Institut
Mines-Telecom

Speech and audio coding

Marco Cagnazzo,
cagnazzo@telecom-paristech.fr

MN910 – Advanced compression

Outline

Outline

Speech signal

- Non-stationary, but locally stationary (20 ms)
- Voiced sounds (vowels, some consonants) non-voiced (some consonants), other (transitions)
- Simple and effective prediction models:
  - Linear filters AR of impulsions for voiced sounds
  - Same filter on white noise for non-voiced sounds

Speech signal

Digitalization

- PCM (Pulse Code Modulation)
  - Bandwidth 200 Hz $\div$ 3400 Hz
  - Enough for understanding
  - Sampling at 8000 Hz : $F_s > 2F_{max}$
  - 8 bits per sample *rightarrow* 64 kbps
- Extended bandwidth
  - 50-200 Hz : more natural
  - and 3.4-7 kHz : better understanding

Speech signal

Coding standards

- G.711 (1972) PCM (no compression), 8 samples per ms codés sur 8 bits : 64 kbps
- G.721 (1984) ADPCM : 32 kbps
- G.728 (1991) CELP (Code Excited Linear Predictive), low latency: 16 kbps
- G.729 (1995) CELP, without latency constraint: 8kbps
- G.723.1 (1995) Encoder at 6.3 kbps

Speech signal

Normes de compression : Mobiles

GSM 06.10 (1988) RPE-LTP : Regular Pulse Excitation Long Term Predictor. 13 (22.8) kbps

GSM 06.20 (1994) “Half-Rate”. 5.6 (11.4) kbps

GSM 06.60 (1996) “ACELP”. 12.2 (22.8) kbit/s

GSM 06.90 (1999) Source/channel coding at variable bit-rate 4.75÷12.2 (11.4÷22.8)kbps (ACELP-AMR : Adaptive Multi Rate)

G.722 Wideband speec coder, rates 6.6, 8.85, 12.65 kbps (AMR-WB) ; further rates 15.85 and 23.85 kbps

Music signal

- Large dynamics: 90dB
- Locally stationary signal
- No simple model

Music signal

Compression

CD: sampling at 44.1 kHz, quantization on 16 bits: 705 kbps (mono)

MP3 : Audio part of MPEG-1. Three layers of increasing complexity, at 192, 128 and 96 kbps

AAC : Audio part of MPEG-2, reputed as the best audio encoder

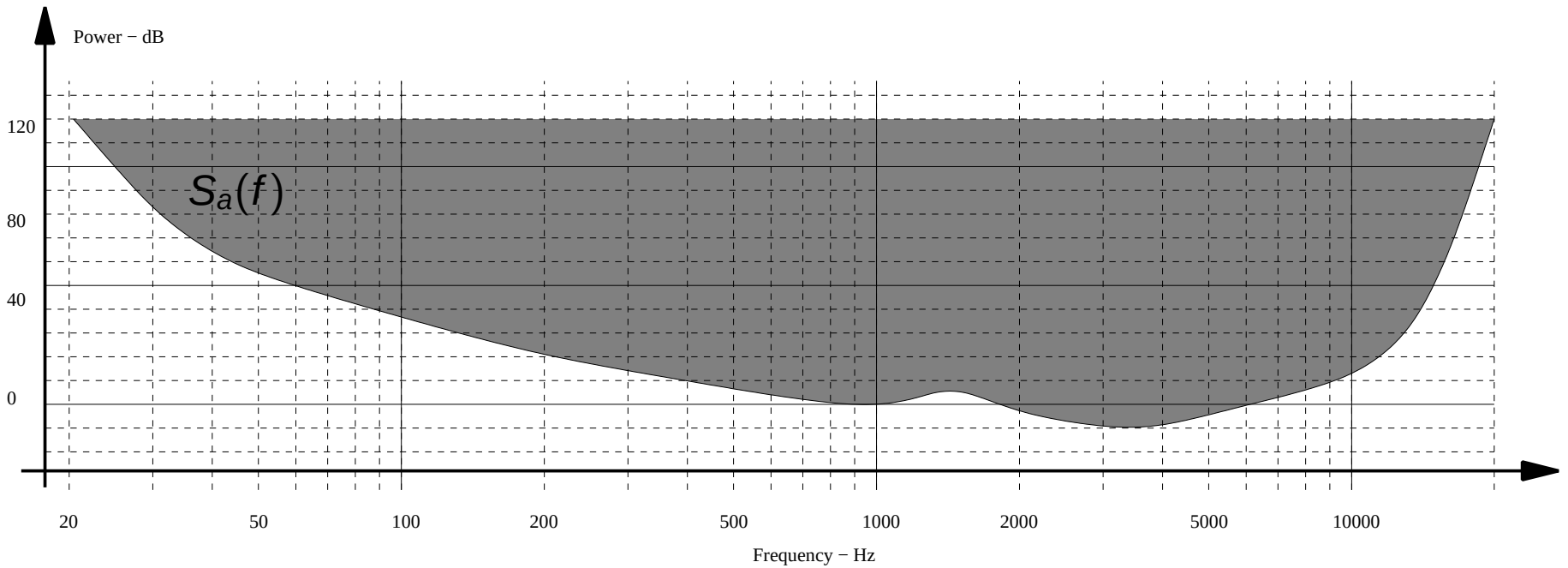
MPEG-4 : Sound object representation

Quality

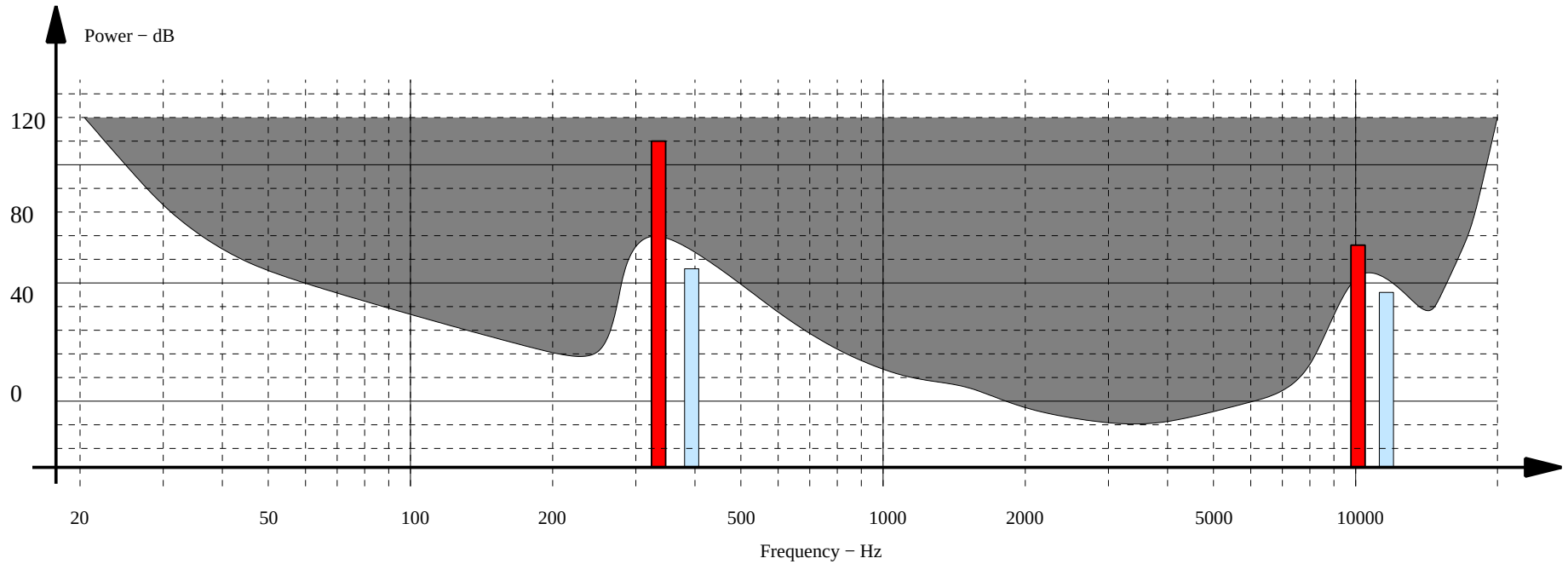
- Objective criteria (MSE) not satisfying
- Subjective stest:
  - Speech : understandability
  - Music: “transparency” criterion. Double blind method with triple stimuli and hidden reference (UIT-T BS.1116)

Outline

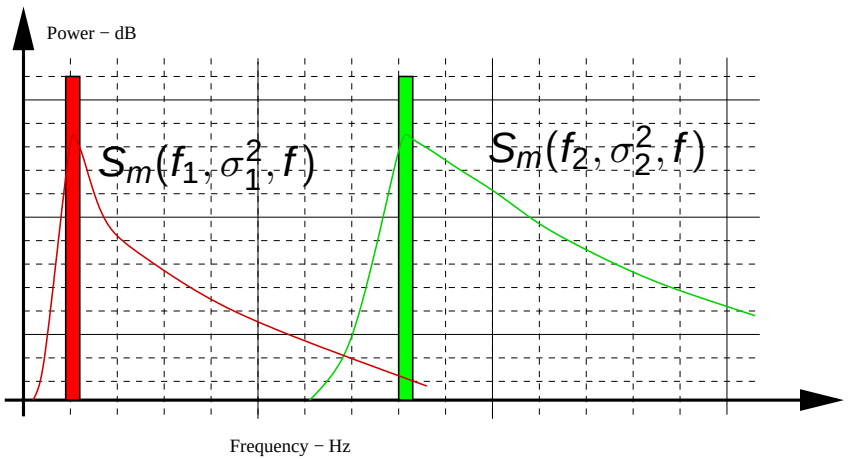
Audition threshold



Masking

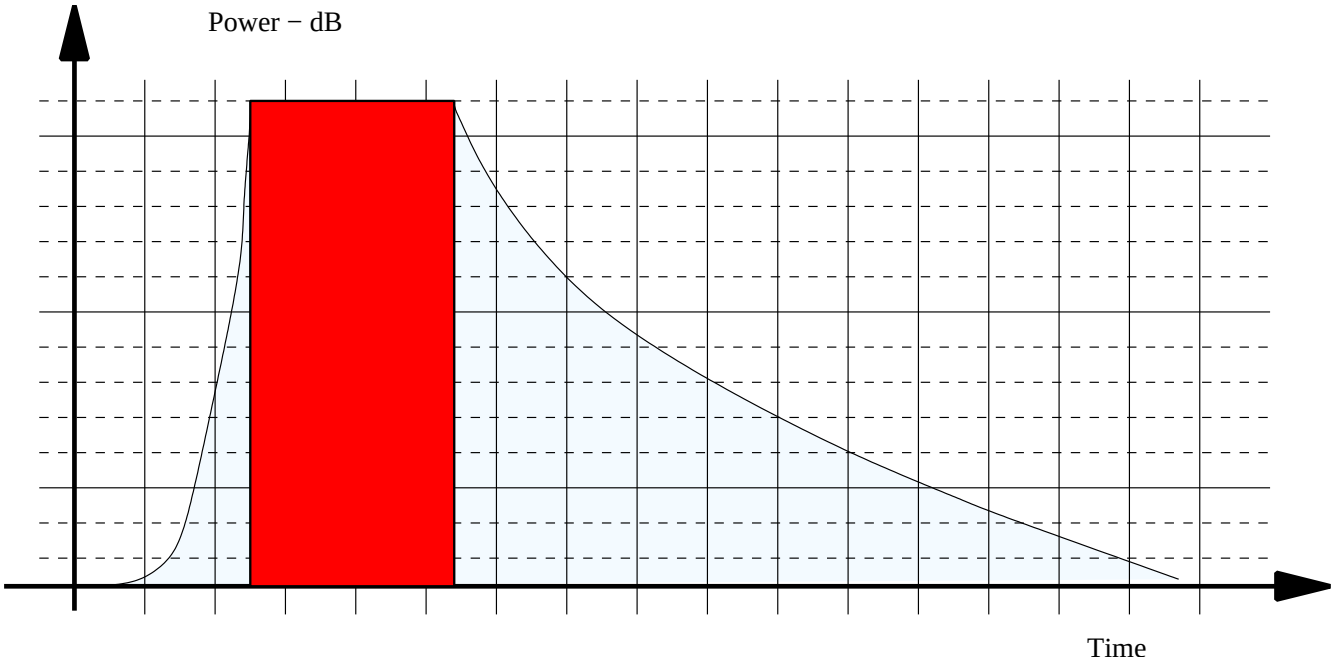

Power − dB
120
80
40
0
20
50
100
200
500
1000
2000
5000
10000
Frequency − Hz

Frequency masking functions $S_m(f_0, \sigma^2, f)$



- For a given f_0 and σ^2, $S_m(f)$ has a triangular shape
- The maximum is in $f = f_0$
- Masking index: $S_m(f, \sigma^2, f) - \sigma^2$

Time Masking



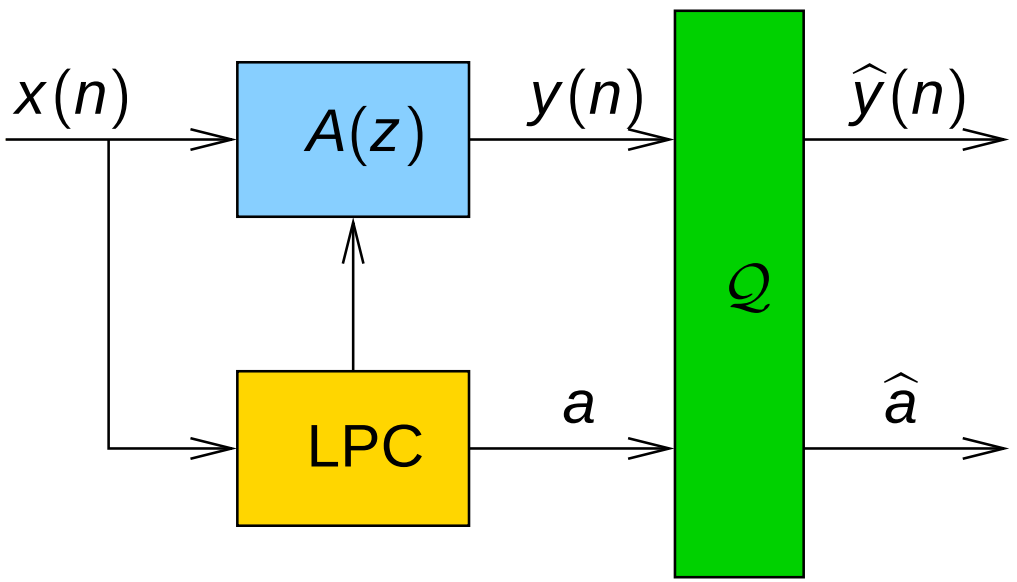
- Pre-masking : 2÷5 ms
- Post-masking : 100÷200 ms

Outline

LPC10 encoder at 2.4 kbps

Linear Prediction Coding avec 10 échantillons

- Only for teaching!
- Window $N = 160$
- Sample prediction using $P = 10$ previous samples
- Scheme



LPC10 encoder at 2.4 kbps

Filter

$$y(n) = x(n) - x_P(n) \qquad x_P(n) = \sum_{k=1}^{P} h_k x(n-k)$$

$$y(n) = \sum_{k=0}^{P} a_k x(n-k) \qquad a_0 = 1 \qquad a_k = -h_k$$

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_P z^{-P} \quad Y(z) = A(z)X(z)$$

A is computed by minimisation of the residual power in the current window

LPC10 encoder at 2.4 kbps

Filter

$$\mathbf{c} = [r_X(1)\, r_X(2)\, \ldots\, r_X(P)]^T$$

$$\mathbf{R} = \begin{bmatrix} r_X(0) & r_X(1) & \ldots & r_X(P-1) \\ r_X(1) & r_X(0) & \ldots & r_X(P-2) \\ \ldots & \ldots & \ldots & \ldots \\ r_X(P-1) & r_X(P-2) & \ldots & r_X(0) \end{bmatrix}$$

$$\mathbf{a} = -\mathbf{R}^{-1}\mathbf{c}$$

Estimation of r_X : using N samples of the current window

$$\widehat{r}_X(k) = \frac{1}{N-k} \sum_{n=k}^{N-1} x(n)x(n-k) \qquad k \in \{0, 1, \ldots, P\}$$

LPC10 encoder at 2.4 kbps

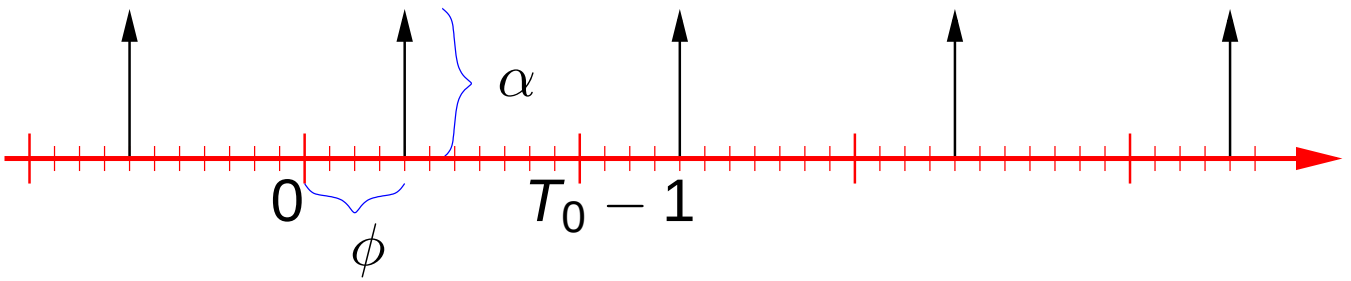
Non-voiced sounds

- If X was deprived of all correlation, Y would be white noise
- We dont send Y: the decoder will filter WN using $1/A(z)$
- Resulting in unaudible phase noise
- WN is a good model only for non-voiced suonds
- For voice sounds we have residual periodicity

LPC10 encoder at 2.4 kbps

Non-voiced sounds : Filtering WN with $1/A(z)$
Voiced sounds : Filtering $1/A(z)$ of a pulse train:

$$\widehat{y}(n) = \alpha \sum_{m \in \mathcal{Z}} \delta\left(n - mT_0 + \phi\right)$$



LPC10 encoder at 2.4 kbps

- Estimation of auto-correlation $\widehat{r}_x(k)$
- $\widehat{r}_x(0)$ estimates opower σ_Y^2 or α
- If $\widehat{r}_x$ decreases towards zero, it is a non-voiced sound
- If $\widehat{r}_x$ is periodical, it is a voiced sound with period T_0

LPC10 encoder at 2.4 kbps

Each 20 ms we send:

- The P filter coefficients $P = 10$ over 3 or 4 bits. $b_P = 36$ bits
- One bit for voiced/non-voiced $b_V = 1$ bit
- For voiced
  - $b_\alpha = 6$
  - The period T_0: $b_{T_0} = 7$
- For non-voiced:
  - $b_{\sigma^2} = 6$ bits

LPC10 encoder at 2.4 kbps

Total

$$\begin{aligned} R &= \left(b_P + b_V + b_{T_0} + b_\alpha\right) / (20\text{ms}) \\ &= (36 + 1 + 6 + 7)/0.02 \\ &= 2500\text{bps} \end{aligned}$$

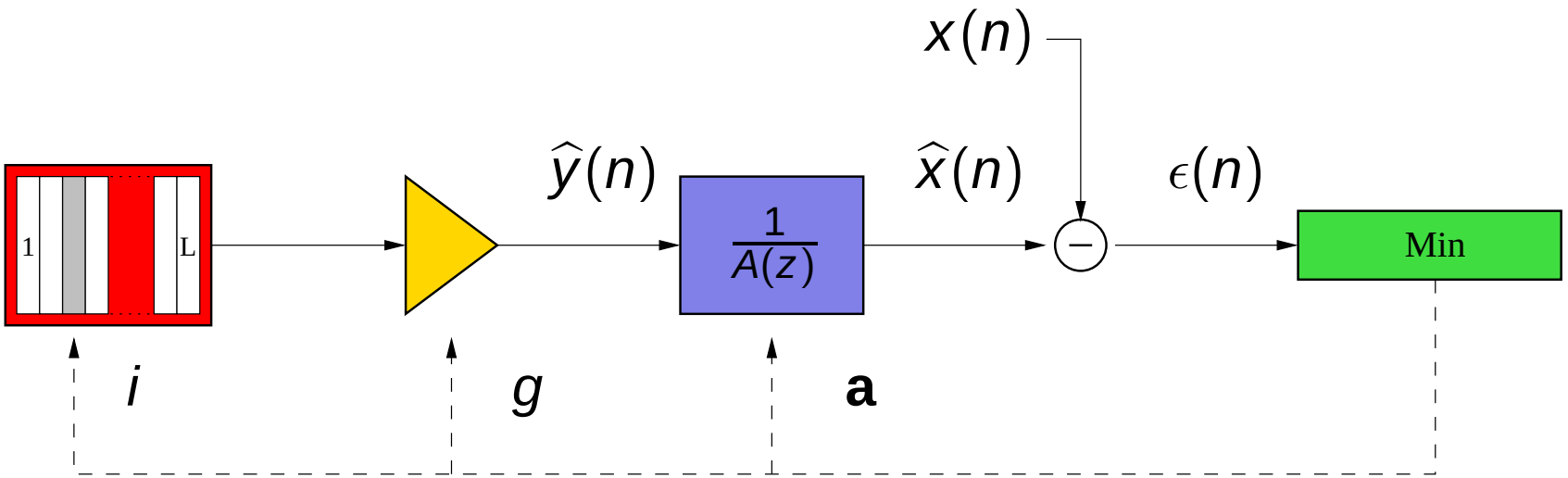
or

$$\begin{aligned} R &= \left(b_P + b_V + b_{\sigma^2}\right) / (20\text{ms}) \\ &= (36 + 1 + 6)/0.02 = 2150\text{bps} \end{aligned}$$

$R = 2.4$ kbps

CELP encoder

- Find a filter and the prediction error signal
- Filter $A(z)$: LPC idea
- Error $y(n)$: from a “GS-VQ” dictionary



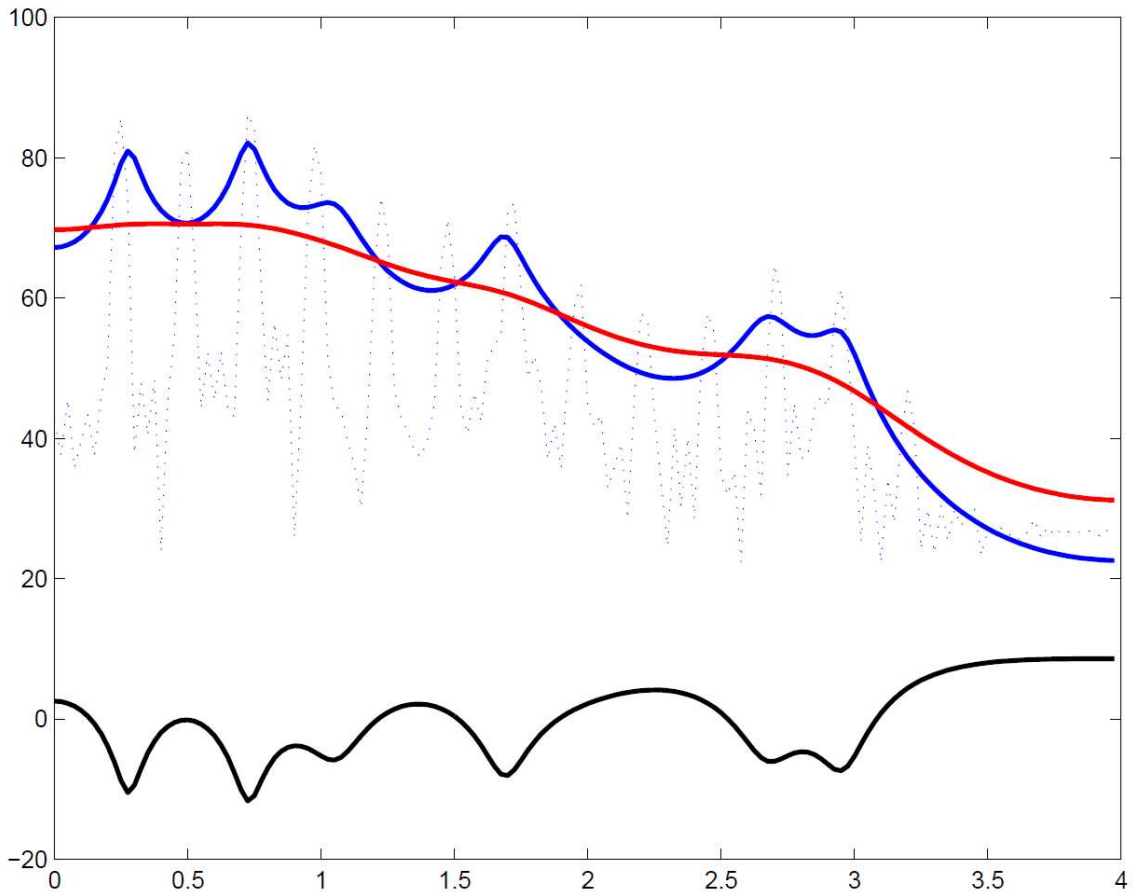
CELP encoder

- Filter coefficient coding
  - *Line Spectrum Pairs*: effective mathematical representation of $A(z)$
- Perceptual weighting function (WF)
  - Noise can be tolored where the signal is strong
  - The WF depends thus on $A(z)$

Perceptual weighting

- We use $W(z) = \frac{A(z)}{A(z/\gamma)}$ to weight the signal
- Noise can be higher in the formantic areas
- The filter $1/A(z)$ has peaks in the formantic areas
- The filter $1/A(z/\gamma)$, with $\gamma \in (0,1)$ has peaks at the sames frequencies, but smaller
  - Let p_i be the poles of $1/A(z)$
  - Thus γp_i are the poles of $1/A(z/\gamma)$, which are thus closer to the center

Perceptual weighting



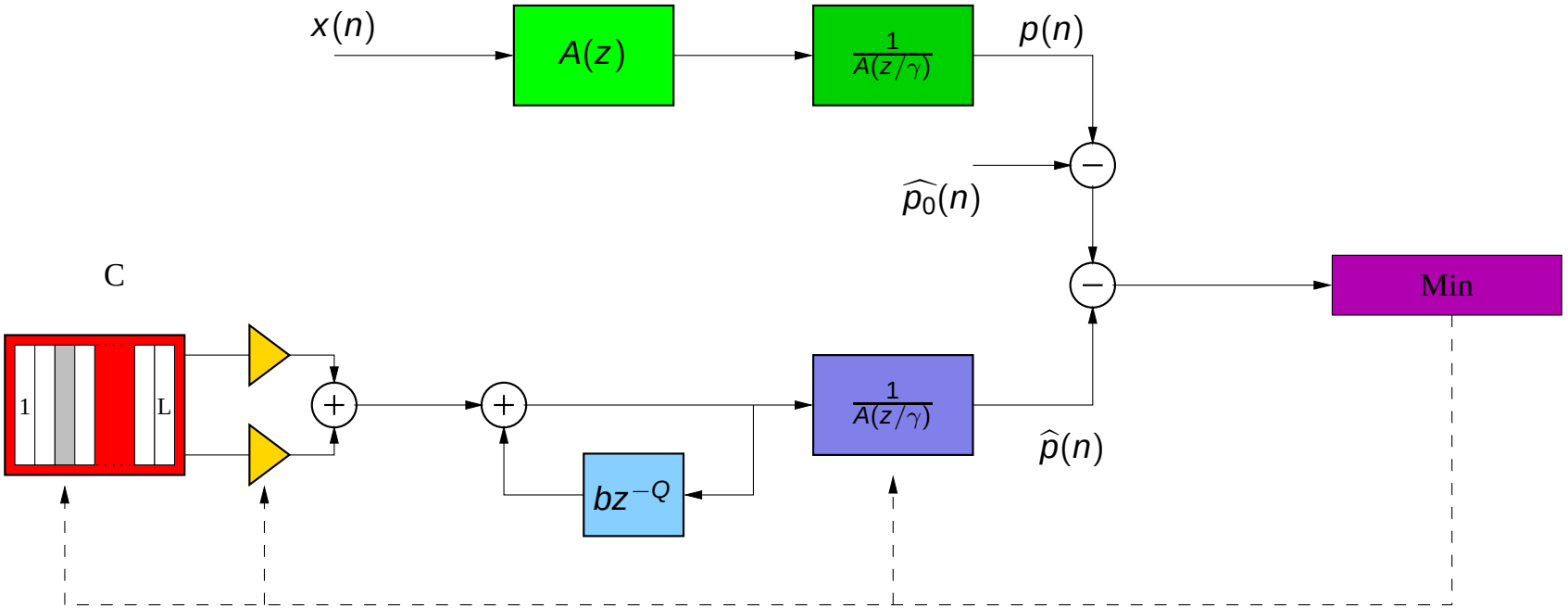
Blue : $1/A(z)$; Red : $1/A(z/\gamma)$; Black: $W(z) = A(z)/1/A(z/\gamma)$

CELP encoder

- Excitation model. It is the sum of $K = 2$ or $K = 3$ vectors
- Vector and gain selection
  - High complexity
  - Sub-optimal algorithms such as the standard iterative algorithm

CELP encoder

Global scheme



CELP encoder

Coding rate

- Filter coefficients, sent once per 10 ms :
  - $P = 10$, $b_P = 18$ bits
- Long-term predictor, sent once per 5 ms
  - Pitch coded on 7 bits
  - Puissance, codée sur 3 bits
- Residual, coded by GS-VQ, once per 5 ms :
  - *Shape* : 17 bits dictionary
  - *Gain* : 4 bits dictionary

$$R = 18/0.010 + (7 + 3 + 17 + 4)/0.005$$
$$= 8 \text{ kbps}$$

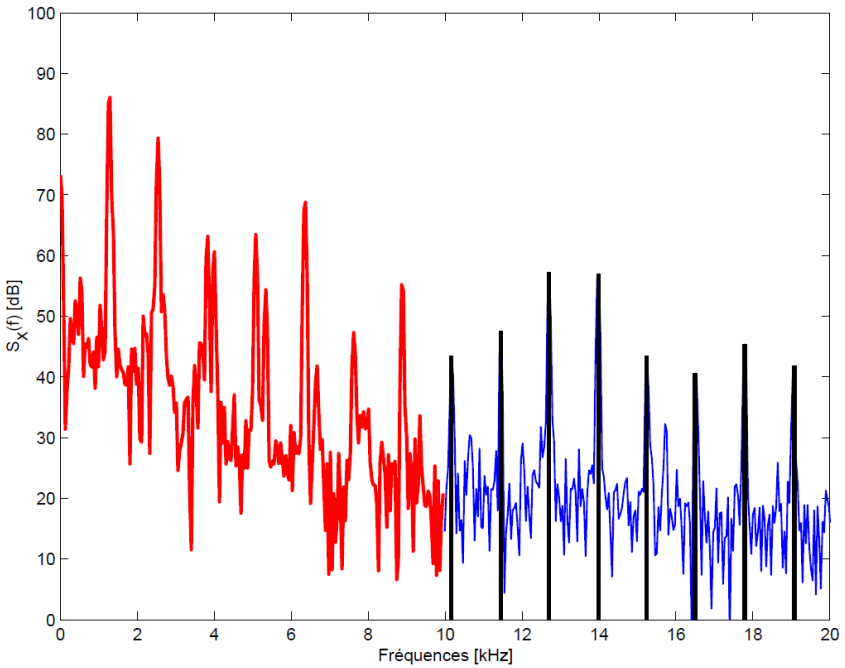
Encoder 3GPP AMR-WB

UIT-T G.722.2

- State of the art in speech coding
- Introduction of 50-200 Hz: more natural, presence effect
- Extension 3.4-7 kHz : better understandability
- ACELP-like encoder with:
  - Modification of perceptual weighting (wide-band)
  - Modification of pitch information
  - very large dictionary
- Rates: 6.6 to 23.85 kbps

Encoder 3GPP AMR-WB

- Harmonic exploitation
- Amplitude modulation



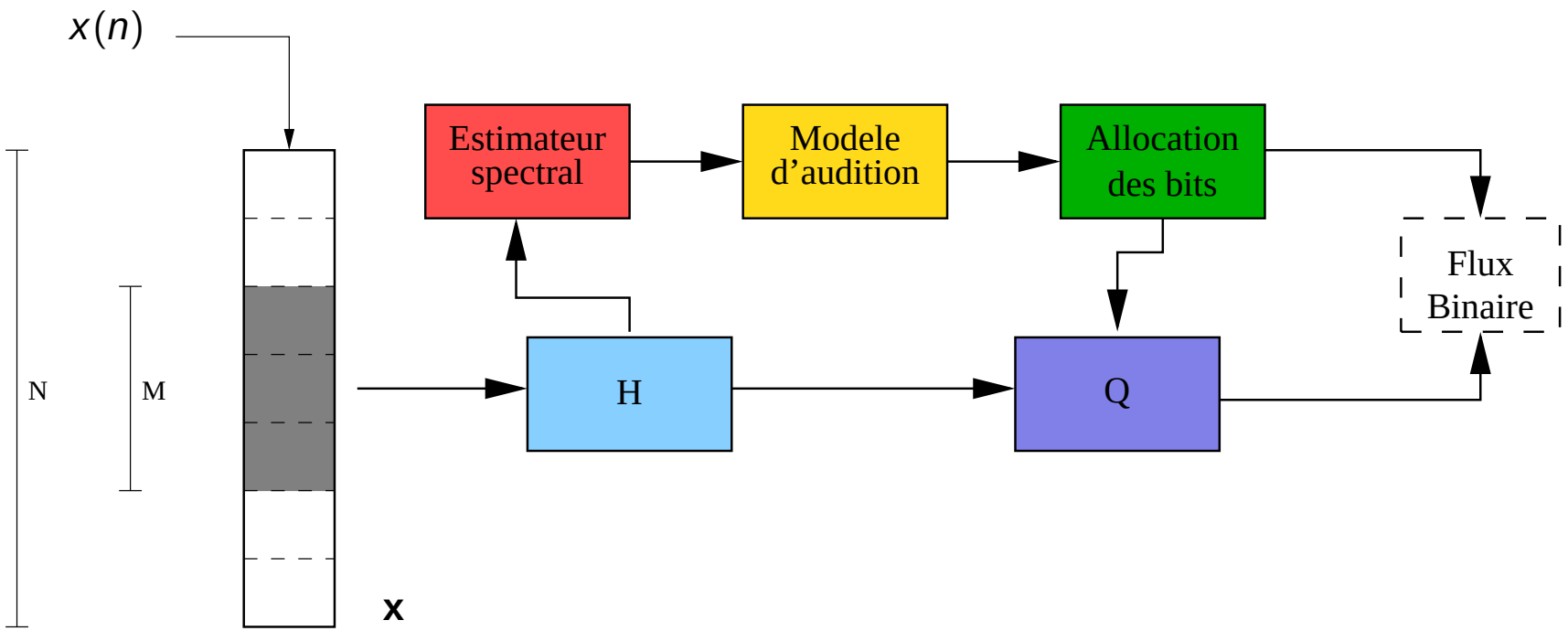
Outline

Perceptual coding

- Overlapping windows
- Buffer of N samples; M new samples are coded at time
- Examples : $M = 32$ et $N = 512$ (MP3) ; $M = 1024$ et $N = 2048$ (AAC)
- Three tools used for encoding the current window
  - Time-frequence analysis
  - Bit allocation (controlled by audion model)
  - Quantization and lossless coding

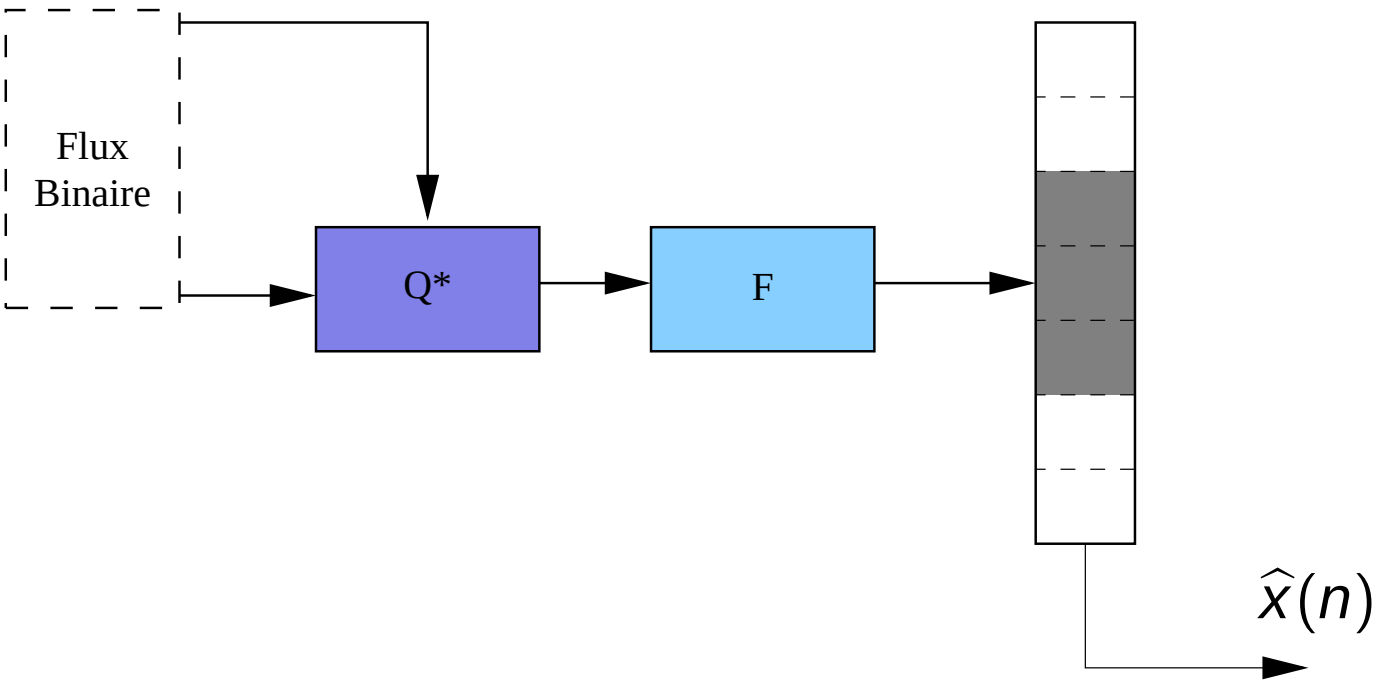
Perceptual coding

Coder



Perceptual coding

Decoder



MPEG-1/MP3

- *Transparent* music encoder: perfect subjective quality
- Three complexity layers
  - MP3 : third layer (max complexity)
- We coniser $f_e = 44.1$kHz

MPEG-1/MP3

TF transform

- 32 filters for the filterbank
- Uniform repartition of frequencies between 0 and 22 kHz
  - 700 Hz per sub-band
- Critical sampling and quasi-perfect reconstruction (SNR$>$90dB)
- A vector of 12 samples from the same subband is encoded jointly
- $\mathbf{y}_k$, it corresponds to $\approx$ 10ms

MPEG-1/MP3

- Subband vectors are normalized

$$\mathbf{y}_k = g_k \mathbf{a}_k$$

- g_k : scale factor, the larges coefficient. It is quantified on 6 bits
- $\mathbf{a}_k$: normalized vector
- The samples x of the current window are also used to compute the perceptual function $\widehat{S}_X(f)$ and the masking threshold $\Phi(f)$ based on psychoacustical models

MPEG-1/MP3

Rate allocation and quantization

- The signal to mask ratio is known for each subband
- The available bits are first given to subbands with the highest ratio, then to others
- For each subband, we can choose the number of bit to be used for the scalar uniform quantizer