Instituto Superior Politécnico de Viseu
Escola Superior de Tecnologia de Viseu
Curso de Engenharia de Sistemas e Informática

Processamento Digital de Sinal

Aula 1,2

4.º Ano – 2.º Semestre

Manuel A. E. Baptista, Eng.º

Programa:

1. Introdução ao Processamento Digital de Sinal
2. Representação e Análise de Sinais
3. Estruturas e Projecto de Filtros FIR e IIR
4. Processamento de Imagem
5. Processadores Digitais de Sinal

Bibliografia:

Processamento Digital de Sinal:

•Sanjit K. Mitra, "**Digital Signal Processing** – A computer based approach", McGraw Hill, 1998
Cota: 621.391 MIT DIG
•Roman Kuc, "**Introduction to Digital Signal Processing**", McGraw Hill, 1988.
Cota: 621.391 KUC INT
•Johnny R. Johnson, "**Introduction to Digital Signal Processing**", Prentice-Hall, 1989.
Cota: 621.391 JOH INT
G. Proakis, G. Manolakis, "**Digital Signal Processing – Principles, Algorithms Applications**", 3ª Ed, P-Hall, 1996.
Cota: 621.391 PRO DIG
•James V. Candy, "**Signal Processing – The modern Approach**", McGraw-Hill, 1988
Cota: 621.391 CAN SIG
•Mark J. T., Russel M., "**Introduction to DSP – A computer Laboratory Textbook**", John Wiley & Sons, 1992.
Cota: 621.391 SMI INT
•James H. McClellan e outros, "**Computer-Based Exercises - Signal Proc. Using Matlab 5**", Prentice-Hall, 1998.
Cota: 621.391 MCC COM

Processamento Digital de Imagem:

•Rafael C. Gonzalez & Richard E. Woods, "**Digital Image Processing** ", Prentice Hall, 2ª Ed., 2002.
Cota: 681.5 GON DIG.
•I. Pittas H. McClellan e outros, "Digital Image Processing Algorithms and Applications", John Wiley & Sons, 2000.
Cota: 621.391 PIT.
•William K. Pratt, "**Digital image processing**", John Wiley, 2ª Ed, 1991.
Cota: 681.5 PRA DIG
•Bernd Jãhne, "**Digital image processing : concepts, algorithms, and scientific applications**", Springer, 1997.
Cota: 681.5 JAH

Avaliação:

A avaliação é composta pela componente teórica e componente prática ponderadas da seguinte forma:

<u>Classificação Final</u> = 80% * Frequência ou exame + 20% * Prática

O acesso ao exame não está condicionado embora não tenha função de melhoria, ou seja, se o aluno entregar a prova de exame, será essa a classificação a utilizar no cálculo da média final independentemente da nota da prova de frequência obtida.

A avaliação prática é constituída por trabalhos laboratoriais a executar em MATLAB

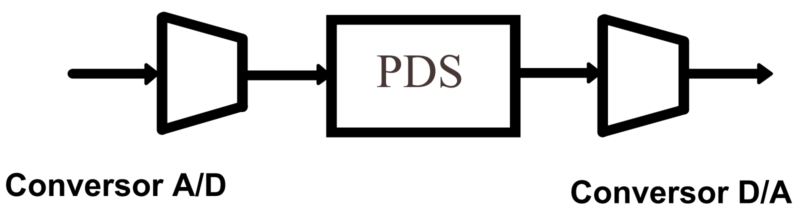
1. Introdução ao Processamento Digital de Sinal

- **Aspectos Gerais**
  - O que é Processamento de um Sinal?
  - Perspectiva Histórica
  - Vantagens e Desvantagens
  - Digital vs Analógico
  - Ferramentas
  - Aplicações
  - Exemplo
- **Sinais**
  - Aquisição e Processamento dum Sinal
  - Tipos de Sinais:
    - Contínuos e Discretos - Periódicos e Não Periódicos
    - Sinais Analógicos e Digitais
  - Processamento no Tempo
  - Processamento na Frequência
- **Conversão A/D**
  - Teorema da Amostragem
  - Parâmetros dum ADC
  - Formatos digitais

Manuel A. E. Baptista
Ernesto R. Afonso

Aspectos Gerais: O que é o Processamento Digital de Sinal?

- Diz-se processamento **digital** de sinal porque normalmente decorre da realização de um sistema discreto que implica, normalmente, a digitalização das amostras dos seus sinais de entrada e de saída.



Manuel A. E. Baptista
Ernesto R. Afonso

Aspectos Gerais: Perspectiva Histórica

- **Modelos matemáticos básicos dos sinais e sistemas contínuos (século XIX):**
  - Transformada de Fourier (1822), por Jean Baptiste Joseph Fourier (1768 – 1830).
  - Transformada de Laplace, por Pierre Simon- Marquês de Laplace (1748- ...).
  - Transformada de Z, por De Moivre em 1730.
- **Advento dos computadores digitais (anos 40) - nascimento de PDS como disciplina.**
- **Anos 50, por Shannon, Bode e Linville: primeiros a equacionar a utilização de computadores de sinal em PDS.**
- **Anos 60, por Kaiser: importantes contribuições para a análise e a síntese de filtros digitais.**
- **1965, a transformada rápida de Fourier (FFT) foi "descoberta" por Cooley e Tukey.**
- **1975, publicação dos livros:**
  - A. V. Oppenheim, R. W. Schafer, *Digital Signal Processing*.
  - L. R. Rabiner, B. Gold, *Theory and Application of Digital Signal Processing*.
  - Marcam o nascimento de PDS como disciplina sendo os autores dos livros considerados como sendo os seus verdadeiros criadores.

Aspectos Gerais: Vantagens e Desvantagens

- **Vantagens**
  - Integrável
  - Flexibilidade
  - Repetibilidade
  - Precisão
  - Processamento de alta complexidade

- **Desvantagens**
  - Requer A/D e D/A
  - Requer filtros de anti-aliasing e reconstrução
  - Limitado em frequência
  - Ruído de quantização

SISTEMAS DE PROCESSAMENTO DIGITAL

Manuel A. E. Baptista
Ernesto R. Afonso

Aspectos Gerais: Processamento Digital vs Processamento Analógico

Digital Signal Processing (DSPing)

Vantagens

- *More flexible.*
- *Often easier system upgrade.*
- *Data easily stored.*
- *Better control over accuracy requirements.*
- *Reproducibility.*

Limitações

- *A/D & signal processors speed: wide-band signals still difficult to treat (real-time systems).*
- *Finite word-length effect.*
- *Obsolescence (analog electronics has it, too!).*

SISTEMAS DE PROCESSAMENTO DIGITAL

Manuel A. E. Baptista
Ernesto R. Afonso

Aspectos Gerais: Objectivo e Ferramentas

Aplicações

- Predicting a system's output.
- Implementing a certain processing task.
- Studying a certain signal.

Hardware

- General purpose processors (GPP), μ-controllers.
- Digital Signal Processors (DSP). } Fast
- Programmable logic (PLD, FPGA). } Faster

Fast / Faster > real-time DSPing

Software

- Programming languages: Pascal, C / C++ ...
- "High level" languages: **Matlab**, Mathcad, Mathematica…
- Dedicated tools (ex: filter design s/w packages).

Manuel A. E. Baptista
Ernesto R. Afonso

Aspectos Gerais:

Aplicações

PDS

- Espaço
  - -Realce de fotografias espaciais
  - -Compressão de dados
  - -Análise inteligente de dados por sondas espaciais
- Medicina
  - -Imagem de diagnóstico (tomografia, ressonância, ultrasons, etc.)
  - -Análise de electrocardiogramas
  - -Aquivo de imagens médicas
- Comercial
  - -Compressão de imagem e som para apresentações multimedia
  - -Efeitos especiais em filmes
  - -Sistemas de video conferência
- Telefone
  - -Compressão de voz e dados
  - -Redução do echo
  - -Multiplexagem de Sinais
  - -Filtragem
- Militar
  - -Radar
  - -Sonar
  - -Ordnance guidance
  - -Comunicações seguras
- Industria
  - -Prospecção de petróleo e minerais
  - -Monitorização e controlo de processos
  - -Testes não destructivos
  - -CAD e ferramentas de design
- Ciência
  - -Registo e análise de tremores de terra
  - -Aquisição de dados
  - -Análise espectral
  - -Modelização e simulação

SISTEMAS DE PROCESSAMENTO DIGITAL
Manuel A. E. Baptista
Ernesto R. Afonso

Aspectos Gerais: Aplicações (2)

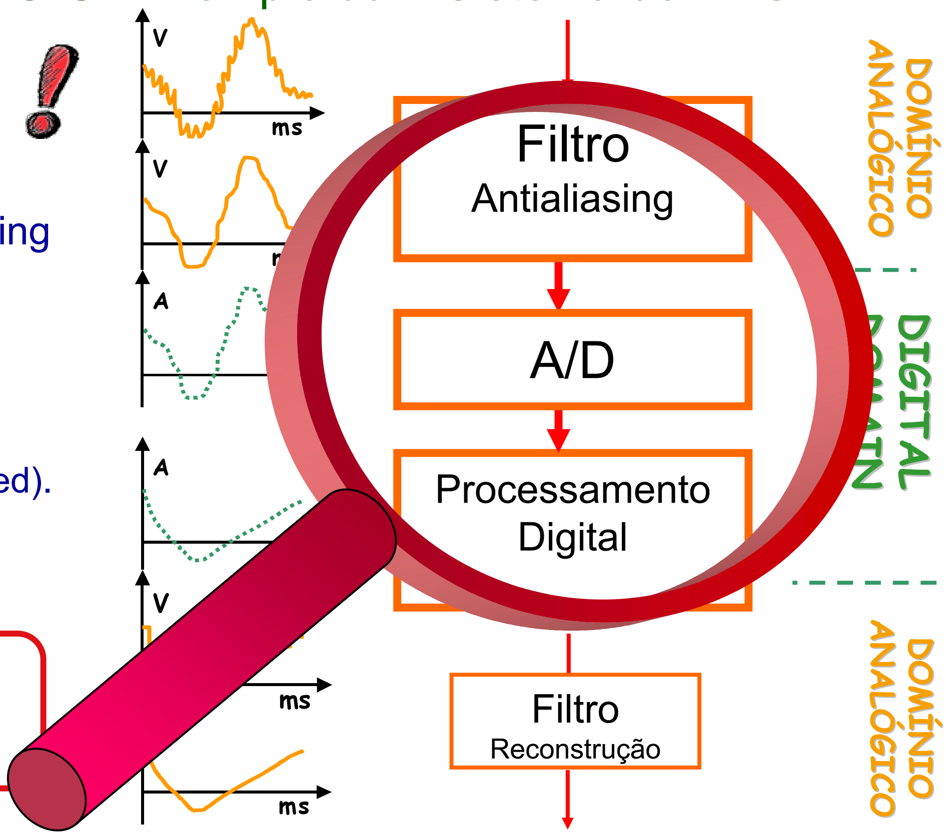
- **Aplicações do processamento de sinais**
  - **Telecomunicações**
    - Radar
    - Compressão de sinais
    - ...
  - **Som**
    - Reconhecimento da fala
    - Síntese de fala
    - Musica [composição e tratamento]
    - ...
  - **Imagem**
    - Tomografia
    - Detecção de movimento
    - ...
  - **Medicina**
    - Electrocardigrafia
    - Electromiografia
    - Electroencefalografia
    - ...

Aspectos Gerais: Exemplo dum Sistema de PDS

Esquema Geral

Sometimes steps missing

- **Filtro + A/D**
  (ex: economics);
- **D/A + filtro**
  (ex: digital output wanted).



Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Aspectos Gerais: Exemplo dum Sistema de PDS (2)

ANALOG INPUT

↓

Antialiasing Filter

↓

A/D

↓

Digital Processing

↓

DIGITAL OUTPUT

KEY DECISION POINTS:

Analysis bandwidth, Dynamic range

- Pass / stop bands.
- Taxa de Amostragem.

1

- No. de bits. Parâmetros.

2

- Formato Digital

3

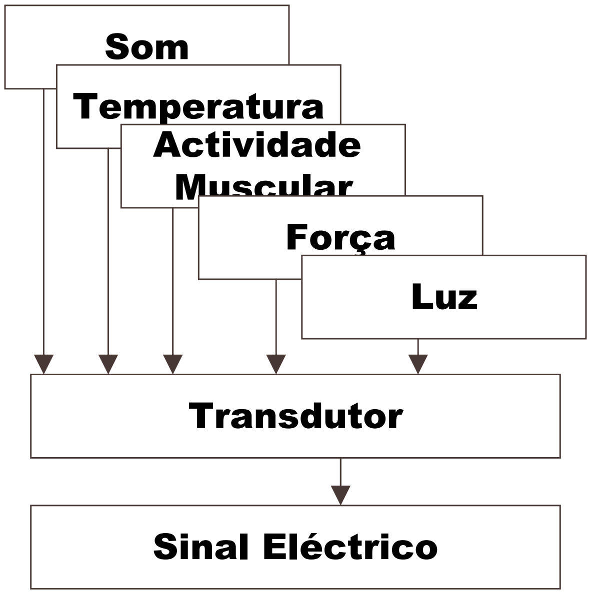
What to use for processing?
See slide “DSPing aim & tools”

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Manuel A. E. Baptista
Ernesto R. Afonso

Sinais: Aquisição e Processamento de um sinal

- Sinal físico

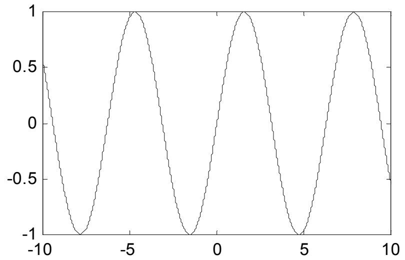
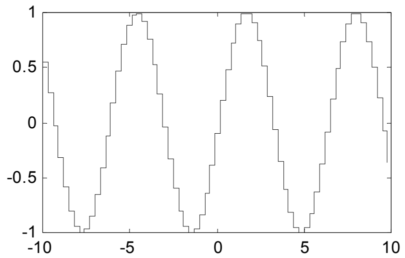


Manuel A. E. Baptista
Ernesto R. Afonso

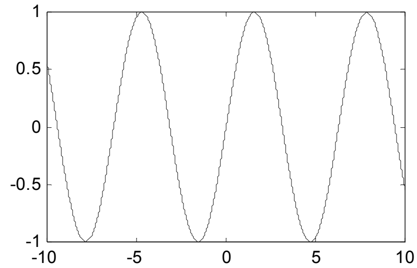
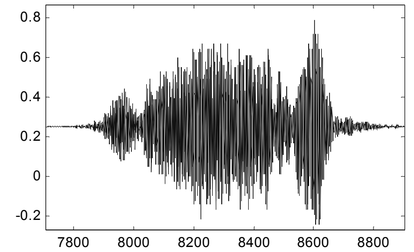
Sinais: Aquisição e Processamento de um sinal (2)

Sinais: Tipos

- Continuo
- Discreto

- Periódico
- Não Periódico

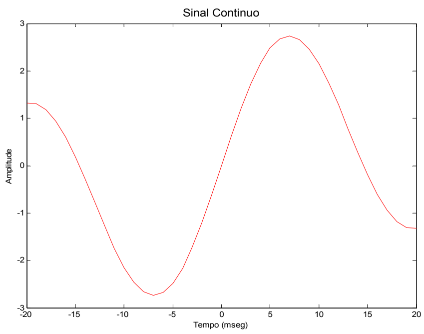



Manuel A. E. Baptista
Ernesto R. Afonso

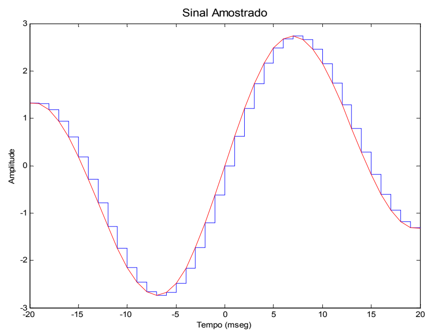
Sinais: Sinais e Sistemas Discretos

- Notação: x[n] → Discreto x(t) → Contínuo

Sinal Contínuo x(t)

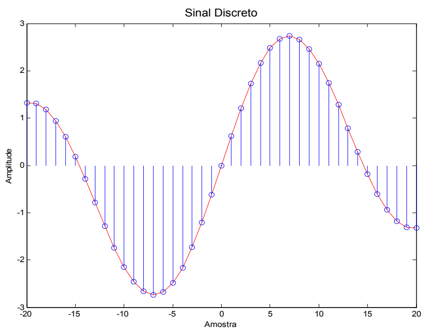


Sinal Amostrado $x_a(t)$

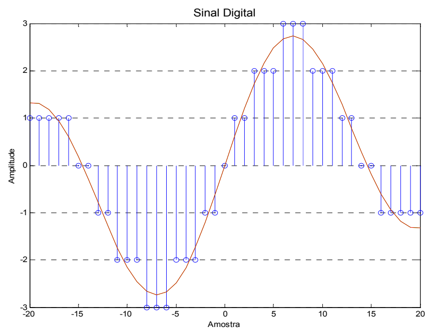


t (ms)

Sinal Discreto x[n]=x(n.T) $n \in Z$



Sinal Digital $x_q[n]$



n(amostras)

Manuel A. E. Baptista
Ernesto R. Afonso

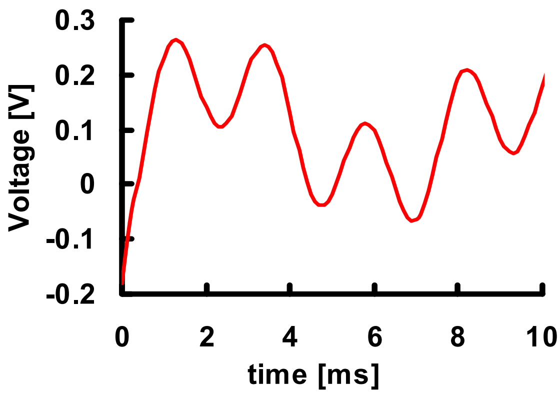
Sinais: analógicos & digitais

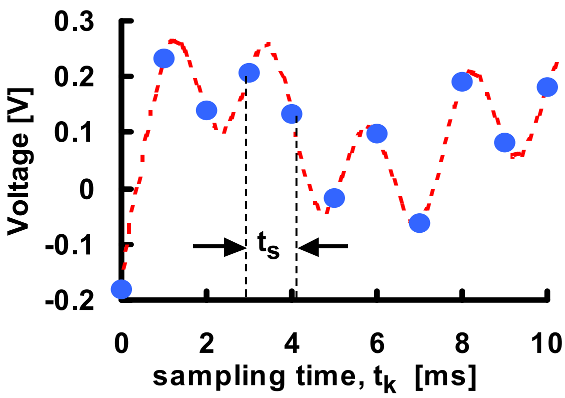
Analógico

Continuous function V of *continuous* variable t (time, space etc) : V(t).

Digital

Discrete function V_k of *discrete* sampling variable t_k, with k = integer: $V_k = V(t_k)$.

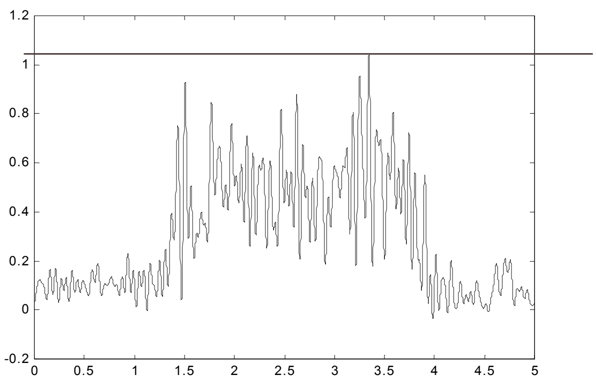


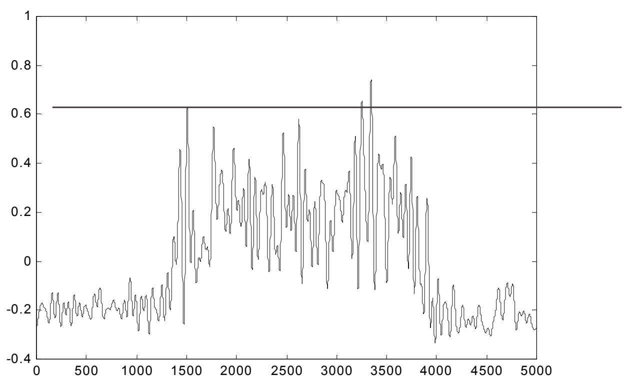


Uniform (periodic) sampling.
Sampling frequency $f_S = 1/ t_S$

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

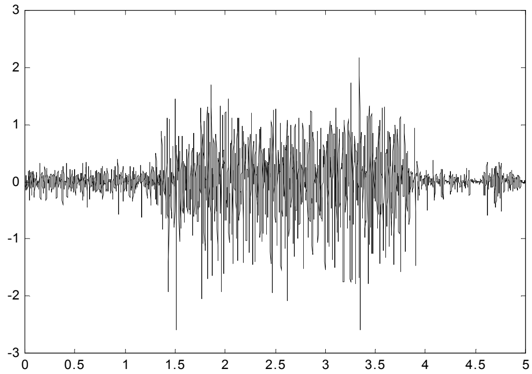
Processamento no Tempo: Remoção da média

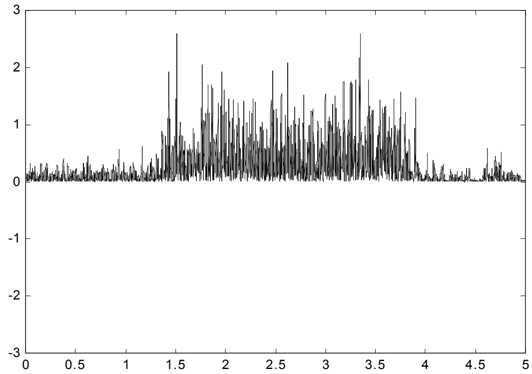




Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

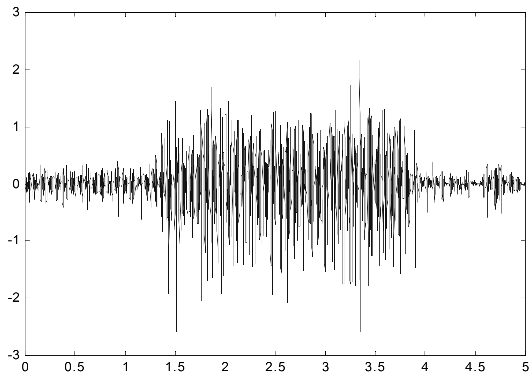
Processamento no Tempo: Rectificação

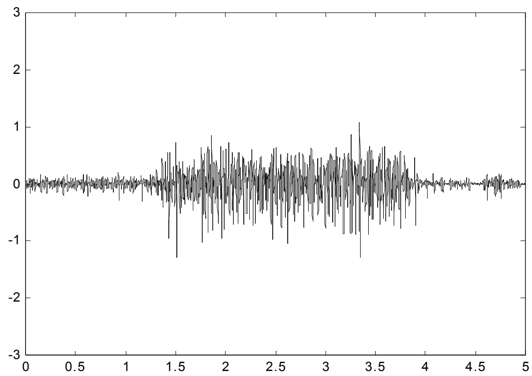




Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

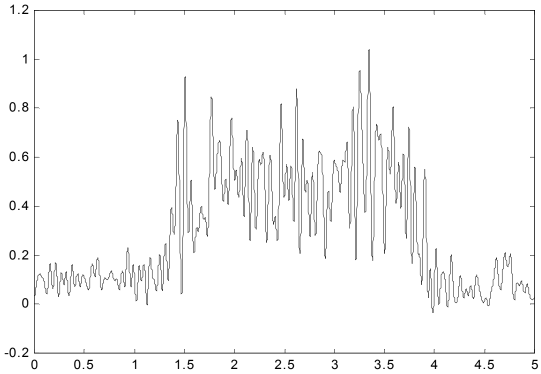
Processamento no Tempo: Amplificação

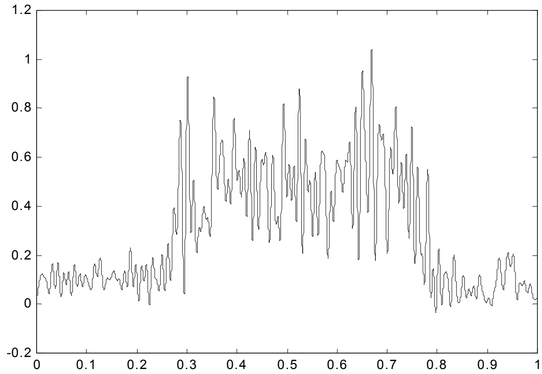




Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Processamento no Tempo: Normalização no tempo

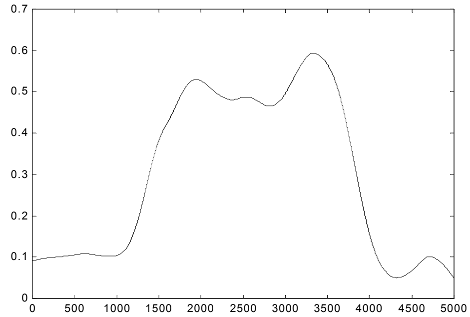




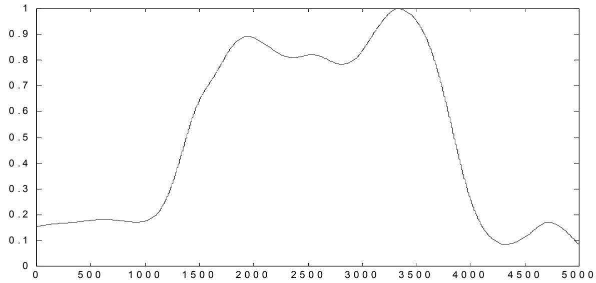
Manuel A. E. Baptista
Ernesto R. Afonso

Processamento no Tempo - Normalização na amplitude

$$y = \frac{x - \mu}{\sigma}$$

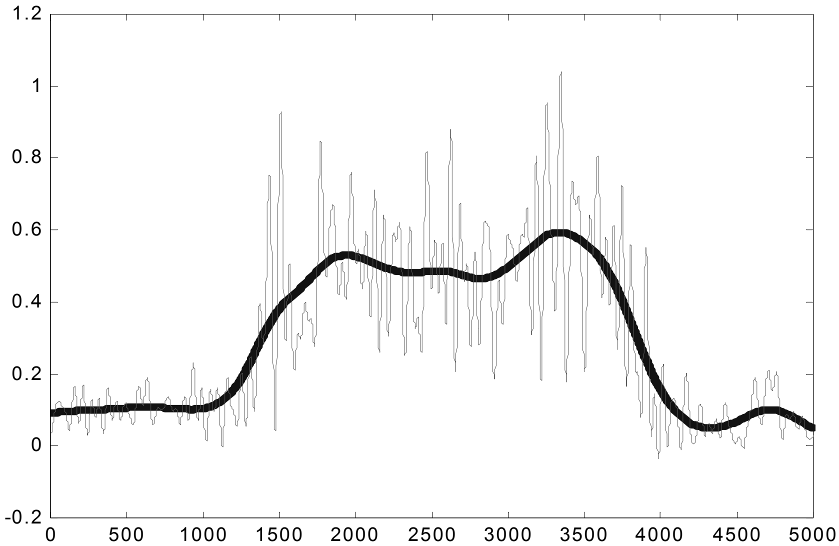


$$y = \frac{x}{\max(x)}$$



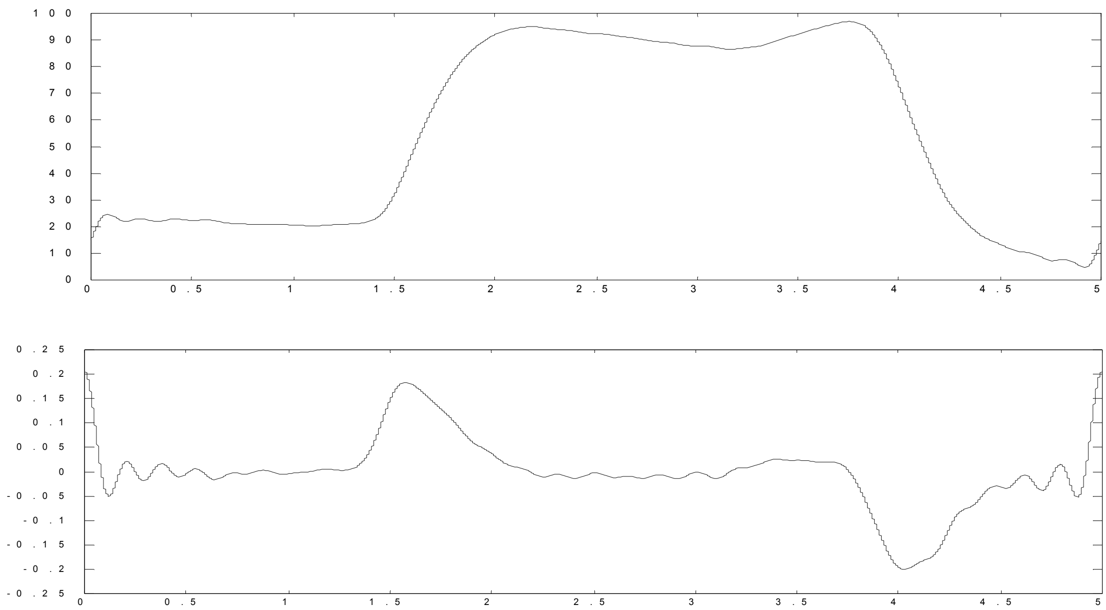
Manuel A. E. Baptista
Ernesto R. Afonso

Processamento no Tempo - Suavização (*smoothing*)



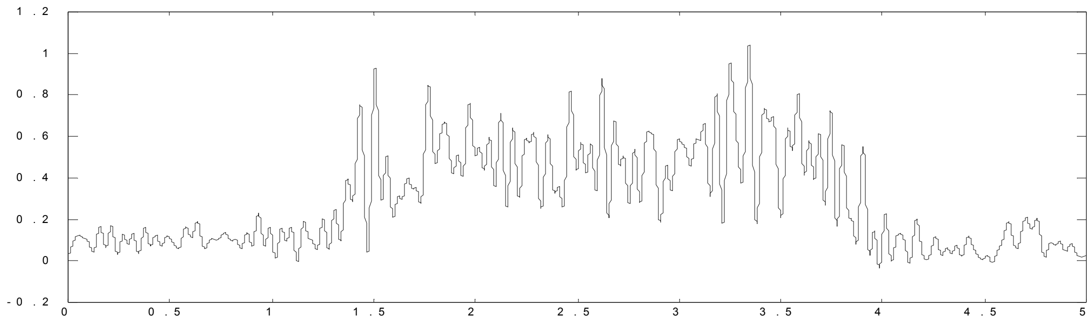
Manuel A. E. Baptista
Ernesto R. Afonso

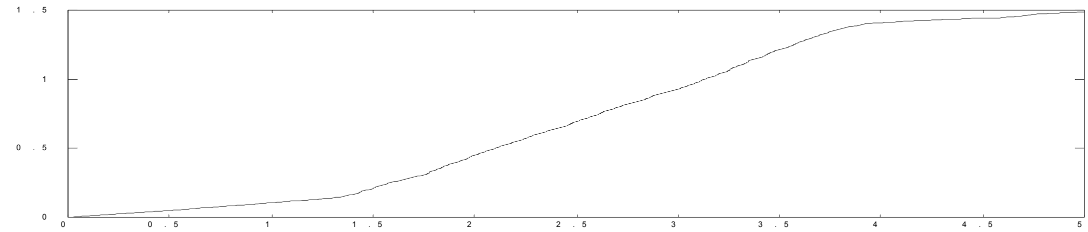
Processamento no Tempo - Derivar



Manuel A. E. Baptista
Ernesto R. Afonso

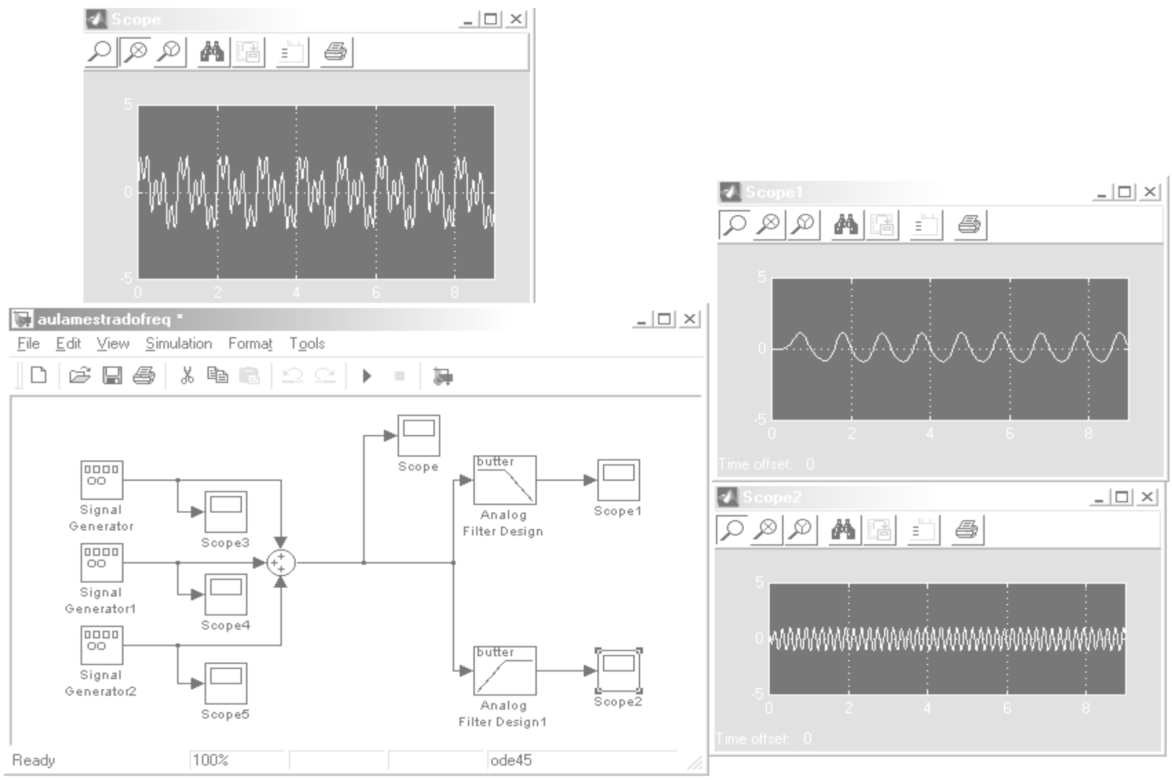
Processamento no Tempo - Integrar





Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

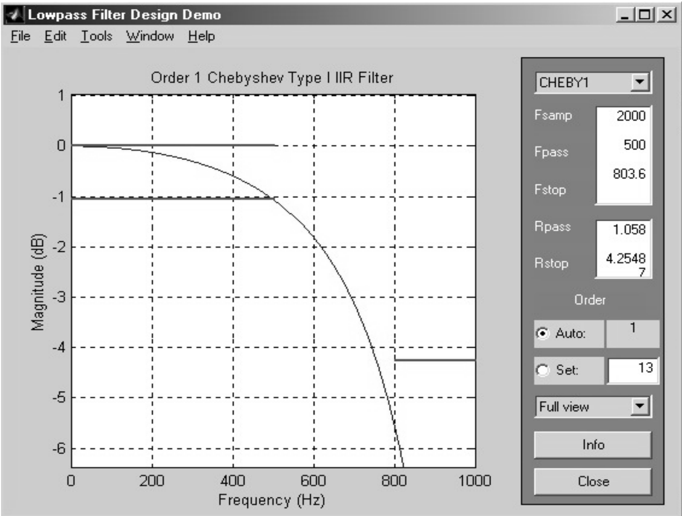
Processamento na Frequência - Filtragem

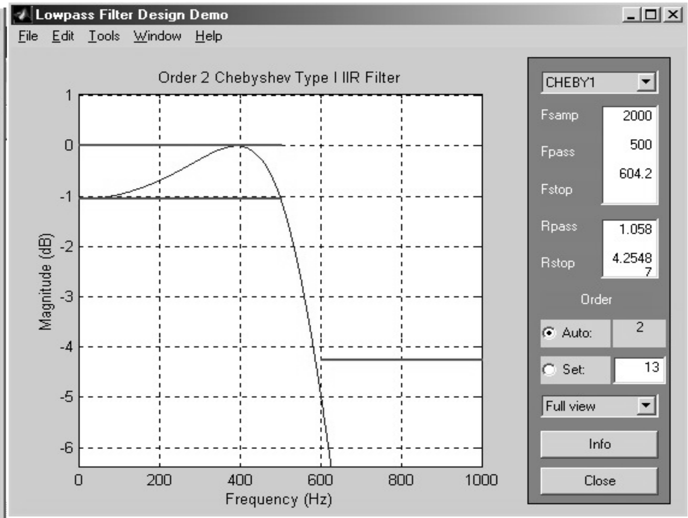


Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Processamento na Frequência - Ordem de um filtro

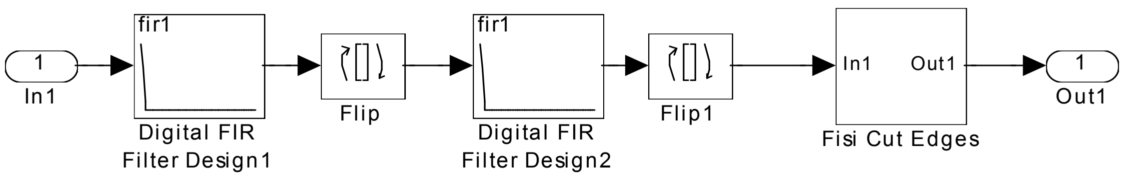
- 1ª Ordem
  - Fpass = 500 Hz
- 2ª Ordem
  - Fpass = 500 Hz





Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Processamento na Frequência - Filtro sem atraso



Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Teorema da Amostragem

How fast must we sample * a continuous signal to preserve its info content?

Ex: train wheels in a movie.

25 frames (=samples) per second.

Train starts ➡ wheels 'go' clockwise.

Train accelerates ➡ wheels 'go' counter-clockwise.

Porquê?

Frequency misidentification due to low sampling frequency.

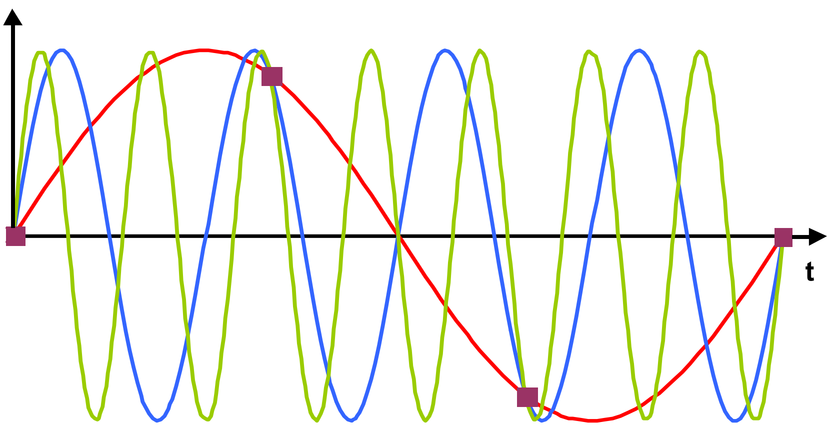
* Sampling: independent variable (ex: time) continuous → discrete.

Quantisation: dependent variable (ex: voltage) continuous → discrete.

Here we'll talk about uniform sampling.

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Teorema da Amostragem (2)



— $s(t) = \sin(2\pi f_0 t)$

■ $s(t)$ @ f_S

$f_0 = 1$ Hz, $f_S = 3$ Hz

— $s_1(t) = \sin(8\pi f_0 t)$

— $s_2(t) = \sin(14\pi f_0 t)$

$s(t)$ @ f_S represents exactly all sine-waves $s_k(t)$ defined by:

$$s_k(t) = \sin(2\pi (f_0 + k f_S) t) \quad , \quad |k| \in \mathbb{N}$$

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Teorema da Amostragem (3)

Theo*

A signal s(t) with maximum frequency f_{MAX} can be recovered if sampled at frequency $f_S > 2\, f_{MAX}$.

* Multiple proposers: Whittaker(s), Nyquist, Shannon, Kotel'nikov.

Naming gets confusing !

Nyquist frequency (rate) $f_N = 2\, f_{MAX}$ *or* f_{MAX} *or* $f_{S,MIN}$ *or* $f_{S,MIN}/2$

Example

$$s(t) = 3\cdot\underbrace{\cos(50\ \pi\, t)}_{F_1} + 10\cdot\underbrace{\sin(300\ \pi\, t)}_{F_2} - \underbrace{\cos(100\ \pi\, t)}_{F_3}$$

Condition on f_S?

$F_1 = 25$ Hz, $F_2 = 150$ Hz, $F_3 = 50$ Hz

f_{MAX}

$\boxed{f_S > 300 \text{ Hz}}$

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Teorema da Amostragem - Domínio da Frequência

- **Tempo & Frequência**: two complementary signal descriptions. Signals seen as "projected' onto time or frequency domains.

Example

Ear + brain act as frequency analyser: audio spectrum split into many narrow bands $\Rightarrow$ low power sounds detected out of loud background.

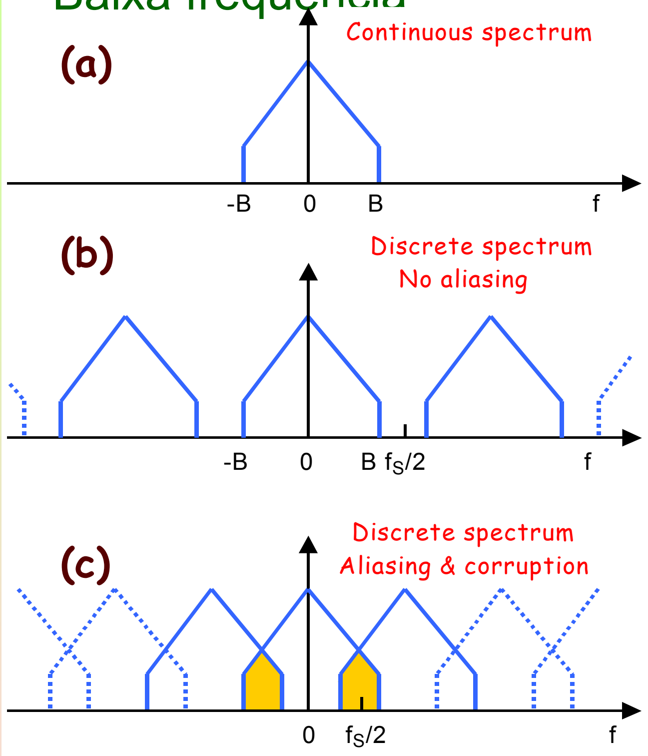
- **Bandwidth**: indicates rate of change of a signal. High bandwidth $\Rightarrow$ signal changes fast.

Atenção: formal description makes use of "negative" frequencies !

minus 50 Hz??

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Teorema da Amostragem – Sinais de Baixa frequência



(a) Band-limited signal:
frequencies in [-B, B] (f_{MAX} = B).

(b) Time sampling ➡ frequency repetition.

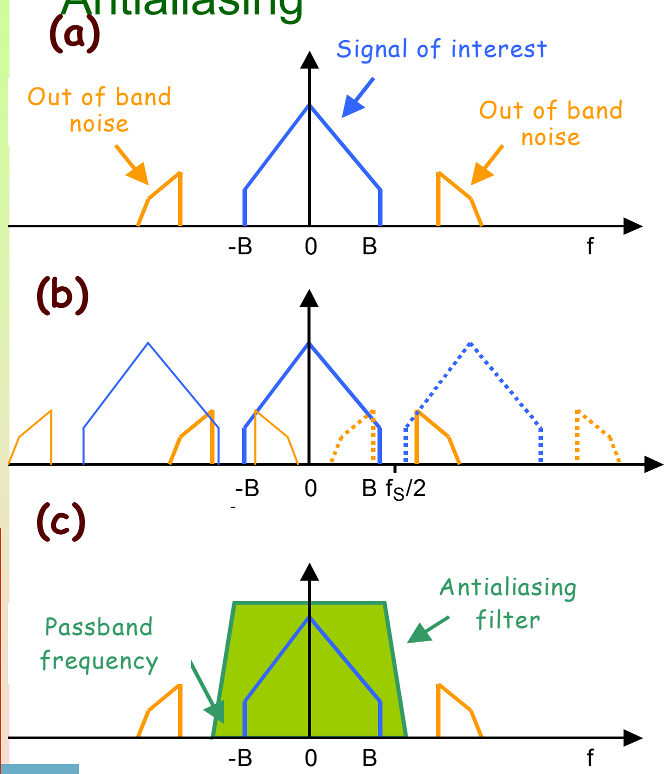
$f_S > 2$ B ➡ no aliasing.

(c) $f_S \leq 2$ B ➡ aliasing !

Aliasing: signal ambiguity in frequency domain

Manuel A. E. Baptista
Ernesto R. Afonso

Conversão A/D: Teorema da Amostragem – Filtro Antialiasing



(a),(b) *Out-of-band* noise can aliase into band of interest. Filter it before!

(c) Antialiasing filter

Passband: depends on bandwidth of interest.

Attenuation A_{MIN} : depends on

- ADC resolution (number of bits N).

$$A_{MIN,\ dB} \sim 6.02\ N + 1.76$$

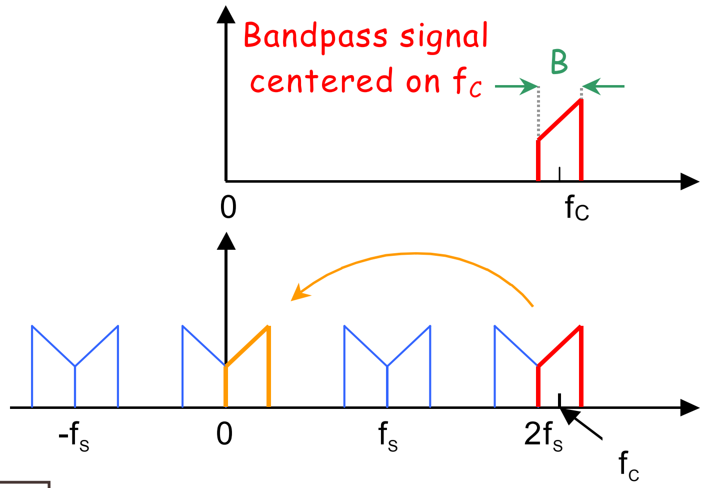
- Out-of-band noise magnitude.

Other parameters: ripple, stopband frequency...

Manuel A. E. Baptista
Ernesto R. Afonso

Conversão A/D: Teorema da Amostragem - *Under-sampling*

Using spectral replications to reduce sampling frequency f_S req'ments.



$$\boxed{\frac{2 \cdot f_C + B}{m+1} \le f_S \le \frac{2 \cdot f_C - B}{m}}$$

$m \in \mathbb{N}$ selected so that $f_S > 2B$

Example

f_C = 20 MHz, B = 5MHz

Without under-sampling f_S > 40 MHz.

With under-sampling f_S = 22.5 MHz (m=1);

= 17.5 MHz (m=2); = 11.66 MHz (m=3).

Advantages

- Slower ADCs / electronics needed.
- Simpler antialiasing filters.

Manuel A. E. Baptista
Ernesto R. Afonso

Departamento de Informática

Conversão A/D: Teorema da Amostragem - *Over-sampling*

Oversampling : sampling at frequencies $f_S >> 2\, f_{MAX}$.

Over-sampling & averaging may improve ADC resolution

(i.e. SNR)

$$\boxed{f_{OS} = 4^w \cdot f_S}$$

f_{OS} = over-sampling frequency,

w = additional bits required.

Each additional bit implies over-sampling by a factor of four.

Caveat

It works for:

- **white noise** with amplitude sufficient to change the input signal randomly from sample to sample by at least LSB.
- Input that can take all values between two ADC bits.

Manuel A. E. Baptista
Ernesto R. Afonso

Departamento de Informática

Conversão A/D: Parâmetros dum ADC

Different applications have different needs.

1. Number of bits N (~*resolution*) ✓
2. Data throughput (~*speed*)
3. Signal-to-noise ratio (SNR) ✓
4. Signal-to-noise-&-distortion rate (SINAD) ✓
5. Effective Number of Bits (ENOB) ✓
6. Spurious-free dynamic range (SFDR)
7. Integral non-linearity (INL)
8. Differential non-linearity (DNL)
9. …

Radar systems

Communication

Imaging / video

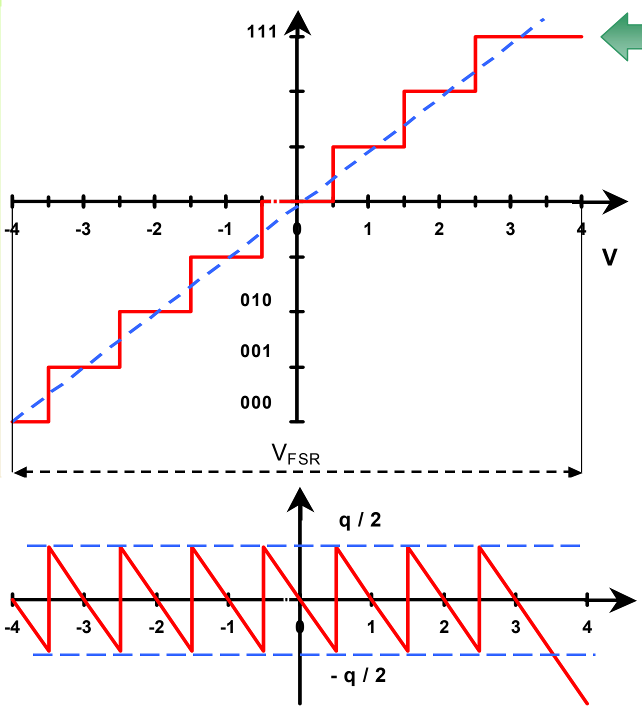
Static distortion

Dynamic distortion

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Parâmetros dum ADC – N.º de bits N

Continuous input signal digitized into 2^N **levels**.



Uniform, bipolar transfer function (N=3)

Quantisation step $q = \dfrac{V_{FSR}}{2^N}$

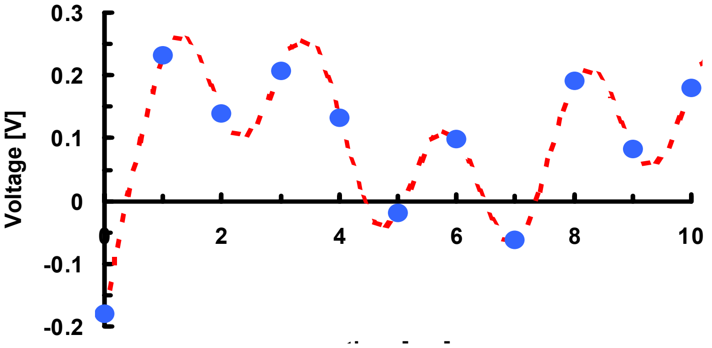
Ex: V_{FSR} = 1V , N = 12 ➡ q = 244.1 µV

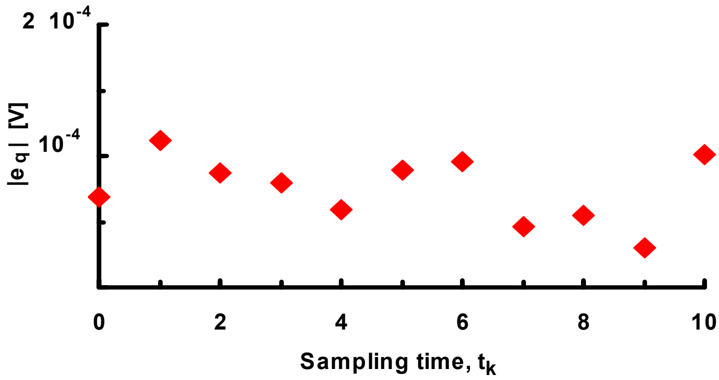
LSB
- Voltage (= q)
- Scale factor (= $1 / 2^N$)
- Percentage (= $100 / 2^N$)

Quantisation error

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Parâmetros dum ADC – Erro de Quantização





- Quantisation Error e_q in [-0.5 q, +0.5 q].
- e_q limits ability to resolve small signal.
- Higher resolution means lower e_q.

QE for
N = 12
V_{FS} = 1

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Parâmetros dum ADC – SNR dum ADC ideal

$$\boxed{\overline{SNR}_{ideal} = 20 \cdot \log_{10}\left(\frac{RMS(input)}{RMS(e_q)}\right)} \quad (1)$$

Also called SQNR
(signal-to-quantisation-noise ratio)

Assumptions

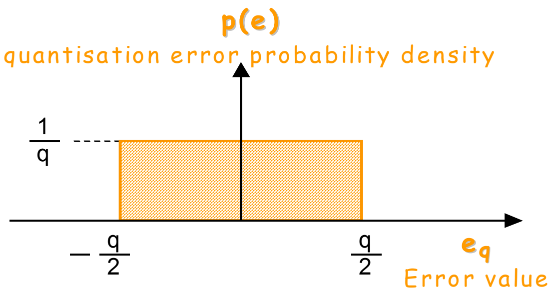
- Ideal ADC: only quantisation error e_q (**p(e)** constant, no stuck bits…)
- e_q uncorrelated with signal.
- ADC performance constant in time.

$$RMS(input) = \sqrt{\frac{1}{T} \cdot \int_0^T \left(\frac{V_{FSR}}{2} \cdot \sin(\omega t)\right)^2 dt} = \frac{V_{FSR}}{2\sqrt{2}}$$

Input(t) = ½ V_{FSR} sin(ω t).

$$RMS(e_q) = \sqrt{\int_{-q/2}^{q/2} e_q^2 \cdot p(e_q) de_q} = \frac{q}{\sqrt{12}} = \frac{V_{FSR}}{2^N \cdot \sqrt{12}}$$

(sampling frequency $f_S = 2\, f_{MAX}$)



Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

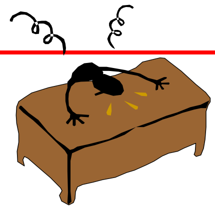
Conversão A/D: Parâmetros dum ADC – SNR dum ADC ideal (2)

Substituting in (1) :

$$\boxed{\overline{SNR}_{ideal} = 6.02 \cdot N + 1.76 \ [dB]} \quad (2)$$

One additional bit ➡ SNR increased by 6 dB

Real SNR lower because:



- Real signals have noise.
- Forcing input to full scale unwise.
- Real ADCs have additional noise (aperture jitter, non-linearities etc).

Actually (2) needs correction factor depending on ratio between sampling freq & Nyquist freq. Processing gain due to oversampling.

Manuel A. E. Baptista
Ernesto R. Afonso

Conversão A/D: Parâmetros dum ADC – Reais

SNR: (sine_in RMS)/(ADC out_noise RMS), with
out_noise = output - (DC + first 5 input harmonics) output components.

SINAD: (sine_in RMS)/(ADC out_noise_2 RMS), with
out_noise_2 = output - (DC output component).

SNR and SINAD often confused in specs.

ENOB : N from (2) when setting SNR ideal = SINAD,
i.e. ENOB = (SINAD – 1.76 dB) / 6.02.
➡ Actual number of bit available to an equivalent ideal ADC

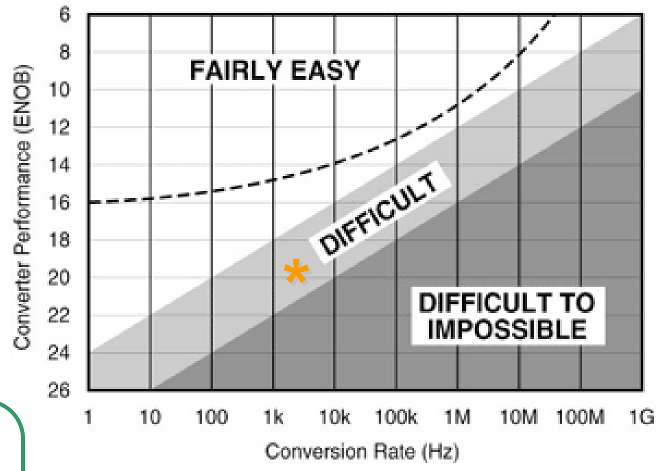
Example

12-bit ADC chip, 68 dB SINAD in specs ➡ 11-bit ideal ADC.

Manuel A. E. Baptista
Ernesto R. Afonso

Conversão A/D: Escolha do tipo de ADC

Speed & resolution:
a tradeoff.



High resolution (bit #)

- Higher cost & dissipation.
- Tailored onto DSP word width.

High speed

- Large amount of data to store/analyse.
- Lower accuracy & input impedance.

* DIFFICULT area moves down & right every year. Rule of thumb: 1 bit improvement every 3 years.

Oversampling & averaging (see 2).
Dithering (= adding small random noise before quantisation).

} may increase SNR.

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Formatos Digitais

Positional number system with base b:

$$\underbrace{[\,..\, a_2\, a_1\, a_0}_{\text{Integer part}} \,.\, \underbrace{a_{-1}\, a_{-2}\, ..\,]_b}_{\text{Fractional part}} = .. + a_2 b^{2} + a_1 b^{1} + a_0 b^{0} + a_{-1} b^{-1} + a_{-2} b^{-2} + ..$$

Important bases: **10 (decimal)**, **2 (binary)**, 8 (octal), 16 (hexadecimal).

Early computers (ex: ENIAC) mainly base-10 machines. Mostly turned binary in the '50s.

Benefits →
a) less complex arithmetic h/w;
b) less storage space needed;
c) simpler error analysis.

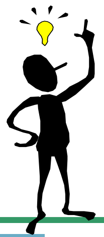
Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Formatos Digitais - Aritmética Digital

Increasing number of applications requires decimal arithmetic. Ex: Banking, Financial Analysis.

MAS

- Common decimal fractional numbers only approximated by binary numbers. Ex: 0.1 → infinite recurring binary fraction.
- Non-integer decimal arithmetic software emulation available *but* often too slow.



IEEE 754, 1985: binary floating point arithmetic standard specified

IEEE 854, 1987: standard expanded to include decimal arithmetic.

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Formatos Digitais – Binário de Virgula Fixa

Ex: 3-bit formats

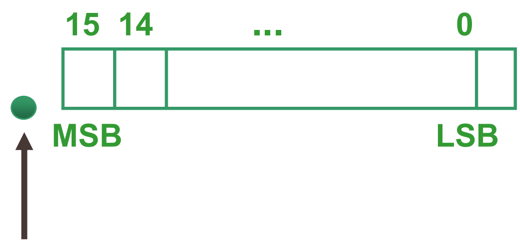
Represent integer or fractional binary numbers.

Unsigned integer		Offset-Binary		Sign-Magnitude		Two's complement	
7	111	4	111	3	011	3	011
6	110	3	110	2	010	2	010
5	101	2	101	1	001	1	001
4	100	1	100	0	000	0	000
3	011	0	011	0	100	-1	111
2	010	-1	010	-1	101	-2	110
1	001	-2	001	-2	110	-3	101
0	000	-3	000	-3	111	-4	100

Decimal equivalent

Binary representation

Sign bit



Fractional point (DSPs)

NB: Constant gap between numbers.

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Formatos Digitais – Binário de Virgula Flutuante

PROBLEM

Wide variety of floating point hardware in ‘60s and ‘70s,

⇨ different ranges, precision and rounded arithmetic.

William Kahan: *“Reliable portable software was becoming more expensive to develop than anyone but AT&T and the Pentagon could afford”*.

IEEE 754 standard

Formats & methods for binary floating-point arithmetic.

Definition of IEEE 754 standard between 1977 and 1985. *De facto* standard before 1985 !

Note: *NOT* the easiest h/w choice!

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Formatos Digitais – Binário de Virgula Flutuante (2)

IEEE 754 standard

Precision

Single (32 bits)

Double (64 bits)

Double-extended (≥80 bits)

31	30 … 23	22 … 0
s	e	f
MSB		LSB

e = exponent, offset binary,-126 < e < 127

s = sign, 0 = pos, 1 = neg

f = fractional part, sign-magnitude + hidden bit

Coded number $x = (-1)^{s} \cdot 2^{e} \cdot 1.f$

Single precision range

Max = $3.4 \cdot 10^{38}$

Min = $1.175 \cdot 10^{-38}$

NB: Variable gap between numbers.

Large numbers ⇨ large gaps; small numbers ⇨ small gaps.

Manuel A. E. Baptista
Ernesto R. Afonso
Departamento de Informática

Conversão A/D: Formatos Digitais – Efeitos do *Finite word-length*

Overflow : arises when arithmetic operation result has one too many bits to be represented in a certain format.

$$\text{Dynamic range}_{dB} = 20 \log_{10} \left[\frac{\text{largest value}}{\text{smallest value}} \right]$$

- Fixed point ~ 180 dB
- Floating point ~1500 dB

High dynamic range ⇒ wide data set representation with no overflow.

NB: Different applications have different needs.
Ex: telecomms: 50 dB; HiFi audio: 90 dB.

Conversão A/D: Formatos Digitais – Efeitos do *Finite word-length (2)*

Round-off: error caused by rounding math calculation result to nearest quantisation level.

⇒ Big concern for real numbers.

Example

0.1 not exactly represented (falls between two floating point numbers).

- *For integers* within ±16.8 million range: single-precision floating point gives no round-off error.
- *Outside* that range, integers are missing: gaps between consecutive floating point numbers are larger than integers.

Round-off error estimate:

Relative error = (floating - actual value)/actual value (depends on base).
The smaller the base, the tighter the error estimate.