysis of the circuits. For this method to be effective, we must block the input voltage on each  $T$  period in order to make it stationary as well.

We illustrate this method by calculating the transfer function of the type 1 integrator (Figure 5.26a) that has an input blocked between two T instants. If we estimate the charge transferred to capacitor  $C_2$ , between two consecutive instants  $(n-1)T$  and  $nT$ , by supposing perfect interruptors and amplifiers (the charge transfers then being instant), we get:

$$C_2 V_2(nT) - C_2 V_2(n-1)T = -C_1 V_1(nT)$$

In taking  $nT$  as the reference instant, we deduce the linked transfer function in  $z$ :

$$C_2 V_2(1-z^{-1}) = -C_1 V_1 \quad \text{or:} \quad \frac{V_2}{V_1} = -\frac{C_1}{C_2} \frac{1}{1-z^{-1}}$$

For a type 2 integrator, a charge is accumulated by  $C_1$  before transfer into  $C_2$ . Taking into account this delay, the inversion of  $C_1$  and the blocked input, the equation for the charge transfer becomes:

$$C_2 V_2(nT) - C_2 V_2(n-1)T = C_1 V_1(n-1)T$$

$$C_2 V_2(1-z^{-1}) = C_1 V_1 z^{-1} \quad \text{or:} \quad \frac{V_2}{V_1} = \frac{C_1}{C_2} \frac{z^{-1}}{1-z^{-1}}$$

This is a transfer function of a non-inverting integrator

### 5.6.3. Synthesis of switched capacitor filters

The technology of switched capacitors plays only a practical part in making integrated filters. The following three application categories have been developed:

- switched capacitor biquadratic cells, available in standard packaging have been developed. These are parametrizable by external resistors. They have a great advantage over the usual equivalent analog devices of being programmable in cut-off frequency by clock frequency;

- manufacturers have developed filters of all types in the form of catalog circuits. This is possible because cut-off frequencies can be adjusted and set by the clock;

- specific circuits have been developed by manufacturers of analog circuits. They also offer a wide range of switched capacitor filters, as well as software necessary to carry out the correct synthesis. This helps us make a filter or include a filtering function in a more complex integrated circuit.

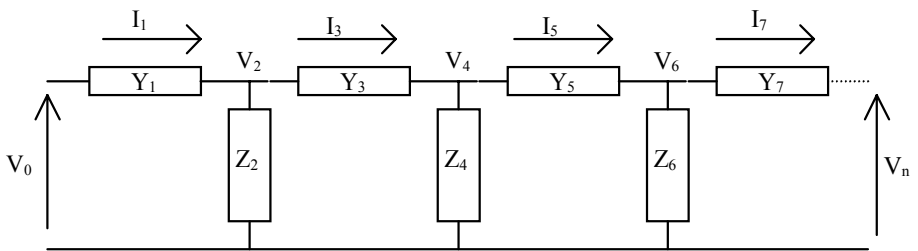
Synthesis of biquadratic cells is done only from state variable cells. Only these cells use nothing but adder-subtractor integrators. As for the last two applications, they use all the operational simulations of L-C filters, in order to take advantage of the low sensitivity of these filters. This method will be described in the next section.

**5.6.4. Operational simulation of an L-C filter (leapfrog simulation)**

It is actually impossible to create a gyrator and an FDNR only with integrators. This is why we use another technique to copy L-C filters: it is an operational simulation called leapfrog. It consists of simulating differential equations that govern the L-C circuit by means of an ensemble of integrators and adder-subtractors. This procedure is similar to that used for creating state variable biquadratic cells. Its principle is illustrated by band pass filter synthesis.

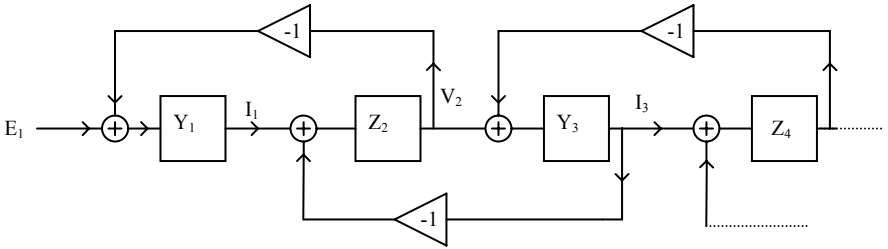
An L-C filter is a ladder filter with impedances of branches in parallel. These are noted as  $Z_i$ , with the admittances of serial branches  $Y_j$  (Figure 5.27). We can express recurrently the node voltages and branch currents as follows:

$$\begin{aligned}
 I_1 &= (V_0 - V_2) Y_1 \\
 V_2 &= (I_1 - I_3) Z_2 \\
 I_3 &= (V_3 - V_4) Y_3 \\
 V_4 &= (I_3 - I_5) Z_4 \\
 \dots &= \dots\dots\dots
 \end{aligned}$$



**Figure 5.27.** Structure of an L-C ladder filter

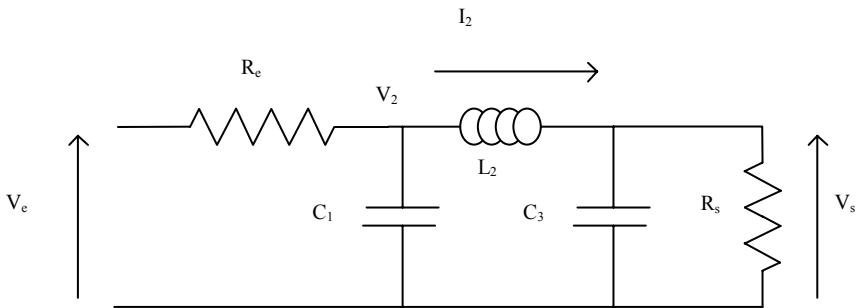
These equations can be symbolized by a graph in leapfrog form, as in Figure 5.28. For a polynomial low pass L-C filter, impedances  $Z_i$  and admittances  $Y_j$  have the respective values of  $1/C_i p$  and  $1/L_j p$ . They can be produced by switched capacitor filters.



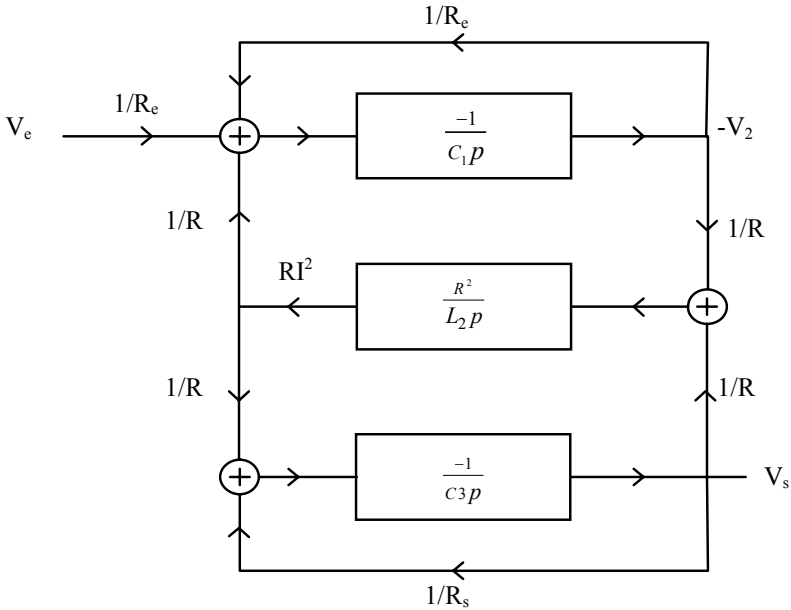
**Figure 5.28.** Leapfrog operational graph

To illustrate this method, we have to use a third order filter, knowing that the process is iterative (Figure 5.29). The iterative equations are modified in order to introduce terminal resistors and to avoid voltages (we multiply the intensities by a scaling resistor of arbitrary value R). We then introduce the appropriate signs for the integrations to end up with a simple schema. We get the following equations:

$$\begin{aligned}
 -V_2 &= \frac{-1}{C_1 p} \left( \frac{V_e - V_2}{R_e} - I_2 \right) \\
 -RI_2 &= \frac{+R^2}{L_2 p} \frac{(V_s - V_2)}{R} \\
 V_s &= \frac{-1}{C_3 p} \left( \frac{V_s}{R_s} - I_2 \right)
 \end{aligned}$$



**Figure 5.29.** Third order polynomial L-C filter



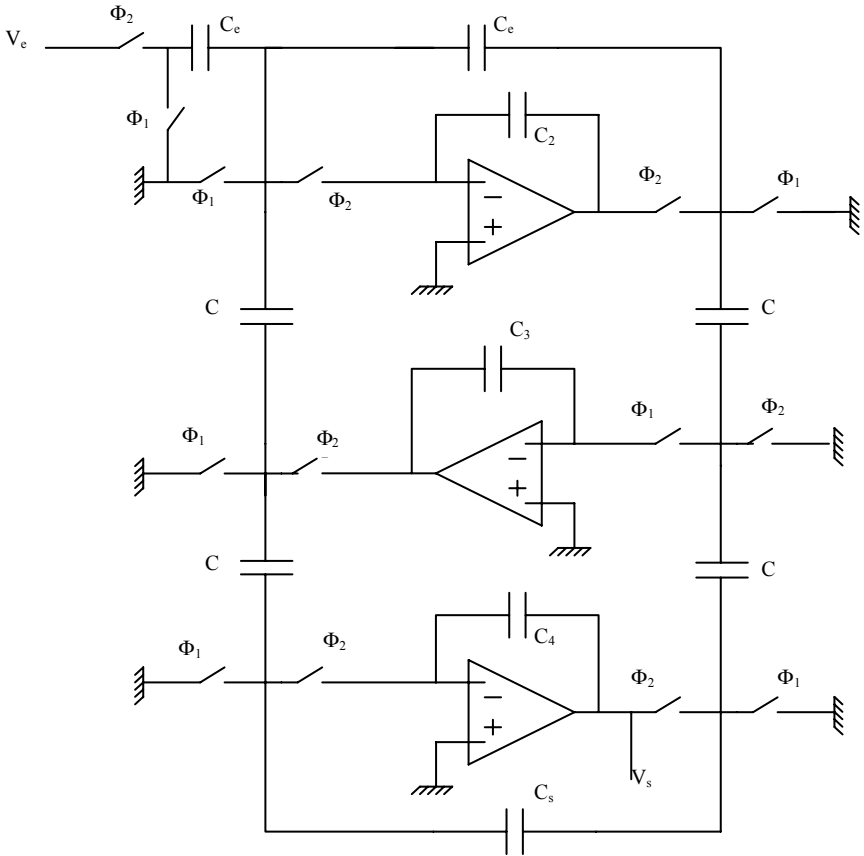
**Figure 5.30.** Operational graph of the L-C filter shown in Figure 5.29

This set of equations is symbolized by the graph shown in Figure 5.30, where there are only integrators and adder-subtractors. Integrators 1 and 3 are type 1 and the central integrator is type 2. The ensemble can be created entirely with switched capacitor integrators. We see that many interruptors have two uses. After suppressing the unnecessary elements, we get the final schema, shown in Figure 5.31.

In this assembly, capacitor values  $C_2$  and  $C_4$  are the same as those in the first schema. The other capacitor values are established after choosing a clock frequency of period  $T$  and a resistor of arbitrary value  $R$ :

$$C = \frac{T}{R} \qquad C_e = \frac{T}{R_e} \qquad C_s = \frac{T}{R_s} \qquad C_3 = \frac{L_2}{R^2}$$

A careful analysis of the assembly, made with the transfer functions in  $z$  introduced before, shows that alternating type 1 and 2 integrators suppresses the imperfections (losses) coming from the discretization of the resistors. We speak of the “exact synthesis” of an L-C filter for this kind of operation, which is notable for its precision and stability.



**Figure 5.31.** Switched capacitor filter deduced from graph in Figure 5.30

### 5.6.5. Switched capacitor biquadratic cells

All manufacturers of analog integrated circuits offer universal cells with switched capacitors. These are usually parametrizable by external resistors, but this can also be done by setting cut-off frequencies to clock frequencies. All these circuits are variants of variable state cells, but they only have two operational amplifiers because they have inverting and non-inverting integrators.

To show this, Figure 5.32 presents a schema of this type of cell. The interruptors shown there are in another form which is also widely used. The reference phase is

that which corresponds to the position of the interruptors on the schema. In the next phase, all interruptors switch. The transfer function of this cell is:

$$\frac{V_2}{V_1} = \frac{DI + z^{-1}(AG - DJ - DI) + z^{-2}DJ}{BD + z^{-1}(AC + AE - 2BD) + z^{-2}(BD - AE)}$$

These switched capacitor filters perform very well from the point of view of flexibility and complete integration possibilities. However, we should remember that their applications are limited in frequency to several hundred kHz. In addition, the filters have been sampled and are subject to aliasing. This means they must have a continuous filter that eliminates this unwanted phenomenon.

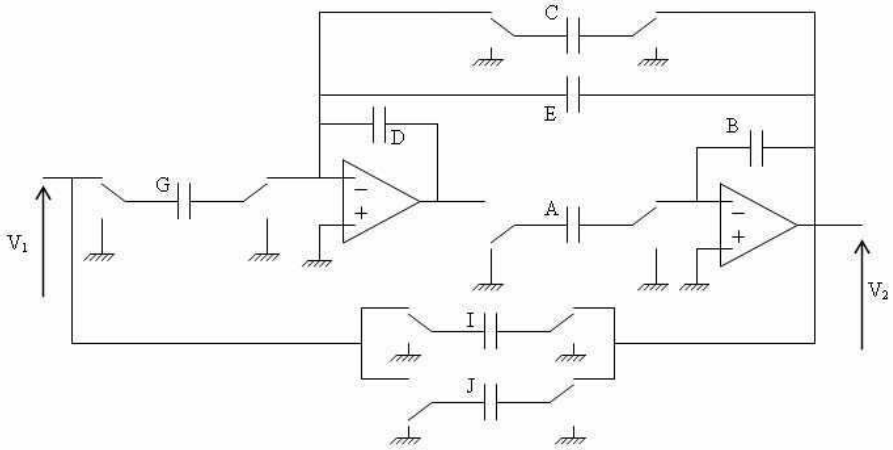


Figure 5.32. Switched capacitor biquadratic cell

### 5.7. Bibliography

[ALL 78] ALLSTOT D.J., BRODERSEN R.W., GRAY P.R. “MOS swiched capacitor ladder filters”, *IEEE Journal on Solid-State Circuits*, 12/1978, p. 806-814.

[BIL 80] BILDSTEIN P., *Filtres actifs*, Editions de la Radio, 3<sup>rd</sup> edition, 1980.

[CAU 41] CAUER W., *Theorie der Linearen Wechselstromschaltungen*, Akademische Vergas, Gesellschaft Becker, 1941.

[DIE 74] DIEULESAINT E., ROYER D., *Ondes élastiques dans les solides*, Monographies d’électronique, Masson, 1974.

[FRI 72] FRIED D.L., “Analog sample data filters”, *IEEE Journal on Solid-State Circuits SC* 7, August 1972.

- [HAS 81] HASLER M., NEIRYNCK J., *Filtres électriques*, vol. XIX, Editions Georgi, 1981.
- [HUE 76] HUELMAN L.P., *Active L-C filters, theory and applications*, Dowden, 1976.
- [SAA 58] SAAL R., ULBRICH E., "On the design of filters by synthesis", *IRE Transactions on Circuit Theory*, vol. CT 5, December 1958.
- [SED 79] SEDRA A.S., BRACKETT P.O., *Filter theory and design, active and passive*, Pitman, 1979.
- [ZOB 23] ZOBEL O.J., "Theory and design of uniform and composite wave filters", *Bell System Technical Journal*, January 1923.



*This page intentionally left blank*

## Chapter 6

# Real-time Data Acquisition and Processing Systems

### 6.1. Introduction

When carrying out numerical analysis in real time, it is not enough to sample the signal to be analyzed at a rhythm that complies with Shannon's criterion. It is also necessary to calculate at a speed compatible with the flow of the incoming samples. If, for a sampling frequency  $f_e$ , we make  $M$  basic operations<sup>1</sup> between each sampling, the necessary calculation power is  $M \cdot f_e$ , expressed in MOPS (millions of operations per second). This calculation power is a datum of the processor being used and is a compromise between the sampling frequency and the sophistication of the chosen analysis. But the lower the sampling frequency, the higher the order of the anti-folding filter;<sup>2</sup> thus it will be difficult to integrate into the numerical system. The popular technique currently used to solve this issue (over-sampling and decimation) is discussed at length in this chapter. Part of this discussion will deal with  $\Sigma - \Delta$  analog-to-digital converters using this technique. This will lead us into a presentation of implanting digital filters in cabled or programmed models (these can be comb or half-band filters).

In this chapter, we will only discuss "real" signals, that is, those with a physical existence. It is also quite often a question of limited spectrum signals. These are signals with a spectrum assumed to be zero outside a band  $[f_{\min} \dots f_{\max}]$ . From a mathematical point of view, limited spectrum "real" signals do not exist. Actually, a

---

Chapter written by Dominique MILLER.

1 This means it is part of the range of instructions of the processor being used.

2 This is always necessary, if only as regards noise.

limited spectrum signal is always of infinite duration, so it not very “physical” in nature. In addition, when we speak of a limited spectrum “real” signal, we mean a signal that has been limited the spectrum occupation to a band containing one sufficient part of the signal energy. This means we have replaced:

$$\int_0^{+\infty} |X(f)|^2 df \tag{6.1}$$

which contains the energy integrality of the signal, with:

$$\int_{f_{\min}}^{f_{\max}} |X(f)|^2 df \tag{6.2}$$

The concept of “sufficient part” obviously depends on the application.

## 6.2. Electronic devices for signal sampling and quantification

### 6.2.1. Nyquist sampling

In real-time, the Nyquist frequency is the minimum frequency that can be used for signal sampling. The first criterion is Shannon’s criterion, which shows that, for a limited spectrum signal  $[0..f_{\max}]$ , we can take<sup>3</sup>:

$$f_e > 2 \cdot f_{\max} \tag{6.3}$$

The second criterion is more constraining than the first. It is the anti-folding filter<sup>4</sup>, and it is indispensable, even if *a priori* the signal to be analyzed is present at input. Here, we must always take noise into account. If we consider that the signal at input is spoiled by a uniform density noise in a band  $[0..K \cdot f_e/2]$ , with K being even, a sampling at  $f_e$  will fold down in the band  $[0..f_e/2]$  the  $K-1$  bands for  $[nf_e \dots (n + \frac{1}{2})f_e]$  for  $-\frac{K}{2} \leq n < \frac{K}{2}$ . This means the noise power in the band  $[0..f_e/2]$  is therefore  $K$  times more significant after sampling than before, degrading the noise-signal ratio accordingly.

3 We will see in section 6.2.4 that there is another, less constraining version in the case of band signals limited around a carrying frequency.

4 Or, in symmetrical fashion, the smoothing filter as a analog-to-digital converter, if the analyzing device has an analog output that functions at the same frequency as the sampling.

The anti-folding filter must:

- “let the band go” through  $[0 \cdots f_{\max}]$ ;
- avoid folding in the band  $[0 \cdots f_{\max}]$  by sufficiently attenuating<sup>5</sup> beyond  $f_e - f_{\max}$ .

It becomes clear that we are not trying to eliminate folding in the band  $[0 \cdots f_e/2]$ . Consequently, there has to be some folding in the band  $[f_{\max} \cdots f_e/2]$ . We assume from the start that the numerical analysis we will carry out will sufficiently attenuate this band.

We see that the width of the transition zone of this filter is  $f_e - 2f_{\max}$ . For a given attenuation, the more narrow the transition zone, the higher the order of the filter; and since this is an analog filter, the higher the phase distortion in the band  $[0 \cdots f_{\max}]$  will be. By way of an example, we sample an order 9 elliptical filter at 256 kHz, and do an 18 bit quantification on a useable band signal heard from 0 to 100 kHz. We understand how difficult it will be to produce such a filter and the stability problems that will occur over time.

This is why another approach is being used increasingly. In order to limit the order of the analog filter, we will sample at a frequency significantly higher than Nyquist’s frequency. This will give us a good flow of samples, which we will reduce by digital filtering operations,<sup>6</sup> called decimations (analyzed in section 6.3.2). We will see in section 6.2.3 that over-sampling yields a better resolution than an analog-to-digital converter. This technique is fundamental to  $\Sigma - \Delta$  converters.

### 6.2.2. Quantification noise

The analog-to-digital converter, as a quantifier, introduces an error term. We can model it as in Figure 6.1. For a converter that rounds off, this term is included between  $-q/2$  and  $q/2$ , where  $q$  is the quantum. *A priori* this quantification error depends on the input signal.

---

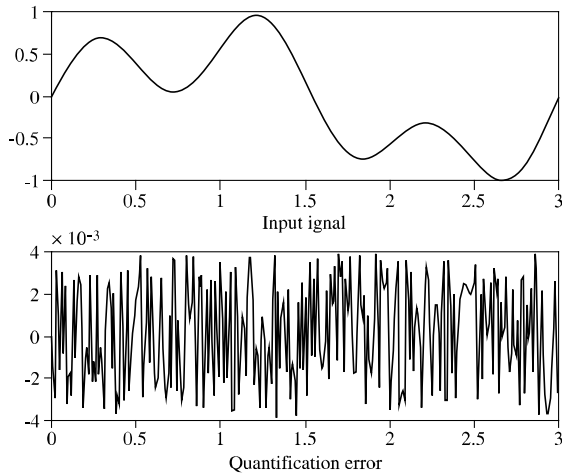
<sup>5</sup> “Sufficiently” here depends, in this case, on the quantum of the analog-to-digital converter being used. The folded data will be of a lower level than the quantum.

<sup>6</sup> These filters, even if they are often of a higher order than their analog equivalent, will be stable over time and more often will be linear-phase filters.



**Figure 6.1.** Modeling an analog-to-digital converter

However, for a high-amplitude signal, this error presents the speed of a noise (Figure 6.2). This means we consider the error to be an independent random input signal called quantification noise. This is characterized by its power spectral density. This approach is only valid if the quantum is small in relation to the maximum value of the input signal. Since this value defines the full range of the converter, the concept of independent quantification of the input signal only has meaning for converters with a fairly high number of bits.



**Figure 6.2.** Quantification error of a 8 bit full-scale CAN

By applying a triangular signal to an amplifier input, we easily show that the power of the quantification error equals  $\sigma_q^2 = q^2/12$  for a quantifier by rounding off. This power becomes the noise power of the quantifier. This concept does not interfere with the sampling frequency. It interferes only in that it guarantees a correct monitoring of the signal; that is, that we do not jump from the quantum between two samplings. Under these conditions, the quantification noise power does

not depend on the signal frequency, and the spectral density can be considered as uniform. In order to give a value to this density, and to respect Shannon's criterion as much as possible, we limit the input signal frequency to the interval  $0$  to  $f_e/2$ . We then over-extend the uniformity domain of the spectral signal throughout  $[0 \cdots f_e/2]$ . With this hypothesis, the spectral density of the quantification noise power equals:

$$S_q = \frac{\sigma_q^2}{\frac{f_e}{2}} = \frac{q^2}{6} \cdot \frac{1}{f_e} \quad [6.4]$$

From here, we establish a reference signal ratio to quantification noise where the reference signal is an amplitude sinusoid equal to the full range  $V_{ref}$  of the converter, whose frequency is in the band  $[0 \cdots f_e/2]$ :

$$\left( \frac{S}{B_q} \right)_{ref} = \frac{V_{ref}^2}{\frac{q^2}{12}} = 6 \left( \frac{V_{ref}}{q} \right)^2 \quad [6.5]$$

The quantum is calculated from  $V_{ref}$  and the  $M$  number of bits:

$$q = \frac{2V_{ref}}{2^M - 1} \approx \frac{2V_{ref}}{2^M} \Rightarrow \left( \frac{S}{B_q} \right)_{ref} \approx \frac{3}{2} \cdot 2^{2M} \quad [6.6]$$

or, expressed in dB, we have the standard relation:

$$10 \log \left( \frac{S}{B_q} \right)_{ref} = 1.76 + 6.02 \cdot M \quad [6.7]$$

### 6.2.3. Over-sampling

#### 6.2.3.1. Acquisition over-sampling

Let us suppose that a signal to be sampled has a Nyquist frequency equal to  $2f_a$ , meaning that its useful band is approximately  $[0 \cdots f_a]$ . We sample this signal at a frequency  $f_e = \alpha \cdot 2f_a$ , where  $\alpha > 1$  is the over-sampling factor. The first goal of this

over-sampling is to enlarge the width of the transition zone of the anti-folding filter that goes through  $2f_a(\alpha-1)$ . This helps lower the order of this filter and improves the phase linearity in the useful band. Again, there will be noise folding in the band  $[f_a \dots f_a(\alpha-1)]$ , but we already know that numerical filtering will eliminate this problem.

The other advantage of over-sampling is that it distributes the quantification noise over a higher band frequency, thereby reducing spectral density:

$$\frac{q^2}{6} \cdot \frac{1}{2f_a} \rightarrow \frac{q^2}{6} \cdot \frac{1}{\alpha \cdot 2f_a} \tag{6.8}$$

The power noise quantification in the only useful band is reduced by a factor  $\alpha$ :

$$\int_0^{f_a} \frac{q^2}{6} \cdot \frac{1}{\alpha \cdot 2f_a} \cdot df = \frac{q^2}{12} \cdot \frac{1}{\alpha} = \frac{\sigma_q^2}{\alpha} \tag{6.9}$$

For a full-scale sinusoid in the band  $[0 \dots f_a]$ , the signal to quantification noise ratio becomes:

$$10 \log \left( \frac{S}{B_q} \right)_{ref} = 1.76 + 6.02 \cdot M + 10 \log \alpha \tag{6.10}$$

We can write:

$$1.76 + 6.02 \cdot M + 10 \log \alpha = 1.76 + 6.02 \cdot M_e \tag{6.11}$$

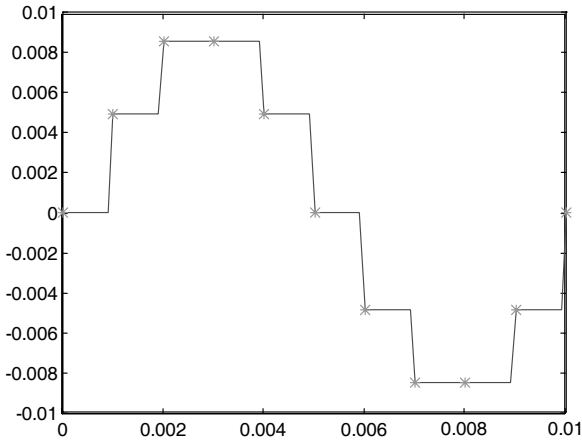
by taking:

$$M_e = M + \frac{1}{2} \log_2 \alpha \tag{6.12}$$

In other words, an analog converter of  $M$  bits, used with an over-sampling factor  $\alpha$ , leads to the same noise quantification power in the useful band as a converter of  $M_e$  bits used at Nyquist frequency. We then say that this converter has a resolution

or an equivalent number of bits of  $M_e$  bits. We gain 1 resolution bit each time we multiply the sampling frequency by four.

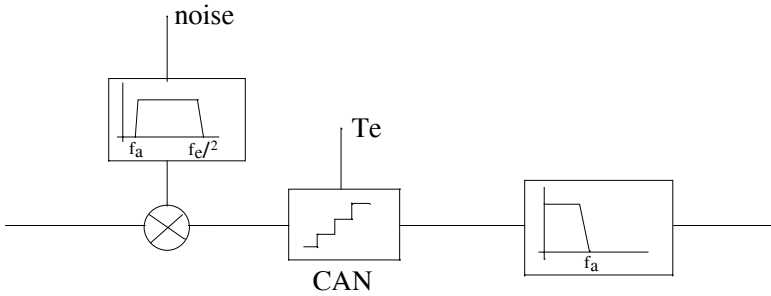
Here, this concept of “resolution” improved by over-sampling needs some explanation. For example, let us consider a 10 bit converter with a full range  $\pm 10$  V. Its quantum is 19.53 mV. We sample a sinusoid of 100 Hz with an amplitude of 9 mV at a frequency of 256 kHz. It becomes obvious that the converter output is always zero. If we then use a 14 bit converter at 1 kHz (whose quantum is 1.22 mV), we get the signal shown in Figure 6.3. The first converter is used with an over-sampling factor of 256 per second, leading to a theoretical resolution augmentation of 4 bits, or 14 equivalent bits. Obviously, this does not lead to the same result.



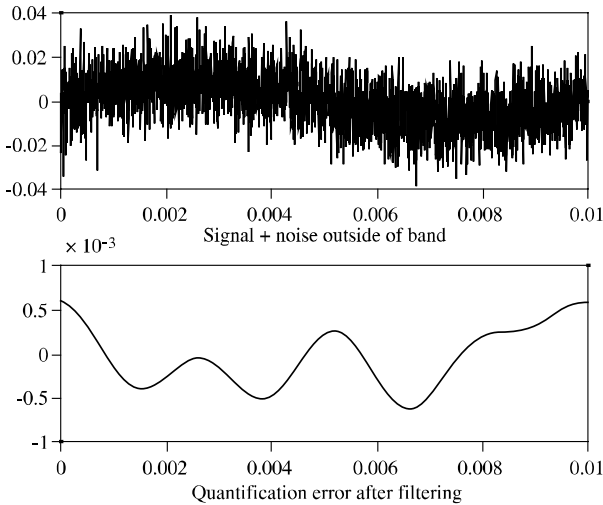
**Figure 6.3.** *Quantification by a 14 bit CAN/10 V with a sinusoid amplitude of 9mV*

Here, the idea of resolution is only defined in terms of the signal-quantification noise ratio in the useful band. However, we can obtain the effective resolution of 14 bits by superimposing on the input signal, before quantification, a noise apart from the useful band. This noise should have an amplitude higher than the quantum of the 10 bit converter. We eliminate this noise by digital filtering, leading to the quantification error shown in Figure 6.5. This operation is called dithering (see Figure 6.4).





**Figure 6.4.** Dithering technique



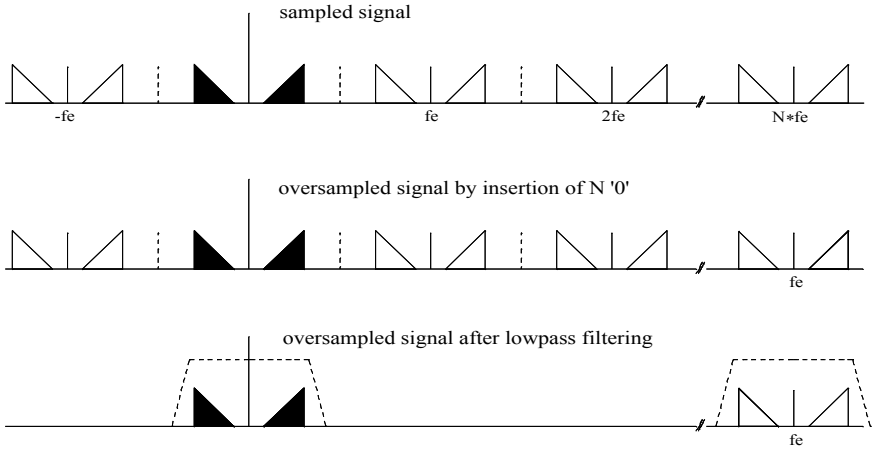
**Figure 6.5.** Quantification error after dithering

6.2.3.2. Over-sampling and reconstruction

If, after over-sampling with Nyquist frequency, we have to reconstruct the analog signal, the problem of the smoothing filter following the analog-to-digital converter is absolutely the same as with anti-folding filter. The over-sampling before conversion allows us to reduce the order of this filter.

Over-sampling of an entire factor  $N$  consists of, for the first time, inserting  $N - 1$  zeros between each sample. This clearly does not change the spectrum of the sampled signal (since we have added only zeros), but it does displace the sampling

frequency of the first image of the spectrum in the basic band to that of the  $N$ th power. A low pass numerical filtering, carried out at the new sampling frequency, helps eliminate all the intermediary images (Figure 6.6).



**Figure 6.6.** Output over-sampling

A sample of  $N$  of the signal to be filtered is not zero. This allows us to introduce this filter in a polyphase form. This reduces the power of the necessary calculation. If the filter is a FIR filter (for phase linearity), of length  $L$ , it must be calculated as follows:

$$y(n) = \sum_{i=0}^{L-1} a_i \cdot x(n-i) \quad [6.13]$$

In general,  $L > N$ . We can take, for example  $L = 2 \cdot N \cdot R + 1$  with  $4 \leq R \leq 10$ . If at this instant  $n$ ,  $x(n)$  is a non-zero sample, the calculation of  $y(n)$  is reduced to:

$$y(n) = \sum_{i=0}^{(L-1)/N} a_{N \cdot i} \cdot x(n - N \cdot i) \quad [6.14]$$

Then, at the instants  $n + 1, n + 2, \dots, n + N - 1$ :

$$\left. \begin{aligned}
 y(n + 1) &= \sum_{i=0}^{(L-1)/N} a_{N \cdot i + 1} \cdot x(n - N \cdot i) \\
 y(n + 2) &= \sum_{i=0}^{(L-1)/N} a_{N \cdot i + 2} \cdot x(n - N \cdot i) \\
 &\vdots \\
 y(n + N - 1) &= \sum_{i=0}^{(L-1)/N} a_{N \cdot i + N - 1} \cdot x(n - N \cdot i)
 \end{aligned} \right\} [6.15]$$

Then the filter initially of length  $L$ , working at a sampling frequency  $Nf_e$ , requires a power calculation  $L \cdot N \cdot f_e$ , which is divided into  $N$  filters of length  $(L-1)/N$ . These are calculated alternately. The necessary power calculation is therefore only  $Lf_e$ .

### 6.2.4. Under-sampling

A discussion of Shannon’s theorem usually concerns a signal to be sampled at limited spectrum in  $[0 \dots f_{\max}]$ , called a baseband signal, but in fact, this theorem is more general. It stipulates that the sampling frequency must be higher than the doubled width of the signal band, that is, for a signal at spectrum in  $[f_{\min} \dots f_{\max}]$ :

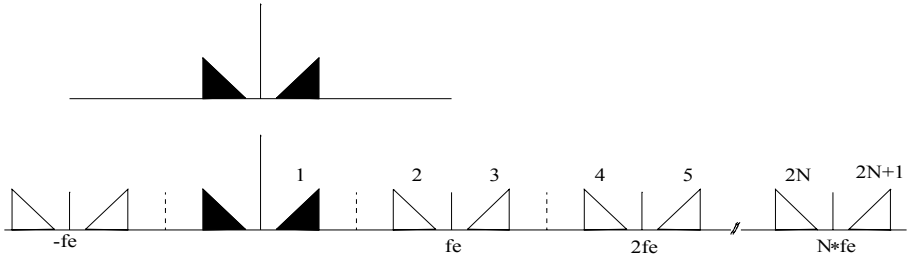
$$f_e > 2(f_{\max} - f_{\min}) \tag{6.16}$$

This does not mean we cannot take a sampling frequency lower than  $f_{\max}$  or  $f_{\min}$ . In general, we speak of under-sampling as soon as the sampling frequency is below  $2f_{\max}$ , this being the minimum frequency for standard sampling.

In Figure 6.7, we give the spectral effect of an ideal sampling of a basic band signal at limited spectrum.

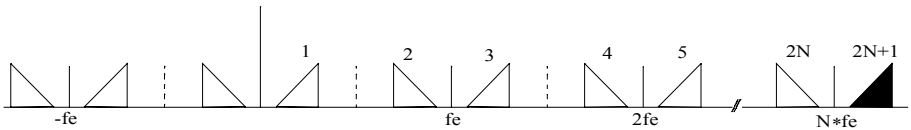
If we number the “Nyquist zones” that have a width  $f_e/2$ , we find that:

- the images of positive frequency are in the odd zones;
- the images of negative frequency are in the even zones.



**Figure 6.7.** Spectra before and after an ideal basic band sampling

If now, for a given sampling frequency, we start from a “high frequency, narrow band”<sup>7</sup> signal (HF-NB) that is located in an odd Nyquist zone, after under-sampling we find absolutely the same spectrum as in the previous case (see Figure 6.8). In zone 1 we find same spectral information as in the signal of origin at an interval ( $f \rightarrow f - Nf_e$ ). In other words, if the signal frequency is not useful information, the under-sampling preserves the information. The major under-sampling application domain is that of transmitting information by amplitude modulation and/or phase, or carrier frequency. This falls within the field of communication, but some sensors function according to the same modulation principle as a measurement carrier<sup>8</sup>.

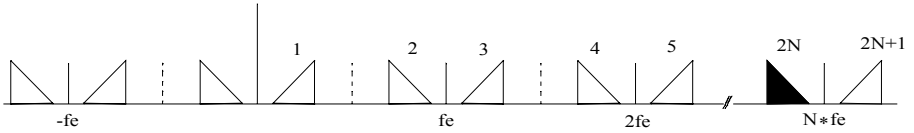


**Figure 6.8.** HF-NB Signal under-sampling in an odd zone

If the HF-NB is in an even zone (see Figure 6.9), we find, after under-sampling, the same spectrum, but with a frequency turn-up ( $f \rightarrow Nf_e - f$ ).

<sup>7</sup> This means the band is still lower by half than the sampling frequency.

<sup>8</sup> For example, sensors using Foucault currents in measurement applications of thickness and distance in front of a moving object. The frequency used is linked to the electromagnetic and geometric properties of the object, and this influences the module impedance and the sensor’s phase. But the development rapidity of this impedance, and thus the band width of the signal, depends mainly on the speed of movement.



**Figure 6.9.** Under-sampling of an HF-NB signal in an even zone

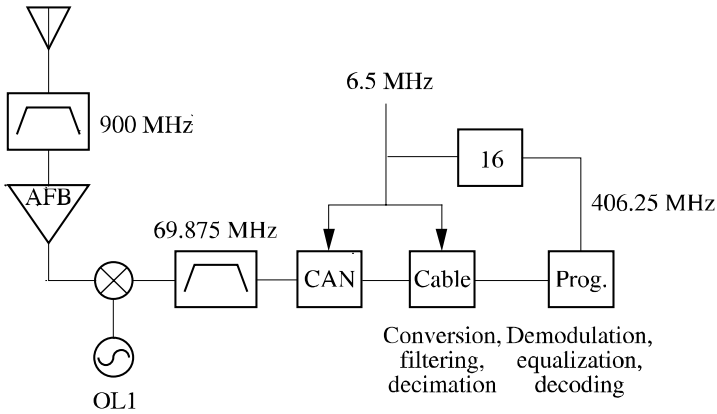
According to both cases, the criteria for under-sampling are as follows:

$$\left. \begin{aligned} N \cdot f_e < f_{\min} \\ \left(N + \frac{1}{2}\right) \cdot f_e > f_{\max} \end{aligned} \right\} \Rightarrow f_e > 2(f_{\max} - f_{\min}) \tag{6.17}$$

$$\left. \begin{aligned} \left(N - \frac{1}{2}\right) \cdot f_e < f_{\min} \\ N \cdot f_e > f_{\max} \end{aligned} \right\} \Rightarrow f_e > 2(f_{\max} - f_{\min}) \tag{6.18}$$

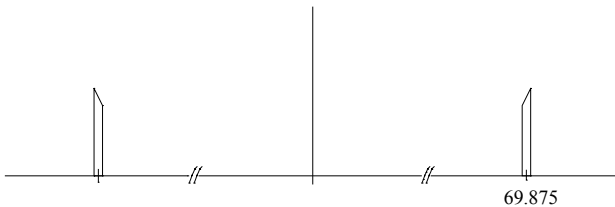
which clearly fits the Shannon criterion of “widened”.

We see that under-sampling can lead to using a sampling frequency much lower than those in the useful field domain. We can thus use an analog-to-digital converter with a fairly long conversion time. However, the sampling must be “ideal”, which means it must be very short as concerns the signal frequency domain. In addition, the sampling clock must have a very weak jitter, always in keeping with the signal frequency domain. This means that under-sampling requires using sampling structures that perform, as well as those of “standard” sampling. The importance is in the fact that the flow of numerical data to be analyzed by the processor may be low and is not connected to the carrying frequency. To demonstrate this (see Figures 6.10 to 6.15), we present the structure and the different analysis steps of a GSM receptor functioning by under-sampling of the first intermediary frequency (69.875 MHz) for a band width of 200 kHz (corresponding to the juxtaposition of 8 channels).

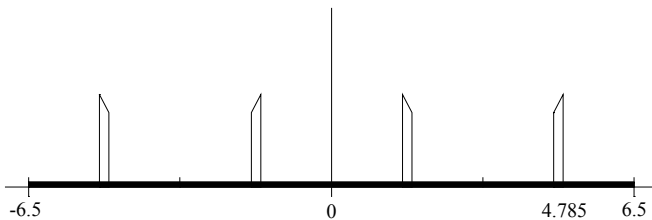


**Figure 6.10.** Structure of a GSM digital receiver

In this structure, the processor finally receives a flood of digital data at 406.25 kHz. We should see that part of the digital analysis (translation, filtering and decimation<sup>9</sup> of a factor 16) of the output of an analog-to-digital converter is carried out by wired structures.

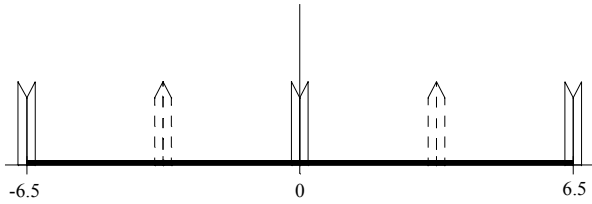


**Figure 6.11.** Signal of origin, at intermediary frequency

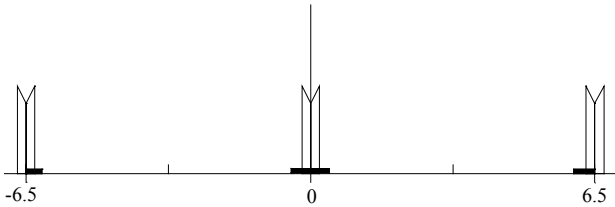


**Figure 6.12.** After sampling at 6.5 MHz, zones 1 and 2 and quantification noise

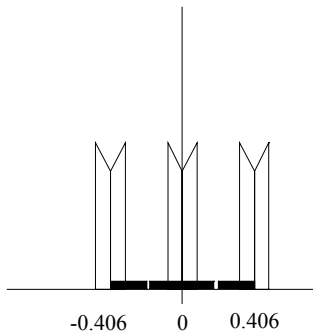
<sup>9</sup> See section 6.3.2 on a discussion of  $\Sigma - \Delta$  converters.



**Figure 6.13.** Changing digital frequency at 100 kHz



**Figure 6.14.** After low pass digital filtering at 200 kHz



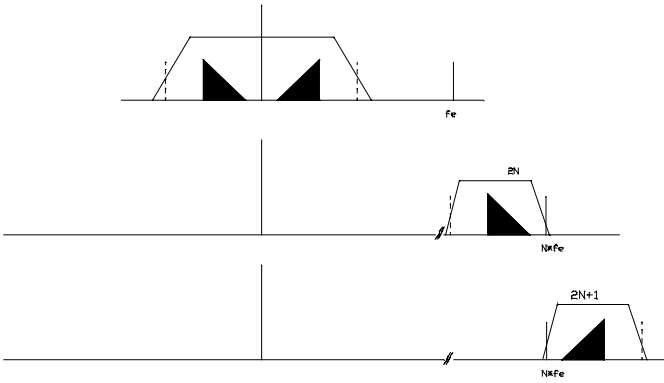
**Figure 6.15.** After decimation by a factor of 16

The anti-folding filter is still necessary, even if only to limit the noise band to  $[f_{\min} \dots f_{\max}]$ . The anti-folding filter then becomes a pass band (see Figure 6.16).

If the pass band filter is symmetrical in relation to its central frequency, it becomes important to center the band in a Nyquist zone:

– in an odd zone: 
$$f_c = \frac{f_{\max} + f_{\min}}{2} = \left(N + \frac{1}{4}\right) \cdot f_e \tag{6.19}$$

– in an even zone:  $f_c = \left(N - \frac{1}{4}\right) \cdot f_e$  [6.20]



**Figure 6.16.** Anti-folding filterings in standard sampling and in under-sampling

We get from this a relation between the sampling frequency, the central frequency of the signal to be sampled, and the number of the Nyquist zone in which we find the signal:

$$\left. \begin{array}{l} \text{even zone} \Rightarrow Z = 2N \\ \text{odd zone} \Rightarrow Z = 2N + 1 \end{array} \right\} \Rightarrow f_e = \frac{4f_c}{2Z - 1} \quad [6.21]$$

In this way we establish the relation between the carrying frequency of 69.875 MHz and the sampling frequency of 6.5 MHz.

### 6.3. Analog-to-digital converters

Since the technologies develop so quickly, in a reference book of this type it is not helpful to discuss all utilization domains of different types of analog-to-digital converters. The following points briefly summarize a few general points:

– there are converters that carry out successive approximations, all-purpose converters, as well as converters with medium resolution and medium sampling frequency converters;

– there are “high speed” converters. These are not defined by their number of bits (that is, by their binary word format), but by a number of effective bits that take imperfections into account, depending on the frequency;



– gradient converters, both the high resolution and low speed varieties, have been supplanted by  $\Sigma - \Delta$  converters.

In this section we have only touched on “new” points, introducing the idea of number of effective bits and, most importantly, the  $\Sigma - \Delta$  converter, which uses the more modern techniques of over-sampling and digital decimation.

### 6.3.1. Features of SINAD<sup>10</sup> and ENOB<sup>11</sup> converters

The first criterion for choosing an analog-to-digital converter is its number of bits  $M$ , determined by the dynamic of the signal  $D^{12}$  to be quantified, as well as by the minimum signal to quantification noise ration  $RSBq_{min}$  that we want to allow. For the full-range signal and a converter of  $M$  bits, the signal to quantification noise ratio is given in equation [6.7]. For the minimum signal, of lower useful value  $D$  dB, the signal-to-noise ratio is only:

$$RSBq_{min} = 6.02 \cdot M + 1.76 - D \tag{6.22}$$

If this quantity is a constraint, we then take:

$$M = \frac{RSBq_{min} + D - 1.76}{6.02} \tag{6.23}$$

From this point, the converter can be characterized in terms of dynamic or static non-linearities, missing codes and so on. These features can be obtained by applying a sinusoid to the converter’s input. This sinusoid must be of very high spectral purity. We then analyze the incoming samples:

- by histogram, that is, mainly in order to detect the missing code;
- by FFT for non-linearities.

---

10 Signal to Noise and Distortion ratio.

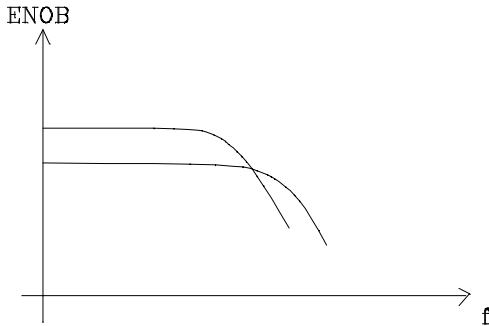
11 Effective Number of Bits.

12 That is, the relation between its maximum value and it minimum useful value, expressed in decibels.

The spectral analysis of an ideal converter gives a single line to the input frequency<sup>13</sup> emerging from a lower limit of quantification noise. With non-linearity, lines will appear at harmonic frequencies. We then call SINAD the ratio, in power, of the fundamental line to the highest disturbance lines or of the ground noise. If there is no distortion, the SINAD is equal to the reference signal-to-noise ratio defined in section 6.2.2, depending only on the number of bits. If not, we define the number of effective bits by:

$$ENOB = \frac{SINAD - 1.76}{6.02} \quad [6.24]$$

The non-linearities depend on the input frequency and increase with it. At low frequencies, the number of effective bits is equal to the number of bits of the converter; but this decreases at higher frequencies. This means an 8 bit converter can have an ENOB of 6.7 bits for an input signal of 100 MHz. A general rule is that we get curves of the type shown in Figure 6.17.



**Figure 6.17.** Development of the ENOB depending on the input frequency, for two converters

At a higher input frequency, the best ENOB does not necessarily correspond to the highest number of low frequency bits.

### 6.3.2. $\Sigma - \Delta$ converters

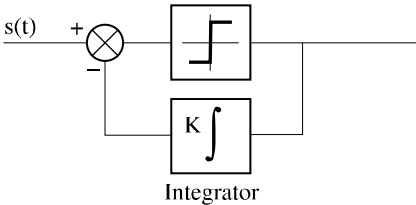
These were first developed by audiodigital technology. They have increasingly supplanted gradient converters for instrumentation purposes. The problem of

---

<sup>13</sup> There must be a precise relationship between the sampling frequency and the sinusoid frequency.

audiodigital technology was this: it used a base Nyquist frequency (44 kHz for a pass band of 20 kHz) in order to reduce the number of samples to be memorized and analyzed. This was done with a quantification of at least 16 bits for 90 dB of desired dynamic and a signal-to-noise ratio of the order of 8 dB, leading to very high order anti-folding filters. This type of filter is not only tricky to produce; in an essentially analog technology it cannot be integrated on the same chip as the rest of the digital part.<sup>14</sup> This resulted in higher costs. To overcome this problem, manufacturers have tried to reduce as much as possible the analog part by transferring most of the analysis towards digital processes. Over-sampling has provided this solution. Today, 64 the most widely used over-sampling factor, bringing the anti-folding filter to an order of 2 to 4. However, there are not (or were not) 16 bit converters at 2.8 MHz (64 times 44 kHz). But we can make use of the fact that over-sampling helps improve a converter's resolution (see section 6.2.3). To improve resolution even more, we proceed to  $\Sigma - \Delta$  modulation that helps reject quantification noise at high frequencies, that is, outside the useful band. We can then significantly lower the number of converter bits, and even use a 1 bit converter (a simple comparator) that has the advantage of being perfectly linear. An inherent part of the converter, digital analysis for filtering and decimation helps return the initial high flow and low resolution to the Nyquist flow with a resolution of 16, 18, ..., 24 or even 28 bits! That is the range of techniques discussed in this section.

The  $\Sigma - \Delta$  modulator is supported in the beginning by the  $\Delta$  modulator (see Figure 6.18) whose output is a signal modulated by impulse length. Its mean value is equal to the derivation of the input signal.



**Figure 6.18.**  $\Delta$  modulator

In order to obtain an output with a mean value directly proportional to an input signal, we integrate the input signal ahead of the  $\Delta$  modulation, hence the name  $\Sigma - \Delta$  filter.

---

<sup>14</sup> The production and purchasing levels reached by the field of audio technology has led to the creation of specific integrated circuits.

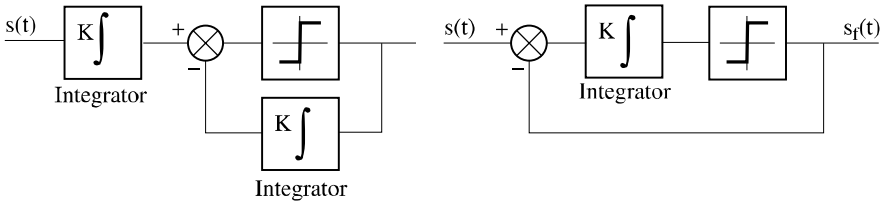


Figure 6.19.  $\Sigma - \Delta$  modulator

In a sampling context, the comparator is actually an analog-to-digital converter (possibly 1 bit, more often  $M$  bits), that we model as a source of quantification noise<sup>15</sup> (see Figure 6.20). In the return loop, an analog-to-digital converter is necessary. It must be modeled by a pure time-lag.

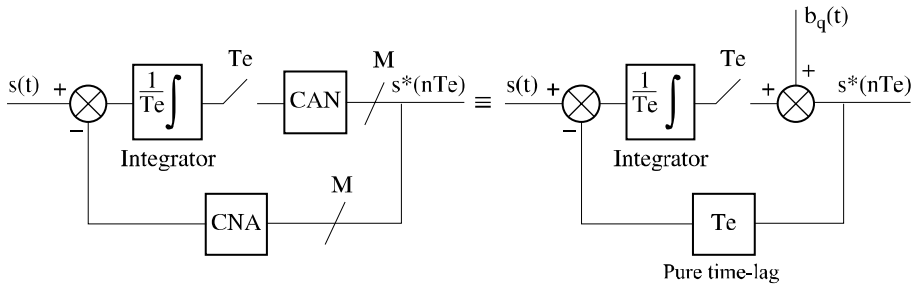


Figure 6.20. Sampled  $\Sigma - \Delta$  modulator

The transfer functions in  $z$  of the integrator and of the time-lag are respectively  $\frac{1}{1-z^{-1}}$  and  $z^{-1}$ . In these conditions, we get:

$$S^*(z) = S(z) + B_q(z)(1 - z^{-1}) \tag{6.25}$$

This is a “pass-through” structure with regard to the signal and a differentiator; that is high pass structure regarding quantification noise. This means it doubles the noise power (if we consider that two successive samplings of the quantification error

---

15 Extrapolating the concept of quantification noise to a 1 bit converter is a bit risky. However, it does give a qualitative approach and the technique of dithering helps us use this model.

are independent), but rejects it at higher frequencies. We say that the noise has been “put into form”. Its power spectral density is obtained by calculating:

$$N_q(f) = \frac{q^2}{6} \frac{1}{f_e} \left| 1 - z^{-1} \right|_{z=e^{j2\pi f/f_e}}^2 = 4 \left| \sin\left(\pi f/f_e\right) \right|^2 \frac{q^2}{6} \frac{1}{f_e} \tag{6.26}$$

If  $\alpha$  is the over-sampling factor and  $q$  is the converter’s quantum, the noise power in the useful band is:

$$\int_0^{f_e/2\alpha} 4 \left| \sin\left(\pi f/f_e\right) \right|^2 \frac{q^2}{6} \frac{1}{f_e} \cdot df = \frac{q^2}{6} \left[ \frac{1}{\alpha} - \frac{1}{\pi} \sin\left(\frac{\pi}{\alpha}\right) \right] \tag{6.27}$$

by supposing  $\frac{\pi}{\alpha} \ll 1$ , this power is written as:

$$\frac{q^2}{6} \cdot \frac{1}{6} \cdot \frac{\pi^2}{\alpha^3} \tag{6.28}$$

We can express the signal-noise ratio with the reference signal according to  $M$  and  $\alpha$ .

$$\left( \frac{S}{B_q} \right)_{ref} = \frac{3}{2} \cdot \frac{3}{\pi^2} \cdot 2^{2M} \cdot \alpha^3 \tag{6.29}$$

$$10 \log \left( \frac{S}{B_q} \right)_{ref} \approx 1.76 + 6.02M + 10 \log \alpha + 10 \log \left( \frac{3\alpha^2}{\pi^2} \right) \tag{6.30}$$

So as soon as  $\alpha > \pi/\sqrt{3} \approx 1.8$ , the signal-to noise ratio increases much faster with  $\alpha$  than in the case of simple over-sampling, without appearing as noise. We gain 9 dB, that is, 1.5 bits of resolution each time that  $\alpha$  doubles.

We can also improve this resolution by using a  $\Sigma - \Delta$  modulator of the order  $N$  (see Figure 6.21). This is expressed by:

$$S^*(z) = S(z) + B_q(z) \left( 1 - z^{-1} \right)^N \tag{6.31}$$

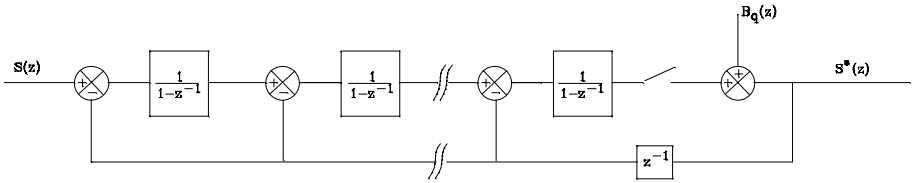


Figure 6.21.  $\Sigma - \Delta$  modulator of order  $N$

The power spectral density of the noise put into form is then:

$$N_q(f) = 2^{2N} \left| \sin\left(\pi \frac{f}{f_e}\right) \right|^{2N} \cdot \frac{q^2}{6} \cdot \frac{1}{f_e} \tag{6.32}$$

Integrating this function is not simple, and we must be content with a first order development. This is reasonable because we integrate from 0 to  $f_e/2\alpha$ , a field for which  $f \ll f_e$ , since we over-sample from a significant factor.

$$\left| \sin\left(\pi \frac{f}{f_e}\right) \right|^{2N} \approx \left| \pi \frac{f}{f_e} \right|^{2N} \tag{6.33}$$

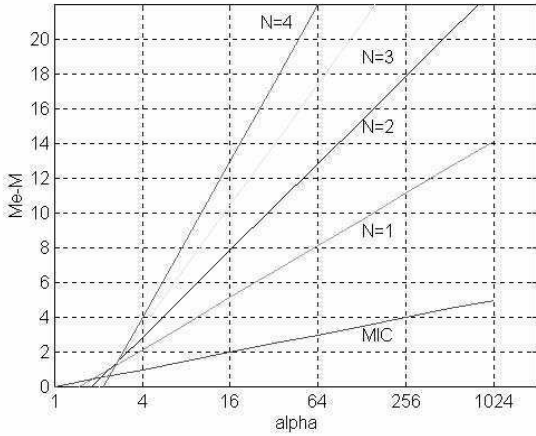
We then get:

$$\left( \frac{S}{B_q} \right)_{ref} = \frac{3}{2} \cdot \frac{2N+1}{\pi^{2N}} \cdot 2^{2M} \cdot \alpha^{2N+1} \tag{6.34}$$

$$10 \log \left( \frac{S}{B_q} \right)_{ref} = 1.76 + 6.02M + 10 \log(\alpha) + 10 \log \left[ (2N+1) \frac{\alpha^{2N}}{\pi^{2N}} \right] \tag{6.35}$$

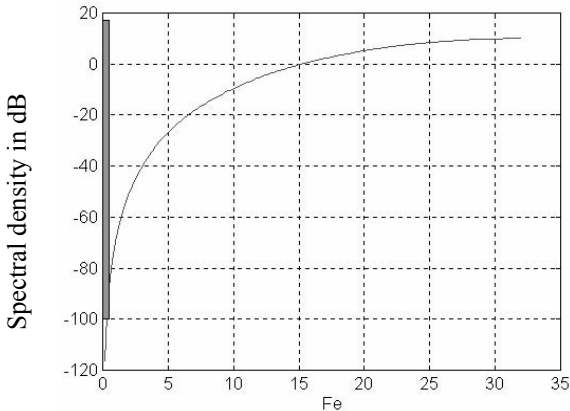
We can give, in monograms, the resolution gain<sup>16</sup> according to  $\alpha$  and  $N$  (Figure 6.22) and compare it with the simple over-sampling (“MIC” in the figure). Theoretically, we get more than 20 bits of resolution with a 1 bit converter, a modulator of order 4, and an over-sampling factor of 64. We should not forget that, in addition to a good resolution, we get a converter with very good linearity. This is because when we use a low-bit converter, the resulting digital analysis will also be linear.

16 That is, the gap between the number of effective bits and the number of converter bits used in the modulator.



**Figure 6.22.** Resolution gain of a  $\Sigma - \Delta$  modulator of  $N$  order

We thus have a significant flow of quantified data with a low number of bits, representative of a basic signal frequency and a significant noise level; rejected beyond, however, the useful band of the signal (Figure 6.23).



**Figure 6.23.** Spectral densities of the signal ( $7 V_{eff}$ ) and of the quantification noise for  $M = 1$ ,  $\alpha = 64$  and  $N = 3$

Digital filtering suppresses this high frequency noise in order to conserve only the useful band  $[0 \dots f_e/2\alpha]$ . Since at the end of this digital filtering, there is nothing left in the band  $[f_e/2\alpha \dots f_e/2]$ , we can only take a sample on  $\alpha$  without losing

information. This brings the sampling frequency to the Nyquist frequency. This ensemble of operations, filtering and lowering the sampling frequency, is called decimation. This process will be discussed in detail.

*A priori*, it is sufficient to set up a digital filter of type FIR to ensure linearity. We can calculate the order of the filter by using a standard empirical formula:

$$2M + 1 = \frac{2}{3} \log \left( \frac{1}{10\delta_p\delta_a} \right) \cdot \frac{f_e}{\Delta f} \quad [6.36]$$

where:

- $\delta_a$  and  $\delta_p$  are undulations in attenuated bands and pass bands;
- $\Delta f$  is the width of the transition band of the filter.

Looking at the example in Figure 6.23, we see that we want to bring the noise level to that in the pass band (–100 dB is the “resolution” of a 16 bit converter) for a transition width of  $22 - 20 = 2$  kHz. This means we need an attenuation of 110 dB. We can choose an undulation in the pass band lower than  $\frac{1}{2}$  quantum. Let this be  $152 \mu\text{V}$  (16 bits, full range 10 V). The sampling frequency is  $64 \cdot 44 \text{ kHz} = 2.816 \text{ MHz}$ . All this leads to a filter order of around 13,000. With a processor capable of analyzing 200 million operations per second, at 2.816 MHz, we cannot produce a filter of an order above 70. This means we have to proceed in several steps, by beginning with a filter that can be made by using a cabled structure: the comb filter.

The comb filter is an FIR filter, with all coefficients equal. It is therefore a linear phase filter, and there is a recursive way of expressing it. For a filter of  $L$  length, we get:

$$h(z) = h_0 \cdot \sum_{k=0}^{L-1} z^{-k} = h_0 \cdot \frac{1 - z^{-L}}{1 - z^{-1}} \Rightarrow y(n) = h_0 \cdot \sum_{k=0}^{L-1} x(n-k) \quad [6.37]$$

The response in frequency of this filter is given by the following equation and is shown in Figure 6.24.

$$h(j\omega) = h_0 \frac{e^{-j\omega \frac{L}{2} T} \sin\left(\omega \frac{L}{2} T\right)}{e^{-j\omega \frac{L}{2} T} \sin(\omega T)} \Rightarrow |h(f)| = h_0 \cdot \frac{\left| \sin\left(\pi \frac{fL}{f_e}\right) \right|}{\left| \sin\left(\pi \frac{f}{f_e}\right) \right|} \quad [6.38]$$



Its continuous gain is therefore  $h_0 \cdot L$ . For a unitary gain we take  $h_0 = 1/L$ . This filter carries out the arithmetic mean of the last  $L$  input samples. If  $L = 2^l$ , the division by  $L$  in turn performs  $l$  intervals to the line of the accumulated total of the samples. Under these conditions, the filter can be made with the help of simple accumulators and registers, that is, by means of a hard-wired logic that can function at a very high sampling frequency.

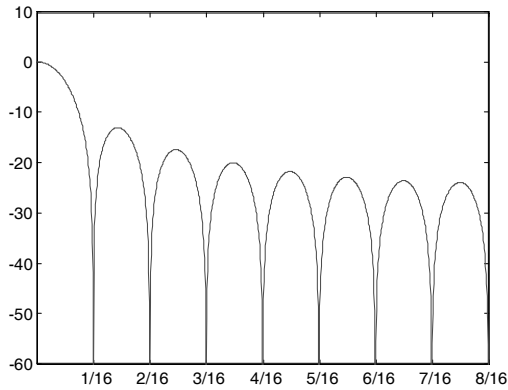


Figure 6.24. 16 length comb filter, response in dB between 0 and  $f_c/2$

The attenuation is infinite throughout  $f_c/L$ , and is significant around these frequencies. This means that if the useful band of the filtered signal is much lower than  $f_c$ , we can proceed to a decimation of order  $L$  (that is, we only sample  $L$ ) without folding in the useful band. Actually, the decimation of order  $L$  will fold the frequencies in the useful band that are close to different multiples of  $f_c/L$ . The comb filter of length  $L$  easily helps us produce a decimator filter of order  $L$  that can function at high frequencies (see Figure 6.25).

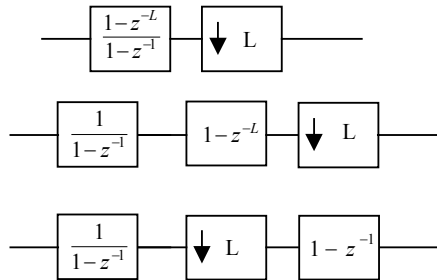
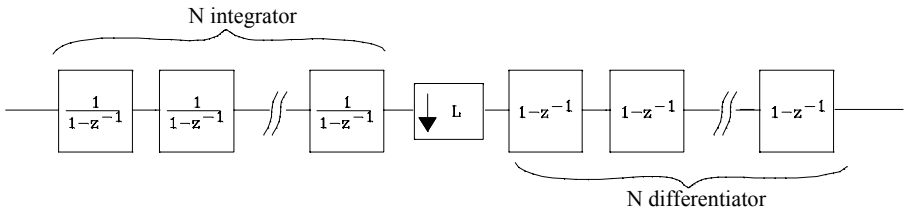


Figure 6.25. Three equivalent ways of producing comb filtering and one decimation

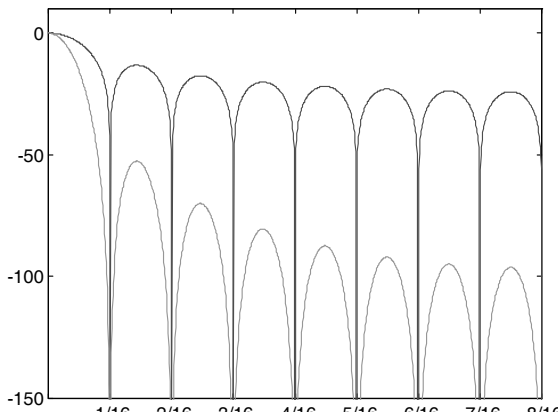
In the last analysis, the filtering-decimation ensemble is made by means of:

- an accumulator  $\frac{1}{1-z^{-1}}$ , functioning at  $f_e$ ;
- taking a sample of  $L$  at the accumulator's output;
- a differentiator  $1-z^{-1}$  functioning at  $f_e/L$ .

To enlarge the zones of high attenuation around all the  $f_e/L$  areas, we increase the filter order by cascading the accumulators<sup>17</sup> and differentiators before and after decimation (Figure 6.26). The frequency response for an order 4 is compared to that of an order 1 in Figure 6.27.



**Figure 6.26.** Order  $N$  comb filter



**Figure 6.27.** Response in frequency (in dB) of comb filters of 16 lengths and of order 1 and 4, between 0 and  $f_e/2$

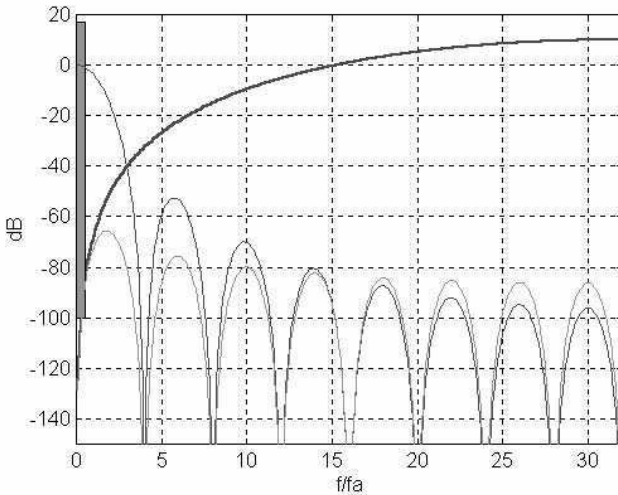
<sup>17</sup> This cascading does not require accumulators with a high number of bits.

This kind of filter is used as the first decimator filter at the output of the  $\Sigma - \Delta$  modulator. It is most frequently of an order 4 and a length 16; this means it has a decimation of factor 16. This filter does not perform well enough to be able to go as low as Nyquist frequency. This phase is completed by FIR programmed filters. For these filters, we must minimize the necessary power calculation.

Looking again at Figure 6.23, we see that the noise remaining after a comb filter of 16 lengths and of order 4, but before decimation of a factor 16, is given in Figure 6.28. We see that, except for the first two lobes, all the others are below  $-80$  dB. The decimation will move back all these lobes for a total level of  $-64$  dB. To proceed further in decimation (there remains a factor 4), we must bring this level to the required  $-100$  dB. By calculating the order of the necessary filter, we get the following parameters:

- 36 dB and 152  $\mu$ V undulation;
- sampling frequency of 176 kHz and transition width of 2 kHz.

This leads to an order of 376 or a more realistic calculation power<sup>18</sup> of 66 MOPS (which can be even somewhat higher).



**Figure 6.28.** *Quantification noise after comb filtering*

---

<sup>18</sup> By assuming that the processor can make a product and an accumulation in one operation.

The decimation, done in two steps with half-band filters, helps reduce the necessary power calculation.

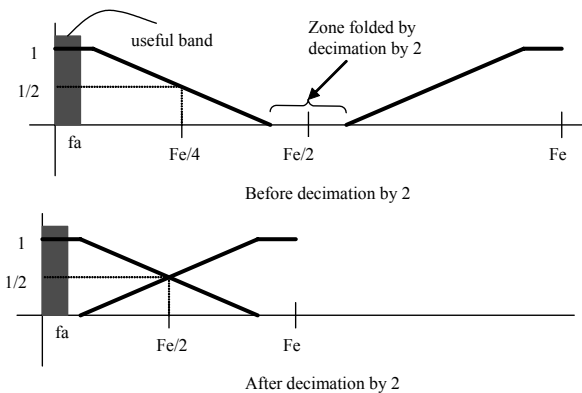
A half-band filter is really a linear phase low pass filter with a unitary gain in the pass band, so that  $H(f)$  presents an odd symmetry around the point  $(\frac{f_c}{4}, 0.5)$ :

$$H\left(\frac{f_c}{4} - f\right) = 1 - H\left(\frac{f_c}{4} + f\right) \tag{6.39}$$

These kinds of filters are of odd order, with even coefficients, apart from the central coefficient, which is zero. For a given order of a filter, the required power calculation is two times lower than for an ordinary FIR filter. In addition,  $\delta_a = \delta_p$ . This kind of filter is especially useful to the decimation of a factor 2, as we see in Figure 6.29. Since the transition width is relatively large, this leads to a slightly higher order. Still for the same example:

- for a first order half-band filter:
  - $\delta_a = \delta_p = 50 \cdot 10^{-6}$  (1/3 of the previous undulation because we will cascade three filters);
  - sampling at 176 kHz and transition width  $(88 - 20) - 20 = 48$  kHz;
- for a second order half-band filter:
  - $\delta_a = \delta_p = 50 \cdot 10^{-6}$ ;
  - sampling at 88 kHz and transition width  $(44 - 20) - 20 = 4$  kHz.

This gives filters of order 19 for 111 for the second. So the total power calculations of  $10 \cdot 176 \cdot 10^3 + 56 \cdot 88 \cdot 10^3 \approx 6.7$  MOPS.



**Figure 6.29.** Half-band filter and decimation

We have to complete the procedure with a standard FIR filter that compensates for the attenuation in the pass band resulting from the comb filter and carries out a last filtering in the transition zone. Of an order of around 100, at 44 kHz, it adds about ten MOPS, which is within the range of all specialized processors.

This concept of over-sampling (with or without being in the form of quantification noise) and decimation is so important that specialized integration circuits have been made that include comb filters and FIR filters.<sup>19</sup>

Today, we find integrated converters<sup>20</sup> that have a modulator, digital filters and decimation capabilities. With the resulting high number of bits, the interface is serial, often connected to a processor unit.

#### 6.4. Real-time digital analysis by a specialized processor

Specialized processors used in signal analysis are basically conceived for linear analyses of the convolution type or the equivalent. These can be expressed by:

$$y = \sum_i a_i \cdot x_i \quad [6.40]$$

Choosing the number of bits used to represent intervening variables in the analysis depends on the same criterion used for choosing the number of bits for an analog-to-digital converter (see section 6.3.1). We need to know the dynamic of these variables and the admissible quantification.

Representation in fixed-point notation (in whole numbers or fractions) leads to a non-uniform quantification error that is more significant for lower values. However, representation in floating point (with mantissa and exponents) will be more or less independent of the quantified value.

A calculation unity expressed in floating point is more complex, and thus more costly<sup>21</sup>; above all, it uses more energy. There are also processors that use both types of notation, but the ones that use fixed-point notation tend to be preferred when energy consumption is a concern.

---

19 For example, HSP50016, HSP43220, HSP43168, etc., from Harris [HAR 95].

20 Often with two in the same circuit, since the audio is stereo.

21 But which especially depends on the production volume.

**6.4.1. Fixed point and floating point analysis**

6.4.1.1. *Fixed point notation*

Here we will discuss only complement notation of 2s, which is used solely for the purpose of carrying out fixed point calculations<sup>22</sup>.

If a whole number is shown in complement of 2s by the  $N$  binary digits  $b_{N-1}b_{N-2}\dots b_0$ , its numerical value is:

$$-b_{N-1} \cdot 2^{N-1} + b_{N-2} \cdot 2^{N-2} + \dots + b_1 \cdot 2^1 + b_0 \cdot 2^0 \tag{6.41}$$

The most weighted bit is representative of the sign, but there is a weighting since:

$$-2^N + 2^{N-1} = 2^{N-1} \cdot (-2 + 1) = -2^{N-1} \tag{6.42}$$

An integer represented with  $N$  bits can be represented by  $N + 1$  bits by recopying the most weighted bits. The extension of the sign is:

$$-b_{N-1} \cdot 2^N + b_{N-1} \cdot 2^{N-1} + \dots + b_1 \cdot 2^1 + b_0 \cdot 2^0 \tag{6.43}$$

We can represent the fractional numbers by arbitrarily putting a decimal point between the bits  $k - 1$  and  $k$ . We then speak of a representation in  $Q_k$ . The representation  $b_{N-1}b_{N-2}\dots b_k \bullet b_{k-1}\dots b_0$  has the value of:

$$\left(-b_{N-1} \cdot 2^{N-1} + b_{N-2} \cdot 2^{N-2} + \dots + b_1 \cdot 2^1 + b_0 \cdot 2^0\right) \cdot 2^{-k} \tag{6.44}$$

For a coding in fractional numbers, the dynamic<sup>23</sup> is always  $(2^{N-1} - 1)$ . The quantification error ( $1/2$  LSB or  $2^{-k-1}$ ) is a constant, and relatively more important for low values.

6.4.1.2. *Floating point notation*

We notate the number in the form  $2^{\text{exp}}$  mantissa, where  $\text{exp}$  is an integer in complement of 2s on  $E$  bits and mantissa is a fractional number coded as  $Q_k$  on  $M = N - E$  bits ( $k \leq M$ ). The number is then written on  $N$  bits:

$$\begin{array}{l} b_{N-1}b_{N-2} \dots\dots\dots b_0 \\ e_{E-1}e_{E-2} \dots e_0 m_{M-1}m_{M-2} \dots m_0 \end{array} \tag{6.45}$$

---

22 Many AN and NA converters use, by successive approximations, the offset binary. We continue with the complement of 2s by complementing the most weighted bit.  
 23 That is, the relation, in absolute values, of the highest value to the non-zero lowest value.

For a reason we will explain below, we exclude the possible values by exposing the value  $-2^{E-1}$ . We then get:

$$-2^{E-1} + 1 \leq \text{exp} \leq 2^{E-1} - 1 \tag{6.46}$$

If we work with positive numbers:

- the highest positive value is  $2^{2^{E-1}-1} \cdot (2^{(M-1)} - 1) \cdot 2^{-k}$  ;
- the lowest non-zero positive value is  $2^{-2^{E-1}+1} \cdot 1 \cdot 2^{-k}$ .

This gives us a dynamic of  $2^{2^E-2} \cdot (2^{M-1} - 1)$ .

The dynamic in floating point notation is thus around  $2^{2^E-2-E}$  times higher than for a fixed point notation on the same number of bits. For 16 bits having 4 exponents and 12 of mantissa, the dynamic is  $3.35 \cdot 10^7$  in floating notation as opposed to  $3.28 \cdot 10^4$  in fractional.

The multiplicity of notation is hard to control. We can impose a condition on the mantissa (similar to that of the scientific notation in the decimal system):

$$1 \leq |mantissa| < 2 \tag{6.47}$$

The value “0” is then represented by  $\text{exp} = -2^{E-1}$ .

The notation of the mantissa is in  $Q_{M-2}$ , but one of the most weighted of the 2 bits is necessarily non-zero:

- the highest positive value is  $2^{2^{E-1}-1} \cdot (2^{(M-1)} - 1) \cdot 2^{-(M-2)}$ ,
- the lowest positive non-zero value is  $2^{-2^{E-1}+1} \cdot 1$ .

So the dynamic is  $2^{2^E-2} \cdot (2 - 2^{-(M-2)})$ . This is around  $2^{(M-2)}$  times lower than in the previous example. For 16 bits, 4 of exponents and 12 of mantissa, the dynamic is of  $3.28 \cdot 10^4$ , practically the same as for fractional notation. In terms of the dynamic, we have lost the advantage of the floating point, but the quantification error is no longer a constant. It equals  $2^{\text{exp}} \cdot 2^{-M+1}$ . This means that for the relative error, we have, no matter what the value is:

$$\frac{2^{-M+1}}{2} = 2^{-M} < \left| \frac{\partial x}{x} \right| \leq \frac{2^{-M+1}}{1} = 2^{-M+1} \tag{6.48}$$

We should keep in mind the following points.

- For positive numbers, the mantissas between 1 and 2 coded as  $Q_{M-2}$  are written 01.xxxxxxx.

– For negative numbers, the mantissas between -1 and -2 coded as  $Q_{M-2}$  are written as 10.xxxxxxx.

As such, 2 bits of the mantissa are always complementary. It is redundant to conserve both. We gain a bit of resolution by making one of them implicit.<sup>24</sup>

There is an IEEE 754 normalization of numbers in floating notation, allowing for the exchange of data between different systems. The format, on 32 bits, is:

SEEEEEEEEMM...M

The exponent on 8 bits is in offset binary of 126. The binary mantissa is signed on 23 bits, with the first “1” implicit. The notated value is then:

$$(-1)^S \cdot 2^{\text{exp}-127} \cdot (01.\text{mantisse}) \quad [6.49]$$

In certain cases:

- 0 if  $\text{exp} = 0$  and mantissa = 0;
- $(-1)^S \infty$  if  $\text{exp} = 255$  and mantissa = 0;
- NaN (Not A Number) if  $\text{exp} = 255$  and mantissa  $\neq 0$ ;
- $(-1)^S \cdot 2^{-126} \cdot (0.\text{mantissa})$  if  $\text{exp} = 0$  and mantissa  $\neq 0$  (denormalization).

If the concepts of *NaN* or of *infinity* are of mathematical importance, we cannot do much in real-time.

#### 6.4.1.3. Comparison between the two notations

If we impose a minimum precision (that is, a maximum relative error), the fixed point dynamic is much lower than in floating point. If we compare the two notations on 32 bits with, for the floating point, 8 bits of exponent, we find the following.

- in floating point, the relative error is always below  $6 \cdot 10^{-8}$ , for a dynamic of the order of  $10^{76}$ ;
- in order to have a relative maximum error below  $6 \cdot 10^{-8}$ , the lowest value that can be notated is therefore  $8.333 \cdot 10^6$ , for a maximum value of  $2.14 \cdot 10^9$  or a dynamic of 256.

Using a processor in fixed point requires special attention in the area of data framing, in order to obtain the best precision by avoiding overflow problems. In floating point, the problem of framing is not of crucial importance. The only remaining issue is the effect on the final result of the variable quantification.

---

<sup>24</sup> In TMS320C3x processors, made by Texas Instruments, the bit of the mantissa sign is kept and the following bit is omitted.



### 6.4.2. General structure of a DSP<sup>25</sup>

Almost by definition, a signal analysis processor is a RISC machine<sup>26</sup>. This is so because we need a machine that can analyze a limited number of operations with maximum efficiency. The limited number of instructions, a maximum resource parallelization and a pipeline of 4 or 6 levels helps us obtain an optimum situation: 1 instruction = 1 word = 1 cycle.

The examples in this section are drawn from processors of TMS320C54x (fixed point) and TMS320C3x (floating point) made by Texas Instruments, the uncontested leader in DSP processors. What we present here is general information: the structures described here can be found in all these processors.

In this section we are only interested in real-time analysis, which generally means that the analysis is done on samples taken by an analog-to-digital converter. In other words, in the expression:

$$y = \sum_i a_i \cdot x_i \tag{6.50}$$

the  $x_i$  are successive samples  $x(n-i)$ ; they are numbers noted in fixed point, coming from a  $M$  bits converter, thus in  $Q_{M-l}$ .

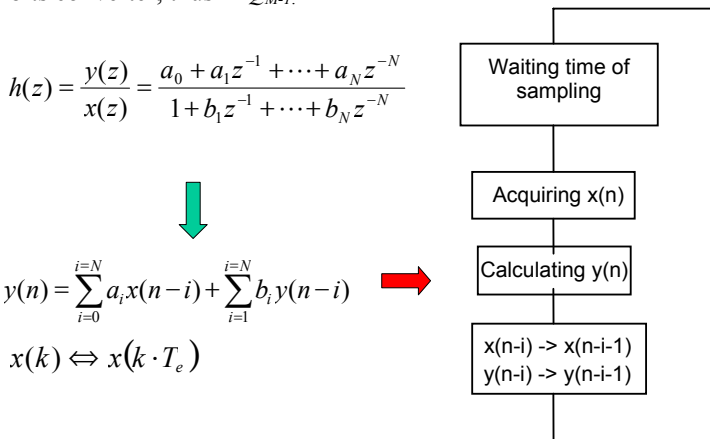


Figure 6.30. Linear real-time digital analysis

<sup>25</sup> Digital Signal Processor: a specialized processor for signal analysis.

<sup>26</sup> Educated Instruction Set computer: a calculator (or processor) with a reduced range of instructions, which allows for a more efficient and rapid wiring circuit, with an instruction carried out by cycle.

Looking at Figure 6.30, we see that for efficient analysis, we must have:

- a multiplication/accumulation structure (see section 6.4.2.1);
- a time lag or data aging structure (see section 6.4.2.2).

Another structure that is not explicitly shown in Figure 6.30, but is nevertheless necessary, for both DSP processors working in fixed point and floating point is data reframing. It will be described in section 6.4.2.3.

#### 6.4.2.1. *Multiplication/accumulation structure*

This operation is the basis of all DSP processors, often indicated by MAC (*Multiply and Accumulate*<sup>27</sup>). The usual structure is shown in Figure 6.31. The multiplication makes two operands play a part, which can be identical or different:

- they can be different for adaptive operations at constant coefficients. The  $a_i$  are the constants and can be arranged as “ROM”, or more generally as program memory, with the  $x(n-i)$  and  $y(n-i)$  being the variables that can be stored as “RAM” (data memory);

- they can be identical for adaptive or correlative filtering operations. All the operands are variable and can be stored as data memory.

In the first case, the operands can be introduced simultaneously if distinct buses exist, and the multiplication/accumulation can be carried out in a single cycle. In the second case, a temporary register  $T$  is necessary, and there must be a supplementary cycle.

We have already discussed multiple buses. These have two types of structures:

- the modified Harvard type of structure with a memory space “program” and a memory space for “data”, all accessible by several bus addresses and data (Figure 6.32). The “program” and “data” spaces can be internal and/or external. This kind of configuration allows, means of a pipeline, an instruction “*fetch*”, two data readings, and a data writing, all in a single cycle (see Figure 6.37). The one condition that must be met for this to happen be that the two “program” and “data” spaces are respectively internal and external;

- a single memory space “simplified”, but with multiplication buses to allow for simultaneous access to the different memory units (Figures 6.34 and 6.35).

---

<sup>27</sup> It is under the name MAC that the aptitude of a general-interest processor or micro-controller is designed to carry out signal analysis.

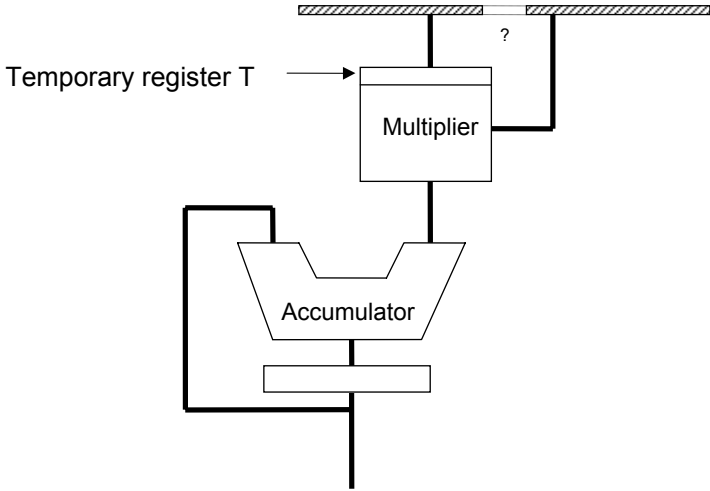


Figure 6.31. Multiplier/accumulator

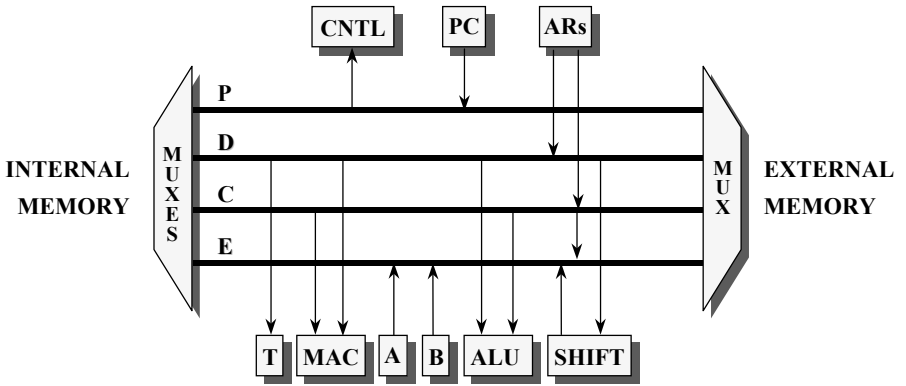


Figure 6.32. Structure of modified Harvard (doc. Texas Instruments C54x)



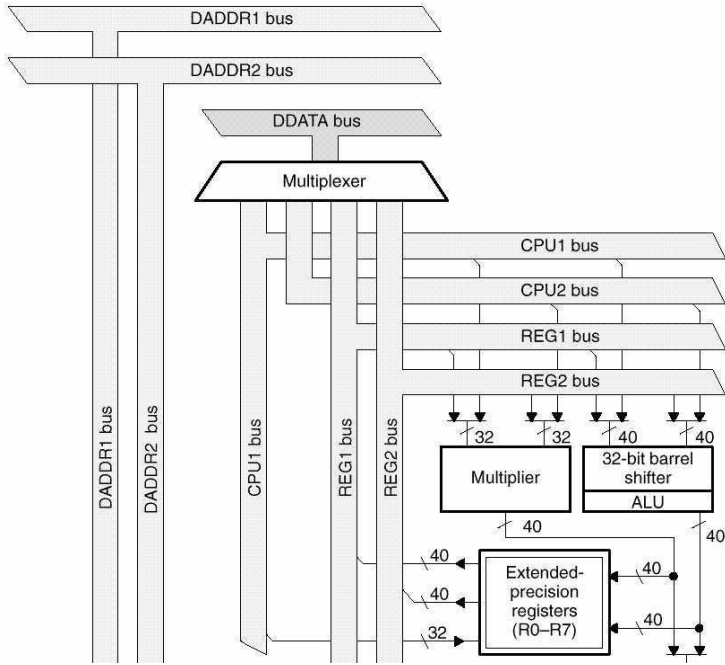


Figure 6.35. C3x MAC unit (doc. Texas Instruments)

6.4.2.2. Time lag structures

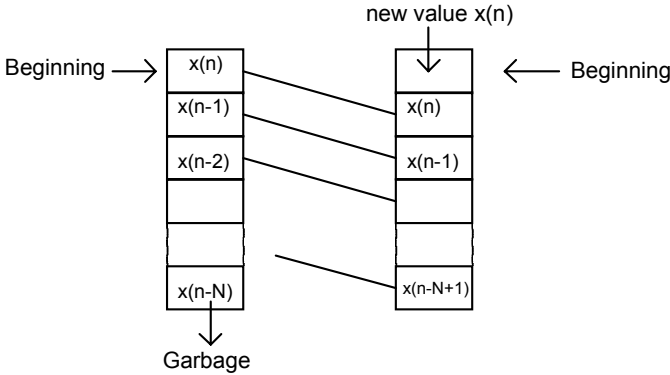
Time lags converts a sample  $x(n-i)$  and  $x(n-i-1)$  for the next convolution. There are two techniques used for carrying out this operation:

- time shifts carried out through memory (see Figure 6.36);
- the use of pointers on a circular table (see Figure 6.38).

Time shifts carried out through memory stimulate high bus activity. However, if the convolution is calculated in “inverse” order, we get:

$$\sum_{i=0}^N a_i \cdot x(n-i) = a_N \cdot x(n-N) + a_{N-1} \cdot x(n-N+1) + \dots + a_0 \cdot x(n) \tag{6.51}$$

Once the sample  $x(n-i)$  is loaded in the multiplier, it can be rewritten at the next address  $x(n-i-1)$ , which already has been used and been shifted. This operation is called MACD (Multiply, Accumulate, and Delay: see Figure 6.37). The modified Harvard structure seen in section 6.4.2.1 helps us produce it in a single cycle if the pipeline is used correctly.

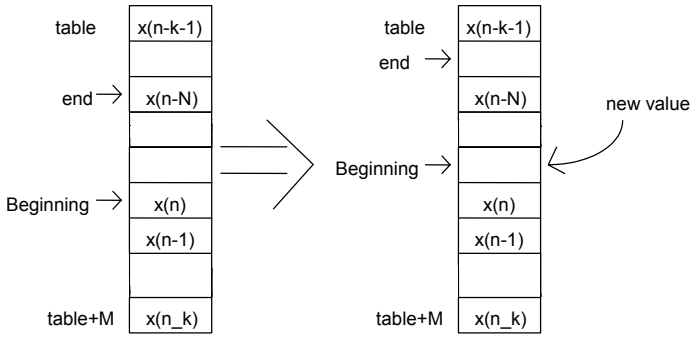


**Figure 6.36.** Time shifting by memory shifts

<p><u>Mult./add. instructions:</u></p> <ul style="list-style-type: none"> <li>•MPY[R] Smem,dst :     (Smem)*(T)-&gt;dst</li> <li>•MPY Xmem,Ymem,dst     (Xmem)*(Ymem)-&gt;dst     (Xmem)-&gt;T</li> <li>•MAC[R] Smem,dst :     (Smem)*(T)+(dst)-&gt;dst</li> <li>•MAC Xmem,Ymem,src[,dst]     (Xmem)*(Ymem)+(src)-&gt;dst     (Xmem)-&gt;T</li> </ul>	<p><u>Time shift instructions:</u></p> <ul style="list-style-type: none"> <li>•DELAY Smem     (Smem)-&gt;Smem+1</li> <li>•LTD Smem     (Smem)-&gt;T     (Smem)-&gt;Smem+1</li> </ul>
<p><u>Multiplication/accumulation and time shifts:</u></p> <p>MACD Smem,pmad,src     (Smem)*(Pmad)+(src)-&gt;src     (Smem)-&gt;T     (Smem)-&gt;Smem+1</p>	

**Figure 6.37.** Arithmetic operations and possible shifts with a modified Harvard structure (extract from instructions of C54x)

Pointers are the only solution to circular addressing (that is, of  $AR_n$  address registers). But these pointers must allow us to carry out modular operations throughout the length of the table. This requires a dedicated arithmetical unit, such as the Address Register Arithmetic Unit (ARAU). This allows for efficient address tables (see Figure 6.39).



**Figure 6.38.** Time shifts done with pointer on a circular table

Mod Field	Syntax	Operation	Description
00000	*+ARn( <i>disp</i> )	addr = ARn + <i>disp</i>	With predisplacement add
00001	*-ARn( <i>disp</i> )	addr = ARn - <i>disp</i>	With predisplacement subtract
00010	*++ARn( <i>disp</i> )	addr = ARn + <i>disp</i> ARn = ARn + <i>disp</i>	With predisplacement add and modify
00011	*--ARn( <i>disp</i> )	addr = ARn - <i>disp</i> ARn = ARn - <i>disp</i>	With predisplacement subtract and modify
00100	*ARn++( <i>disp</i> )	addr = ARn ARn = ARn + <i>disp</i>	With postdisplacement add and modify
00101	*ARn--( <i>disp</i> )	addr = ARn ARn = ARn - <i>disp</i>	With postdisplacement subtract and modify
00110	*ARn++( <i>disp</i> )%	addr = ARn ARn = circ(ARn + <i>disp</i> )	With postdisplacement add and circular modify
00111	*ARn--( <i>disp</i> )%	addr = ARn ARn = circ(ARn - <i>disp</i> )	With postdisplacement subtract and circular modify

**Figure 6.39.** Extract from addressing mode by register of C3x (doc. Texas Instruments)

Apart from the problem of shifting, addressing by register (or indirect addressing) is useful because it makes instruction coding possible with a single word, whether with one, two, or three operands.

### 6.4.2.3. Reframing structures

Suppose we want to carry out, on a fixed point DSP, an analysis of the type:

$$y(n) = \sum_i a_i \cdot x(n - i) \tag{6.52}$$

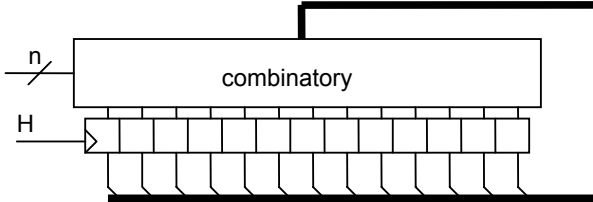
Here,  $y(n)$  and the  $x(n-i)$  are samples directly represented as whole numbers, since they come from or will go into AN or NA converters. The  $a_i$  are coefficients that can be determined by one of several techniques, and are therefore real numbers that we will represent as  $Q_k$ :

$$a_i \rightarrow A_i = \text{round} \left( a_i \cdot 2^k \right) \quad [6.53]$$

The processor will calculate:

$$Y(n) = \sum_i A_i \cdot x(n-i) = 2^k \cdot y(n) \quad [6.54]$$

At each sampling period,  $y(n)$  will be obtained by a cropping<sup>28</sup> of  $Y(n)$ . A *barrel shifter* is the structure that allows us to shift a binary word to the right or left of a quantity. If the principle structure (Figure 6.40) is simple, the number of ports of the combination unit increases significantly.



**Figure 6.40.** Barrel shifter

If we now consider a DSP working in floating point mode, the problem of reframing appears when two numbers are added (Figure 6.41) when we must bring one of the exponents to the same value as the other.

<sup>28</sup> A multiplication by  $2^{-k}$  is carried out by  $k$  shifts, to the right or left according to the sign of  $k$ .



$$x_1 = m_1 \cdot 2^{\text{exp}_1}, 1 \leq |m_1| \leq 2$$

$$x_2 = m_2 \cdot 2^{\text{exp}_2}, 1 \leq |m_2| \leq 2$$

$$\text{exp}_2 = \text{exp}_1 + k, k > 0$$

$$x = x_1 + x_2 = (m_2 + 2^{-k} m_1) \cdot 2^{\text{exp}_2} = m \cdot 2^{\text{exp}}$$

$$\text{if } |m| \geq 2, x = m/2 \cdot 2^{\text{exp}_2+1}$$

Figure 6.41. Addition of two numbers in floating point mode

6.4.2.4. Resource parallelization

All the resources described above are for the most part parallelized (see Figures 6.42 and 6.43).

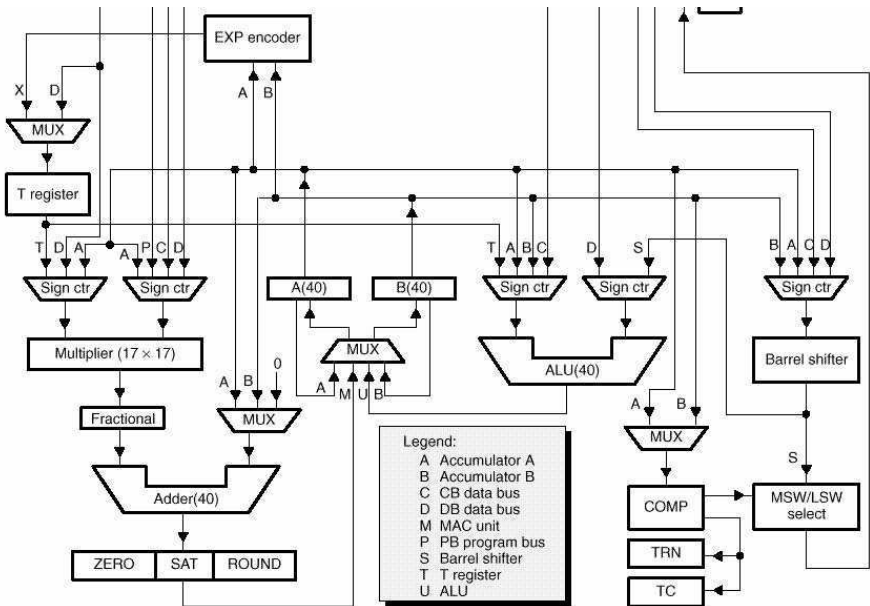
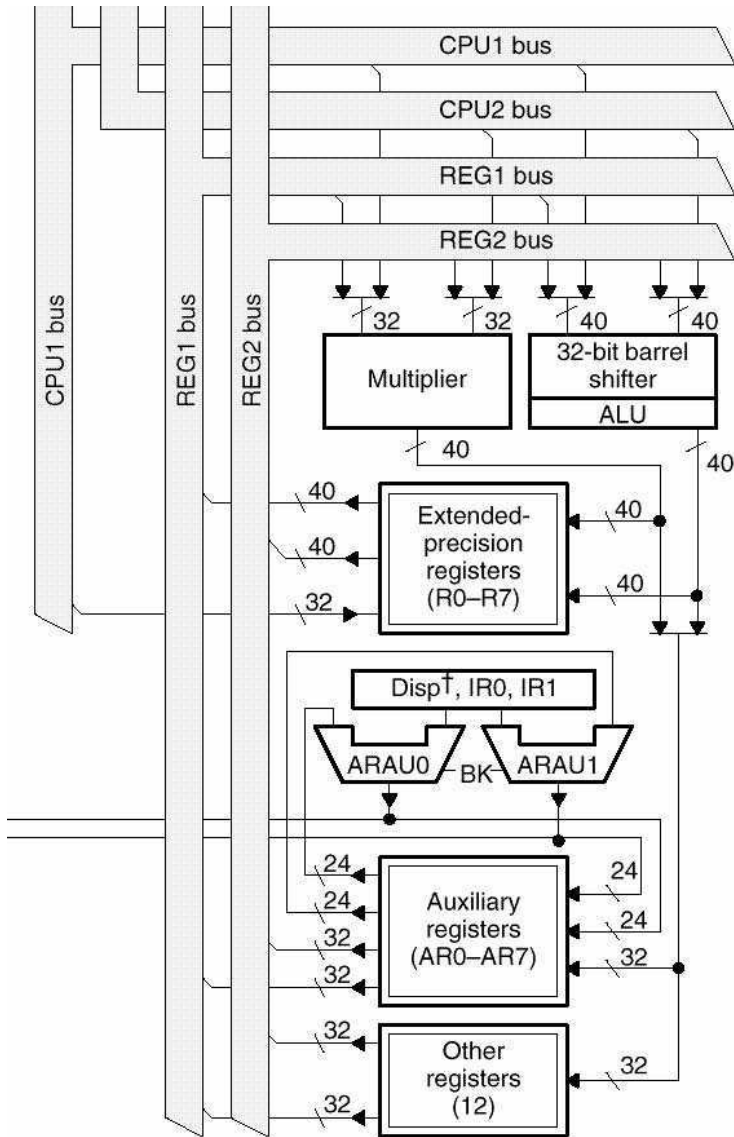


Figure 6.42. Resource parallelization in a C54x (doc. Texas Instruments)



**Figure 6.43.** Resource parallelization in a C3x (doc. Texas Instruments)

This structure is linked to an intensive pipeline. In most cases, this allows for functioning according to the following rule: one instruction = one cycle (Figures 6.44 and 6.45).

<b>P</b>	<b>Drive address of instruction</b>	→	<b>P<sub>A</sub></b>
<b>F</b>	<b>Collect instruction</b>	←	<b>P<sub>D</sub></b>
<b>D</b>	<b>Interpret instruction, plan job</b>		<b>ctrl</b>
<b>A</b>	<b>Set up pointers, Calc data address</b>	→	<b>D<sub>A</sub>/C<sub>A</sub></b>
<b>R</b>	<b>Collect operand</b>	←	<b>D<sub>D</sub>/C<sub>D</sub></b>
	<b>Calculate Write address</b>	→	<b>E<sub>A</sub></b>
<b>X</b>	<b>Execute operation</b>		<b>*,+</b>
	<b>Send result</b>	→	<b>E<sub>D</sub></b>

Figure 6.44. Decomposition of a C54x instruction (doc. Texas Instruments)

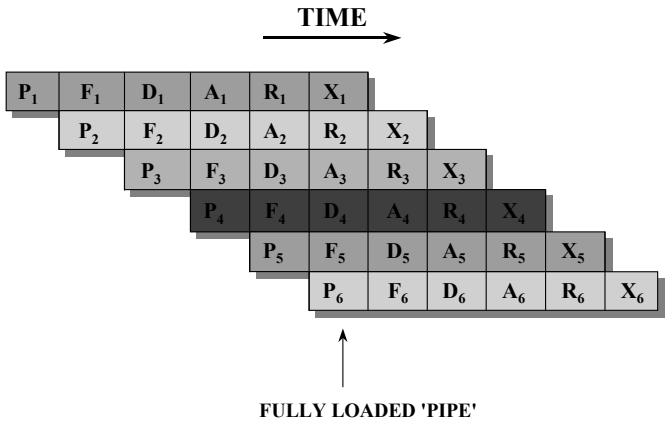


Figure 6.45. C54x pipeline (doc. Texas Instruments)

It should be remembered here that, because of the pipeline, the instructions giving the parallel operations calculate with the register values before the beginning of the instructions. This means that a MAC instruction accumulates a product that has already been calculated. It then calculates a new product which will then be accumulated subsequently.

**6.4.3. Using standard filtering algorithms**

6.4.3.1. General structure of a real-time filtering program

In a digital filtering operations, there is input and output of analog variables via the converters. Since parallelization is a concern, DSPs have input/output ports.

Usually these are serial, and interface directly with the analog interfaces of the same family. These interfaces have converters, anti-folding filters and switched smoothing capacities. The counters help determine the filters' sampling frequencies and break frequencies (see Figure 6.46).

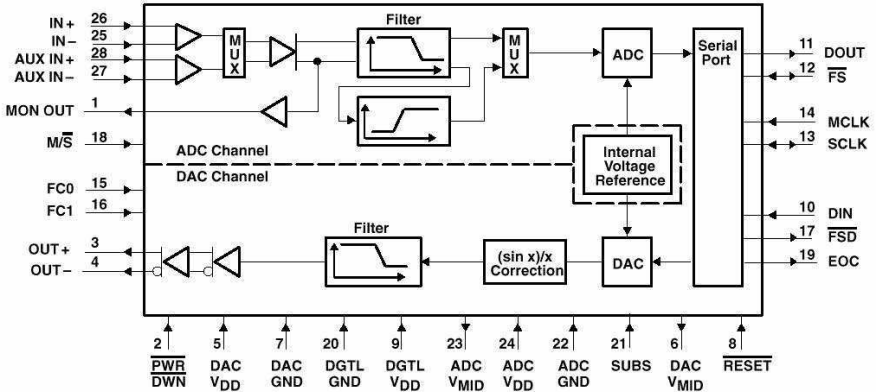


Figure 6.46. TCL32040 analog interface (doc. Texas Instruments)

The interfaces also ensure data transfer to and from the DSP. This makes the exchanges completely transparent for the programmer. Almost without interruption, an input register can be read and an output register can be filled in when a transfer has taken place. The main program resumes at initializations and while waiting for interruptions (see Figure 6.47) and analysis is done during the interruptions (Figure 6.48).

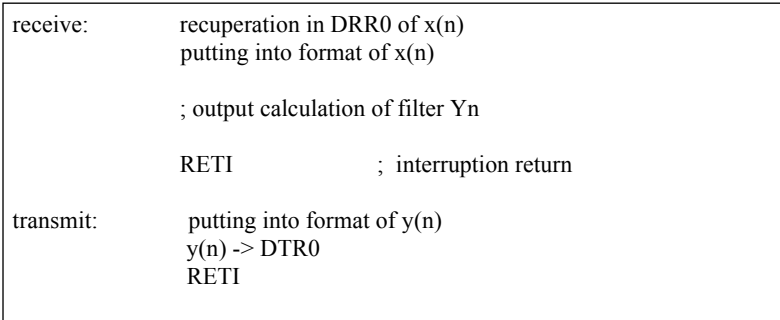
The system is configured so there will be interruptions:

- each time the CAN carries out a conversion (DRRO full)
- each time the system carries out a conversion (DRRO empty)

```
.text
; different initializations and pointers

wait_and_see:  idle          ; IT waiting
               NOP          ; continuous loop
               NOP
               b  wait_and_see
```

Figure 6.47. Main program (C3x syntax)

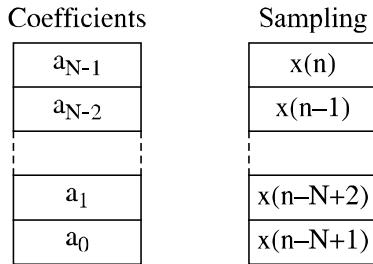


**Figure 6.48.** Analysis of emission and reception interruptions

6.4.3.2. The FIR filter and simple convolutions

The transfer function in  $z$  of these filters being a simple polynomial, they are always stable. Quantifying coefficients, especially in fixed point, can be expressed by a slightly different frequency response, but this does not lead to instability. However, there is no specific structure for setting up these filters. We can choose the memory-based structure shown in Figure 6.49 for a filter of length  $N$ . The relevant algorithm is given in Figure 6.50. We see that there is a setting for the instruction repetition RPTS that uses a loop counter and provides for the optimum functioning of the pipeline in the *fetch* phase. Use of the pointers is as follows:

- the pointer on the coefficients undergoes  $N$  incrementations modulo  $N$ . It thus returns to the initial situation;
- the pointer on the samples undergoes  $N + 1$  decrements modulo  $N$ . It shifts at each FIR carrying out. This brings about the time shift.



**Figure 6.49.** Memory-based structure for an FIR filter

R4 contains $x(n)$ AR1 point $a(n-N+1)$ , AR2 point $x(n-N+1)$ The circular address is made modulo N		
FIR:	stf R1,*AR2--(1)%	storage $x(n)$ and time shift
	ldf 0,R2	RAZ sum total
	mpyf *AR1++(1)%,*AR2--(1)% ,R0	first product
	RPTS N-2	repetition N-1
	mpyf *AR1++(1)%,*AR2--(1)% ,R0	
	addf R0,R2,R2	MAC
	addf R0,R2,R2	final product
the result is in R2		

**Figure 6.50.** FIR filter algorithm (C3x syntax)

Only six words are necessary, whatever the filter order.  $11 + (N - 1)$  cycles must be anticipated.

If we use a fixed-point DSP with a quantification closest to the coefficients (see section 6.4.1), the accumulation can temporarily exceed the maximum representable value. This can force us to a quantification as  $Q_k$  with  $k$  lower; that is, with a relatively higher quantification error. However, most accumulators or data registers have guard bits (8 for 32 bit registers) and these help solve the problem of overflow. Only the final result must be in format. This happens if the filter does not carry the gain. But we can still configure the ALU so that if overflow occurs,<sup>29</sup> we can obtain a saturation-type functioning.

FIR filters in linear phase have a central coefficient symmetry, so that the number of products can be divided by two:

$$y(n) = \sum_{i=0}^{L-1} a_i \cdot x(n-i) = \sum_{i=0}^{(L-1)/2} a_i \cdot [x(n-i) + x(n-L+i+1)] \quad [6.55]$$

Each calculation step requires an addition, a product and an accumulation. Without an instrument to aid in carrying out these three operations in a single cycle,

<sup>29</sup> This means that an error results from a 2 complement.

we gain nothing from the symmetry feature, since it requires two instructions and thus two cycles. The C54x makes use of the FIRS instruction:

- total of the accumulation A in the accumulation B;
- multiplication of the accumulation A by a coefficient of a memory-based program;
- addition of the two samples of a memory-based program in the accumulation A.

### 6.4.3.3. IIR filters

An IIR filter is characterized by a transfer function in  $z$  of a rational fraction type:

$$h(z) = \frac{y(z)}{x(z)} = \frac{a_0 + a_1z^{-1} + \dots + a_Nz^{-N}}{1 + b_1z^{-1} + \dots + b_Nz^{-N}} \tag{6.56}$$

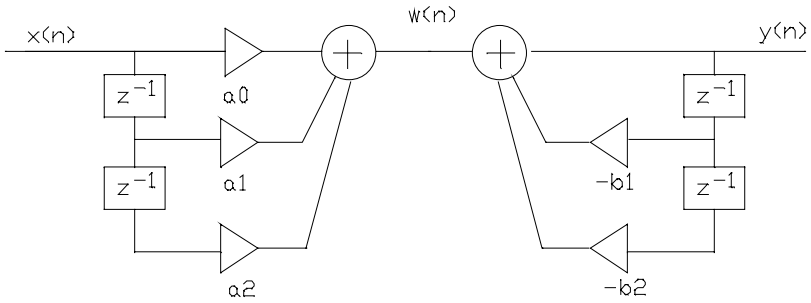
linked to a direct recurrent equation:

$$y(n) = \sum_{i=0}^N a_i \cdot x(n-i) - \sum_{i=1}^N b_i \cdot y(n-i) \tag{6.57}$$

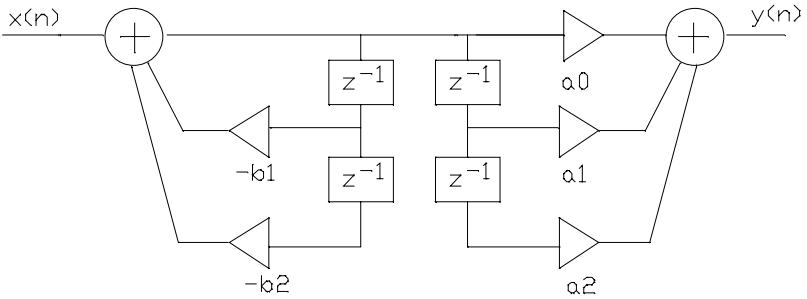
Studies on the quantification effect of coefficients show that we reduce calculation errors by decomposing the transfer function through a cascade of second order cells. We then pair the poles and the zeros:

$$h(z) = \prod_i \frac{a_{0i} + a_{1i}z^{-1} + a_{2i}z^{-2}}{1 + b_{1i}z^{-1} + b_{2i}z^{-2}} \tag{6.58}$$

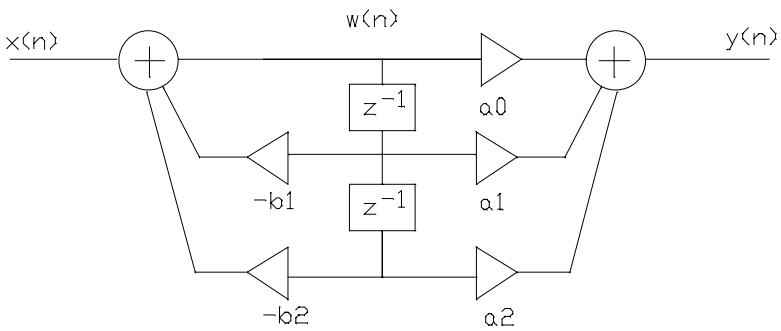
The most straightforward way of producing the effect described above on second order cells is shown in Figure 6.51, but since the analyses are linear, we can exchange them (see Figure 6.52), leading to the effect shown in Figure 6.53, which has the advantage of reducing the number of samples to be memorized.



**Figure 6.51.** Direct form of a second order cell



**Figure 6.52.** Equivalent form



**Figure 6.53.** Canonic form

If this solution is recommended by floating-point processors, it can be used with precaution for fixed-point processors. This is because the input of a cell can be



found outside format. In this case, we use a scale factor before each cell (Figure 6.54), such that:

- for the first cell:  $SF_1 = \max(|h_1(z)|_{z=e^{j\omega T_e}})$
- for the second cell:  $SF_2 = \max\left(\left|\frac{1}{SF_1} \cdot h_1(z) \cdot h_2(z)\right|_{z=e^{j\omega T_e}}\right)$

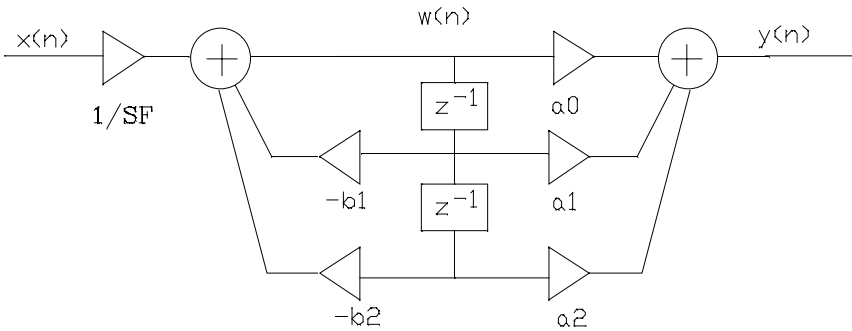


Figure 6.54. Second order cell with scale factor

The memory structure for a second order cell is given in Figure 6.55. Setting up the circular address (here, modulo 3) with masking creates the alignment of an address table to the addresses of the power of 2 (here 4).

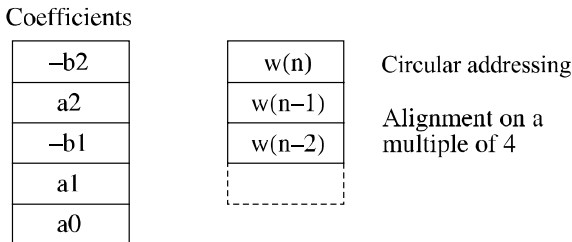


Figure 6.55. Memory structure for a second order cell

This structure allows for a simultaneous calculation of the products of the numerator and denominator (see Figure 6.56). We no longer use circular addressing for the coefficients, since the tables are not of the same length. This means we must reinitialize the pointer at each call.

```

; R2 contains x(n), AR1 point -b2, AR2 point w(n-2)
the circular addressing is done modulo N

IIR2:  mpyf *AR1,*AR2,R0          ; -b2*w(n-2) -> R0
        mpyf *++AR1(1),*AR1--(1)%,R1  ; a2*w(n-2) -> R1

        mpyf *++AR1(1),*AR2,R0        ; -b1*w(n-1) -> R0
        ||addf R0,R2,R2                ; x(n)-b2*w(n-2) -> R2

        mpyf *++AR1(1),*AR2--(1)%,R0  ; a1*w(n-1) -> R0
        ||addf R0,R2,R2                ; x(n)-b2*w(n-2) -b1*w(n-1) -> R2

        mpyf *++AR1(1),R2,R2          ; w(n)*a0 -> R2
        ||stf  R2,*AR2++(1)%          ; memorization w(n)
                                           ; and time shift

        addf R0,R2                    ; a0*w(n)a1*w(n-1) -> R2
        addf R1,R2,R0                 ; a0*w(n)a1*w(n-1)+a2*w(n-2) -> R0

; result in R0

```

**Figure 6.56.** Algorithm for a second-order IIR cell ( $C3x^{30}$  syntax)

This sequence requires 7 words and 11 cycles.

For the cascading operation, we can use the structure shown in Figure 6.57. The empty compartment is necessary for the alignment of the multiple addresses of 4 tables. After calculating a cell, the pointer on the coefficients jumps by 4 in order to go from one table to the next.

<sup>30</sup> The assembler of the C3x makes the idea of parallel instrumentation explicit, here a “MAC” written as mpyf (operands) || addf (operands). But this corresponds to one instruction, coded on one word and carried out in one cycle, if the pipeline functions well.

Coefficients

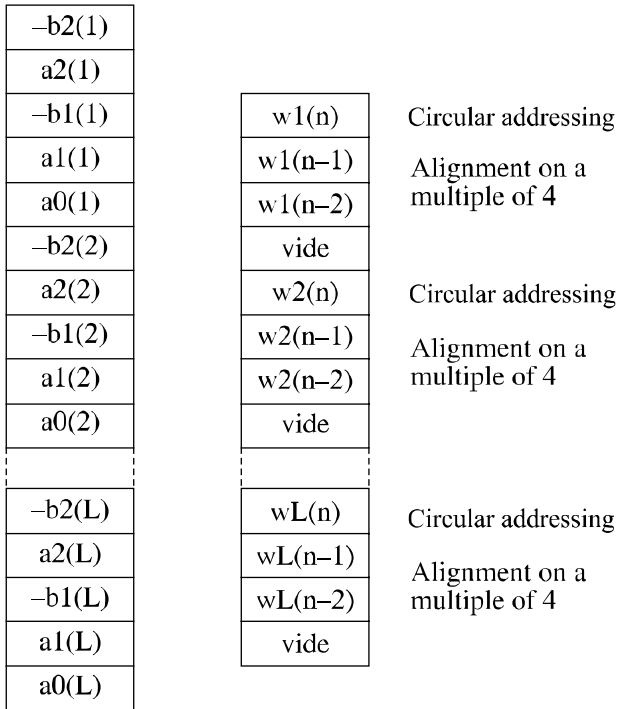


Figure 6.57. Memory structure for a cascade of second order cells

6.5. Conclusion

In this chapter, we have tried to show the efficiency and relative simplicity of digital signal analysis as carried out by modern converters and specialized processors. Their use for instrumentation purposes will continue to grow because manufacturers are increasingly offering “standard” microcontrollers with added signal analysis features. We thus have mechanisms with input/output that sometimes have defects on DSPs. As well, the increased capacities of FPGAs allow us to set up filters with order that are too high and coefficients coded on 8 or 12 bits. In this way, they can function at fairly high sampling frequencies, but we should not forget integrated digital filters that work with coefficients and data with significant word capacities.

## 6.6. Bibliography

- ANALOG DEVICE, Practical Analog Design techniques, Compte rendu du séminaire 1995.
- ANALOG DEVICE, Sigma-delta ADCs and DACs, Note d'application AN-283.
- AZIZ P.M. *et al.*, "An overview of sigma-delta converters", *IEEE signal processing magazine*, January 1996.
- BAUDOING G., *Les processeurs de traitement de signal, famille 320C5x*, Dunod.
- DE FATTA D.J. *et al.*, *Digital Signal Processing: A System Design Approach*, Wiley.
- HARRIS Semiconductor, A brief introduction to Sigma Delta Conversion, Note d'application AN9504, 1995.
- HERSCH R.D., *Informatique industrielle*, Presses polytechniques et universitaires romandes.
- IFEACHOR E.C. *et al.*, *Digital Signal Processing: A Practical Approach*, Addison Wesley.
- KUC R., *Introduction to Signal Processing*, McGraw-Hill.
- KUMAR M.S., *Digital Signal Processing: A Computer Based Approach*, McGraw-Hill.
- MARVEW C. *et al.*, *A Simple Approach to Digital Signal Processing*, Texas Instruments.
- MOTOROLA, Principles of Sigma-Delta Modulation for Analog to Digital conversion, Note d'application APR8/D, 1990.
- PARKS T. *et al.*, *Digital Filter Design*, Wiley.
- SENN P. *et al.*, "Convertisseurs analogique numérique CMOS à haute résolution pour les circuits VLSI audio", *L'écho des Recherches*, no. 153, 1993.

*This page intentionally left blank*

## Chapter 7

# The Contribution of Microtechnologies

### 7.1. Introduction

#### 7.1.1. *The vehicle: a system of complex, interdependent parts*

Optimizing the performances and respecting qualitative and quantitative rules of a complex system comprising a vehicle, its drivers and its passengers means acquiring and exploiting a large amount of dedicated data. These include the following:

- data having to do with passive security features, such as airbags, seatbelts with pretensioners, de-mistifiers, door closing indicators, levels, and quality of wheel-to-ground contact;
- data having to do with active security, including automatic driving and monitoring, anti-collision radar, ABS, and turning adjustment;
- passenger comfort data, such as temperature control, hygrometric degree and air quality;
- data relevant for control and transmission, including parametrization by cartography, electronic steering and steering wheel functions, servo-control of oil and water temperatures, and anti-pollution devices.

In this sense, these shared acquisition and analysis structures are superior to equivalent centralized systems in at least three ways:

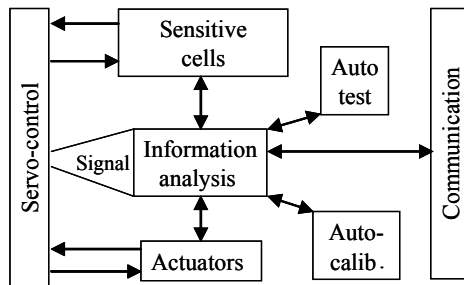
- in terms of reliability, because of redundancy and reduction of connective wiring;
- response time;
- flexibility.

These shared structures are based on data capturing, carried out as closely as possible to their source, accompanied by a local conditioning and pre-analysis. The goal of these operations is to make a quick decision, most often of the reflexive kind. The exchanges with the central unit, made through a serial transmission bus, are then reserved for slowly changing data analysis (door closing indicator, trunk, hood, tire pressure, different alarms, to name a few features). Some of these features require fairly complex calculations.

**7.1.2. Microtechnologies and microsystems**

The “neurons” of these shared systems must have perception mechanisms (sensitive elements) and eventually must be able to respond to the environment. They have the ability to analyze local information (conditioning, digitization, communication, auto-test and autocalibration, among others). They have access to an energy source (energy supply unit, cell and battery units, telesupply, latent energy microsource, to name several). These are very small in dimension, and are called *microsystems* (Figure 7.1). Ideally, *microsystems* are the natural result of extending microelectronic methodology to the collective manufacture of sensitive elements, called actuators and to the hybrid or monolithic integration of relevant electronics.

In practice, depending on a manufacturer’s abilities, microsystems are the final result of a goal of fulfilling market needs within economic constraints.



**Figure 7.1.** A microsystem or intelligent system

This is certainly true of the automotive sector, in which electronic innovation is a priority. Research costs must be absorbed incrementally even if they lead to fairly high production expenses. If the vehicle proves to be popular, the market for the product may extend to medium and low-end buyers, meaning lower production costs even if research costs climb, since these will be absorbed over a larger number of sales.

According to this schema, the first generation of a microsystem depends on the following:

- sensitive elements;
- more or less miniaturized actuators;
- an electronic mechanism made of discrete components;
- a programmable circuit that can be a microcontroller.

The above may be assembled on a primed circuit or on a CMS. *This constitutes an intelligent mesosystem.* Later, when the innovation has been absorbed and diffused, the sensitive elements and actuators must be optimized, miniaturized and integrated in hybrid or monolithic manner with the electronics. Then the ensemble can be called a *microsystem*. This new system can benefit from the inherent advantages of microelectronic or non-microelectronic technologies, such as:

- miniaturization;
- reliability;
- low production costs;
- insertion compatibility with traditional microelectronic units.

### **7.1.3. *Appropriate architectures for electronic microsystems***

The transition from intelligent mesosystems to microsystems is facilitated by using electronic architectures, adapted to these *specific development steps*.

These architectures must be linked to significant material adaptability (primed circuits, CMS, hybrid circuits, specific monolithic circuits, etc.). It should be unnecessary to question, at each change of technology, the underlying principles of these architectures. In other words, the many forms of electronics must remain identical in terms of system description.



These properties help us use the *experience gained from previous versions and apply it to the newer technologies*. They lead to higher profits for the manufacturer and better quality for the consumer through lower production and advertizing costs.

#### **7.1.4. Which examples should be chosen?**

To support the concepts presented in this long introduction, we will emphasize two strategies that are completely symmetrical. These concern microtechnology and electronic microsystems.

The technological aspects will be developed using the highly representative example of the car. We will explain how pressure microsensors that integrated with its MOS electronic device work, and will briefly describe how they are manufactured.

The electronic aspects will be discussed, along with some of the relevant architectures for the conditioning of signals coming from three different sensitive elements (capacitive and pressure piezoelectric cells, and capacitive acceleration cells). These are important because of their diversity and complementarity.

All the examples we have chosen come from the automotive field.

## **7.2. Microtechnologies**

Car motors have been affected by microtechnological advances since the 1970s, when monitoring pressure sensors appeared, and since the 1980s, with the development of accelerometers that released airbags. Today, apart from these two applications, microsystems are no longer employed in this general way. However, within the next decade their use may increase greatly, both in the replacement of older technologies and in new applications.

The general history of new automotive microsystems can be summarized as follows:

- 1960-1970: development and commercialization of first pressure sensors with a layer of silicon;
- 1970-1980: large-scale production by photolithography, extended use of micromachining, and Si/glass sealing;
- 1980-1990: new micromachining applications, related to both surface and scale, as well as development of new functions (for example, inertial

electromechanical microsystems (IEMS), temperature sensors, electromagnetic field, flow);

- 1990-2000: monolithic or hybrid integration of microsensors and electronics;
- 2000-2010: large-scale use of microsystems for automotive applications in the following systems:
  - security (airbags, night vision, anti-collision radar, etc.),
  - motor control (combustion, cylinder pressure, etc.),
  - transmission control (rotation speed, road condition information, etc.),
  - comfort control (humidity, microphone for vocal instructions, etc.).

These and other microsystems have developed because of the growing use of electronics, which has the higher calculation power necessary for the car of the future. In order to use the increasingly miniaturized electronic systems throughout a car, we also must be able to send these systems adequate information through sensitive elements or very small, even miniature, sensors.

However, a harsh environment can present a problem with using sensors in a car. High temperatures, shocks, vibrations, humidity, conditions that cause corrosion, electromagnetic interferences and radio frequencies can cause problems. This type of environment makes more demands on design and manufacturing. As well, production volume must be high (usually one million units or more) in order to absorb research and manufacturing costs and follow the market demand for new vehicles. The lifespan of vehicles must be at least 10 years/250,000 km and their prices must remain low. Generally, cars need the hardy qualities of military vehicles, at mass-market prices. These qualities (high production volume, low prices, reliability, and durability) are inherent to microsensors and microsystems.

The mass manufacture of many units at the same time, all with the same Si layers, functioning like integrated units, leads to very high production volume, overall low price and high reproducibility. The reliability of these microsystems is due to:

- expertise in manufacturing processes;
- using materials (especially silicon) with well-known (mechanical, thermal, electronic) properties;
- relatively simple assemblies with few units and few or no mobile parts.

Microtechnologies allow for integration within the same casing, even in monolithic form (that is, on the same Si layer) of the sensor and the electronic

mechanism of analog and/or digital analysis. These capacities can even extend to monolithic integrated circuits merging DSP and Si microsensors.

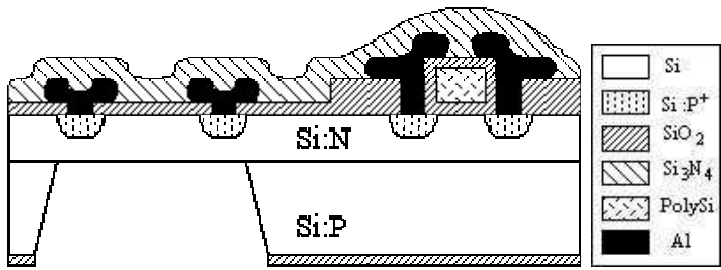
There are drawbacks to microsensors and microsystems. They are relatively costly and time-consuming to develop, and must be of high quality to be used in cars. This is why pressure microsensors take almost 40 years to develop and refine before they can be integrated monolithically and used in automotive applications.

In summary, microsystems are used in cars to reduce sensitive cell size, to lower their production costs, to improve their performances, and to integrate them simultaneously with their electronics and/or other microsensors. This is to enlarge the application of the cells, leading to the creation of an “intelligent car”.

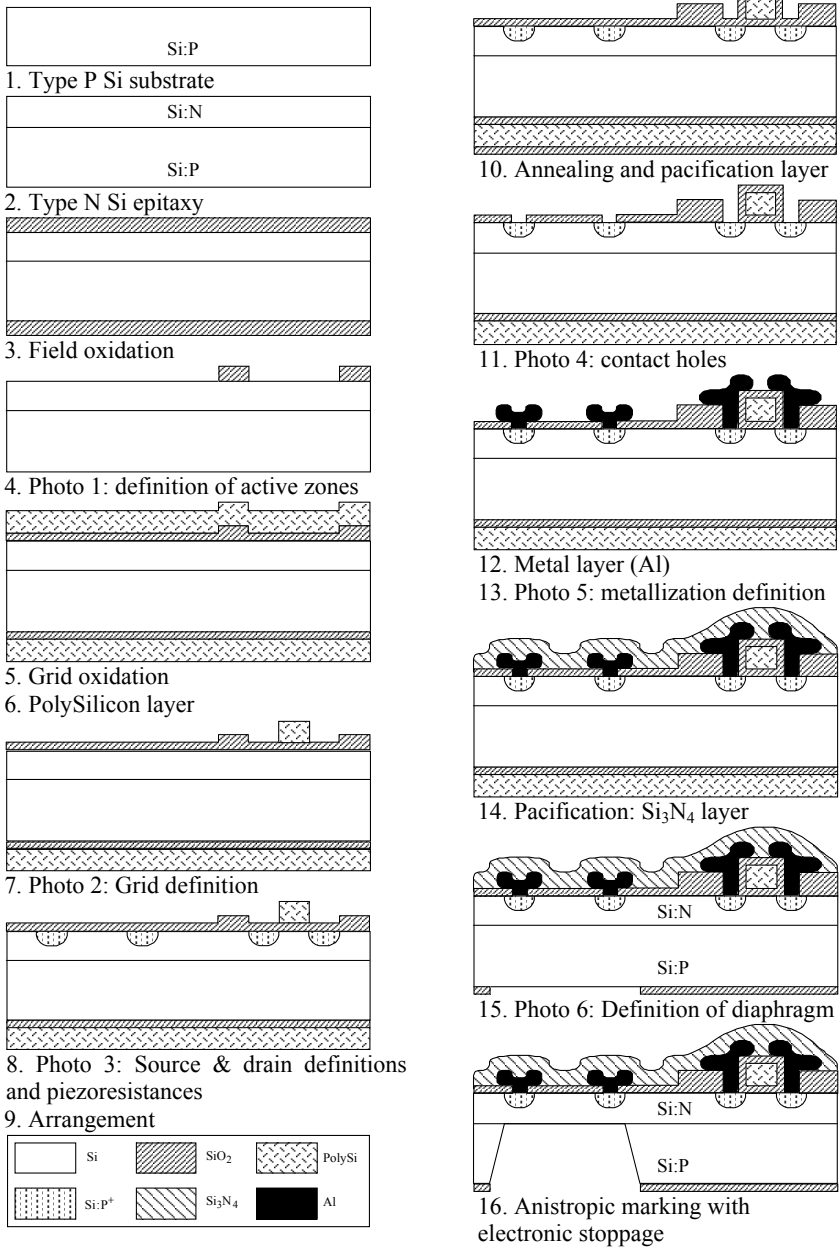
The technology behind the manufacturing of microsystems mostly comes from microelectronics. At each stage of the manufacturing of integrated circuits, the specific stages have been developed to build microstructures into the silicon, mainly by micromachining. Assembling and housing sensors can be complicated in comparison to integrated circuits because the sensor must be in contact with the external environment it measures. However, new technologies are being developed using materials other than silicon to extend the application field of microsystems.

As an example for introducing the different steps and technologies of microsystem manufacturing, we will discuss the main points of the steps of producing a piezoresistive Si pressure microsensor that is integrated with its type MOS electronics. For reasons of clarity, the manufacturing processes we discuss have been simplified and modified.

The Si pressure microsensor we describe has been schematized according to Figure 7.2. This is made of a thin Si layer of some piezoresistive microns that convert the mechanical signal (deformation of the layer due to pressure) into an electrical signal. The resistances change and a tension appears on a Wheatstone bridge.



**Figure 7.2.** Design of the principle (not to scale) of a piezoresistive microsensor in micro-machined silicon, integrated monolithically with a MOS transistor



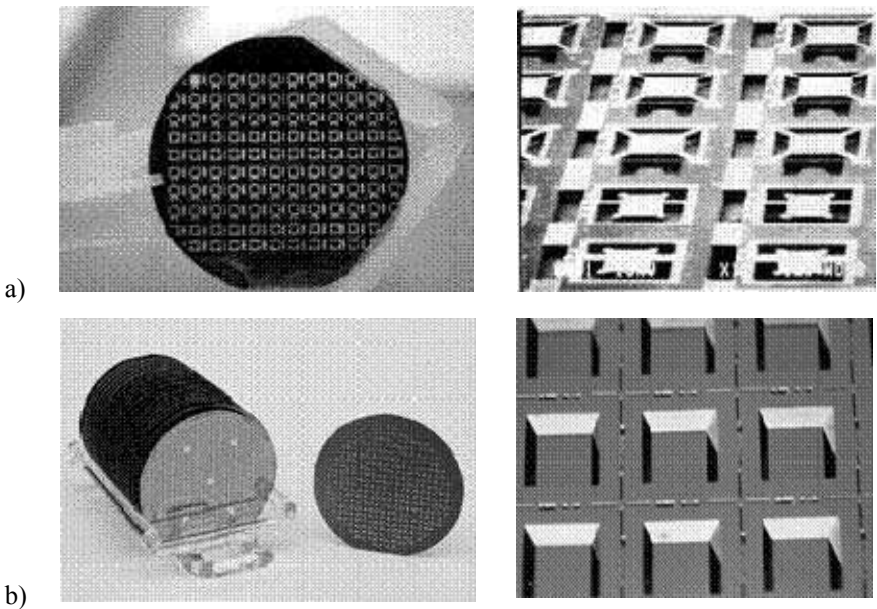
**Figure 7.3.** Principle of creating a piezoresistive pressure silicon microsensor integrated with its MOS electronics

Figure 7.3 shows all the necessary steps needed to produce this integrated pressure microsensor with its electronics. As the succession of steps show, the principle of producing integrated circuits and microsystems is to repeat, as many times as is necessary, the following two basic steps:

- remove or modify the thickness of the structure or material layer;
- use photolithography to establish the geometry of shapes.

The devices are made on an Si layer by successive stacking of layers with bidimensional patterns that are different but aligned with each other. By repeating these patterns all along the layer, many units can be made in parallel, all with the same specifications.

To illustrate this concept of mass manufacturing, Figure 7.4 shows how accelerometers and pressure microsensors are manufactured with Si layers.



**Figure 7.4.** Mass manufacture of a) micro-accelerators and b) pressure microsensors by using silicon (a: courtesy of ESIEE; b: courtesy of Auxitrol)

In the following sections, we will discuss, in order:

- technologies derived from microelectronics (section 7.2.1);

- technologies specific to Si microstructures (section 7.2.2);
- technologies developed for using materials other than Si in microsystems (section 7.2.3).

Our discussion will be limited to the general principles and essential points of these technologies. For a more detailed discussion, the reader can consult more specialized texts, including:

- for microelectronics [SZE 81], [GHA 94];
- for microstructures [SZE 94], [RIS 94], [GAR 94], [ELW 98] and [FUK 98].

### **7.2.1. Technologies derived from microelectronics**

#### *7.2.1.1. Si substrate*

Integrated circuits and silicon microsensors are mass-produced in monocrystalline substrate in the form of a disc (see Figure 7.3, step 1 and Figure 7.4). The Si layers are made from a monocrystalline bar obtained by Chzochalski's method. The bar is sawed, then polished to obtain a finished mirror.

According to the technologies used by different companies, the Si layers can be of varying diameters and different compositions. The diameters of layers vary from four inches or 100 mm in a research laboratory to 200 mm, even 300 mm in the production of integrated circuits. Their thickness is usually between 300 mm and 500  $\mu\text{m}$ . For integrated circuits, the layers are of type 100. For certain microstructures, we can use other types, often the 110. As well, the Si layers are doped with either type P or N in a regular and homogeneous way throughout their volume. The doping levels can vary from  $10^{13}$  atoms/cm<sup>3</sup> to  $10^{19}$  atoms/cm<sup>3</sup>, typically between  $10^{14}$  and  $10^{16}$  atoms/cm<sup>3</sup>.

The Si substrate provides:

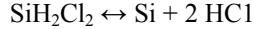
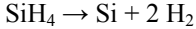
- mechanical and thermal support;
- active material for semiconductor devices;
- active material for microsensors and microactuators.

#### *7.2.1.2. Si epitaxy*

The Si monocrystalline substrate can be continued by the ordered or epitaxial layering of a thin Si layer (see Figure 7.3, step 2). Epitaxy helps control the active layer in which the devices are made. In addition, epitaxy enables the stacking of Si

layers of different dopings, since these cannot be created by diffusion or arrangement (see section 7.2.1.7).

Si epitaxy proceeds by a Chemical Vapor Deposition, or CVD by silane pyrolysis ( $\text{SiH}_4$ ) or chlorisilane decomposition ( $\text{SiH}_{4-n}\text{Cl}_n$ ):

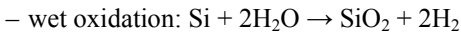
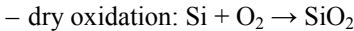


The Si atoms deposited on the surface must then organize themselves following the substrate atoms in order to form a perfect crystalline film. In general, this step is difficult to perfect technologically, and therefore is rather costly. In this case, we can use mechanisms to attain better performances. For these reasons, Si epitaxy is used mainly in bipolar and BiCMOS technologies, and optionally in some CMOS industries.

### 7.2.1.3. *Si thermal oxidation*

The widespread use of Si as a material in integrated circuits is due to several factors, including its abundance, thermal and mechanical properties, and in particular, its natural oxide,  $\text{SiO}_2$ .

When the Si layer is exposed to an oxidizing atmosphere at a high temperature, Si oxidizes on the surface, forming a layer of  $\text{SiO}_2$  according to one of the two following reactions:



Wet oxidation occurring with water steam ( $\text{H}_2\text{O}$  obtained by  $\text{H}_2$  with  $\text{O}_2$ ) is faster and helps us obtain thicker oxides. It can provide lateral isolation between the different mechanisms of a circuit (see Figure 7.3, step 3).

Dry oxidation with  $\text{O}_2$  is slower and results in layers of oxide that are thinner but of better quality. These oxides are used as grid dielectrics in MOS transistors, the building blocks of very high density integrated circuits such as VLSI-ULSI (Figure 7.3, step 5).

As microsystems,  $\text{SiO}_2$  is used as an isolation layers or for chemical properties different from that of Si.

#### 7.2.1.4. Photolithography

In order to manufacture integrated circuits and microsystems on a significant scale, the patterns that make up mechanisms are transferred to the wafer by photolithographic techniques whose main steps are shown in Figure 7.5.

In the manufacture of our pressure microsensor integrated with its MOS electronic unit, there are at least five steps of photolithography necessary to carry out across the sensor and circuit (Figure 7.3, steps 4, 7, 8, 11 and 13).

For each of these steps, a glass mask with opaque patterns in chrome has been designed from a computer-generated layout.

#### 7.2.1.5. Polycrystalline silicon layer

As soon as grid oxidation occurs with dry oxidation, a layer of Si is deposited to form the metallic grid of the MOS transistor (Figure 7.3, step 6). Because this silicon is deposited on the amorphous SiO<sub>2</sub>, the resulting layer is composed of a mosaic of small, random crystals. This polycrystalline silicon is called polysilicon (PolySi).

This PolySi layer is deposited by a technique called *Low Pressure CVD* by silane pyrolysis. It is also possible to deposit PolySi with types N or P doping by introducing certain gases: phosphine (Ph<sub>3</sub>) and diborane (B<sub>2</sub>H<sub>6</sub>) respectively.

#### 7.2.1.6. Etching

To transfer defined photoresin patterns on the Si wafer, photoresine acts as a protective mask for the parts of the wafer we want to retain (thick oxide in step 4 and PolySi in step 7). The rest of the wafer that is not protected is removed by etching.

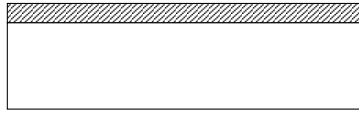
Two types of etching are currently used. These are listed below:

- Moist etching, an operation in which a liquid agent marks the different layers selectively, in relation to both photoresine and lower existing layers (see Figure 7.3, steps 4, 11, and 15, SiO<sub>2</sub> marking by HF; step 13, Al marking).

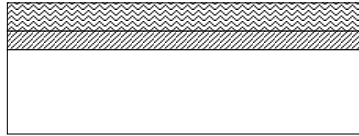
- Dry etching, an operation in which a plasma is used to mark the layer by physical effect (chemical reaction and conversion of the layer) or by a combination of two effects of reactive pulverization called *Reactive Ion Etching* (RIE) (see Figure 7.3, step 7 for PolySi etching; step 13 shows Al etching).



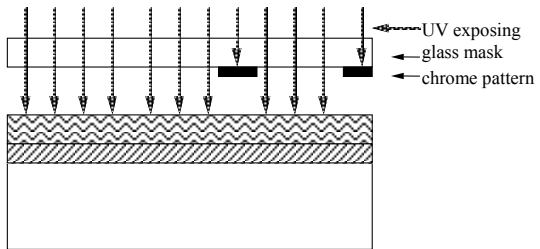
a. Starting structure: thick oxide for lateral isolation between circuit components



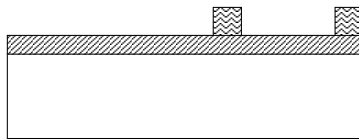
b. Photoresist application to the spin coater



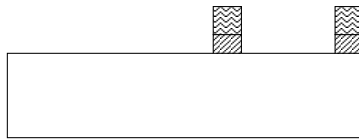
c. Aligning the glass mask with the chrome patterns and UV exposing



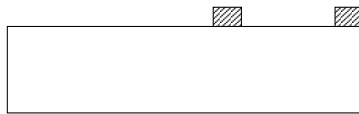
d. Development of exposed photoresist. The patterns of the mask are transferred into the photoresine



e. Patterns are transferred on to the wafer. By SiO<sub>2</sub> etching not protected by photoresist



f. Removal of the photoresist



**Figure 7.5.** Details of photolithography steps. Definitions of active zones for a circuit's MOS transistors and for a pressure microsensors (see also Figure 7.3, step 4)

Dry etching gives better control of the dimensions of etched patterns, as well as high reproducibility. Because of this, it is used in VLSI and ULSI technologies.

### 7.2.1.7. Doping

With conductors, we can modify and control their electric conductivity with the addition of miniscule amounts of impurities. This is called doping. When a semiconductor is doped type N, the current transfer is made by electrons, while with a semiconductor doped type P, the charge carriers are the positive holes. By juxtaposing zones of different dopings, we can make semiconductive electronic mechanisms such as PN junction diodes, bipolar transistors and MOS, thyristors, and optoelectrics such as light transmitting diodes, lasers, photodetectors and many others.

There are two doping techniques:

- The first technique is called ionic insertion (see Figure 7.3, step 9). The doped atoms are first ionized in an ion source, then extracted and accelerated electrostatically with energies of between a few keV to several MeV. They are then incorporated into the Si wafer (in a way analogous to a bullet being shot into a wall). The ionic insertion is a relatively violent physical phenomenon that creates many defects in the Si crystal. Therefore, an annealing at a high temperature (800 – 1,100°C) is necessary to repair the crystal and activate the doping. However, ionic insertion allows for very good control of the quantity, dose and profile of the dopage incorporated into the Si, respectively by measurement of incident current, selection of mass and ion energy. This explains why, despite its relatively high costs, insertion is widely used in VLSI and ULSI technologies.

- Predisposition is the second doping technique. During predisposition, the dopants are incorporated into the Si surface by placing the wafers in a high temperature environment that contains the dopants. To incorporate the dopants more deeply into the Si, the predisposition is followed by a diffusion annealing. Historically, predisposition was the first doping technique that made integrated circuits possible, and it is still used in certain industries, usually for microsystems. While predisposition is a very simple technique compared to ionic implantation, its drawback is its poor reproducibility, especially if the surface and/or the operating conditions are not well-controlled.

Once introduced, the doping profile can be modified by using the following steps. At high temperatures ( $T > 800^{\circ}\text{C}$ ), impurities diffuse significantly in the Si. Depending on the situation, this diffusion can be desirable to obtain an adequate profile for the diffusion annealing and/or oxidant (see Figure 7.3, step 10). Or this diffusion may be an inevitable interference that we try to limit as much as possible following a strict thermal regime.

### 7.2.1.8. Deposit of thin metallic and dielectric layers

In our example of a pressure microsensor and its MOS electronics, the mechanisms are made entirely in the Si at the end of step 10 in Figure 7.3. These mechanisms need to be protected from the exterior environment (from humidity or contamination) in order to ensure linkages and to have electronic access.

This protection or passivation is made with deposits of thin layers of dielectrics such as  $\text{SiO}_2$  (Figure 7.3, step 10) and  $\text{Si}_3\text{N}_4$  (Figure 7.3, step 14), or even other inorganic isolating materials (TiN and others), or organic materials such as polyimides.

There are two main techniques of depositing dielectric layers:

- Low-pressure CVD deposit (LPCVD) or carried out by a plasma (*Plasma Enhanced CVD*).
- Pulverization, in which a target of the material to be deposited is eroded by argon ions generated in a plasma. The atoms of the target are pulverized and settle on the circuit wafer.

To ensure the connections between the different parts of the circuit (or microsystems), and the environment, the metallic lines are defined by the wafer (Figure 7.3, steps 12 and 13). Several levels of metallic interconnections are sometimes necessary for complex circuits, in which case an electric layer is inserted between each level of metal for isolation.

Different metals can be used:

- for interconnecting lines, aluminum (Al) and its alloys (Al-Si-Cu), and more recently, copper (Cu);
- for local metallic contacts, tungsten (W) and metallic compounds such as silicides ( $\text{TiSi}_2$  and others).

The main depositing techniques for metals are the following:

- Pulverization, which has already been explained.
- Evaporation under vacuum, in which a vapor of atoms of metal is generated in a vacuum tank. The atoms condense on the “cold” substrate so that the water vapor condenses on a cold pane of glass.
- A small number of CVD deposits may also be used.

### 7.2.2. Technologies specific to microstructures

In our example of a pressure microsensor and its integrated MOS electronics, we stopped at step 14 of Figure 7.4 showing the creation of a circuit facing the Si wafer. Now a deformable membrane is necessary.

#### 7.2.2.1. Double face photolithography

Step 15 of Figure 7.3 is a step of “double face” photolithography. It allows us to align the back facing patterns with the front facing patterns. This leads to the next step of creating an Si membrane with correctly placed piezoresistances, thus ensuring maximum sensitivity.

#### 7.2.2.2. Volume micromachining

Liquid anisotropic etching was the first technique developed for making Si microstructures.

This kind of etching has Si erosion speeds that are highly dependent on the directions of the crystalline planes. This means the planes  $\langle 100 \rangle$  and  $\langle 110 \rangle$  are etched much more rapidly than the planes  $\langle 111 \rangle$ , which stay almost intact. The chemical agents that stimulate this anisotropic etching are inorganic alkalines such as KOH or organic solutions like EDP (ethylene diamine pyrocatechol), to name a few of the best-known. Table 7.1 gives these two anisotropic etching agents the etching speed of  $\text{SiO}_2$ . Figure 7.6 shows the details of the different steps of the process and the resulting microstructures.

Solution alkaline	Temperature (°C)	Etching speed in $\mu\text{m/h}$			
		Si $\langle 100 \rangle$	Si $\langle 110 \rangle$	Si $\langle 111 \rangle$	$\text{SiO}_2$
KOH: $\text{H}_2\text{O}$	80	66	132	0.33	$\leq 0.008$
EDP	110	51	57	1.25	0.004

**Table 7.1.** Main features of liquid anisotropic etchings for Si and  $\text{SiO}_2$  (from [GAR 94])

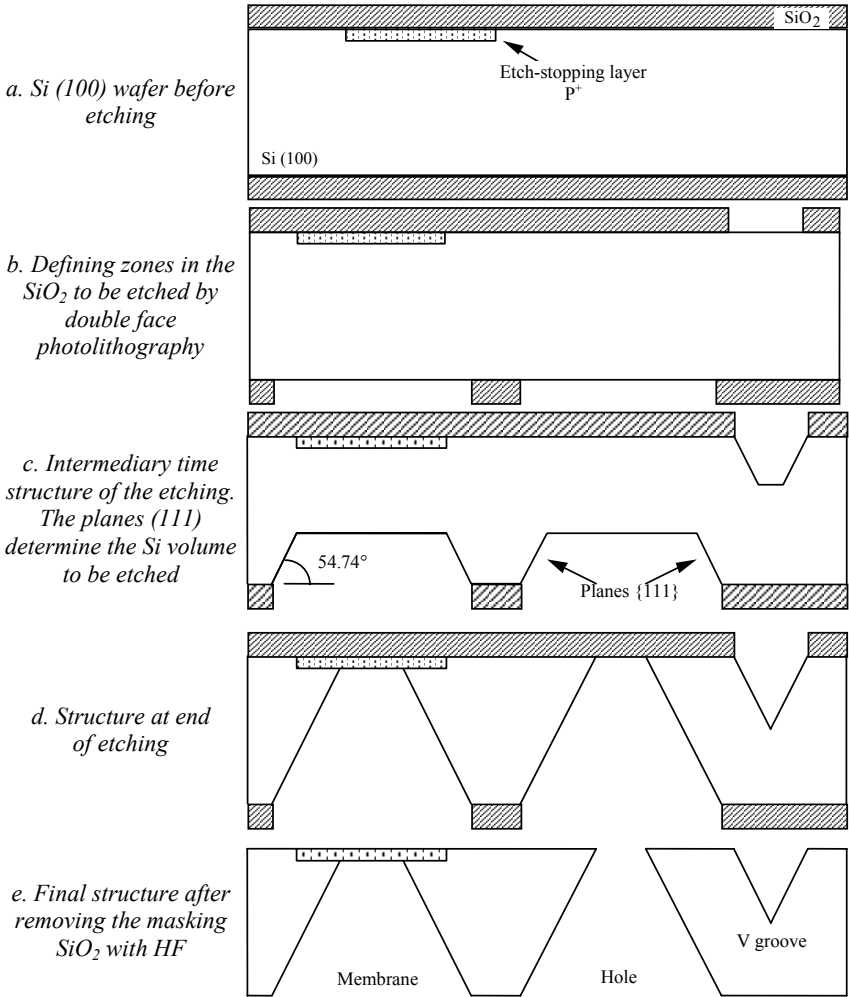
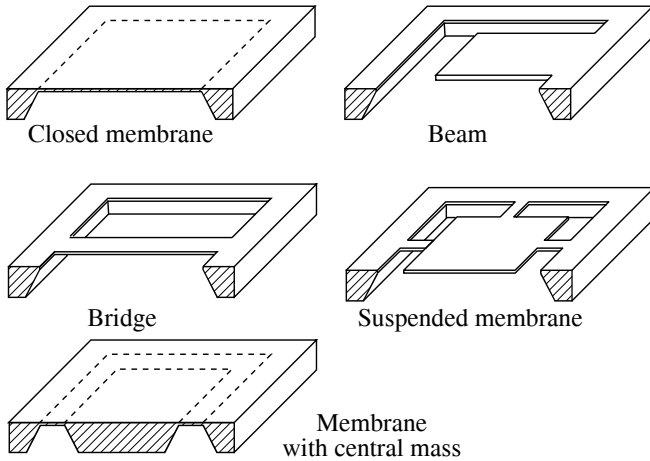


Figure 7.6. Details of anisotropic etching steps

To determine the Si zones to be etched, an SiO<sub>2</sub> layer can be used as a mask. However, in certain conditions, other materials must be used as masking, such as Si<sub>3</sub>N<sub>4</sub> or metals like chrome, which are not at all affected by KOH or EDP. According to the patterns established by double face photolithography, different structures can be micromachined and these are shown by Si <100> liquid anisotropic etching in Figure 7.7.



**Figure 7.7.** Examples of structure made possible by liquid anisotropic etching of Si (100)

To form the Si membranes, bridges and beams, we must not etch the entire thickness of the wafer. Instead the etching must be controlled, then stopped to leave the desired fine layer of Si.

The simplest way of controlling the anisotropic etching is to stop it after a certain time period, but because the structures of Si wafers are inherently inhomogeneous, requiring different etching speeds from one point to another, this technique is not sufficiently reproducible and reliable. In practice, four etch-stopping techniques are used. We will discuss them here.

Etch-stopping on an Si layer that has been highly doped with boron or Si:P<sup>+</sup> (see Figure 7.6 for how the membrane is made). The liquid anisotropic etching speed drops quickly for the Si doped with boron, with concentrations above  $3 \cdot 10^{19}$  atoms/cm<sup>3</sup>. The problem of layers that have been stopped with Si:P<sup>+</sup> is that the high boron concentration induces significant voltage constraints in the Si, modifying the mechanical properties of the membrane. It also stops the creation of semiconductive mechanisms in the layer, even if these are simple piezoresistances.

Electrochemical etch-stopping at PN junctions (Figure 7.3, step 15). The Si wafer is plunged into an electrochemical cell containing KOH or EDP. By adequate polarization of the contact potential of the junction (0.6 V), the N zone of the Si is protected by electrochemical potential. The anisotropic etching stops at the PN junction. This technique avoids the problems of mechanical constraints associated with high boron dopings and is compatible with certain procedures relating to integrated circuit technologies.

Etch-stopping at  $\text{SiO}_2$  or  $\text{Si}_3\text{N}_4$  protected by the opposite facing layer (see Figure 7.6d before HF piercing). The membranes or multi-layered beams of  $\text{SiO}_2/\text{Si}_3\text{N}_4$ ,  $\text{SiO}_2/\text{PolySi}/\text{SiO}_2$  and so on can be created in opposite layer after the volume of the subjacent Si wafer has been etched.

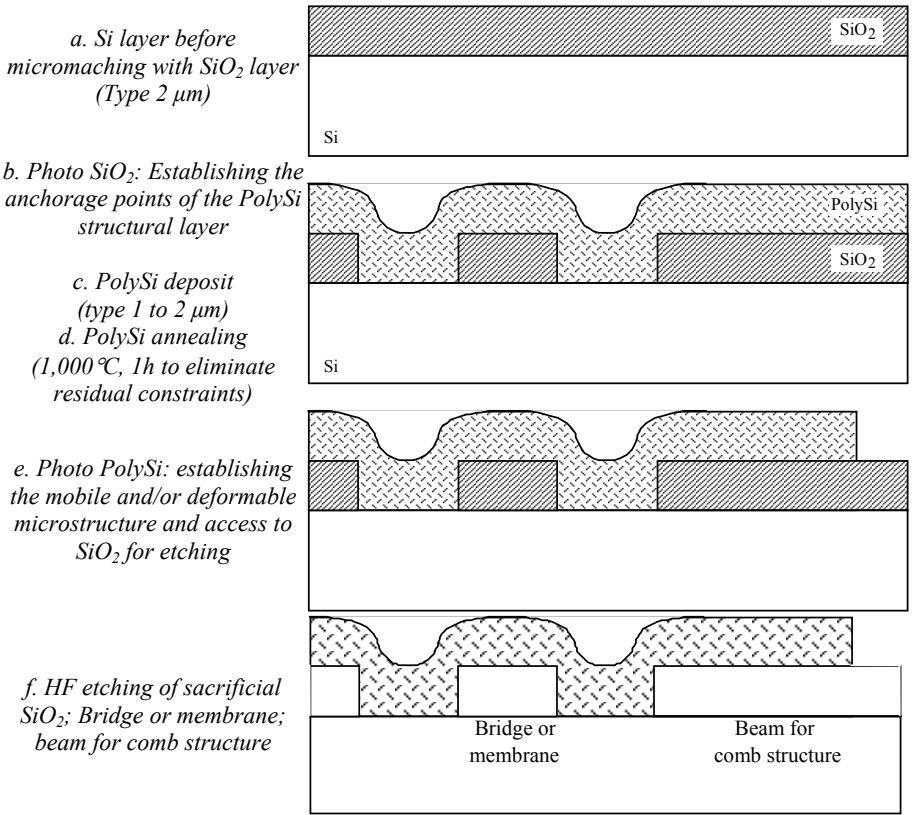
Etch-stopping at the  $\text{SiO}_2$  layer buried in an Si wafer on an isolator (*Silicon On Insulator*, SOI). Using different techniques, it is possible to make Si wafers in which an Si monocrystalline layer is separated from the rest of the substrate by a layer of buried  $\text{SiO}_2$ . These SOI wafers allow for the creation of Si microcrystalline microstructures and their electronics. These have high design flexibility and very good performance value, both mechanically and electronically. However, they need a technological network of integrated SOI circuits in which to develop.

### 7.2.2.3. Surface micromachining

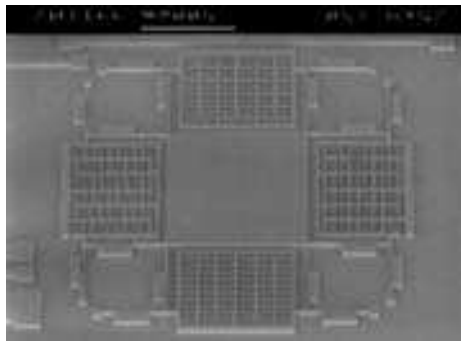
Large-scale micromachining by liquid anisotropic etching of Si is not always compatible or easy to integrate with the manufacture of electronic circuits. At the beginning of the 1980s, a micromachining technique was developed using only manufacturing processes such as depositing, etching, all parts of VLSI technologies.

The main principles of this surface micromachining technique is shown in Figure 7.8. It depends on etching a sacrificial layer, usually  $\text{SiO}_2$ , to free a mobile and/or deformable microstructure that is usually PolySi. By using several sacrificial structures and structures stacked on top of each other with adequate patterns defined by photolithography, it is possible to make microstructures with one, two or three consecutive layers. We see that it is also possible to use SOI wafers in which the buried oxide and the surface Si constitute the structural layer.

Accelerometers and above all other inert sensors (like the microgyrometer shown in Figure 7.9) can be created on the surface of the Si wafer with their electronic command and analysis integrated monolithically. In general, the principle of detection and/or excitation is capacitive, the different Si or PolySi levels being conductors. These microstructures are called *micro electro-mechanical systems* (MEMS).



**Figure 7.8.** Surface micromachining using sacrificial layers



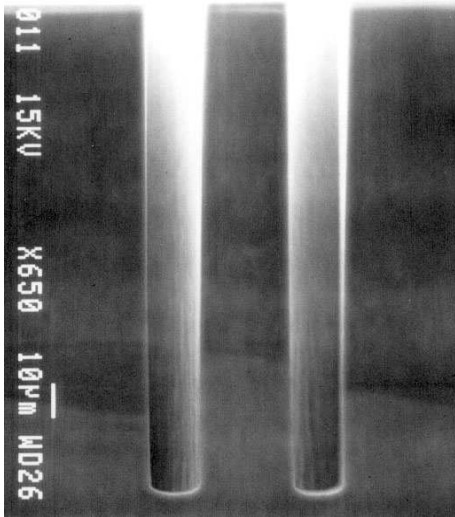
**Figure 7.9.** Microgyrometer made by surface micromachining (source: IEF [VER 99])



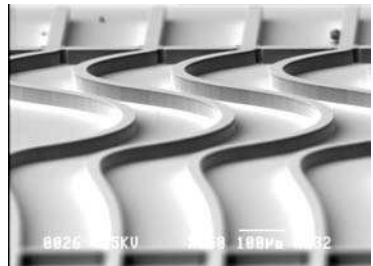
7.2.2.4. Micromachining by deep anisotropic dry etching

Recently, following the example of VLSI etching technologies, dry etching techniques have been developed that are used with or instead of moist etching.

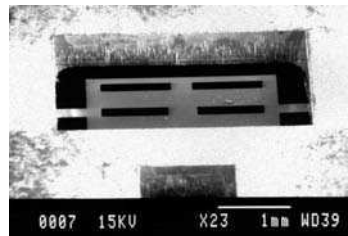
In particular, deep anisotropic etching devices have appeared (Deep Reactive Ion Etching (DRIE)). With these, a dense plasma and specific operating conditions (having to do with the nature of gases, wafer temperature, etching sequencing) are conducive to high-speed etching Si carried out vertically and very deeply (Figure 7.10, example a). The form of these etched structures no longer depends on patterns defined by photolithography; and this leaves complete freedom in designing the microdevice (Figure 7.10, example b). In using the masking layers and adequate stopping layers, the Si wafer can be locally etched, either partially or completely, as is shown in the examples of Figure 7.10.



a. Test structure for deep anisotropic etching of holes 15  $\mu\text{m}$  in diameter at a depth of 150  $\mu\text{m}$



b. Etched microstructures in Si on a thickness of 40  $\mu\text{m}$



c. Supported and pierced membrane of 15  $\mu\text{m}$  thickness created by etching across the entire Si wafer

**Figure 7.10.** Examples of deep anisotropic dry etching (source: ESIEE)

### 7.2.2.5. Heterogenous assemblies

In general, microstructures created in or on an Si wafer are not part of a device. This means that in our example of a pressure microsensor, if we want to an absolute pressure sensor, we must close the cavity created by the anisotropic etching of the Si by making this cavity into a sufficiently large empty space. For pressure microsensors with Si membranes of capacitive type, we assemble a counter-electrode on a rigid support that is linked with the Si wafer.

Depending on the application field, different heterogenous assembling technologies can be used for the complete assembling and usage of the microsensor. The main features are summarized in Table 7.2.

<b>Bonding</b>	<b>Temperature</b>	<b>Bonding the Si wafer with</b>
by collage	low 130-350°C	a layer of polymer resin, polyimides, epoxy, etc.
by forming an eutectic or metal alloy	400°C	a layer of gold or other metal
by glass, Si low temperature	100-600°C	a thin layer of phosphorus glass, with boron or fritted with Pb
anodic glass-Si	average 350-800°C	a substrate of sodium glass under polarization of  400 to 700 V
direct Si-Si	average ≥ 300°C	another Si wafer

**Table 7.2.** Principle characteristics of heterogenous assembling technologies and Si welding (from [GAR 94] and [SZE 94])

From these five technologies, the glass-Si anodic bonding was developed specifically within the framework of microsystems.

The glass-Si anodic bonding helps us assemble an Si wafer with a sodium glass wafer that can be micromachined, or it can have metallic patterns for the electrodes. After having connected the two wafers, the ensemble is placed at a moderate temperature (350-500°C), at which the sodium ions become mobile in the glass. By

applying a voltage of between 400 and 700 V according to the temperature, the sodium ions leave the interface zone. This produces a significant electrical field and generates high electrostatic pressures that are sufficient to establish very close contact between the two wafers, even when their flatness is not perfect. The bonding takes several minutes, usually with the formation of a very fine layer of  $\text{SiO}_2$  as a connecting layer. Today, the reliability and reproducibility of anodic bonding means it can be used in industrial settings, especially in the assembling of Si pressure sensors. However, the use of high voltages is not always compatible with the presence of integrated circuits on the same wafer.

Direct Si bonding is a more recent development that makes it possible to put together two Si wafers without an intermediate layer. The two Si wafers are cleaned according to correct procedures and then brought into contact. At this stage, the wafers have already been adhered using weak Van der Waals forces. They are then transferred to a furnace set at a relatively high temperature ( $700^\circ\text{C}$ ) to carry out the final bonding, which probably occurs due to Si-O bonds. Direct Si-Si bonding is especially useful for constructing Si microstructures and is one of the techniques used in making SOI layers. Its disadvantage is that the high temperature required for the bonding process means integrated circuits cannot be present on the layers.

### 7.2.3. *Beyond silicon*

To end this discussion on microsystems, we will briefly mention other possible technologies:

- these can be based on materials other than Si, either in the substrate or in the thin layers. If Si is still used, it is usually as a mechanical substrate with the possibility of the electronic circuit being integrated with it;

- techniques other than integrated circuit technologies can be used, such as techniques used with more “macroscopic” materials.

In addition, other substrates can be used. Some of these are listed below:

- silicon-carbon (SiC) can be used in creating microsystems in harsh environments (high pressure, high temperature, corrosion);

- gallium-arsenide (GaAs) and, less often, indium phosphide (InP). Both are III-V mixed semiconductors used for HF and electronic devices and circuits.

New microsystem functions can be created by using active materials, often deposited in thin layers, such as:

- piezoelectric materials like ZnO, AlN and PZT;

- mechanical materials like metals (Cr and W, among others) and shape memory alloys (Ti/Ni and others);
- magnetic materials (NiFe, CoFe and others);
- thermoelectric materials (SiGe, Bi<sub>2</sub>Te<sub>3</sub>);
- inorganic chemical materials (SnO<sub>2</sub>, metallic oxides and others) and organic chemical materials (polymide, polypyrroles and phtalocyanines).

Lastly, a range of micromachining techniques enable us to make microstructures and microdevices. Some of these are:

- the LIGA process, a micronic scale molding technique;
- lazer-beam micromachining;
- micromachining by focalized ion beams (Focused Ion Beam Milling (FIBM));
- micromachining by electrostatic discharges (Electro-Discharge Machining (EDM) [FUK 98]).

### 7.3. Electronic architectures and the effects of miniaturization

#### 7.3.1. Overall trends

Before beginning the description of selected architectures, we will remind the reader of certain general ideas that will help in understanding the reasons for our selection.

First of all, whatever the mechanical, physical, chemical or biological variables to be measured, and whatever conversion mechanisms are used, *the output signals of the sensitive elements are almost always electrical* – voltage, current, charge, variation resistance or capacity – with a marked tendency towards *capacitive variation*, linked to miniaturization.

Moreover, the robust architectures, as regards the material variability, must have the following characteristics.

- they must not be significantly affected by the length of their connections to the sensitive cells and by the majority of the interference capacities. This suggests that low impedance inputs of the virtual mass type;
- they must be sensitive to the single performances of a minimum number of components, especially to the sensitive element. This means that a *feedback loop* must be used systematically in order to reduce significantly the influence of the component variability that is part of the direct chain. Then we proceed to

*measurement by zero method.* In the absence of nonlinearities, the output signal or electrical image of the quantity to be measured is then proportional to the relation between the value of the measured quantity and a reference value. The proportionality coefficient is a stable electrical quantity.

The architectures must proceed to an early digitization of the measurement signal. This means the A/N conversion is incorporated into the feedback loop that is part of the basic structures of  $\Sigma$ - $\Delta$  modulators.

This last concept is only valid for transducers with a relatively slow response speed (limited to around 10 kHz). As for the how the process is carried out, this value rarely is an obstacle. In a vehicle, however, this limitation prohibits the use of a  $\Sigma$ - $\Delta$  modulator in applications requiring a dynamic measurement of the steering wheel angle. Conversions of the “flash” or “weighted” types that occur outside the loop are, in this case, preferred.

In addition, these architectures must be dynamically reconfigured so that functioning modes, such as measurement, calibration and autotest, may be changed on demand or automatically.

*Calibration* is a programming operation whose purpose is the storing of numerical values specific to each microsystem. Initially carried out collectively in order to reduce cost prices, this procedure can be renewed, for certain microsystems, throughout the life of the vehicle.

*Autotest*, either total or partial (we will look at the accelerometer for an example of this difference), must be at all times superimposed on the measurement, which itself must remain fully operational.

In the case of a disfunctioning microsystem (an intelligent mesosystem), the central computer of the vehicle must be informed of the problem in order to proceed to the implementation of a degraded functioning mode for the ensemble. An airbag failure, for example, must be signaled to the driver, who can then adjust to this new situation.

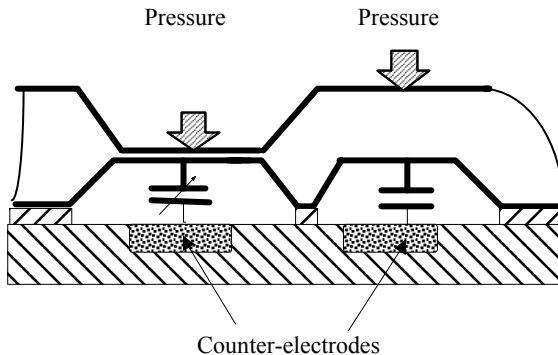
On a completely conceptual order, the electronic architectures must either overcome the effects of miniaturization or use them. This is especially true of electronics associated with sensitive cells for inert variables (acceleration, angular speed, to name a few), with capacitive detection. The electrostatic forces brought about by the processing of the measurement signal are of the same order as for cells of mechanical origin.

As for noise, *miniaturization degrades the signal-to-noise ratio of sensitive cells* in relation to that of larger cells. The noise of sensitive cells in microsystems remains low and requires low noise architectures to conserve reasonably good resolutions.

Lastly, taking into account these important objectives: material flexibility; low-cost integration of analog and digital parts on the same chip with CMOS or BICMOS technology; low sensitivity to supply voltages; and good low noise frequency functioning. Without a doubt, switching capacity techniques are the best means of carrying out these architectures.

### 7.3.2. Conditioning electronics for capacitive cells that are sensitive to absolute pressure

A capacitive cell sensitive to pressure has two capacities. One is variable according to the pressure  $C_{mes}(P)$ , and the other,  $C_{ref}$ , is not. A possible process (see Figure 7.11) consists of linking two identical counter-electrodes that are both of millimetric dimensions. These are then diffused in a silicon substrate of reversed doping to a thickened, electrically conductive membrane to the right of one of them. This membrane is connected to the substrate by means of a *spacer*, of thickness  $d_0$ , made of isolating material that surrounds the counter-electrodes. The holes formed by this are empty of atmosphere, and constitute a capacitive cell that is sensitive to absolute pressure. This is because, being far from the spacer, the thin membrane deforms under the effect of pressure. Independent input signals can be applied to the counter-electrodes that are, practically speaking, electrically isolated from the substrate by well-polarized PN junctions. The output signal can be extracted from the membrane. This type of sensitive cell is used in measuring barometric pressure for regulating motors.



**Figure 7.11.** Capacitive pressure sensor with a differential structure

7.3.2.1. Measurement principle

In the absence of pressure:  $C_{ref} = C_{meso} = \frac{\epsilon o \cdot S}{do}$

An approximation of the piston plan (see Figure 7.12a)  $C_{mes}(P)$  is expressed:

$$C_{mes}(P) = \frac{C_{meso}}{1 - \frac{P}{P_{max}}}$$

where  $P_{max}$  is the pressure at which the thin membrane touches the substrate;  $P$  is in principle always below  $P_{max}$ .

The voltage:  $V_s = V_{ref} \left[ 1 - \frac{C_{ref}}{C_{mes}(P)} \right] = V_{ref} \frac{P}{P_{max}}$

is then a “good” expression of the pressure to be measured.

However, this model is not very realistic, since it does not take into account the embedment effect that immobilizes the thin membrane, keeping it to the limits of the enclosure. This gives us the new expression of  $C_{mes}(P)$  (see Figure 7.12b):

$$C_{mes}(P) = C_{offset} + \frac{C_o}{1 - \frac{P}{P_{max}}} \quad \text{with:} \quad C_{offset} + C_o = C_{ref} \quad [7.1]$$

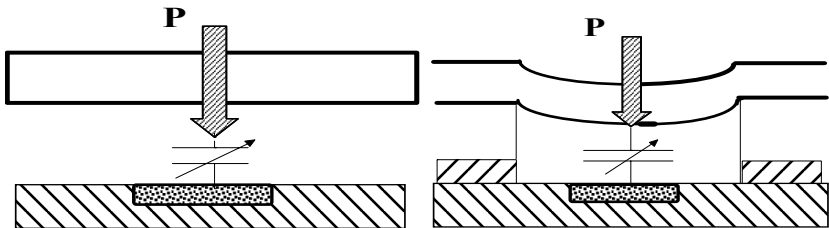


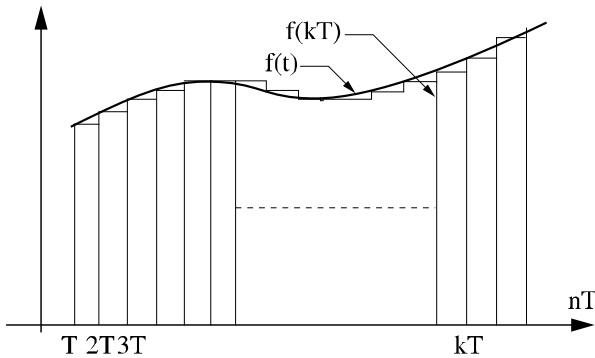
Figure 7.12. Effect of pressure on the membrane: two models

We will keep the formula shown in [7.1] as the definitive expression of the capacity dependent on the pressure. To obtain a measurement voltage linearly linked to the pressure, it must be capable of calculating:

$$V_{mes} = V_{ref} \left[ 1 - \frac{C_{ref} - C_{offset}}{C_{mes}(P) - C_{offset}} \right] = V_{ref} \frac{P}{P_{max}} \quad [7.2]$$

switching capacity techniques allow us to easily carry out the above formula.

We note here that this technique is based on analysis of analog signals sampled periodically (Figure 7.13). It requires discrete time circuits made of capacities, switches, and amplifiers sequenced at the sampling frequency  $F_e$  by a clock  $\Phi$  (in general of  $\frac{1}{2}$  cyclic ratio) of period  $T$ .



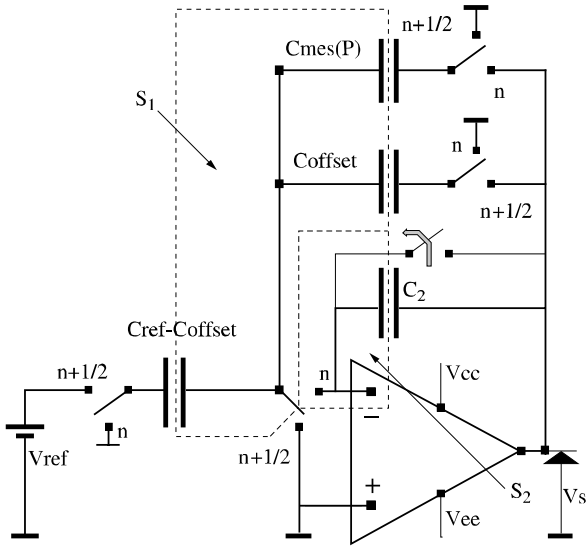
**Figure 7.13.** Discretization of an analog signal

### 7.3.2.2. The analog version

The electrical schema with two supply sources (+Vcc, -Vee) is represented in Figure 7.14. Equation [7.2] simply uses the analysis of this structure by expressing, step by step, the sharing of charges and their relations with reference voltages to the common point of energy supplies. As well,  $\Phi$  and  $\bar{\Phi}$  are two complementary clocks and are non-recoverable. They are assumed to be active at high states (switch closing demand):

- high state of  $\Phi$ : ]  $nT, (n + 1/2)T$  [; low state of  $\Phi$ : ]  $(n + 1/2)T, (n + 1)T$  [
- low state of  $\bar{\Phi}$ : ]  $nT, (n + 1/2)T$  [; high state of  $\bar{\Phi}$ : ]  $(n + 1/2)T, n(n + 1)T$  [





**Figure 7.14.** *Switching capacities structure, analog version*

Here, the amplifier is ideal (infinite gain in open loop and infinite pass band); the charges are always calculated on the plaques that can be part of electrostatically isolated systems. To be able to apply the Z conversion,  $C_{mes}(P)$  will be assumed to be invariant in time [BAI 94-1].

In the architecture shown above, the two capacities  $\{C_{ref} - Coffset\}$  and  $Coffset$  are very easily produced. This is done by dividing the counter-electrode from the capacitive structure that is insensitive to pressure into two electrically independent parts (see Figure 7.11). There are two subsystems:  $S1$ , which has the three capacities  $Coffset$ ,  $\{C_{ref} - Coffset\}$ ,  $C_{mes}(P)$ ; and  $S2$ , which is made of  $C_2$ .

7.3.2.2.1. Switching capacities integrator: the first phase (see Figure 7.15)

The switch that short-circuits  $C_2$  only has an initialization function. In normal functioning, it stays permanently open.

Interval ]  $(n - 1/2)T, nT$  [:

$$(Q_{S1})_{(n-1/2)T} = -(C_{ref} - Coffset)V_{ref} - Coffset \cdot V_S((n - 1/2)T)$$

$$(Q_{S2})_{(n-1/2)T} = -C_2 \cdot V_S((n - 1/2)T)$$

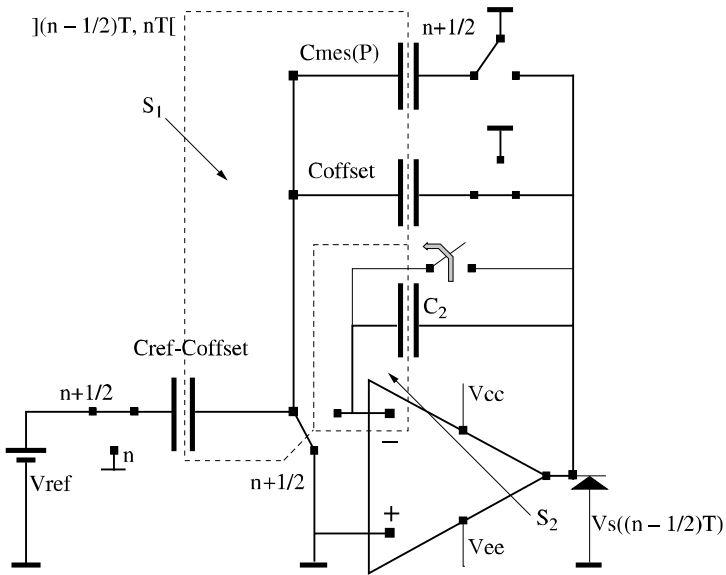


Figure 7.15. Precharge

7.3.2.2.2. Switching capacities integrator: second phase (see Figure 7.16)

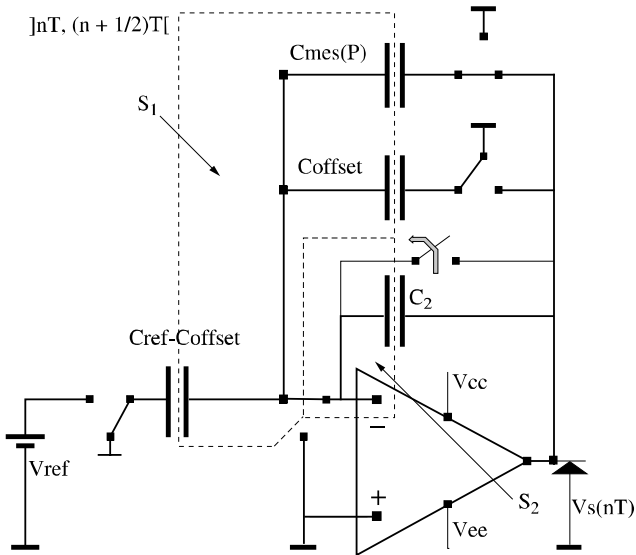


Figure 7.16. Transfer

Interval ] nT, (n + 1/2)T [:

$$(Q_{S1})_{nT} + (Q_{S2})_{nT} = (Q_{S1})_{(n+1/2)T} + (Q_{S2})_{(n+1/2)T}$$

Or, after explicit development:

$$-Cmes(P)(V_S(nT)) - C_2 \cdot V_S(nT) = -Coffset \cdot V_S((n - 1/2)T) - C_2 \cdot V_S((n - 1/2)T) - (Cref - Coffset)Vref$$

Interval ] (n + 1/2)T, (n+1)T [:  $V_S((n + 1/2)T) = V_S(nT)$

Also, the recurrent equation of the system is expressed as:

$$V_S((n+1)T) \left( 1 + \frac{Cmes(P)}{C_2} \right) = V_S(nT) \left( 1 + \frac{Coffset}{C_2} \right) + \left( \frac{Cref - Coffset}{C_2} \right) Vref \tag{7.3}$$

From the transfer function in “Z”, we deduce the recurrence equation:

$$\frac{V_S(Z)}{Vref(Z)} = H(Z) = \frac{Cref - Coffset}{C_2 + Cmes(P) + (C_2 + Coffset)Z^{-1}}$$

It is easy to establish the stability condition of such a system, knowing that:

$$Coffset < Cmes(P) \tag{7.4}$$

a constraint which is always satisfied when the substrate is maintained at a constant potential.

If the value of  $C_2$  conditions the response time of the device, the sensitivity of the sensor, on the other hand, is completely independent of the choice, since the relation that links the voltage  $V_S$  to the pressure is expressed in stabilized regime:

$$V_S((n+1)T) = V_S(nT)_{n \rightarrow \infty} = Vref \left( \frac{Cref - Coffset}{Cmes - Coffset} \right) = Vref \cdot (1 - P/Pmax) \tag{7.5}$$

Here, we see the advantage of the feedback loop that eliminates the variability influence of the components inside the loop. It also enables the use of a zero method

that generates a signal in the form of a product of a stable value of the chosen electrical variable, here voltage, using the relation between the measured variable value and the reference value. By taking  $V_S$  from  $V_{ref}$ , we get  $V_{mes}$  (equation [7.2]). This subtraction operation, not shown in the schema, is easily carried out with a switching capacities amplifier. The voltage  $V_{mes}$  can then be digitized with an A/N converter used at the end of the chain.

7.3.2.3. Basic first order  $\Sigma$ - $\Delta$  modulator with a one-bit quantifier

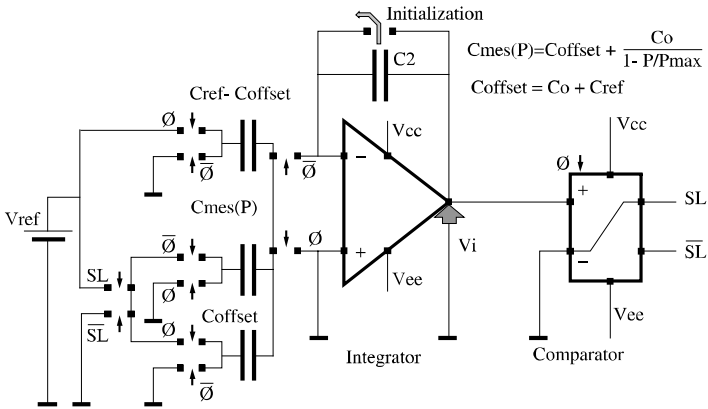


Figure 7.17.  $\Sigma$ - $\Delta$  modulator used for pressure measurement

Let us look again at the example shown above, substituting the discrete analog feedback  $\{-V_S(nT)(C_{mes} - C_{offset})\}$  with a quantified feedback:  $\{0, -V_{ref}(C_{mes} - C_{offset})\}$ , driven by a conditional clock (see Figure 7.17). As we will see, we keep the advantages of the zero method by proceeding to digitization by using some of the properties of first order  $\Sigma$ - $\Delta$  modulators with one bit [BAI 96].

Our discussion will be in two stages:

- first, in order to clarify principles, we will discuss a qualitative approach based on a hydraulic analogy;
- then we will present a quantitative approach based on recurrent equations that show the advantages of this technique in terms of precision and “precision/response time” compromises.

7.3.2.3.1. The qualitative approach

Employing one-bit  $\Sigma$ - $\Delta$  modulators for pressure measurements with the help of capacitive sensors makes use of a hydraulic analogy. With this analogy, we can

easily understand the relevant principle. In effect, we compare (see Figure 7.17) the capacities  $C_{ref} - C_{offset}$  and  $C_{mes} - C_{offset}$  to the tub  $v_{ref}$  at a liquid level. The integrator then becomes a tub, and the comparator (a one-bit quantifier) is a measurer of the logical output level ( $SL = 0/1$ ). In this *direct mode*, at each clock cycle, ( $\Phi$ ) the tub is refilled with the help of  $C_{ref} - C_{offset}$ , which is always smaller than  $C_{mes} - C_{offset}$ . When the liquid level in the tub rises above a certain level, it is partly emptied with the help of  $C_{mes} - C_{offset}$  by means of commanded commutators, especially by the conditional cycle  $SL(\Phi)$ . During  $N$  clock cycles, the *number* of times it makes use of  $C_{mes} - C_{offset}$ :

$$\sum_{i=0}^{i=N-1} SL(i)$$

thus also expressing the maintenance of a constant level in the tub:

$$N \cdot V_{ref} (C_{ref} - C_{offset}) - \sum_{i=0}^{i=N-1} \{SL(i) \cdot V_{ref} (C_{mes} - C_{offset})\} = 0$$

The mean  $\frac{1}{N} \sum_{i=0}^{i=N-1} SL(i)$  represents this quantity with a fractional number between

{0 and 1}, which can be linked to the *numerical pressure measurement*:

$$number_d = \frac{1}{N} \sum_{i=0}^{i=N-1} (1 - SL(i)) = \frac{1}{N} \sum_{i=0}^{i=N-1} \overline{SL(i)} = 1 - \frac{C_{ref} - C_{offset}}{C_{mes} - C_{offset}} = \frac{P}{P_{max}} \tag{7.6}$$

to the near quantification noise.

There can be errors when this schema is being carried out. These can occur when the result of the transferred charges in the integrator by charge injection (of liquid, for example) due to the clocks, by the interference capacities that are part of the circuit is not zero “ $number_d$ ”. To eliminate this systematic error, we can proceed to a new sequence in *reverse mode*. In this case, the tub is emptied at each clock cycle ( $\Phi$ ) by  $C_{ref} - C_{offset}$  and is refilled, under the effect of the conditional clock  $SL(\Phi)$  with the help of the capacity  $C_{mes} - C_{offset}$ . By alternating the two modes, we see:

$$number_d = \frac{P}{P_{max}} + err \dots \dots \dots number_i = \frac{P}{P_{max}} - err \tag{7.7}$$

The error is naturally eliminated by the addition of the two direct and reverse modes. It disappears easily through a simple decimation filtering.

Quite often, we must be content with a lower quality correction, with the goal of simplifying the architecture. The cause of errors can be assimilated to a capacity that is added to or subtracted from  $Cref - Coffset$ . This means it is enough to add or subtract a physical capacity of the same value to find the quantity “number” we need to find.

$$number = 1 - \left[ \frac{\{(Cref - Coffset \pm Cerr) \mp Cerr\}}{\{Cmes - Coffset\}} \right] = number_d - err = \frac{P}{P_{max}}$$

### 7.3.2.3.2. The quantitative approach

In this discussion, we have not discussed interference effects. By using the schema shown in Figure 7.17, sequenced in direct mode, we get successively:

– interval ]  $nT, (n + 1/2)T$  [:

$$(Q_S)_{nT} = -C_2 \cdot V_S(nT) - SL(nT) \cdot Coffset \cdot Vref + (Coffset - Cref) Vref$$

– interval ]  $(n + 1/2)T, (n + 1)T$  [:

$$(Q_S)_{(n+1/2)T} = -C_2 \cdot V_S((n + 1/2)T) - SL((n + 1/2)T) \cdot Cmes((n + 1/2)T) \cdot Vref$$

$$SL((n + 1/2)T) = SL(nT)$$

In this expression,  $Cmes$  is a variable that constantly varies over time according to the measured pressure but slowly with *the recurrence frequency of the clock  $\Phi$  (Fe)*, or “over sampling” frequency chosen to avoid problems of spectrum folding. That is why a certain degree of leeway is possible in selecting the associated sampled variable; so we write:  $Cmes((n + 1/2)T) = Cmes(nT)$ .

$$\text{Interval ] } (n + 1)T, (n + 1 + 1/2)T \text{ [: } V_S((n + 1/2)T) = V_S((n + 1)T)$$

and:

$$C_2 \cdot V_S((n + 1)T) = C_2 \cdot V_S(nT) + (Cref - Coffset) Vref - SL(nT) (Cmes(nT) - Coffset) Vref \tag{7.8}$$

with:  $\frac{Cref - Coffset}{Cmes(nT) - Coffset} = 1 - \frac{P(nT)}{P_{max}}$

To create  $number(nT)$ , we can set up, at each clock cycle, the accumulation of consecutive N values of  $SL$ :

$$number(nT) = \frac{1}{N} \left\{ \sum_{i=0}^{i=N-1} (1 - SL(n - i)T) \right\} = \frac{1}{N} \left\{ \sum_{i=0}^{i=N-1} \overline{SL}((n - i)T) \right\} \tag{7.9}$$

The above equation becomes:

$$number(nT) = \frac{1}{N} \left( \sum_{i=0}^{i=N-1} \left( \frac{P((n-i)T)}{P \max} \right) + \frac{1}{Vref} \left( \frac{C_2 \cdot V_S((n+1)T)}{(Cmes - Coffset)_{nT}} - \frac{C_2 \cdot V_S((n-N+1)T)}{(Cmes - Coffset)_{(n-N)T}} \right) \right) \quad [7.10]$$

so that  $Cmes((n-i)T)$  is very close to  $Cmes(n-i-1)T$ , which is even coherent with the principle of “oversampling”.

Formula 7.10 shows, in a certain way, the increase in resolution produced by the averaging operation. However, since it contains two unknowns, this equation, like equation [7.8], does not have any predictive value. To obtain this value, we must eliminate an unknown by expressing  $V_S$  according to SL. They are connected by the nonlinear relation:

$$VS(n+1) > 0 \Rightarrow SL(n+1) = 1$$

$$VS(n+1) < 0 \Rightarrow SL(n+1) = 0$$

This implication, added to equation [7.8], completely describes the behavior of the  $\Sigma$ - $\Delta$  modulator, but does not help us precisely quantify the resolution increase suggested by equation [7.10].

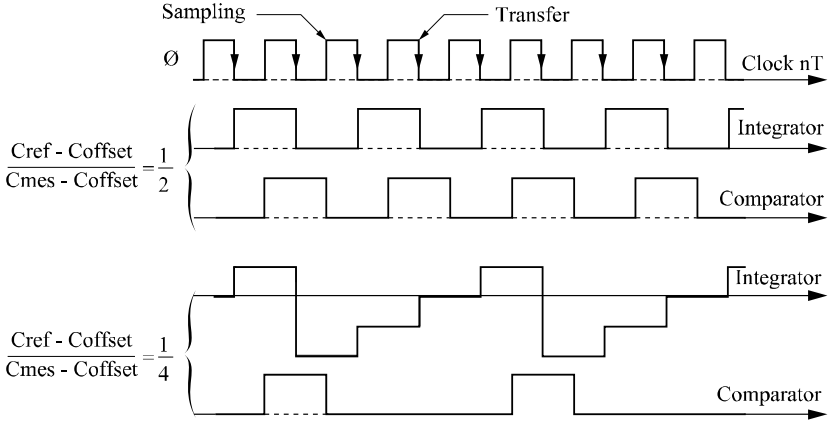
A double approximation, based on linearization on a range of limited pressure and the introduction of quantification noise, resolves this problem.

For a stabilized value of the observed measurement capacity, the average voltage of integrator output  $\langle V_S \rangle$ , is established at:

$$\langle V_S \rangle \approx Vref \left( \frac{2(Cref - Coffset) - \langle Cmes - Coffset \rangle}{2C_2} \right) \quad [7.11]$$

This approximate expression (the rigorous equation is not linear) comes from the simple observation of the output voltage of the integrator for diverse values of  $(Cmes - Coffset)$  (see Figure 7.18). The relation between  $V_S(nT)$  and  $SL(nT)$  is then established easily. By combining equations [7.6] and [7.9], equation [7.11] becomes:

$$\frac{C_2 \langle V_S \rangle}{Vref \langle Cmes - Coffset \rangle} \approx \langle SL \rangle - \frac{1}{2}$$



**Figure 7.18.** Chronogram of voltages for diverse pressure values

In instantaneous value, this is translated into:

$$\frac{C_2 \cdot V_S((n+1)T)}{V_{ref}(C_{mes}((n+3/2)T) - C_{offset})} = (SL((n+1)T) - \delta S((n+1)T)) - \frac{1}{2} \tag{7.12}$$

$$\frac{C_2 \cdot V_S((n+1)T)}{(C_{ref} - C_{offset})V_{ref}} \approx \left[ \frac{P_{max}}{P_{max} - P((n+3/2)T)} \right] \left[ (SL((n+1)T) - \delta S((n+1)T)) - \frac{1}{2} \right]$$

where  $\delta S(nT)$  represents the instantaneous error resulting from quantification. Actually, the only logical values of the comparator output (with a one-bit quantifier),  $SL(nT)$ , are “0” or “1”, while the fraction of the first member has a spectrum of continuous values.

Then, after taking into account the results and rearrangement, equation [7.8] becomes:

$$SL((n+1)T) = \left( \frac{C_{ref} - C_{offset}}{C_{mes}(nT) - C_{offset}} \right) + \delta S((n+1)T) - \delta S(nT) \tag{7.13}$$

$$\overline{SL}((n+1)T) = \left( \frac{P(nT)}{P_{max}} \right) - (\delta S((n+1)T) - \delta S(nT))$$

This equation also assumes, in an approximate way, that  $C_{mes}((n+1)T)$  is very close to  $C_{mes}(nT)$ . This assumption, which is coherent with the “oversampling” principle, is the recurrence equation of a first order  $\Sigma$ - $\Delta$  modulator.



It is usual to link this recurrent equation two transfer functions in  $Z$ :

– The transfer function of the corrected useful signal:

$$\overline{SL}(Z) = Z^{-1} \left[ \frac{P(Z)}{P_{\max}} \right]$$

$$\overline{SL}(Z) = H_{SL}(Z) \left[ \frac{P(Z)}{P_{\max}} \right] \Rightarrow H_{SL}(e^{i2\pi f / Fe}) = e^{-i2\pi f / Fe} \quad [7.14]$$

with:  $\left( \frac{C_{ref} - C_{offset}}{C_{mes} - C_{offset}} \right) (Z) = \left( 1 - \frac{P}{P_{\max}} \right) (Z)$

– The transfer function of the quantification error:

$$br_{SL}(Z) = (1 - Z^{-1})\delta S(Z) \quad [7.15]$$

*Why digital filtering is necessary*

For a usual range of relation values  $P(t)/P_{\max}$  (1/4, 3/4), which covers all the cases encountered experimentally, and as far as the input signal is itself noisy and variable, the instantaneous quantification error  $\delta S(nT)$  has, very approximately, the properties of a sampled white noise, with a spectral density that can be of the magnitude of:

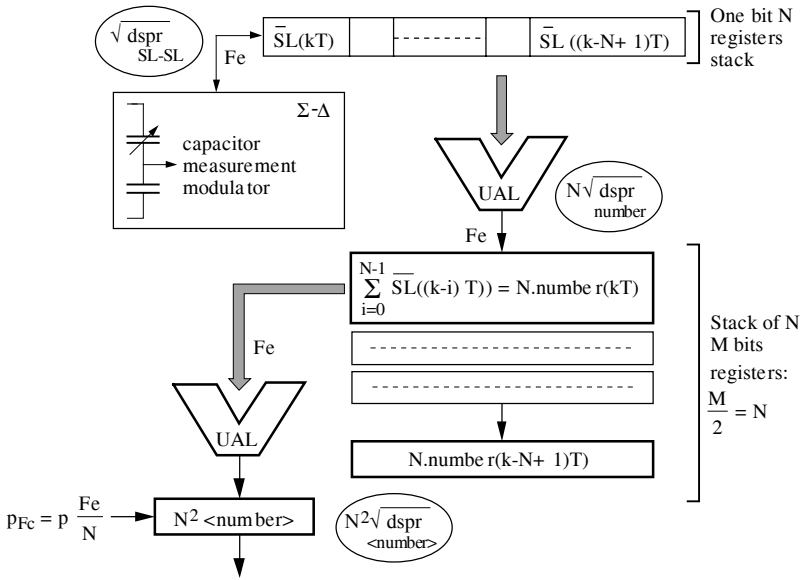
$$dspbr_S = \frac{\Delta^2}{12Fe} \quad \text{with: } -\frac{Fe}{2} \leq f \leq \frac{Fe}{2} \quad [7.16]$$

with  $\Delta$ , the quantification step (which is equal to 1, since we evaluate on the basis of the capacitive ratios and, by extension, the pressure ratios included between  $\{0,1\}$ ) and that the quantifier has one bit. A noise power corresponds to this density:

$$Pbr_S = \frac{\Delta^2}{12}$$

Relations [7.13] and [7.15] show that the  $\Sigma$ - $\Delta$  modulator carries out the discrete differentiation of the quantification noise of the comparator. This operation transforms the “white noise” of the one-bit quantifier to a colored noise concentrated in high frequencies (Figure 7.19). It is characterized by a spectral density of power:

$$dspbr_{SL} = \frac{\Delta^2}{12Fe} 4 \sin^2 \left( \frac{\pi f}{Fe} \right) \quad \text{with: } -\frac{Fe}{2} \leq f \leq \frac{Fe}{2} \quad [7.17]$$



Pressure measurement by means of  $\Sigma$ - $\Delta$  modulator

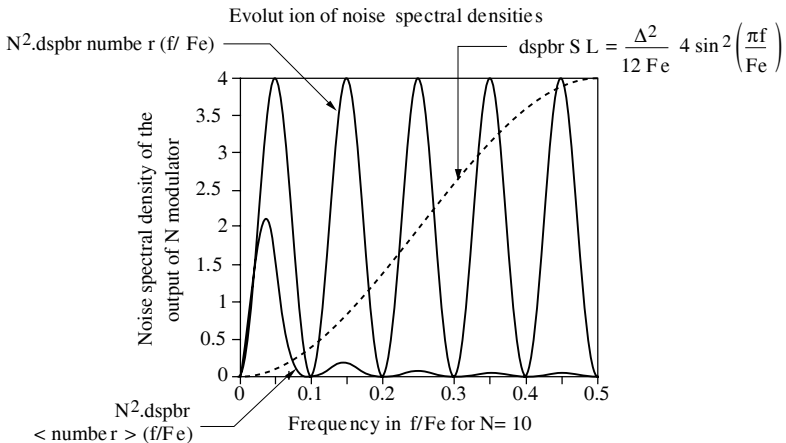
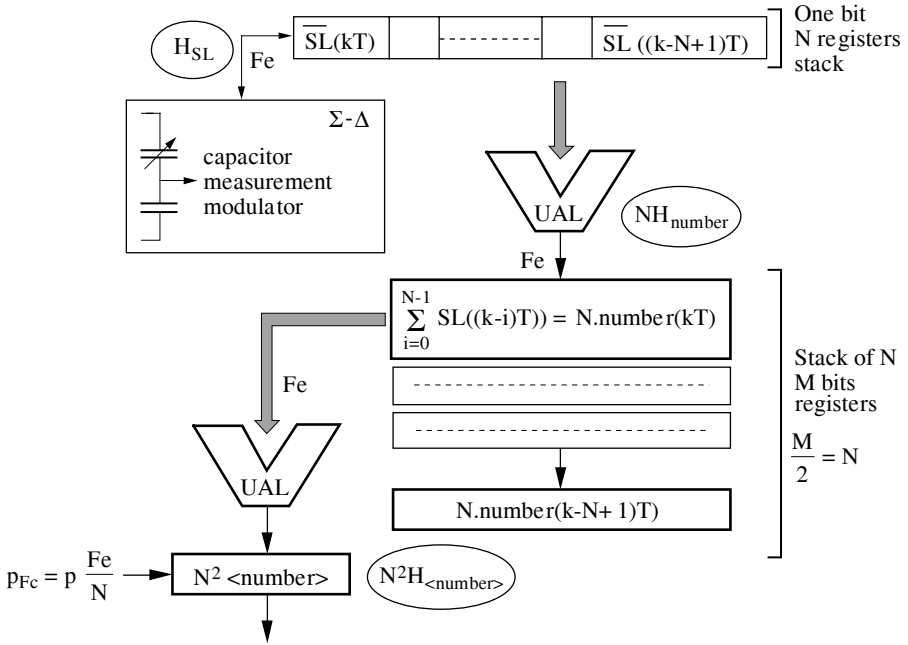


Figure 7.19. Effect of digital filtering on quantification noise

The effects of this transformation of quantification noise are the following:

- a multiplication by almost two of the quantification noise power;

– a drastic reduction of the amplitude of the spectral density of the quantification noise power at low frequencies (Figure 7.19). It is advisable to follow the  $\Sigma$ - $\Delta$  modulator with low pass digital filtering.



Pressure measurement by means of  $\Sigma$ - $\Delta$  modulator

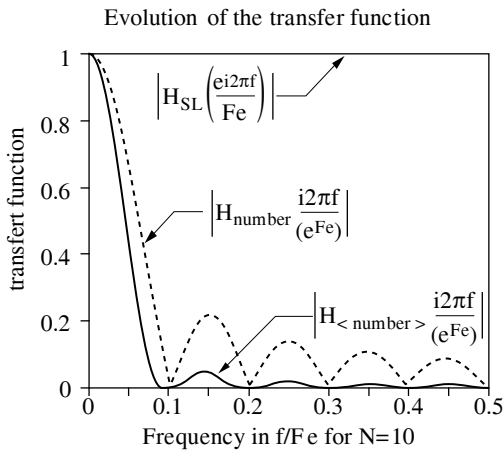


Figure 7.20. Transfer functions

This filter must combine three fundamental features:

- It must ensure a pass band that is compatible with performance expectations.
- It must guarantee a signal-to-noise ratio compatible with expected precision.
- It must allow for the slowest possible output sampling in accord with the pass band.

With this modulator, the choices of an oversampling frequency  $F_e$  and the numbers of samplings  $N$  constitute the only required parameters.

The first filter is a simple averaging filter that helps establish the *number*. Its schema of principle is shown in Figure 7.20. It has a shift register that contains an adder and its accumulator on  $M$  bits:  $2^M = N$  at each instant  $kT N$ , consecutive values of  $SL$  ( $SL(kT)$  to  $SL(k-N+1)T$ ). These are sequenced at the oversampling frequency  $F_e$ . This filter is regulated by the recurrent equation shown below, deduced from equations [7.9], [7.10] and [7.13]:

$$\begin{aligned}
 number((n+1)T) &= \frac{1}{N} \sum_{i=0}^{i=N-1} \left( \frac{P((n-i)T)}{P_{\max}} \right) \dots \\
 &\dots - \frac{\{\delta S((n+1)T) - \delta S((n-(N-1))T)\}}{N}
 \end{aligned}$$

to which we can link the transfer function for the corrected useful signal of the quantification error:

$$\begin{aligned}
 number(Z) &= \frac{Z^{-1}}{N} \left( \frac{1-Z^{-N}}{1-Z^{-1}} \right) \left( \frac{P}{P_{\max}} \right) (Z) \\
 &= H_{number}(Z) \left( \frac{P}{P_{\max}} \right) (Z)
 \end{aligned}$$

and the transfer function for the quantification noise:

$$br_{number}(Z) = \frac{(1-Z^{-N})}{N} \delta S(Z) = \frac{1}{N} \left( \frac{1-Z^{-N}}{1-Z^{-1}} \right) br_{SL}(Z)$$

For the useful signal, this filter presents a frequency response more or less in cardinal sinus:

$$H_{number} \left( e^{\frac{i2\pi f}{F_e}} \right) \cong e^{-i\pi f(N+1)/F_e} \left( \frac{\sin c(\pi f N / F_e)}{\sin c(\pi f / F_e)} \right) \text{ with: } \sin c(x) = \frac{\sin(x)}{x} \quad [7.18]$$

The module of the above equals the unity at zero frequency, which has transmission zeros at all the  $\pm \frac{Fe}{N}$  and presents a pass band:  $\{-Fe/2N \text{ to } Fe/2N\}$ , shown in Figure 7.20.

As for the quantification noise, it is characterized by a power spectral density:

$$dspbr_{number} = \frac{\Delta^2}{12Fe} 4 \sin^2 \left( \frac{\pi f}{Fe} \right) \left( \frac{\sin c(\pi f N / Fe)}{\sin c(\pi f / Fe)} \right)^2 \text{ with: } -\frac{Fe}{2} \leq f \leq \frac{Fe}{2} \quad [7.19]$$

At each step of the filtering, the signal-to-noise ratio is already improved by a fraction of the factor N, as is shown by the closer relations [7.17] and [7.19]. However, the output signal must be sampled at the oversampling frequency Fe, since the spectral density of the noise still extends throughout the band  $\{-Fe/2 \leftrightarrow +Fe/2\}$  (see Figure 7.19).

The second filter is an averaging filter of the same type as described above, but it has as input *number* and for output  $\langle number \rangle$ :

$$\begin{aligned} \langle number(nT) \rangle &= \frac{1}{N} \left( \sum_{i=0}^{i=N-1} number((n-i)T) \right) \quad \text{Therefore:} \\ \langle number((n+1)T) \rangle &= \left[ \begin{aligned} &\frac{1}{N^2} \left\{ \sum_{i=0}^{i=N-1} \sum_{j=0}^{j=N-1} \left( \frac{P((n-i-j)T)}{P \max} \right) \right\} - \frac{1}{N^2} \times \\ &\left\{ \sum_{j=0}^{j=N-1} \delta S((n+1-j)T) - \sum_{j=0}^{j=N-1} \delta S((n-(N-1)-j)T) \right\} \end{aligned} \right] \quad [7.20] \end{aligned}$$

Its schema principle includes the accumulator presented above, and (N - 1) identical registers. It thus constitutes a “pipeline” structure sequenced to the oversampling frequency for reasons already shown. The result is accumulated with the help of a second adder. The ensemble is regulated by the following recurrent equations:

$$\begin{aligned} \langle number(Z) \rangle &= \frac{Z^{-1}}{N^2} \left( \frac{1-Z^{-N}}{1-Z^{-1}} \right)^2 \left( \frac{p}{P \max} \right) (Z) \\ &= H_{\langle number \rangle} (Z) \left( \frac{P}{P \max} \right) (Z) \\ br_{\langle number \rangle} (Z) &= \frac{1}{N} \left( \frac{1-Z^{-N}}{1-Z^{-1}} \right) br_{number} (Z) = \frac{1}{N^2} \left( \frac{1-Z^{-N}}{1-Z^{-1}} \right)^2 br_{SL} (Z) \end{aligned}$$

For the useful signal, this filter presents a frequency response approximately in squared cardinal sinus:

$$H_{\langle number \rangle} \left( e^{\frac{i2\pi f}{Fe}} \right) \cong e \left( -i2\pi fN / Fe \right) \left( \frac{\sin c(\pi fN / Fe)}{\sin c(\pi f / Fe)} \right)^2 \quad [7.21]$$

which clearly attenuates the useful signal (see Figure 7.20) outside the pass band  $\approx \{-Fe/2N \leftrightarrow +Fe/2N\}$ .

As for the quantification noise, it has a power spectral density:

$$dspbr_{\langle number \rangle} \cong \frac{\Delta^2}{12Fe} 4 \sin^2(\pi f / Fe) \left( \frac{\sin c(N\pi f / Fe)}{\sin c(\pi f / Fe)} \right)^4 \quad -\frac{Fe}{2} \leq f \leq \frac{Fe}{2} \quad [7.22]$$

At this filtering level, the signal-to-noise ratio gains another factor  $\sqrt{N}$  in relation to the previous step. This is because its non-zero spectral density (see Figure 7.19) is more or less concentrated from:  $\sim \{-Fe/2N$  to  $Fe/2N\}$ ; that is, a band  $N$  times more narrow than before the previous filtering.

After all calculations have been made, we get:

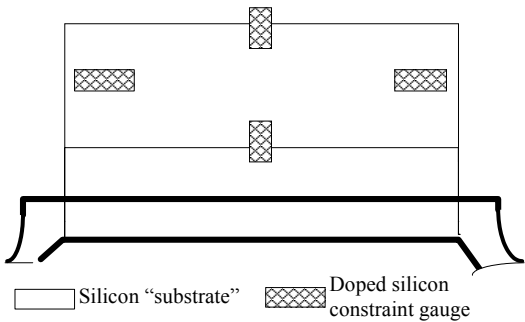
$$\sqrt{Pbr_{\langle number \rangle}} \cong \frac{\pi}{N^{3/2}} \sqrt{Pbr_S} \quad [7.23]$$

The output sampling frequency can be greatly reduced without, however, reaching the limits of *NYQUIST*: *pass band*:  $\{-Fe/2N$  to  $Fe/2N\}$ ; *sampling frequency*:  $Fc = Fe/N$ . In fact, because of the low reliable attenuation outside the pass band of this filtering, a reduction of this order of the sampling frequency can cause irreversible damage (linked to the spectrum replenishment) to the useful signal.

### 7.3.3. Electronic conditioning for piezoresistive cells sensitive to differential pressure

A simple thin membrane chamber manufactured in a silicon substrate constitutes a pressure sensitive cell. It is sensitive to absolute pressure if the membrane delimits an empty space, or to differential pressure if the substrate has been pierced. The

piezoelectric elements, arranged as a Wheatstone bridge, are placed on the edge of the membrane and are subject to maximum constraints for a given pressure shift (Figure 7.21).

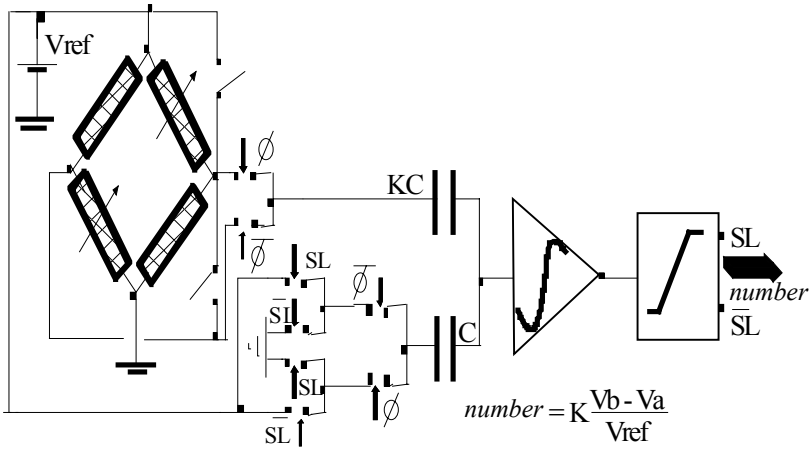


**Figure 7.21.** Piezoresistive cells sensitive to differential pressure

The temperature range that can support this type of device can be increased by placing the piezoresistive elements in oxide shells. This eliminates any escape currents. These sensitive cells are used in electronic speed boxes or they are used to measure oil pressure (20-50 bars, more than 200 degrees).

Figure 7.22 shows a principle schema of how a  $\Sigma$ - $\Delta$  modulator adapts to a one-bit quantifier measuring the disequilibrium voltage of the gauge bridges. The integrator, the comparator, and the CMOS switches are of the same type used in the previous application. The scale factor is fixed by the relation:  $K \cdot C_{ref} / C_{ref}$ . This helps us measure the low disequilibrium voltage of the bridge, with the help of a reference voltage that is higher by two orders of magnitude, all without using a preamplifier [ANA 95], [BAI 97].

However, the sequencing differs from that used in the capacitive pressure sensor, since here we measure a differential pressure that can be positive, negative, or zero. From this we see that at each clock cycle ( $\Phi$ ), the quantity of charge  $C_{ref} \cdot V_{ref}$ , must be brought to or sampled by an integrator according to the state of the comparator. According to the principle schema, the state " $SL = 1$ " of the comparator is a sampling and counting order, while the state " $SL = 0$ " constitutes a carrying and deducting order. As well, during N clock cycles of the clock  $\Phi$ , the balance of charge transfers is established as:



**Figure 7.22.** Application of  $\Sigma$ - $\Delta$  modulator for the measurement of low voltages

$$K \cdot N \cdot C_{ref} (V_b - V_a) - p_{(SL=1)} \cdot C_{ref} \cdot V_{ref} + n_{(SL=0)} \cdot C_{ref} \cdot V_{ref} = 0$$

$$N \cdot number = p_{(SL=1)} - n_{(SL=0)} = K \cdot N \left( \frac{V_b - V_a}{V_{ref}} \right); \text{ with: } p_{(SL=1)} + n_{(SL=0)} = N \quad [7.24]$$

“Number” is a measurement of the voltage of the bridge disequilibrium in relation to the quantification noise and to the charge injection errors of the clocks.

By linking the results to the constraint effects created by the differential pressure of the membrane, we get:

$$N \cdot number = K \cdot N \left( \frac{\delta R}{4R} \right) = K \cdot N \cdot \gamma \cdot \Delta P \quad [7.25]$$

to the quantification noise and to close to the charge injection errors of the clocks.

The factor  $\gamma$  is the result of the combination of piezoelectric properties of the gauges and the elasticity of the membrane.

It is notable that in equation [7.24], as in equation [7.25], only the factor “K” remains; the values of  $V_{ref}$  and  $C_{ref}$  have no effect from a metrological point of view. The improvement of the modulator the proceeds by introducing two functioning modes, “direct” and “reverse”, which, as in the previous application,



eliminate the effect of injection. And by using a reconfigured mode (shown by the dotted line in the schema), the scale factor  $K$  can be measured precisely (the exact value of other capacities are without importance from a metrological point of view). Then  $K$  can be stored for subsequent measurements. This functioning mode can be activated at any time; the modulator is able to performing autocalibration. By reducing the impact of interference effects, integration, in the form of symmetrical architecture, leads to very good performances (16 bits of resolution for a pass band of 1,000 Hz, an oversampling frequency of 2 MHz for a first order modulator, to cite two examples).

### 7.3.4. *Electronic conditioning for cells sensitive to acceleration*

Accelerators are being used more and more inside of vehicles. Recently, airbags have been equipped with similar microsystems; these are the basis of the inflation that occurs in case of accidents. As well, accelerometers functioning as inclinometers are used for controlling the suspension of some vehicles. These mechanisms also function in automatic shock absorption systems, used more or less often according to the state of the road or the car.

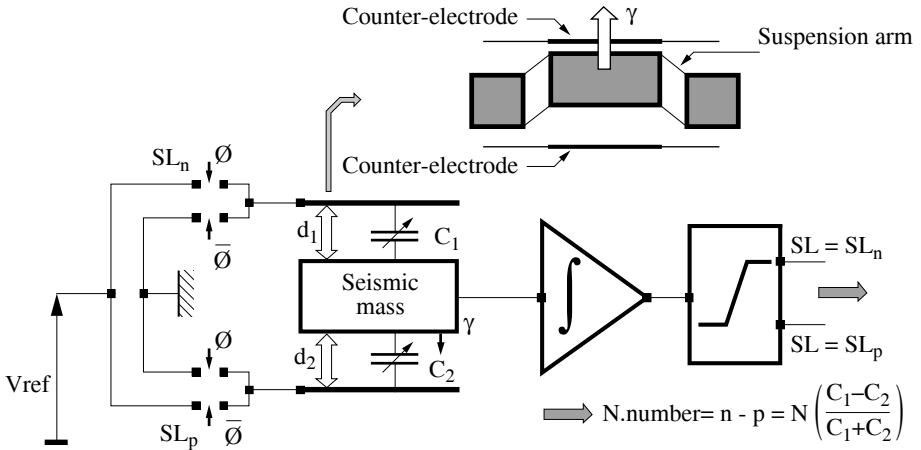
Even though these capacitive structures are mostly used in low-precision piezoresistive structures, they are can also be adapted to servo-control accelerometers.

#### 7.3.4.1. *Direct applications of first-order $\Sigma$ - $\Delta$ modulators to 1 bit quantifiers*

Figure 7.23 shows the schema for the principle of adapting a  $\Sigma$ - $\Delta$  modulator to a one-bit quantifier for measuring acceleration with a differential capacitive sensor that has been optimized to function in open loop. This means that from zero frequency to the “pseudo-frequency” of resonance of the system of a buffered spring connected to a mass, the response will be almost flat and is regulated by the equation:

$$d_2 - d_1 = \frac{m}{k} \gamma \quad [7.26]$$

where  $\gamma$  is the acceleration and  $m$  the seismic mass. The distances  $d_2$  and  $d_1$  measure, are, respectively, the gaps (polegaps) that separate the seismic mass from the lower and higher counter-electrodes.



**Figure 7.23.** Schema principle of a modulator for an accelerometer

This kind of device is much smaller when it is micromachined on silicon. It has a mass of several micrograms to several nonograms, a pole gap of several microns, lateral dimensions of several hundred microns, and reduced stiffness. A fluid such as atmospheric air pressure is an excellent accelerometer buffer of the range of between 0 and 100 g. Moreover, the seismic mass, along with the counter-electrodes, makes up two flat capacities:

$$C_1 = \frac{\epsilon_0 S}{d_1} \text{ and } C_2 = \frac{\epsilon_0 S}{d_2} \tag{7.27}$$

where  $S$  represents the surface of the seismic mass as relates to the counter-electrodes, and  $\epsilon_0$  the permittivity of the void. These two capacities are not independent and are linked by the relation:

$$\frac{1}{C_1} + \frac{1}{C_2} = \frac{d_1 + d_2}{\epsilon_0 S} = 2 \frac{d_0}{\epsilon_0 S} = \frac{2}{C_0} \tag{7.28}$$

where  $d_0$  represents the distance of the seismic mass to one or the other of the counter-electrodes in the absence of acceleration, and  $C_0$  represents the corresponding capacity. The introduction of relations [7.2] and [7.28] to the interior

of the expression in [7.26] suggests that a measurement principle has been completely adapted to the potentialities of the  $\Sigma$ - $\Delta$  architecture. So we get:

$$\frac{d_2 - d_1}{2d_o} = \frac{m}{2d_o \cdot k} \gamma = \frac{d_2 - d_1}{d_2 + d_1} = \frac{C_1 - C_2}{C_2 + C_1} \Rightarrow \gamma = \frac{2d_o \cdot k}{m} \left( \frac{C_1 - C_2}{C_2 + C_1} \right) \quad [7.29]$$

The modulator that enables this acceleration measurement is identical to the device which produced the pressure measurement with the help of the capacitive cell, and is close to the modulator that modified the clock regulator. By using the hydraulic analogy again, we understand the measurement process. If the tub level is too low, it must be replaced with the help of  $C_1$  under the effect of the conditional clock  $SL_n \Phi$ . In the opposite case, it is emptied with the help of  $C_2$  under the effect of the complementary clock  $SL_p \Phi$  ( $SL_n = \overline{SL_p}$ ). On  $N$  count rates, the balance of calls to  $C_2, p$ , subtracted from  $C_1, n$ , equals:

$$n + p = N$$

$$nC_1 \cdot Vref - pC_2 \cdot Vref = 0 \quad \text{with} \quad n = \sum_0^N SL_n(i) \dots p = \sum_0^N SL_p(i) \quad [7.30]$$

to the quantification noise and to near to the clocks' charge injection errors.

Also:

$$n - p = N \left( \frac{C_1 - C_2}{C_1 + C_2} \right) = N \left( \frac{d_2 - d_1}{d_1 + d_2} \right) = N \left( \frac{m}{2kd_o} \right) \gamma \quad [7.31]$$

with the same error sources as in [7.30].

All techniques used for eliminating the effects of injection by a combination of direct and reverse modes, as well as the control of the combination of pass band/precision by digital filtering discussed above in detail in capacitive pressure sensor applications, are rigorously reusable without any prior adaptations.

*7.3.4.2. Producing an accelerometer in true open loop by eliminating the effects of electrostatic forces*

Here, we are close to the formula in [7.3]. It ignores the impact of electrostatic forces whose results are not zero by principle with the functioning mode of the  $\Sigma$ - $\Delta$

architecture. An acceptable compromise can be attained by reducing the extent of the voltage applied to the electrodes ( $V_{ref}$ ). This is because the electrostatic forces decrease with the square of the electrodes.

However, when we must try to find linearity and a very good signal-to-noise ratio in order to conserve the functioning of the open loop of the capacitive sensitive cells, we must use other architectures that in principle eliminate this problem.

Miniaturization aggravates this problem. In fact, if all the dimensions, including the pole gap, develop at the same rhythm, the mechanical force decreases at a fixed acceleration rate along with the seismic mass volume, while the electrostatic force remains constant.

$$F_e = \frac{\varepsilon \cdot L^2 \cdot V^2}{2 \cdot d^2} \propto V^2 \Leftrightarrow F_m = \rho \cdot L^3 \cdot \gamma \propto L^3 \cdot \gamma$$

$\rho = \text{massic density}$

These forces are of the same order of magnitude for the potential value compatible with the microelectronics in the ranges of acceleration measurements usual when the pole gap has an order of several micrometers for mass dimensions of the seismic mass, which are millimetric. These values are typical of cells that can be produced through current silicon microtechnologies.

It is always desirable to *eliminate electrostatic forces*. Doing so means bringing the seismic mass to a certain potential in relation to two counter-electrodes in order to cancel the effects of the forces (see Figure 7.24).

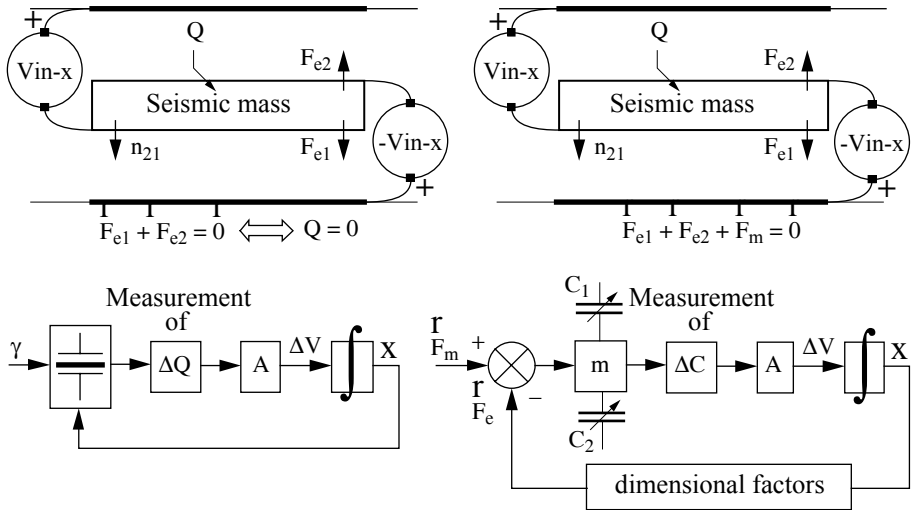
$$\vec{F}e_1 + \vec{F}e_2 = 0$$

$$\vec{F}e_1 = \frac{Q_1^2}{2\varepsilon S} \cdot \vec{n}_{21} \Leftrightarrow \vec{F}e_2 = \frac{Q_2^2}{2\varepsilon S} \cdot \vec{n}_{12} \quad [7.32]$$

where  $Q_1$  and  $Q_2$  are the shared charges on the surfaces of the seismic mass in relation to the two counter-electrodes. The equation is fulfilled under two specific conditions:

$$Q_1 = Q_2 \quad \text{or} \quad Q_1 = -Q_2 \Rightarrow Q = 0 \quad [7.33]$$

Only the second condition can lead to a relevant measurement of the technique of the switching capacities. We will know if the verification of the sum of the charges contained by the mass is actually zero.



**Figure 7.24.** Illustration of the two principles of acceleration measurement

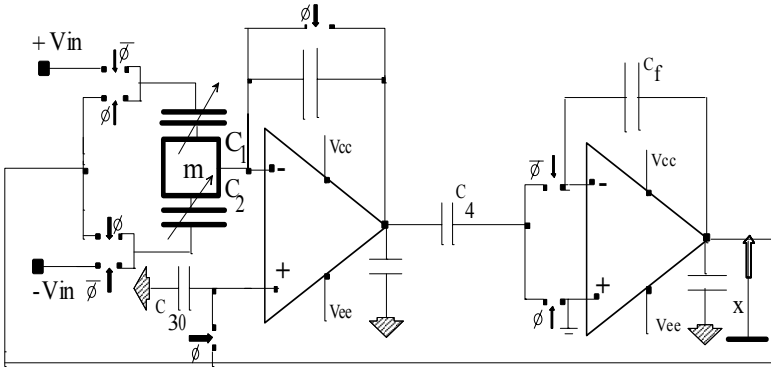
We can interpret these results by using dimensional parameters, the potentials  $V_{in}$ ,  $-V_{in}$ , and  $x$ , the two counter-electrodes, and the seismic mass. We do this by assuming as a hypothesis that the seismic mass displaces in parallel position to itself. So we get (see Figure 7.24);

$$\begin{aligned}
 Q &= Q_1 + Q_2 = \frac{\epsilon S}{d_1}(x - Vin) + \frac{\epsilon S}{d_1}(x + Vin) = C_1(x - Vin) + C_2(x + Vin) \\
 \Rightarrow Q = 0 &\Leftrightarrow x = \left(\frac{C_1 - C_2}{C_1 + C_2}\right) Vin
 \end{aligned}
 \tag{7.34}$$

Since  $V_{in}$  is stable, the potential  $x$  of the seismic mass of the effects of the electrostatic forces constitutes an acceleration measurement, since from equation [7.31], we get:

$$x = Vin \cdot \left(\frac{C_1 - C_2}{C_1 + C_2}\right) = Vin \cdot \left(\frac{d_2 - d_1}{d_1 + d_2}\right) = Vin \cdot \left(\frac{m}{2kd_o}\right) \gamma
 \tag{7.35}$$

The circuit shown in Figure 7.25 ensures the required functionality [LEU 90].



**Figure 7.25.** Schema principle of the analog architecture for eliminating effects of electrostatic forces

During the intervals  $]nT, (n + 1/2)T[$ , the switches switched by the clock  $\Phi$  are closed. All the nodes of the charge amplifier, the counter-electrodes, and the seismic mass are brought to the same potential. The seismic mass is subject only to mechanical forces.

During the intervals  $](n + 1/2)T, (n + 1)T[$ , the node “+” of the charge amplifier does not change in potential, but the output voltage varies according to the disequilibrium measurement of the shared charges on the surface of the seismic mass, since:

$$V_{I(n+1/2)} - x_n = \frac{(C_1 + C_2)x_n + (C_2 - C_1)V_{in}}{C_3} \tag{7.36}$$

The integrator accumulates the effects of these disequilibriums, then cancels them in static regime, thanks to the feedback loop; at the same time it eliminates the electrostatic effects. We then have:

$$x_{n \rightarrow \infty}(nT) = V_{in} \cdot \left( \frac{C_1 - C_2}{C_1 + C_2} \right)_{n \rightarrow \infty} = V_{in} \cdot \left( \frac{m}{2kd_o} \right) \gamma_{static} \tag{7.37}$$

With this result, we must take note of two phenomena:

- This system includes a feedback loop. This means it is subject to stability conditions that impose certain constraints on the capacity values:

$$0 < \frac{C_4}{C_f} \left( \frac{C_1 + C_2}{C_3} \right) < 2$$

- In variable regime, equation [7.36], for practical purposes, is satisfied in so far as the sampling frequency remains as fast as the variation speed of acceleration. In other words, the sampling frequency must be high in relation to the pass band of the sensitive cell.

Here we again see the inherent advantage of feedback structures. These help us obtain a measurement quality (apart from response times) that is not sensitive to component variability, with the exception of the sensitive cell itself.

### 7.3.4.3. Servo-control of an accelerometer using balanced mechanical forces through electrostatic forces

In the previous solutions, the analog or digital measurement representing the acceleration of an approximate scale factor contained the quotient:

$$m / (d_0 \cdot k)$$

The value of the stiffness “*k*” is a parameter that is highly sensitive to dispersive effects related to dimension reduction and to certain thermal steps of manufacturing processes. In addition, this value can vary according to conditions of usage. Although the small pole gap “*d*<sub>0</sub>” is not identical to a sample taken from the same series, but its value remains stable. The seismic mass is, however, a well-identified physical object of some size even in the most miniaturized sensitive cells and is time-invariant. “*d*<sub>0</sub>” and “*m*” can thus be memory-stored objects.

The theoretic static scale factor now depends on only two parameters if the measurement method is based on the balancing of the mechanical force by the electrostatic forces.

#### 7.3.4.3.1. Analog solution

In the cell with two counter-electrodes (Figure 7.24), the equation expressing this principle:

$$\vec{F}_m + \sum \vec{F}_{elec} = 0 \Rightarrow \gamma_{stat} = \frac{\sum \vec{F}_{elec}}{m}$$

expresses at equilibrium:

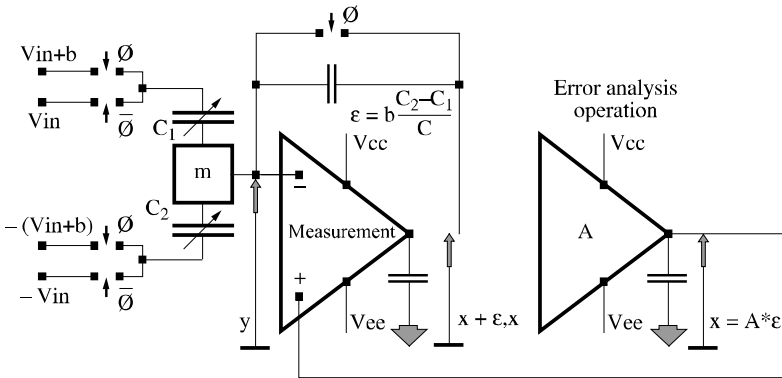
$$\begin{aligned} \vec{F}m + \vec{F}e_1 + \vec{F}e_2 = 0 &\Leftrightarrow C_1 + C_2 = C_0 \\ \frac{C_0(V_{in} - x)^2}{2d_0} \bar{n}_{21} + \frac{C_0(V_{in} + x)^2}{2d_0} \bar{n}_{12} + m \cdot \gamma_{stat} \cdot \bar{n}_{21} &= 0 \end{aligned} \quad [7.38]$$

In addition:

$$x_{theoretical} = \frac{m \cdot d_0}{2 C_0 \cdot V_{in}} \gamma_{stat} = \frac{m \cdot d_0^2}{2 \varepsilon \cdot S \cdot V_{in}} \gamma_{stat} \quad [7.39]$$

The measurement structure is a looped electromechanical system in which the error signal is the gap between the values of two capacities  $C_1$  and  $C_2$  (see Figure 7.24).

The switching capacities circuit (Figure 7.26) fulfills the desired functionality; the potential difference “ $b$ ” brings about the gap between the two capacities  $C_1$  and  $C_2$ .



**Figure 7.26.** Principle schema of the analog architecture of equilibrium forces

Under the effect of constant acceleration, the signal this circuit transmits is expressed as:

$$\begin{aligned} x_{real} &= \frac{m \cdot d_0}{2 C_0 V_1} \frac{\gamma_{stat}}{1 + \frac{C \cdot d_0^2}{2b \cdot V_1 \cdot C_0^2} \frac{k}{A_{stat}}} \\ V_1 &= V_{in} + b/2 \end{aligned} \quad [7.40]$$



This relation shows the dependence between the static scale factor and the stiffness that occurs when the loop gain is not infinite. However, to further develop this analysis, the dynamic characteristics of the sensitive cell must be taken into account. The seismic mass is subject to the linking of four forces: the effects of suspension stiffness; acceleration; viscous absorption; and the result of electrostatic forces.

In the absence of “small signal” electrostatic forces, the function of the harmonic transfer of the sensitive cell is expressed as:

$$\frac{Y(\omega)}{\Gamma(\omega)} = H(\omega) = \frac{m}{k} \frac{1}{1 + i \frac{\lambda}{k} \omega - \frac{m}{k} \omega^2} \tag{7.41}$$

where  $Y(\omega)$ ,  $\Gamma(\omega)$  and  $H(\omega)$  are respectively the harmonic expressions of the displacement of the seismic mass, of acceleration, and of the transfer function; and “ $\lambda$ ” is the coefficient of viscous absorption.

As soon as the sampling frequency  $Fe$  is high enough, the results of the discretization can be ignored. The transfer function describing the “small signal” behavior of the entire system is easily determined by explicitly inserting  $H(\omega)$  into the block schema (see Figure 7.27):

$$\frac{X_{real}(\omega)}{\Gamma(\omega)} = \frac{m \cdot d_0}{2 C_0 \cdot V_1} \frac{1}{1 + \left( \frac{C \cdot d_0^2}{2b \cdot V_1 \cdot C_0^2} \right) \left( \frac{m}{H(\omega) A(\omega)} \right)} \tag{7.42}$$

$$V_1 = V_{in} + b/2$$

This expression shows that the error analysis operator,  $A(\omega)$  cannot be chosen independently of the sensitive cell. To clarify this point, we will look at two cases: the cell with dominant absorption and the cell with optimum absorption.

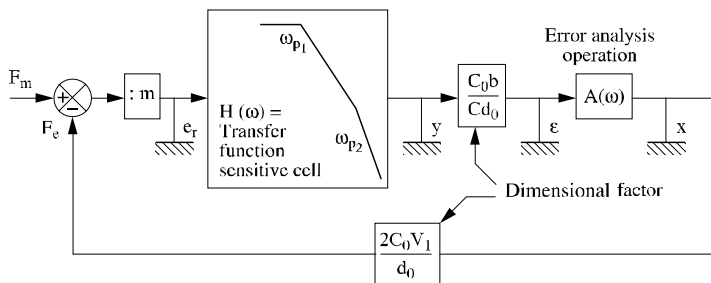


Figure 7.27. Block schema

7.3.4.3.2. Analog solution: sensitive cell with dominant absorption

In this kind of cell, the stiffness “ $k$ ” is very low, so that the transfer function of the cell is close to:

$$H(\omega) \cong \frac{m}{i\lambda\omega(1+i\frac{m}{\lambda}\omega)}$$

To avoid any instability in the looped system, the error analysis operator is a limited-gain amplifier. Expression [7.42] then becomes:

$$\frac{X_{real}(\omega)}{\Gamma(\omega)} = \frac{m \cdot d_0}{2C_0 \cdot V_1} \frac{1}{1+i\left(\frac{C \cdot d_0^2}{2b \cdot V_1 \cdot C_0^2}\right)\left(\frac{\lambda\omega(1+i(m/\lambda)\omega)}{A}\right)}$$

$$V_1 = V_{in} + b/2 \tag{7.43}$$

$$A < \frac{C \cdot d_0^2}{2b \cdot V_1 \cdot C_0^2} \frac{\lambda^2}{m}$$

The signal to mechanical noise ratio of this type of sensitive cell is potentially the best. But these cells are also the most fragile because of the low stiffness of the seismic mass suspension; and they also withstand miniaturization poorly.

7.3.4.3.3. Analog solution: sensitive cells with optimum absorption

The response of these cells is nearly flat up to their pseudo-pulsation resonance  $\Omega_R$ :

$$H(\omega) \cong \frac{m}{k}$$

$$0 < \omega < \Omega_R = \sqrt{\frac{k}{m}}$$

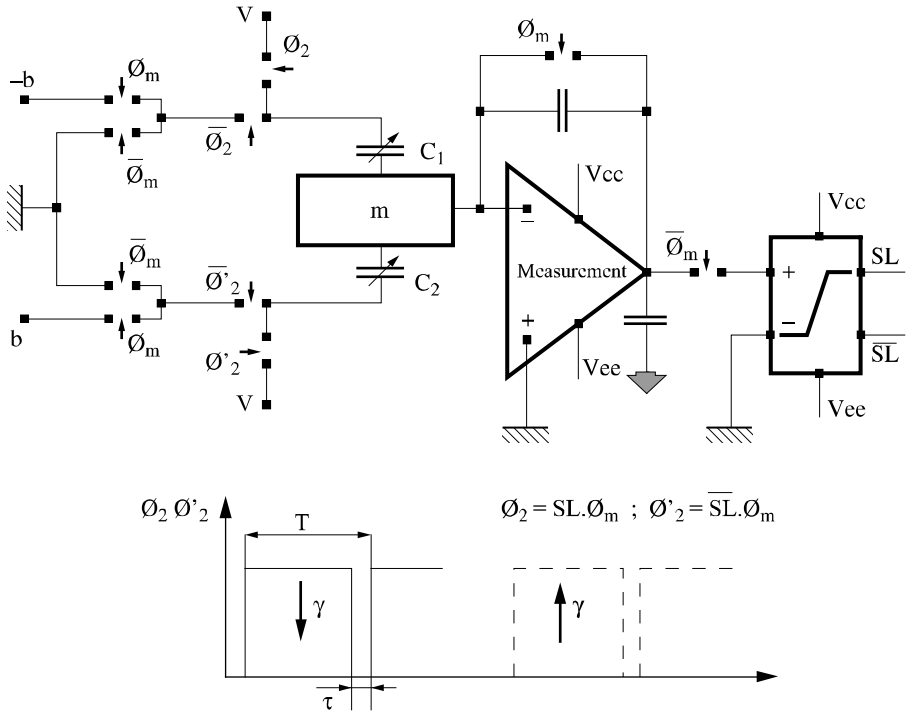
The error analysis operator can then be an integrator with a limited *transition pulsation* “ $\omega_T$ ”. Expression [7.42] then becomes:

$$\frac{X_{real}(\omega)}{\Gamma(\omega)} = \frac{m \cdot d_0}{2C_0 \cdot V_1} \frac{1}{1+i\left(\frac{C \cdot d_0^2}{2b \cdot V_1 \cdot C_0^2}\right)\left(\frac{k\omega}{\omega_T}\right)}$$

$$V_1 = V_{in} + b/2 \tag{7.44}$$

$$\omega_T < \left(\frac{C \cdot d_0^2}{2b \cdot V_1 \cdot C_0^2}\right) k\Omega_R$$

Micromachined accelerometers used to measure high accelerations typically have this kind of behavior.



**Figure 7.28.** Chronogram and digital architecture for a sensitive cell with dominant absorption

### 7.3.4.3.4. Digital solutions

Structures derived from  $\Sigma$ - $\Delta$  modulators can also be used in servo control accelerometers. Figures 7.28 and 7.29 are shown in block schemas with two structures of this kind. The first is suitable for accelerometers with dominant absorption; the second is for accelerometers that are close to optimum absorption [ZIM 95], [BAI 94b].

$$\gamma = \frac{C_0 V^2}{2d_0 \cdot m} \left( \frac{T - \tau}{T} \right) \left( \frac{p - n}{N} \right) \dots \dots \quad [7.45]$$

with:

$$\begin{aligned}
 p + n &= N \\
 n(kT) &= \sum_{i=0}^{N-1} SL((k-i)T) \dots \dots \dots p(kT) = \sum_{i=0}^{N-1} \overline{SL}((k-i)T) \\
 N \cdot number_{\gamma}(kT) &= p(kT) - n(kT)
 \end{aligned}$$

$$\gamma = \frac{C_0 \cdot V^2}{2 \cdot d_0 \cdot m} \left( \frac{\tau_n - \tau'_n}{T} \right)$$

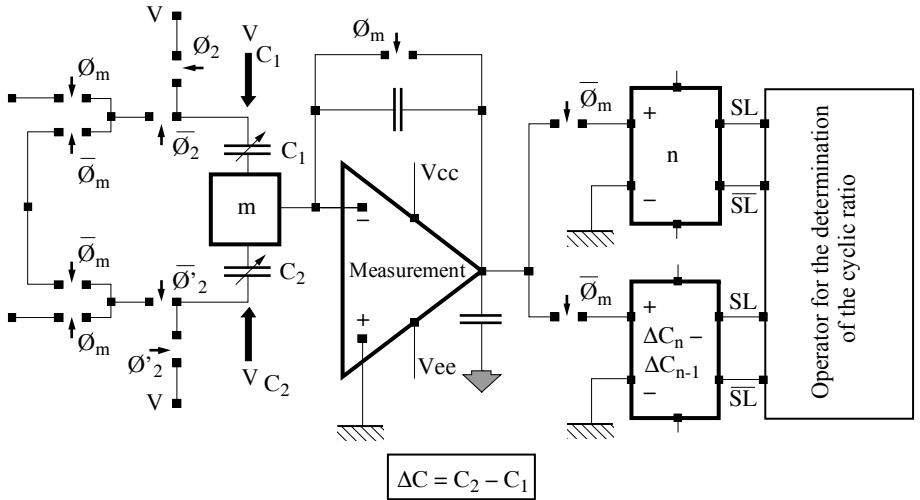
Truth table for the operator: *R* for rapid, *L* for slow

$$\tau \geq 0 \Leftrightarrow \gamma \geq 0 \qquad \tau' > 0 \Leftrightarrow \gamma < 0$$

$\Delta C_n > 0 \quad \Delta C_n - \Delta C_{n-1} > 0 \Rightarrow \tau_{n+1} = \tau_n + \delta R$	$\Delta C_n < 0 \quad \Delta C_n - \Delta C_{n-1} < 0 \Rightarrow \dot{\tau}_{n+1} = \dot{\tau}_n + \delta R$
$\Delta C_n > 0 \quad \Delta C_n - \Delta C_{n-1} < 0 \Rightarrow \tau_{n+1} = \tau_n + \delta L$	$\Delta C_n < 0 \quad \Delta C_n - \Delta C_{n-1} > 0 \Rightarrow \dot{\tau}_{n+1} = \dot{\tau}_n + \delta L$
$\Delta C_n < 0 \quad \Delta C_n - \Delta C_{n-1} > 0 \Rightarrow \tau_{n+1} = \tau_n - \delta R$	$\Delta C_n > 0 \quad \Delta C_n - \Delta C_{n-1} > 0 \Rightarrow \dot{\tau}_{n+1} = \dot{\tau}_n - \delta R$
$\Delta C_n < 0 \quad \Delta C_n - \Delta C_{n-1} < 0 \Rightarrow \tau_{n+1} = \tau_n - \delta L$	$\Delta C_n < 0 \quad \Delta C_n - \Delta C_{n-1} < 0 \Rightarrow \dot{\tau}_{n+1} = \dot{\tau}_n - \delta L$

[7.46]

Accelerometers, and more generally inert systems, have been the subjects of much research; the final performances of these microsystems are very sensitive to using electronics within technological parameters. However, it would be outside the scope of this chapter to treat this topic in depth, since here we are limiting our discussion to initiation technologies



Superposition of the servo-control signal and measurement

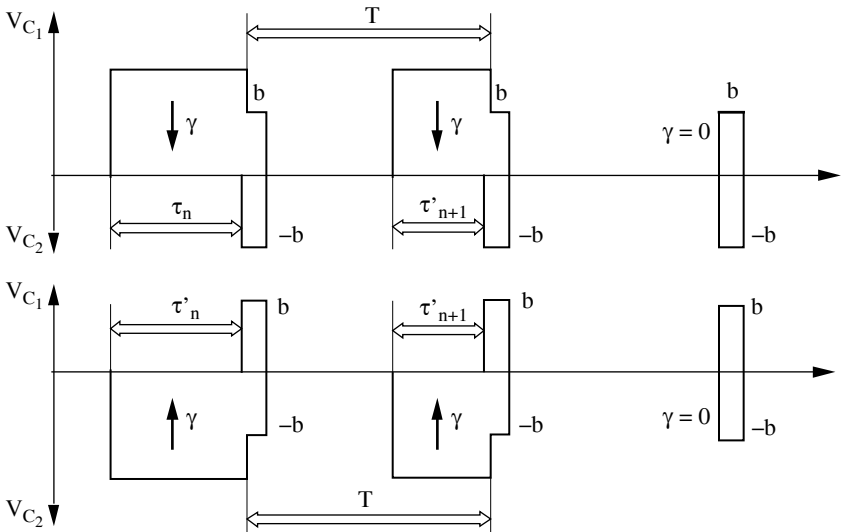


Figure 7.29. Digital architecture and chronogram of a sensitive cell with optimum absorption

### 7.3.5. Energy sources in microsystems

In a certain number of microsystem applications, energy sources can prove problematic. This is the case with nomad systems and even more with abandoned systems without their own sources.

Much research has been undertaken on electronics with low consumption levels, microsensors (electrothermal, electrodynamic and photovoltaic), Hz and optical telesupplying, among other areas.

Up to now, this research has not had much impact in the area of automotive technology devoted to the battery. However, telesupplying has already appeared in mobile parts microsystems responsible for chassis-road connections (such as tire pressure), remote control, and engine immobilization.

#### 7.4. Bibliography

- [ANA 95] Analog Devices, 1995.
- [BAH 95] BAHER H., AFIFI E., "A fourth order switched capacitors cascade structure for  $\Sigma\Delta$  converters", *International journal of circuit theory and applications*, vol. 23(1), p. 3-21 (1995).
- [BAI 94a] BAILLIEU F., BLANCHARD Y., *Signal analogique et capacités commutées*, Dunod, 1994.
- [BAI 94b] BAILLIEU F., MARTY J., MOREAUX C., DELPOUX A., BLANCHARD Y., ZIEMMERMANN L., EBERSOHL J.P.H., LÉ HUNG F., "Low cost differential + - 2G capacitive accelerometer: Technology and electronics design based on fuzzy logic", *Proceedings of EUROSENSOR VIII*, 1994.
- [BAI 96] BAILLIEU F., BLANCHARD Y., LOUMEAU P., PETIT H., PORTE J., *Capacités commutées et applications*, Dunod, 1996.
- [BAI 97] BAILLIEU F., "Les microsystèmes, des microtechnologies et une nouvelle approche de l'électronique", *Forum ADEMIS*, 1997.
- [ELW 98] ELWENSPOEK M., JANSEN H., *Silicon Micromachining*, Cambridge University Press, 1998.
- [FUK 98] FUKUDA T., MENZ W. (ed.), "Micro Mechanical Systems: Principles and technology", volume 6 from *Handbook of Sensors and Actuators*, Elsevier, 1998.
- [GAR 94] GARDNER J.W., *Microsensors: Principles and Applications*, Wiley, 1994.
- [GHA 94] GHANDHI S.K., *VLSI Fabrication Principles: Silicon and Gallium Arsenide*, Wiley, 1994.
- [LEU 90] LEUTHOLD H., RUDOLPH R., "An ASIC for high resolution micro accelerometer", *Sensors and actuators*, A21/A23 (1990).
- [NAG 86] NAGARAJ K., VLACH J., VISWANATHAN T.R., SINGHAL K., "Switched capacitor integrator with reduced sensitivity to amplifier gain", *Electronics letters*, vol. 22 (1986).
- [RIS 94] RISTIC L., *Sensor Technology and Devices*, Artech House, 1994.
- [SZE 81] SZE S.M., *VLSI Technology*, Wiley, 1981

- [SZE 94] SZE S.M. (ed.), *Semiconductor Sensors*, Wiley, 1994.
- [VER 99] VERJUS F., QUEMPEL J.M., BOUROUINA T., BELHAIRE E., DUFOUR-GERGAM E., PÔNE J.F., GILLES J.P., “FEM analysis of resonance frequency matching by Joule heating, Application to a vibratory micro-gyroscope”, *Eurosensor XIII*, September 12-15, 1999, The Hague, The Netherlands.
- [WAN 98] WANG B., KAJITA T., SUN T., THEMES G., “New high-precision circuits for on-chip capacitor ratio testing and sensor read-out”, *IEEE Transactions on Circuits and Systems*, March 1998.
- [ZIM 95] ZIMMERMANN L., EBERSOHL J.P.H., LÉ HUNG F., BERRY J.P., BAILLIEU F., REY P., DIEM B., RENARD S., CAILLAT P., “Airbag application: a microsystem including a silicon capacitive accelerometer, CMOS switched capacitors electronics and true self test capability”, *Sensors and Actuators*, A 46/47 (1995).

## Chapter 8

# Instruments and Measurement Chains

In general, studying a physical phenomenon is done with a computerized system and a central computer guiding a range of measurement devices. Each of these devices is connected to a type of sensor, as well as to one or more buses for the exchange of information. Between the operator and the computer is the user interface. The idea of a measuring device should be understood in a broad sense, whether the device exists separately (battery or mains operated) or is part of a measurement card in a rack or a computer. In this chapter, our goal is to present an overall view of the principle measurement instruments, their functioning principles and the kinds of measurement that can be obtained by using them. Then connecting buses will be discussed, as well as ways to create a measurement chain; that is, an instrumentation system.

### **8.1. Measurement devices**

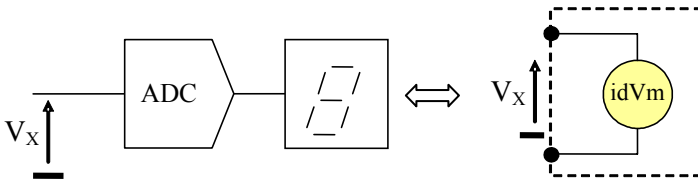
In this section, we will discuss standard measurement devices and their recent, digital, developments. Recently digitization is done throughout the system before the overall measurement process begins. This means it is carried out increasingly with sensors.



### 8.1.1. Multimeters

#### 8.1.1.1. Measurement principles

Generally speaking, a multimeter links a voltmeter, an amperemeter and an ohmmeter. We limit ourselves to these three kinds of measurement because they can be organized easily around the same unit: the analog-to-digital converter (ADC). The three measurements listed above are often based on the numerization of one voltage. The measurement results are converted and displayed in the form of digital values. The ADC unit and its display constitute what we call the ideal voltmeter, “idVm”, shown in Figure 8.1.



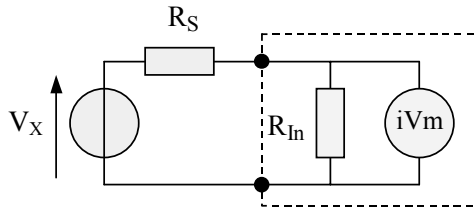
**Figure 8.1.** Principle of the multimeter and voltage measurement

Analog-to-digital converters generally operate by successive approximations, but there are some converters that operate on the multiple ramp system.

Displays are carried out with a certain number of numerals or digits. We find devices called “3 ½ digits” or “6 ½ digits”. In this designation, the ½ digit is the first character to be displayed, with 0 or 1 or even 2 (it is a half digit because it does not take all the possible values). The whole number (here a 3 or a 6) represents the following digits. Thus, the displayed number (without taking into account possible decimal points) can have values of between 0 and the maximum displayable or number of points. A 3 ½ multimeter can display from 0 to 1,999 and is called a 2 million point device. The HP34410A multimeter made by Agilent Technologies™ is a 6 ½ digit or 2 million point multimeter.

#### 8.1.1.2. Input resistance influence

The first source of errors in measuring voltage comes from the non-infinite resistance of the voltmeter. We can design a digital voltmeter by joining the ideal voltmeter (idVm) to an input resistance  $R_{in}$ . A systematic measurement error appears when the equivalent Thévenin resistance of the dipole we want to study is relatively high compared to the input resistance of the multimeter.



**Figure 8.2.** Influence of input resistance

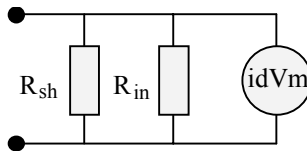
Here, the measurement error (in percentages) is expressed by the following formula.

$$\delta m = \frac{100 R_S}{R_S + R_{in}} \quad [8.1]$$

The input resistance is typically 10 M $\Omega$  but can reach much higher values. For example, on the HP34401 multimeter, we can choose on the ranges of 0.1 V, 1 V and 10 V as an input impedance of 10 M $\Omega$  or even above G $\Omega$ , whereas on the 100 and 1,000 V ranges it is worth 10 M $\Omega$ .

#### 8.1.1.3. Intensity measurements

An intensity measurement is obtained by converting the intensity into a voltage through a shunt resistance. The voltmeter then measures the voltage to the resistance limits.

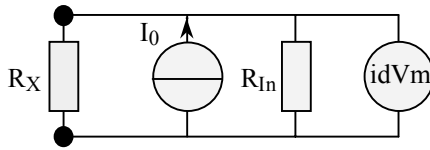


**Figure 8.3.** Principle of the multimeter: intensity measurement

The important parameter to consider in estimating the quality of the measurement is the voltage drop to the limits of the shunt resistance.

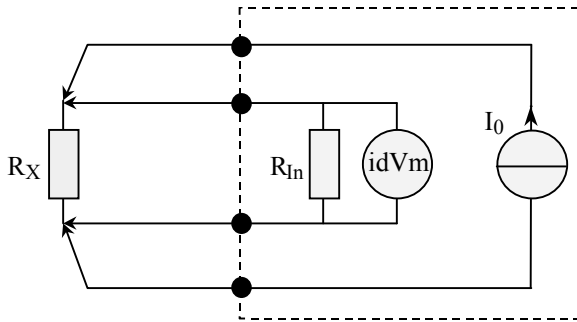
#### 8.1.1.4. Resistance measurements

The measurement of a resistance is usually obtained in a multimeter by crossing it with a known current and measuring the voltage at its limits. It is important to have a good calibrated current source.



**Figure 8.4.** Principle of the multimeter: resistance measurement

A problem occurs when measuring low resistances. In this case, the resistance of the measurement wires is of the same order of magnitude and introduces a significant error (drop of potential in the wires). Here we use the “4 point” method.



**Figure 8.5.** Multimeter principle: resistance measurement by the 4 point method

8.1.1.5. Two types of multimeters

In situations of variable voltage and intensity measurements, two types of multimeters are used: averaged response multimeters (sensitive to the averaged response of the corrected signal) and *root mean square* (RMS) multimeters with true, effective responses that are sensitive to the square of the signal. Multimeters sensitive to the averaged value are calibrated to display the effective value for a sinusoidal signal. This means that with other signals there is a measurement error to correct; this is done by taking the form factor into account.

One basic feature to consider is the passband of the device, which for multimeters is a fairly general base. For the HP34401A multimeter, the passband has 100 kHz of voltage and 5 kHz of intensity (for an accuracy below 1%).

The measurement of the effective value takes into account the effective value of the alternating current  $V_{f-ac}$  and that of the direct current  $V_{f-cc}$ . We especially need to

separate the two in cases of an end face alternative value of a component remaining high. In this case, we get the true effective value.

$$V_f = \sqrt{V_{f-cc}^2 + V_{f-ac}^2} \quad [8.2]$$

Signals having high comb factors (relation between the comb value and the effective value) set at relatively high frequencies pose error risks, due to the presence of fairly high harmonics.

#### 8.1.1.6. *Measurement accuracy*

Other measurement error sources are often due to ground loops, to thermoelectric effects due to contacts between different kinds of metals, to problems of common mode rejection (isolation of the LO limit in relation to the ground) and to sector noise in the case of a device in the sector.

Of course, the fact that a digital value is displayed is not a sign of absolute accuracy. Measurement accuracy is not directly accessible with a device, as was the case with magnetoelectric instruments with moving coils indicating the class of the instrument. It is necessary now to refer to the device's instructions for information concerning accuracy, taking into account a percentage of the measurement, of the range, and even the number of digits. Accuracy is expressed as:

$$\pm (\% \text{ of the measurement} + \% \text{ of the range}) \quad [8.3]$$

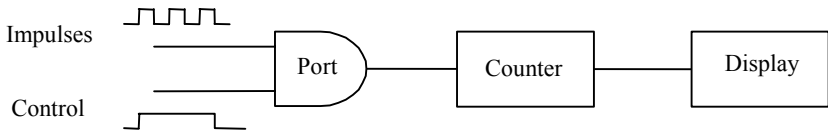
Still using the example of the HP34401A multimeter, to measure the direct voltage on a caliber 1 V, the accuracy for a year is given as  $0.0040 + 0.0007$  or  $0.0047\%$  full-scale maximum. This means an accuracy of  $4.7 \cdot 10^{-5}$  V for a 1 V measurement.

Accuracies are less precise measuring resistances or currents than in measuring voltages, and are less precise measuring variable signals than direct signals.

#### 8.1.2. *Frequency meters*

A frequency meter is organized around an accuracy time base and a counter. A good time base has a quartz oscillator for a reference. This oscillator can be at ambient temperature (we speak of *room temperature crystal oscillator* (RTXO)). This gives a stability of up to 2.5 ppm. It can be temperature compensated

(*temperature compensated crystal oscillator (TCXO)*); in this case the stability drops to 0.5 ppm. The oscillator also can be stabilized in a temperature-controlled enclosure, and here we speak of an *oven controlled crystal oscillator (OCXO)*. The stability of an OCXO can reach values of 0.01 ppm.



**Figure 8.6.** *Synoptic of a counter frequency meter or periodmeter*

Whether frequency or periods are being measured, one synoptic can be given for the instrument being used. This synoptic is shown in Figure 8.6. In this figure, the synoptic counts the number of impulses during the high state period of the control signal.

To measure the frequency of a periodic signal means we refer to the definition of the frequency by counting the number of times the signal is reproduced during an interval of known time (which can be a second or some other established interval). In this case, the signal we want to measure is found after it is formatted as an input “Impulses”. The time base gives the “Control” signal shown in Figure 8.6. This is the frequency meter functioning mode.

For measuring high frequencies, we can introduce a frequency divider to determine the number of times a reference signal (a clock) is reproduced identically. This mode is used for measuring very low frequencies. In this case, the time base (clock) creates the “Impulse” signal and the input signal, which serves as the “Control” signal shown in Figure 8.6. This is functioning in periodmeter mode.

Reciprocal counting functioning combines the advantages of the two other modes. A counting window is opened by the input signal, and it stays open during a time  $T_{CK}$  that is set by the time base. During the period  $\Delta T_F$  of this window, a first counter sets the number  $N_p$  of events of the input signal. A second counter determines the number  $N_{CK}$  of periods of the time base (clock). Since  $\Delta T_F = N_{CK} \cdot T_{CK} = N_p \cdot T$ , we get the signal’s period:

$$T = \frac{T_{CK} \cdot N_{CK}}{N_p} \tag{8.4}$$

The input impedance is usually  $1\text{ M}\Omega$  in low frequencies or  $50\ \Omega$  beyond 50 MHz. We find a direct DC coupling on inputs or an AC coupling that suppresses the direct component.

Here we mention a little-known but useful instrument of the same type. It is an analyzer of time intervals or frequencies that helps to visualize the temporal developments of frequencies, of periods, and phases of signals. It carries out a measurement by counting the time interval between each zero crossing of the input signal. We can then follow the behavior of a signal whose frequency develops over time. We can also note the stability and dynamic behavior of different components or systems such as a VCO (variations of the frequency according to temperature), and the dynamic behavior of a phase-locked loop.

### 8.1.3. Oscilloscopes

#### 8.1.3.1. Introduction

Today, since oscilloscopes are usually digital oscilloscopes, we will only discuss this type. The synoptic of a oscilloscope is shown in Figure 8.7.

After selecting the coupling mode as input (DC for direct current and AC for the part of the signal without a direct current), the principle chain is connected to a preamplifier (G), a sample and hold (Samp.B), an analog-to-digital converter and a memory storage device. The converter usually has an 8-bit accuracy.

The criteria which fix the frequency of the oscilloscope are the analog pass band of the input stages and the sampling frequency. Rise time is also a possible criterion but is seldom used anymore.

Sampling can be carried out either in real time (or in single-cycle mode) or repetitively (in sequential or random mode). In the first case, in order to respect Shannon's theorem, we must have  $f_s > 2.BW$  where  $f_s$  represents the sampling frequency and BW the length of the signal band. In the second case, we must implement a reconstruction procedure and we have  $f_s > k_R.BW$  where  $k_R$  is equal to 2.5 for a sinusoid, around 4 for a transient regime, or even 10 when there is no reconstruction. We find oscilloscopes with sampling frequencies of up to 20 MEch/s (as with series HP54600 made by Agilent Technologies<sup>TM</sup>) to 2 GEch/s (the HP54615 and Infinium).

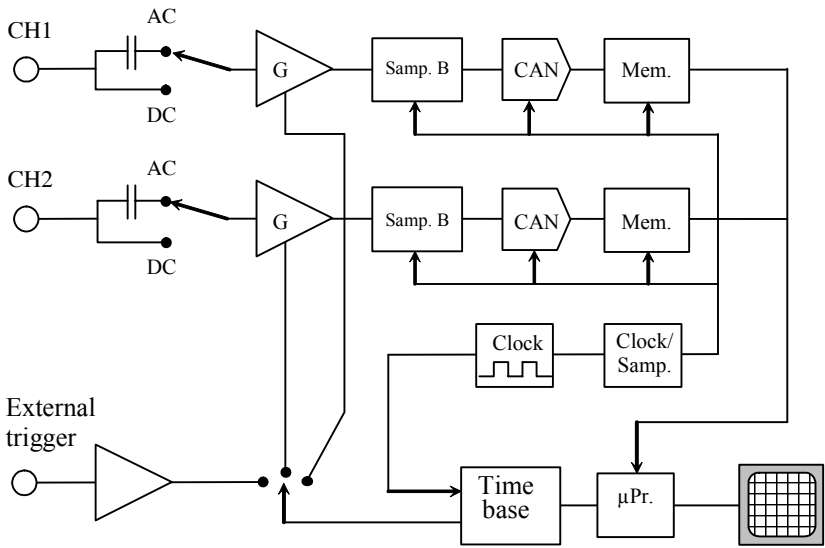


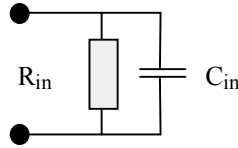
Figure 8.7. Synoptic of a digital oscilloscope

After the analog-to-digital conversion, the data are stored in a memory unit that currently can be up to 1 Mega-samplings. Recent instruments almost always have an interface (serial or IEEE488) that facilitates data transfer to a calculator.

There are several visualization modes for signals: by points or vectorial (the points obtained are linked). We can also carry out an averaging of the data to improve the signal-to-noise ratio on in general 8, 64 or even 256 acquisitions. The visualization of modulated signals (amplitude or angular modulation) can cause some problems with display. This is due to the fact that multiple frequencies occur during reconstruction. With such cases we must be careful to avoid taking a vectorial display so as not to place too much importance on falsely correlated phenomena. A single-cycle sampling can be most practical here, since the passband is strictly limited by Shannon’s theorem.

8.1.3.2. Input impedance and measurement

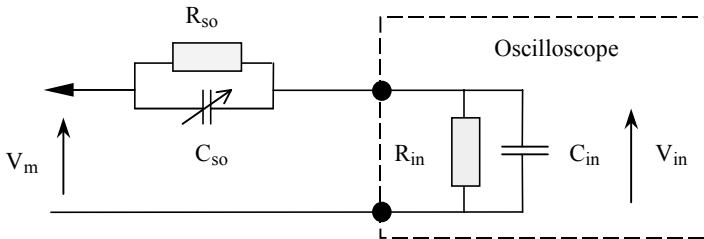
The standard model of an oscilloscope’s input impedance is shown in Figure 8.8. The input resistance of an oscilloscope is generally equal to 1 MΩ (normalized value). With oscilloscopes that have high bandwidths (above around 500 MHz), we can also choose an input impedance of 50 Ω.



**Figure 8.8.** Model of an oscilloscope's input impedance

The input capacitor has a capacity with an order of magnitude of 15 pF, but this value is not normalized (and varies from around 10 to 20 pF). If we carry out the measurements with a standard coaxial cable (with a line capacity of 100 pF/m), the capacitor brought to the measurement point is therefore of 15 pF + d\*100 pF where d is the cable length. For a coaxial cable of 1 m this can come to 115 pF, which can prove to be counter-productive. To resolve this problem, we can use measurement probes.

The model of the probe used with the input stage of an oscilloscope is shown in Figure 8.9. For this model, we can establish the corresponding transfer function that links the voltage analyzed by the oscilloscope (written as  $V_{in}$ ) to the voltage to be measured (written  $V_m$ ).



**Figure 8.9.** Modelization of the measurement probe used with an oscilloscope

$$\frac{V_{in}}{V_m} = \frac{R_{in}}{R_{in} + R_{so}} \frac{1 + jR_{so}C_{so}\omega}{1 + j\frac{R_{so}R_{in}}{R_{so} + R_{in}}(C_{so} + C_{in})\omega} \quad [8.5]$$

Regulating the probe to obtain an all-pass filter means adjusting the capacitor  $C_{so}$  so that:

$$R_{so}C_{so} = R_{in}C_{in} \quad [8.6]$$



The impedance brought to the measurement point is then:

$$Z_{eq} = \frac{R_{so} + R_{in}}{1 + jR_{in}C_{in}\omega} \quad [8.7]$$

This impedance is equivalent to a resistance  $R_{so} + R_{in}$  in parallel with a capacitor of the capacity  $C' = \frac{R_{in}}{R_{so} + R_{in}} C_{in}$  and is therefore lower than  $C_{in}$ , the total capacity of the coaxial cable, which is itself smaller than that of a impedance cable of about  $50 \Omega$ .

So, with  $R_{in} = 1 \text{ M}\Omega$ ;  $R_{so} = 9 \text{ M}\Omega$ ;  $C_{in} = 12 \text{ pF}$ . In this case,  $C_{so}$  must be adjusted to  $1.33 \text{ pF}$ , which gives an input impedance probe + oscilloscope equivalent to  $10 \text{ M}\Omega$  in parallel with  $1.2 \text{ pF}$ . The capacity of the cable being used for the probe is often around  $7$  to  $8 \text{ pF}$ , which gives the ensemble a capacity below  $10 \text{ pF}$ ! But of course, even the smallest disturbance brought to the assembly is obtained to the detriment of an attenuation by  $10$  of the voltage to be measured. There are also active probes that present very high input impedances and very low disturbance capacities (the measurement element is a MOS transistor).

### 8.1.3.3. *Measurements done by an oscilloscope*

Once the signal displayed on a screen is digitized, it is relatively easy to introduce measurement functions into an oscilloscope. These functions have different aspects. In voltage measurements, we most often find measurements of peak voltage, of effective or average voltage, the final value (high or low), or a squared signal. With temporal measurements we find the period, the frequency, the rise time and fall time of a signal. Related to these measurement functions we also find mathematical functions that are applied to signals: sum; difference; product; FFT (fast Fourier transform, the calculation of the coefficient differential or integrant).

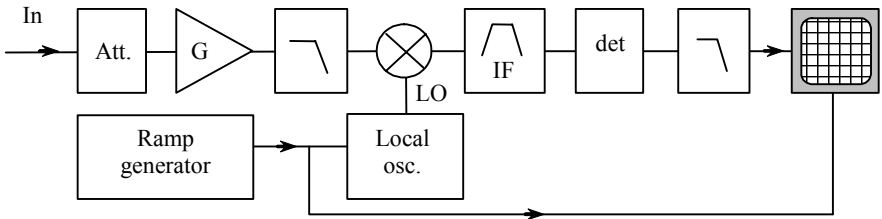
## 8.1.4. *Spectrum analyzers*

There are mainly two methods of carrying out a frequency analysis of a signal. With low frequencies (up to a few  $100$  of  $\text{kHz}$ ), we used an analyzer based on the calculation of the Fourier transform of the signal, which has been digitized beforehand. For radiofrequencies, high frequencies and microwaves, we use a sweeping analyzer (a technique using several  $100 \text{ kHz}$  up to  $100 \text{ GHz}$ ).

### 8.1.4.1. *Sweeping analyzers*

The principle of a sweeping spectrum analyzer is more or less the same as that of a heterodyne radio receptor.

Instead of regulating a tunable filter to cover the signal we want to measure, it is better to pass the spectrum through a fixed filter. We then combine (or multiply) the signal with another signal that carries out a frequency sweeping. The frequencies obtained by structure (sum of frequencies) are detected when they are equal to the central frequency of the selective filter (also called the intermediate frequency (IF)). The principle is shown in Figure 8.10.



**Figure 8.10.** *Schema principle of the sweeping spectrum analyzer*

We see that the most important unit in this process is the mixer, the signal to be studied having been applied after formatting on the RF input and the sweeping signal having been formatted on the LO input. Since the signals we apply on the RF input of a mixer must come up to certain levels, the input has an attenuation stage and an amplifier. It can have several successive conversion stages before optimally adapting the frequency range of the signal to be studied to the frequency of the selective filter. Each stage has a mixer, a local oscillator and an intermediary frequency filter. With recent instruments, we use a digital filter for the last IF stage that facilitates high stability in the filter, even at very low resolutions like 1 Hz. In this case, the final section is digital, including the peak detection and the display control.

After the selective filter, there is a peak detector for finding the amplitude of selected lines, then a filtering before beginning the visualization process. Recently, in some microprocessing variations, the signal is digitized after the video filtering and a microprocessor controls the local oscillator and display functions.

The main parameters to regulate are the central frequency, the frequency span, the resolution (length of the selective filter or resolution bandwidth, written as  $BW_{res}$ ) and the sweep rate, written as  $SW_r$ . However, we should remember that the higher the response, the more selective the passband filter. We have to “wait” for the output signal to go through transient regime before being able to correctly carry out peak detection. So, if we want to improve resolution (that is, separate the close lines), we must increase the sweep time so that the span parameters, sweep rate and resolution cannot be regulated independently. The resolution values usually can be

regulated by sequences 1, 3 (100 Hz, 300 Hz, 1 kHz, etc.). We can give an approximate formula between sweep rate and resolution:

$$SW_r = \frac{BW_{res}^2}{k_F} \tag{8.8}$$

where  $k_F$  is a kind of form factor dependent on the type of filter used (for a Gaussian filter we have  $k_F \approx 2$ ).

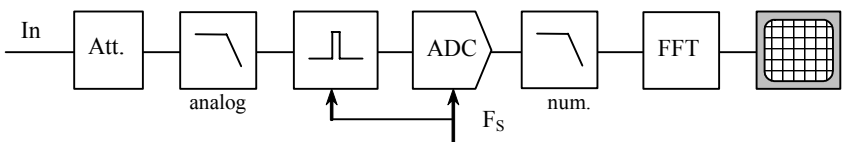
For the adjustments concerning detected amplitude, we find the reference, that is the maximum value displayed at the top of the screen, the unit chosen for the display and the scale (quantity displayed by division). The unit of amplitude can be in volts (or its sub-multiples), in dB (with different variants dB, dB<sub>v</sub>, dB<sub>m</sub>, and so on, according to the reference chosen for the calculation of decibels). We see that display in dB (logarithmic) is the most widely used, taking into account the very wide gaps possible between amplitudes of different peaks.

Sweeping can be done by a analog VCO, but today it is more often carried out by frequency synthesis devices with frequencies that are calculated according to the span and number of measurement points. With these devices it is possible to average the measurements carried out at each of the frequencies; this increases the signal-to-noise ratio.

Spectrum analyzers with sweeping functions have another important option: tracking generators. These are generators that supply a sinusoidal signal, of constant amplitude, that varies linearly over the course of time (but are synchronized on sweeping by the frequency of the analyzer itself). We can, for example, apply this signal to the quadrupole input. The spectrum measurement of the output signal of this quadrupole directly shows the response curve in amplitude of the quadrupole being studied.

#### 8.1.4.2. FFT analyzers

The schema principle for this instrument is shown in Figure 8.11.



**Figure 8.11.** Schema principle for a spectrum analyzer using FFT

The signal is first attenuated or amplified (according to its amplitude), then filtered to avoid spectrum aliasing. Then the signal is sampled and digitized (with an ADC sample and hold) to a frequency written as  $F_S$ . Lastly, the system calculates the Fourier transform of the number obtained by a rapid algorithm called the FFT. Shannon's criterion must also be respected, so we get:

$$F_S > 2F_{\max} \quad [8.9]$$

One important advantage of the FFT analyzer is its rapidity, since it establishes all the components of the spectrum in frequency in one time; the measurement speed is, at equal resolution, well above that of a sweep analyzer. Another advantage is that it also allows us to obtain a good resolution even at low frequencies (as low as Hz), which would be impossible with a sweep method.

It is important to remember one of the problems basic to the calculation principle of the Fourier transform. The calculation is applied to the digitized signal on a finite interval of time. This means it is seen in a certain temporal window. The obtained result is thus the convolution product of the Fourier transform of the signal, which itself is the product of the Fourier transform of the window. A simple rectangular or uniform window will then appear from the sinus functions  $(x)/x$  at each peak with relatively high lobes (Gibbs phenomenon) that risk flooding lower neighboring peaks. We then have the possibility, according to the quality criterion selected, of using different forms of windows to improve our results.

The windows proposed by FFT analyzers are the following: uniform; Bartlett (or triangular); Hanning (in cosine); Hamming; Blackman; Kaiser; and Flattop. Each type of window has certain advantages (such as fewer secondary lobes, a good respect for the maximum value, among others), but these are to the detriment of the length at half-maximum or to the measurement accuracy of the amplitude. Here, we give the expressions of some of these windows, by the function  $w(n)$ , defined as  $0 \leq n \leq N - 1$ , and which is zero outside,  $n$  representing the number of the sample [OPP 74].

$$\text{Uniform: } w(n) = 1 \quad [8.10a]$$

$$\begin{aligned} \text{Bartlett: } w(n) &= \frac{2n}{N-1} \quad \text{for } 0 \leq n \leq \frac{N-1}{2} \\ w(n) &= 2 - \frac{2n}{N-1} \quad \text{for } \frac{N-1}{2} \leq n \leq N-1 \end{aligned} \quad [8.10b]$$

$$\text{Hanning: } w(n) = \frac{1}{2} \left[ 1 - \cos\left(\frac{2\pi n}{N-1}\right) \right] \tag{8.10c}$$

$$\text{Hamming: } w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{8.10d}$$

$$\text{Blackmann: } w(n) = 0.42 - 0.5 \cos\left(\frac{2\pi n}{N-1}\right) + 0.08 \cos\left(\frac{4\pi n}{N-1}\right) \tag{8.10e}$$

$$\text{Kaiser: } w(n) = \frac{I_0 \left[ \omega_a \sqrt{\left(\frac{N-1}{2}\right)^2 - \left[n - \left(\frac{N-1}{2}\right)\right]^2} \right]}{I_0 \left[ \omega_a \left(\frac{N-1}{2}\right) \right]} \tag{8.10f}$$

The Kaiser is defined from the Bessel function of order zero, of the first type. The parameter  $\omega_a$  allows us to adjust the compromise between width of the central lobe and amplitude of the secondary lobes.

### 8.1.4.3. Principles of possible measurements

Here we give some of the main applications of spectrum analyzers.

Of course, we must mention direct spectral studies of a signal, for example the signal delivered by an oscillator or modulated signals of type AM or FM. These help us establish the spectral dimension around the carrier. Some spectrum analyzers even include demodulation functions (one example is the ESA1500 made by Agilent Technologies<sup>TM</sup>).

When we study the output signal of an amplifier, the spectrum analyzer allows a certain number of measurements. The measurement of harmonic distortion rate is worth mentioning here. This is an indicator of the relation of the energy contained in all the harmonics and the energy contained in the fundamental; the results are usually shown in percentages. The measurement is limited to a few harmonics, with the analyzer offering the possibility of choosing the number (here we cite the example of the FFT analyzer SR760 made by Stanford Research Systems<sup>TM</sup>).

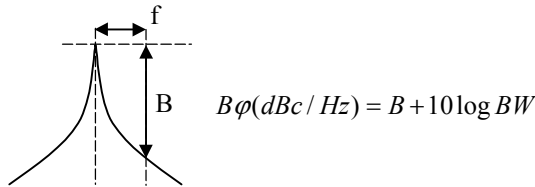
$$THD = \frac{\sqrt{\sum_{2..i..} V_i^2}}{V_1} \tag{8.11}$$

A very important measurement concerns the *power spectral density* (PSD). This is the amplitude normalized to 1 Hz of passband (expressed as  $V/\sqrt{Hz}$  or in  $dB/\sqrt{Hz}$ ). This measurement gives us a result independent of the span (an example of this type of signal analyzer is the HP89410 made by Agilent Technologies™).

The study of oscillators requires another type of measurement carried out with a spectrum analyzer: the measurement of phase noise. This measurement lets us encode the spectral purity of an oscillator, or the sharpness of the line corresponding to the oscillation frequency. If the resolution of the spectrum analyzer filter is written as BW and if we carry out the measurement at a distance  $f_x$  of the central line or carrier  $f_c$ , the phase noise is then:

$$L(f)(dBc / Hz) = P(f_x)(dBm) - P(f_c)(dBm) - 10 \log(BW / 1Hz) \quad [8.12]$$

Figure 8.12 shows the principle of the measurement of phase noise done with a spectrum analyzer. The measurement is carried out with a gap  $f$  in frequency in relation to the central frequency of the oscillator (or of the carrier in a transmission system).



**Figure 8.12.** Measurement of phase noise done by a spectrum analyzer

### 8.1.5. Network analyzers

#### 8.1.5.1. S parameters

With high frequencies (radio frequencies or hyper frequencies), it is imperative to take into account the propagation effects on the transmission lines, and even, in some cases, on the components. In the case of lines (bifilar, coaxial, microstrip, and waveguide), we define the incident and reflected waves as a line plane ( $V_i, V_r$ ) the reflection coefficient  $\rho = \frac{V_r}{V_i}$ , the characteristic impedance  $R_C$ , the reduced

impedance  $\underline{z} = \frac{Z}{R_C}$ .

We define the incident traveling wave by  $a = \frac{V_i}{\sqrt{R_c}}$  and the reflected traveling wave by  $b = \frac{V_r}{\sqrt{R_c}}$ .

Remembering that fundamental formulae link  $\rho$  and  $z$ :

$$\rho = \frac{z-1}{z+1} \text{ and } z = \frac{1+\rho}{1-\rho} \tag{8.13}$$

The chart allows  $\rho$  to go to  $z$ , as well as the reverse, according to the previous formulae. This is Smith’s chart.

The matrix S is made of scattering parameters or S parameters representing the way energy enters into a multiport system and is shared at output at the level of these ports:

$$(b) = [S](a) \tag{8.14}$$

or also:

$$b_i = \sum_j S_{ij} a_j \tag{8.15}$$

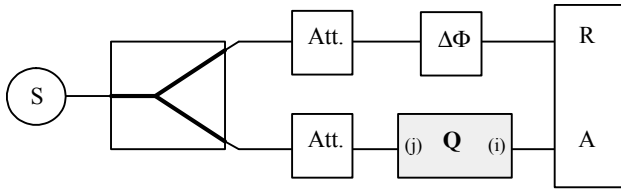
$S_{ii} = \left( \frac{b_i}{a_i} \right)_{a_j=0}$  is therefore the reflection coefficient of the port (i) when the ports (j) are matched to  $R_c$ .

$S_{ij} = \left( \frac{b_i}{a_j} \right)_{a_k=0}$  is the transmission coefficient of the port (j) towards the port (i) all the ports (k ≠ j) being matched to  $R_c$ .

### 8.1.5.2. Measuring S parameters

Measuring the S parameters of a quadrupole is done with a network analyzer called a vectorial analyzer when it carries out an amplitude and a parameter phase measurement. We say that we are using a system that helps us compare amplitude and phase of two incident waves, as two inputs written as R and A.

8.1.5.2.1. Measuring  $S_{ij}$



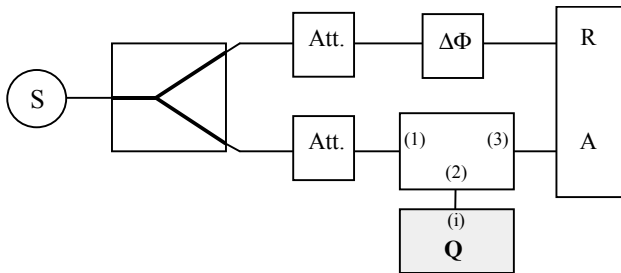
**Figure 8.13.** Measurement of a reflection coefficient

At first, an all pass replaces Q. The attenuators and the phase shifter are adjusted to regulate  $R = A$ . Then we place Q, R is unchanged and  $A' = S_{ij}^* A$  so that we get  $S_{ij} = \frac{A'}{R'}$ , and from this the module and the phase of the transmission coefficient.

8.1.5.2.2. Measuring  $S_{ii}$

We replace Q with a reflectometer bridge, characterized by its matrix S:

$$S = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{pmatrix} \tag{8.16}$$



**Figure 8.14.** Measurement of a reflection coefficient

Port (3) is matched, so  $a_3 = 0$ . Port (2) is in open circuit. We adjust  $R = A$  as before, and we have  $R = A = \frac{1}{4} a_1$ . Then we place the port (i) of the relevant



quadrupole at port (2) of the reflectometer.  $R = R'$  and  $A' = S_{ii} \frac{1}{4} a_1$  and so  $S_{ii} = \frac{A'}{R'}$ , from which we get the module and the coefficient reflection phase.

Because of coupling possibilities between the paths R and A, and of defects at the connector levels, as well as defects of the bidirectional coupler, it is necessary to proceed to calibration, that is to pre-measurements, with all the elements perfectly known (open circuit, short-circuit, adapted charge and all pass). The instrument deduces from the error vectors, taking into account the error sources and making corrections during later measurements.

It is very important to define the components' measurement plans as clearly as possible during calibration.

The S-parameter test set is an essential part of a network analyzer and can be integrated either to the device itself or in casing. This unit ensures all successive connections of the source, of the quadrupole, of the reflectometer to ports (i) and (j) alternately, and allows us to determine the four parameters  $S_{ii}$ ,  $S_{ij}$ ,  $S_{ji}$  and  $S_{jj}$ .

The results obtained can be shown on the analyzer screen unit in the form of a Smith chart, mainly for the reflection coefficients or in the module and phase form for the transmission coefficients. From the transmission coefficients  $S_{ii}$  we establish the input impedance at the level of port (i).

### 8.1.6. Impedance analyzers

#### 8.1.6.1. Method using a self-equilibrated bridge

This method, or at least its principle, can be described by the somewhat simplified schema shown in Figure 8.15.

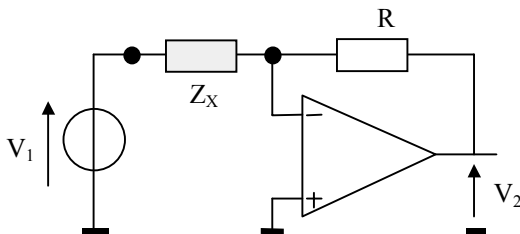
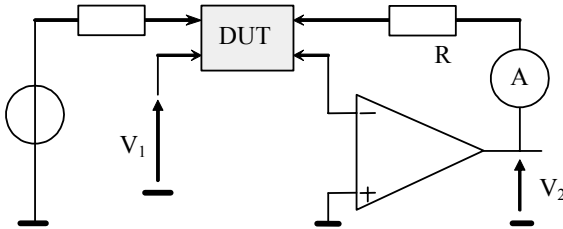


Figure 8.15. Measurement of impedance, self-equilibrated method

The impedance to be measured  $Z_X$  is obtained by the transfer function of the assembly:

$$\underline{Z}_X = -R \frac{V_1}{V_2} \quad [8.17]$$

There is also a four-point method for very low impedances, which is quite similar to that used for multimeters.



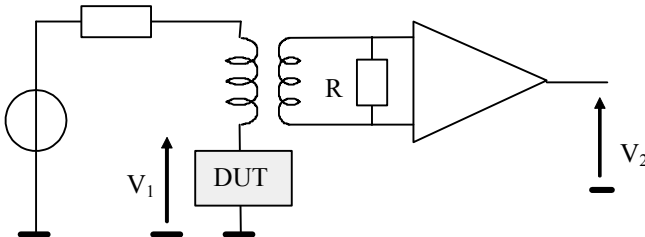
**Figure 8.16.** *Measurement of impedance, self-equilibrated four point method*

We get a high accuracy (up to 0.05%) during resistance, conductance, inductances, capacity, quality factor and loss angle measurements. This method has its upper limit at around 100 MHz. It is this method that is used by the HP4192A analyzer manufactured by Agilent Technologies™ that functions from 1 MHz to 13 MHz.

#### 8.1.6.2. RF I-V method

This method is used from around 1 MHz to 2 GHz, with an accuracy of about 0.8%. The range covered is relatively large, from 0.1 to 50 kΩ.

In a way similar to the voltamperometric method of measuring resistances, we find test plates of the “long shunt” type for high impedances, or “short shunt” for low impedances (see Figure 8.17).



**Figure 8.17.** *Impedance measurement by RF I-V method*

The impedance to be measured is given by the following formula:

$$\underline{Z} = R \frac{V_2}{V_1} \tag{8.18}$$

This method is used by the HP4291B analyzer made by Agilent Technologies™.

8.1.6.3. Measurement with a network analyzer

Measurement of impedances can also be done with a network analyzer at high frequencies, from 1 MHz, but more appropriately above 1 to 2 GHz. However, these methods are not recommended for high impedances, since the measurement accuracy is around 3% for impedances close to 50 Ω.

The methods described below are in accord with E1A512 recommendations.

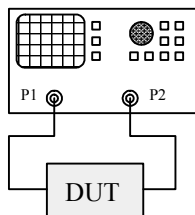
8.1.6.3.1. Measurement by reflection

This method measures the S<sub>11</sub> parameter and deduces the impedance from it by the basic relation linking impedance with the reflection coefficient.

$$\underline{Z}_X = R_C \times \frac{1 - S_{11}}{1 + S_{11}} \tag{8.19}$$

8.1.6.3.2. S parameters method

It is also possible to deduce the impedance of a complete measurement of S parameters (two port method).



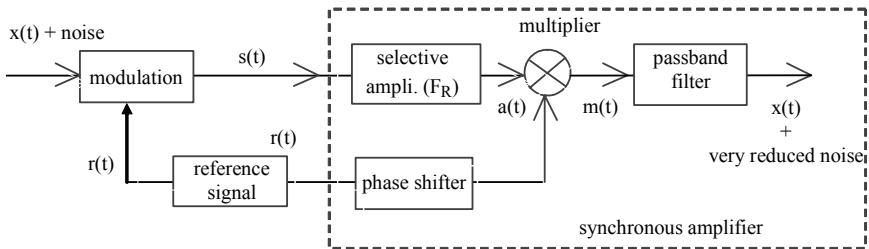
**Figure 8.18.** Measurement of impedance with a network analyzer

$$\underline{Z}_X = R_C \times \frac{(1 + S_{11})(1 + S_{22}) - S_{21}S_{12}}{2 \times S_{21}} \tag{8.20}$$

With all these methods it is important to closely observe the quality of the fixture systems of the components being tested (in particular for CMS components) and then proceed to a calibration step of the measurement system. This last step controls all stages leading up to and including measurement planning.

### 8.1.7. Synchronous detection

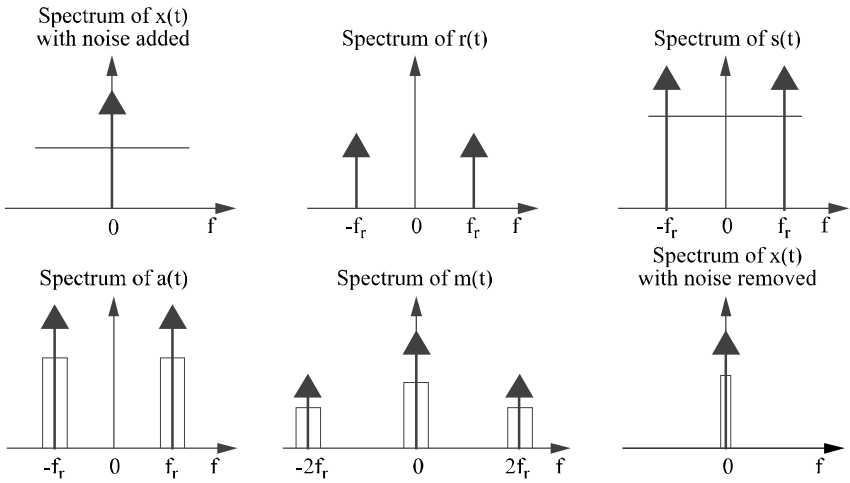
Using a synchronous amplifier (called a *lock-in amplifier*, which carries out detection by synchronous modulation) helps us eliminate  $1/f$  noise in circuits, as well as noise close to  $f = 0$ , on the condition that the direct or slowly varying useful signal  $x(t)$  can be modulated in amplitude to the source by a reference signal  $r(t)$ . Synchronous detection allows for a reduction in the signal-to-noise ratio.



**Figure 8.19.** Schema principle of synchronous demodulation detection

In fact, with the useful signal spectrum  $x(t)$  concentrated close to  $f = 0$ , the modulation of  $x(t)$  by the reference signal of the frequency  $f_r$  leads to the creation of a signal  $s(t)$  whose spectrum is that of  $x(t)$  but is shifted from  $f = 0$  to  $f = \pm f_r$ . The signal  $s(t)$  is then filtered by an amplifier tuned to the frequency  $f_r$ , which eliminates all the noise found outside the passband of the tuned amplifier. A synchronous detection using a multiplier, a phase shifter and a passband filter restores the useful signal  $x(t)$ .

To illustrate this method, let us consider a signal  $x(t)$  provided with a white noise. We suppose that the reference signal is purely sinusoidal. Figure 8.20 shows the spectrums obtained at the output of the different analysis units shown in Figure 8.19.



**Figure 8.20.** Schema principle of detection by synchronous demodulation

We see that this type of detection helps us measure the phase shifting between two signals of the same frequency, as well as the separation of real and imaginary components from a signal in the complex ensemble.

One of the important parameters of the *lock-in* amplifier is the time constant, which is directly linked to the cut-off frequency of the passband filters. These filters eliminate the  $2 f_r$  component, and also help reduce noise by diminishing the bandwidth. For example, the SR510/SR530 amplifiers made by Stanford Research<sup>TM</sup> have two passband filters. The first has time constants that are adjusted from 1 ms to 100 s for the first stage, and from 1 s to 0.1 s for the second stage. The bandwidth of the amplifier is of 100 kHz.

To improve measurement accuracy, we find the input of band reject filters set to the frequency of the sector. We then double this frequency.

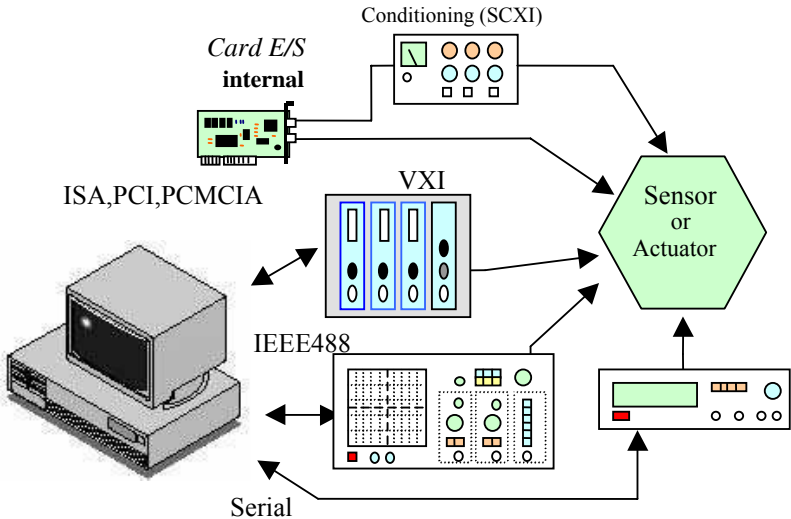
There are also digital lock-in amplifiers (one example is the SR810/830 made by Stanford Research<sup>TM</sup>). With these instruments, the synchronous detection is carried out digitally. The signal is sampled (after antialiasing filtering) at the maximum frequency of 256 kHz, then a DSP carries out the demodulation, that is, the “digital” multiplication of the signal sampled by the reference signal. The signal is then filtered. As an example, the DSP can carry out 16 million multiplications and additions per second on 24 bits. The time constant is adjustable from 10  $\mu$ s to 30 ks, the input passband going from 1 mHz to 102 kHz.

## 8.2. Measurement chains

### 8.2.1. Introduction

Previously we have seen the description of a number of measurement instruments which are currently in use. The user/instrument interface is reduced in many cases to its simplest expression; that is, to some buttons and a display. We can see why the microcomputer has become, over the years, an essential tool in a measurement chain. Linked to an acquisition card or driving an instrumentation bus, a computer gives its user numerous control and analysis capabilities.

In the field of instrumentation, we will look at the most up-to-date solutions available to the consumer. Figure 8.21 summarizes data acquisition systems.



**Figure 8.21.** Measurement chains connected to a microcomputer

Depending on the type of measurement and the environment in which the system functions, the designer can choose between different options. There is no one solution to any given problem, since today there are so many possibilities from which to choose. The new communication interfaces available to the public are reflected in the field of instrumentation, and so offer new possibilities. The serial bus USB is one of the latest examples.

### 8.2.2. *Communication buses PC/instruments*

Communication buses connecting microcomputers and instruments fall mainly into two categories: parallel and serial buses.

#### 8.2.2.1. *The parallel bus IEEE488*

Parallel transmission is the most usual way to transfer data between two devices in binary form. The high number of lines that are required, as well as the connecting technology needed, make this option relatively expensive, and its usage is limited to shorter distances.

However, this kind of transmission is quite suitable for measurement banks or for distances of several meters. For these applications, the IEEE488 bus, a standard instrumentation bus widely used today, was developed.

There are many devices (those previously mentioned) and almost 250 manufacturers (among them Agilent Technologies™, Tektronics™, National Instrument™) that offer interfaces and programs for this bus.

At the international level, the IEEE488 bus has different names:

- HPIB (Hewlett-Packard Interface Bus);
- GPIB (General Purpose Interface Bus);
- IEEE BUS, ASCII BUS or PLUS BUS.

##### 8.2.2.1.1. Specifications for the IEEE488 bus

The IEEE488 norm completely defines the electrical features and mechanics of this bus, as well as the exchange protocols.

This is a parallel bus with asynchronous communication (bit-serial parallel bytes), and is directed by a handshake system in which the slowest unit imposes its rhythm. It can achieve a maximum transmission speed of 1 Mbit/s.

The ensemble can be linked by starfish or chain connections, or can use a combination of the two. In all cases, for reasons of transfer speed, it is imperative to respect the following criteria:

- the distance between two devices must not exceed 2 m;
- the total length of the bus must be below 20 m.

We can connect a maximum of 15 devices, including the PC controller, and half of these devices must be powered on.

### 8.2.2.1.2. Architecture of the IEEE488 bus

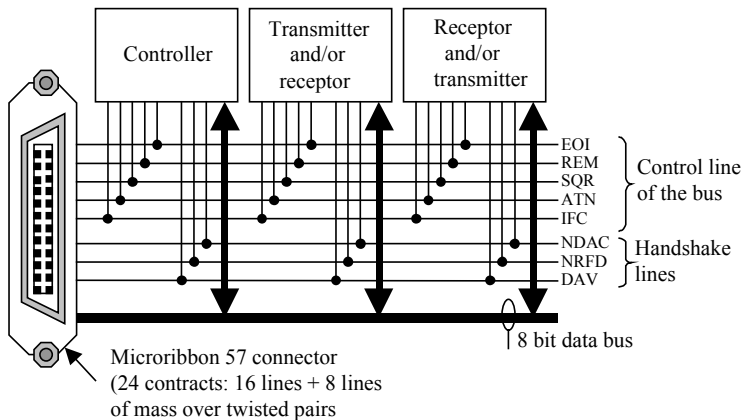
As shown in Figure 8.22, the instruments connected to the bus can have the following functions:

- They can have control functions. This is usually the card inserted into the microcomputer. Although many controllers also can be used in this way, they cannot be used at the same time as controllers. The controller organizes exchanges, configures the ensemble of devices, and ensures proper sequencing of operations throughout the measurement chain.

- They can have “talker” functions. The instrument transmits messages (continuation of binary words) to other instruments, most often towards the controller. This is true of all measurement instruments (multimeters, oscilloscopes, etc.).

- They can have “listener” functions. The instrument receives messages from the controller or from other instruments. This occurs during the configuration of a measurement instrument, as well as with graphic tracers.

An instrument is usually both transmitter and receptor, but the transmission direction is determined by the controller. Each device on the bus is identified by an address that is sent on the bus for each new transfer.



**Figure 8.22.** Architecture of the IEEE488 bus

The signals used for transmission are divided into three main levels, with subdivisions according to functions.

- there are eight data lines. Each is a bidirectional bus ensuring word, transmission, addresses and ASCII data;



- there are three handshake lines:
  - NDAC (*No Data ACcept*): the receiving device indicates if it will accept or reject received data,
  - NRFD (*Not Ready For Data*): this indicates if the device is ready to receive data,
  - DAV (*DAta Valid*): the transmitter confirms if the data presented over the bus are valid;
- there are five control line for the bus:
  - ATN (*ATtention*), or the control mode of the bus, depending on the mode of the data,
  - IFC (*InterFace Clear*) indicating the interface is in inactive mode,
  - REN (*Remote ENable*) authorizes the instruments to function in remote mode, meaning they are piloted by the bus,
  - SRQ (*Service ReQuest*) warns the controller that an instrument needs its attention,
  - EOI (*End Or Identifty*) indicates the last byte in a message.

Electrical signals have levels compatible with TTL standards and work in negative logic.

SRQ, NRFD and NDAC lines only use open collectors. For other lines, we find both open collectors and three-state buffers that help us obtain transfer speeds above 250 Ko/s.

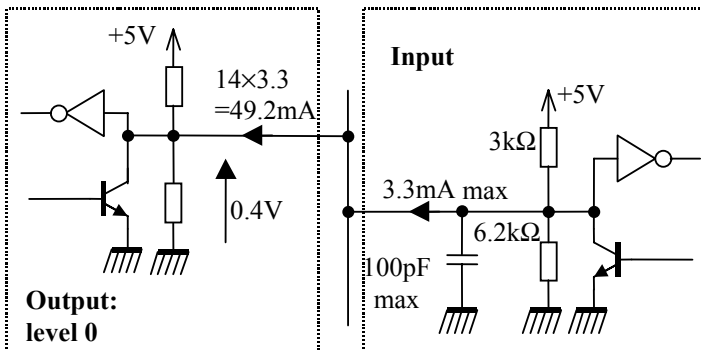


Figure 8.23. Electric linkage over the IEEE488 bus

### 8.2.2.2. Serial buses

In this transmission mode, the data pass sequentially bit by bit over the same conductor, their rhythm set by a transmitter clock. The main advantage of this method is the limited number of wires connected to small connectors, meaning it takes up less space. Serial buses are mostly used for transmission distances of between several meters and around 1 km.

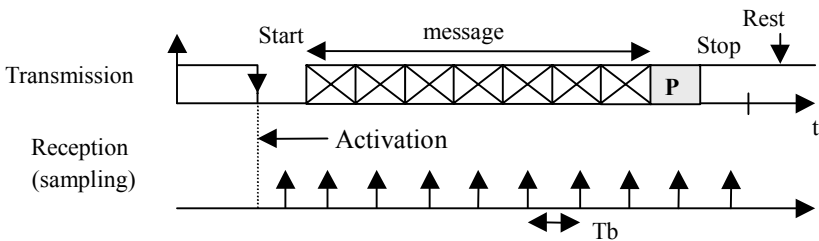
#### 8.2.2.2.1. Transmission modes

A line is called simplex if the transfer is always in the same direction; it is a duplex line for bidirectional communications. In a linked series, there are two data transmission modes corresponding to the synchronicity of the receiver clock in relation to that of transmission.

#### *Asynchronous mode*

The ensemble of bits to be transmitted is organized as a frame of about 10 bits maximum. Each bit has a duration  $T_b$  in time. This frame contains the following features:

- a low-level start bit;
- a message containing  $N$  bits, usually a type ASCII code, with  $N$  having between 5 to 8 bits;
- a control bit (*Checksum*) more or less equal to the transmitted message;
- one or two high-level stop bits.



**Figure 8.24.** *Asynchronous serial transmission*

Signal recuperation begins when the detection at the beginning ends as transmission begins. Reading the transmitted bits is done by sampling each bit in the middle of each basic period  $T_b$ . This process requires a good knowledge of the format being used and of the transmission flow ( $1/T_b$ ), expressed in bauds.

Controlling the message reception can be done by using specialized components of type UART (*Universal Asynchronous Receiver Transmitter*) or more often by a microcontroller.

This very simple transmission mode is widely used. However, it does not allow for significant flows (lower than 19,200 bauds, with a corresponding  $T_b$  period of 52  $\mu$ s). In the majority of cases, the transmitter is faster than the receiver, usually because the receiver takes longer to analyze data. However, the reception buffer is generally very limited. This means the signal receiver used with the transmitter must use a handshake system. This system can be material (one example is the *Data Terminal Ready* (DTR) protocol) or it can consist of software (such as the *Xon Xoff* protocol).

### *Synchronous mode*

Here, the transmitter and the receiver are synchronized, which means their transmission and reception clocks of identical frequencies and phases. We then can obtain very long and high speed message transmissions.

The transmission clock can be transmitted to the message or can also be combined with it. In this case, we use transcoders that, in a logical manner, convert a signal and the data to be transmitted from a clock. One of the most widely used of these codes is the Manchester code. The transmitted signal comes from the XOR of both clock and initial data signals. In this system, transmitted data can appear on leading and trailing edges over all clock periods. It then becomes easy to produce a receiver clock by synchronizing clocks by edge detection.

#### 8.2.2.2.2. Electric interfaces

Whatever occurs in transmission and receiving operations, electric signals are most often TTL compatible. This does not give good transmission conditions for distances longer than 1 meter. We then have to use electric interfaces that change and adapt the logical levels to be transmitted.

We can speak of two structures for driver connections and line receivers:

- Unbalanced structures, in which one conductor wire is used for transmitting a logical signal, as well as a ground conductor wire that can be shared when several transmission lines are necessary.

- Balanced structures, a mode in which two conductor wires are needed to transmit a logical signal. As with the unbalanced structure, one ground conductor wire is used. The major advantage of this structure is that it is relatively insensitive to environmental noise. For this reason it is widely used in industrial applications. As well, it allows larger flows than unbalanced structures do.

At this time, there are four important standards used in serial transmission systems. There are drivers and receivers of connected lines corresponding to each norm. Figure 8.25 shows the type of wiring for each norm needed for the transmission of logical signals (apart from the TIA/EIA485 norm, which permits bidirectional transmission).

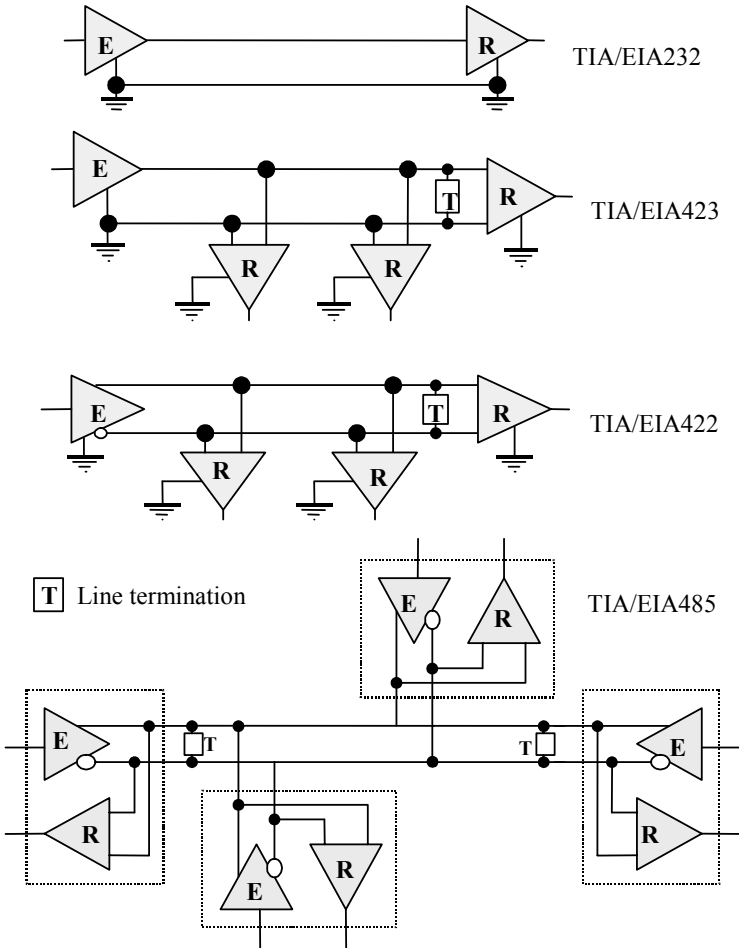


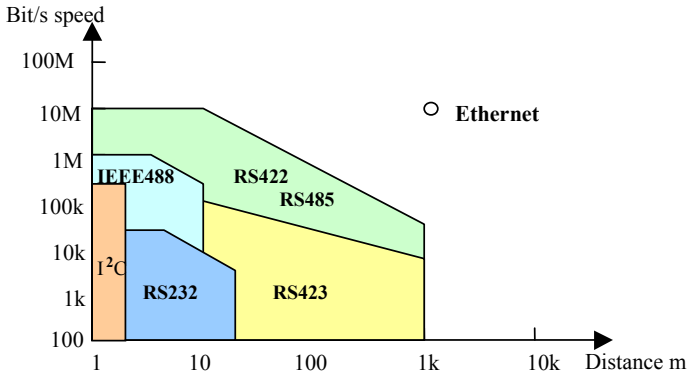
Figure 8.25. Standard transmission interfaces

As we can see by looking at Table 8.1, an essential element of buses is the speed-distance product. Figure 8.26 summarizes this feature for different types of serial and parallel transmission.

	EIA/TIA232	EIA/TIA423	EIA/TIA422	EIA/TIA485
Structure	Asymmetrical	Asymmetrical	Symmetrical	Symmetrical
Maximum line length	“Capa” line < 2,500 pF	1,200 m	1,200 m	1,200 m
Maximum flow	20 kbps	100 kbps	10 Mbps	10 Mbps
Maximum number of transmitters	1	1	1	32
Maximum number of receivers	1	10	10	32

**Table 8.1.** Essential features of serial transmission systems

The voltages we find are basically due to the line capacities used for carrying signals.



**Figure 8.26.** Speed/distance for different communication buses

### 8.2.3. Internal acquisition cards

These are electronic cards placed in extensions of computers dedicated to instrumentation. These are not only used for the acquisition of analog and/or logical signals but can also deliver control signals or command variables.

Their use is relatively simple and their costs are moderate, considering their capabilities and performances. The major advantage of these cards is their good transfer speed for data in terms of measurement and control. This means we can use real-time analysis.

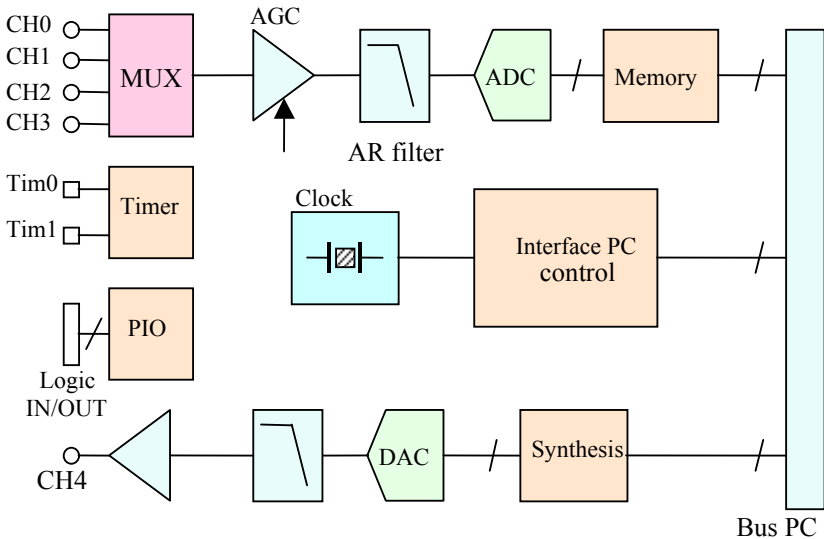
In general, the number of inputs/outputs is fairly limited and can pose some problems when a computer is not close to the process.

### 8.2.3.1. Description of inputs/outputs and associated conditioning

Depending on the process to be measured or controlled, it is vital to know:

- the nature of the information (analog or digital variables and their electric characteristics);
- the type of analysis (amplification, scaled, filtering, analog-to-digital conversion, memorization, etc.).

Figure 8.27 gives us a glimpse of a “universal” acquisition card showing a number of inputs and outputs, as well as the electronics used for analysis.



**Figure 8.27.** Universal acquisition card

The main features of this kind of acquisition card are as follows.

- For analog input/output:
  - bipolar or unipolar mode,
  - analog-to-digital converter (ADC) of 8, 12, or 16 bits,
  - digital-to-analog converter (DAC) of 8, 10, or 14 bits,

- a maximum sampling frequency of 1 Go/s,
  - a programmable gain with ratios of 1, 2, 4, 8, and/or 1, 10, 100, or 1,000,
  - number of channels covered by using a multiplexer (MUX),
  - depth of maximum memory of about 100 Mo.
- For logical input/output:
- input/output levels TTL/CMOS,
  - all or nothing input with the use of a optocoupler,
  - input counting or period measurement.

If signal conditioning is still hard to carry out, we can use external signal conditioning modules, such as the SCXI (*Signal Conditioning eXtension for Instrumentation*).

#### 8.2.3.2. Description of PC buses

According to the acquisition system to be inserted, transfers between the instrumentation card and the microprocessor of the PC can be done another way. There is a method called polling. In polling, the microprocessor is completely dedicated to the task of acquisition, which limits overall analysis time. This method uses interruptions. These interrupt the microprocessor only for the acquisition and memorization of data, the rest of the time being devoted to control and applications. By far the best-performing technique is the *Direct Memory Access* (DMA). In this method, there are specific controllers that direct acquisitions and transfers towards memory without the microprocessor and PC being involved. This leads to better performances, since the PC deals solely with applications.

The cards developed for instrumentation use the standard principles of computer extension cards.

– The first is the *Industry Standard Architecture* (ISA) bus. Although it is older, it is still used for instrumentation purposes. This is an asynchronous 16 bit bus set to a rhythm of 8.33 MHz, with a transmission rhythm that does not exceed 1 to 2 Mo/s because of cycles and interruptions. However, there are cards in ISA format that allow for sampling of signals at frequencies of the order of several tens of Mech/s that must be integrated with memory. The transfer towards the PC is then done according to a lower rhythm and does not allow for a real-time analysis.

– The second is the *Peripheral Component Interconnect* (PCI) bus, which was developed by Intel in 1993 and has been widely used since 1995. This is a 32 bit bus set at 33 MHz, allowing for a theoretical maximum flow of 132 Mo/s. This is higher than the flows allowed by the ISA bus, which explains its popularity for users

needing rapid acquisition cards. It also has “plug and play” features, eliminating the cordless plugs needed for cards of the ISA formats. Again, differing from the ISA system, the cards of PCI format integrate the DMA controller on the card itself, so the bus can have autonomous control; thus, these are called master cards. There are also “slave” cards that cannot integrate DMA functions in this way.

## **8.2.4. External acquisition cards: the VXI system**

### *8.2.4.1. Functions of the VXI bus*

The VXI system, developed from the VME, is widely used in modular instrumentation. VXI is an abbreviation of “*VMEbus eXtensions for Instrumentation*”. The specifications of this bus, which is dedicated to instrumentation, detail the technical constraints for compatible VXI systems; we find certain requirements in terms of chassis, signals used in load pockets, energy feed and modules that can be connected.

This system was first developed in 1987 by five manufacturers of electronic devices wishing to create a new standard for the industry. Their goal was to create an instrumentation system with the advantages of the VME bus and increase the capacities of the IEEE488 bus. They did this by increasing rapid activation between devices.

### *8.2.4.2. Description of the VXI bus*

The major advantage of the VXI is that it shares much of the format of the VME bus. This means it can be easily used in industrial applications. The two A and B card formats of the VME bus are kept, as well as the P1 connector and the center row of the P2 connector as shown in Figure 8.28. In order to offer the largest possible range of equipment, the VXI system allows for the addition of higher C and D formats, as well as a P3 connector that is specific to the VXI system.

Specifications for the VXI bus completely describe the P2 and P3 connectors. The P2 bus is made of several elements. It has a VME bus, a 10 MHz clock, activation lines with ECL and TTL logical levels, an analog summation line, a module identification line and a local bus. The P3 connector has supplementary lines for a local bus structure, a clock line set at 100 MHz, and high performance activation lines in ECL logic.

This design gives activation signals above 50 MHz and local data transfers of 100 Mb/s.



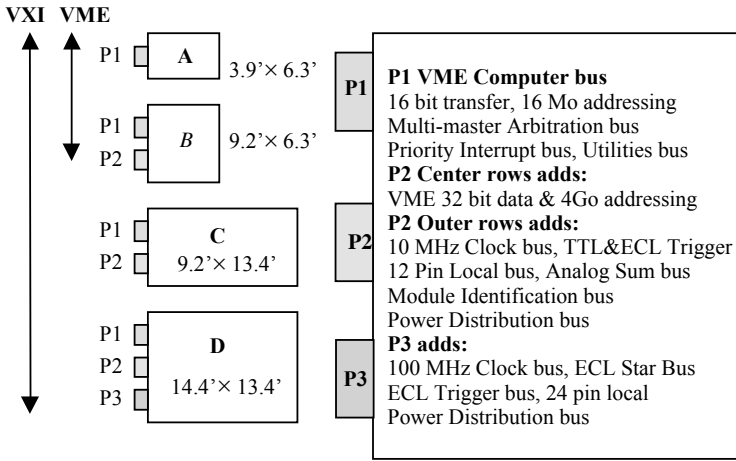


Figure 8.28. Connector specifications

A system developed with a VXI bus foundation can support up to 256 instruments, of which one or several are control and arbitration modules. These modules are usually found in the slot0 of the VXI chassis. Figure 8.29 shows the standard VXI chassis with 13 modules. In the basket of the chassis are P1, P2 and P3 connectors. The chassis integrates a feed that supplies different voltages offered in VXI specifications (+5 V, +12 V, +24 V, -2 V, -5.2 V, -12 V, -24 V) with different available powers.

The VXI system is very complete and offers a standard of performance for the field of instrumentation. However, it is still relatively expensive and therefore is found only in top-of-the-line acquisition and automatic test systems.

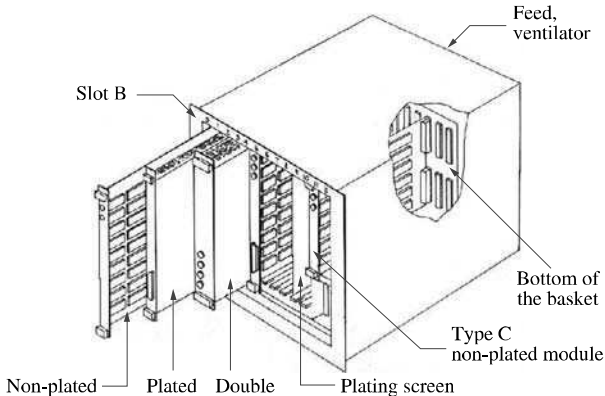


Figure 8.29. VXI chassis

### 8.3. Bibliography

- [ASC 99] ASCH G., *Acquisition de données, du capteur à l'ordinateur*, Dunod, 1999.
- [COM 96] COMBES P.F., *Micro-ondes*, volume 1, *Lignes guides et cavités*, Dunod, 1996.
- [COM 97] COMBES P.F., *Micro-ondes*, volume 2, *Circuits passifs, propagation, antennes*, Dunod, 1997.
- [COO 95] COOMBS C.F., *Electronic Instrument Handbook*, McGraw-Hill, 1995.
- [DIE 99] DE DIEULEVEULT F., *Electronique appliquée aux hautes fréquences*, Dunod, 1999.
- [IMH 90] *The Impedance Measurement Handbook*, Hewlett-Packard, 1990.
- [OPP 74] OPPENHEIM A.V., SCHAFFER R.W., *Digital Signal Processing*, Prentice Hall, 1974.
- [SEI 99] Scientific and Engineering Instruments, Stanford Research Systems, 1998-1999.
- [TMC 99] Test and Measurement Catalog 2000, Agilent Technologies, Dec. 1999.
- [WHI 93a] WHITE R.A., *Electronic Test Instruments – Theory and Applications*, Prentice Hall, 1993.
- [WHI 93B] WHITE R.A., *Spectrum & Network Measurements*, Prentice Hall, 1993.

*This page intentionally left blank*

## Chapter 9

# Elaboration of Models for the Interaction Between the Sensor and its Environment

### 9.1. Modeling a sensor's interactions with its environment

The focus of this chapter will be to describe the relation between a sensor's output variables and the physical variable applied to its input, called the measurand.

This relation can also take into account the role played by other variables, *a priori* external, that can cause variations in the output signal (some examples are the sensor's feed tension and the temperature).

#### 9.1.1. *Physical description of the model*

The best approach is first to analyze what a sensor does and then to understand and completely describe the physical processes of transduction. We then convert these into mathematical forms by using physical laws. The result is an equation that links output variables to input variables (creating a knowledge model).

This is a difficult task, requiring a complete understanding of all the phenomena involved. In general, this process is long, especially for complex phenomena, but this approach does have the advantage of being easily transposable to other, similar systems.

### 9.1.2. Phenomenological approach

This is an experimental method that consists of collecting data. The values of the output signals are compared to the values taken by the input variables under a set of given conditions. The study of the structure of these data helps us collect results in the form of mathematical relations (dependency models) that explain the observations.

This experimental approach is improved by well-organized experiments, with the goal of reducing the number of such experiments in order to keep a relevant meaning. This approach does have the disadvantage of not being easily transposable to other systems.

### 9.1.3. Adjustment

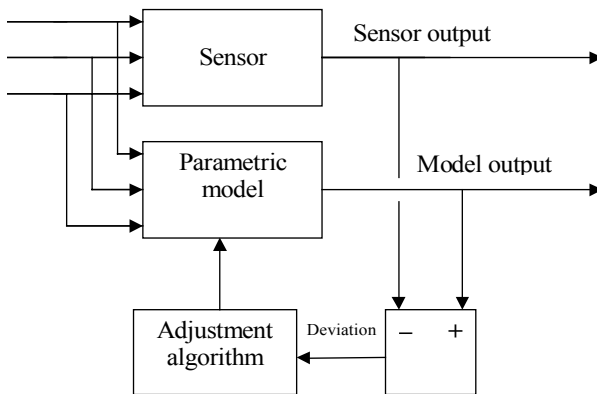


Figure 9.1. Data adjustment between a process and its model

In both cases, the method uses the following steps (see Figure 9.1):

- using either physical laws or a close observation of data, we establish a model (equation) that has a certain number of unknown parameters;
- by setting an optimization criterion (the least squares, for example), we look for parameter values that “at least” adjust the observed deviations between the sensor’s output signal and the output signal of the model;
- the study of deviations between the data and the adjusted variables allows us, for the first time, to verify the adequacy of the model and then, a second time, to estimate the limits of this adequacy in terms of variability.

## 9.2. Researching the parameters of a given model

### 9.2.1. The least squares method

Let us assume  $Y$ ,  $X_1$ ,  $X_2$  are the variables represented by the output signal of a sensor according to the input variables (signal, limiting quantity and so on). After a physical analysis, we know that there is a mathematical relation between these quantities, written as:

$$Y = f(X_1, X_2, \dots, \theta_0, \theta_1, \dots, \theta_k) \quad [9.1]$$

where  $\theta_0$ ,  $\theta_1$ , ...,  $\theta_k$  represent the parameter relation.

We can also observe the measurements carried out with this sensor and make the hypothesis that there is an expression such as the one in equation [9.1].

In both cases, the problem consists, apart from observations, of calculating the values of the parameter model. For example, the relation between the input signal  $X$  and the output signal  $Y$  can be a polynomial form of degree  $k$ :

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \dots + \theta_k X^k \quad [9.2]$$

where we estimate the  $k+1$  parameters from the  $n$  pairs of observations  $(x_1, y_1)$ ,  $(x_2, y_2)$ , ...,  $(x_n, y_n)$ .

If the number of pairs is equal to the number of parameters to be estimated, we have a linear parameter system to be estimated, with  $n$  equations for  $n$  unknowns. If the number of point pairs is lower than the number of parameters to be estimated, the values of some of these parameters can be chosen arbitrarily to resolve the problem.

The situation discussed here is when the number of points is strictly higher than the number of parameters to be assessed. Under these conditions, the  $n$  equations representing the  $n$  measurements cannot be resolved simultaneously. We then have to analyze the system:

$$y_1 = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \dots + \theta_k x_1^k + e_1 \quad [9.3]$$

$$y_2 = \theta_0 + \theta_1 x_2 + \theta_2 x_2^2 + \dots + \theta_k x_2^k + e_2 \quad [9.4]$$

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \dots + \theta_k x_i^k + e_i \quad [9.5]$$

$$y_n = \theta_0 + \theta_1 x_n + \theta_2 x_n^2 + \dots + \theta_k x_n^k + e_n \quad [9.6]$$

where the quantities  $e_1, e_2, \dots, e_n$  represent the deviations between the supposed or theoretical model and the effected measurements. The value of each of these deviations varies according to the optimization criterion adapted to analyze this problem.

The criterion used here will be the least squares method of Gauss, who described it as follows: “The estimator of the parameters  $\theta_0, \theta_1, \dots, \theta_k$  are the specific values that reduce to minimum the sum of the squared deviations between the experimental observations and the corresponding values predicted by the adopted theoretic model.”

This means we must form the quantity

$$Q(\theta_0, \theta_1, \dots, \theta_k) = \sum_{i=1}^n e_i^2 \quad [9.7]$$

which is a function of  $\theta_0, \theta_1, \dots, \theta_k$  and find the values of these parameters that minimize this function.

### 9.2.2. Application to estimate a central value

An example of a simple application of this principle may be provided by looking for the estimator of the central value parameter of a series of measurements.

Let  $x_1, x_2, \dots, x_n$  be the results obtained during  $n$  independent repetitions of the measurement of the same variable under the same experimental conditions. We make the hypothesis that these results are  $n$  specific numerical values of an expected variable  $\mu$  that we want to estimate (the expectation represents the ideal value that the variable has under ideal conditions, that is, without random disturbance). The obtaining conditions being identical, we can formulate the hypothesis that these results have the same variance  $\sigma^2$ . The  $n$  measurement results are translated by the system of the following  $n$  equations:

$$x_1 = \mu + e_1 \quad [9.8]$$

$$x_2 = \mu + e_2 \quad [9.9]$$

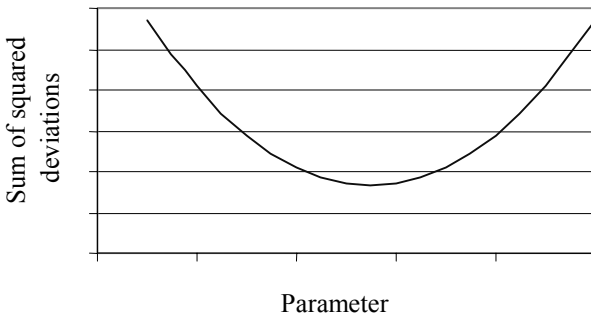
$$x_i = \mu + e_i \quad [9.10]$$

$$x_n = \mu + e_n \tag{9.11}$$

Applying the Gaussian criterion leads to forming the quantity:

$$Q(\mu) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (x_i - \mu)^2 \tag{9.12}$$

The graph representing the variation of  $Q(\mu)$  according to  $\mu$  is a parabola. (Figure 9.2).



**Figure 9.2.** Variation of the sum of squared deviations according to parameter

The minimum of this function is attained for the value  $\hat{\mu}$  of  $\mu$ , which is the solution of the equation obtained by writing that:

$$\frac{dQ}{d\mu} = 0 \tag{9.13}$$

which immediately gives us:

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} \tag{9.14}$$

In other terms, the estimator of the central value of the measurement results, in the sense of the least squares, is simply the average arithmetic  $\bar{x}$  of the values obtained. We know that this estimator is not biased; that is, its expectation is equal to  $\mu$ . The variance of this estimator is then:



$$V(\hat{\mu}) = V(\bar{x}) = \frac{\sigma^2}{n} \quad [9.15]$$

### 9.2.3. Introduction to weighting

Let us look again at the example of finding the estimator of the central value parameter of a measurement series.

Let  $x_1, x_2, \dots, x_n$  be the results obtained during  $n$  independent repetitions of the measurement of the same value. We now make the hypothesis that certain experimental conditions have varied during the measurements, so that the variance associated with each result varies from one measurement to another. This leads to:

$$V(x_i) = \sigma_i^2 \quad [9.16]$$

However, the expectation of each result is the same; thus, we let  $\mu$  be the expectation of this variable. The  $n$  measurement results are expressed by the same system as before:

$$x_1 = \mu + e_1 \quad [9.17]$$

$$x_2 = \mu + e_2 \quad [9.18]$$

$$x_i = \mu + e_i \quad [9.19]$$

$$x_n = \mu + e_n \quad [9.20]$$

To take in consideration the fact that the measurements do not have the same variance, we weigh each square of the deviance with a weighting (or weight) coefficient  $g_i$ . We then look at the expression:

$$Q(\mu) = \sum_{i=1}^n g_i e_i^2 = \sum_{i=1}^n g_i (x_i - \mu)^2 \quad [9.21]$$

The minimum of this function is attained for the value  $\hat{\mu}_p$  of  $\mu$ , which is the solution of the equation obtained by writing that:

$$\frac{dQ}{d\mu} = 0 \quad [9.22]$$

which gives us:

$$\hat{\mu}_p = \frac{\sum_{i=1}^n g_i x_i}{\sum_{i=1}^n g_i} = \sum_{i=1}^n w_i x_i \tag{9.23}$$

where  $w_i$  is a normed weight such that  $\sum_{i=1}^n w_i = 1$ . [9.24]

The normalization condition imposed on  $w_i$  automatically means that  $\hat{\mu}_p$  is an unbiased estimator of  $\mu$ . We can also try to determine the weights that minimize the variance of  $\hat{\mu}_p$ . The variance of the weighted estimator is written:

$$V(\hat{\mu}_p) = \sum w_i^2 \sigma_i^2 \tag{9.25}$$

We can try to determine the form of the weights that bring about the smallest variance for  $\hat{\mu}_p$ . Looking for the minimum of this function, taking into account the constraint on the sum of the weights, gives us:

$$g_i = \frac{1}{\sigma_i^2} \tag{9.26}$$

and

$$w_i = \frac{1/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2} \tag{9.27}$$

In other words, the weight obtained by imposing the minimum variance condition of the estimator is inversely proportional to the result of the variance being considered. Under these conditions, the expression of the variance of  $\hat{\mu}_p$  is written as:

$$V(\hat{\mu}_p) = \frac{1}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} \tag{9.28}$$

### 9.3. Determining regression line coefficients

We measure  $n$  pairs of joined values  $x_i, y_i$  that, in a plan brought to an axial system  $(Ox, Oy)$ , are represented by  $n$  points. A model explaining the form of the scatter plot has a linear tendency that can be understood by two possible models:

- the proportional model  $Y = \theta.X$  that depends on one parameter  $\theta$ ;
- the affine model  $Y = \theta_0 + \theta_1.X$  that make use of two parameters.

With a sensor, the parameter factor of  $X$  is the sensitivity of the sensor, with the scale gap being a constant parameter that is either optional when the instrument has an expanded scale or is required in the case of an inopportune zero gap. We discuss the two situations by presenting the following hypotheses:

- H1, with values  $x_i$  that are perfectly known ( $E(x_i) = x_i$  and  $V(x_i) = 0$ );
- H2, with values of  $Y$  made without systematic errors;
- H3, with the variable  $Y$  measured with a constant variance, so  $V(y_i) = \sigma^2$  constant;
- H4, with independent measurements of  $Y$ , that is  $\text{cov}(y_i, y_j) = 0$  when  $i \neq j$ .

#### 9.3.1. A proportional relation

The system representing  $n$  measured points is written:

$$y_1 = \theta.x_1 + e_1 \quad [9.29]$$

$$y_2 = \theta.x_2 + e_2 \quad [9.30]$$

$$y_i = \theta.x_i + e_i \quad [9.31]$$

$$y_n = \theta.x_n + e_n \quad [9.32]$$

The terms  $e_1, e_2, \dots, e_n$  express the gaps between the observed values of  $Y$  and the values predicted by the chosen model. Hypothesis H2 immediately shows that each of these gaps is, on average, zero. So  $E(e_i) = 0$  whatever the index  $i$  of the measurement.

By applying the Gaussian criterion, we form the quantity:

$$Q(\theta) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (x_i - \theta.x_i)^2 \quad [9.33]$$

The graph representing the variation of  $Q(\theta)$  according to  $\theta$  is also a parabola. The minimum of this function is reached for the value  $\hat{\theta}$  of  $\theta$ , which is the solution of the equation obtained by writing:

$$\frac{dQ}{d\theta} = 0 \tag{9.34}$$

which gives the normal equation:

$$\sum_{i=1}^n \hat{\theta} \cdot x_i^2 = \sum_{i=1}^n x_i y_i \tag{9.35}$$

whose solution is:

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \tag{9.36}$$

This estimator is expressed according to random variables. It is therefore a random quantity. Building on the hypotheses previously formulated, this unbiased estimator, that is,  $E(\hat{\theta}) = \theta$ .

In addition, the variance calculation is considerably simplified by hypotheses H3 and H4. Its expression is given by:

$$V(\hat{\theta}) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2} \tag{9.37}$$

For each measured point of the abscissa  $x_i$ , we call the residuals  $r_i$  the gap existing between the measured value  $y_i$  and  $Y$  of the corresponding value  $\hat{y}_i = \theta \cdot x_i$  of the model:

$$r_i = y_i - \hat{y}_i = y_i - \hat{\theta} \cdot x_i \tag{9.38}$$

If we look at the normal equation, we see that  $\sum_{i=1}^n r_i \cdot x_i = 0$  [9.39]

This property allows us, *a posteriori*, to verify the validity of the numerical value of  $\theta$ .

As well, the residuals help us obtain an unbiased estimator  $\hat{\sigma}^2$  of the variance, with which the measurements of Y are carried out, providing the chosen model allows it:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-1} \quad [9.40]$$

Using the least squares line equation helps us calculate the interposed value  $\hat{y}$  of Y that corresponds to an undetermined abscissa of the field validity of the model:

$$\hat{y} = \hat{\theta}.x \quad [9.41]$$

The variance of this quantity is written:

$$V(\hat{y}) = x^2 V(\hat{\theta}) = \frac{\sigma^2 x^2}{\sum_{i=1}^n x_i^2} \quad [9.42]$$

It is possible to estimate this by replacing the variance  $\sigma^2$  with its estimator. We can see that the variance of  $\hat{y}$  is in origin zero (this is normal, the point being absolutely fixed by the chosen model), and that it increases as the square of  $x$  for all other abscissas.

### 9.3.2. Affine relations

We will look at the previous schema again by adapting it to the case of an affine model in order to explain the measurements. The system representing the measured n points is written:

$$y_1 = \theta_o + \theta_1 .x_1 + e_1 \quad [9.43]$$

$$y_2 = \theta_o + \theta_1 .x_2 + e_2 \quad [9.44]$$

$$y_i = \theta_o + \theta_1 \cdot x_i + e_i \tag{9.45}$$

$$y_n = \theta_o + \theta_1 \cdot x_n + e_n \tag{9.46}$$

The terms  $e_1, e_2, \dots, e_n$  express the deviations between the observed values of  $Y$  and the values predicted by the chosen model. Hypothesis H2 leads us immediately to the conclusion that each of the gaps are on average zero, so  $E(e_i) = 0$  whatever the index  $i$  of the measurement. By applying the Gaussian criterion, we get the quantity:

$$Q(\theta_0, \theta_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (x_i - \theta_0 - \theta_1 \cdot x_i)^2 \tag{9.47}$$

In a three-dimensional space brought to the axes  $0 \theta_0, 0 \theta_1$  and  $0 Q$ , the surface represented by the previous equation is an elliptic paraboloid that is intersected by a vertical plane. The intersection is a parabola, while the intersection by a horizontal plane (when it exists) is an ellipsis. As a general rule, the surface is not of a revolution, which means the axes of the ellipsis are not parallel to the axes of the coordinates.

Even though no relation exists between  $\theta_0$  and  $\theta_1$  (that is, they can vary independently from one another), the minimum value of this surface is reached for the value  $\hat{\theta}_0$  and  $\theta_0$  and the value  $\hat{\theta}_1$  of  $\theta_1$  that are the solutions of the usual equation systems obtained when we write:

$$\frac{\partial Q}{\partial \theta_0} = 0 \tag{9.48}$$

$$\frac{\partial Q}{\partial \theta_1} = 0 \tag{9.49}$$

which gives us the normal equation systems:

$$n\hat{\theta}_0 + \hat{\theta}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \tag{9.50}$$

$$\hat{\theta}_0 \sum_{i=1}^n x_i + \hat{\theta}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \tag{9.51}$$

whose solution is written:

$$\hat{\theta}_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i \cdot y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \sum_{i=1}^n y_i \left[ \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{9.52}$$

$$\hat{\theta}_1 = \frac{n \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n y_i \cdot (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{9.53}$$

For each of the solutions, the first form, directly deduced from normal equations, is the fastest for carrying out calculations. This is because the different sums that are part of it are calculated as input data arrives. However, the second form, which requires *a priori* calculation of the average of the values of X and Y, interposes the differences into these averages and is less sensitive to rounding errors of the calculation systems.

These estimators are expressed according to the random variables. These are therefore random quantities. Taking into account hypothesis H2, these are unbiased estimators, which means that  $E(\hat{\theta}_0) = \theta_0$  and that  $E(\hat{\theta}_1) = \theta_1$ .

Despite hypotheses H3 and H4, calculating the variance of each of these estimators requires several precautions regarding the basic relations to be used. In particular, the second form given for each estimator simplifies the calculations. We get the following expressions:

$$V(\hat{\theta}_0) = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \tag{9.54}$$

$$V(\hat{\theta}_1) = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{9.55}$$

Also, we see that the obtained estimators are usually correlated, so their covariance expression is as follows:

$$\text{cov}(\hat{\theta}_0, \hat{\theta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{9.56}$$

Later we will give an explanation for this correlation and why the covariance is zero when the arithmetic mean of the values of  $X$  are zero.

For each measured point, we call residuals  $r_i$ , the gap existing between the value measured  $y_i$  of  $Y$  and the corresponding value  $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 \cdot x_i$  of the model:

$$r_i = y_i - \hat{y}_i = y_i - \hat{\theta}_0 - \hat{\theta}_1 \cdot x_i \tag{9.57}$$

The first of the normal equations shows that:

$$\sum_{i=1}^n r_i = 0 \tag{9.58}$$

This property shows that the residuals are positive for some, negative for others, and that overall, the sum cancels itself out: the line of the least squares goes from “the middle” of the scatter of measured points and can be either above or below the line.

The second of the normal equations give us:

$$\sum_{i=1}^n r_i \cdot x_i = 0 \tag{9.59}$$

*A posteriori*, these properties make it possible to verify the validity of the numerical value of the estimators.

The residuals also help to obtain an unbiased estimator  $\hat{\sigma}^2$  of the variance. The residuals make up the measurements of  $Y$ , on condition that the chosen model is pertinent:



$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-2} \tag{9.60}$$

We see that the denominator, corresponding to the number of degrees of flexibilities associated with this estimator, is formed by the number of measured points, and diminished by the number of estimated parameters (here two, one for the gradient, the other for the ordinate of origin).

The first of the normal equations shows that the point whose coordinates are:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n} \tag{9.61}$$

belongs to the least squares line. This point is linked to the measurements made by the system and not to the adjusted line.

In other terms, the relation:

$$\bar{y} = \hat{\theta}_0 + \hat{\theta}_1 \cdot \bar{x} \tag{9.62}$$

is always satisfied.

This has two consequences. The first is that we can now calculate one of the estimators (in practice,  $\hat{\theta}_0$ ) from knowing the other one (in practice,  $\hat{\theta}_1$ ), and the values of  $\bar{x}$  and  $\bar{y}$ . The second is that it qualitatively explains the correlation between  $\hat{\theta}_0$  and  $\hat{\theta}_1$ . Since the line must go through the point of the coordinates  $\bar{x}$ ,  $\bar{y}$  being fixed for a data set, any attempt to modify the gradient, for example, can only be done by rotation around this point.

Let us suppose that the average of X is strictly positive. Under these conditions, the covariance between  $\hat{\theta}_0$  and  $\hat{\theta}_1$  is negative. We then see graphically that augmenting the gradient means a diminishment of the ordinate source. This explains the covariance sign. A completely similar conclusion can be obtained when the average of X is negative. In the specific case when the average of X is zero, this particular point is on the ordinates axis. The rotation of the line does not mean source ordinate modification, and the covariance is zero.

Using the least squares line equation allows us to calculate the interposed value  $\hat{y}$  of Y corresponding to an indeterminate abscissa  $x$  of the validity domain of the model:

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 \cdot x \tag{9.63}$$

The variance of this quantity is written as:

$$V(\hat{y}) = V(\hat{\theta}_0) + x^2 \cdot V(\hat{\theta}_1) + 2x \operatorname{cov}(\hat{\theta}_0, \hat{\theta}_1) \tag{9.64}$$

By replacing the variances of the estimators and the covariance between the estimators with their respective expressions, we get the forms:

$$V(\hat{y}) = \sigma^2 \left[ \frac{1}{n} + \frac{n(x - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \right] = \sigma^2 \left[ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \tag{9.65}$$

We see that the variance of  $\hat{y}$  varies according to the value of  $x$  (see Figure 9.3). It presents a minimum of  $x = \bar{x}$ , where its value is:

$$[V(\hat{y})]_{x=\bar{x}} = \frac{\sigma^2}{n} \tag{9.66}$$

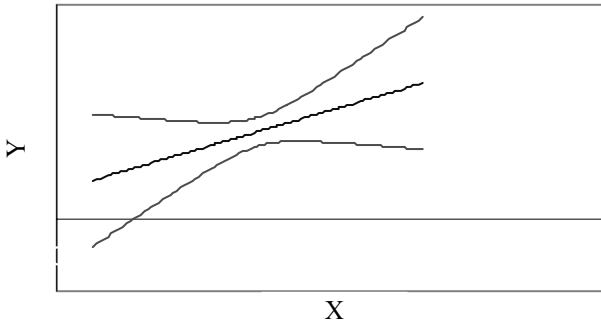
This is a logical result, since  $\hat{y} = \bar{y}$

We also see that for  $x = 0$ , we also find:

$$[V(\hat{y})]_{x=0} = V(\hat{\theta}_0) \tag{9.67}$$

The estimated ordinate variance of the line of least squares increases according to the lengthening function at  $\bar{x}$ .

All these expressions can be evaluated by replacing the variance  $\sigma^2$  with its estimator, which has been obtained from the residual sum of squares.



**Figure 9.3.** *Approximate gap envelope around the least squares line*

We can also reverse the least squares line method by calculating the abscissa  $\hat{x}$  that corresponds to an ordinate  $\hat{y}$ :

$$\hat{x} = \frac{\hat{y} - \hat{\theta}_0}{\hat{\theta}_1} \tag{9.68}$$

Applying the variance composition law to this expression leads to the following conclusion:

$$V(\hat{x}) = \frac{V(\hat{y})}{\hat{\theta}_1^2} \tag{9.69}$$

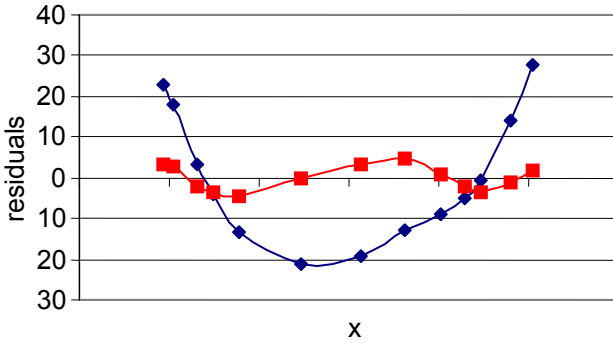
There is a direct correspondence between the gap type of the abscissa and the gap type of the ordinate throughout the gradient of the least squares line.

All these expressions can be estimated by replacing the variance  $\sigma^2$  by its estimator. This is shown in equation [9.60].

We should remember that the least squares criterion does not, when used alone, allow us to test the validity of the chosen linear model.

An examination of the graph representing the distribution of residuals according to the values of X allows us to make a zoom around the line. This shows that one or several points are abnormally far from the model, a concavity or inflection that can

be explained by a too low polynomial degree, or in the opposite situation, by a satisfactory distribution (random) of the points (Figure 9.4).



**Figure 9.4.** Distribution of residuals according to the model being used

We call the coefficient signification  $R$  the square of the correlation coefficient between the values of  $X$  and the values of  $Y$ , so:

$$R = \frac{\left[ \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y}) \right]^2}{\left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right] \left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right]} \tag{9.70}$$

This coefficient can also be expressed in one of the following forms:

$$R = \frac{\hat{\theta}_1 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n r_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{9.71}$$

We can easily see that if the experimental points are perfectly aligned, that is, if the dispersion of the values of  $Y$  are completely explained by the chosen theoretical model, the residuals are zero, so  $R = 1$ . However, if the values of  $Y$  are independent of the values of  $X$ , that is, if the gradient of the model is zero, then  $R = 0$ . Aside from these two extreme cases (which are rarely found in practice), we must be

careful to draw conclusions solely from the value of  $R$ , since different forms of point scatters can lead to the same value of the correlation coefficient.

### 9.3.3. Weighting application

#### 9.3.3.1. Calculation hypotheses

We will look again at adjustment problems with linear models by modifying hypothesis H3 only. Now we are in a situation where the measurement uncertainty of the variable  $Y$  can be different from one value to another, so that:

$$V(y_i) = \sigma_i^2 \quad [9.72]$$

This is the case when measurements are made with a type of constant relative gap  $\sigma_i/y_i$ . In these conditions, we use a weighting coefficient that converts the more or less high proximity of the passage of the line near to the point according to the uncertainty function that is being affected.

#### 9.3.3.2. Weighting and proportional relations

The quantity we want to minimize is the sum of the squared gaps, each gap having a weight  $g_i$ :

$$Q(\theta) = \sum_{i=1}^n g_i e_i^2 = \sum_{i=1}^n g_i (y_i - \theta \cdot x_i)^2 \quad [9.73]$$

The minimum of this function is reached for the value  $\hat{\theta}_p$  of  $\theta$ , which is the solution of the equation obtained by writing that:

$$\frac{dQ}{d\theta} = 0 \quad [9.74]$$

which gives the normal equation:

$$\sum_{i=1}^n \hat{\theta}_p g_i x_i^2 = \sum_{i=1}^n g_i x_i y_i \quad [9.75]$$

whose solution is:

$$\hat{\theta}_p = \frac{\sum_{i=1}^n g_i x_i y_i}{\sum_{i=1}^n g_i x_i^2} \tag{9.76}$$

Taking into account the hypotheses that have already been formulated, the estimator is always unbiased, which means that  $E(\hat{\theta}_p) = \theta$ .

In addition, the variance of  $\hat{\theta}_p$  is given by the general expression:

$$V(\hat{\theta}_p) = \frac{\sum_{i=1}^n g_i^2 x_i^2 \sigma_i^2}{\left[ \sum_{i=1}^n g_i x_i^2 \right]^2} \tag{9.77}$$

This is the function of values that we can give to the weighting coefficients.

We can find the weighting values that allow us to obtain a minimal value for  $V(\hat{\theta}_p)$ , that is, for the solutions obtained by writing:

$$\frac{\partial V(\hat{\theta}_p)}{\partial p_j} = 0 \tag{9.78}$$

for all the values of  $j$  between 1 and  $n$ .

We come to the condition:

$$g_1 \sigma_1^2 = g_2 \sigma_2^2 = \dots = g_j \sigma_j^2 = \dots = g_n \sigma_n^2 \tag{9.79}$$

This means that we find the fact that the weight that minimizes the variance is inversely proportional to the variance of the quantity it weights:

$$g_i = \frac{1}{\sigma_i^2} \tag{9.80}$$

In these conditions, the gradient estimator and its variance respectively take the following expressions:

$$\hat{\theta}_p = \frac{\sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2}} \quad [9.81]$$

$$V(\hat{\theta}_p) = \frac{1}{\sum_{i=1}^n \frac{x_i^2}{\sigma_i^2}} \quad [9.82]$$

In the specific case of a variance being identical for each measurement, these expressions result in the same relations as those shown for the non-weighted case.

For the measurement point of an abscissa  $x_i$ , the estimated ordinate is  $\hat{y}_i = \hat{\theta}.x_i$ , and the residual is worth:

$$r_i = y_i - \hat{y}_i = y_i - \hat{\theta}.x_i \quad [9.83]$$

The normal equation shows that:

$$\sum_{i=1}^n g_i r_i .x_i = 0 \quad [9.84]$$

*A posteriori*, this property allows us to verify the validity of the numerical value of  $\hat{\theta}_p$ .

Since the idea of weighting requires knowledge of weights, and therefore of variances linked to each result, estimating a variance from the sum of squared residuals is not relevant to this discussion.

### 9.3.3.3. Weighting and affine relations

Here we again look at the previous schema adapted to a situation of choosing an affine model to explain measurements. The quantity to be minimized is the sum of the squared gaps weighted by the weight  $g_i$ , so

$$Q(\theta_0, \theta_1) = \sum_{i=1}^n g_i e_i^2 = \sum_{i=1}^n g_i (y_i - \theta_0 - \theta_1 .x_i)^2 \quad [9.85]$$

The minimum of this function is attained for the value  $\hat{\theta}_{0p}$  of  $\theta_0$  and the value  $\hat{\theta}_{1p}$  of  $\theta_1$ , solutions of the equation obtained by writing:

$$\frac{\partial Q}{\partial \theta_0} = 0 \tag{9.86}$$

$$\frac{\partial Q}{\partial \theta_1} = 0 \tag{9.87}$$

which gives us the normal equation system:

$$\hat{\theta}_{0p} \sum_{i=1}^n g_i + \hat{\theta}_{1p} \sum_{i=1}^n g_i x_i = \sum_{i=1}^n g_i y_i \tag{9.88}$$

$$\hat{\theta}_{0p} \sum_{i=1}^n g_i x_i + \hat{\theta}_{1p} \sum_{i=1}^n g_i x_i^2 = \sum_{i=1}^n g_i x_i y_i \tag{9.89}$$

for which the solution is written:

$$\hat{\theta}_{0p} = \frac{\sum_{i=1}^n g_i x_i^2 \sum_{i=1}^n g_i y_i - \sum_{i=1}^n g_i x_i \sum_{i=1}^n g_i x_i y_i}{\sum_{i=1}^n g_i \sum_{i=1}^n g_i x_i^2 - \left( \sum_{i=1}^n g_i x_i \right)^2} \tag{9.90}$$

$$\hat{\theta}_{1p} = \frac{\sum_{i=1}^n g_i \sum_{i=1}^n g_i x_i y_i - \sum_{i=1}^n g_i x_i \sum_{i=1}^n g_i y_i}{\sum_{i=1}^n g_i \sum_{i=1}^n g_i x_i^2 - \left( \sum_{i=1}^n g_i x_i \right)^2} \tag{9.91}$$

Taking into account the previously formulated hypotheses, these estimators are unbiased, that is:

$$E(\hat{\theta}_{0p}) = \theta_0 \tag{9.92}$$

$$E(\hat{\theta}_{1p}) = \theta_1 \tag{9.93}$$



Whatever the weighting coefficients being used, this results in the “autonormalization” of these.

By dividing the two members of the first normal equation by  $\sum_{i=1}^n g_i$ , we get:

$$\bar{y}_p = \hat{\theta}_{0p} + \hat{\theta}_{1p} \cdot \bar{x}_p \tag{9.94}$$

by positing:

$$\bar{x}_p = \frac{\sum_{i=1}^n g_i x_i}{\sum_{i=1}^n g_i} \quad \text{and} \quad \bar{y}_p = \frac{\sum_{i=1}^n g_i y_i}{\sum_{i=1}^n g_i} \tag{9.95}$$

which are respectively the weighted average of the values of X and Y. The point of the coordinates  $\bar{x}_p, \bar{y}_p$  belongs to the least squares line. Consequently, it is possible to obtain a new set of relations that give estimators by using a new axial system, parallel to the initial axes but centered on the point of coordinates  $\bar{x}_p, \bar{y}_p$ . Here, the line intersects the origin and we find a proportional form, with the same leading coefficient, so:

$$\hat{\theta}_{1p} = \frac{\sum_{i=1}^n g_i (x_i - \bar{x}_p) \cdot (y_i - \bar{y}_p)}{\sum_{i=1}^n g_i (x_i - \bar{x}_p)^2} = \frac{\sum_{i=1}^n g_i y_i \cdot (x_i - \bar{x}_p)}{\sum_{i=1}^n g_i (x_i - \bar{x}_p)^2} \tag{9.96}$$

The value of  $\hat{\theta}_{0p}$  is expressed by:

$$\hat{\theta}_{0p} = \bar{y}_p - \hat{\theta}_{1p} \cdot \bar{x}_p \tag{9.97}$$

As was the case before, this second set of solutions, even though requiring a prior calculation of the weighted averages of X and Y, produce values that are less sensitive to calculation errors. This is because the expressions only relate to gap values measured in relation to these averages.

These estimators are expressed according to random variables. This means they are random quantities. Taking into consideration hypothesis H2, these are unbiased estimators, so that  $E(\hat{\theta}_0) = \theta_0$  and  $E(\hat{\theta}_1) = \theta_1$ .

The following expressions give the variances of each estimator, as well as their covariances.

$$V(\hat{\theta}_{0p}) = \frac{\sum_{i=1}^n g_i x_i^2}{\sum_{i=1}^n g_i \sum_{i=1}^n g_i x_i^2 - \left(\sum_{i=1}^n g_i x_i\right)^2} = \frac{1}{\sum_{i=1}^n g_i} + \frac{\bar{x}_p^2}{\sum_{i=1}^n g_i (x_i - \bar{x}_p)^2} \quad [9.98]$$

$$V(\hat{\theta}_1) = \frac{\sum_{i=1}^n g_i}{\sum_{i=1}^n g_i \sum_{i=1}^n g_i x_i^2 - \left(\sum_{i=1}^n g_i x_i\right)^2} = \frac{1}{\sum_{i=1}^n g_i (x_i - \bar{x}_p)^2} \quad [9.99]$$

and

$$\text{cov}(\hat{\theta}_0, \hat{\theta}_1) = \frac{-\sum_{i=1}^n g_i x_i}{\sum_{i=1}^n g_i \sum_{i=1}^n g_i x_i^2 - \left(\sum_{i=1}^n g_i x_i\right)^2} = \frac{-\bar{x}_p}{\sum_{i=1}^n g_i (x_i - \bar{x}_p)^2} \quad [9.100]$$

The variance values of the obtained estimators are functions of the values attributed to weighting coefficients.

As before, we can show that the weights that minimize these quantities are inversely proportional to the gap variance, so:

$$g_i = \frac{1}{\sigma_i^2} \quad [9.101]$$

Using the least squares line equation allows us to calculate the interposed value  $\hat{y}$  of Y corresponding to an indeterminate abscissa  $x$  of the validity field of the model:

$$\hat{y} = \hat{\theta}_{0p} + \hat{\theta}_{1p} \cdot x \quad [9.102]$$

The variance of this quantity is written:

$$V(\hat{y}) = V(\hat{\theta}_{0p}) + x^2 \cdot V(\hat{\theta}_{1p}) + 2x \text{cov}(\hat{\theta}_{0p}, \hat{\theta}_{1p}) \quad [9.103]$$

By replacing the variance of the estimators and the covariance between the estimators with their respective expressions, we get the form:

$$v(\hat{y}) = \frac{1}{\sum_{i=1}^n g_i} + \frac{(x - \bar{x}_p)^2}{\sum_{i=1}^n g_i (x_i - \bar{x}_p)^2} \tag{9.104}$$

**9.3.4. The least measured-squares line: when two measured variables contain uncertainties**

The previous sections do not discuss the problem of determining line coefficients when there are uncertainties only for variables represented as ordinates.

Practically, the quantities represented as X and Y are both likely to include uncertainties. We can refer back to the previous examples if we wish to show that the uncertainty of the variable represented on the axis of the abscissas, projected as an uncertainty on the axis of the ordinates, is small compared to the uncertainty of Y itself; that is:

$$V(y_i) \gg \theta_1^2 V(x_i) \tag{9.105}$$

If this inequality is not resolved, we can resolve the problem by using the method developed by Williamson.

This is part of the framework of the general hypotheses that follow.

The measured variables X and Y are connected by a formal relation:

$$y = \theta_0 + \theta_1 x \tag{9.106}$$

We measure n pairs of values  $(x_i, y_i)$ , each of these measurements being seen as a random variable, with the following variances:

$$V(x_i) = p_i \tag{9.107}$$

$$V(y_i) = q_i \tag{9.108}$$

It is not possible to bring in covariance *a priori*, since the variables X and Y are results of different experimental processes.

The quantity to be minimized is the sum of the squared weighted distances between each experimental point  $M_i$  of the coordinates  $(x_i, y_i)$  and the corresponding point  $M'_i$  of coordinates  $(X_i, Y_i)$  belonging to the theoretic equation line  $Y = \theta_0 + \theta_1 X$ , so:

$$Q = \sum_{i=1}^n \left[ \left( \frac{x_i - X_i}{p_i} \right)^2 + \left( \frac{y_i - Y_i}{q_i} \right)^2 \right] \tag{9.109}$$

or again:

$$Q = \sum_{i=1}^n \left[ \left( \frac{x_i - X_i}{p_i} \right)^2 + \left( \frac{y_i - \theta_0 - \theta_1 X_i}{q_i} \right)^2 \right] \tag{9.110}$$

which is also written:

$$Q = \sum_{i=1}^n \left[ \left( \frac{x_i - X_i}{p_i} \right)^2 + \left( \frac{v_i}{q_i} \right)^2 \right] \tag{9.111}$$

by proposing:

$$v_i = y_i - \theta_0 - \theta_1 x_i \tag{9.112}$$

We again mark the variances of  $x_i$  and  $y_i$  as the denominators of this expression, which weighted each square of the gap, as we have previously seen.

During a preliminary phase, it is necessary to determine the values  $X_i$  and  $Y_i$  of the point of the line, then make the adjustments in relation to these.

Williamson handles the problem by minimizing each quantity written inside the bracket by proposing that:

$$\frac{\partial Q}{\partial X_i} = 0 \tag{9.113}$$

We get:

$$Q = \sum_{i=1}^n g_i \cdot v_i^2 \quad [9.114]$$

by proposing:

$$g_i = \frac{1}{q_i + \theta_1^2 p_i} \quad [9.115]$$

We see that the proposed method means considering the squared gap, measured parallel to the axis of the ordinates, between the experimental point and the theoretical line. This quantity is modified by a weighted coefficient projected onto the axis of the ordinates of the variance of  $y_i$ , and then by the component projected onto the axis of the ordinates of the variance of  $x_i$ . This means that we come back to the standard situation analyzed above: the calculation of  $v_i$  as well as of  $g_i$  requires knowledge of the theoretical line that we are trying to measure. We thus immediately know that resolving the problem requires going through an iterative process or phase.

The first phase, in which we establish the estimators of  $\theta_0$  and of  $\theta_1$ , is done in the standard way by finding the solution of the equations system obtained by writing that:

$$\frac{\partial Q}{\partial \theta_0} = 2 \sum_{i=1}^n g_i v_i \frac{\partial v_i}{\partial \theta_0} = 0 \quad [9.116]$$

$$\frac{\partial Q}{\partial \theta_1} = \sum_{i=1}^n \left( v_i^2 \frac{\partial g_i}{\partial \theta_1} + 2g_i v_i \frac{\partial v_i}{\partial \theta_1} \right) = 0 \quad [9.117]$$

Taking into account the relations existing between  $v_i$  and  $g_i$  on the one hand, and between  $\theta_0$  and  $\theta_1$  on the other, we get:

$$\frac{\partial v_i}{\partial \theta_0} = -1 \quad [9.118]$$

$$\frac{\partial v_i}{\partial \theta_1} = -x_i \quad [9.119]$$

$$\frac{\partial g_i}{\partial \theta_1} = -2p_i \theta_1 g_i^2 \tag{9.120}$$

We also get the system of normal equations:

$$\sum_{i=1}^n g_i v_i = 0 \tag{9.121}$$

$$\sum_{i=1}^n g_i v_i (g_i v_i p_i \hat{\theta}_1 + x_i) = 0 \tag{9.122}$$

By replacing  $g_i$  and  $v_i$  with their expressions in the first of the normal equations, and by proposing:

$$\bar{x}_\pi = \frac{\sum_{i=1}^n g_i x_i}{\sum_{i=1}^n g_i} \quad \text{and} \quad \bar{y}_\pi = \frac{\sum_{i=1}^n g_i y_i}{\sum_{i=1}^n g_i} \tag{9.123}$$

we end up with the formula:

$$\hat{\theta}_0 + \hat{\theta}_1 \cdot \bar{x}_\pi = \bar{y}_\pi \tag{9.124}$$

This easily shows that the estimated line intersects with the point of the coordinates  $\bar{x}_\pi, \bar{y}_\pi$ , a result we have already seen with more restrictive hypotheses.

As for the second normal equation, with the following:

$$x'_i = x_i - \bar{x}_\pi \tag{9.125}$$

$$y'_i = y_i - \bar{y}_\pi \tag{9.126}$$

and:

$$z_i = x'_i g_i q_i - g_i p_i \hat{\theta}_1 y'_i \tag{9.127}$$

This can be expressed as:

$$\sum_{i=1}^n g_i z_i (y'_i - \hat{\theta}_1 \cdot x'_i) = 0 \tag{9.128}$$

From this we immediately get the gradient estimator:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n g_i z_i y'_i}{\sum_{i=1}^n g_i z_i x'_i} \tag{9.129}$$

This format is less satisfactory than it seems, since  $g_i$  and also  $z_i$  are functions of  $\hat{\theta}_1$  and therefore of the solution!

Practically, we calculate by iteration from an initial value of  $\theta_1$  written as  ${}_0\hat{\theta}_1$  (obtained from a graphic estimation or by means of a brief preliminary calculation). From this value we calculate the following quantities:

$${}_0g_i = \frac{1}{q_i + {}_0\hat{\theta}_1^2 \cdot p_i} \tag{9.130}$$

$${}_0\bar{x}_\pi = \frac{\sum_{i=1}^n {}_0g_i \cdot x_i}{\sum_{i=1}^n {}_0g_i} \tag{9.131}$$

$${}_0\bar{y}_\pi = \frac{\sum_{i=1}^n {}_0g_i \cdot y_i}{\sum_{i=1}^n {}_0g_i} \tag{9.132}$$

$${}_0x'_i = x_i - {}_0\bar{x}_\pi \tag{9.133}$$

$${}_0y'_i = y_i - {}_0\bar{y}_\pi \tag{9.134}$$

and:

$$z_i = x'_i \cdot 0 \cdot g_i \cdot q_i - 0 \cdot g_i \cdot p_i - 0 \cdot \hat{\theta}_1 \cdot 0 \cdot y'_i \tag{9.135}$$

We get a new value of the gradient estimator:

$${}_1 \hat{\theta}_1 = \frac{\sum_{i=1}^n 0 \cdot g_i \cdot z_i \cdot y'_i}{\sum_{i=1}^n 0 \cdot g_i \cdot z_i \cdot x'_i} \tag{9.136}$$

which helps us calculate new values for  $g_i, \bar{x}_\pi, \bar{y}_\pi, x'_i, y'_i, z_i$  and thus for  $\hat{\theta}_1$ , etc., continuing up to the convergence towards the gradient value.

As with all iterative problems, it is important to start from a value close to the solution, both to minimize the number of iterations and to guarantee the convergence towards the desired value, even though here this last point is not an issue.

Once we have  $\hat{\theta}_1$ , we get  $\hat{\theta}_0$  from the relation:

$$\hat{\theta}_0 = \bar{y}_\pi - \hat{\theta}_1 \cdot \bar{x}_\pi \tag{9.137}$$

This technique also helps us obtain the variances of the estimators:

$$V(\hat{\theta}_1) = T^2 \sum_{i=1}^n g_i \cdot (x_i'^2 q_i + y_i'^2 p_i) \tag{9.138}$$

$$V(\hat{\theta}_0) = \frac{1}{\sum_{i=1}^n g_i} + 2 \cdot (\bar{x}_\pi + 2\bar{z}_\pi) \bar{z}_\pi T + (\bar{x}_\pi + 2\bar{z}_\pi)^2 V(\hat{\theta}_1) \tag{9.139}$$

expressions in which

$$T = \frac{1}{\sum_{i=1}^n g_i \cdot \left[ \frac{x'_i \cdot y'_i}{\hat{\theta}_1} + 4 \cdot (z_i - \bar{z}_\pi) \cdot (z_i - x'_i) \right]} \tag{9.140}$$



and:

$$\bar{z}_\pi = \frac{\sum_{i=1}^n g_i \cdot z_i}{\sum_{i=1}^n g_i} \tag{9.141}$$

All these relations restore values already seen when uncertainties only affect the variable Y, whether in a constant way as a variable according to the measurement index.

### 9.4. Example of a polynomial relation

Now we come to a situation where the adjustment model is written in the form of a polynomial development of X. We get:

$$Y = f(X) = \theta_0 \cdot f_0(X) + \theta_1 \cdot f_1(X) + \theta_2 \cdot f_2(X) + \theta_k \cdot f_k(X) \tag{9.142}$$

$f_0(X), f_1(X), f_2(X), \dots, f_k(X)$  being polynomials of X, of 0, 1, 2, ..., k respectively.

We try to find the estimators of the parameters  $\theta_0, \theta_1, \theta_2, \dots, \theta_k$  of the model from the measurement of n pairs of values  $(x_i, y_i)$  when  $n > k + 1$ .

The ordinate form:

$$Y = f(X) = \theta_0 + \theta_1 \cdot X + \theta_2 \cdot X^2 + \theta_k \cdot X^k \tag{9.143}$$

is a specific example of this model.

#### 9.4.1. A simple example

This situation will be discussed in the following hypotheses, already formulated for the line:

- H1: the values are perfectly known ( $E(x_i) = x_i$  and  $V(x_i) = 0$ );
- H2: the measurements of Y are obtained without systematic errors;
- H3: the variable Y has the same variance whatever its value, so  $V(y_i) = \sigma^2$  constant;
- H4: the measurements of Y are independent, that is,  $cov(y_i, y_j) = 0$  when  $i \neq j$ .

The measurement ensemble is expressed by the system of n equations:

$$\theta_0 \cdot f_0(x_1) + \theta_1 \cdot f_1(x_1) + \theta_2 \cdot f_2(x_1) + \theta_k \cdot f_k(x_1) + e_1 = y_1 \quad [9.144]$$

$$\theta_0 \cdot f_0(x_2) + \theta_1 \cdot f_1(x_2) + \theta_2 \cdot f_2(x_2) + \theta_k \cdot f_k(x_2) + e_2 = y_2 \quad [9.145]$$

$$\theta_0 \cdot f_0(x_n) + \theta_1 \cdot f_1(x_n) + \theta_2 \cdot f_2(x_n) + \theta_k \cdot f_k(x_n) + e_n = y_n \quad [9.146]$$

We can obtain a more compact expression by using the matrix formalism. By introducing the following vectors:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{bmatrix} \quad [9.147]$$

the above system is written:

$$A \cdot \theta + e = Y \quad [9.148]$$

The matrix:

$$A = \begin{bmatrix} f_0(x_1) & f_1(x_1) & \dots & f_k(x_1) \\ f_0(x_2) & f_1(x_2) & \dots & f_k(x_2) \\ \dots & \dots & \dots & \dots \\ f_0(x_n) & f_1(x_n) & \dots & f_k(x_n) \end{bmatrix} \quad [9.149]$$

is sometimes called the conditioning matrix of the system.

The sum of the squared gaps becomes:

$$Q = \sum_{i=1}^n e_i^2 = e^T \cdot e \quad [9.150]$$

where  $e^T$  represents the transposed vector  $e$ .

This sum is minimum when the derivations of Q in relation to each of the model parameters are simultaneously zero. This leads us to the system of normal equations:

$$\hat{\theta}_0 \sum_{i=1}^n f_0^2(x_i) + \hat{\theta}_1 \sum_{i=1}^n f_0(x_i)f_1(x_i) + \dots + \hat{\theta}_k \sum_{i=1}^n f_0(x_i)f_k(x_i) = \sum_{i=1}^n y_i f_0(x_i) \quad [9.151]$$

$$\hat{\theta}_0 \sum_{i=1}^n f_0(x_i)f_1(x_i) + \hat{\theta}_1 \sum_{i=1}^n f_1^2(x_i) + \dots + \hat{\theta}_k \sum_{i=1}^n f_1(x_i)f_k(x_i) = \sum_{i=1}^n y_i f_1(x_i) \quad [9.152]$$

$$\hat{\theta}_0 \sum_{i=1}^n f_0(x_i)f_k(x_i) + \hat{\theta}_1 \sum_{i=1}^n f_1(x_i)f_k(x_i) + \dots + \hat{\theta}_k \sum_{i=1}^n f_k^2(x_i) = \sum_{i=1}^n y_i f_k(x_i) \quad [9.153]$$

In this system,  $\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k$ , solutions of this system, are the estimators of the parameters  $\theta_0, \theta_1, \dots, \theta_k$  in the sense of the least squares.

By writing the vector of the estimator:

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \dots \\ \hat{\theta}_n \end{bmatrix} \quad [9.154]$$

the system of normal equations takes the form:

$$A^T . A . \hat{\theta} = A^T . y \quad [9.155]$$

From this we get the solution to the problem:

$$\hat{\theta} = (A^T . A)^{-1} A^T . y \quad [9.156]$$

We see that the matrix:

$$A^T . A = \begin{bmatrix} \sum_{i=1}^n f_0^2(x_i) & \sum_{i=1}^n f_0(x_i)f_1(x_i) & \dots & \sum_{i=1}^n f_0(x_i)f_k(x_i) \\ \sum_{i=1}^n f_0(x_i)f_1(x_i) & \sum_{i=1}^n f_1^2(x_i) & \dots & \sum_{i=1}^n f_1(x_i)f_k(x_i) \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^n f_0(x_i)f_k(x_i) & \sum_{i=1}^n f_1(x_i)f_k(x_i) & \dots & \sum_{i=1}^n f_k^2(x_i) \end{bmatrix} \quad [9.157]$$

is a squared symmetrical matrix with lines and columns that are equal to the number of estimated parameters.

We can calculate the matrix given the variances and the covariances of the obtained estimators.

We get the following general result:

$$\begin{aligned}
 V(\hat{\theta}) &= \begin{bmatrix} V(\hat{\theta}_0) & \text{cov}(\hat{\theta}_0, \hat{\theta}_1) & \dots & \text{cov}(\hat{\theta}_0, \hat{\theta}_k) \\ \text{cov}(\hat{\theta}_0, \hat{\theta}_1) & V(\hat{\theta}_1) & \dots & \text{cov}(\hat{\theta}_1, \hat{\theta}_k) \\ \dots & \dots & \dots & \dots \\ \text{cov}(\hat{\theta}_0, \hat{\theta}_k) & \text{cov}(\hat{\theta}_1, \hat{\theta}_k) & \dots & V(\hat{\theta}_k) \end{bmatrix} \\
 &= (A^T .A)^{-1} A^T .V(y) .A (A^T .A)^{-1} \tag{9.158}
 \end{aligned}$$

where  $V(y)$  is the variances-covariances matrix of  $y$ .

With hypotheses H3 and H4, this takes the form of:

$$V(y) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{9.159}$$

and the variances-covariances matrix of the estimators takes the simpler form:

$$V(\hat{\theta}) = \sigma^2 (A^T .A)^{-1} \tag{9.160}$$

Near to the factor  $\sigma^2$ , this is simply the inverse matrix of  $A^T \cdot A$ . Even though it is symmetrical, this matrix generally is expressed in non-zero terms outside the diagonal principle. This becomes a general rule: the obtained estimators are correlated.

As needed, it is always possible to obtain an estimator of  $\sigma^2$  from calculating the residuals  $r$  by using the relation:

$$\sigma^2 = \frac{\sum_{i=1}^2 r_i^2}{n - p} \tag{9.161}$$

where  $p$  is the number of estimated parameters. When a polynomial form is complete,  $p = n + 1$ .

We can use the coefficients obtained to calculate the value  $\hat{y}$  of the polynomial corresponding to a given value of  $x$ , so:

$$\hat{y} = \hat{\theta}_0 \cdot f_0(X) + \hat{\theta}_1 \cdot f_1(X) + \hat{\theta}_2 \cdot f_2(X) + \hat{\theta}_k \cdot f_k(X) \tag{9.162}$$

If we introduce the vector:

$$\mathbf{x} = \begin{bmatrix} f_0(x) \\ f_1(x) \\ \dots \\ f_k(x) \end{bmatrix} \tag{9.163}$$

we also get:

$$\hat{y} = \mathbf{x}^T \cdot \hat{\theta} \tag{9.164}$$

from which we have the variance of  $\hat{y}$ :

$$V(\hat{y}) = \mathbf{x}^T \cdot V(\hat{\theta}) \cdot \mathbf{x} = \sigma^2 \cdot \mathbf{x}^T \cdot (\mathbf{A}^T \cdot \mathbf{A})^{-1} \cdot \mathbf{x} \tag{9.165}$$

This solution done with a least squares matrix contains the results already obtained in the example of the line.

### 9.4.2. An example using weighting

Here, we look once more at a situation in which the values of  $Y$  are not correlated but are obtained with a different variance for each measurement. Under these conditions, hypothesis H3 is written:

$$V(y_i) = \sigma_i^2 \tag{9.166}$$

so that the variance-covariance matrix of  $y$  takes the form:

$$V(y) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \tag{9.167}$$

The system describing the measurements always has the form:

$$A\theta + e = Y \tag{9.168}$$

However, we need to know the sum of the squared gaps weighted by a weight  $g_i = 1/\sigma_i^2$ , so:

$$Q = \sum_{i=1}^n g_i e_i^2 \tag{9.169}$$

Finding the minimum of this quantity leads us to the system of normal equations which, in matrix form, is written:

$$A^T . g . A \hat{\theta} = A^T . g . y \tag{9.170}$$

From this we get the solution of the problem:

$$\hat{\theta} = (A^T . g . A)^{-1} A^T . g . y \tag{9.171}$$

In these expressions, the weighting matrix  $g$  is the inverse of the variances-covariances matrix of  $y$ :

$$g = [V(y)]^{-1} \tag{9.172}$$

The matrix giving the variances and covariances of the obtained estimators is written:

$$V(\hat{\theta}) = (A^T . g . A)^{-1} \tag{9.173}$$

This matrix solution of the least squares contains the results already found in the example of the line.

### 9.4.3. Examples with correlated variables

The matrix notation of the least squares provides the simpler calculation solutions than the algebraic form.

For example, let's look at the following example: we are measuring pairs of values  $(x_i, y_i)$  whose representation in a system of axes  $Ox, Oy$  gives points that are fairly well-aligned. We then use the following representation model:

$$y = \theta_0 + \theta_1 .x \tag{9.174}$$

In relation to the examples analyzed above, we make the hypothesis that the measurements of  $y$  are made with the same variance and are correlated. This occurs fairly often in practice, if only because of the uncertainties introduced by the measurement instrument.

If we suppose that the covariance between the values of  $y$ , taken two by two, remain the same, the variance-covariance matrix of the vector  $y$  takes the form:

$$V(y) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \dots & \dots & \dots & \dots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \dots & \dots & \dots & \dots \\ \rho & \rho & \dots & 1 \end{bmatrix} \tag{9.175}$$

by writing that  $\rho$  is the correlation coefficient between the two measurements of  $Y$ .

Since the measurements have the same variance, it is not necessary to weight the results; and the solution of the least squares retains the usual form:

$$\hat{\theta} = (A^T .A)^{-1} A^T .y \tag{9.176}$$

However, the variance-covariance matrix of these estimators is obtained by using the complete form:

$$V(\hat{\theta}) = (A^T .A)^{-1} A^T .V(y).A.(A^T .A)^{-1} \tag{9.177}$$

which leads to the expressions:

$$V(\hat{\theta}_0) = \sigma^2 \left[ \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} (1 - \rho) + \rho \right] \tag{9.178}$$

$$V(\hat{\theta}_1) = \frac{n\sigma^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} (1 - \rho) \tag{9.179}$$

$$\text{cov}(\hat{\theta}_0, \hat{\theta}_1) = \frac{-\sigma^2 \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} (1 - \rho) \tag{9.180}$$

All these values obviously depend on the correlation coefficient. When this takes the value 0, we find the same results as obtained before.

When the correlation coefficient takes the value + 1, we get:

$$V(\hat{\theta}_0) = \sigma^2 \tag{9.181}$$

$$V(\hat{\theta}_1) = 0 \tag{9.182}$$

$$\text{cov}(\hat{\theta}_0, \hat{\theta}_1) = 0 \tag{9.183}$$

These paradoxical results can be explained as follows: when the correlation coefficient equals + 1, the experimental points are rigorously aligned. As well, the least squares line is parallel to the theoretical line. This means the gradient has no random features, from which we get a zero value for its variance (as well as for the covariance between the gradient and the ordinate of source). The only random quantity is the ordinate of origin, which is equal to the variance of the values of Y.

The ordinate  $\hat{y}$ , which is on an interposed point of the abscissa  $x$ , is expressed by:

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 .x \tag{9.184}$$



has a variance:

$$V(\hat{y}) = \sigma^2 \left\{ \rho + (1 - \rho) \left[ \frac{1}{n} + \frac{n(x - \bar{x})^2}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \right] \right\} \tag{9.185}$$

When  $\rho = 0$ , we find the normal value, but when  $\rho = + 1$ , we get:

$$V(\hat{y}) = \sigma^2 \tag{9.186}$$

This variance is constant and equal to the measurement variance of Y.

We should remember that, for a measurement series, all the correlation coefficients cannot be simultaneously negative.

### 9.5. A simple example

The least squares method also applies to situations when the model chosen to represent the dependence between the variables X and Y results in a non-linear normal equation system. Here, we can linearize the equation in several ways: by changing the variable; by developing the function serially near to the representative measurement points; or by working numerically and finding the function of the sum of the squared gaps.

#### 9.5.1. Linearizing the function

There are functions linking X and Y that lead to non-linear systems expressed in parameters to be estimated, but which, by changing a variable, can be relevant to this example. This means they can be analyzed by the standard methods. By way of example, we cite the following cases:

- $U = A \cdot \exp(BT)$  which refers to the linear example  $Y = \theta_0 + \theta_1 X$  by proposing  $Y = \ln(U)$  and  $X = T$ , from which we derive  $\theta_0 = \ln(A)$  and  $\theta_1 = B$ .

- $U = A + B \cdot \ln(T)$ , which refers to the linear example  $Y = \theta_0 + \theta_1 X$  by proposing that  $Y = U$  and  $X = \ln(T)$ , from which we derive  $\theta_0 = A$  and  $\theta_1 = B$ .

- $U = A \cdot T^B$ , which refers to the linear example  $Y = \theta_0 + \theta_1 X$  by proposing that  $Y = U$  and  $X = \ln(T)$ , from which  $\theta_0 = \ln(A)$  and  $\theta_1 = B$ .

–  $U = \frac{T}{A.X + B}$  which refers to the linear example  $Y = \theta_0 + \theta_1 X$  by proposing that  $Y = \frac{1}{U}$  and  $X = \frac{1}{X}$ , from which we derive  $\theta_0 = A$  and  $\theta_1 = B$ .

We should be aware of the fact that even if the values of the transformed variable  $U$  have the same variance  $V(U)$ , changing the variable usually involves unequal variances for the values of the resulting variable  $Y$ , since the variance  $Y$  is expressed as:

$$V(Y) = \left( \frac{\partial Y}{\partial U} \right)^2 V(U) \tag{9.187}$$

Calculating the coefficients of the model thus involves weighted forms.

When changing variables proves impossible, we can resolve the system in an approximate way by carrying out a limited development. So, for example:

$$Y = f(X, \theta_0, \theta_1, \dots, \theta_k) \tag{9.188}$$

The equation transforms the relations between the values of  $X$  and the measurements of  $Y$ . It depends on the values of  $k + 1$  parameters  $\theta_0, \theta_1, \dots, \theta_k$ .

We measure  $n$  pairs of values  $(x_i, y_i)$  so that  $V(y_i) = \sigma_i^2$ . Applying the principle of the least squares gives us the function:

$$Q = \sum_{i=1}^n g_i [y_i - f(x_i, \theta_0, \theta_1, \dots, \theta_k)]^2 \tag{9.189}$$

with the weighting coefficient  $g_i = \frac{1}{\sigma_i^2}$ . [9.190]

The estimators we are trying to find are the solution of the normal equations systems:

$$\frac{\partial Q}{\partial \theta_0} = \sum_{i=1}^n -2 \cdot g_i \frac{\partial f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_0} [y_i - f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k)] = 0 \tag{9.191}$$

$$\frac{\partial Q}{\partial \theta_1} = \sum_{i=1}^n -2 \cdot g_i \frac{\partial f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_1} [y_i - f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k)] = 0 \quad [9.192]$$

$$\frac{\partial Q}{\partial \theta_k} = \sum_{i=1}^n -2 \cdot g_i \frac{\partial f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k)}{\partial \hat{\theta}_k} [y_i - f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k)] = 0 \quad [9.193]$$

In general, this system is not linear, which makes its resolution difficult. We can resolve it from the initial values of the written solutions  ${}_0\theta_0, {}_0\theta_1, \dots, {}_0\theta_k$  and giving each of them an increase  $e_0, e_1, \dots, e_k$ , so that:

$$\hat{\theta}_0 = {}_0\theta_0 + e_0 \quad [9.194]$$

$$\hat{\theta}_1 = {}_0\theta_1 + e_1 \quad [9.195]$$

$$\hat{\theta}_k = {}_0\theta_k + e_k \quad [9.196]$$

We can then write:

$$f(x_i, \hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k) = f(x_i, {}_0\theta_0, {}_0\theta_1, \dots, {}_0\theta_k) \sum_{j=0}^k \frac{\partial f(x_i, {}_0\theta_0, {}_0\theta_1, \dots, {}_0\theta_k)}{\partial \theta_j} \quad [9.197]$$

The normal equations are then of the type (what we have written here is only that of the index parameter j):

$$\sum_{i=0}^n g_i \frac{\partial f(x_i, {}_0\theta_0, {}_0\theta_1, \dots, {}_0\theta_k)}{\partial \theta_j} \left[ y_i - f(x_i, {}_0\theta_0, {}_0\theta_1, \dots, {}_0\theta_k) - \sum_{j=0}^k \frac{\partial f(x_i, {}_0\theta_0, {}_0\theta_1, \dots, {}_0\theta_k)}{\partial \theta_j} e_j \right] = 0 \quad [9.198]$$

We find a linear system as  $e_j$ . From this solution, and from the initial values given to the parameters, we get a new set of values that can help iterate the calculation.

### 9.5.2. Numerical search for the minimum of the function of the sum of the squared gaps

Another approach is to state that the quantity:

$$Q = \sum_{i=1}^n g_i [y_i - f(x_i, \theta_0, \theta_1, \dots, \theta_k)]^2 \quad [9.199]$$

is a function of the parameters of the model. This is an equation of a surface in an axial system formed by the parameters and by  $Q$ . In a situation of an affine relation, we have an equation of an elliptic paraboloid, so-called because its intersection with a plane parallel to the axis  $OQ$  produces a parabola. Moreover, its intersection with a plane parallel to the parameter plane is an ellipsis (when this intersection exists).

As a general rule, the form of this surface can vary, but close to the solution (corresponding to a surface summit), we find an elliptic paraboloid.

The concept is as follows: from the initial values, we calculate a first value of  $Q$ . We then give an increase to each of the values of the parameters and observe the variation of  $Q$ . If  $Q$  increases, we move away from the surface summit, so that the increases are in the wrong direction. In the opposite situation, the sense of displacement is correct, and we continue until converging on the solution. To put it another way, the representative point of the parameter values is displaced on the surface until it joins its summit again.

Carrying out the method can be somewhat difficult when the number of parameters is high. There are methods that allow us to systematize finding the solution and the speed of the convergence towards this solution (the Maquard method, for example).

Certain precautions must be taken if we use graphic methods. When using these methods, the following conditions must be met:

- The departure point must be sufficiently close to the solution in order to have a quick maximum convergence.

- The variations given to the parameters must not be so significant that oscillation from one part to another of the solution occurs and no solution is achieved.

- Due to the local curvature of the surface, the rapidity of convergence may be different depending on whether the solution is reached by larger or smaller values.

– Some functions that lead to a surface with a representation that presents several minima. In this case, we must ensure that the found solution is well-researched.

### 9.6. Examples of multivariable models

The principle of the least squares also applies to finding the parameters of a multivariable model, that is, to describing the development of an explained variable  $Z$  according to the explicative variables.

This is an example of the least squares plane when the relation between  $Z$  and the variables  $X$  and  $Y$  is expressed as:

$$z = \theta_0 + \theta_1 x + \theta_2 y \tag{9.200}$$

The problem consists of finding the estimators  $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2$  of  $\theta_0, \theta_1, \theta_2$  from the measurement of  $n$  triplets of the paired values  $x_i, y_i, z_i$  which, in an  $n$  plane brought to an axial system  $(Ox, Oy, Oz)$ , is represented by  $n$  points.

The system of  $n$  equations transforming these measurements is written:

$$\theta_0 + \theta_1 x_1 + \theta_2 y_1 + e_1 = z_1 \tag{9.201}$$

$$\theta_0 + \theta_1 x_2 + \theta_2 y_2 + e_2 = z_2 \tag{9.202}$$

$$\theta_0 + \theta_1 x_n + \theta_2 y_n + e_n = z_n \tag{9.203}$$

or again, by proposing:

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \tag{9.204}$$

and:

$$A = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \dots & \dots & \dots \\ 1 & x_n & y_n \end{bmatrix} \tag{9.205}$$

The previous system is written:

$$A\theta + e = z \tag{9.206}$$

Looking again at the expressions in the examples of polynomial relations, we see that the solutions given there are still valid, whatever these are for the vector of the estimators or for the variance-covariance matrix.

Here we give as examples the expressions of the coefficients of the least squares plane:

$$\hat{\theta}_0 = \text{num}(\hat{\theta}_0) / \text{den} \tag{9.207}$$

$$\hat{\theta}_1 = \text{num}(\hat{\theta}_1) / \text{den} \tag{9.208}$$

$$\hat{\theta}_2 = \text{num}(\hat{\theta}_2) / \text{den} \tag{9.209}$$

with:

$$\text{den} = n \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2 + 2 \sum_{i=1}^n x_i \sum_{i=1}^n y_i \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i^2 \left( \sum_{i=1}^n y_i \right)^2 - \left( \sum_{i=1}^n x_i \right)^2 \sum_{i=1}^n y_i^2 - n \left( \sum_{i=1}^n x_i y_i \right)^2 \tag{9.210}$$

$$\begin{aligned} \text{num}(\hat{\theta}_0) &= \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2 \sum_{i=1}^n z_i + \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \sum_{i=1}^n y_i z_i + \sum_{i=1}^n y_i \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i z_i \\ &- \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i \sum_{i=1}^n y_i z_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i^2 \sum_{i=1}^n x_i z_i - \left( \sum_{i=1}^n x_i y_i \right)^2 \sum_{i=1}^n z_i \end{aligned} \tag{9.211}$$

$$\begin{aligned}
 num(\hat{\theta}_1) &= n \sum_{i=1}^n y_i^2 \sum_{i=1}^n x_i z_i + \sum_{i=1}^n y_i \sum_{i=1}^n z_i \sum_{i=1}^n x_i y_i + \sum_{i=1}^n x_i \sum_{i=1}^n y_i \sum_{i=1}^n y_i z_i \\
 &- n \sum_{i=1}^n x_i y_i \sum_{i=1}^n y_i z_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i^2 \sum_{i=1}^n z_i - \left( \sum_{i=1}^n y_i \right)^2 \sum_{i=1}^n x_i z_i \quad [9.212]
 \end{aligned}$$

$$\begin{aligned}
 num(\hat{\theta}_2) &= n \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i z_i + \sum_{i=1}^n x_i \sum_{i=1}^n y_i \sum_{i=1}^n x_i z_i + \sum_{i=1}^n x_i \sum_{i=1}^n z_i \sum_{i=1}^n x_i y_i \\
 &- n \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i z_i - \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i \sum_{i=1}^n z_i - \left( \sum_{i=1}^n x_i \right)^2 \sum_{i=1}^n y_i z_i \quad [9.213]
 \end{aligned}$$

The multivariable forms also help us in cases where sensors have sensitivities that vary according to a variable; that it is external to measurement (influence variables).

Let us look, for example, at the response  $Y$  of a sensor to a variable  $X$ . If the sensor is linear, its response can be modeled by the form:

$$y = \theta_0 + Sx \quad [9.214]$$

where  $S$  is the sensitivity of the sensor. If this sensitivity is itself a function of a parameter  $t$ , which we write:

$$S = \theta_1 + \theta_2 t \quad [9.215]$$

$\theta_1$  being the sensitivity of the sensor when  $t = 0$  and  $\theta_2$  describing the development of the sensitivity according to  $t$ . We then get:

$$y = \theta_0 + (\theta_1 + \theta_2 t).x = \theta_0 + \theta_1 .x + \theta_2 .x.t \quad [9.216]$$

We recognize the plane equation, the variable  $y$  being described according to the function of the variable  $x$ , and  $xt$ .

## 9.7. Dealing with constraints

### 9.7.1. Presentation of the method

The least squares method consists of finding the minimum of the parameter function of the model  $Q(\theta_0, \theta_1 \dots \theta_k)$  formed by writing the sum of the squared gaps. This minimum is reached when:

$$dQ = \frac{\partial Q}{\partial \theta_0} .d\theta_0 + \frac{\partial Q}{\partial \theta_1} .d\theta_1 + \dots + \frac{\partial Q}{\partial \theta_k} .d\theta_k = 0 \quad [9.217]$$

In the absence of any constraint (that is, of an exterior relation between  $\theta_0, \theta_1 \dots \theta_k$ ), the differential elements  $d\theta_0, d\theta_1, \dots, d\theta_k$  can be chosen arbitrarily and, for the above relation to be resolved, it must be enough that:

$$\frac{\partial Q}{\partial \theta_0} = 0 \quad [9.218]$$

$$\frac{\partial Q}{\partial \theta_1} = 0 \quad [9.219]$$

$$\frac{\partial Q}{\partial \theta_k} = 0 \quad [9.220]$$

This is the standard normal equation system.

Now we come to a situation with a constraint relation between the parameters  $\theta_0, \theta_1 \dots \theta_k$ , that is, they are connected in an equation that we can write as:

$$g(\theta_0, \theta_1, \dots, \theta_k) = C \quad [9.221]$$

where  $C$  is a constant. Under these conditions, it is clear that we can no longer arbitrarily choose the differential elements  $d\theta_0, d\theta_1, \dots, d\theta_k$  since they are linked by the equation:

$$dg = \frac{\partial g}{\partial \theta_0} .d\theta_0 + \frac{\partial g}{\partial \theta_1} .d\theta_1 + \dots + \frac{\partial g}{\partial \theta_k} .d\theta_k = 0 \quad [9.222]$$



The result is that we can express one of the elements (for example,  $d\theta_k$ ) according to the others:

$$d\theta_k = -\frac{1}{\frac{\partial g}{\partial \theta_k}} \left( \frac{\partial g}{\partial \theta_0} .d\theta_0 + \frac{\partial g}{\partial \theta_1} .d\theta_1 + \dots + \frac{\partial g}{\partial \theta_{k-1}} .d\theta_{k-1} \right) \tag{9.223}$$

By bringing back this value to equation [9.217], and by writing that the quantity as a factor of each differential element is zero, we end up with:

$$\frac{\partial Q}{\partial \theta_0} - \frac{\partial Q}{\partial \theta_k} \frac{\partial g}{\partial \theta_0} \frac{1}{\frac{\partial g}{\partial \theta_k}} = 0 \tag{9.224}$$

$$\frac{\partial Q}{\partial \theta_1} - \frac{\partial Q}{\partial \theta_k} \frac{\partial g}{\partial \theta_1} \frac{1}{\frac{\partial g}{\partial \theta_k}} = 0 \tag{9.225}$$

$$\frac{\partial Q}{\partial \theta_{k-1}} - \frac{\partial Q}{\partial \theta_k} \frac{\partial g}{\partial \theta_{k-1}} \frac{1}{\frac{\partial g}{\partial \theta_k}} = 0 \tag{9.226}$$

The solution of these gives the estimators  $\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_{k-1}$ .

The last estimator  $\hat{\theta}_k$  is obtained by applying the constraint relation, so:

$$g(\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_k) = C \tag{9.227}$$

### 9.7.2. Using Lagrange multipliers

We can systematize our search for solutions by applying the method of Lagrange multipliers.

Let us look at the system formed by the sum of the squared gaps and the constraint condition:

$$Q(\theta_0, \theta_1, \dots, \theta_k) = \sum_{i=1}^n e_i^2 \tag{9.228}$$

$$g(\theta_0, \theta_1, \dots, \theta_k) = C \tag{9.229}$$

Then we form the quantity:

$$Q_L(\theta_0, \theta_1, \dots, \theta_k) = \sum e_i^2 + \lambda[g(\theta_0, \theta_1, \dots, \theta_k) - C] \tag{9.230}$$

where  $\lambda$  is the Lagrange multiplier. We thus add an unknown and the system giving the solution of the problem is formed by writing the following  $k + 2$  equations:

$$\frac{\partial Q_L}{\partial \theta_0} + \lambda \frac{\partial g}{\partial \theta_0} = 0 \tag{9.231}$$

$$\frac{\partial Q_L}{\partial \theta_1} + \lambda \frac{\partial g}{\partial \theta_1} = 0 \tag{9.232}$$

$$\frac{\partial Q_L}{\partial \theta_k} + \lambda \frac{\partial g}{\partial \theta_k} = 0 \tag{9.233}$$

$$\frac{\partial Q_L}{\partial \lambda} = g(\theta_0, \theta_1, \dots, \theta_k) - C = 0 \tag{9.234}$$

We can either resolve the system of unknown  $k + 2$  equations or extract the value of the multiplier  $\lambda$  of one of the equations and bring it to the others in order to decrease the order of the system. This methodology is applicable to situations where several constraint relations exist simultaneously.

## 9.8. Optimizing the search for a polynomial model

### 9.8.1. System resolution

When looking for model parameters:

$$Y = \theta_0 \cdot f_0(X) + \theta_1 \cdot f_1(X) + \theta_2 \cdot f_2(X) + \dots + \theta_k \cdot f_k(X) \tag{9.235}$$

We have seen that the resolution of the least squares method with the help of matrix formalism works by means of matrix inversion:

$$A^T \cdot A = \begin{bmatrix} \sum_{i=1}^n f_0^2(x_i) & \sum_{i=1}^n f_0(x_i)f_1(x_i) & \dots & \sum_{i=1}^n f_0(x_i)f_k(x_i) \\ \sum_{i=1}^n f_0(x_i)f_1(x_i) & \sum_{i=1}^n f_1^2(x_i) & \dots & \sum_{i=1}^n f_1(x_i)f_k(x_i) \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^n f_0(x_i)f_k(x_i) & \sum_{i=1}^n f_1(x_i)f_k(x_i) & \dots & \sum_{i=1}^n f_k^2(x_i) \end{bmatrix} \quad [9.236]$$

We see that the calculations are considerably simplified if all the terms outside the principle diagonal are zero; that is, if:

$$\sum_{i=1}^n f_j(x_i) \cdot f_h(x_i) = 0 \quad [9.237]$$

This condition is a condition of orthogonal polynomials  $f_j(X)$  and  $f_h(X)$  on the defined ensemble by measurements. As a general rule, polynomials do not contain this orthogonality condition. We then make the hypothesis that the polynomial form:

$$Y = \theta_0 \cdot f_0(X) + \theta_1 \cdot f_1(X) + \theta_2 \cdot f_2(X) + \dots + \theta_k \cdot f_k(X) \quad [9.238]$$

is rewritten in the *equivalent form*:

$$Y = \Phi_0 \cdot P_0(X) + \Phi_1 \cdot P_1(X) + \Phi_2 \cdot P_2(X) + \dots + \Phi_k \cdot P_k(X) \quad [9.239]$$

Here we impose orthogonality throughout the ensemble of experimental points on the polynomials  $P_0(X), P_1(X), P_2(X), \dots, P_k(X)$ , taken two by two. This basic change results in a change of the parameters that describe the model.

The least squares resolution remains unchanged, the only modification being the matrix to be inverted, which takes the following form:

$$\mathbf{B}^T \cdot \mathbf{B} = \begin{bmatrix} \sum_{i=1}^n P_0^2(x_i) & \sum_{i=1}^n P_0(x_i)P_1(x_i) & \dots & \sum_{i=1}^n P_0(x_i)P_k(x_i) \\ \sum_{i=1}^n P_0(x_i)P_1(x_i) & \sum_{i=1}^n P_1^2(x_i) & \dots & \sum_{i=1}^n P_1(x_i)P_k(x_i) \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^n P_0(x_i)P_k(x_i) & \sum_{i=1}^n P_1(x_i)P_k(x_i) & \dots & \sum_{i=1}^n P_k^2(x_i) \end{bmatrix} \quad [9.240]$$

or, taking into account the condition of orthogonality imposed on the polynomials:

$$\mathbf{B}^T \cdot \mathbf{B} = \begin{bmatrix} \sum_{i=1}^n P_0^2(x_i) & 0 & \dots & 0 \\ 0 & \sum_{i=1}^n P_1^2(x_i) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sum_{i=1}^n P_k^2(x_i) \end{bmatrix} \quad [9.241]$$

The matrix is diagonal and the resolution of the problem no longer poses a problem. This is because the equations of the system are independent. We thus obtain the general form of the equation:

$$\hat{\Phi}_j = \frac{\sum_{i=1}^n y_i P_j(x_i)}{\sum_{i=1}^n P_j^2(x_i)} \quad [9.242]$$

The general form of the variances is:

$$V(\hat{\Phi}_j) = \frac{\sigma^2}{\sum_{i=1}^n P_j^2(x_i)} \quad [9.243]$$

The orthogonality has the secondary effect of making the estimators non-correlated, so that:

$$\text{cov}(\hat{\Phi}_j, \hat{\Phi}_h) = 0 \quad \forall h \neq j \tag{9.244}$$

**9.8.2. Constructing orthogonal polynomials using Forsythe’s method**

In books dealing with this subject, we find many polynomial forms that have orthogonal characteristics. We can start here with trigonometric polynomials with orthogonality features used to calculate the coefficients of the Fourier development of a periodic function. There are also Lagrange and Legendre polynomials that present constraints on the values situated on the axes of abscissas (these values are between -1 and +1 and/or equidistant values of x). The polynomials used by Forsythe do not have that constraint. They are written:

$$P_0(x) = 1 \tag{9.245}$$

$$P_1(x) = (x - \alpha_1).P_0(x) \tag{9.246}$$

$$P_2(x) = (x - \alpha_2).P_1(x) - \beta_2.P_0(x) \tag{9.247}$$

$$P_j(x) = (x - \alpha_j).P_{j-1}(x) - \beta_j.P_{j-2}(x) \tag{9.248}$$

In these expressions the coefficients  $\alpha_j$  and  $\beta_j$  are expressed by:

$$\alpha_j = \frac{\sum_{i=1}^n x_i . P_{j-1}^2(x_i)}{\sum_{i=1}^n P_{j-1}^2(x_i)} \quad \text{and} \quad \beta_j = \frac{\sum_{i=1}^n P_{j-1}^2(x_i)}{\sum_{i=1}^n P_{j-2}^2(x_i)} \tag{9.249}$$

We immediately notice the recurrent nature of these expressions. The polynomial of degree j and the coefficients are expressed according to the polynomials (already known) of degree j - 1 and j - 2.

In order to make these calculations more exact, Forsythe recommends establishing norms for numerical values of  $x_i$  of between -2 and +2. Not following

this recommendation nevertheless leads to polynomial values and to coefficients that are completely acceptable. In the case of a line adjustment:

$$Y = \Phi_0 \cdot P_0(X) + \Phi_1 P_1(X) \tag{9.250}$$

The calculations are carried out using the following sequence.

a) *Calculation of  $\Phi_0$*

The polynomial  $P_0(X)$  is, by definition, worth 1.

$$\text{Consequently, } \hat{\Phi}_0 = \frac{\sum_{i=1}^n y_i}{n} = \bar{y} \text{ and } V(\hat{\Phi}_j) = \frac{\sigma^2}{n} \tag{9.251}$$

b) *Calculation of  $\Phi_1$*

$$\text{We have } P_1(X) = (x - \alpha_1) \text{ with } \alpha_1 = \frac{\sum_{i=1}^n x_i}{n} = \bar{x} \tag{9.252}$$

$$\text{Consequently, } \hat{\Phi}_1 = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ and } V(\hat{\Phi}_j) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{9.253}$$

Finally, the relation is expressed in the form:

$$y = \bar{y} + \Phi_1 (x - \bar{x}) \tag{9.254}$$

This can be attained if we remember the fact that the least squares line intersects the coordinate point  $(\bar{x}, \bar{y})$ .

**9.8.3. Finding the optimum degree of a smoothing polynomial**

The recurrent nature of the Forsythe polynomials has the advantage of easily increasing the degree of the adjustment polynomial.

Generally, we can show that the sum of the squares of the residuals  $R_k$  after an adjustment of the degree  $k$ , is expressed according to the sum of the squares of the residuals  $R_{k-1}$  that correspond to an adjustment by the polynomial of degree  $k - 1$  by the relation:

$$R_k = R_{k-1} - \hat{\Phi}_k^2 \sum_{i=1}^n P_k^2(x_i) \tag{9.255}$$

This quantity decreases as  $k$  increases. In other words, the fact of increasing the degree of the adjustment polynomial acts to constrain the polynomial from coming closer to the experimental points. This can carry the risk of being unnecessary because these points have a variance that defines an uncertainty field. Therefore, it is physically sufficient that the polynomial goes into the interior of the uncertainty field without intersecting the experimental points. In addition, this constraint has no effect on the measured points. Outside these points, no constraint applies, even if the polynomial might oscillate with an amplitude well above that of the desired adjustment gain.

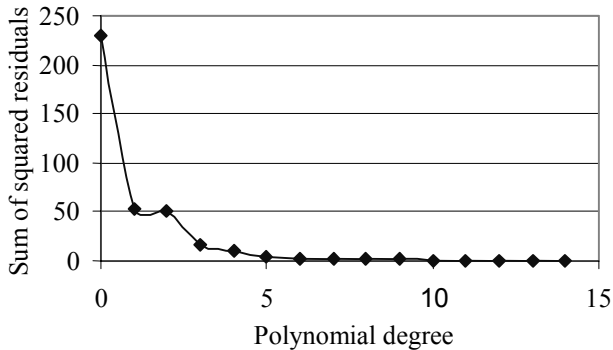
There are two simple methods for determining the optimum degree of the smoothing polynomial.

The first method calculates, after adjusting the degree  $k$ , the quantity:

$$\frac{R_k}{n - (k + 1)} \tag{9.256}$$

This is an estimator of the variance linked to the variable  $Y$  until this estimation is coherent with a predetermined value.

The second method traces the graphic representation of the development of  $R_k$  according to the function of  $k$  (see Figure 9.5). This graph presents, in general terms, a gradient rupture showing that the degree increase of the polynomial degree no longer tells us much about the model's experimental adequacy. This break occurs for an adjustment degree that coincides with the optimum degree of the polynomial. According to the conditions in the figure, this rupture will be more obvious in a system of Cartesian axes, which are systems of semi-logarithmic or logarithmic axes.



**Figure 9.5.** Representation of the sum of squared residuals showing an optimum polynomial degree equal to 3 (the polynomial of degree 2 here introduces no real reduction of the residuals)

## 9.9. Bibliography

- [FOR 77] FORSYTHE G.E., “Generation and use of orthogonal polynomials for data fitting with a digital computer”, *J. Soc. Indus. Appl. Math.*, vol. 5, no. 2, 1957.
- [JAF 96] JAFFARD P., *Méthodes de la statistique et du calcul des probabilités*, Masson Editions, 1996.
- [NIE 98] NIELSEI L., “Least-squares estimation using Lagrange multipliers”, *Metrologia*, vol. 35, no. 2, 115-118, 1998.
- [SAP 90] SAPORTA G., *Probabilités, analyse de données statistiques*, Technip Editions, 1990.
- [WIL 45] WILLIAMSON J.H., “Least square fitting of a straight line”, *Can. J. Phys*; vol. 46, 1845-1847, 1968.



*This page intentionally left blank*

## Chapter 10

# Representation and Analysis of Signals

### 10.1. Introduction

Generally, a physical phenomenon is observed through a signal carrying information. We want to extract this information and to convert it so it can be exploited. The signal processing tools and methods used in measuring a physical variable and the associated instrumentation depend on the exploitation that is carried out. An instrumentation chain consists of analog or digital electronic devices, according to the objectives for the analysis of the measured signals and the properties of these signals. Actually, rapid variations of *high frequency* signals are sometimes difficult to observe with a digital measurement chain; a purely analog measurement chain is preferable in this case. However, the development of microprocessors has meant that very sophisticated functions can be used in the processing algorithms. This means that a digital processing chain helps to solve much more complex problems than are possible for an analog processing chain.

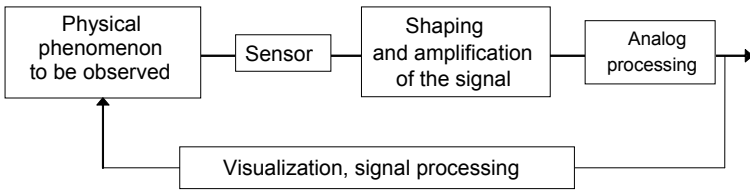
The goal of this chapter is twofold. We will first present some basic mathematical tools necessary for analyzing analog and digital signals that are present in instrumentation chains. Then we will provide some examples of signal processing methods that can lead to adaptations, according to the applications being considered. In this chapter we will especially emphasize mathematical tools for digital signal processing and time-frequency representations that are useful for extracting continuous information in signals that may not be stationary, a usual situation in practice. Any reader wanting to gain a more in-depth knowledge of basic signal representations may consult, for example, the following books: [CHA 90];

[COT 97]; [COU 84]; [DEL 91]; [GAS 90]; [MAX 89]; [MAX 96]; [PIC 89, 93, 94, 95]; and [ROD 78]. In addition, some methods of analog processing are discussed in Chapters 4 and 5.

## 10.2. Analog processing chain

### 10.2.1. Introduction

Sensors used to observe a physical phenomenon provide a signal that carries information. This signal follows, in this case, a continuous time law and thus it is a purely analog instrumentation chain throughout. It can undergo various processing by purely analog electronic devices, including: amplification; filtering; modulation-demodulation; clipping, correlation; synchronous demodulation. The synoptic schema of this kind of processing chain is shown in Figure 10.1.



**Figure 10.1.** *Analog measurement chain*

In certain applications, for example, in control systems, the signal obtained after processing can be reinjected into the process input in the form of a control law that will eventually modify the behavior of the observed physical phenomenon. For example, in active vision, the analyzed signal can serve to follow the trajectory of an object; and, if necessary, correct the position, the direction and the settings of the sensors that are following the scene.

Signals can be analyzed in the time domain or in the frequency domain. The following section gives some definitions that are often used in the analysis of analog signals.

### 10.2.2. Some definitions and representations of analog signals

#### 10.2.2.1. Deterministic signals

A signal can be described by a mathematical model. For example, a sinusoidal signal is determined by its magnitude, its pulsations or its frequency, and its phase at

time origin. Practically, a deterministic model is only partially known, and its unknown parameters introduce random behaviors that are more or less unpredictable. A deterministic signal has little importance in real situations, since it does not carry information, apart from its presence or absence. However, this kind of signal can act as an excitation signal that indirectly obtains information about a physical variable through the interaction that this signal can have with the physical system being analyzed. A signal carrying information of uncertain nature is a random signal. Some statistical properties of this signal allow us to describe it simply or even to evaluate our knowledge of this signal by relating it to a known model.

In practice, a sensor is activated from a given time  $t_0$ , very often chosen as a time origin. The signals which are observed and processed are then considered as zero signals up to the time  $t_0 = 0$ . These are called causal signals. In addition, signals are observed during a finite period  $T$ . By commodity, especially of calculation, we very often represent an observed signal by a periodic signal; or we construct a periodic auxiliary signal from the observed signal. The rest of this section will provide some fundamental descriptions of signals and present some of their usual features.

An analog signal represented in the time domain by a scalar function  $x(t)$  of the continuous variable  $t$  can be characterized in different ways. Subject to the existence of integrals, we have the following definitions:

- mean value  $m_x = \frac{x(t)}{T} = \lim_{T \rightarrow +\infty} \frac{1}{2T} \int_{-T}^T x(t) dt$  ;
- energy  $E_x = \lim_{T \rightarrow +\infty} \int_{-T}^T |x(t)|^2 dt$  ;
- instantaneous power  $p_x(t) = |x(t)|^2$  ;
- mean power  $P_x = \frac{|x(t)|^2}{T} = \lim_{T \rightarrow +\infty} \frac{1}{2T} \int_{-T}^T |x(t)|^2 dt$  ;
- mean power of the signal fluctuations around its mean  $\sigma_x^2 = \frac{|x(t) - m_x|^2}{T} = \frac{|x(t)|^2}{T} - |m_x|^2$  .

The time representation of a signal in its form  $x(t)$  is the most natural. It directly shows the magnitude variation of the signal according to time. However, there are other representations. The remainder of this section will discuss the frequency representation that indicates the variation frequency of the signal magnitude. This

representation is ensured by the Fourier transform. A more general discussion of representations is given in section 10.6.

10.2.2.1.1. The Fourier transform

Under certain conditions, always verified by physical signals, a signal has an equivalent representation that is a function  $X(f)$  of the frequency  $f$  or even a continuous function  $X(\omega)$  of the pulsation  $\omega$  called the Fourier transform. This is defined by:

$$X(f) = \int_{-\infty}^{+\infty} x(t) \exp(-j2\pi ft) dt \tag{10.1}$$

and:

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) \exp(-j\omega t) dt \tag{10.2}$$

A sufficient condition of existence of this representation is that the energy of the signal must be finite. For clarity, from now on,  $X(f) = \text{TF}[x(t)]$  represents the function defined by equation [10.1] and  $X(\omega)$  will represent the function defined by equation [10.2]. The inverse Fourier transform is expressed by the relation:

$$x(t) = \int_{-\infty}^{+\infty} X(f) \exp(j2\pi ft) df = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) \exp(j\omega t) d\omega \tag{10.3}$$

If  $x(t)$  is a real signal, its Fourier transform is a complex function of even module and odd argument:  $X(-f) = X^*(f)$ . In addition to the linearity of the Fourier transform, we see several other properties: the time reversal property:  $\text{TF}[x(-t)] = X(-f)$ ; the conjugation property  $\text{TF}[x^*(t)] = X^*(-f)$ ; and the delay theorem  $\text{TF}[x(t-\tau)] = \exp(-j2\pi f\tau)X(f)$ . The convolution product of two signals  $x(t)$  and  $y(t)$ , denoted by  $(x*y)(t)$  is defined by:

$$(x * y)(t) = (y * x)(t) = \int_{-\infty}^{+\infty} x(u)y(t-u) du$$

and this expression may be written more easily in the frequency domain than in the time domain. Indeed, we have  $\text{TF}[(x*y)(t)] = \text{TF}[x(t)]\text{TF}[y(t)]$ . This property, which is also called the Plancherel theorem, is very useful for the linear filtering of signals and for calculating correlation functions.

## 10.2.2.1.2. Correlation and spectral density

The energy of the signal  $x(t)$  is also expressed in the frequency domain:

$$E_x = \int_{-\infty}^{+\infty} |x(t)|^2 dt = \int_{-\infty}^{+\infty} |X(f)|^2 df = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(\omega)|^2 d\omega$$

This means the energy of a signal does not depend on the chosen representation. The function  $\Phi_{xx}(f) = |X(f)|^2$  is called the energy spectral density of the signal  $x(t)$  and is a frequency representation of the energy. This quantity is, by definition, always real and positive. The inverse Fourier transform of  $\Phi_{xx}(f)$ , written  $\gamma_{xx}(\tau)$ , is called the energy autocorrelation function of the energy signal  $x(t)$ . By writing that  $|X(f)|^2 = X(f)X^*(f)$  and by using the property  $X^*(f) = \text{TF}[x^*(-t)]$  we deduce from this:

$$\gamma_{xx}(\tau) = \int_{-\infty}^{+\infty} \Phi_{xx}(f) \exp(j2\pi f\tau) df = \int_{-\infty}^{+\infty} x(t)x^*(t-\tau) dt$$

$\gamma_{xx}(\tau)$  expresses the resemblance between  $x(t)$  and  $x^*(t-\tau)$  and produces the continuous autosimilarities in the signal. If the signal  $x(t)$  is real, its autocorrelation function is real, even, and maximum at the time origin.

These results can be applied to two signals  $x(t)$  and  $y(t)$ . First of all, we have:

$$\int_{-\infty}^{+\infty} x(t)y^*(t) dt = \int_{-\infty}^{+\infty} X(f)Y^*(f) df$$

This relation constitutes the Parseval theorem. The quantity  $\Phi_{xy}(f) = X(f)Y^*(f)$  is called the cross-energy spectral density of the signals  $x(t)$  and  $y(t)$ . The inverse Fourier transform of this quantity, written  $\gamma_{xy}(\tau)$ , is the energy cross-correlation function of the signals  $x(t)$  and  $y(t)$ . Then we have the relation:

$$\gamma_{xy}(\tau) = \int_{-\infty}^{+\infty} \Phi_{xy}(f) \exp(j2\pi f\tau) df = \int_{-\infty}^{+\infty} x(t)y^*(t-\tau) dt$$

which constitutes the Wiener-Kintchine theorem. If the signal  $y(t)$  is also obtained by a linear filtering of the signal  $x(t)$ , then we have  $\Phi_{yx}(f) = G(f)\Phi_{xx}(f)$  or  $\Phi_{yy}(f) = |G(f)|^2 \Phi_{xx}(f)$ ,  $G(f)$  being the frequency response of the filter. These relations correspond to the interference formula.

We will now discuss a specific example dealing with periodic signals.

10.2.2.1.3. Periodic signals

Obtaining a spectral representation of a signal requires knowing this signal throughout its entire time domain. Practically, we only know the signal  $x(t)$  on a finite time support, for example  $[0, T]$ . Outside this interval, we consider that the signal is equal to zero in order to calculate the previously introduced quantities. It is also possible to define an auxiliary periodic signal  $x_T(t)$  that is equal to  $x(t)$  on the interval  $[0, T]$  and is of period  $T$ . Always under certain conditions, the representation of  $x_T(t)$  in the form of a Fourier series is:

$$x_T(t) = \sum_{n=-\infty}^{+\infty} c_n \exp(j2\pi n t/T) \text{ with } c_n = \frac{1}{T} \int_0^T x(t) \exp(-j2\pi n t/T) dt$$

where  $1/T$  is the fundamental frequency and  $c_n$  the amplitude of the harmonic of rank  $n$ . With this new representation, integrating the signal is done on a finite interval, but we have to calculate an infinite number of coefficients. This is why another representation will be defined later for the analysis of digital signals that are defined by a finite number of harmonics (see section 10.6).

By introducing the distribution formalism [ROD 78], we can extend the definition of the Fourier transform to periodic signals, and connect this to Fourier’s serial decomposition. Actually, by introducing the impulse signal that is expressed by  $\text{TF}[\exp(j2\pi t/T)] = \delta(f - 1/T)$ , we firstly get:

$$X_T(f) = \sum_{n=-\infty}^{+\infty} c_n \delta(f - n/T)$$

The mean power  $P_T$  of the periodic signal  $x_T(t)$  is written in the form:

$$P_T = \frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 dt = \sum_{n=-\infty}^{+\infty} |c_n|^2 = \int_{-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} |c_n|^2 \delta(f - n/T) df$$

The function  $\Phi_T(f) = \sum_{n=-\infty}^{+\infty} |c_n|^2 \delta(f - n/T)$  is the power spectral density of the periodic signal  $x_T(t)$ . The inverse Fourier transform of this quantity is the power autocorrelation function of this signal. It is defined by:

$$\gamma_T(\tau) = \frac{1}{T} \int_{-T/2}^{T/2} x_T(t) x_T^*(t - \tau) dt = \sum_{n=-\infty}^{+\infty} |c_n|^2 \exp(j2\pi n \tau/T)$$

and is also periodic.

### 10.2.2.2. Random signals

In many applications, we are only interested in statistical properties of first and second orders of random signals: mean, correlation, power and signal-to-noise ratio. This section presents these ideas.

Strictly speaking, a random signal is stationary if all its statistical properties are invariant by changing the origin of time. If we limit this property only to the first and second statistical times, we say that the signal is in a wide sense stationary. This last feature is commonly accepted as the starting hypothesis for processing methods that use the statistical properties of the signal to be analyzed. The random process fluctuates and the observed signal corresponds only to a special realization of the random process  $x(t, u)$  called trajectory. In absolute terms, in order to know the statistical variables of the process, we must carry out the same experiment many times. This is obviously not possible in most situations and we usually assume that the nature of the information conveyed by the time behavior of the signal is the same as that which is conveyed by carrying out the process a number of times. The stationary random process is called ergodic if all the statistical means coincide asymptotically towards the time means, in particular that is if:

$$E[x^n(t, u)] = \int_{-\infty}^{+\infty} x^n f(x) dx = \lim_{T \rightarrow +\infty} \frac{1}{2T} \int_{-T}^T x^n(t, u) dt = \underline{x^n(t, u)}$$

where  $f(x)$  is the probability density of the random signal  $x(t, u)$ . The correlation function of a stationary signal is defined by:

$$\gamma_{xx}(\tau) = E[x(t, u)x^*(t - \tau, u)]$$

For this kind of signal,  $\gamma_{xy}(\tau)$  does not depend on the time  $t$ . Moreover, if the signal is ergodic, we get:

$$\gamma_{xx}(\tau) = \lim_{T \rightarrow +\infty} \frac{1}{2T} \int_{-T}^T x(t, u)x^*(t - \tau, u) dt$$

In practice, the random signal is observed on a finite time support and we therefore exploit windowed time information.

The power spectral density  $\Phi_{xx}(f)$  of a stationary signal is the Fourier transform of the autocorrelation function (the Wiener-Khinchine theorem):  $\Phi_{xx}(f) = \text{TF}[\gamma_{xx}(\tau)]$ . For stationary and ergodic signals, the time means and the



statistical means coincide, the quadratic mean is directly related to the power spectral density, and we have:

$$P_x = \overline{|x(t,u)|^2} = E\left[|x(t,u)|^2\right] = \gamma_{xx}(0) = \int_{-\infty}^{+\infty} \Phi_{xx}(f) df$$

As an example, we look at a signal that is both stationary and ergodic: that of a white noise  $w(t)$  characterized by a zero mean and a uniform spectral density:  $\Phi_{ww}(f) = c$ . Its autocorrelation function is thus  $\gamma_{ww}(\tau) = c \delta(\tau)$ . A white noise is often used in a random process with a power spectral density that is constant up to a frequency much higher than the maximum frequency intervening in modelization of the signal or of the system being studied – even higher than the passband of the process chain.

The cross-correlation function of two random signals  $x(t, u)$  and  $y(t, u)$ , which are jointly stationary and ergodic, is defined by:

$$\gamma_{xy}(\tau) = \lim_{T \rightarrow +\infty} \frac{1}{2T} \int_{-T}^T x(t,u)y^*(t-\tau,u) dt = E[x(t,u)y^*(t-\tau,u)]$$

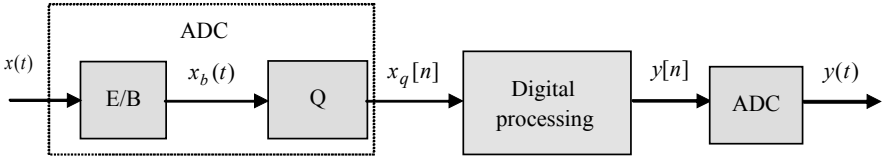
and their cross-power spectral density is shown by  $\Phi_{xy}(f) = TF[\gamma_{xy}(\tau)]$ . The signals  $x(t, u)$  and  $y(t, u)$  are said to be uncorrelated if their correlation is zero (whatever  $\tau$  may be). If the signal  $y(t, u)$  is obtained by a linear filtering of the signal  $x(t, u)$ , the interference formula is written  $\Phi_{yx}(f) = G(f)\Phi_{xx}(f)$  or  $\Phi_{yy}(f) = |G(f)|^2 \Phi_{xx}(f)$ ,  $G(f)$  being the filter’s frequency response.

Many analog circuits allow us to carry out, in an approximate manner, the functions described above. Recent developments in digital signal processing have led to methods that are more sophisticated than those obtained with analog electronics. The following section will discuss the tools and representations of digital signals.

### 10.3. Digital processing chain

#### 10.3.1. Introduction

Digital functions process series of numbers that usually come from sampling an analog signal, the amplitude being quantified. Figure 10.2 shows part of a synoptic of a digital instrumentation chain. We note by  $x_b(t)$  the sampled and hold signal and  $x_q[n]$  the digital signal.



**Figure 10.2.** *Digital processing chain*

In an instrumentation chain, in addition to the principal elements that make up the analog chain seen above, we add an analog-to-digital converter (ADC) and if, after digital processing, we want to observe or exploit the signal in an analog form, a digital-to-analog converter (DAC). An ADC converter has a sampling/holding function (S/H) and a quantizer (Q). The sampling/holding function takes instantaneous samples of the time signal  $x(t)$  and keeps them constant to the input of the quantizer during a period necessary in order to convert the sampled signal into a digital signal. The study and processing of digital signals require us to first bear in mind some basic concepts, including the sampling theorem and quantization.

### 10.3.2. Sampling and quantization of signals

#### 10.3.2.1. The Fourier transform and sampling

Periodic sampling is the acquisition of values or samples of the analog signal  $x(t)$  at time  $t_n = nT_e$ ,  $T_e$  being the sampling period. The choice of the sampling period depends on the spectral content of the signal  $x(t)$ . The informational content will remain intact if we theoretically have the capability to exactly reconstruct the analog signal  $x(t)$  from the sampled signal  $x[n] = x(nT_e)$ . The sampling theorem establishes a criterion for the preservation or alteration of the quantity of information contained in the sampled signal and expresses the conditions for a good restitution of the original signal.

##### 10.3.2.1.1. The discrete time Fourier transform

The Fourier transform of the sampled signal  $x[n]$  is defined by:

$$X_{TD}(f) = \sum_n x[n] \exp(-j2\pi fn) \quad [10.4]$$

or also:

$$X_{TD}(\omega) = \sum_n x[n] \exp(-j\omega n) \quad [10.5]$$

In the following presentation,  $X_{TD}(f)$  will be the function defined by equation [10.4] and  $X_{TD}(\omega)$  the function defined by equation [10.5]. We point out that the variable  $f$  intervening in the definition of  $X_{TD}(f)$ , which is the Fourier transform of the sampled signal, is without dimension. While it is involved in the definition of  $X(f)$ , the Fourier transform of an analog signal  $x(t)$  has a dimension: if the variable  $t$  represents the time and is expressed in seconds, then  $f$  is expressed in Hertz. The function  $X_{TD}(f)$  is periodic (of period 1) and is expressed, as we will see, according to  $X(f)$ . To establish this result, we first of all introduce the impulse train  $\delta_{T_e}(t) = \sum_n \delta(t - nT_e)$  then the ideal sampled signal:

$$x_e(t) = x(t) \delta_{T_e}(t) \tag{10.6}$$

If the signal  $x(t)$  is continuous, we have:

$$x_e(t) = \sum_n x[n] \delta(t - nT_e) \tag{10.7}$$

Thus, the signals  $x_e(t)$  and  $x[n]$  are equivalent. First, taking the Fourier transform of equation [10.7], we establish that:

$$X_e(f) = \sum_n x[n] \exp(-j2\pi f n T_e) = X_{TD}(f T_e)$$

$X_e(f)$  is thus periodic of period  $F_e = 1/T_e$ . Starting from Fourier series  $\delta_{T_e}(t)$ , by taking the transform of this decomposition, we establish that:

$$\text{TF}[\delta_{T_e}(t)] = \frac{1}{T_e} \sum_n \delta(f - n/T_e) = \frac{1}{T_e} \delta_{1/T_e}(f)$$

This tells us that the Fourier transform of an impulse train is still an impulse train. The Fourier transform of equation [10.6] is then written:

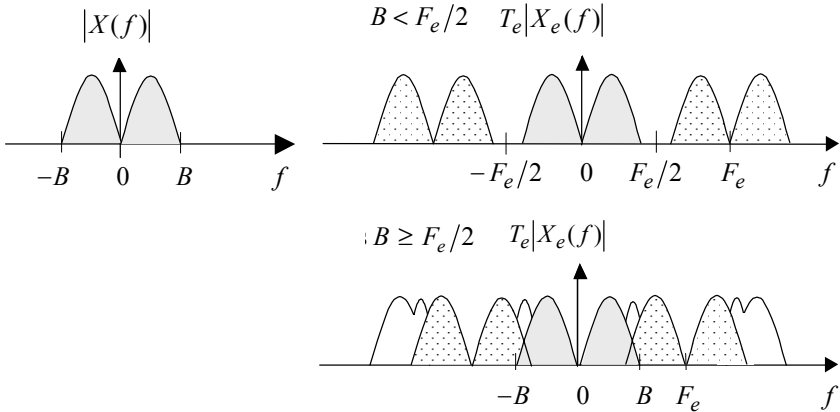
$$X_e(f) = \frac{1}{T_e} (X * \delta_{1/T_e})(f)$$

The impulse function being the neutral element of the convolution product, we have deduced from this fact that:

$$X_e(f) = X_{TD}(f T_e) = \frac{1}{T_e} \sum_n X(f - n/T_e)$$

The Fourier transform of the sampled signal is obtained from that of the analog signal by periodic replication. Figure 10.3 illustrates this point for a bandlimited

signal; that is, that  $B$  exists so that  $|X(f)| = 0 \quad \forall |f| \geq B$  or that all the frequency representation of  $x(t)$  is inside the band  $]-B, B[$ . Two examples are then considered:  $B < F_e/2$  or  $B \geq F_e/2$ .



**Figure 10.3.** Sampling and aliasing

If  $BT_e < 0.5$ ,  $X(f)$  and  $X(f - n/T_e)$  do not overlap for all  $n \neq 0$ ;  $X(f)$  can then be obtained from  $X_e(f)$  by simple multiplication by the frequency response  $H(f)$  of a ideal lowpass filter defined by:

$$H(f) = \begin{cases} T_e & \text{if } -F_e/2 \leq f \leq F_e/2, \\ 0 & \text{otherwise.} \end{cases}$$

The original signal  $x(t)$  can then be reconstructed from the ideal sampled signal  $x_e(t)$  or then from the sampled signal  $x[n]$ . After filtering the signal  $x_e(t)$  by the lowpass filter, the information is intact.

If  $BT_e \geq 0.5$ , the supports of  $X(f)$  and of  $X(f - n/T_e)$  ( $n \neq 0$ ) overlap, which leads to a spectrum aliasing, as shown in Figure 10.3. Thus we cannot come back to  $X(f)$  from  $X_e(f)$ . This brings us to a representation of the sampling theorem.

#### 10.3.2.1.2. Sampling theorem

An analog bandlimited signal  $x(t)$  exists with  $B$  so that  $|X(f)| = 0 \quad \forall |f| \geq B$  can be reconstructed from the samples  $x[n] = x(nT_e)$  without loss of information if the sampling frequency  $F_e$  is higher than  $2B$ .

10.3.2.1.3. Interpolation

The interpolation formula is obtained by going back to the time domain of the relation  $X(f) = H(f)X_e(f)$ . We then have  $x(t) = (h * x_e)(t)$ , with:

$$h(t) = T_e \int_{-F_e/2}^{+F_e/2} e^{j2\pi ft} df = \text{sinc}(F_e t)$$

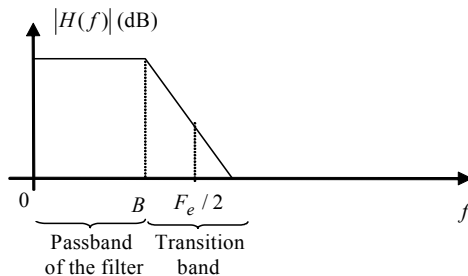
where  $\text{sinc}(x) = \sin(\pi x)/\pi x$ . By using the properties of the impulse function, we deduce from it that:

$$x(t) = \sum_n x(nT_e) \text{sinc}((t - nT_e)/T_e)$$

The duality of the time representation of a signal with its frequency representation helps establish an equivalent theorem for the sampled signals in the frequency domain. Sampling the Fourier transform of a signal of finite support signal (for example,  $x(t) = 0, \forall t \notin [0, T[$ ) does not lead to any information loss if the Fourier transform is sampled with a step  $\Delta f \leq 1/T$  (see section 10.6).

These results lead us to several remarks concerning practical implementation. First of all, the interpolation formula requires that we calculate an infinite sum of terms. Practically, only a finite number of terms can be calculated, and the reconstruction of the signal  $x(t)$  can only be approximated.

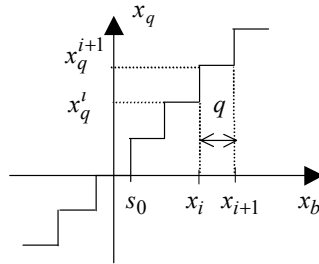
Secondly, the sampling condition  $F_e \geq 2B$  cannot always be achieved and the inputs of the digital measuring devices must have anti-aliasing filters. This kind of filter guarantees the sampling condition independently of the applied input signal. The transition band of this lowpass filter is not inconsiderable (see Figure 10.4), and thus we must always take it into account when choosing a sampling rate.



**Figure 10.4.** Choice of the pass-band of an anti-folding filter

### 10.3.2.2. Quantization

We are only looking at real-valued signals. Representing numbers on a calculator requires approaching their values by whole numbers coded on a given number of bits. The quantization models the operation that carries out this approximation. The input-output non-linear characteristic represents the nature of the approximation (Figure 10.5).



**Figure 10.5.** *Input-output characteristic of a quantizer*

We write  $x_q^i$  as the output values of the quantizer. With the example shown in Figure 10.5, all the quantization steps are equal: quantization is then uniform and  $x_q^{i+1} = x_q^i + q$ ,  $q$  being the quantization step. The quantity  $s_0$ , shown in Figure 10.5, models a threshold that characterizes the nature of the quantization. Two examples are current today: quantization by truncation ( $s_0 = q$ ) and quantization by rounding off ( $s_0 = q/2$ ).

Quantization of a discrete time signal introduces a precision limitation. Even if the quantization operation is nonlinear, by commodity to lead the calculations, we modelize the quantization operation by a linear relation:  $x_q(t) = x(t) + \eta_q(t)$  where the signal  $\eta_q(t)$  which represents the quantization noise is presumed to be independent of  $x(t)$ . In many applications, if we do not know *a priori* the statistical properties of this noise, we presume that its values are uniformly distributed throughout their extreme values. Consequently, we regard the quantization noise as being a random signal of uniform probability density function  $f(a)$ . In the case of a rounded off quantization discussed here, the extreme values are  $-q/2$  and  $q/2$  (see Figure 10.6).

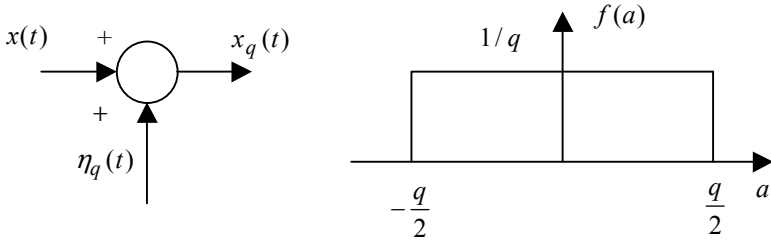


Figure 10.6. Modelization of quantization noise

Using these simplified hypotheses, we can calculate the statistical properties of the second order of the quantization error and a signal-to-noise ratio. The error mean statistic is zero and its variance or its power is:

$$\sigma_{\eta}^2 = \int_{-\infty}^{+\infty} a^2 f(a) da = \frac{1}{q} \int_{-q/2}^{q/2} a^2 da = \frac{q^2}{12}$$

If the input signal is a triangular signal that is not clipped, the quantization noise is a sawtooth signal of period  $T$  with a power expressed by:

$$P_{\eta} = \frac{1}{T} \int_{-T/2}^{+T/2} \left( \frac{-qt}{T} \right)^2 dt = \frac{q^2}{12}$$

We find the same result as before. If the input signal has any form, the power of the quantization error must be calculated by considering both its probability density and time representation. However, if the quantization step and sampling period are sufficiently low, it is justifiable to conserve the uniform law hypothesis. With this hypothesis, for a sinusoidal signal of maximum amplitude  $A$  and a quantization operation on  $n$  bits with a dynamic  $[-A, A]$ , the quantization step is  $q = 2A/2^n$  and the error variance  $\sigma_{\eta}^2 = A^2 / (3 \times 2^{2n})$ . The power of the signal being equal to  $A^2/2$ , the signal-to-noise ratio is  $RSB = 3 \times 2^{2n-1}$ . In decibels, this becomes a widely used relation in technical documentation [AZI 96]:

$$RSB = 10 \log(3/2) + 20 n \log 2 = 1.76 + 6.02 n$$

We must keep in mind that the signal-to-noise ratio increases by 6 dB each time we increase the converter capacity by one bit. In Chapter 6, we described the principles of sigma-delta converters that can significantly increase the signal-to-noise ratio with a looped system. In the remainder of this chapter, no more distinction will be made between  $x[n]$  and  $x_q[n]$ .

The concepts of correlation and spectral density relating to analog signals seen in section 10.2 can easily be generalized to discrete time signals; the time integrals are then replaced by sums.

## 10.4. Linear digital filtering

Only digital filters will be considered in this chapter. For information on producing analog filters, we direct the reader to Chapter 5, and for details of mathematical tools, to the general texts cited at the beginning of this chapter. We will now discuss analysis tools used for digital filters.

### 10.4.1. The $z$ transform

We call the  $z$  transform of the sequence of  $x[n]$ , the complex function defined by:

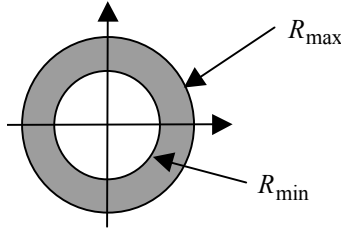
$$X(z) = \text{TZ}\{x[n]\} = \sum_n x[n] z^{-n}$$

We can show that this series converges if the complex number  $z$  belongs to a ring of complex plane that is delimited by two concentric circles centered on the origin of the radius  $R_{\min}$  and  $R_{\max}$ , that is, for  $|z| \in D_x = ]R_{\min}, R_{\max}[$  (see Figure 10.7). If the signal is causal as well, the convergence domain becomes  $D_x = ]R_{\min}, +\infty[$ . For certain signals, the  $z$  transform converges in  $|z| = R_{\min}$ ; for example, the  $z$  transform of the unit impulsion unit  $x[n] = \delta[n]$  where  $\delta[n]$  is the Kronecker symbol ( $\delta[n] = 1$  for  $n = 0$  and  $\delta[n] = 0$  for  $n \neq 0$ ), exists whatever the value of  $z$  and the convergence ring is the entire complex plane:  $D_x = [0, +\infty[$ .

Some properties of the  $z$  transform are:

- delay:  $\text{TZ}\{x[n - p]\} = z^{-p} X_z(z)$  ;
- differentiation:  $\text{TZ}\{n x[n]\} = -z \frac{dX_z(z)}{dz}$  ;
- convolution:  $\text{TZ}\left\{ (x * y)[n] = \sum_k x[k] y[n - k] \right\} = X(z) Y(z)$ .





**Figure 10.7.** *Convergence domain*

The inverse z transform is:

$$x[n] = \frac{1}{j2\pi} \oint_C X(z) z^{n-1} dz$$

where  $C$  models a closed contour directed in the direct trigonometric direction belonging to the convergence ring and surrounding the origin. The direct calculation of this integral uses the residual theorem [BEL 87] and is often difficult. We prefer to carry out a simple partial expansion to cause, as much as possible, basic  $z$  transforms for which we know the corresponding signals, with the help of tables. We deduce the signal  $x[n]$  from linearity. We point out here that knowledge of the convergence domain is required, since two different signals can give the same  $z$  transform, as we will see. Actually, the  $z$  transforms of the causal signal  $x[n] = 1$  if  $n \geq 0$  and  $x[n] = 0$  otherwise, and of the anticausal signal  $y[n] = -1$  if  $n \leq -1$  and  $y[n] = 0$  otherwise are equal:  $X(z) = Y(z) = (1 - z^{-1})^{-1}$  but with different convergence domains:  $D_x = ]1, +\infty[$  and  $D_y = ]0, 1[$ .

### 10.4.2 Filtering applications

We are going to apply certain properties of the  $z$  transform for digital filtering. The kind of digital processing described here concerns linear time and invariant systems, that is, linear filters, for which the input signal  $x[n]$  is related to that of the output  $y[n]$  by the constant-coefficient difference equation:

$$y[n] = \sum_{i=0}^p a_i x[n-i] - \sum_{j=1}^q b_j y[n-j] \text{ with } p > 0 \text{ and } q > 0 \tag{10.8}$$

This kind of filter is causal. This relation is written in the form of a discrete convolution equation:

$$y[n] = \sum_k h[k] x[n-k]$$

The causality of the filter means that  $h[n] = 0$  for  $n < 0$ . When the coefficients  $b_j$  are not all zero, the filter is called recursive and its impulse response is infinite (IIR). However, when the coefficients  $b_j$  ( $j = 1, \dots, q$ ) are zero, the coefficients  $a_i$  ( $i = 0, \dots, p$ ) are the  $p + 1$  non-zero coefficients of the impulse response of the filter and we say that the filter has a finite impulse response (FIR).

By taking the  $z$  transform from the relation shown in equation [10.8], and by using the delay property of the  $z$  transform, we deduce that:

$$Y(z) \left( 1 + \sum_{j=1}^q b_j z^{-j} \right) = \sum_{i=0}^p a_i z^{-i} X(z)$$

then the transfer function of the system:

$$H(z) = \frac{Y(z)}{X(z)} = \frac{a_0 + a_1 z^{-1} + \dots + a_p z^{-p}}{1 + b_1 z^{-1} + \dots + b_q z^{-q}}$$

This transfer function is a rational fraction. The filter is then called a rational filter. The roots of  $a_0 + a_1 z^{-1} + \dots + a_p z^{-p} = 0$  are called zeros of  $H(z)$  and the roots of  $1 + b_1 z^{-1} + \dots + b_q z^{-q} = 0$  poles of  $H(z)$ . In the specific example when all the coefficients of  $a_i$  and  $b_j$  are real, we have the property  $H(z^*) = H^*(z)$ . The poles and zeros of  $H(z)$  are then real or complex and conjugate. If the unit circle ( $z = \exp(j2\pi f)$ ) belongs to the convergence domain, the transfer function of the filter  $H(z)$  allows for the expression of the complex gain  $G(f)$ . We then have  $G(f) = H(\exp(j2\pi f))$ . The unit circle can be either graduated according to angle or frequency (between  $-1/2$  and  $1/2$ ).

### *Causality and stability*

The impulse response of a linear filter expresses generally as a sum of exponential signals. The stability condition  $\sum_n |h[n]| < +\infty$  is expressed by the fact that the unit circle must belong to the convergence ring. The convergence ring of a causal linear filter being of the form  $]R_{\min}, +\infty[$ , the stability thus imposes  $R_{\min} \leq 1$  which means that the poles of the transfer function are inside the unit circle.

The stable rational and causal filters constitute a specific class of filters called *dynamic filters*. *Minimum phase* filters are dynamic filters with zeros that are also inside the unit circle. These filters are especially useful because their inverse is also a minimum phase filter. The inverse filter is also stable.

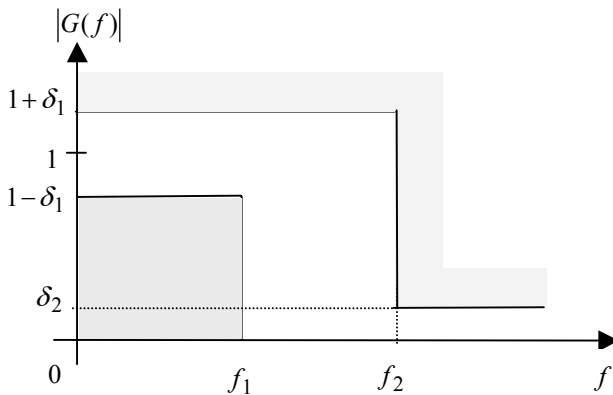
*Stability of FIR and IIR filters*

The transfer function of a causal FIR filter being of the form  $H(z) = a_0 + a_1z^{-1} + \dots + a_pz^{-p}$ , this filter is always stable whatever the  $a_i$  coefficients. However, the transfer function of a recursive filter being a  $z$  rational fraction, the stability of this filter requires that all its poles are inside the unit circle.

*Synthesis of digital filters*

The synthesis of a digital filter consists of determining its transfer function or its impulse response from specifications defined in the time or frequency domain. The specifications can, for example, be represented by the frequency specifications of a lowpass filter shown in Figure 10.8.

Synthesis consists of finding a filter that fulfills the specifications. Choosing an FIR or IIR filter depends on the implementation constraints. The advantages or disadvantages of each structure must be considered in the choice. For example, linear phase filters are used with FIR filters. For a frequency specification, the required calculation charge for an FIR filter is generally much higher than that needed for a recursive filter. The synthesis of FIR filters is not discussed in this chapter. We advise the interested reader to consult [BEL 87], [FON 81], [LAB 88] and [OPP 75].



**Figure 10.8.** Example of a lowpass filter specification envelope

### 10.4.3. Synthesis of IIR filters

We here present two synthesis methods for IIR filters. The first method involves approaching the transfer function of an analog filter by a discrete time transfer function. The second method calculates the coefficients of the filter by an optimization process. Analog filters are usually Butterworth, Bessel, Tchebycheff or elliptic filters.

#### 10.4.3.1. Methods using an analog reference filter

The impulse invariance method consists of making a digital filter with an impulse response which coincides with the impulse response of a given analog filter at sampling times. This method has several disadvantages, It requires long calculations, introduces distortions in frequency responses due to spectrum aliasing, and does not conserve the DC gain. This is why we prefer to carry out a direct transform of the transfer function  $H_a(p)$  of the analog filter according to a transfer function  $H_n(z)$  of the digital filter by directly replacing  $p$  with a function  $f(z)$ . The most often used transform is the bilinear transform which, at every  $M(z)$  point of affix  $z$ , makes the  $M'(p)$  point correspond to the affix  $p$  by the passing relation:

$$p = \frac{2}{T_e} \frac{1 - z^{-1}}{1 + z^{-1}}$$

This transform is justified by the approximation of the integral of a signal with the trapezoidal method, so:

$$y((n+1)T_e) = \int_{-\infty}^{(n+1)T_e} x(t) dt = y(nT_e) + \int_{nT_e}^{(n+1)T_e} x(t) dt$$

can be approximated by:  $y((n+1)T_e) = y(nT_e) + \frac{T_e}{2} (x((n+1)T_e) + x(nT_e))$ .

In this way, the transfer function  $\frac{T_e}{2} \frac{1+z^{-1}}{1-z^{-1}}$  corresponds in an integration in the Laplace domain of the transfer function  $1/p$ . With this transform, we link a point of the imaginary axis of affix  $p = j\omega_a$  with  $\omega_a \in ]-\infty, +\infty[$  and the point of  $z = \exp(j\omega_n T_e)$  with  $\omega_n = \frac{2}{T_e} \text{artg}(\omega_a T_e / 2)$  and  $\omega_n \in ]-\pi/T_e, \pi/T_e[$ . This transform introduces a frequency response distortion due to its non-linearity. If  $\omega_a$  is low compared to the sampling frequency,  $\omega_n$  is close to  $\omega_a$ , so the distortion introduced by the bilinear transform for the low frequencies is low. If  $T_e$  increases, the non-linear relation that links  $\omega_a$  and  $\omega_n$  involves a frequency deformation of the

axis and, consequently, a specification deformation. A scale factor  $k$  in part compensates the distortion in order to fit the frequency response of the analog filter with the frequency response of the digital filter within a given frequency range. The transform is then as follows:

$$p = k \frac{1 - z^{-1}}{1 + z^{-1}}$$

A suitable choice of the factor  $k$  is to impose the same frequency responses  $\omega_a = \omega_{oa}$  and  $\omega_n = \omega_{on}$ . For this, it is enough to choose  $k = \omega_{oa} / \text{tg}(\omega_{on} T_e / 2)$ . If  $\omega_c$  is the cut-off frequency of the analog filter, the choice  $\omega_{oa} = \omega_{on} = \omega_c$  helps conserve the cut-off frequency. The choice  $k = 2/T_e$  helps to conserve the behavior at low frequencies. With the bilinear transform, a low-pass or pass-band analog filter is converted to a digital filter of the same type; only the characteristic frequencies are modified. But the linear property of the phase of a filter will not be conserved. We mention here that there are other transforms capable of carrying out synthesis [LAB 88].

#### 10.4.3.2. Methods of synthesis by optimization

The synthesis methods discussed above introduce a distortion of the frequency response. The development of computer-based techniques have led to the development of synthesis methods which make use of optimization algorithms of a cost function. This approach consists of imposing *a priori* the structure of the digital filter (of a recursive filter, for example), then adjusting the coefficients of this filter to approach as closely as possible an impulse response, a frequency response, or satisfy specifications established by, for example, a specification envelope.

Several criteria can be chosen to carry out approximation in the best possible way. In general, these criteria lead to resolving a non-linear optimization issue with constraints. This means we need a method to determine the coefficients of the filter; these coefficients will optimize the criterion with the goal of determining the optimal filter in a given family. For example, we look for the coefficients of a filter with a given structure that will minimize the distance between the frequency response of this filter and the required frequency response for the frequencies  $f_i (i = 1, \dots, N)$ .

The least squares method, initially conceived to study the movement of planets, is widely used and is basic to many estimation techniques. As we will see, its popularity is due to the fact that the choice of a quadratic parameter criterion leads to an explicit solution of the parameters we want to find. The principle of this method will now be discussed.

We try to minimize the mean-squares gap  $Q$  between the given coordinate points  $(x_i, y_i)$  ( $i = 1, \dots, N$ ) and the values  $(x_i, g_i(\theta))$  given by a model that depends on the parameters vector  $\theta = [\theta_1, \dots, \theta_p]^T$ . This gap is written:

$$Q = \sum_{i=1}^N |y_i - g_i(\theta)|^2 = \|y - g(\theta)\|^2$$

with  $y = [y_1, \dots, y_N]^T$ . To simplify the presentation, we assume that  $g(\theta)$  is real.

If the model is affine in its parameters  $g(\theta) = G\theta + c$ , the mean-squares gap is then quadratic in its parameters:  $Q = (y - G\theta - c)^T (y - G\theta - c)/2$ . The optimal value  $\hat{\theta}$  of  $\theta$ , which minimizes  $Q$ , is again said to be estimated in the least squares sense. It is obtained by writing that the gradient of  $Q$ , written  $\nabla_Q(\theta)$  in  $\theta = \hat{\theta}$  is zero. Now,  $\nabla_Q(\theta) = -G^T(y - G\theta - c)$ ; so, if the matrix  $G^T G$  is reversible, there is one solution:  $\hat{\theta} = (G^T G)^{-1} G^T (y - c)$ . If the model is not quadratic in its parameters,  $Q$  is not quadratic in its parameters. We then can come back to problem that can be processed by the least squares method by a first order Taylor approximation  $g(\theta)$  at  $\hat{\theta}_k$ :

$$g(\theta) \approx c + G\Delta\theta, \text{ with } c = g(\hat{\theta}_k), \quad G = \left. \frac{\partial g}{\partial \theta^T} \right|_{\hat{\theta}_k} \text{ and } \Delta\theta = (\theta - \hat{\theta}_k)$$

We then apply the previous results with this linearized function to obtain the new estimated value of  $\theta$ . After calculations, we get  $\hat{\theta}_{k+1} = \hat{\theta}_k + \Delta\hat{\theta}$  with  $\Delta\theta = (G^T G)^{-1} G^T (y - c)$ . This new recursive formula is also expressed according to the gradient  $\nabla_Q(\theta)$  of the criterion  $Q$  and of its approximate Hessian formula  $\tilde{H}_Q(\theta)$  (in this formula, the terms on which the second derivatives depend are ignored). So we have  $\nabla_Q(\hat{\theta}_k) = -G^T (y - c)$  and  $\tilde{H}_Q(\hat{\theta}_k) = G^T G$ ; the estimated value of  $\theta$  is updated according to the recursive formula  $\hat{\theta}_{k+1} = \hat{\theta}_k - \tilde{H}_Q^{-1}(\hat{\theta}_k) \nabla_Q(\hat{\theta}_k)$ . However, the decrease of the criterion  $Q$  is not guaranteed. This is why we introduce a coefficient  $\lambda_k$  that allows the algorithm converge, at least to a local minimum:  $\hat{\theta}_{k+1} = \hat{\theta}_k - \lambda_k \tilde{H}_Q^{-1}(\hat{\theta}_k) \nabla_Q(\hat{\theta}_k)$ . This is called a Gauss-Newton algorithm and can be obtained by a second order Taylor approximation of the criterion. There are methods that avoid the inverse calculation of the approximate Hessian formula. The algorithms are based on a first order Taylor approximation of the criterion, only using a gradient but converging more slowly. For more details on optimization methods, we direct the reader to [MIN 83] and [WAL 94].

### 10.5. Examples of digital processing

The goal of this section is to use several concepts discussed earlier to solve common signal processing problems that lead to finding an optimal filter. Even though signals can be complex, in the remainder of this section, in order to simplify formulae, we will suppose that they are real.

#### 10.5.1. Matched filtering

From observing a signal drowning in noise, we want to learn from linear filtering if this effective (noiseless) signal, whose waveshape is known, is present or absent. The matched filter that maximizes the signal-to-noise ratio at its outset at the time of our decision. This problem is found, for example, in radar applications when we need to determine the presence or absence of a target from measuring the received signal.

Let us suppose that  $s(t)$  is the effective signal of known waveshape that we assume is buried in an additive noise  $b(t)$ . A digital version of an matched filter is then introduced, and an analog filter can be used in a similar way. The impulse response of the filter is noted  $h[n]$  and its frequency response  $H(f)$ . The time we decide if the effective signal is present (P) or absent (A) is written as  $t_0 = n_0T_e$ , with  $T_e$  the sampling period. The process performed by the matched filter is shown in Figure 10.9.

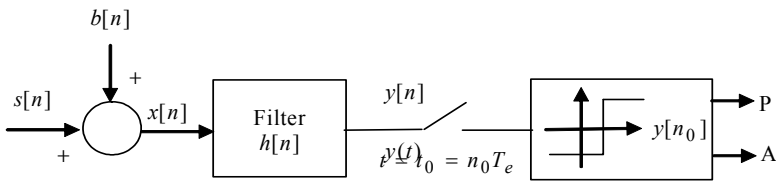


Figure 10.9. Detecting a signal buried in additive noise

We also assume that the noise  $b[n]$  is white, of a power  $\sigma_b^2$ . The mean power of the output noise of the filter is then  $P_{hb} = E[(h * b)[n]^2] = \sigma_b^2 \sum_k h[k]^2$ . The instantaneous power of the effective output signal of the filter is:

$$p_{hs}[n_o] = ((h * s)[n_o])^2 = \left( \sum_k h[k]s[n_o - k] \right)^2$$

and the signal-to noise ratio of the filter at the time of detection  $n_0 T_e$  is  $\text{RSB}[n_0] = p_{hs}[n_0]/P_{hb}$ . From the Schwartz inequality we get:

$$p_{hs}[n_0] \leq \left( \sum_k h[k]^2 \right) \left( \sum_k s[n_0 - k]^2 \right) = \left( \sum_k h[k]^2 \right) \left( \sum_k s[k]^2 \right)$$

the inequality being satisfied if and only if  $h[k] = c s[n_0 - k]$ , which leads to

$$\text{RSB}[n_0] \leq \frac{1}{\sigma_b^2} \sum_k s[k]^2 = \frac{\gamma_{ss}[0]}{\sigma_b^2} \text{ where } \gamma_{ss}[n] \text{ is the energy autocorrelation of}$$

the effective signal  $s[n]$ . The maximum output signal-to-noise ratio therefore only depends on the energy of the effective signal  $s[n]$  and on the noise power  $b[n]$ . The impulse response of the optimal filter ( $h[n] = c s[n_0 - n]$ ) is a reversed and shifted duplicate of the signal  $s[n]$ . This filter is said to be matched to the signal  $s[n]$ . Its response to the effective signal  $s[n]$  is  $y_s[n] = c \sum_k s[n_0 - k] s[n - k]$  and again, to a

close multiplicative coefficient, the energy autocorrelation of the effective signal:  $y_s[n] = c \gamma_{ss}[n - n_0]$ . The matched filter is a correlator. For example, the response of the matched filter to a rectangular impulse is a triangular signal. The choice of  $n_0$  depends on the application and on a causality constraint related to a real-time process. This causality constraint can be strictly fulfilled if the effective signal  $s[n]$  has a finite time structure.

### 10.5.2. Optimum filtering

Now we will look at the problem of estimating a signal  $x(t)$  by causal linear filtering from a signal  $y(t)$  that is correlated with  $x(t)$ . We consider here a digital filter. The signal  $y(t)$  is then sampled at the frequency  $F_e = 1/T_e$  and we want to know the estimation of  $x[n] = x(nT_e)$  according to samples  $y[k] (k \leq n)$ . We write  $\hat{x}[n]$  as the estimated value of  $x[n]$ .

#### 10.5.2.1. Wiener filtering

The Wiener filter is a causal linear filter that minimizes the mean square error  $P[n] = E[(x[n] - \hat{x}[n])^2]$ . We first adjust the causal filter so that it is at the finite impulse response of length  $M$ . We write  $h[n] (n = 0, \dots, M-1)$  the non-zero coefficients of the impulse response. We then look for  $\hat{x}[n]$  in the form

$$\hat{x}[n] = \sum_{k=0}^{M-1} h[k] y[n - k]. \text{ The optimal filter is the one whose impulse response}$$

minimizes  $P[n]$ ; so it verifies the relations:



$$\frac{\partial P[n]}{\partial h[k]} = -2 E[y[n-k](x[n] - \hat{x}[n])] = 0, \quad \forall k = 0, \dots, M-1$$

If the signals are stationary, these relations are rewritten according to the correlation functions of the signals  $x[n]$  and  $y[n]$ :  $\gamma_{xy}[k] - \sum_{i=0}^{M-1} h[i]\gamma_{yy}[k-i] = 0, \forall k = 0, \dots, M-1$ . The causal FIR optimal filter thus satisfies the equations:

$$\gamma_{xy}[k] = (h * \gamma_{yy})[k] \quad \forall k = 0, \dots, M-1 \tag{10.9}$$

and it is necessary to know the correlation functions.

We can show that this relation can also apply to a causal filter that is not necessarily of finite impulse response:

$$\gamma_{xy}[k] = \sum_{i \leq 0} h[i]\gamma_{yy}[k-i] = (h * \gamma_{yy})[k] = 0 \quad \forall k \leq 0 \tag{10.10}$$

and that it is also validated with an analog filter [LIF 81]; [MAX 89]; [PIN 95]. Equation [10.10] is called the Wiener-Hopf equation.

An optimal linear filter without constraints satisfies the relation  $(h * \gamma_{yy})[k] = \gamma_{xy}[k] \quad \forall k$ . In taking the Fourier transform of this relation, we deduce that the frequency response of the optimal filter (digital or analog) without constraints is then expressed according to the spectral densities of the signals:  $H(f) = \Phi_{xy}(f) / \Phi_{yy}(f)$ .

In the specific case where  $y(t) = (h_0 * x)(t) + b(t)$  with  $b(t)$ , which is a white noise independent of  $x(t)$ , we get:

$$H(f) = \frac{H_0^*(f)\Phi_{xx}(f)}{|H_0(f)|^2 \Phi_{xx}(f) + \Phi_{ww}(f)} = \frac{1}{H_0(f)} \frac{1}{1 + \Phi_{ww}(f) / \Phi_{xx}(f)}$$

which means that  $1/H_0(f)$  is weighted by a coefficient which is all the closer to 1 since the noise is low.

The causality constraint leads to a mean square error above that obtained with an optimal filter without constraints. The resolution, shown in equation [10.10] (or its equivalent with an analog filter), leads to the Wiener filter; and this is fairly complicated [LIF 81]; [PIN 95]. This is why often we prefer using a causal linear filter with finite impulse response, in order to solve equation [10.9].

### 10.5.2.2. Matched filtering

To obtain the best estimation possible of the signal  $x[n]$ , we take into account all previous available observations  $y[1], \dots, y[n]$  ( $y[1]$  being the first sample). We thus choose for  $M$  the highest value:  $M = n$ . The length of the impulse response of the filter depends on the time and is equal to  $n$ . The filter thus defined is called adaptive because its impulse response is time variant. To further explain this filter, we propose:  $\mathbf{y}_n = [y[n] \cdots y[1]]^T$ ,  $\mathbf{h}_n = [h[0] \cdots h[n-1]]^T$ ,  $\boldsymbol{\gamma}_n = [\gamma_{xy}[0] \cdots \gamma_{xy}[n-1]]^T$  and  $\Gamma_n$  the squared and symmetrical Toeplitz matrix of  $n$  dimension where the element of the  $i^{\text{e}}$  line and  $j^{\text{e}}$  column is  $\Gamma_{ij} = \gamma_{yy}[i-j]$ . The relation in equation [10.9] with  $M = n$ , is then written  $\Gamma_n \mathbf{h}_n = \boldsymbol{\gamma}_n$  and this leads to  $\mathbf{h}_n = \Gamma_n^{-1} \boldsymbol{\gamma}_n$  and then  $\hat{x}[n] = \boldsymbol{\gamma}_n^T \Gamma_n^{-1} \mathbf{y}_n$ . The disadvantage of this formula is that it requires inverted a matrix of dimension  $n$  that increases over time. The next section shows how to avoid calculating the inverse of this matrix.

### 10.5.2.3. Kalman filtering

If the signals are described by an internal representation, the Kalman filter allows us, as we will see, to obtain a recursive formula for the estimation problem seen above. This algorithm avoids inverting a matrix with dimensions that increase over time; also, it is no longer necessary to know the correlation of signals and this applies to time variant systems.

To introduce the Kalman filter, we still keep in mind the previous problem, but add that the system is described by state space representation:

$$\begin{cases} x[n+1] = ax[n] + v[n], \\ y[n] = cx[n] + w[n], \end{cases}$$

the noises  $v[n]$  and  $w[n]$  being white and uncorrelated to each other and to the initial state  $x[0]$ . We respectively write  $R_v$  and  $R_w$  as the covariances of  $v[n]$  and  $w[n]$ . The signal  $x[n]$  is called autoregressive of order one. We can generalize this to a filter of a higher order by introducing a state vector of equal dimension to the order of the filter.

The best causal linear estimator (in the sense of mean square error) of  $x[n]$  according to all available observations  $y[1], \dots, y[n]$  is now written as  $\hat{x}[n/n]$  and its quadratic error  $P[n/n] = E[(x[n] - \hat{x}[n/n])^2]$ . From what has been previously shown, we now have:  $\hat{x}[n/n] = \mathbf{h}_{n/n}^T \mathbf{y}_n$  with  $\mathbf{h}_{n/n} = \Gamma_n^{-1} \boldsymbol{\gamma}_n$ .

We now introduce a new observation  $y[n+1]$  and look for the estimator. We will establish that  $\hat{x}[n+1/n+1]$  can be obtained in a recursive manner in two steps. For that, we write  $\hat{x}[n+1/n]$  as the best linear estimator of  $x[n+1]$  according to all previous observations  $y[1], \dots, y[n]$ :  $\hat{x}[n+1/n] = \mathbf{h}_{n+1/n}^T \mathbf{y}_n$ . This quantity is called a

one-step prediction. To establish the equations of the predictive filter, we write that  $\mathbf{h}_{n+1/n}$  minimizes  $P[n+1/n] = E[(x[n+1] - \hat{x}[n+1/n])^2]$ , which leads to:

$$\frac{\partial P[n+1/n]}{\partial \mathbf{h}_{n+1/n}} = E[\mathbf{y}_n(ax[n] + v[n] - \mathbf{h}_{n+1/n}^T \mathbf{y}_n)] = \mathbf{0}$$

Taking into account the statistical properties of noises, this relation is rewritten  $\Gamma_n \mathbf{h}_{n+1/n} = \alpha \gamma_n$ . We deduce from this that  $\mathbf{h}_{n+1/n} = \alpha \mathbf{h}_{n/n}$ , then  $\hat{x}[n+1/n] = \alpha \hat{x}[n]$ . By expressing  $x[n+1]$  according to  $x[n]$  in the expression  $P[n+1/n]$  and in considering the statistical properties of  $v[n]$ , we deduce that  $P[n+1/n] = a^2 P[n/n] + R_v$ .

Next, in order to produce the recurrence relation, we decompose  $\hat{x}[n+1/n+1]$  in two terms, the first corresponding to the past and the second corresponding to the present:  $\hat{x}[n+1/n+1] = \beta_n^T \mathbf{y}_n + k_{n+1} y[n+1]$ . We express  $x[n+1]$  and  $y[n+1]$  according to  $x[n]$  in the formula  $P[n+1/n+1]$ , then write that the optimal estimator corresponds to the cancellation of the gradient of this quantity with respect to  $\beta_n$ . We deduce from this that  $\beta_n = a(1 - k_{n+1}c) \Gamma_n^{-1} \gamma_n$  which leads to  $\hat{x}[n+1/n+1] = \hat{x}[n+1/n] + k_{n+1}(y[n+1] - \hat{y}[n+1/n])$  where  $\hat{y}[n+1/n] = c \hat{x}[n+1/n]$  is the one-step prediction of  $y[n+1]$ . So we correct  $\hat{x}[n+1/n]$  with a term that depends on the prediction error of the measurement. We then write that the derivative of  $P[n+1/n+1]$  with respect to  $k_{n+1}$  must also be zero. After calculations, we get:

$$\begin{cases} k_{n+1} = c P[n+1/n] (c^2 P[n+1/n] + R_w)^{-1}, \\ P[n+1/n+1] = (1 - k_{n+1}c) P[n+1/n]. \end{cases}$$

These relations no longer require a matrix inversion or signal correlation.

By following a similar line of reasoning, we show that these equations can extend to vectors and a time variant linear system. The equations of the system are then as follows:

$$\begin{cases} \mathbf{x}[n+1] = \mathbf{A}_n \mathbf{x}[n] + \mathbf{B}_n \mathbf{u}[n] + \mathbf{v}[n] \\ \mathbf{y}[n] = \mathbf{C}_n \mathbf{x}[n] + \mathbf{w}[n] \end{cases}$$

with  $\mathbf{x}[n] \in \mathfrak{R}^p$ ,  $\mathbf{y}[n] \in \mathfrak{R}^q$ , and  $\mathbf{A}_n, \mathbf{B}_n, \mathbf{C}_n$  the known matrixes,  $\mathbf{u}[n]$  a known input,  $\mathbf{v}[n]$  and  $\mathbf{w}[n]$  the white noises of covariance  $\mathbf{R}_v[n]$  and  $\mathbf{R}_w[n]$  known. We also assume that the noises are uncorrelated to each other and to the initial state  $\mathbf{x}[0]$ . By applying what we see with the matched filter and adding a term when looking for

an optimal estimator to take into account  $\mathbf{u}[n]$  (we are then looking for  $\hat{\mathbf{x}}[n/n]$  in the form  $\hat{\mathbf{x}}[n/n] = \mathbf{h}_{n/n}^T \mathbf{y}_n + \alpha_n$ ), we obtain the recurrence relations that are those of the Kalman filter. Finding the estimator is carried out in two steps. The first prediction step leads to:

$$\begin{cases} \hat{\mathbf{x}}[n+1/n] = \mathbf{A}_n \hat{\mathbf{x}}[n/n] + \mathbf{B}_n \mathbf{u}[n] \\ \mathbf{P}[n+1/n] = \mathbf{A}_n \mathbf{P}[n/n] \mathbf{A}_n^T + \mathbf{R}_v[n] \end{cases}$$

and the correction step to:

$$\begin{cases} \hat{\mathbf{x}}[n+1/n+1] = \hat{\mathbf{x}}[n+1/n] + \mathbf{K}_{n+1} (\mathbf{y}[n] - \mathbf{C}_{n+1} \hat{\mathbf{x}}[n+1/n]), \\ \mathbf{K}_{n+1} = \mathbf{P}[n+1/n] \mathbf{C}_{n+1}^T (\mathbf{C}_{n+1} \mathbf{P}[n+1/n] \mathbf{C}_{n+1}^T + \mathbf{R}_w[n+1])^{-1}, \\ \mathbf{P}[n+1/n+1] = (\mathbf{I} - \mathbf{K}_{n+1} \mathbf{C}_{n+1}) \mathbf{P}[n+1/n]. \end{cases}$$

This algorithm must be initialized. If we have the first and second order of the initial state  $\mathbf{x}[0]$ , we can choose  $\hat{\mathbf{x}}[0/0] = \mathbf{E}[\mathbf{x}[0]]$  for initialization and  $\mathbf{P}[0/0] = \mathbf{E}[(\mathbf{x}[0] - \hat{\mathbf{x}}[0])(\mathbf{x}[0] - \hat{\mathbf{x}}[0])^T]$ ; we then show that this estimator is non-biased. Otherwise, we can choose  $\hat{\mathbf{x}}[0/0] = 0$  and  $\mathbf{P}[0/0] = \alpha \mathbf{I}$  with  $\alpha$  being rather high. There are other ways (least squares, Bayesian approach) to obtain these equations and we direct the reader to [AND 79], [JAZ 70] and [SCH 91]. We note here that if the state noises and observations are Gaussian, the obtained estimator is the best estimator without imposing a linearity constraint. The Kalman filter is relatively robust and is being used more and more frequently. For non-linear systems, the previous equations can be applied after a first order approximation of the system around the current estimated state.

## 10.6. Frequency, time, time-frequency and wavelet analyses

As we have seen before, a signal coming from a sensor can be represented in several ways. There are actually almost an infinite number of representations. The best-known and most natural is time representation  $x(t)$ . Another representation that is currently being used is frequency representation, written  $X(\omega)$ , the passage of one representation to another being ensured by the Fourier transform given in equations [10.2] and [10.3]. The particular benefit of frequency representation is linked not only to the relevance of its content but also to its properties for certain operations as the convolution product or cross-correlation products. The time representation of a signal directly indicates the variation in the time of the amplitude, while the

frequency representation demonstrates the frequency at which these variations have taken place. Even if these two representations are equivalent, since for the physical signals we assume the Fourier transform always exists and is perfectly reversible, the nature of the information that is directly accessible differs radically. The essential feature of these two representations is their global nature, which is expressed by the infinite length of the integration field:  $x(t)$  integrates all the time information while  $X(\omega)$  contains all frequency information.

Generally, to make the presentations of different transforms uniform, we assume that a representation is obtained by projecting a time signal on an analysis signal that somehow is a *basic tile* or atom of the time-frequency plane:

$$Tx(t, \omega) = \langle x, \psi_{t, \omega} \rangle = \int_{-\infty}^{+\infty} x(s) \psi_{t, \omega}^*(s) \, ds$$

This projection operation is linear, so that projecting one linear combination of signals is the linear combination of projections.

We can define the orders of magnitude of the dimensions of this atom from the variance of the energy of the analysis function and from its Fourier transform so that, for an analysis function of normed energy, we have  $\sigma_t^2 = \int_{-\infty}^{+\infty} s^2 |\psi_{t, \omega}(s)|^2 \, ds$  and  $\sigma_\omega^2 = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \theta^2 |\Psi_{t, \omega}(\theta)|^2 \, d\theta$ . These variables are shown in Figure 10.10.

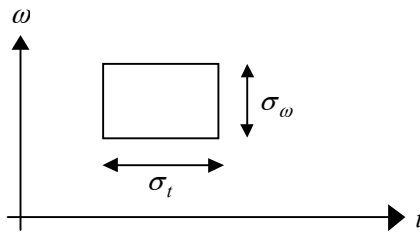


Figure 10.10. Tiling of the time-frequency plane

It is easy to demonstrate that the surface of the atom (for an analysis function normalized to 1) cannot be below a limit given by the Heisenberg-Gabor inequality:  $\sigma_t \sigma_\omega \geq 1/2$ . This result fixes the resolution limits of the time-frequency representations. It is important to note that this limit is attained for an analysis function of the Gaussian model:

$$|\psi_{t, \omega}(x)| = \pi^{-1/4} \exp(-(x - t)^2 / 2)$$

For the Fourier transform (frequency representation), the analysis function only depends on the pulsation and the atom is a band of infinite length and of zero height parallel to the time axis:  $\psi_{\omega}(s) = \exp(j\omega s)$ ,  $\Psi_{\omega}(\theta) = \delta(\omega - \theta)$ ,  $\sigma_t = \infty$  and  $\sigma_{\omega} = 0$ . In addition, we can use the same expression for the time representation:  $\psi_{t_s}(s) = \delta(t - s)$  and  $\Psi_{t_s}(\theta) = \exp(-j\theta)$ . The atom being analyzed in this case is a band of zero thickness parallel to the frequency axis  $\sigma_t = 0$  and  $\sigma_{\omega} = \infty$ .

In general, the atom of analysis contains the essential of the representation energy. It therefore represents the time-frequency resolution of the energy. Time and frequency representations correspond to specific *degenerated* examples of time-frequency representation. Most processed signals processed in instrumentation are generally not the most appropriate. A true time-frequency representation is necessary. These methods are currently done from the standard toolbox of a signal processor. In the following sections we present the tools that help us obtain the most current time-frequency representations: the sliding window Fourier or short-term transform; the wavelet transfer; and bilinear transforms (in particular, the group of Wigner-Ville transforms). We will see that the wavelet transform leads more often to a time-scale representation than to a time-frequency representation. Our presentation will be schematic; the reader wanting a deeper knowledge of these questions may consult the works of Patrick Flandrin [FLA 93] and Stéphane Mallat [MAL 99].

These various representations can also be used with digital signals and a certain number of efficient processing algorithms have been developed. In this regard we emphasize the importance of the fast Fourier transform (FFT) and the discrete wavelet transform.

### 10.6.1. Frequency analysis

Frequency analysis of a signal is given by its frequency representation. This can be obtained, under a set of conditions specific to signals processed in an instrumentation chain (see the beginning of this chapter), by the Fourier transform or by one of its variants such as the cosine transform or Hartley transform (these last two are real).

#### 10.6.1.1. Continuous transforms

The most important of these transforms is the Fourier transform whose definition and main properties were discussed in section 10.2.2. There are many variants of the

Fourier transform. Among them, the Hartley transform stays within the domain of real functions, since its nucleus is real:  $\psi_{\omega}(s) = \cos \omega s + \sin \omega s$ . It is defined by:

$$H(\omega) = \int_{-\infty}^{+\infty} x(t) (\cos \omega t + \sin \omega t) dt$$

The inversion is obtained the same way because the analysis function is real and forms an orthogonal base:

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} H(\omega) (\cos \omega t + \sin \omega t) d\omega$$

The Hartley transform can easily be linked to the Fourier transform by the intermediary of its even part  $H_p(\omega) = (H(\omega) + H(-\omega))/2$  and of its odd part  $H_i(\omega) = (H(\omega) - H(-\omega))/2$  which are respectively the cosine transform and the sinus transform: so we have  $X(\omega) = H_p(\omega) - jH_i(\omega)$ . The Hartley transform is usually deduced from the Fourier transform, its real part providing the even part, and its imaginary part, the odd part. Although this transform is not much used today, it can be advantageous because the analysis is real and the inversion formula is identical to direct transform. The cosine transform, seldom used as a frequency analysis tool, has a particular importance in another context; its discrete version is a good approximation of the Karhunen-Loève transform [BEL 87].

### 10.6.1.2. Discrete Fourier transform

In practice, we often calculate a discrete transform rather than a continuous transform, either because the signal to be analyzed is sampled, or because computer-based methods seem more practical than usual analog methods. In both cases, it is useful to examine the consequences of this technique for interpreting results in terms of frequency representation of the signal (see [DUV 91], [KUN 84] and [PRO 92]).

Calculating a discrete Fourier transform (DFT) means not only sampling (of period  $T_e$ ) but also windowing the signal. Windowing is a constraint due to the material impossibility of carrying out an infinite number of calculations. A sampling in the frequency domain becomes necessary.

Signal windowing is imposed by the choice of the transform length, that is, by the number  $N$  of samples retained for the calculation. The width of the window is then  $T = NT_e$ . The truncated signal is written  $x_f(t) = x(t) \text{rect}_T(t)$  where  $\text{rect}_T(t)$  is the carried function equal to 1 on the observation horizon and to 0 elsewhere. The

Fourier transform of the truncated signal becomes  $X_f(\omega) = \frac{1}{2\pi}(X * \text{Rect}_T)(\omega)$  with, if the observation horizon is centered at the origin in time,  $\text{Rect}_T(\omega) = T \text{sinc}(\omega T/2)$ . The frequency representation  $X(f)$  of the analog signal is thus convoluted with a cardinal sinus.

In such a fashion that the transform is reversible and non-redundant, the sampling in the frequency domain should retain  $N$  samples per period. The positions of the samples also correspond to the zeros of the cardinal sinus function. Under these conditions, the discrete Fourier transform of a discrete signal of length  $N$  is a discrete signal of length  $N$  that is written:

$$X[n] = \sum_{k=0}^{N-1} x[k] \exp(-j2\pi k n/N) \quad [10.11]$$

and we have  $X[n] = X_{f\text{TD}}(n/N)$ . The reversed transform is expressed by:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \exp(j2\pi k n/N)$$

Applying an inversion formula leads to a periodic signal  $x[n]$  of period  $N$  identical to the initial signal of a period. We see from what follows that  $x[n]$  and its transform are periodic. This transform is also used as a frequency analysis tool for digital signals. The properties of the discrete Fourier transform are very close to those of the discrete time transform. Plancherel's theorem is applied by considering the circular convolution  $z[n]$  that is only represented for the periodic groups  $x[n]$  and  $y[n]$  or periodized with the same period  $N$ :

$$z[n] = \sum_{k=0}^{N-1} x[k] y[n-k] = x \tilde{*} y[n]$$

and  $Z[n] = X[n] Y[n]$ . The Fourier transform of the circular correlation:

$$\gamma_{xy}[n] = \sum_{k=0}^{N-1} x[k] y^*[k-n]$$

is  $R_{xy}[n] = X[n] Y^*[n]$ . Plancherel's theorem becomes:

$$E_x = \sum_{n=0}^{N-1} |x[n]|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |X[n]|^2$$



10.6.1.3. Algorithm of the fast Fourier transform

Calculating the DFT ([BEL 87]; [KUN 84] and [TRU 97b]) by the direct application of equation [10.11] requires  $N^2$  complex multiplications and  $N(N-1)$  complex additions. This is a complex algorithm  $O(N^2)$ . In 1965, Cooley and Tukey developed a method that lead to a series of rapid algorithms that could be adapted to a large number of unitary transforms, such as Hadamard, Haar, Fourier, cosine and Hartley. These algorithms, which are of complexity  $O(N \log N)$ , have the generic name of FFT. These are recursive algorithms based on the implementation of a basic transform on two data, so that these algorithms have maximum efficiency if the total number of data are of a power of  $N = 2^m$ .

The basic operation, called the butterfly operation, is simply a DFT on two samples, one of even range and the other of odd range:

$$\begin{cases} X[n] = X_1^p[n] + \exp(-j 2\pi m/N)X_1^i[n] \\ X[n + N/2] = X_1^p[n] - \exp(-j 2\pi m/N)X_1^i[n] \end{cases}$$

For a complete transform, there are  $mN/2$  butterfly operations. Since  $m = \log_2 N$ , the total number of operations is  $(N/2)\log_2 N$  complex multiplications and  $N \log_2 N$  complex additions. We estimate the gain  $G$  of the fast algorithm in relation to the direct calculation by only considering that the most complex operation is the complex multiplication that represents four real multiplications and two real additions. We then have  $G = 2N/\log_2 N$  (Figure 10.11).

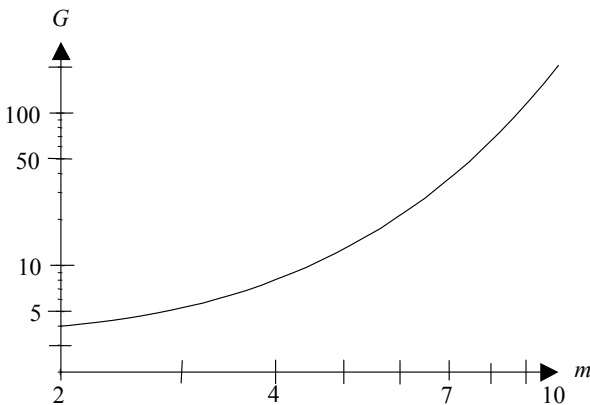


Figure 10.11. FFT gain according to m

## 10.6.2. Sliding window or short-term Fourier transform

### 10.6.2.1. Continuous sliding window Fourier transform

As we saw in the introduction to this section, the representation given by the Fourier transform has a zero time resolution. The simplest way to improve the precision of the analysis is to proceed to a windowing of the function to be analyzed before carrying out its frequency analysis. This is what occurs during the sliding window Fourier transform (SWFT), otherwise called the short-term Fourier transform [FLA 93]. If the windowing function is written  $g(t)$ , the analysis function of the representation becomes  $\psi_{t,\omega}(s) = g(s - t) \exp(j\omega s)$  and:

$$T_{fg}x(t, \omega) = \langle x, \psi_{t,\omega} \rangle = \int_{-\infty}^{+\infty} x(s)g^*(s - t) \exp(-j\omega s) ds$$

This transform is reversible if the resolution relation of identity is verified, which, in this case, is equivalent to a normality condition of the energy of the envelope function:  $\int_{-\infty}^{+\infty} |g(t)|^2 dt = 1$ . In these conditions, the reconstruction is ensured by:

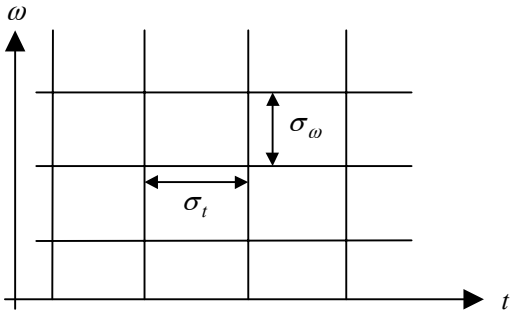
$$\begin{aligned} x(t) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_{fg}x(s, \omega) \psi_{s,\omega}(t) ds d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_{fg}x(s, \omega) g(s - t) \exp(-j\omega t) ds d\omega \end{aligned}$$

We see that the reconstruction can, in a more general way, be conducted with the help of a function  $\psi'$  satisfying  $\int_{-\infty}^{+\infty} \psi(t) \psi'^*(t) dt = 1$ . This representation conserves the energy; and we have a theorem analogous to that of Parseval's:

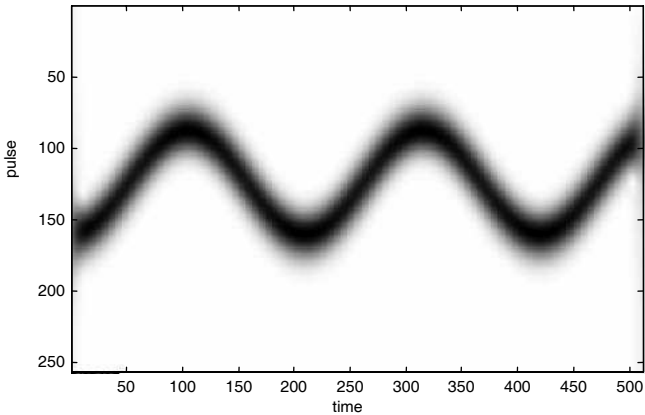
$$E_x = \int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_{fg}x(t, \omega) T_{fg}^*x(t, \omega) dt d\omega$$

Each atom of the time-frequency representation is defined following the method shown in the introduction. The tiling of the time-frequency plane (see Figure 10.12) thus obtained is regular, the dimension of the tiles is constant at all points of the plane. Intuitively it is clear that, in these conditions, the resolution will not be optimum. In fact, an atom in a low-frequency domain will contain only a low number of time oscillations and will thus be badly estimated, while an atom in a high-frequency domain contains many more time oscillations than is necessary for a good estimation. This leads to a sub-optimal time localization in this situation.

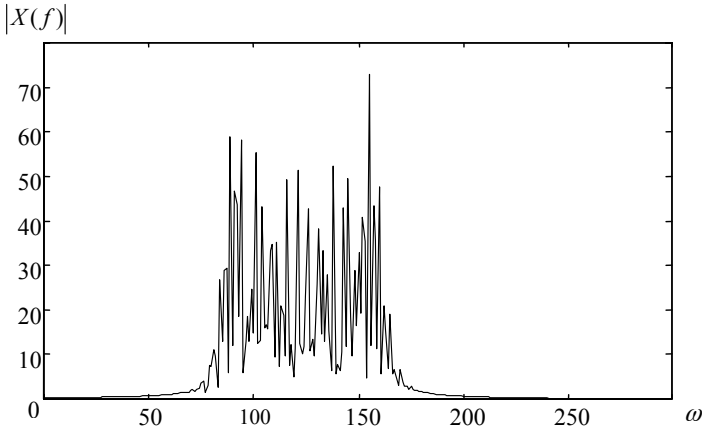
We have seen the importance of choosing a Gaussian window that helps in obtaining a minimal surface for the basic tile ( $\sigma_t \sigma_\omega = 1/2$ ). The transform obtained in this case is sometimes called a Gabor transform and the analysis functions are called gaborlets:  $\psi_{t,\omega}(s) = \pi^{-1/4} \exp(-(s - t)^2/2) \exp(j\omega s)$ . Figure 10.13 shows the time-frequency representation given by the Gabor transform with a phase-modulated signal. In this model, the intensity represents the module of the SWFT. The modulating sinusoidal signal is clearly identifiable, which is not the case when the frequency representation of this signal is that given by the Fourier transform (see Figure 10.14).



**Figure 10.12.** Tiling of the time-frequency plane by SWFT



**Figure 10.13.** Time-frequency representation of a phase modulated signal



**Figure 10.14.** Fourier transform of a phase modulated signal

### 10.6.2.2. Discrete sliding window Fourier transform

The representation given by the SWFT is continuous and the tiling of the time-frequency plane is obviously very redundant. A time sampling (of step  $t_0$ ) and frequency (of step  $\omega_0$ ) limits this redundancy. However, a discrete base cannot be constructed and an exact reconstruction is impossible.

$$x(t) = \sum_n \sum_m T_{fg} x[n, m] \psi_{nm}(t) \text{ with } \psi_{n,m}(s) = \exp(-(s/t_0 - n)^2/2) \exp(jm\omega_0 s) / (\pi\omega_0^2)^{1/4}$$

but the analysis does not allow us to find the coefficients of this discrete representation:  $T_{fg} x[n, m] \neq \int_{-\infty}^{+\infty} x(t) \psi_{n,m}^*(t) dt$ .

### 10.6.3. Wavelet transforms

A tiling of the time-frequency plane at a constant overvoltage can bring about an optimum resolution of the representation throughout the plane. This idea led to the development of the wavelet transform (see [DAU 92], [GAS 90], [MAL 99] and [TRU 97a]). The analysis function has a time area inversely proportional to its spectral area. We easily obtain this result by dilating and translating a mother function:  $\psi_{a,b}(t) = \psi((t-b)/a) / \sqrt{a}$  or  $\Psi_{a,b}(\omega) = \sqrt{a} \Psi(a\omega) \exp(-j\omega b)$ . The variable  $a$ , which has the dimension of the inverse of a frequency is called a *scale* variable, and  $b$  is the translation factor. The term in  $1/\sqrt{a}$  conserves the energy of the analysis function despite variation of scale.

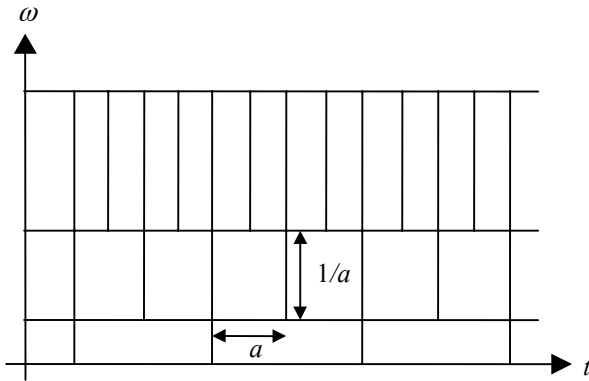
If  $\Delta t$  and  $\Delta\omega$  are the dimensions of the time-frequency atom of the representation of the scale  $a = 1$ , it is easy to show that for whatever scale  $a$ , these become

$\sigma_t = a\Delta t$  and  $\sigma_\omega = \Delta\omega/a$ . This means that the surface of the atom being analyzed remains constant throughout the scales but that the tiling of the plane respects the conditions of an optimum analysis: a time spread that is inversely proportional to the frequency. This is called a time-scale representation.

10.6.3.1. *Continuous wavelet transforms*

The continuous wavelet transform (CWT) of the signal  $x(t)$  is written:

$$T_{oc}x(a, b) = \langle x, \psi_{a,b} \rangle = \int_{-\infty}^{+\infty} x(t) \frac{1}{\sqrt{a}} \psi^* \left( \frac{t-b}{a} \right) dt$$



**Figure 10.15.** Time-frequency tiling for the wavelet transform

and the admissibility condition or identity resolution:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \psi_{ab}(t) \psi_{ab}^*(t) \frac{da db}{a^2} = \delta(t - t')$$

which, after the Fourier transform, becomes:

$$\int_{-\infty}^{+\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega = 1$$

This condition is not very restrictive and a function of normalized energy and of zero mean ( $\Psi(0) = 0$ ) suitably located around the source will generally be

admissible. We see that, as was the case with SWFT, reconstruction is possible with another function  $\psi'(t)$  if it verifies the condition:

$$\int_{-\infty}^{+\infty} \frac{\Psi(\omega) \Psi'^*(\omega)}{|\omega|} d\omega = 1$$

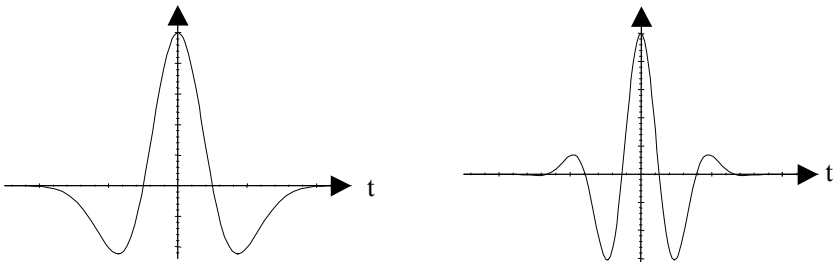
If this admissibility condition is met by the function analysis, the transform is reversible using the following formula:

$$x(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} T_{oc} x(a, b) \psi_{a,b}(t) \frac{da db}{a^2}$$

The time-scale representation conserves the energy:

$$E_x = \int_{-\infty}^{+\infty} |x(t)|^2 dt = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |T_{oc} x(a, b)|^2 \frac{da db}{a^2}$$

We can cite two examples of functional analyses used for the continuous wavelet transform. One is called the *Mexican hat*:  $\psi(t) = 2(1-t^2) \exp(-t^2/2) / (\sqrt{3}\pi^{1/4})$ . This is simply a second derivation of the Gaussian and the Morlet wavelet formula  $\psi(t) = \exp(-t^2/2) \exp(j\omega_0 t) / \pi^{1/4}$  which itself came from gaborets; however, gaborets are not, strictly speaking, admissible (see Figure 10.16).



**Figure 10.16.** Mexican hat and Morlet wavelet (real part)

Practically, this transform is never calculated analytically and must be estimated digitally. In all cases this means it is sampled. This being the case, the analysis is continuous and thus redundant. However, contrary to the situation of discrete SWFT, the discrete wavelet transform allows for an exact reconstruction; the bases for this do exist.

10.6.3.2. Discrete wavelet transforms

Discretizations in time and frequency are often done by a discrete dyadic analysis that aids in constructing orthogonal bases. In these situations, the relation between two successive scales is 2 and the translation step will be arbitrarily unitary with the scale  $a = 1$   $a = 2^i$  and  $b = n2^i$ . From this, we have the discrete wavelet transform (DWT):

$$T_{od}x(i, n) = \langle x, \psi_{i,n} \rangle = \int_{-\infty}^{+\infty} x(t) 2^{-i/2} \psi^*(2^{-i}x - n) dx$$

Y. Meyer [MEY 90] and S. Mallat [MAL 99] propose an especially efficient calculation algorithm from this transform and its inverse in the general field of multiresolution analysis (MRA). We will present a very general overview.

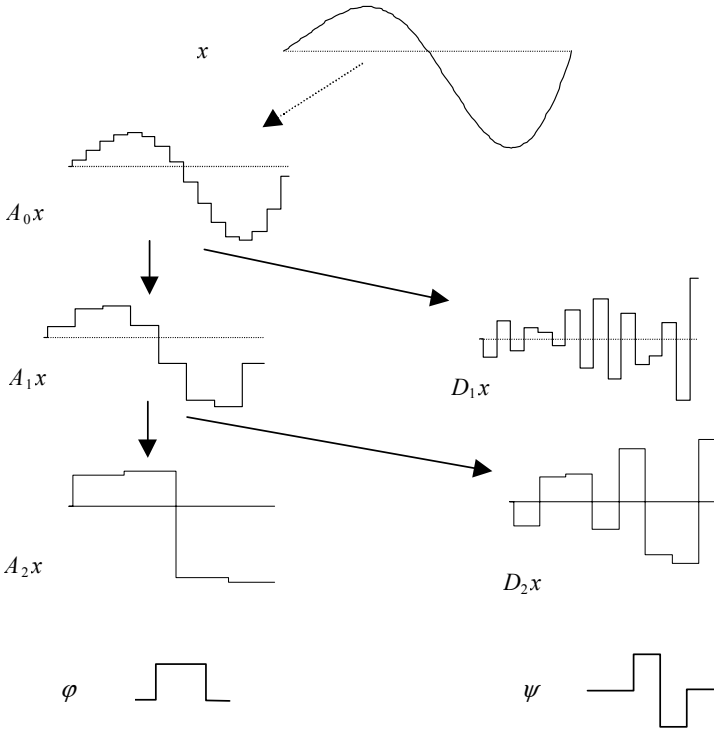
A multiresolution analysis of a function  $x(t)$  of  $L^2(R)$  is made of projections  $A_i x(t)$  of this function by a series of infinite spaces of approximations  $V_i$  enclosed in each other and filling  $L^2(R)$ . These subspaces are constructed by a simple dilation of a subspace  $V_o$  that is invariant by translation throughout the entire step. Each subspace  $V_i$  is completed in its immediately higher container by an orthogonal subspace  $W_i$  so that the approximation  $A_{i-1}x(t)$  can be expressed according to the immediately larger approximation of the projection  $D_i x(t)$  of the function on  $W_i$ :  $A_{i-1}x(t) = A_i x(t) + D_i x(t)$ . The following properties must be verified:

$$\left\{ \begin{array}{l} \dots \subset V_1 \subset V_0 \subset \dots \subset V_i \subset V_{i-1} \subset \dots, \\ \overline{\bigcup_{i \in Z} V_i} = L^2(R), \quad \bigcap_{i \in Z} V_i = \{0\}, \quad V_{i-1} = V_i \oplus W_i, \quad V_i \perp W_i, \quad \forall i \in Z, \\ x(t) \in V_i \Leftrightarrow x(2t) \in V_{i-1}, \quad \forall i \in Z, \quad x(t) \in V_0 \Leftrightarrow x(t-k) \in V_0, \quad \forall k \in Z. \end{array} \right.$$

Orthonormed bases of the subspaces  $V_o$  and  $W_o$  are made of two families of functions obtained by translating the integer step of two functions  $\varphi(t)$  and  $\psi(t)$  called, respectively, the scale function and the wavelet function. The bases of other subspaces are made by dilation of the mother functions.

The  $\varphi_{i,n}(t) = 2^{-i/2} \varphi(2^i t - n)$  with n integer form an orthonormal basis of  $V_i$ : we see that these functions are not admissible wavelets.  $\psi_{i,n}(t) = 2^{-i/2} \psi(2^i t - n)$  with an integer forms an orthonormal basis of  $W_i$ ; we see that these functions are admissible wavelets.

All  $W_i$  spaces are, by construction, 2 by 2 orthogonal, the direct sum of all these subspaces is equal to  $L^2(R)$ . This means the ensemble of  $\psi_{i,n}$  for  $i$  and  $n$  integers forms an orthonormal basis of  $L^2(R)$ . We therefore have a discrete wavelet basis and coefficients  $d_n^i$  of the projection of  $x(t)$  on the subspaces  $W_i$  constituting the discrete wavelet transform of  $x(t)$ :  $A_i x = \sum_n \langle x, \varphi_{i,n} \rangle \varphi_{i,n}$ ,  $D_i x = \sum_n \langle x, \psi_{i,n} \rangle \psi_{i,n}$ ,  $a_n^i = \langle x, \varphi_{i,n} \rangle$  and  $d_n^i = \langle x, \psi_{i,n} \rangle = T_{od} x(i, n)$ .

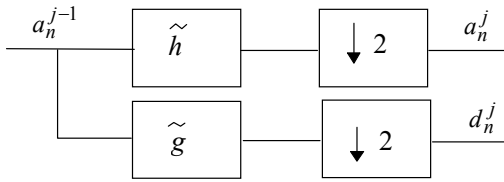


**Figure 10.17.** Schema for multiresolution analysis

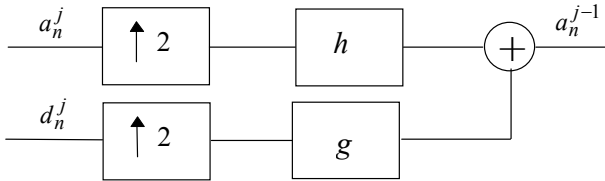
The calculation algorithm of these coefficients, proposed by Mallat, is recursive. It makes use of linear digital filtering operations. We define two filters by their impulse responses  $h[n]$  and  $g[n]$ :  $h[n] = \langle \varphi_{0,n}, \varphi_{-1,n} \rangle$  and  $g[n] = \langle \psi_{0,n}, \varphi_{-1,n} \rangle$ . Taking into account the subspaces and their bases, these two filters form a pair of quadratic mirror filters:  $G(z) = -z^{-1}H(-z^{-1})$  and  $g[n] = (-1)^n h[1-n]$ . By representing these returned filters  $\tilde{h}[n] = h[-n]$  and  $\tilde{g}[n] = g[-n]$ , we show that the



algorithms of analysis and reconstruction use the filtering operations and processes of over-and under-sampling presented in Figures 10.18 and 10.19.



**Figure 10.18.** Recursive algorithm of a multiresolution analysis



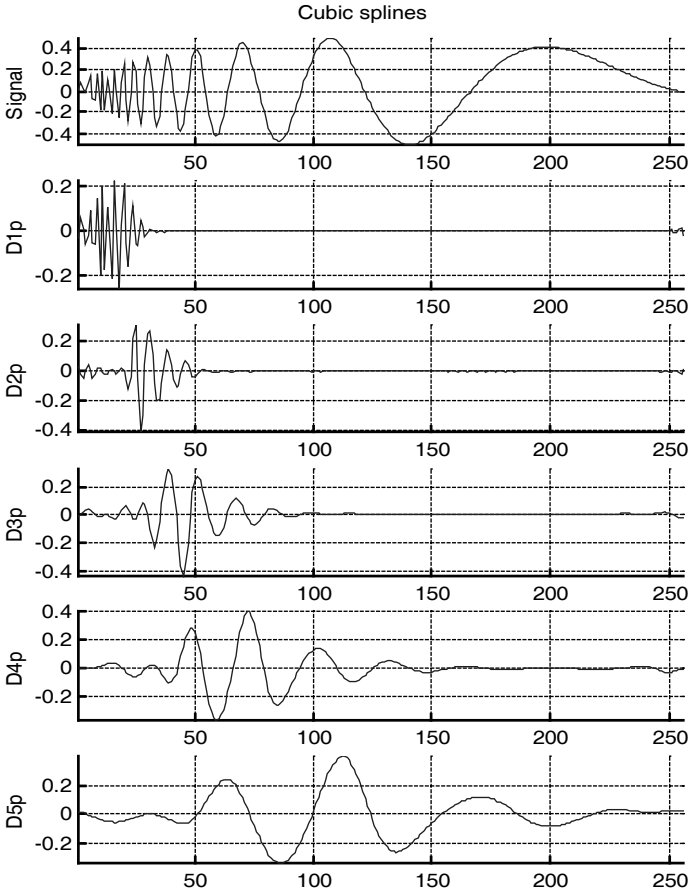
**Figure 10.19.** Recursive algorithm of a MRA

The iterated application of the analysis algorithm helps determine as many wavelet coefficients as necessary – provided we use an initial approximation of the signal to be analyzed. This first approximation is not, in general, precisely accessible and often we must be content with assimilating the digital signal deduced from sampling  $x(t)$  following  $a_n^0$ .

We can show that, in the case of 1 – D, the choice of an orthogonal multiresolution analysis leading to linear-phase filters with finite impulse responses does not work: if the filters are linear-phase, they will be IIR; if they are FIR, then the phase-linearity will not be respected. We see that it is possible to satisfy the two constraints if we use non-orthogonal multiresolution analyses. We can, in particular, by choosing biorthogonal bases, obtain perfect reconstruction analyses that are well-localized in the time-scale, space linear-phase, leading to filters with finite impulse responses. In this situation, the function analysis is different from the reconstruction function and the interpretation of the time-scale analysis will be less clear.

By way of example, we cite two families of filters used in orthogonal multiresolution analyses. The first is phase-linear and the filters are IIR. The analysis functions are constructed by the orthogonalization of a base of functions B-splines, the parameter is of the order  $N$  of these splines [TRU 97a]. If  $N = 0$ , we

find the Haar analysis (very badly frequently localized); if  $N = 3$ , we obtain Battle-Lemarié wavelets; and the higher  $N$  is, the better the frequency localization is. Doppler signal analysis on 5 scales is discussed later as an application example (see Figure 10.20). The signals presented are the approximations to different scales constructed from the transform coefficients. We can clearly see the development of the resolution in the time-frequency plane consecutive to a constant overvoltage analysis; the time localization decreases when the frequency analysis decreases.



**Figure 10.20.** Example of analysis of a Doppler signal by cubic spline wavelets

The second family has been developed by I. Daubechies [DAU 92] and is made of an FIR filter. The analysis functions have a finite structure ( $2N$ ), and the family is parametered by  $N$ . As size increases, the analysis function becomes regular, and

analysis becomes better localized in the time-scale plane. We see that the linear-phase condition can be met in this context for a sufficiently large structure ( $N > 10$ ).

Lastly, we point out that the discrete wavelet transform is not invariant in translation.

**10.6.4. Bilinear transforms**

In some applications, the relevant variable is not the signal itself but its energy. In such cases, we may want time-frequency representations of this variable. Energy is by its nature a quadratic variable and transforms produce these representations using bilinear combinations of the signal (when it is real). These transforms are called bilinear transforms. In the following sections, we discuss several of these, the best known being the group of transforms of the Cohen class: the spectrogram; the scalogram; the Wigner-Ville transform; and the pseudo-Wigner-Ville transform.

10.6.4.1. *The spectrogram*

The simplest way to obtain an energetic time-frequency representation is to use the results of signal representations. The spectrogram is a bilinear transform obtained from the sliding window Fourier transform, which gives a real positive representation:

$$T_{fg} x(t, \omega) T_{fg}^* x(t, \omega) = \left| \int_{-\infty}^{+\infty} x(s) g^*(s - t) \exp(-j\omega s) ds \right|^2$$

We can see that, when a signal is of limited duration, the transform does not preserve the width of the structure, which has been enlarged by the analysis window. As well, the spectrogram is a non-reversible transform, so some loss of information is at the source of this phenomenon. However, in spite of this problem, this representation is one of the most widely used and the range of apodization windows used for calculating the Fourier transform (such as Blackman, cosinusoidal, Gaussian, Hamming, Hanning and Kaiser) influence the choice of a window function.

### 10.6.4.2. The scalogram

The same simple concept is applicable to the wavelet transform. The scalogram is the equivalent of the spectrogram for a time-scale representation:

$$T_{oc}x(a,b)T_{oc}^*x(a,b) = \left| \int_{-\infty}^{+\infty} x(t) \frac{1}{\sqrt{a}} \psi^* \left( \frac{t-b}{a} \right) dt \right|^2$$

The representation is equally positive and real; it also is non-reversible.

### 10.6.4.3. The Wigner-Ville transform

The Wigner-Ville transform (WVT) is a bilinear transform that makes possible another energetic time-frequency representation of the signal [FLA 93]:

$$W_x(t, \omega) = \int_{-\infty}^{+\infty} x(t + \tau/2) x^*(t - \tau/2) \exp(-j\omega\tau) d\tau$$

If we see that the autocorrelation is in fact the mean of the instantaneous correlation  $x(t + \tau/2)x^*(t - \tau/2)$ , we observe that the Wigner-Ville transform is the Fourier transform of this instantaneous correlation. The Wigner-Ville transform can thus be interpreted as a sliding window Fourier analysis in which the window is simply the signal itself reversed in time. It is a natural auto-adaptation of the analysis window. Consequently, with a low structure function, the representation preserves the structure, and there is no time spread. In addition, we see that the instantaneous correlation of a real signal is an even function, so the Wigner-Ville transform is also real. This transform is reversible and satisfies the marginal properties:

$$\int_{-\infty}^{+\infty} W_x(t, \omega) dt = |X(\omega)|^2$$

and:

$$1/2\pi \int_{-\infty}^{+\infty} W_x(t, \omega) d\omega = |x(t)|^2$$

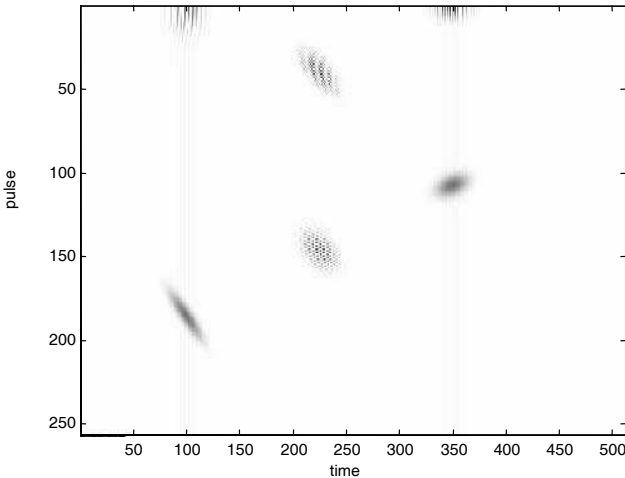
always preserving the total energy:

$$E_x = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W_x(t, \omega) dt d\omega = \int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(\omega)|^2 d\omega$$

However, the Wigner-Ville is not positive and its energetic interpretation is therefore limited. It corresponds more to a quadratic superimposition principle that generates the intermodulation terms in the representation. These terms, which are called interferences, disrupt the smoothness of the time-frequency plane:

$$W_{x+y}(t, \omega) = W_x(t, \omega) + W_y(t, \omega) + 2 \operatorname{Re}[W_{xy}(t, \omega)]$$

with  $W_{xy}(t, \omega) = \int_{-\infty}^{+\infty} x(t + \tau/2) y^*(t + \tau/2) \exp(-j\omega\tau) d\tau$ . These interferences are oscillatory terms that are concentrated in the time-frequency plane in the middle of the segment that joins the representation centers of  $x(t)$  and  $y(t)$ . The phenomenon is illustrated in the figure below that shows the Wigner-Ville transform of the sum of the two signals localized in time (100 and 350) and in frequency. We will see that the interferences also affect the negative frequency terms.



**Figure 10.21.** *The Wigner-Ville transform of a sum of two signals*

We can decrease the number of these disturbance terms in this way: we use the analytic signal deduced from the real signal by presetting to zero the negative frequency components of the Fourier transform of the signal. Another way to

attenuate the interferences is to smooth the oscillations of the Fourier transform with a lowpass filter.

Discretizing the Wigner-Ville transform produces a supplementary constraint; the discrete transform obtained is periodic of period  $\pi$ , so in order to avoid spectrum folding, the sampler must work at a period that is half the signal sampling period required by Shannon's theorem:  $W_x[n, \omega] = 2 \sum_k x[n+k] x^*[n-k] \exp(-j2\omega k)$

#### 10.6.4.4. The pseudo-Wigner-Ville transform

As indicated in the previous section, filtering the WVT can attenuate the interference terms and also improve the smoothness of the time-frequency representation. The new transform obtained is called the pseudo-Wigner-Ville transform (PWVT). This frequency smoothing is equivalent to a time windowing so that the PWVT is obtained by a sliding window Fourier transform of instantaneous autocorrelation:

$$PW_x(t, \omega) = \int_{-\infty}^{+\infty} g(\tau) x(t + \tau/2) x^*(t - \tau/2) \exp(-j\omega\tau) d\tau$$

The filtering window is usually positive and real, so we can propose:  $g(\tau) = h(\tau/2)h^*(\tau/2)$  which allows us to introduce a *windowed* version of  $x(t)$ :  $x_g(t + \tau) = h^*(\tau)x(t + \tau)$  and the PWVT is written:

$$PW_x(t, \omega) = \int_{-\infty}^{+\infty} x_g(t + \tau/2) x_g^*(t - \tau/2) \exp(-j\omega\tau) d\tau$$

This transform is simply the WVT of a *windowed* signal, and so, contrary to the spectrogram, the PWVT preserves the structure of a compact signal. The other side of this advantage is that the sampling, as is the case with the WVT, must be at the double frequency of the minimum required by Shannon for  $x(t)$  to avoid spectrum folding.

## 10.7. A specific instance of multidimensional signals

Processing multidimensional signals is a vast subject which requires long developments that are incompatible with the strongly synthetic nature of our presentation. Here, we will limit ourselves to the few paragraphs that follow, providing some facts and possible research paths.

Multidimensional sensors, especially those producing images (2D) are being used more and more in instrumentation. CCD technology is important in these developments (see [GON 92] and [HOR 93]). However, the processing capabilities brought by the miniaturization of electronic components, and the wide availability of microprocessors, have also contributed to the use of these methods. Data furnished by ultrasonic sensors, thermal sensors, magnetic probes and olfactory sensors are *spatialized*. Three-dimensional (3D) images are also produced by topographical systems used widely in medical contexts, but also in industrial settings (X rays, PET-scan, ultrasounds, scanners and MRI, to give some examples). We will discuss  $nD+1$  for time sequences of spatial signals: time, by its irreversible nature, appears as a variable. These developments have repercussions on processing methods, and the simple transposition of techniques developed for standard time signals are insufficient for processing the specific properties of multidimensional data.

Among the fundamental difference between the two types of signals, we must first note the impossibility of good planning for data  $nD$ . This has obvious consequences for transposing recursive algorithms; the causality constraints, strictly speaking, disappear but implementing the algorithms must take into account the order in which the operations are possible (see the problem of recursive 2D filters). This absence of a natural order of data reading emphasizes the need to use algorithms that respect this symmetry; the problem of phase linearity of the filters used for imaging is a good example of this constraint.

Multidimensional signals are often intrinsically low. CCD sensors, for example, furnish finite dimension images that are known in advance, with the signal to be processed usually available in its whole in the image memory. The problems must be taken into account and a processing can never be completely invariant for translation. Image sensors often work through spatial sampling of data before they are measured. As such, we can say, in many instances, that the *primary* signal is digital and the means of processing the signal must be approached in this perspective.

We also find the usual problems of multidimensional signal processing sometimes transposed: finding differences becomes looking for contours, and the concept of segmentation [COC 95] is at the heart of imaging problems.

In certain cases, 1D processing tools can be transposed directly and naturally. We mention here separable filters in which the impulse response is simply the product of two monovaryable functions. Here, the 2D filtering operation is conducted twice in the form of 1D filtering following lines, then columns. The Gaussian filter is a good example of a separable filter. Correlation function calculations are also separable, as well as time-frequency analyses (Fourier, wavelets, etc.).

However, there are operations and techniques which specifically deal with multidimensional processing [KUN 93]. There are non-separable filters, geometric transforms, segmentation methods by active contour [BLA 97], mathematic morphology [COS 89] and Markovian field modelization [GUY 93], among others. Here, we briefly mention the specific case of signal sequencing processing;  $n$  spatial dimensions more than time. Specific tools are used in this case, since the time dimension has particular properties.

## 10.8. Bibliography

- [AND 79] ANDERSON B.D.O., MOORE J.B., *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, 1979.
- [AZI 96] AZIZ P.M., SORENSEN H.V. and SPIEGEL J.V.D., “An Overview of Sigma-Delta Converters”, *IEEE Signal Processing Magazine*, January 1996.
- [BEL 87] BELLANGER M., *Traitement Numérique du Signal*, Masson, Paris, 1987.
- [BLA 87] BLAKE A., ZISSERMAN A., *Visual Reconstruction*, MIT Press, Cambridge-MA, 1987.
- [CHA 90] CHARBIT M., *Eléments de Théorie du Signal: les Signaux Aléatoires*, Ellipses, Paris, 1990.
- [COC 95] COCQUEREREZ J.P., PHILIPP S., *Analyse d’Images: Filtrage et Segmentation*, Masson, Paris, 1995.
- [COS 89] COSTER M., CHERMANT J.L., *Précis d’Analyse d’Images*, CNRS, Paris, 1989.
- [COT 97] COTTET F., *Traitement des signaux et acquisition de données*, Dunod, Paris, 1997.
- [COU 84] DE COULON F., *Théorie et Traitement des Signaux*, Dunod, 1984.
- [DAU 92] DAUBECHIES I., *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- [DEL 91] DELMAS J.P., *Eléments de théorie du signal: les signaux déterministes*, Ellipses, Paris, 1991.
- [DUV 91] DUVAUT P., *Traitement du signal – Concepts et Applications*, Hermès, Paris, 1991.
- [FLA 93] FLANDRIN P., *Temps-fréquence*, Hermès, Paris, 1993.
- [FON 81] FONDANECHÉ P., GILBERTAS P., *Filtres Numériques, Principes et Réalisations*, Masson, Paris, 1981.
- [GAS 90] GASQUET C., WITOMSKI P., *Analyse de Fourier et Applications – Filtrage, Calcul Numérique et Ondelettes*, Masson, Paris, 1990.
- [GON 92] GONZALES R.C., WOODS R.E., *Digital Image Processing*, Addison-Wesley, 1992.
- [GUY 93] GUYON X., *Champs Aléatoires sur un Réseau*, Masson, 1993.



- [JAZ 70] JAZWINSKI A., *Stochastic Processes and Filtering Theory*, Academic Press, San Diego, 1970.
- [HOR 93] HORAUD R., MONGA O., *Vision par Ordinateur*, Hermès, Paris, 1993.
- [KUN 84] KUNT M., *Traitement Numérique des Signaux*, Dunod, Paris, 1984.
- [KUN 93] KUNT M., GRANLUND G., KOCHER M., *Traitement Numérique des Images*, Presses Polytechniques et Universitaires Romandes, 1993.
- [LAB 88] LABARRERE M., KRIEF J.P., GIMONET B., *Le Filtrage et ses Applications*, collection sup'aéro, 1988.
- [LIF 81] LIFERMANN J., *Les principes du traitement statistique du signal: les méthodes classiques*, Masson, Paris, 1981.
- [MAL 99] MALLAT S., *A Wavelet Tour of Signal Processing*, 2<sup>nd</sup> ed., Academic Press, 1999.
- [MAX 89] MAX J., *Méthodes et Techniques de Traitement du Signal et Applications aux Mesures Physiques*, vol. 1 and 2, Masson, Paris, 1989.
- [MAX 96] MAX J., LACOUME J.L., *Méthodes et Techniques de Traitement du Signal et Applications aux Mesures Physiques*, 5<sup>th</sup> ed., Masson, Paris, 1996.
- [MEY 90] MEYER Y., *Ondelettes et Opérateurs I, Ondelettes*, Hermann, Paris, 1990.
- [MIN 83] MINOUX M., *Programmation mathématique: théorie et algorithmes*, vol. 1, Dunod, Paris, 1983.
- [OPP 75] OPPENHEIM A.V., SCHAFER R.W., *Digital Signal Processing*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1975.
- [PIC 89] PICINBONO B., *Théorie des signaux et des systèmes avec problèmes résolus*, Dunod, Paris, 1989.
- [PIC 93] PICINBONO B., *Signaux aléatoires, tome 1: Probabilités et variables aléatoires*, Dunod, Paris, 1993.
- [PIC 94] PICINBONO B., *Signaux aléatoires, tome 2: Fonctions aléatoires et modèles avec problèmes résolus*, Dunod, Paris, 1994.
- [PIC 95] PICINBONO B., *Signaux aléatoires, tome 3: Bases du traitement statistique du signal avec problèmes résolus*, Dunod, Paris, 1995.
- [PRO 92] PROAKIS J.G., MANOLAKIS D.G., *Digital Signal Processing*, 2<sup>nd</sup> ed., Macmillan, New York, 1992.
- [ROD 78] RODDIER F., *Distributions et transform de Fourier*, McGraw-Hill, Paris, 1978.
- [SCH 91] SCHARF L.L., *Statistical Signal Processing: Détection, Estimation, and Time Series Analysis*, Addison Wesley, 1991.
- [TRU 97a] TRUCHETET F., *Ondelettes pour le signal numérique*, Hermès, Paris, 1997.
- [TRU 97b] TRUCHETET F., *Traitement linéaire du signal numérique*, Hermès, Paris, 1997.
- [WAL 94] WALTER E., PRONZATO L., *Identification de modèles paramétriques*, Masson, Paris, 1994.

## Chapter 11

# Multi-sensor Systems: Diagnostics and Fusion

### 11.1. Introduction

Generally, sensors are used to acquire relevant information about an environment for purposes of knowledge and control. In this chapter, the “control” aspect refers to a sensor’s capacity to respond to specification needs of surveillance and command. The development of this capacity leads to the creation of an acquisition chain in the sensor itself which, although it is an essential element, becomes just one basic element in the overall system.

Actually, beyond obtaining an “image” or “cartography” of an environment, resolving the problem requires diagnostics. This can include, for example, a classification of the situations, then a decision leading to an action taken on the environment. Here we refer to the environment in the broadest sense, not only the external elements but also the internal elements of the sensors themselves.

As for the term “diagnostics”, usually rather broadly defined, it refers to classification issues that focus on three important problems:

- finding weak or degraded system modes;
- signal segmentation (or sometimes detecting a specific transient element in a signal);
- signature classification or, in a broader sense, formula recognition.

Different steps appear in the conception and implementation of a diagnostic or control system; they require choices and optimizations associated with evaluation criteria.

Choosing the right sensor is obviously of crucial importance; the quality of the obtained cartography depends on it. The image of the environment must have sufficient precision and reliability to be used in subsequent processing. To meet this requirement, a network of sensors called a *multisensor* is the best way to increase and improve the overall performance of an observation system.

The first solution consists of using redundancy, by multiplying the number of sensors of the same modalities (with the same variables being observed) working in parallel. Three sensors, observing the same variable, after a “majority vote” decision make up the basis of this approach, which increases reliability.

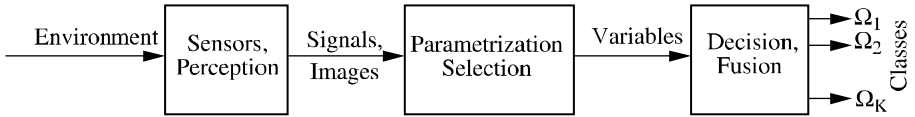
On the other hand, increasing the relevance and the quality of the cartography usually leads to linking sensors of different modalities (different observed variables). For example, in a situation where we want to predict the presence of ice on roads, knowing the temperature and the hygrometric degree of the best predictions will not help us learn anything about the temperature.

Once defined by the transducer or transducers used for learning about the physical phenomenon or, more generally, the environment, we must go on to analyze the information they provide to extract the variables that are relevant to the diagnostic task. The term “information” is used here in its broadest sense, since it can refer to signals, images, numerical or symbolic content, to mention a few usages. The work of parameterization generates and chooses variables, which then help us decide:

- if a system is or is not functioning in degraded mode;
- if a transient is or is not present in a signal;
- if the perceived object belongs or does not belong to a specific class.

If the information is available from precise, complete measurements, the decision phase will effectively integrate techniques for combining data, and the decisions made will lead to reconfiguration actions, of moving or changing sensors, even to an active modification of the environment.

We can see that a pragmatic approach is necessary, centered on sensor choice, variables, decision and data fusion techniques. The nature of the problem is summarized in Figure 11.1 and encompasses three disciplines: signal processing; artificial intelligence; and statistics.



**Figure 11.1.** *Diagnostic processing chain*

This chapter will present these themes in the following order:

- representation space, parametrization and selection;
- Bayesian and non-Bayesian classifications;
- probabilistic and possibilistic fusion.

We cannot cover all existing techniques in depth, but rather offer a range of solutions with comparisons between these solutions whenever possible. We direct the interested reader to reference works for more detailed implementation techniques.

## 11.2. Representation space: parametrization and selection

### 11.2.1. Introduction

The goal of parameterization (see Figure 11.1) is to find a set of variables extracted from raw data that have both a high descriptive potential of signals and a high insensitivity to certain transformations recorded as invariants of the problem (homotheties if the gains are modified, translation if offsets appear; or possibly symmetries). This variable set constitutes the representation space.

In diagnostics, where we must generally classify the observed situations into different categories, other constraints guide the choice of this space: the observations belonging to the same class must be grouped as much as possible in the representation space. Inversely, the observations coming from different classes must be situated in regions separated from each other. In spite of a good level of initial expertise with a given application, sometimes a certain number of variables considered as “obvious” turn out to be relatively irrelevant to the expected diagnostic.

The choice of a representation space also requires a thorough consideration of the practical problems of constructing test bases, which are necessarily of limited size. This is related to the “curse of dimensionality” described by the mathematician Bellman [BEL 61] who showed that the number of observations required for developing and perfecting a program grow exponentially with the dimension  $p$  of the

representation space. Practically, we often must limit  $p$  in the representation space by using feature selection so that the  $n$  observations at our disposal can carry out a reasonably good tiling of the space. This limitation leads to diagnostic algorithms that require less calculation time, an especially important parameter for reactive systems.

The work is thus divided into two stages: parameterization, then variable selection, but these stages are in fact joined. After a brief summary of the parameterization currently in use, the following sections will offer a range of variable selection methods.

### 11.2.2. Signal parametrization

Generating descriptive variables is greatly influenced by whether or not we understand and can control the type of physical phenomenon being observed and the type of sensor being used.

When there is significant *a priori* knowledge, heuristic parameterizations are often used; these are chosen for their known relevance. These parameterizations can group together variables of very different types, such as peak-to-peak values, useful values, Kurtosis factors, mid-height widths, power spectral densities or wavelet transforms [ZWI 95]. In the case of image information, these variables can be contrast measurements, entropy measurements, first or second order histograms, measurements of local curves or air [THE 99]. These heuristic parameterizations compress the initial information of a certain number of features judged relevant by the expert.

Unfortunately, this kind of expertise is not always possible – especially for a new problem – and it is often difficult to predict in advance which variable will be important. Compressing information *a priori* cannot really be done, at least in the same terms. It is better to use more complete modelizations of initial information which will help in processing unknown degraded modes or a new class of objects.

Processing signals and images can be done using many different techniques, among them being: AR, MA and ARMA modelization; Fourier descriptors; time-frequency analyses; polynomial approximations; splines; and Prony models. All these techniques provide, for each observation, a variable set, such as coefficients of ARMA filters, serial Fourier coefficients, or polynomial coefficients. The dimensions of this variable set are usually significant, but before trying to select from among these variables, certain expressions must be represented.

The observation basis will be written  $X$ . This includes  $n$  lines corresponding to  $n$  observations  $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$  and  $p$  columns corresponding to  $p$  variables  $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_p$  obtained after signal parameterization:

$$X = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_n \end{bmatrix} = \begin{array}{ccc} \begin{matrix} \text{variable} \\ \downarrow \end{matrix} & & \\ \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \ddots & \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} & \leftarrow \text{observation} & \\ & & = [\underline{X}_1, \underline{X}_2 \cdots \underline{X}_p] \end{array} \quad [11.1]$$

For each of these observations of value in  $\mathbb{R}^p$ , we have one desired output  $y_i$  of the decision system. This output corresponds to a class designation and takes values of between 0 and  $K - 1$  for a problem general to  $K$  classes  $\Omega_0, \Omega_1, \dots, \Omega_{K-1}$ :

$$\underline{Y} = [y_1, y_2 \dots y_n]^t \text{ with } y_i \in \{0, 1 \dots K-1\} \quad [11.2]$$

First and second order statistics will later be utilized. The center of gravity in the scatter plot is a vector of  $p$  components:

$$\underline{m} = [m_1, m_2 \dots m_p] \text{ with } m_i = E[\underline{X}_i] = \frac{1}{n} \sum_{k=1}^n x_{ki} \quad [11.3]$$

The covariance matrix will be written  $V$ :

$$V = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{1p} \\ \sigma_{21} & \ddots & \\ \vdots & & \ddots \\ \sigma_{p1} & & \sigma_{p^2}^2 \end{bmatrix} \text{ with } \sigma_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - m_i)(x_{kj} - m_j) \quad [11.4]$$

This matrix of dimensions  $p \times p$  is symmetrical and positive. Its values are therefore real positive or negative. They will be written as  $\lambda_i$ .

The total variance of the scatter plot is represented by:

$$\sigma^2 = \frac{1}{n} \sum_{k=1}^n \|\underline{x}_k - \underline{m}\|^2 \quad [11.5]$$

We can also show that this total variance is equal to the trace of  $V$  [SAP 90]

$$\sigma^2 = \sum_{k=1}^p \sigma_k^2 \quad [11.6]$$

We see that this quantity is independent of the coordinates system chosen and that in the particular axial system made of vectors specific to  $V$ , this total variance is expressed by:

$$\sigma^2 = \sum_{k=1}^p \lambda_k^2 \tag{11.7}$$

In order to suppress irrelevant variables that may appear, or variables that are over-coordinated between themselves, and also to limit the curse of dimensionality (see section 11.2.1), we must choose a subset of a variable set  $\underline{X}_1, \underline{X}_2, \dots, \underline{X}_p$ . The following sections will present in some detail several good selection methods in a supervised context in which a database, which we assume to be exhaustive, is available for designing a diagnostic system.

### 11.2.3. Principle component analysis

This method of information analysis was first proposed by K. Pearson in 1901. Basically, principle components analysis or PCA determines the  $p_r$  axes of a subspace or  $\mathbb{R}^p$  that best represents the basic data  $X$  after projection. The “representivity” criterion is an inertia type criterion:

$$J = \frac{1}{n} \sum_{i=1}^n \|\underline{x}_i - \text{proj}(\underline{x}_i)\|^2 \tag{11.8}$$

In PCA, we try to make  $J$  minimum, so that the projected scatter plot is as undeformed as possible; that is, the variance of the projected scatterplot is then maximum.

We can easily show [SAP 90] that the subspace we are looking for is generated by the  $p_r$  vectors proper to  $\underline{U}_i$  of the variance-covariance matrix  $V$  (see equation [11.4]) related to its first  $p_r$  proper values  $\lambda_i$  ranged in decreasing order. The axes defined by these proper vectors are called inertia axes or principle axes.

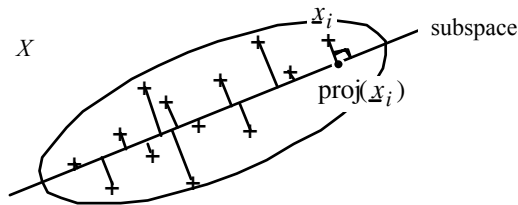


Figure 11.2. Two-dimensional illustration of the PCA

From the example of the 2D illustration in Figure 11.2, we can easily see that 1D space that least deforms the initial scatterplot is the linear regression line that merges with the first inertia axis. In any dimension space, the PCA determines through linear combinations the initial  $p$  variables  $\underline{X}_i$  of the new centered variables  $\underline{U}_j$  of the maximum and uncorrelated variance between them.

The PCA can be used for two purposes: to reduce the representation space of the data by projecting the data in a space of reduced dimensions, or simply to aid in visualizing the observation bases of large dimensions. In analyzing data, it is important to visualize the correlations between the initial variables and the variables coming from the PCA that “summarize” a large part of the basic inertia. The correlation between the initial variable  $\underline{X}_i$  and the principle component  $\underline{U}_j$  is calculated with the following expression [LEB 97]:

$$R(\underline{X}_i, \underline{U}_j) = \frac{\sqrt{\lambda_j} U_j(i)}{\sigma_i} \quad [11.9]$$

where  $U_j(i)$  is the  $i^{\text{th}}$  component of the vector related to  $\lambda_i$ .

The visualization of these correlations is traditionally done with correlation circles in which each initial  $\underline{X}_i$  variable is represented by a coordinate point:

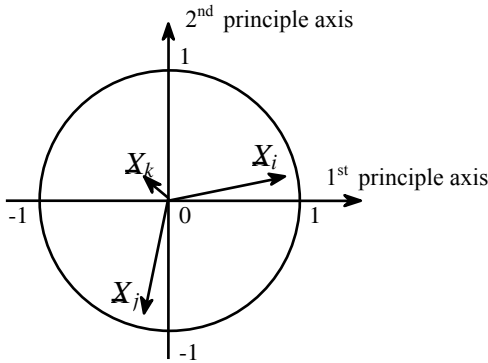
$$\left( R(\underline{X}_i, \underline{U}_j), R(\underline{X}_i, \underline{U}_k) \right)$$

This visualization helps us understand the links between each  $\underline{X}_i$  variable with the principle inertia axes (the principle plane  $j = 1, k = 2$  is the most used). These points are contained in a center circle 0 and of radius 1, whatever  $i$  and  $j$  are:

$$-1 \leq R(\underline{X}_i, \underline{U}_j) \leq 1$$

Looking at the example in Figure 11.3, we observe that  $\underline{X}_i$  is correlated to the first principle axis,  $\underline{X}_j$  is anticorrelated to the second principle axis, and  $\underline{X}_k$  has no strong correlation with these two axes.

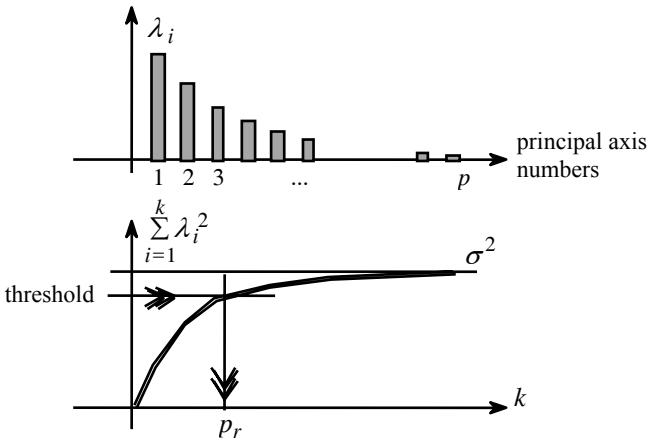




**Figure 11.3.** PCA correlation circle

The PCA is a relatively powerful data analysis method, allowing many ways of visualizing information [SAP 90]. With the PCA, we can verify the independence of descriptive signal variables. It offers a new representation base made by a linear combination of initial variables. The axial hierarchy is established by using an inertia criterion that favors variables presenting the highest variances.

If we want to reduce the number of initial variables by using the PCA, we then face the problem of choosing the  $p_r$  dimension of the representation subspace.

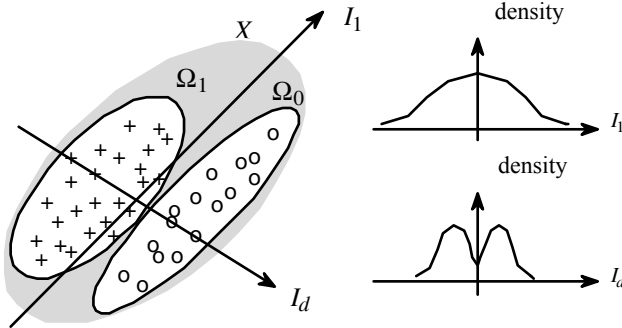


**Figure 11.4.** Choosing a dimension of the projection subspace using a PCA

Usually this choice is made by visualizing the range of values belonging to the matrix of variance-covariance ranged in decreasing order. By using equation [11.7], each value raised to the square “explains” part of the total variance of the cloud of data points  $X$ . Choosing a threshold percentage of the curve represents the accumulated sum of the squares of the values that help us obtain the value of  $p_r$  (see Figure 11.4). The existence of a break in the value set also helps us set the threshold.

As we can observe, this choice is made without taking into account the class  $Y$  labeling that is available to us. This is doubtlessly the biggest drawback to using the PCA as a method for selecting variables for diagnostic purposes. Actually, nothing shows that the retained variables will be the most relevant ones for separating classes. With the example shown in Figure 11.5, the basis of data  $X$  is divided in two groups (corresponding, for example, to the class of correct modes and to the class of faulty modes). The principle axis  $I_1$  (the one possessing the highest inertia) is not essentially the most important axis for distinguishing between the two classes; choosing the axis  $I_d$  seems much better.

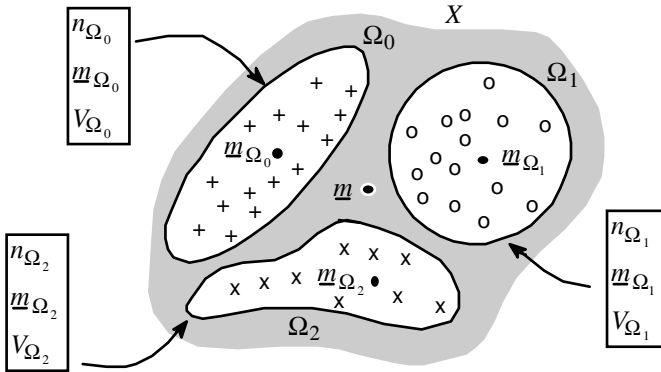
These considerations lead to implementing other technologies that take into account class labeling.



**Figure 11.5.** Principle inertia axis and class separation

#### 11.2.4. Discriminate factorial analysis

The work of Fischer and Mahalanobis (1936) first explored this statistical method. Discriminate factorial analysis (DFA) is both a descriptive method and a method of classifying data. In this section, we will only discuss the first method.



**Figure 11.6.** Centre of gravity and variance matrix internal to classes

To present the basic principles of DFA, some equations must be represented. For each  $\Omega_i$  of  $X$ , it is possible to define a center of gravity and a variance-covariance matrix by using equations [11.3] and [11.4]. These are expressed, respectively,  $\underline{m}_{\Omega_i}$  and  $V_{\Omega_i}$ .  $n_{\Omega_i}$  is the number of observations of  $X$  belonging to the class  $\Omega_i$  (see Figure 11.6).

Using the variance-covariance matrices belonging to each class, we can represent the idea of an interclass variance matrix:

$$V_{in} = \frac{1}{n} \sum_{i=0}^{K-1} n_{\Omega_i} V_{\Omega_i} \tag{11.10}$$

The interclass variance matrix produces the average of the matrices proper to each class. It can be interpreted as an overall measurement of the concentration of classes around their center of gravity. If all the classes were reduced to their center of gravity, their matrix trace of interclass variance would be zero.

Inversely, the interclass variance matrix measures the dispersion of the centers of gravity of classes:

$$V_{ex} = \frac{1}{n} \sum_{i=0}^{K-1} n_{\Omega_i} (\underline{m}_{\Omega_i} - \underline{m})(\underline{m}_{\Omega_i} - \underline{m}) \tag{11.11}$$

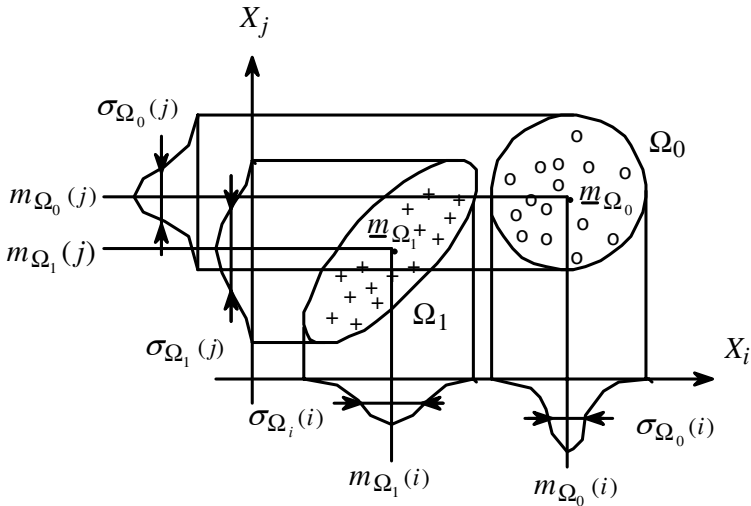
The closer the  $K$  centers of gravity are to each other, the weaker the interclass matrix traces are.

Analogous to Huygen's theorem in mechanics, we show that the total variance is equal to the average of the variances plus the variance of the averages [GAI 83]:

$$V = V_{in} + V_{ex} \quad [11.12]$$

We can easily understand the importance of these two matrices for diagnostic problems. The DFA tries to find a projection into the subspace of  $\mathbb{R}^p$  that minimizes  $V_{in}$  (concentrated classes) while maximizing  $V_{ex}$  (classes far from each other). The main result of this method [SAP 90] is that the subspace we are looking for is generated by the  $p_r$  vectors belonging to the matrix  $V_{in}^{-1}V_{ex}$  linked to its first values ranged in decreasing order. The choice of  $p_r$  is made in exactly the same way as with the PCA method. However, we can show that the ranking of  $V_{ex}$  is at most  $K-1$  [GAI 83]; this means we must limit  $p_r$  to  $p_r \leq K-1$ .

A more detailed analysis also shows that the DFA is nothing less than a PCA on the  $K$  centers of gravity of classes having a non-Euclidian metric (the Mahalanobis distance:  $\|u\|^2 = u^t V_{in}^{-1}u$  [GAI 83]).



**Figure 11.7.** Illustration of the Fischer criterion

A simplified version of this method, called the Fischer criterion, consists of measuring the compactness of classes not on the principle axes of the matrix  $V_{in}^{-1}V_{ex}$

but on the source axes by making the hypothesis that they are orthogonal. For a problem of two classes, for example, we calculate the quantities:

$$F(X_i) = \frac{(m_{\Omega_0}(i) - m_{\Omega_1}(i))^2}{n_{\Omega_0}\sigma_{\Omega_0}^2(i) + n_{\Omega_1}\sigma_{\Omega_1}^2(i)} \tag{11.13}$$

The larger this quantity, the better we can discriminate it from the axis  $i$ . For example, in Figure 11.7, the variable  $X_i$  is more discriminate than the variable  $X_j$ .

This sub-optimal method does not allow us to combine the initial variables to obtain new, more relevant variables. Selecting variables using this criterion is therefore very simple to do; a version of equation [11.13] can be extended to a situation in which  $K$  classes also exist [DOC 81]. We should avoid using this simple method when the complete variance-covariance matrix structure is not close to being a diagonal structure.

DFA is often more useful than the PCA for diagnostic problems and for formula recognition. This is because it takes into account the fact that elements of the observation base appear in classes. However, the interclass variance matrix used is an averaged matrix that only imperfectly represents the internal variance matrices of each class, especially if the classes are very disparate. Added to this theoretical difficulty is a practical one that occurs during the estimation of matrices  $V_{in}$  and  $V_{ex}$ , mainly if the observation basis contains few examples.

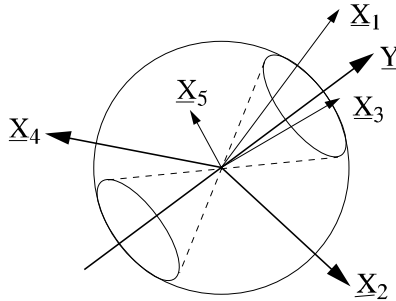
### 11.2.5. Selection by orthogonalization

The selection method described in this section uses the linear regression formalism. With the representations proposed in section 11.2.1, choosing a linear regression model imposes the following matrix relation between the input and output variables:

$$\underline{Y} = X\underline{P} + \underline{\varepsilon} \tag{11.14}$$

$\underline{P}$  contains the regressors obtained by orthogonal projection of  $\underline{Y}$  on the information base  $X$ . This projection is carried out in the least squares sense in order to minimize the modelization error  $\underline{\varepsilon}$ . We remember that the output vector  $\underline{Y}$  contains the class labeling of the base examples.

The relevance of each variable – that is, of each  $X$  column – is estimated by measuring the output vector in a space of dimension  $n$ , with  $n$  being the number of base examples [CHE 89]. The lower the angle between  $\underline{X}_i$  and  $\underline{Y}$ , the better this variable “explains” the output. In the example shown in Figure 11.8, the variable  $\underline{X}_3$  is the most relevant.



**Figure 11.8.** Illustration of the OFR method for  $n=3$  and  $p=5$

The selection method by orthogonalization (or the Orthogonal Forward Regression (OFR)) works by a classification of initial variables, from the most relevant to the least relevant. However, so as to not count the same information several times, we iteratively eliminate the variables remaining to be classed. The Gram-Schmidt iterative orthogonalization procedure is used for this.

During the first iteration, we choose the variable  $\underline{X}_{i_1}$ , the most colinear to  $\underline{Y}$ :

$$\cos(\underline{X}_{i_1}, \underline{Y})^2 = \max_{1 \leq k \leq p} \left[ \frac{(\underline{X}_k^t \underline{Y})^2}{\|\underline{X}_k\|^2 \|\underline{Y}\|^2} \right]$$

During the second iteration, we first orthogonalize all the remaining variables, as well as the output so it is perpendicular to  $\underline{X}_{i_1}$

$$\underline{Y}^{(2)} = \underline{Y} - \frac{\underline{X}_{i_1}^t \underline{Y}}{\underline{X}_{i_1}^t \underline{X}_{i_1}} \underline{X}_{i_1} \quad \text{and} \quad \underline{X}_k^{(2)} = \underline{X}_k - \frac{\underline{X}_{i_1}^t \underline{X}_k}{\underline{X}_{i_1}^t \underline{X}_{i_1}} \underline{X}_{i_1}$$

with  $1 \leq k \leq p$  and  $k \neq i_1$

The procedure ends at the iteration  $p$ , when all the variables have been classed.

The choice of the reduced dimension  $p_r$  of the representation subspace is carried out, as with the PCA method, by choosing a threshold for the estimation of the contributions of the subspaces successively constructed. This contribution is written:

$$o(j) = \sum_{k=1}^j \cos(\underline{X}_{i_k}^{(k)}, \underline{Y}^{(k)})^2$$

The OFR method is very simple to implement. The underlying linearity hypothesis is quite complex, but the variables retained by the method are pertinent [STO 97]. In addition, the curve visualization  $o(j)$  (see Figure 11.9) easily gives us an idea of the significance of the non-linearities of the problem as formulated; if the problem is completely linear,  $o(p) = 1$ . Observing a sufficiently high value of  $o(p)$  ( $0.75 < o(p) < 1$ , for example), *a posteriori* justifies using this method.

The reader will find very complete developments in the variable selections in texts such as [DUD 73], [FUK 72], [KIT 86], and [KRI 82].

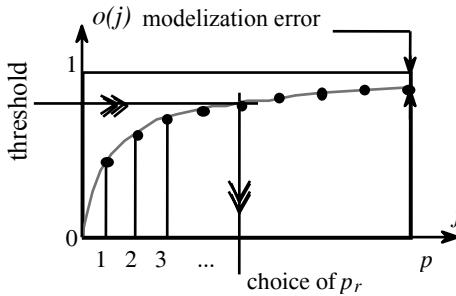


Figure 11.9. Choice of reduced dimension in the OFR method

### 11.3. Signal classification

#### 11.3.1. Introduction

From the parameterization/selection phase, we use observations belonging to an optimized representation space, from which we implement the diagnostic procedures. The labeling of a base being assumed known, section 11.3 will only present techniques of supervised classification.

After having introduced the Bayesian classification approach, first we will discuss in some detail a Bayesian parametric method, then k methods and Parzen nuclei.

The last two methods to be presented come from the domain of Bayesian classification; they are first of all decision trees and neuron networks, of either multilayered perceptron types or radial base functions.

### 11.3.2. Bayesian classification

#### 11.3.2.1. Optimum Bayes classifier

Statistical classification has a long history that dates back to the work of Thomas Bayes (1763). His formula helps in calculating the probability of creating a class by knowing the observation  $\underline{x}$  is created in a given space. The Bayes formula, given below, is a consequence of definitions and properties of conditional probability laws. Its practical importance is significant:

$$p(\Omega_i | \underline{x}) = \frac{p(\underline{x} | \Omega_i) p(\Omega_i)}{\sum_{j=0}^{K-1} p(\underline{x} | \Omega_j) p(\Omega_j)} \quad (\text{Bayes formula}) \quad [11.15]$$

$p(\Omega_i | \underline{x})$  is the *a posteriori* probability that the observation represented by  $\underline{x}$  that belongs to  $\Omega_i$ ,  $p(\Omega_i)$  is the *a priori* probability of the class  $\Omega_i$  and  $p(\underline{x} | \Omega_i)$  is the probability that the observation equals  $\underline{x}$  knowing that it belongs to the class  $\Omega_i$ .

Applying the Bayes formula assumes that the range of classes is complete, or that – since we are dealing with classification – each observation only belongs to one class and that the group of classes entirely covers the representation space  $R^p$ .

*Example: what is the probability that every person measuring 1.60 meters in height is a woman?*

Responding to this question brings us back to estimating the following conditional probability:  $p(\Omega_F | x = 1.60)$ . If the person is chosen randomly from a population, and if we suppose that this population is half women and half men, applying the Bayes formula gives us:

$$p(\Omega_F | 1.60) = \frac{p(1.60 | \Omega_F) \times 0.5}{p(1.60 | \Omega_F) \times 0.5 + p(1.60 | \Omega_H) \times 0.5}$$



With this simplified model of the distribution of the heights of men in France, a Gaussian centered at 1.75 meters, and a deviation of 0.15 meters, and for women a Gaussian centered at 1.65 meters and a deviation of 0.15 meters, applying the Bayes formula finally gives us:

$$p(\Omega_F | 1.60) = \frac{1.61 \times 0.5}{1.61 \times 0.5 + 2.52 \times 0.5} = 60.9\%$$

If now, we choose a person from among, say, a national legislative body, and not from among the French population, the probabilities change, if we assume that this legislative body is composed of 10% women and 90% men. Applying the Bayesian rule in this case gives us:

$$p(\Omega_F | 1.60) = \frac{1.61 \times 0.1}{1.61 \times 0.1 + 2.52 \times 0.9} = 14.8\%$$

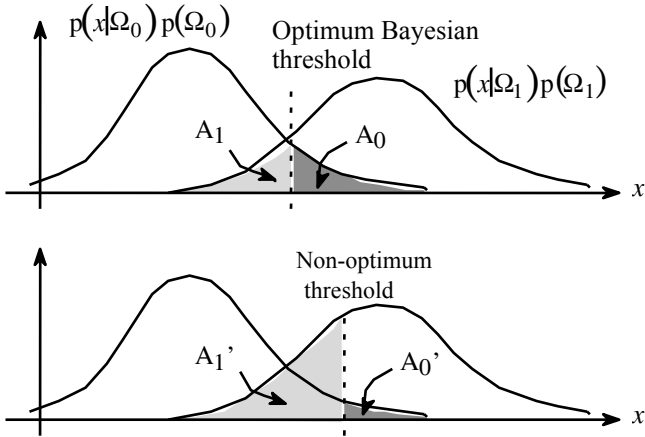
We notice that the answer to the question is significantly influenced by the probabilities of classification. With the formula, we can even calculate that a legislator has 50% chance of being a woman, if this legislator is less than 1.20 meters tall. Of course, these models are very simplified.

To construct a classifier, we must complete the Bayes formula with a decision rule. The decision rule of Bayes leads to an optimum classifier and is expressed as follows:

$$\underline{x} \in \Omega_i \quad \text{such that} \quad \text{Max}_{j=0 \dots K-1} \left[ p(\Omega_j | \underline{x}) \right]$$

The final choice of class is carried out by comparing the *a posteriori* probabilities of belonging to all the classes and by choosing the highest one. By observing that the denominator of equation [11.15] is the same for all the conditional probabilities, irrespective of class, the Bayes rule can also be formulated in the following way:

$$\underline{x} \in \Omega_i \quad \text{such that} \quad \text{Max}_{j=0 \dots K-1} \left[ p(\underline{x} | \Omega_j) p(\Omega_j) \right] \quad (\text{Bayes rule}) \quad [11.16]$$



**Figure 11.10.** Illustration of the Bayesian decision rule for two classes

Figure 11.10 shows the Bayesian rule in one dimension. On the first curve, we see that the decision threshold is placed at the equality point of probabilities *a posteriori*, as the Bayes rule shows. The line of this threshold (relative to the left), the observation will be modified to class 1 (relative to class 0). The poorly modified observations of class 0 (relative to class 1) are regrouped in the area  $A_0$  (respectively in the area  $A_1$ ). With a threshold differently arranged, as in the second curve in Figure 11.10, the sum of these two areas can only increase (see areas  $A'_0$  and  $A'_1$ ). This explains that the Bayes rule must also be called a minimum cost rule.

This decision rule gives an equivalent weight to all classification errors. Practically, we perhaps penalize certain errors more harshly (for economic or security reasons, for example). We direct the interested reader to more specialized texts for variants of the Bayesian decision: [DUB 90] and [THE 99].

The Bayesian classifier cannot always be used directly; it requires knowing the probabilities of belonging to classes, as well as the internal probabilities densities of classes. To resolve these difficulties, many classification methods have been developed. Some of these, including derivation methods, parametric and non-parametric methods, will be discussed at the end of this chapter. The Bayesian decision rule still has great theoretic importance; it provides a standard for comparison for all these methods.

11.3.2.2. *Parametric Bayesian classification*

This method provides a model for probability densities and *a priori* probabilities with parameters that we adjust with a learning base. The most up-to-date model is that of the classes distributed according to the Gaussian multidimensional laws:

$$p(\underline{x}|\Omega_i) = \frac{1}{(2\pi)^{p/2} |V_{\Omega_i}|^{1/2}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{m}_{\Omega_i})^t V_{\Omega_i}^{-1} (\underline{x} - \underline{m}_{\Omega_i})\right)$$

The parameters  $\underline{m}_{\Omega_i}$  and  $V_{\Omega_i}$  are determined on the learning base. Lacking complementary information, the probabilities will be chosen equal:

$$p(\Omega_i) = \frac{n_{\Omega_i}}{n} \tag{11.17}$$

In this specific case of a parametric model, applying the rule of Bayes' decision (see equation [11.16]) will lead to [DUB 90]:

$$\underline{x} \in \Omega_i \text{ such that } \underset{j=0\dots K-1}{\text{Min}} \left[ (\underline{x} - \underline{m}_{\Omega_j})^t V_{\Omega_j}^{-1} (\underline{x} - \underline{m}_{\Omega_j}) + \text{Log}(|V_{\Omega_j}|) - 2\text{Log}(p(\Omega_j)) \right]$$

The first term expresses the removal of the observation to the center of the class *j* with a Mahalanobis metric. The second term is a corrective term linked to the dispersion of of the class *j*. The final term takes into account the *a priori* probability of the class *j* in the decision rule. If the classes are equiprobable and of the same dispersion, the Bayes rule is reduced to an attempt to find the minimum distance between the classes' centers.

11.3.2.3. *Method of the k-nearest neighbor*

This non-parametric method was introduced by Fix and Hodges in 1951. Instead of using a probability density model, here we try to locally estimate these features by observing the nearness of each observation. We determine the volume *v* centered on  $\underline{x}$  which incorporates *k* observations of the learning base. Once we have determined this volume, we count the number *k<sub>i</sub>* of the neighbors belonging to each class. The probability density of the class *i* is then estimated locally by:

$$\hat{p}(\underline{x}|\Omega_i) = \frac{k_i}{n_{\Omega_i} \times v(\underline{x})} \quad i = 0\dots K-1 \quad \text{with} \quad \sum_{i=0}^{K-1} k_i = k$$

Choosing a specific metric influences the form of the volume  $v$ ; the volume is a sphere with a Euclidian distance, a cube with a Manhattan distance ( $\|u\| = \sum |u_i|$ ) and an ellipsoid with a Mahalanobis distance.

If we choose, *a priori*, a probability like the one in [11.17], the Bayes decision rule then becomes very simple:

$$\underline{x} \in \Omega_i \quad \text{such that} \quad \text{Max}_{j=0 \dots K-1} [k_j]$$

The observation is allocated to the class that is most represented among the closest  $k$  neighbors. We can demonstrate that the error rate of the method tends towards that of the Bayesian classifier if  $k$  tends to infinity (the error is two times higher if  $k = 1$ ) [KRI 82].

Unfortunately, this method, which is very easy to use, takes a long time to calculate; with each new observation to be classed, we must calculate the distances between each observation and all the observations of the learning base.

#### 11.3.2.4. Parzen nuclei

A dual method of the method described above was developed more recently by Rosenblatt in 1956 and Parzen in 1962. It fixes a given volume around the observation and counts the number of examples of the learning base that it contains, class by class. The simplest volume is a hypercube of side  $h$  with a volume  $h^p$ . In this way, we directly estimate the probability densities with the formula:

$$\hat{p}(\underline{x}|\Omega_i) = \frac{k_i(\underline{x})}{n_{\Omega_i} \times h^p} \quad i = 0 \dots K-1 \quad [11.18]$$

Applying this method often leads to obtaining very noisy probability densities, the results of densities that are too low in the learning base in certain zones of the representation space. Parzen proposed smoothing these densities by using nuclei [PAR 62]; these “gently” modify the term  $k_i(\underline{x})$  shown in equation [11.18], instead of an all-or-nothing counting:

$$\hat{p}(\underline{x}|\Omega_i) = \frac{1}{n_{\Omega_i} \times h^p} \sum_{\underline{x}_k \in \Omega_i} \varphi\left(\frac{\underline{x} - \underline{x}_k}{h}\right) \quad [11.19]$$

$\varphi$  is the nuclei function we must verify:  $\varphi(\underline{x}) \geq 0$  and  $\int_{\mathbb{R}^p} \varphi(\underline{x}) d\underline{x} = 1$ .

Finally, by the bias of this function, each observation of the class I intervenes in estimating  $p(\underline{x}|\Omega_i)$ , not those situated in the immediate proximity of  $\underline{x}$ . Gaussian nuclei are most often used. The adjustment parameter  $h$  then plays the role of the gap type of the nuclei; the more  $h$  increases, the more the density estimation will be smoothed.

By supplementing a choice concerning *a priori* probabilities (for example, in equation [11.17]), the decision rule in equation [11.16] can then be applied for modifying the class.

### 11.3.3. Decision trees

The first work on this non-Bayesian method dates from the 1960s. According to experts, tree classification brings about a series of interleaved tests with a learning phase that helps define the structure. These tests operate successively on each descriptive variable and divide the representation space into the most homogeneous regions possible relative to the classes. Figure 11.1 gives an example of a tree structure in a situation with two dimensions and three classes.

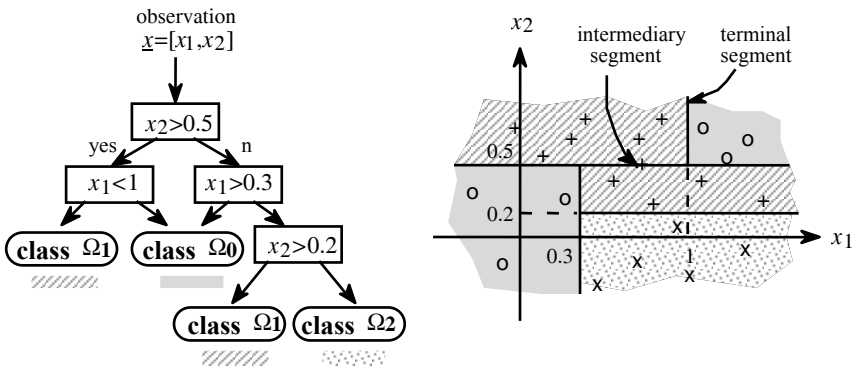


Figure 11.11. Example of a decision tree in two dimensions

There are many ways to construct this type of tree [GUE 88]. Their construction principles are often the same. Using the complete learning group, we must look for all the variables, with the best thresholds separating the base into two groups that are as homogenous as possible relative the classes. We then have the threshold – and its linked variable – which produces the best among the best separations. We then develop the two branches obtained. With each of them, we apply the same procedure

as before, but applied to the part of the learning base that corresponds to the threshold as defined above.

Each branch develops this way up to the point when the observations verify all the tests that already belong to the same class. This sub-group is then termed “pure”.

To measure the efficiency of a separation, entropic criteria or “rate of impurities” criteria can be used. For example, the impurity measurement of a group E can be obtained by using equation [11.20] where  $p(\Omega_i|E)$  corresponds to the proportion of representatives of the class  $i$  in the group E. We can verify that the impurity measurement of a pure group is zero:

$$I(E) = \sum_i \sum_{j, i \neq j} p(\Omega_i | E) \times p(\Omega_j | E) \quad [11.20]$$

In our case, the placing of a threshold  $\alpha$  on a variable  $x_i$  leads to measuring the linked impurities of the two subsets issues from the test:

$$I(\alpha) = \sum_r \sum_{\substack{s \\ r \neq s}} p(\Omega_r | x_i > \alpha) \cdot p(\Omega_s | x_i > \alpha) + p(\Omega_r | x_i \leq \alpha) \cdot p(\Omega_s | x_i \leq \alpha) \quad [11.21]$$

For classification, the procedure is very simple; we carry out a new observation on the majority class represented in the extremity of the branch where the observation is located.

The main importance of this method is that the order of the tests allows us to take a decision in an optimized way according to the zones of the space where the observation is located. Modifying a particular class only requires a low number of tests, and only a few variables need be introduced; but for other classes, the set of tests can be longer. Overall, the calculation time required is low.

What is more, knowing all the descriptive variables is not necessary if the observation is in a zone where the modification is simple. Here, the parametrization time is less (but in industrial control settings, the cost of certain controls is quite high).

The problems with implementing decision trees lie in choosing the development level of each branch. If the sub-groups obtained are not pure, the tree structure can be developed up to the point where each extremity only contains a single observation! The correct classification rate is then 100% for the learning base but of course, the generalization capacity of the tree is poor.

In the domain of decision trees, current research is being done on “pruning” techniques applied to trees whose branches are too long, while looking for more efficient subtrees [LEB 97].

### 11.3.4. Neural networks

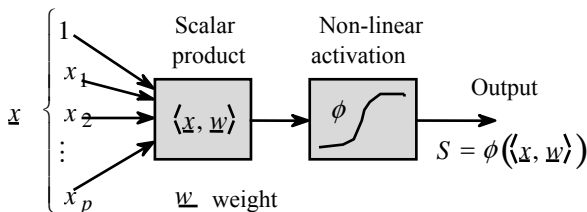
Neural networks (NN) have undergone important developments since the mid-1980s. The work of Rumelhart and Le Cun [RUM 86] is fundamental to this new trend, since with it we can implement efficient and fast algorithms for learning multilayered NNs. However, as early as the 1940s, McCulloch and Pitts had begun working on elementary NNs.

The “biological analogy” is the source of the name, but today, NNs are no longer seen as simple mathematical operators. Rather, their popularity is due more to their universal approximate and inexpensive qualities [BIS 95] than to their potential to reproduce the functioning of the human brain!

Several families exist. In this chapter, we will discuss the most widely used, supervised-mode structures: the multilayer perceptron (MLP) and the Radial Basis functions network (RBF). We direct the reader to other texts dealing with other, less widely used networks [HAY 94] [HEY 94].

#### 11.3.4.1. Basic neurons

Neural networks are assemblies of basic blocks; we will discuss two types of these. The first type is a basic neuron called a “scalar product”.



**Figure 11.12.** “Scalar product” type of basic neuron

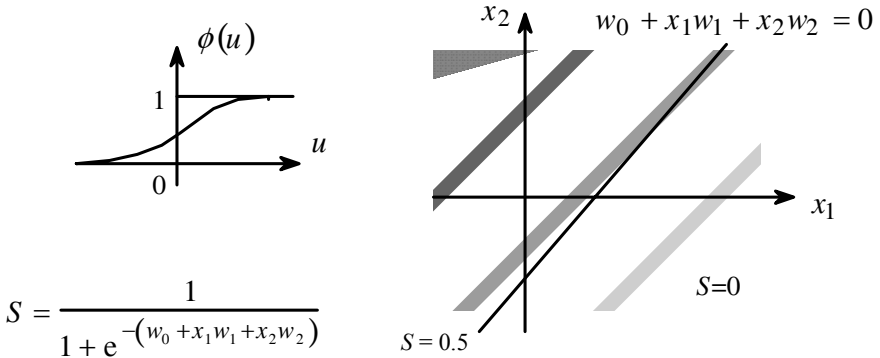
This neuron carries out two operations. The first operation is the scalar product between the input vector and a vector  $w$  called the weight (we see that the input vector includes a constant so that an adjustable bias may be introduced).

The second operation is an activation throughout of a non-linear function with many variances: a threshold function (Rosenblatt's perceptron, 1960); linear (Adaline de Widrow's model, 1960); and a hyperbolic tangent. The sigmoidal function is also a widely used function (see Figure 11.13).

This "scalar product" neuron carries out a linear separation of the input space in two regions, whether or not the activation function being used is linear or non-linear.

An example is given in Figure 11.13 in two dimensions. Here we see the line of the plane for which  $S = 0.5$  and which defines two half-planes. The neuron output is indicated in gray. We can understand the importance of using this type of operator in a classification problem of two classes; for example:

$$S < 0.5 \Rightarrow \underline{x} \in \Omega_0 \quad \text{and} \quad S > 0.5 \Rightarrow \underline{x} \in \Omega_1$$



**Figure 11.13.** Linear separation in two dimensions

The second type of neuron is called a basic "distance" neuron. This time, the neuron (or nucleus) calculates the distance, with a metric  $A$ , between the input vector and a center  $\underline{C}$ , before injecting the result into the activation function, which is usually Gaussian. The possible partition of the space is then quadratic. An example is given in Figure 11.15 in two dimensions.



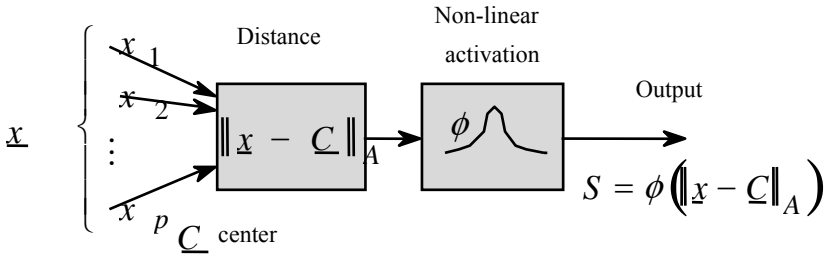


Figure 11.14. Basic “distance” neuron

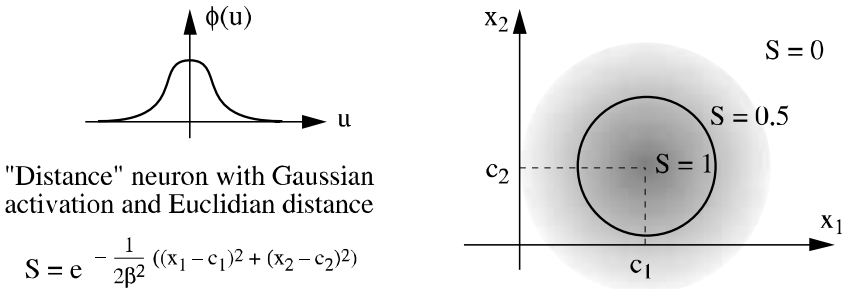
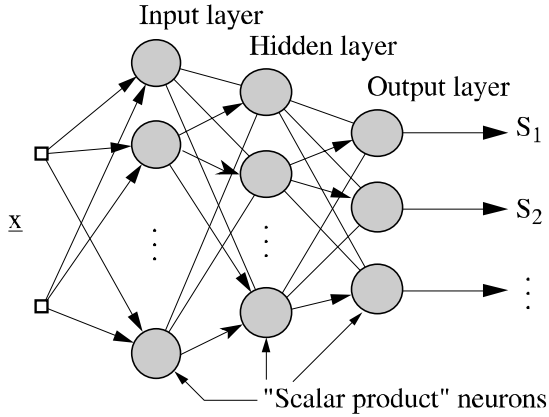


Figure 11.15. Quadratic separation in two dimensions

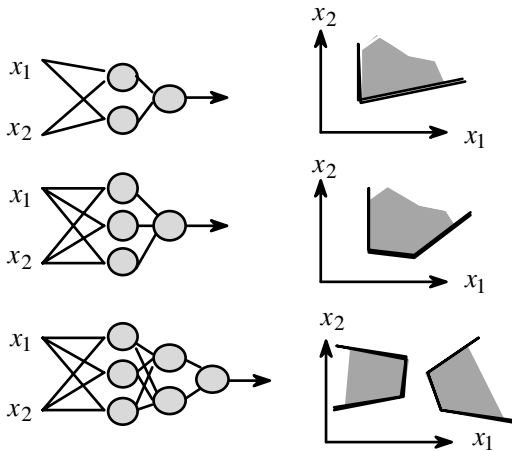
The space partitions remain in a rudimentary form when one neuron is introduced. More complex partitions necessarily occur when several neurons are linked in a network.

11.3.4.2. *Multilayered perceptrons*

The linkage most often used is called a multilayered perceptron. This structure combines the neurons of the “scalar product” into several interconnected layers, as shown in Figure 11.16.



**Figure 11.16.** Architecture of multilayer perceptrons



**Figure 11.17.** Partition examples in two dimensions with multilayer perceptrons

Regulating the weights of all the layers is done in supervised mode, thanks to the retropropagation algorithm of the gradient which uses the derivation of the prediction error of the outputs calculated on the group of the learning base, and propagated in layers from the output to the input [HEY 94].

The parameters of the network's architecture that require adjustment are the number of layers and the number of neurons per layer. This means that the space

partitions brought about by using these networks are no longer simple hyperplanes (2D lines), but much more complex forms.

Figure 11.17 gives examples in two dimensions of possible partition speeds with two or three layers and one output.

11.3.4.3. Radial base function networks

This second architecture, called a radial base function network, connects basic “distance” neurons to an input layer and the “scalar” product” neurons to an output layer (see Figure 11.18).

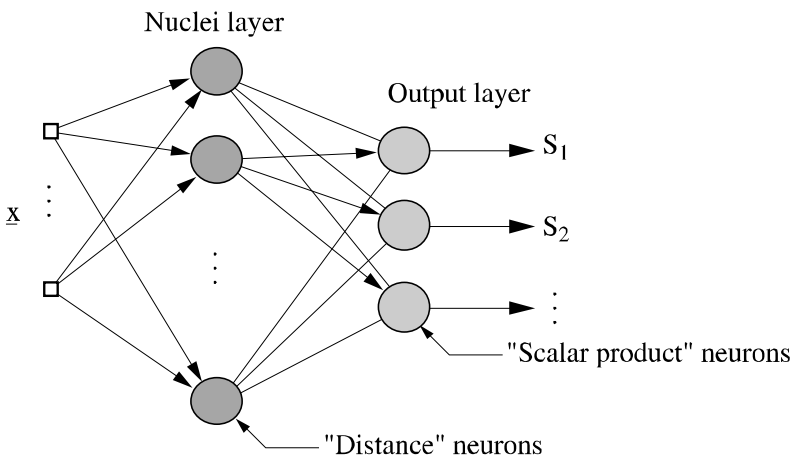
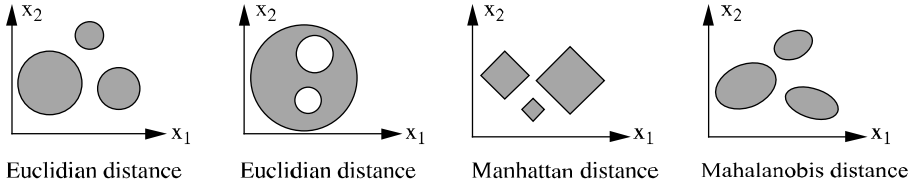


Figure 11.18. Architecture of a radial base function network

Adjusting this kind of network is more complicated than with multilayer perceptrons [OUK 99]. Adjustments must deal with the number and position of the nuclei of the first layer, the widths of the Gaussians linked to these nuclei, as well as the weights of the output layer. The type of distance must also be selected (Euclidian, Manhattan, Mahalanobis, among others).

Even with few nuclei, the partitions of the space are of widely varying speeds, depending both on the type of distance being used and the weights of the output layer. Figure 11.19 gives several examples of partitions that are possible with three nuclei.



**Figure 11.19.** Partition in two dimensions with radial base function networks

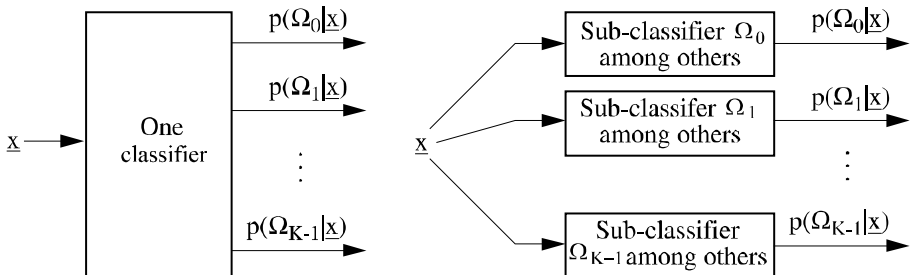
Choosing a type of neural network must obviously be done by considering the general form of the classes in the representation space, combining linear separations (see Figure 11.17) adapted to multilayer perceptrons, and combining quadratic forms (see Figure 11.19) that are adapted to radial base function networks.

#### 11.3.4.4. Neural networks and classification

Contrary to the Bayesian approaches, NNs used for classification purposes tend to *a posteriori* directly approach the probabilities of belonging to classes. At the moment of learning, the outputs representing the classification of base examples.

For situations in which there are more than two classes, two approaches are possible. Even if the network has as many outputs as there are classes, each output in the end represents the *a posteriori* probability of belonging to the class  $i$  knowing  $\underline{x}$ :  $p(\Omega_i|\underline{x})$ . This approach is called global classification. The retained class will be the one with the highest associated output.

Let us assume that we subdivide the global problem of  $K$  classes into two classes and construct sub-classifiers dedicated to each sub-problem. Several subdivision techniques are possible [PRI 94]; the simplest is shown in Figure 11.20, where  $K$  sub-classifiers are represented by the separation of each class among the  $K - 1$  others. In this simple subdivision instance, the retained class will also be the one with the highest output of the associated sub-classifier.



**Figure 11.20.** Global approaches and by partition of a  $K$  class classification

The partition approach is very useful in practice. Following the aphorism “divide and rule”, the complete problem is divided and an individualized adjustment of each sub-classifier is possible according to the relative difficulty of its sub-problem. Learning sub-classifiers are independent of each other, which often improves convergence speeds. This approach even allows personalized selection of input variables for each sub-classifier, since some variables prove to be relevant for the separation of a specific class [OUK 98].

We should remember that the partition approach can be used for all types of classifiers and not only for neural networks.

After having presented some supervised classification methods (Bayesian and non-Bayesian), in section 11.4 we will discuss some issues dealing with information fusion.

## 11.4. Data fusion

### 11.4.1. Introduction

Up to now, we have discussed information coming from sensors, and sensors themselves, without taking into account their intrinsic qualities. Statements such as “variable No. 1 is more certain than variable No. 2”, “it is more certain”, “sensor A breaks down more often than sensor B” express these intrinsic qualities.

The first objective of data fusion is to help manipulate this kind of knowledge, always taking into account the changing nature of a dynamic environment. Data fusion may be seen as a process which helps us integrate information coming from multiple sources so as to produce more specific and relevant data about an entity, an activity or an occurrence.

#### 11.4.1.1. *Modelizing imperfections and performances*

A key point here is the characterization and modelization of knowledge in terms of imprecision and incompleteness, as well as the level of the sensor and of the diagnostics.

Imprecision characterizes a quantity. “The security distance between two vehicles on the highway must be large” or “The distance must be around 200 meters to 100 km per hour” are both imprecise statements of information. A quantification of this imprecision on the content of the information, of the knowledge, or of the measurement is obviously necessary.

Uncertainty characterizes a quality. “The road might be icy” is a uncertain statement that qualifies the information in relation to its truth. Information can be both imprecise and uncertain as in the statement “The road may be icy about 10 km from here”.

Incompleteness characterizes the fact that knowledge, information or a fact is either not inaccessible or not especially accessible. The absence of a hygrometric measurement to estimate the presence of road ice or the masking of obstacles in a radar view are examples of incomplete information.

Finally, these are performances of an overall system that matter. Among them, reactivity plays a basic role in “real time” systems when the reaction time of the system to a change in the environment is a very significant parameter. Thus, with the car, a system that detects obstacles must be able to react to a pedestrian in its field of vision; that is, to provide information to the driver in less than 100 ms. The reliability of the system is also a factor of utmost importance. In our example we give an example of limiting or cancelling the false alarm rate.

How can we improve performances of these systems? In particular, how can we improve the reliability of information, the quality of the “cartography”, diagnostic and decision-making precision, all the time respecting cost and reactivity constraints?

#### 11.4.1.2. *Different fusion techniques and levels*

Several levels of fusion can be defined according to the nature and semantic of the information processed. This means that we would not process the values coming from a tachymeter the same as we would symbolic knowledge such as that relative to signalization panel for vehicles.

There are three main types of this fusion.

- low-level fusion of basic data. An example of this is image retiming;
- numerical fusion. This has to do with quantified data or information. An example of this is infrared laser telemetry used in radar;
- symbolical fusion, used in knowledge and forms of expertise. An example of this is the introduction of expert advice that has a strong added semantic value.

Whatever its level, the issue is to know what should be fused. What are the sensors or information sources? How these should be fused, and which techniques should be implemented?

Among data fusion techniques, we can very generally distinguish three large classes:

- standard probabilistic techniques (Bayesian fusion) and non-standard techniques (the evidence theory of Dempster-Shafer);
- least-squares techniques (PPV, PDAF, optimization);
- techniques that are not part of the above two (flow and principle possibilities theory).

All these theories (probabilistic, evidence theory, possibilities theory) modelize the knowledge of a source of information on a basic referential composed of the group of important hypotheses. The models that manipulate the data, as well as the fusion techniques used, depend on the type of imperfections that we need to take into account. This means the fusion will not be the same with imprecise data or with uncertain data; the formal scope will be different.

We point out two stages in fusional processing of data: the fusion itself and the decision. As with our discussion in earlier sections, here we will only speak of the decision or the diagnostic procedures concerning the hypotheses (class homologues) constructed on the data or measurements (signal homologues).

We will only discuss in detail three techniques of fusion that present a gradation in possible modelization of the imprecision and uncertainty [HEN 88]. These techniques also can be used in low level data processing as well as in symbolic processing.

### 11.4.2. *The standard probabilistic method*

Data fusion very often uses the Bayesian method of decision (see section 11.3.2). Here we use a well-established formalism that has theoretical and experimental advantages [PEA 88].

The problem consists of determining the configuration of the environment observed in  $K$  configurations (hypothesis or occurrences) possible. More exactly, we must decide on the most possible hypothesis, taking into account the imprecisions of different redundant and/or complementary measurements. Using the same equations as in the earlier sections, we write  $\Omega_i$  the  $i^{\text{th}}$  hypothesis among the  $K$  listed and  $\underline{x}$  the measurement or observation carried out.

#### 11.4.2.1. *Modelization, decision and hypothesis choice*

In a situation with one sensor, choosing a hypothesis is made by opting for one that maximizes its *a posteriori* probability  $p(\Omega_i|\underline{x})$  [11.16]. In other words, having observed the measurement  $\underline{x}$  provided by the sensor, we choose the most probable hypothesis – the one that gives the highest value of  $p(\Omega_i|\underline{x})$ . If we assume the group of

$K$  hypotheses to be exhaustive in providing configurations of the environment, we can use the Bayes formula (equation [11.15]), written again below, to calculate the  $p(\Omega_i|\underline{x})$

$$p(\Omega_i|\underline{x}) = \frac{p(\underline{x}|\Omega_i)p(\Omega_i)}{\sum_{j=0}^{K-1} p(\underline{x}|\Omega_j)p(\Omega_j)}$$

In this approach, we modelize the sensor by the probability  $p(\underline{x}|\Omega_i)$  that the observation equals  $\underline{x}$ , knowing that the hypothesis  $\Omega_i$  is verified. We then speak of the likelihood that the measurement  $\underline{x}$  is conditional to the hypothesis  $\Omega_i$ . A Gaussian distribution is often used to modelize the sensor's measurement.

For example, let us assume a detection system using an ultrasonic sensor with binary response, is triggered in the presence of obstacles in its field of vision:  $x = 0$  (no triggering) or  $x = 1$  (triggering). Assuming that we have two hypotheses:  $\Omega_0 =$  no obstacles,  $\Omega_1 =$  presence of an obstacle. So, let us lastly assume that the sensor has a false alarm rate of 1% ( $p(1|\Omega_0) = 0.01$ ) and a non-detection rate of 5% ( $p(0|\Omega_1) = 0.05$ ). The sensor will then be modelized by the matrix of the conditional probabilities shown in Table 11.1. The importance of choosing an example with two hypotheses with a measurement that can take only two states is that it allows us to calculate the group of related conditional probabilities.

$p(x \Omega_i)$	$\Omega_0$	$\Omega_1$
$x = 0$	0.99	0.05
$x = 1$	0.01	0.95

**Table 11.1.** Matrix of conditional probabilities  $p(x|\Omega_i)$

Let us now assume that an *a priori* knowledge of the environment in which the vehicle moves helps us to estimate the probability of the presence of an obstacle up to 0.1% ( $p(\Omega_1) = 0.001$ , thus  $p(\Omega_0) = 0.999$ ). Constructing the conditional probability matrix using the Bayes formula gives us Table 11.2.

$p(\Omega_i x)$	$\Omega_0$	$\Omega_1$
$x = 0$	<b>0.999</b>	$5.10^{-5}$
$x = 1$	<b>0.913</b>	0.087

**Table 11.2.** Conditional probability matrix  $p(\Omega_i|x)$



We then observe that the decision rule (equation [11.16]) which chooses the hypothesis giving the highest value of  $p(\Omega_i|x)$  leads to retaining, whatever the observation, the “non-obstacle” hypothesis! We see how Bayesian information fusion can modify this result.

11.4.2.2. *Multisensor Bayesian fusion*

Suppose we now have two independent sensors that verify:

$$p(x_1, x_2 | \Omega_i) = p(x_1 | \Omega_i) \times p(x_2 | \Omega_i)$$

The decision is made this time by choosing the hypothesis that *a posteriori* maximizes the conditional probability. We know that  $x_1$  and  $x_2$  are calculated with equation [11.22]. Under these conditions, the two sensors both have an information fusion coming from them both:

$$p(\Omega_i | x_1, x_2) = \frac{p(x_1 | \Omega_i) \times p(x_2 | \Omega_i) \times p(\Omega_i)}{\sum_{j=0}^{K-1} p(x_1 | \Omega_j) \times p(x_2 | \Omega_j) \times p(\Omega_j)} \tag{11.22}$$

Now we look at the above example but use two identical ultrasonic sensors. This time, we will obtain the conditional probability matrix shown in Table 11.3.

$p(\Omega_i x_1,x_2)$	$\Omega_0$	$\Omega_1$
$(x_1,x_2) = (0,0)$	<b>0.999</b>	0.001
$(x_1,x_2) = (0,1)$	<b>0.995</b>	0.005
$(x_1,x_2) = (1,0)$	<b>0.995</b>	0.005
$(x_1,x_2) = (1,1)$	0.1	<b>0.9</b>

**Table 11.3.** *Conditional probability matrix  $p(\Omega_i|x_1,x_2)$*

The decision rule shown in equation [11.16] chooses the same conclusion as before (no obstacle), except when there is concomitant triggering in the two sensors  $(x_1, x_2) = (1, 1)$ . The effective fusion of information of the two sensors modifies the result significantly.

Applying this method is simple but assumes:

- *a priori* knowledge of probabilities  $p(\Omega_i)$  that can be re-actualized during processing);

- statistical independence of the measurements provided by the sensors;
- knowledge of their probability model  $p(\underline{x}|\Omega_i)$ .

This poses a problem. The measurement of these probability laws is on one hand far from being easy, and on the other hand can lead to different results in a learning context than in a reality context. It also assumes that we have an exhaustive knowledge of the group of possible hypotheses.

If Bayesian fusion analyzes sensor imprecisions, it does not allow us to take into account their reliability or, more generally, the uncertainties in their model.

In addition, if we have a sensor incapable of distinguishing between two hypotheses  $\Omega_0$  and  $\Omega_1$ , its probabilistic model will be  $p(\underline{x}|\Omega_0 \text{ or } \Omega_1)$ . Using the Bayesian approach to fusion means equally distributing the probabilities between the two hypotheses and then choosing  $p(\underline{x}|\Omega_0) = p(\underline{x}|\Omega_1)$ . In conditional cases, only the probability values that are *a priori*  $p(\Omega_i)$  will decide the hypothesis choice, the sensor response no longer intervening in this choice. To avoid this difficulty, the Dempster-Shafer theory of evidence introduces masses that allow us to characterize the confidence degree of a measurement, of a model, of a hypothesis, and of a group of hypotheses.

### 11.4.3. A non-standard probabilistic method: the theory of evidence

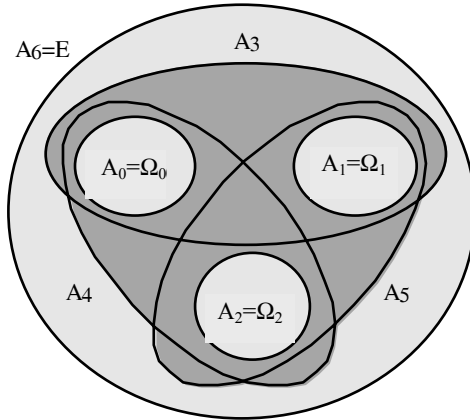
This theory is based on the work of A.P. Dempster and was formalized by G. Shafer [SHA 76]. It uses as a starting point the group E made of K hypotheses  $\Omega_i$  considered as exclusive and exhaustive. It also assumes that one of these hypotheses corresponds to each observed situation. The group E is called the *frame of discernment*.

#### 11.4.3.1. Mass sets of a source

To take into account the uncertainties of the information provided by a sensor and its possible inability to perceive certain hypotheses, we introduce the group  $2^E$  of all the possible combinations of the hypotheses, combinations made by the intermediary of the connector “or”, which is also written  $\cup$ . The group consists of  $2^K - 1$  elements  $A_i$ . Thus, with a group E of three hypotheses  $\Omega_0$ ,  $\Omega_1$  and  $\Omega_2$  (see Figure 11.21), we construct the following group  $2^E$ :

$$\begin{array}{lll}
 A_0 = \Omega_0 & A_1 = \Omega_1 & A_2 = \Omega_2 \\
 A_3 = \Omega_0 \cup \Omega_1 & A_4 = \Omega_0 \cup \Omega_2 & A_5 = \Omega_1 \cup \Omega_2 \\
 A_6 = E = \Omega_0 \cup \Omega_1 \cup \Omega_2 & & 
 \end{array}$$

$A_0$ ,  $A_1$  and  $A_2$  are called singletons.



**Figure 11.21.** Ensemble of hypothesis combination for a 3D discernment frame

The theory of evidence introduces a mass function  $m(\cdot)$  representing the likelihood of hypothesis combinations. This function is defined for the group  $2^E$  and is of value in the interval  $[0,1]$ . It verifies:

$$m(\emptyset) = 0 \text{ with } \emptyset \text{ the empty group}$$

$$\sum_{j=0}^{2^K-2} m(A_j) = 1 \tag{11.23}$$

In the previous example,  $m(A_3)$  will represent the likelihood that can be attributed to the hypothesis  $\Omega_0 \cup \Omega_1$ , without discernment possible between  $\Omega_0$  and  $\Omega_1$ .

We call elements *focal*,  $A_i$  having a non-zero mass. We say that if all the focal elements are singletons, then the function  $m(\cdot)$  corresponds to a probability.

Practically, the mass sets generated each time the sensor is tested easily allows us to express the indiscernability between hypotheses. When there are three hypotheses, if  $m(E) = 1$ , then  $m(A_0) = m(A_1) = m(A_2) = 0$ . This extreme situation expresses total uncertainty: the sensor is incapable of discerning between the different hypotheses of the discernment frame.

The major advantage of this theory is that it allows for conjoint evaluation of hypotheses, modelization of uncertainties and the indiscernability between hypotheses – all factors that the Bayesian theory cannot handle well. Uncertainty cannot be expressed in Bayesian theory, except by an equidistribution of probabilities for the different hypotheses.

#### 11.4.3.2. Example of mass set generation

Using the theoretical framework discussed above helps us quantify the related uncertainty to a sensor's functioning. For each measurement  $\underline{x}$  delivered, a distribution of masses, which take into account the uncertainty of its functioning, will be represented. Later in this chapter, we will systematically express equations in one dimension to simplify them ( $\underline{x} = x$ ). Obviously, all the results shown can be extended, for a given representation space, into any dimension.

Let us look at a binary response sensor  $C_1$ ,  $x = 0$  (no triggering) or  $x = 1$  (triggering) in the context of two hypotheses:  $\Omega_0 =$  no obstacle,  $\Omega_1 =$  obstacle. The sensor is assumed to function perfectly. When  $x = 1$ , we use, for example, the following mass set:

$$m_{C_1}(\Omega_0) = 0.2 \quad m_{C_1}(\Omega_1) = 0.8 \quad m_{C_1}(\Omega_0 \cup \Omega_1) = 0$$

The value  $m_{C_1}(\Omega_0 \cup \Omega_1) = 0$  characterizes the certainty as to the sensor's  $C_1$  functioning. The distribution of the two other masses and  $m_{C_1}(\Omega_0)$  et  $m_{C_1}(\Omega_1)$  characterize the precision of the measurement  $x = 1$ .

Let us now assume a second sensor  $C_2$  whose functioning is less reliable. We introduce, to characterize this uncertainty, a coefficient  $\alpha$  to the interval value  $[0, 1]$ . We then attribute to the mass  $m_{C_2}(\Omega_0 \cup \Omega_1)$  the value  $1 - \alpha$ . With  $\alpha = 0.9$  and still using the example of a response  $x = 1$ , we will use the following mass set:

$$m_{C_2}(\Omega_0) = 0.18 \quad m_{C_2}(\Omega_1) = 0.72 \quad m_{C_2}(\Omega_0 \cup \Omega_1) = 0.1$$

The value  $m_{C_2}(\Omega_0 \cup \Omega_1) = 0.1$  characterizes the uncertainty of the sensor's functioning. The distribution of the two other masses  $m_{C_2}(\Omega_0)$   $m_{C_2}(\Omega_1)$  remains, in our example, in the same proportions as before.

It is important to remember that the mass sets are not fixed entities and they can be changed at each new measurement.

The fusion itself of the mass sets coming from several sensors will be discussed in section 11.4.3.4.

We can also adjust the modelization of a sensor of its functioning is assumed to be certain in a given environment and uncertain in another environment. This means we can know perfectly the conditional distribution function  $p(x|\Omega_i)$  of the measurements in the framework of hypothesis  $\Omega_0$  and imperfectly in the framework of  $\Omega_1$ . We then construct two different mass sets following the underlying hypothesis [APR 91]. These mass sets can also be fused.

11.4.3.3. *Credibility and plausibility*

For a given mass set, we should bear in mind that  $m(A_i)$  does not express the likelihood of  $A_i$  particularly well. The group of elements of  $2^E$  included in  $A_i$  also contributes to our knowledge of  $A_i$ . In Figure 11.21, we see that the likelihood of  $A_5$  must take into account the mass  $m(A_5)$  but also the masses  $m(A_1)$  and  $m(A_3)$ , since  $A_1 \subset A_5$  and  $A_3 \subset A_5$ . For that, we introduce the *credibility* function defined on the group  $2^E$  and of interval value  $[0, 1]$ , such as:

$$Cr(A) = \sum_{B \subseteq A} m(B) \tag{11.24}$$

The credibility expresses the minimum likelihood veracity of hypothesis  $A$ .

On the contrary, it is good to have an idea of the maximum likelihood that is possible for the veracity of hypothesis  $A$ . For that, we introduce the *plausibility* function represented for the group  $2^E$  of interval value  $[0, 1]$ , such that:

$$Pl(A) = 1 - Cr(\neg A) = \sum_{B \cap A \neq \emptyset} m(B) \tag{11.25}$$

where  $\neg A$  represents the complementary group of  $A$  for the group  $2^E$ . We see that if the only elements  $A_i$  having a non-zero mass (focal elements) are the singletons, then:

$$Cr(\Omega_i) = Pl(\Omega_i) = m(\Omega_i) = p(\Omega_i)$$

Finally, we should remember that generating a mass set linked to an information source is a complex task, requiring *a priori* knowledge of the nature of the source and its environment. This generation should take into account uncertainties and the underlying hypotheses of the context, as discussed in section 11.4.3.2.

11.4.3.4. *Fusion of mass sets*

Fusion by the Dempster-Shaffer combination rule [11.26] allows us to reinforce the increase associated with hypotheses whose sources are in accord, and to attenuate the increases associated with hypotheses whose sources are in disaccord. If

we have two sources, and thus the mass sets  $m_{C_1}$  and  $m_{C_2}$ , the fusion is expressed by the following rule:

$$m(A_k) = m_{C_1}(A_k) \oplus m_{C_2}(A_k) = \frac{1}{1-H} \sum_{A_i \cap A_j = A_k} m_{C_1}(A_i) m_{C_2}(A_j) \quad [11.26]$$

$1/(1-H)$  represents a normalization coefficient in which  $H$  characterizes the degree of conflict between the sources and is expressed by the relation:

$$H = \sum_{A_i \cap A_j = \emptyset} m_{C_1}(A_i) m_{C_2}(A_j)$$

$H = 0$  shows that the two sources are never in contradiction. This normalization allows the mass set obtained by fusion to verify the definition [11.23]. This rule generalizes when there are several or many sources and/or sensors, always respecting the properties of commutability and associability that are indispensable to all fusion mechanisms (see [JAN 96] and [SHA 76]).

Below we will define the decision rule constructed from the fused mass set.

#### 11.4.3.5. Decision rule

The decision phase is based on the definitions given in section 11.4.3.3. We will retain the singleton hypothesis  $\Omega_1$  as the most plausible among the  $K$  singletons of the discernment frame:

$$\Omega_i \text{ such that } Pl(\Omega_i) = \max_{k=0 \dots K-1} \{Pl(\Omega_k)\} \quad [11.27]$$

#### 11.4.3.6. Example

Let us look again at the example of the two binary response sensors working in the framework of two hypotheses  $\Omega_0 =$  no obstacles;  $\Omega_1 =$  an obstacle. Assuming that they independently analyze the same scene with different modalities, one of the sensors is ultrasonic, the other infrared. The ultrasonic sensor is assumed to have reliable functioning. The infrared sensor's functioning is assumed to be less well known ( $\alpha \neq 1$ ). Once the mass set and response functions of the sensors are set up, and taking into account the value attributed to  $\alpha$  for the infrared sensor, it is possible to fuse the sets coming from the two sensors and to make a decision by retaining the most possible hypothesis. Assuming that the ultrasonic sensor provides the response  $x_{us} = 1$  (an obstacle is present). We then construct the mass set:

$$m_{us}(\Omega_0) = 0.2 \quad m_{us}(\Omega_1) = 0.8 \quad m_{us}(\Omega_0 \cup \Omega_1) = 0$$

We assume that the infrared sensor provides the response  $x_{ir} = 0$  (no obstacle) and that the reliability of its functioning is characterized by  $\alpha = 0.9$ . We then construct the following mass set:

$$m_{ir}(\Omega_0) = 0.72 \quad m_{ir}(\Omega_1) = 0.18 \quad m_{ir}(\Omega_0 \cup \Omega_1) = 0.1$$

The two sensors give contradictory results but the mass fusion allows us to decide. Applying the rule in [11.26] and the representation [11.25] provides the result given in Table 11.4 that chooses the “presence of an obstacle” hypothesis.

	$\Omega_0$	$\Omega_1$	$\Omega_0 \cup \Omega_1$
US $x_{us} = 1$	0.2	0.8	0
IR $x_{ir} = 0$	0.72	0.18	0.1
Fusion	0.42	0.58	0
Plausibility	0.42	<b>0.58</b>	0

**Table 11.4.** *Specific example of the fusion of two mass sets*

We see that the mass set of the infrared sensor with uncertainties (and which responds to “no obstacle”) cannot counterbalance the response of the ultrasonic sensor. We also see that, in this simple situation, plausibility is directly equal to the mass set resulting from fusion.

Tables 11.5 and 11.6 show the results of the fusion obtained for the two previous sensors in all possible response configurations.

$m_{us} / m_{ir}$	$\Omega_0$	$\Omega_1$	$\Omega_0 \cup \Omega_1$
$x = 0$	0.75 / 0.72	0.25 / 0.18	0 / 0.1
$x = 1$	0.2 / 0.2	0.8 / 0.7	0 / 0.1

**Table 11.5.** *Mass sets of two sensors*

$m_{us} \oplus m_{ir} = Pl$	$\Omega_0$	$\Omega_1$	$\Omega_0 \cup \Omega_1$
$(x_{us}, x_{ir}) = (0,0)$	<b>0.9</b>	0.1	0
$(x_{us}, x_{ir}) = (0,1)$	<b>0.53</b>	0.47	0
$(x_{us}, x_{ir}) = (1,0)$	0.42	<b>0.58</b>	0
$(x_{us}, x_{ir}) = (1,1)$	0.09	<b>0.91</b>	0

**Table 11.6.** *Fusion of two mass sets*

We notice that in cases where there is contradiction, the ultrasonic sensor does not impose its choice because of the distributions of the masses on singletons.

Simple situations with two hypotheses, shown in the examples, imperfectly illustrate the complexity of the credibility-plausibility approach. In this theoretic framework, working out more complex cases (in terms of number of hypotheses) helps us imagine very elaborate decision strategies that can lead to rapid choices of non-singleton elements of the hypotheses' spaces.

An extension of these fusion methods is also possible with a non-exhaustive discernment frame [JAN 96]. A possible problem is the constraint on the mass sum, which must be equal to 1; this can be the case when we have very little information from the source.

To open up the range of fusion possibilities by limiting constraints, we can, while still using the methods described above, use the theory of fuzzy groups and extend it to fusion: the theory of possibilities.

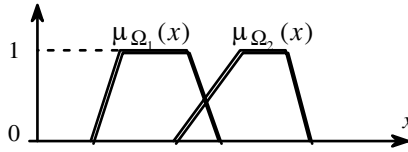
#### **11.4.4. *Non-probabilistic method: the theory of possibilities***

The theory of possibilities was introduced by D. Dubois and H. Prade [DUB 87]. It is based on the fuzzy subgroups of L.A. Zadeh [ZAD 65] [ZAD 78]. It allows us to leave the probabilistic framework of many fusion approaches. It is particularly well-adapted to a situation where we know very little about the information sources (sensors, experts, etc.).

Modelization of imprecise or uncertain information is based on the functions represented in the observation space and the interval value  $[0, 1]$ . These are called ownership functions; they express the available knowledge of the different propositions or hypotheses of the discernment frame.

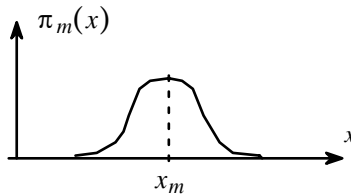
an ownership function  $\mu_{\Omega_i}(x)$  will be linked to the hypothesis  $\Omega_i$ . This function is represented on the group of values possible for the variables. The value  $\mu_{\Omega_i}(x_1)$  characterizes the veracity of the hypothesis  $\Omega_i$  for the observation  $x_1$ . Figure 11.22 shows, in a 1D case, examples of ownership function for two hypotheses.





**Figure 11.22.** *Examples of ownership functions*

The imperfections of a sensor and how well the person performing the experiment knows the sensors can also be modeled by functions of the same type. Here, we speak of a distribution of possibilities. We assume that a sensor delivers the measurement  $x_m$  of a real variable  $x_r$ . The information we have about the value of the measured variable is expressed by a distribution of possibilities written as  $\pi_m(x)$ : that the variable is equal to  $x$ , knowing that the measurement is  $x_m$  (see Figure 11.23).



**Figure 11.23.** *Distribution of possibilities linked to a measurement*

A completely imprecise sensor is expressed by:  $\pi_m(x) = 1 \forall x \in \mathbb{R}$ . The existence of an ideal sensor and the knowledge of this perfection is expressed by the distribution of possibilities:  $\pi_m(x_m) = 1$  and  $\pi_m(x) = 0 \forall x \neq x_m$ .

The analogy with a probability  $p(x|x_m)$  is only graphic, since no constraint is imposed on the probability distribution.

11.4.4.1. *Operations on ownership functions and possibility distributions*

We can define operations on functions  $\mu_{\Omega}(x)$  or  $\pi_m(x)$  such as inclusion, intersection, union and complementing. This helps express logical operations on hypotheses or on information relative to variables. The “and” and the “or” or their corresponding symbols  $\cup$  and  $\cap$  can be expressed, for example, by [11.28]:

$$\begin{aligned}
 \mu_{A \cup B}(x) &= \max(\mu_A(x), \mu_B(x)) \\
 \mu_{A \cap B}(x) &= \min(\mu_A(x), \mu_B(x)) \\
 \mu_{\neg A}(x) &= 1 - \mu_A(x)
 \end{aligned}
 \tag{11.28}$$

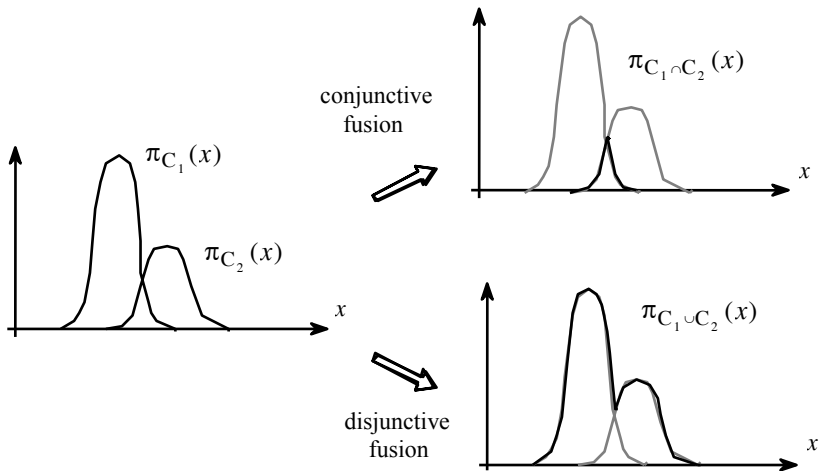
#### 11.4.4.2. Possibilistic multisensor fusion

Let us look at a situation where there are reliable but imprecise sensors all observing the same variable. Each will be modeled by a distribution of possibilities relative to this variable. The fusion of information will then take the form of an intersection of distributions of possibilities, using the operator  $\cap$ .

If these sensors are not completely reliable, especially if the information they provide is discordant, we can take this discordance and uncertainty of this information into account by using the conjunction of distributions using the operator  $\cup$ .

Figure 11.24 illustrates these two types of fusion when there are two sensors.  $\pi_{C_1}(x)$  and  $\pi_{C_2}(x)$  are the distributions of possibilities of the values of the variable  $x$  observed by two sensors  $C_1$  and  $C_2$ .

All intermediary forms of fusion are certainly possible [BLO 96], which justifies sophisticated supervision techniques, according to a *a priori* knowledge.



**Figure 11.24.** Possibilistic multisensor fusion

As for hypotheses, the same types of fusion can be used. This means that the ownership function  $\mu_{\Omega_1 \cup \Omega_2}(x)$  characterizes the veracity of the hypothesis  $\Omega_1 \cup \Omega_2$ .

#### 11.4.4.3. Diagnostics and fusion

From a distribution of possibilities, obtained from a sensor or from a fusion operation between sensors, it is possible to characterize the information we have

about a variable. To do this, we introduce two functions, the possibility degree  $\Pi$  and the necessity degree  $N$  represented on a subgroup  $S$  of the possible variable values:

$$\begin{aligned} \Pi(S) &= \text{Sup}_{x \in S} (\pi(x)) \\ N(S) &= \text{Sup}_{x \in S} (1 - \pi(x)) = 1 - \Pi(\neg S) \end{aligned} \tag{11.29}$$

These functions evaluate the confidence we have in the statement: “The variable has a value belonging to the group  $S$ .”

If now a hypothesis  $\Omega_i$  can be expressed by an ownership function, it is easy to construct, by extension, a possibility degree linked this time to a hypothesis  $\Omega_i$ , knowing the distribution of possibilities of the variable. Formula [11.30] shows the calculation of  $\Pi(\Omega_i)$ .

$$\Pi(\Omega_i) = \text{Sup}_x \left\{ \min(\mu_{\Omega_i}(x), \pi_m(x)) \right\} \tag{11.30}$$

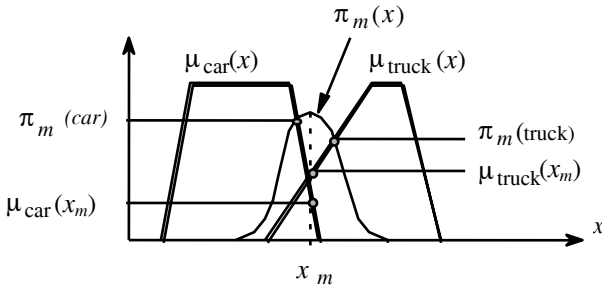
If  $\Pi(\Omega_i)$  is obtained from a possibilistic multisensor fusion, its value evaluates the confidence that we have in the veracity of hypothesis  $\Omega_i$ , taking into account the measurements carried out and fused.

A diagnostic operation is thus possible by discriminating the hypothesis corresponding to the maximum possibility degree:

$$\Omega_i \text{ such that } \Pi(\Omega_i) = \text{Sup}_{k=0 \dots K-1} \{ \Pi(\Omega_k) \} \tag{11.31}$$

To illustrate this, let us imagine an obstacle detection situation in a road or highway. We assume these obstacles are of two types: “car” or “truck”. To validate one or the other of these hypotheses, a sensor (or a fusion of sensors) gives us a measurement (for example, of volume of vehicles). A distribution of possibilities will be linked to the measurement; we express it as:  $\pi_m(x)$ . In addition, our two classes have their own ownership functions:  $\mu_{\text{truck}}(x)$  and  $\mu_{\text{car}}(x)$ .

We can evaluate the degree of possibility of each hypothesis (truck or car) and finally decide by using the rule of maximum possibilities (equation [11.31]). Figure 11.25 shows this mechanism.



**Figure 11.25.** Example of fusion and possibilistic diagnostics

We should observe that, on the figure, the maximum possibility degree rule gives a “car” result different from the “truck” result which would have been obtained from the raw measurement  $x_m$  by using as a decision rule the maximum of the ownership function. So:

$$\pi_m(car) > \pi_m(truck) \quad \text{while} \quad \mu_{truck}(x_m) > \mu_{car}(x_m)$$

Probabilistic fusion allows us to analyze the measurements provided by sensors, cameras [DEV 94], and temporal information about sequences of events [NIF 97] directly. The advantage of this theoretical framework is that it gives us many choices among fusion mechanisms that can lead to supervised control on a hierarchic level of the symbolic type [DUB 97] [ZAD 92]. The drawback is in the difficulty of mastering the obtained results. This means that fusion laws cannot be associative, cannot even be used in cases where there are more than two sources. Decisions coming from them may not agree with the operator’s intuition.

### 11.4.5. Conclusion

Data fusion can take various forms. The reader will find a more thorough coverage of different fusion techniques in [ABI 92], [TS 94] and [TS 97].

Generally, we can say that if only imprecision has to be controlled, we can choose a Bayesian fusion. If fusion must deal with both uncertainties and imprecisions, but if we have information of a symbolic nature, we can choose the theory of evidence.

On the other hand, if the data are numeric, we choose possibilistic fusion; in particular if we do not have a reliable probabilistic modelization of the sensors.

Constraints are relaxed when we go from Bayesian fusion to theory of evidence fusion, then to theory of possibilities fusion, but it can also lead to systems in which it is difficult to regulate parameters.

### 11.5. General conclusion

This chapter has presented some of the aspects of diagnostic problems with the help of multisensor systems: choosing a representation space of signals and in particular the reduction of its dimension; Bayesian and non-Bayesian classification techniques; and fusion of probabilistic and possibilistic data.

All these topics have been extensively researched and developed. The recent refinement of connectionist techniques, logical flow and the theory of evidence, to name several, have all made their contribution.

We have also noticed a recent convergence of interests, even of methods, on the part of a scientific community that had previously been divided into areas of specialization such as data analysis (in statistics), form recognition (in artificial intelligence) and neural networks (in signal processing). This synergy has resulted in many new theoretical and practical developments such as vocal and cursive handwriting recognition, aids to detecting driving problems, analysis of malfunctions in industrial settings, and medical imagery.

A certain number of problems encountered during the implementation of a diagnostic system can be mentioned here. Other chapters of this book have covered themes strictly related to those covered in this one. So, the geographic distribution of these sensors, whether the system being used is in an industrial workshop or a system that perceives specific targets, poses the problem of choosing software and material architecture. It can be centralized or distributed. If distributed, we need sensors that integrate an “intelligence” that helps them make local diagnostics. Chapter 12 will discuss these problems.

### 11.6. Bibliography

- [ABI 92] ABIDI M.A., GONZALEZ R.C., *Data Fusion in Robotics and Machine Intelligences*, Academic Press, 1992.
- [APR 91] APPRIOU A., “Probabilités et incertitudes en fusion de données multi-senseurs”, *Revue Scientifique et Technique de la Défense*, no. 11, p. 27-40, 1991.
- [BEL 61] BELLMAN R., *Adaptive Control Process: A Guided Tour*, Princeton University Press, 1961.
- [BIS 95] BISCHOP C.M., *Neural Networks for Pattern Recognition*, Clarendon Press, 1995.

- [BLO 96] BLOCH I., "Information Combination Operators for Data Fusion: A comparative Review with Classification", *IEEE Trans on SMC*, vol. 26, no. 1, 1996.
- [CHE 89] CHEN S., BILLINGS S.A., LUO W., "Orthogonal least squares methods and their application to non-linear system identification", *Int J. Control*, vol. 50, no. 5, p. 1873-1896, 1989.
- [DEV 94] DEVEUGHELE S., DUBUISSON B., "Adaptabilité et combinaison possibiliste: application à la vision multicaméra", *Revue Traitement du Signal*, vol. 11, no. 6, p. 559-568, 1994.
- [DOC 81] DOCTOR P.J., HARRINGTON T.P., DAVIS T.J., MORRIS C.J., FRALEY D.W., "Pattern recognition methods for classifying and sizing flaws using eddy current data", *Eddy Current Characterization of Materials and Structures*, ASTM, p. 464-483, 1981.
- [DUB 87] DUBOIS D., PRADE., "Une approche ensembliste de la combinaison d'informations imprécises ou incertaines", *Revue d'intelligence artificielle*, vol. 1, no. 4, p. 23-42, 1987.
- [DUB 90] DUBUISSON B., *Diagnostic et reconnaissance des formes*, Hermès, 1990.
- [DUB 97] DUBOIS D., PRADE H., YAGER R.R., *et al.*, *Fuzzy Information Engineering*, Wiley, 1997.
- [DUD 73] DUDA R., HART P., *Pattern Recognition and Scene Analysis*, Wiley & sons, 1973.
- [FUK 72] FUKUNAGA K., *Introduction to Statistical Pattern Recognition*, Academic Press, 1972.
- [GAI 83] GAILLAT G., *Méthodes statistiques de reconnaissance de formes*, Cours ENSTA, 1983.
- [GUE 88] GUEGEN A., NAKACHE J.P., "Méthode de discrimination basée sur la construction d'un arbre de décision binaire", *Revue Stat. Appl.*, 36 (1), p. 19-38, 1988.
- [HAY 94] HAYKIN S., *Neural networks, A Comprehensive Foundation*, Prentice Hall, 1994.
- [HEN 88] HENKIND S.J., HARISSON M.C., "An Analysis of Four Uncertainty Calculi", *IEEE Trans SMC*, vol. 18, no. 5, p. 700-714, 1988.
- [HEY 94] HEYRAULT J., JUTTEN C., *Réseaux neuronaux et traitement du signal*, Hermès, 1994.
- [JAN 96] JANEZ F., APPRIOU A., "Théorie de l'évidence et cadres de discernement non exhaustifs", *Revue Traitement du Signal*, vol. 13, no. 3, 1996.
- [KIT 86] KITTLER J., *Feature Selection and Extraction. Handbook of Pattern Recognition and Image Processing*, Academic Press, 1986.
- [KRI 82] KRIHNAIAH *et al.*, "Classification, pattern recognition and reduction of dimensionality", *Handbook of Statistics*, vol. 2, North-Holland, 1982.
- [LEB 97] LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimensionnelle*, Dunod, 1997.

- [NIF 97] NIFLE A., REYNAUD R., “Classification des comportements fondée sur l’occurrence d’événements en théorie des possibilités”, *Revue Traitement du Signal*, vol. 14, no. 5, p. 523-533, 1997.
- [OUK 98] OUKHELLOU L., AKNIN P., “Optimisation de l’espace de représentation dans un problème de classification par réseaux de neurons”, *Journal Européen des Systèmes Automatisés*, vol. 32, no. 7-8, p. 915-938, 1998.
- [OUK 99] OUKHELLOU L., AKNIN P., “Hybrid training of radial basis function networks in a partitioning context of classification”, *Neurocomputing*, vol. 28, no. 1-3, p. 165-175, 1999.
- [PAR 62] PARZEN E., “On the estimation of a probability density function and mode”, *Ann. Math. Stat.*, vol. 33, p. 1065-1076, 1962.
- [PEA 88] PEARL J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, 1988.
- [PRI 94] PRICE D., KNEER S., PERSONNAZ L., DREYFUS G., *Pairwise neural network classifiers with probabilistic outputs*, Neural Information Processing Systems, 1994.
- [RUM 86] RUMELHART D.E., HINTON G.E., WILLIAMS R.J., “Learning internal representations by error propagation”, *Parallel Distributed Processing: Explorations in Microstructure of Cognition*, vol. 1, p. 318-362, MIT Press, 1986.
- [SAP 90] SAPORTA G., *Probabilités, Analyse des données et Statistique*, Technip, 1990.
- [SHA 76] SHAFER G., *Mathematical Theory of Evidence*, Princeton University Press, 1976.
- [STO 97] STOPPIGLIA H., Méthodes statistiques de sélection de modèles neuronaux, PhD Thesis, University of Paris VI, 1997.
- [THE 99] THEODORIDIS S., KOUTROUMBAS K., *Pattern Recognition*, Academic Press, 1999.
- [TS 94] *Revue Traitement du Signal, numéro spécial Fusion de données*, vol. 11, no. 6, 1994.
- [TS 97] *Revue Traitement du Signal, numéro spécial Fusion de données*, vol. 14, no. 5, 1997.
- [ZAD 65] ZADEH L.A., “Fuzzy sets”, *Information and Control*, vol. 8, p. 338-353, 1965.
- [ZAD 78] ZADEH L.A., “Fuzzy sets as a basis for a theory of possibility”, *Fuzzy sets and systems*, vol. 1, p. 3-28, North-Holland Publishing, 1978.
- [ZAD 92] ZADEH L.A., KACPRZYK J., *Fuzzy Logic for the management of uncertainty*, Wiley, 1992.
- [ZWI 95] ZWINGELSTEIN G., *Diagnostic des défaillances*, Hermès, 1995.

## Chapter 12

# Intelligent Sensors

### 12.1. Introduction

Since the end of the 1980s, many articles have appeared in scientific [BER 87], [GIA 86] and technical [BLA 87], [JOR 87] books using the term “intelligent”, a word often associated with sensors, transmitters, actuators and instrumentation. In this chapter, we will define an intelligent sensor as the linkage of one or several measurement chains to a computing “machine”. Its main function is providing reliable, useful information. However, sometimes we read that the sensor is the weakest link in the measurement chain.

As shown in Figure 12.1, we remember that the purpose of a sensor is essentially to provide information, that is, to measure a variable and to communicate representative information of the value of this variable. A sensor is always part of an action/reaction loop:

– This action/reaction loop may be “closed” in the technological sense through a certain number of components such as controllers, calculators, Programmable Logic Controllers (PLC) or actuators. The sensor is then an element in an automation loop that provides a representative signal of a physical variable: for example, it provides a signal to a regulator or to a PLC that will itself provide a command signal to be transmitted to an actuator. In the automotive field, a typical example is that of a rotation speed sensor placed on each of the wheels which transmits this rotation “speed” information to the central Anti-Breaking System (ABS), which makes the “right” decision by comparing the rotation speeds of the same axle.



– This action/reaction loop could also be closed by a human operator. Still in the automotive context, it is the driver who usually observes the speedometer and either brakes, accelerates or reduces gear to adjust the speed of the car to the speed limits. Obviously in this context the speed sensor must be precise and reliable.

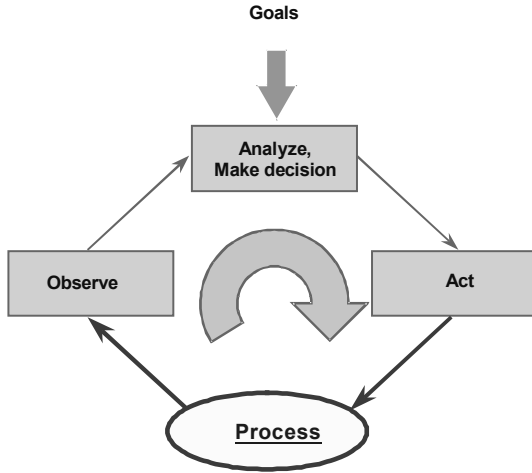


Figure 12.1. Sensor in an action/reaction loop

## 12.2. Users’ needs and technological benefits of sensors

Whatever the context, sensor users want products that perform well; that is, sensors that are reliable from the first performance or, to be more technologically exact, that are accurate<sup>1</sup>. In complementing this accuracy, it is more particularly the credibility<sup>2</sup> of the information which is required by the users of the systems. This is especially true in the automotive context, with users who tend to have absolute confidence in their vehicles and would be angered by erroneous warnings generated by a faulty surveillance system.

Together with this need for accurate measurements, integration strategies require mutual exchanges of information between the automation units. The development of these techniques and automation methods has led to less centralized processing, made possible by new communication networks. These networks are called

---

1 Accuracy of a measurement instrument: the ability of the measurement instrument to give close indications of the true value of a measured variable (ISO norm, NF X07-001, December, 1984, International Organization for Standardization).

2 The quality, capability or power to elicit belief (OED).

fieldbuses [CIA 99] and the appearance, by the 1980s, of “smart” or “intelligent” equipment.

In industry, current strategies aim to improve the quality of this equipment by increasing reliability and improving productivity. Automation is part of this improvement and requires the collection of relevant data to be used and processed by a production system. This data must be conveyed to decision-making units which then undertake a range of actions that culminate in predefined objectives. Measurements, which are the result of a collection of information, are thus inherently linked to automation, which itself integrates the following procedures:

- automation integrates the control-command features, which provide “real time” information about the value of strategic variables that are indispensable to production. These variables qualitatively or quantitatively describe the product or they provide information about the state of the production system;
- automation integrates the maintenance; in this area, the measurements detect or anticipate any deterioration in the system that might adversely affect its functioning;
- automation coordinates safety procedures, usually cross-integrated, that protect operators, equipment, and the environment;
- automation regulates production, helps establish guidelines and expected results, and helps adjust product and energy flow;
- automation improves technical management, giving information about processing availability.

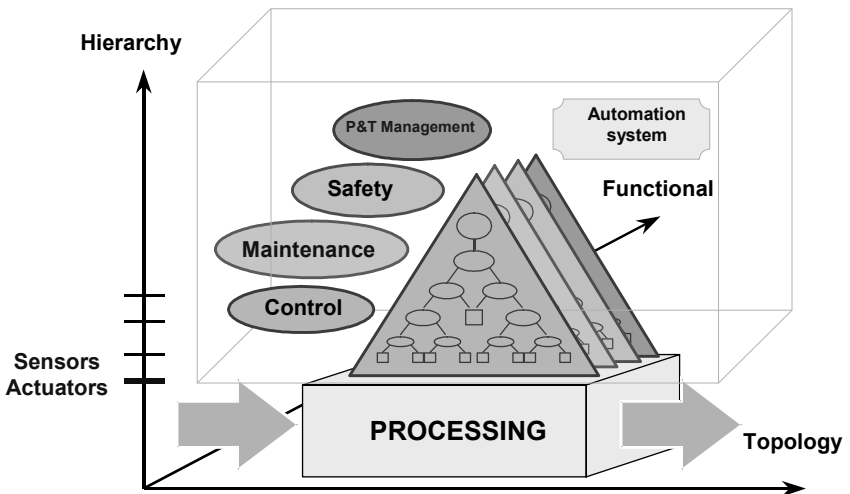


Figure 12.2. Four automation units (DEL 90)

In the automotive context, automation is used in the following instruments or systems:

- it is used in the command-control of a car. It coordinates everything to do with speed regulation, since the car must be able to move safely. It also guides the vehicle and regulates the interior temperature;

- automation helps maintain the car, integrating diagnostics, on-off engine functions, engine temperature, tire pressure and compression measurements, among other functions;

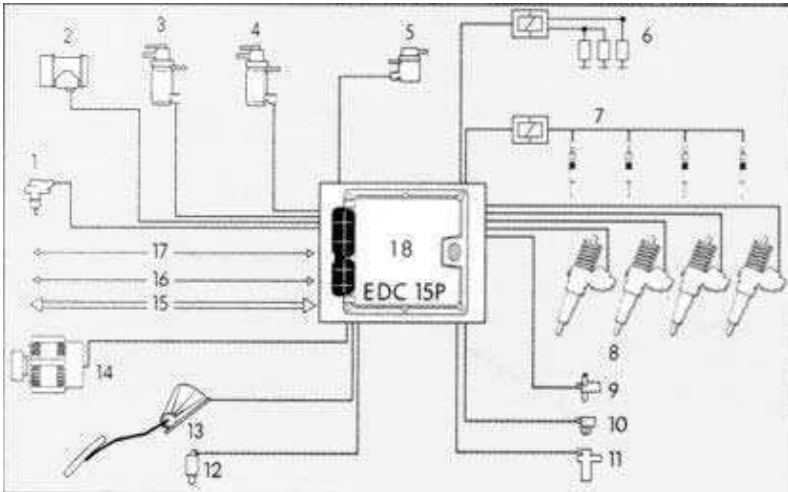
- automation integrates safety functions, including the essential function of protecting drivers and passengers. This includes inflatable airbags, ethylometers, which can stop a driver from driving under the potentially dangerous influence of alcohol, as well as anti-theft systems;

- automation coordinates data concerning fuel consumption. This data usually tells the driver how much fuel is available and, indirectly, the distance that can be covered with this fuel. This kind of information comes from a gas tank sensor that determines from the volume of fuel available, taking into account the geometry of the tank and conditions in which the vehicle is being used. These factors are provisionally parameterized by conditions such as urban or highway driving that an intelligent sensor can provide, given the distance to be traveled. The information the sensor provides is in an adapted form the driver can use;<sup>3</sup>

- automation coordinates technical regulation of the system itself; there are sensors that give information about tires and their condition. The company Continental Teves has developed a system called Sidewall Torsion Sensor (SWT) that has produced Smart Tyres [CRO 99].

---

<sup>3</sup> We will come back to this example of “intelligent distance sensors” after discussing the concept of “intelligent sensors”.



**Figure 12.3.** Schema principle for the direct injection system with an injection pump

- 1) Intake pressure sensor
- 2) Airflow meter
- 3) Regulated control of supercharging pressure
- 4) Control for gas recirculation system (EGR)
- 5) Control of supercharging cut-off
- 6) Reheating of cooling system
- 7) Preheating spark plugs
- 8) Pump injectors
- 9) Fuel temperature sensor
- 10) Brake camshaft sensor
- 11) Crankshaft sensor
- 12) Speed sensor
- 13) Pedal accelerator sensor
- 14) Alternator
- 15) Multiplexer circuit
- 16) Diagnostics connection
- 17) Anti-theft system connection
- 18) Electronic control-command unit engine unit

We see that the car can be described as a system and also can be broken down into subsystems; the motor control and anti-lock brake systems are themselves systems. In looking at the example of Figure 12.3, taken from [GUY 99], we see there are no less than seven sensors that provide information to the injection system, which, in this particular case, leads to an increase of 21.3% of the engine torque to 1,900 tr/mm.

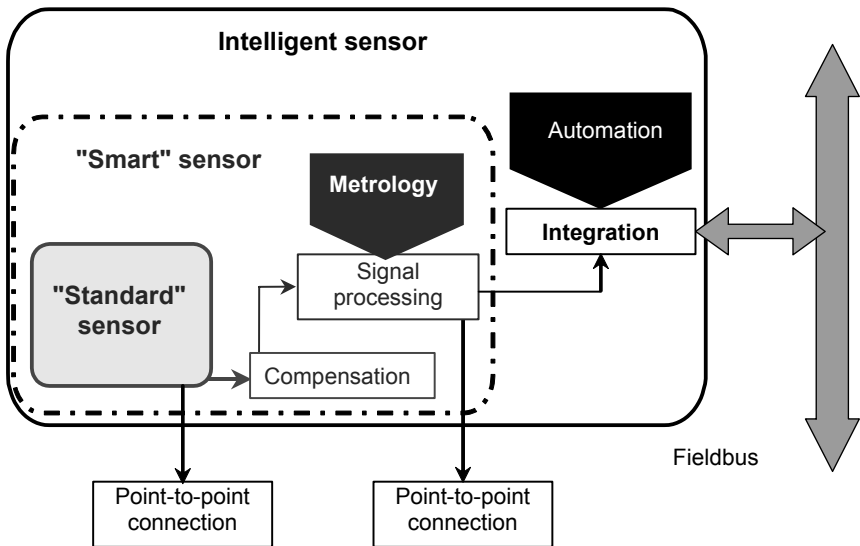
**12.2.1. A short history of smart sensors**

The first “smart” sensors were developed by internationally well-known manufacturers like Honeywell, Fuji and Control Bailey, who had in a sense already developed precursors to these sensors, at least in terms of concept. These first-generation smart sensors were most often dedicated to numerical control-command systems. In France, work on users’ needs was gathered in 1987 within the publication of a *White Book* [CIA 87], and completed by a census of services proposed by intelligent sensors in automated production systems [ROB 93].

**12.2.2. Smart or intelligent?**

We are basing our ideas on those in [KLE 91] in distinguishing between a *smart sensor* and an *intelligent sensor* (see Figure 12.4):

- a *smart sensor* has functionalities that improve its metrological performances by using numerical processing;
- an *intelligent sensor* integrates functions that allow it to fully participate in the goal of an automated system, which then becomes a distributed automated system. Mechanisms are implanted in this system and exchange information through the dedicated communication system. This system is the backbone of a true real-time database [BAY 95].



**Figure 12.4.** Smart and intelligent sensors

Although these names are now in common usage, we prefer the term “digital sensor with processing and communication capacities”, which specifies that the system is a measurement device, that it is created by digital technology, that it has bidirectional communication means, and processing capacities. An intelligent sensor is then seen as a fully functional system with its own processing abilities, one that can take part in more complex systems.

### 12.2.3. Architecture of an intelligent system

Figure 12.5 gives an example of material architecture of an intelligent sensor. This architecture includes one or several transducers connected to the conditioners:

- basic sensor components that convert the measurand into an electric signal, which is usually analog;
- a numerical processing chain of information, including the following elements:
  - an interface of the measurement (a multiplexer, amplifier, ADC, sample and hold);
  - a calculation unit (a microcontroller, microprocessor, DSP) and linked peripherals (memory units);
  - a communication interface that ensures bidirectional communication with the automation system through a fieldbus;
- an energy feed that is usually integrated with the intelligent sensor.

This architecture clearly is more complex than that of a standard sensor. It links one or several measurement chains and the equivalent of a computer machine [BRI 96]. The related processing possibilities improve metrological performances in terms of reliability and sensor availability.

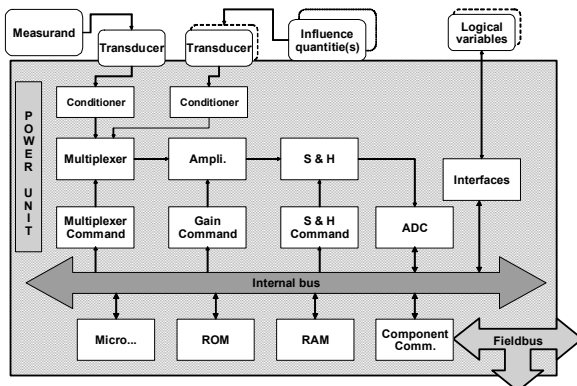


Figure 12.5. Example of intelligent sensor architecture

### 12.3. Processing and performances

#### 12.3.1. *Improving performances with sensors*

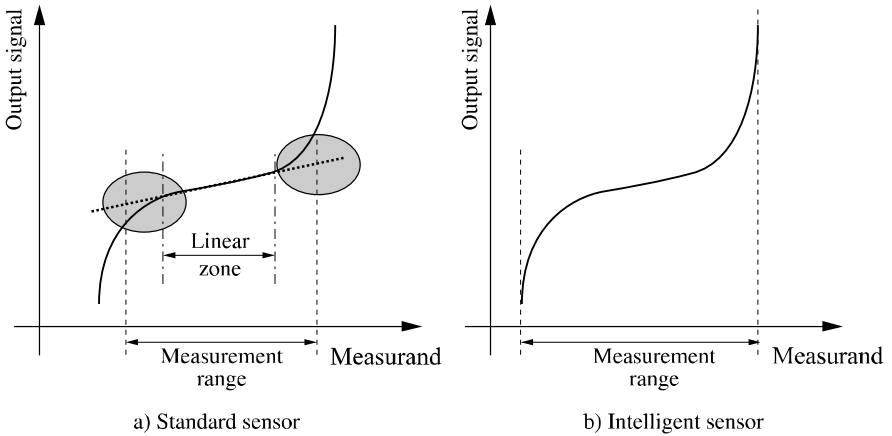
Among the essential qualities of sensors [AFN 94] are the following: freedom from bias error, fidelity, accuracy, rangeability, sensitivity, linearity, sharpness or keenness, rapidity, resolution, traceability, repeatability and reproducibility. Accuracy is the attribute which is most often considered by users.

Information processing carried out by intelligent sensors improves accuracy by the following means:

- processing compensates for influence quantities, which must be taken into account as “normal” variables, leading to a multisensor approach to data fusion;
- processing uses signal processing algorithms, from the simple mobile average to, for example, the implementation of deconvolution procedures that return to the excitor signal of the sensor. If its transfer function is known, fairly sophisticated numerical filters can be used.

This improvement of accuracy can make the sensor more adaptable and improve its range, so it can be used in a variety of situations and applications.

For example, the same type of temperature sensor can acquire information about the interior temperature of a car and about the interior of a cylinder, always retaining its precision. Obviously, modelizing the transfer function signal output =  $f(\text{measurand})$  feature goes beyond simple linear transformations. In addition, the linearity requirements need no longer be respected, since the information provided is quantified, digitized and transmitted according to a range of different codings. These codings allows us to link the corresponding physical unity, freeing us from fixed rule of linearity imposed by using an intermediary variable such as the 4-20 mA, or 0-10 V for reasons of interoperability between sensors, actuators, regulators or recorders.



**Figure 12.6.** Rangeability, linearity and exactitude

In other words, these improvements in metrological performances have a downside. This downside is mainly connected with modelization which takes place during conception stages and static or dynamic modelization of the transducer(s), which may be completed by individualized calibration procedures [NOI 97], [ROB 99].

### 12.3.2. Reliability and availability of information

Technical writing abounds with examples of failures in production systems [VIL 98] caused by sensors transmitting incorrect measurements and information to operators or to automation systems. The decisions made from this information based on measurements sometimes have had serious consequences.

Thus, the overriding need for reliability was behind the early development of intelligent sensors. An intelligent sensor must be able to provide valid information leading to the dependability of the application, even if the application itself is not of optimal reliability, availability, safety, maintainability or durability.

These goals are met through validation procedures that are completed by:

- auto-tests and auto-diagnostics;
- stored memory of the last delivered values;
- alarm systems used when failures are detected;
- configuration re-readings;



- network reconfigurations.

These procedures are described in detail in [BRI 94] and [CIA 87].

In the context of cars, the need for reliable, certain and validated information is obvious. No consumer will accept an airbag that doesn't inflate; no driver will accept false warnings about engine problems.

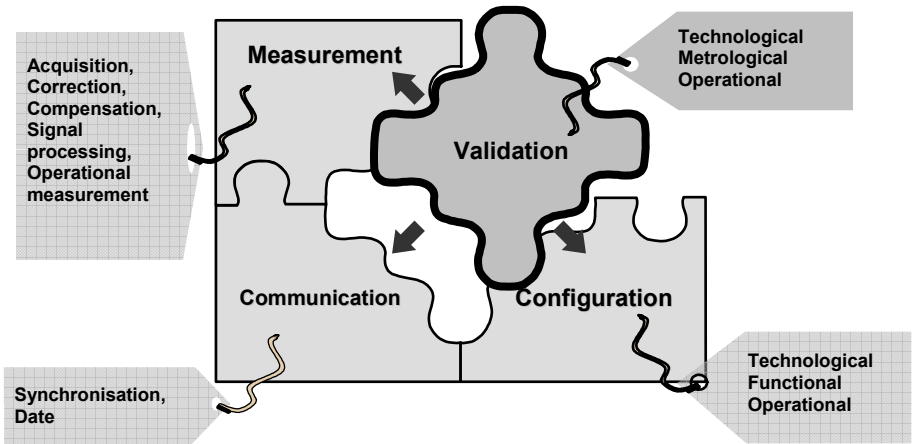


Figure 12.7. Validation and functionalities

Validation functions provide informal descriptions of the idea of the intelligent sensor, as shown in Figure 12.7.

There is a crucial need for intelligent sensors which can provide data and generate validated information. Indeed, the issue of validation is crucial to intelligent sensors.

Validation on many levels, including technological, functional, metrological and operational, is discussed in detail in [ROB 93]. Validation is based on material or analytical redundancies and is integrated at different levels. However, we must keep in mind that a material and conceptual limitation does not allow us to recursively validate all information necessary to carry out and record a measurement.

However, the basic function of a sensor is to provide measurements. The processes it undergoes allow it to transform the principle quantity or quantities into an operational measurement (see Figure 12.8) that results from the consideration of metrological compensations through the influence quantity or quantities, and of

validations made through technological means. Obtaining this “high value-added” measurement requires:

- knowing and exploiting the behavior models of the transducer(s);
- regulation of internal time, which helps in dating the operational measurement;
- regulation of diverse data, such as:
  - validation thresholds;
  - anterior values to the operational measurement;
  - measurements coming from other sensors.

All these are re-grouped in an equivalent database.

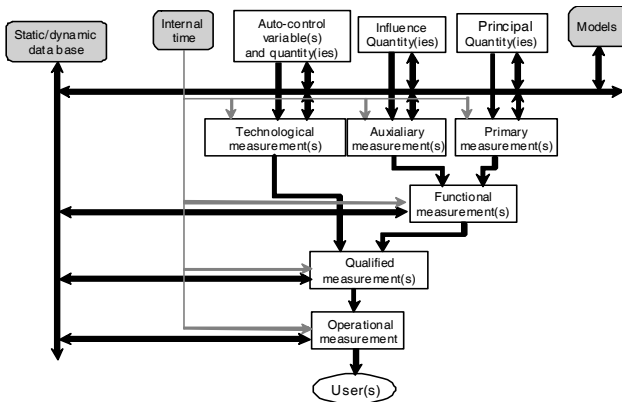


Figure 12.8. Obtaining an operational measurement

## 12.4. Intelligent distance sensors in cars

In the following paragraphs we will illustrate these concepts by showing how the level of gasoline in a tank is converted to relevant information about how far a car can travel with this gasoline.

The *principle quantity* here is obviously the *level of fuel* in the gas tank.

The *temperature* inside the gas tank can be seen as an *influence quantity* that through the laws of liquid dilation influences the quantity of energy actually available. The influence quantities are intrinsically linked to metrological and physico-chemical characteristics of the proof body, allowing us to carry out the first measurement. We see that the chemical characteristics of the fuel (the octane rate can vary according to where the gasoline was bought) can be taken into account to make the prediction more exact.

The *autocontrol variables* are, for example, the *pressure* and *temperature* in the gas tank. These help to validate the nominal functioning of the sensor of the level used and provide information relating to safety. We see that the variable “temperature inside the gas tank” appears as both an influence quantity and an autocontrol variable; it can be used in many validation steps. On another level, the temperature of the processing unit, the supply voltage of the electronic module, and the supply voltage of any sensor can also be variables requiring surveillance.

The *primary measurement* is an electronic signal coming from the height sensor, a signal which must be validated technologically. Then it must be verified that the frequency and amplitude of the supply voltage conditioner linked to the capacitive transducer being used are close to nominal values, that the output of this conditioner is coherent with the geometric characteristics of the gas tank.

The *auxiliary measurement* is the tension delivered by a thermistor that has been traversed by a known or measured current that represents the temperature of the tank. This measurement is obtained through a model specific to the thermistor that describes the temperature relation of the range of potential utilization of this proof body. So the temperature =  $f(\text{tension, current})$ . This model must be constructed and stored in the intelligent sensor. We point out here that this auxiliary measurement can be compensated by taking into account auto-heating phenomena, for example, or validated by measuring the current going through the thermistor; so validation has a recursive or returning aspect.

*Technological measurements* are:

- the signal delivered by the proof-body and pressure-conditioner that have been placed in the sensor for safety reasons;
- the signal delivered by the temperature sensor;
- the various supply voltages that must be watched.

The group of *primary* and *auxiliary measurements* leads to the production of a *functional measurement*. In our example of a car, the initial measurement of the fuel level can be converted into information showing the volume of available fuel. The information is produced by taking into account the geometry of the fuel tank, which varies according to the vehicle. The information is then clarified during a configuration step and then is stored in the intelligent sensor in the “static database” zone. This data about fuel volume can then be corrected by data about volume of fuel available at a “normalized” temperature; that is, by consideration of laws of fluid dilation, which will then also be memorized in the “static database”.

The *validated* or *qualified measurement* is the information pertaining to the available fuel volume available at a normalized temperature relative to the

technological measurements requiring that the nominal conditions for primary and auxiliary measurements be combined at the moment a measurement is made. If this is not the case, the processing can be done following several different strategies:

- if the producer of the primary measurement is faulty, then an estimation of the measurement can be obtained from measurements previously made that have been stored on a specific temporal horizon in the “dynamic database” by taking into account the distance traveled since the problem began;

- if the metrological conditions are not combined by the faulty temperature sensor, for example, the estimation of the distance traveled will be less precise and this estimation error may be transmitted to the user;

- in other circumstances, a foldback value that has previously been parameterized according to the application may be produced. Typically, the foldback value can be the last produced operational measurement;

- in addition, alarm systems can be produced; the receiving systems or units must have be able to make decisions to regulate these unusual situations which, however, have been foreseen in the design step of the system.

Lastly, the *operational measurement* is the information concerning the distance that the vehicle can travel; this is the truly useful information. It will then be transmitted to the “final” user who may be the driver, or to a control guidance system that will even lead the driver to a local, open service station. This operational measurement is expressed in a configurable unit for the user in miles or kilometers. It integrates a margin of error that takes into account the exactitude of each of the first mechanisms that provide information: the primary sensors. This measurement is obtained by looking at the instantaneous consumption that reflects how the vehicle is being used. The fuel consumption can be measured by a flow sensor that itself produces validated information which it transmits to an intelligent sensor showing the distance that can be traveled.

The physical variable or quantity “time” can also be measured or more or less produced within the system itself, in order to represent the refresh period of the operational measurement. This physical variable “time” can be considered as an implicit physical quantity.

This purely academic example shows us that an intelligent sensor:

- can integrate local processing, based on models;
- requires parameterization or configuration that allows it to be dedicated to a given application;
- requires information coming from other components.

An example of the third point is a temperature measurement that must be tested by a control mechanism of the vehicle that guards against overheating or fires, possibly by triggering a fire extinguisher.

This transmission of information in a vehicle is not perfect [MEN 99]. It often depends on a speed sensor, for example, put in an ABS system that transmits information about speed to a dashboard, which then can override the rotation speed sensor initially part of the output of the transmission shaft.

This last example underlines the need for integration and precise representations of relevant data, specifications for related processing, and efficient exchanges of data within a vehicle. It brings up the potential problems of optimal data distribution and the related processing of a system.

## 12.5. Fieldbus networks

Using intelligent sensors or, more generally, a range of automation components means digital communication between producers and consumers, and also using dedicated communication networks, since operating constraints in real time or critical time are not the same as those of usual computer networks.

For production systems, more than 50 types of fieldbuses are currently on the market, many of which are described in [CIA 99], [FAG 96], [PET 96] and [TER 97].

In the automotive domain, the Controller Area Network (CAN) [PAR 96, 99] is currently the reference fieldbus, as well as another automotive multiplexing network, the Vehicle Area Network (VAN), which has mainly been used by PSA and its subsidiaries [ABO 97].

CAN was first developed by Bosch and Intel for automotive applications. It has had good reliability and low costs. Many car manufacturers use or are preparing to use the CAN network. The association of utility vehicles of the USA have also adopted the fieldbus as a standard, as have most industrial manufacturers, mostly because of the buses' wide availability and competitive costs. At this time, according to [MEN 99], sales for CAN components for cars have surpassed sales to the industry.

Schematically, the CAN uses a bus topology and belongs to the class of multimaster networks of the producer/consumer type. In these networks, different levels of information are transmitted according to the diffusion principle, and regulated by implanting the protocol CSMA/CR (Carrier Sense Multiple

Access/Collision Resolve). The nominal rate of communication depends on the physical length of the bus; typically, a length of 40 m is associated with a transmission rate of 1 Mbit/s.

Many other multiplexing systems or, more generally, exchange systems in cars are now being developed by automotive manufacturers or other consortia. Among these are the J1850 or the ITS (Intelligent Transport System), the Data Bus made by the SAE, Society of Automotive Engineers, and the OSEK/VDE or the Open systems and Interfaces for Distributed Electronics in Cars/Vehicle Distributed Executive. The latter is a consortium of European manufacturers and scientists that aims to establish operating system norms and communication protocols in the field of electronics.

Other projects connected to automotive embedded electronics are using fiber optic networks that enable flows of more than 10 Mbit/s (the AMINC consortium, standing for the Automotive Multimedia Integration Consortium).

## 12.6. Towards a system approach

The previous sections show that, in the world of automotive electronics, the current trend is towards developing interacting, multiplexing systems. This approach is very important to the field of intelligent sensors, which are integrated into a distributed automation architecture since, in addition to the characteristics discussed above, an intelligent sensor must have the following features;

- it must be interoperable. This means it must cooperate with other automation components in and with a specific application. This, at the very least, means using a common standard of communication to allow the exchange of information and also means that the two components conform to the same interpretation of data;

- its components must be interprocessible. This feature, which is not easily differentiated from the first one, is especially important for the equipment and its integration with a automation system with distributed architecture (ASDA) dedicated to an industrial process;

- it must be interchangeable. This means that the equipment of one manufacturer can be replaced with that of another manufacturer without changing the components.

The above concepts [STA 96] are both an advance and a check to any large-scale diffusion of intelligent sensors in industry. They imply normalization procedures which could appear to be constraints. However, some studies ([INC 93]; [LEE 96]) have proposed a standard, or at least an aid, to producing intelligent instruments that might disregard interface communication. In the near future, these works might lead to a total or partial modelization of the abilities of intelligent sensors and also lead to

creating tools to aid in designing intelligent sensors, thus verifying specifications and performances and ensuring that interoperability and inter-processing criteria are respected.

This is one of the objectives of the LARII project for “Software of Assistance to the Creation of Intelligent Interfaces for Sensors and Actuators” which profited from the financial support of the French authorities and which must make it possible to diffuse near the French companies manufacturing of the sensors, mainly PMI-PME of the software libraries, comprising functional blocks based on the standard IEC 1131 (typing of the data, definition of the functions, programming language, etc.) directly usable in the design of intelligent sensors. These libraries will be integrated in the strategy of design which should no longer remain on a solely technological approach [ LEM 99].

### 12.7. Perspectives and conclusions

In spite of industry’s stated enthusiasm for the idea of intelligent sensors, there are clear drawbacks to their present widespread use:

- the wait-and-see policy of designers, manufacturers and potential users, who must choose between the current wide range of communication networks and fieldbuses (more than 50 fieldbuses are currently on the market ([TER 97, PET 98]));
- advances in industrial instrumentation are basically marketing tools of manufacturers who develop new techniques before there is a real need for them;
- the integration of intelligent automation components is linked to the issue of an optimal distribution of data and processing [CON 99]; [HER 98].

However, academic studies have seemed to develop in two directions:

- integration of information processing techniques using fuzzy logic, neural networks, etc. [END 97];
- connection on the same sensitive element structure and of the processing electronics using microtechnologies or nanotechnologies [MUL 95].

In addition, catalogues of automation systems tend to promote “intelligent” products to improve sales.

Especially in the automotive context, Figure 12.9 shows the variety of sensors which can be integrated in a vehicle that is used on a daily basis.

An article from the end of the last century [GRA 99] mentioned the sale of sensors used in vehicles: these sales are estimated at \$5.18 billion in 1997, while

sales predictions for 2002 are estimated at \$6.86 billion. Even though the technical works often combine sensors, associated electronics, associated software and multiplexing, the sale of electronic chips for cars was estimated at \$8.25 million in 1998; predictions for 2000-2001 have been estimated at \$13.3 million [VER 99]. Experts' predictions say that automotive electronics will, in the short term, represent 20% of the finished product, which may well surpass the cost of the mechanics. According to [GRA 99], the "salability threshold" for a vehicle sensor may be around \$5-7, including the cost of signal processing. This means using digital technologies of the MEMS (Microelectrochemical systems) or MST (Microsystems technologies) type.

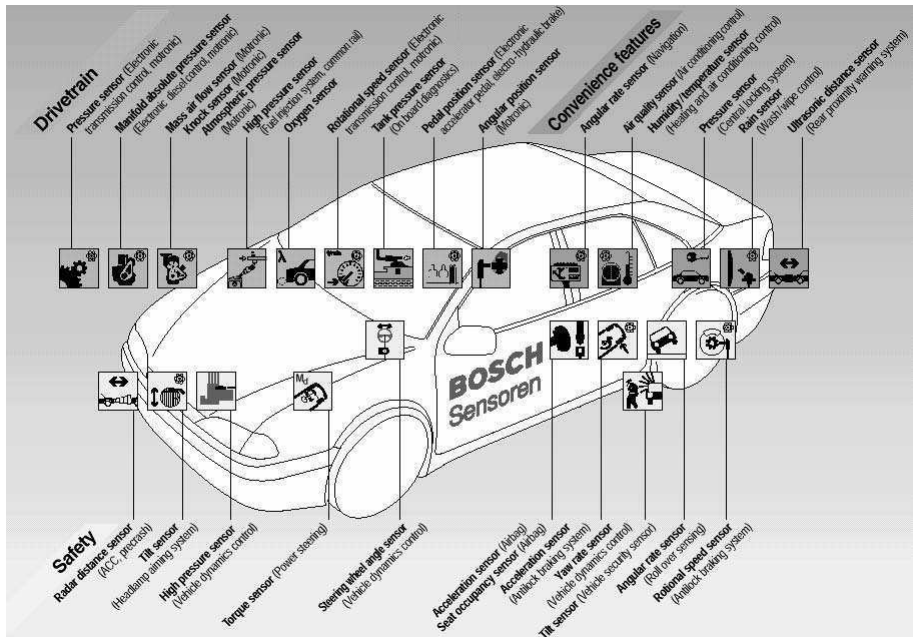


Figure 12.9. Car sensors (image courtesy of Robert Bosch GmbH)

Aside from the optimal running of the engine and the comfort of the driver and passengers, the key word in the automotive field is safety. We can see this simply by looking at acronyms such as ESP (Electronic Stability Program), EBS (Electronic Braking System), TCS (Traction Control System) and ASR (Acceleration Slip Regulation), which imposes calculation powers comparable to those of a 1982 A310 Airbus [VER 99].



We also note that on the margins of automotive manufacturing, the demand for sensors and associated systems is also important in the field of vehicle testing, both in the design phase and in the test-run phases [ERW 99].

## 12.8. Bibliography

- [ABO 97] ABOU B., MALVILLE J., *Le Bus Van (Vehicle Area Network). Fondements du protocole*, Dunod, 1997.
- [AFN 94] AFNOR, *Vocabulaire fondamental des termes fondamentaux et généraux de métrologie*, Norme AFNOR NF X07-001, December 1994.
- [AFN 96] *Maintenance industrielle*, Recueil de normes françaises, réf. 321 61 40, Paris, Ed. AFNOR, 1996.
- [BAY 95] BAYART M., SIMONOT F., Rapport Convention MESR 92-P-0239: Impact de l'émergence des réseaux de terrain et de l'instrumentation intelligente dans la conception des architectures des systèmes d'automatisation de processus. Rapport Final. Résultats et Perspectives, Ministère de l'Enseignement Supérieur et de la Recherche, January 1995.
- [BER 87] BERRY, "Distributed intelligence in Process Control", *Control Engineering*, May 87, p. 62-64, 1987.
- [BLA 87] BLADOU, "Comment l'intelligence vient aux capteurs", *Electronique Industrielle*, no. 127, 1987.
- [BRI 94] BRIGNELL J., TANER A., "Aspects of smart sensor reconfiguration", *Proceedings of Eurosensors VIII conference*, Toulouse, p. 525-529, 25-28 September 1994.
- [BRI 96] BRIGNELL J., *Intelligent Sensor Systems*, Institute of Physics Publishing, Sensors Series, 1996.
- [CIA 87] CIAME, *Livre blanc: les Capteurs Intelligents*, réflexions des utilisateurs, CIAME AFCET, 1987.
- [CIA 99] CIAME (Collectif), *Réseaux de terrain. Description et critères de choix*, Hermès, 1999.
- [CON 99] CONRARD B., Contribution à l'évaluation quantitative de la sûreté de fonctionnement des systèmes d'automatisation en phase de conception, PhD Thesis, University of Henri Poincaré-Nancy I, 24 September 1999.
- [CRO 99] CROSSE J., "Smart Tyres, The shape of things to come", *Financial Times Automotive World*, p. 45, October 1999.
- [DEL 90] DELCUVELLERIE J.-L., "L'impact de FIP sur la conception des architectures", *Revue Générale de l'Electricité*, p. 57-64, 7 July 1990.
- [END 97] ENDRES H.E., "Signal evaluation of gas sensors with artificial neural nets", *Proceedings of SICICA '97*, p. 399-404, IFAC, 9-11 June 1997.

- [ERW 99] ERWIN F., "Improvements in Vehicle-Testing Technology", *Sensors*, vol. 16, no. 12, December 1999.
- [FAG 96] FAGES G., *Les Bus de Terrain*, Technical collection, Schneider, 1996.
- [GIA 86] GIACHINO, "Smart Sensors", *Sensors & Actuators*, no. 10, p. 239-248, Elsevier Science S.A., 1986.
- [GRA 99] GRACE R., "The Growing Presence of MEMS and MST in Automotive Applications", *Sensors*, vol. 16, no. 9, September 1999.
- [GUY 99] GUYOT R., "Injecteur pompe Diesel", *Auto Concept*, no. 25, April-May 99.
- [HER 97] HERMANN F., THIRIET J. M., ROBERT M., "Task allocation for the design of distributed architecture in automation systems: application to a pilot thermal process using a fieldbus network", *Proceedings of SICICA '97*, Annecy, IFAC, p. 459-464, 9-11 June 1997.
- [INC 93] Projet Eureka EU 666 INCA Interface Normalisée pour Capteurs et Actionneurs, Tâche 0B, Rapport final des travaux, July 1993.
- [JOR 87] JORDAN, "Sensor technologies of the future", *GEC Review*, vol. 3, no. 1, p. 23-32, 1987.
- [KLE 91] KLEINSCHMIDT P., SCHMIDT F., "How many sensors does a car need?", *Proceedings of EuroSensors V Conference*, Rome, p. 1-13, 2 October 1991.
- [LEE 96] LEE K.B., SCHNEEMAN R., "A standardized approach for transducer interfacing: implementing IEEE-P1451 Smart transducer interface standards", *Proceedings of Sensors Conference*, Philadelphia, p. 87-100, 22-24 October 1996.
- [LEM 99] LEMAIRE E., Spécification et Vérification Fonctionnelle et comportementale d'un Equipement Intelligent, PhD Thesis, University of Sciences and Technologies, Lille, 14 December 1999.
- [MEN 99] MENARD C., "L'électronique embarquée se cherche une architecture et une fiabilité", *l'Usine Nouvelle*, no. 2703, p. 76-83, 30 September 1999.
- [MUL 95] MULLINS M.A., VAN PUTTEN A.F.P., BAYFORD R., BUTCHER J.B., "Potential for a smart sensor based on an integrated silicon anemometer", *Sensors and Actuators (A)*, no. 46-47, p. 342-348, Elsevier Science S.A., 1995.
- [NOI 97] NOIZETTE J.L., ROBERT M., RIVIERE J.M., "Intelligent sensor calibration complexity, methodology and consequences", *Proceedings of IEEE Instrumentation and Measurement Technology Conference*, Ottawa, p. 948-952, 19-21 May 1997.
- [PAR 96] PARET D., *Le bus CAN, Description, De la théorie à la pratique*, Dunod, 1996.
- [PAR 99] PARET D., *Le bus CAN, Applications CAL, CANopen, DeviceNet, OSEK, SDS*, Dunod, 1999.

- [PET 98] PETERSON G.C., "Selecting the right Industrial Network", *Control Engineering International*, p. 43-46, January 1998.
- [ROB 93] ROBERT M.J., MARCHANDIAUX M., PORTE M., *Capteurs Intelligents et Méthodologie d'Evaluation*, Hermès, 1993.
- [ROB 99] ROBERT M.R., Contribution au développement d'une méthodologie d'étalonnage de capteurs : application à un transmetteur intelligent de température, Doctorat de l'Université Henri Poincaré-Nancy 1, 18 February 1999.
- [RUT 91] RUTLEDGE D.N., DUCAUZE C.J., "An iterative method for determining the linear zone of a detector response", *Chemometrics and Intelligent Laboratory Systems*, no. 12, p. 5-19, 1991.
- [STA 96] STAROSWIECKI M., BAYART M., "Models and languages for the interoperability of smart instruments", *Automatica*, vol. 32, no. 6, June 1996.
- [TER 97] *Terrain*, "Le marché des bus de terrain", no. 13, p. 18-22, May-June 1997.
- [VER 99] VERNAY J.P., "La voiture branchée est électronique", *l'Usine Nouvelle*, no. 2703 30, p. 72-75, September 1999.
- [VIL 98] VILLEMEUR A., *Sûreté de fonctionnement des systèmes industriels*, Eyrolles, 1998.

## List of Authors

Patrice AKNIN  
INRETS, Arcueil

François BAILLIEU  
ESIEE, Noisy-le-Grand

Paul BILDSTEIN  
ESIEE, Noisy-le-Grand

Cécile DURIEU  
Ecole Normale Supérieure de Cachan

Bernard JOURNET  
Ecole Normale Supérieure de Cachan

Michel LECOLLINET  
CNAM, Paris

François LEPOUTRE  
CNAM, Paris  
ONERA/DMSE

Thierry MAURIN  
Ecole Normale Supérieure de Cachan  
IEF, Paris-Sud University

Dominique MILLER  
Ecole Normale Supérieure de Cachan

Mustapha NADI  
LIEN  
Henri Poincaré University  
Nancy

Dominique PLACKO  
Ecole Normale Supérieure de Cachan

Stéphane POUJOULY  
IUT Cachan

Denis PRÉMEL  
Ecole Normale Supérieure de Cachan

Michel ROBERT  
Henri Poincaré University  
Nancy

Eduardo SANTANDER  
Ecole Normale Supérieure de Cachan

Frédéric TRUCHETET  
Le2i  
University of Bourgogne  
Le Creusot

Olivier VANCAUWENBERGHE  
ESIEE, Noisy-le-Grand

# Index

## A

- amplifier 66, 147
  - instrumentation 160
  - isolation 162
  - logarithmic 163
  - operational 147, 153, 192
- axis
  - inertia 468
  - principle 468

## C

- cell
  - capacitive 291
  - Fleisher-Tow 198
  - Friend 194
  - piezoresistive 307
  - photoconductor 94
  - Sallen-Key 193
  - Tow-Thomas 196
- converter
  - analog-to-digital 229
- current
  - darkness 82
  - gap 156
  - polarization 156

## E

- effect
  - load 12, 18
  - Peltier 131
  - photoelectric 94
  - Seeback 129
  - Thomson 130, 131

## F

- filter 223
  - active 168, 191, 199
  - analog 167, 169
  - anti-folding 215, 228
  - corrective 170
  - dynamic 432
  - FIR 258, 431
  - half-band 215, 241
  - IIR 260, 431
  - Kalman 439
  - low pass ladder 179
  - passive 168, 177
  - Wiener 437

**M**

- measurement
  - auxiliary 520
  - functional 520
  - operational 519, 521
  - primary 520
  - technological 520

**N**

- noise
  - electronic 138
  - thermal 138
  - white 422

**P**

- processing
  - analog 137, 416
  - digital 422, 436
  - signal 464, 506

**Q**

- quantities
  - influence 11, 61, 518, 519
  - interfering 11
  - modifying 11

**S**

- sensor
  - intelligent 509
- signal
  - analog 29, 416
  - causal 417
  - deterministic 416
  - ergodic 421
  - periodic 420
  - quantization of 423, 427
  - sampled 423
  - stationary 421
- signal-to-noise ratio 145, 306
- synthesis
  - cascade 175
  - Darlington analytic 181

**T**

- theorem
  - Parseval 419
  - Plancherel 418
  - Wiener-Kintchine 419, 421
- transform
  - bilinear 456
  - Fourier 418, 420, 423, 443, 444, 446, 447, 449
  - Gabor 448
  - Hartley 444
  - Wigner-Ville 456, 457, 459