SI350
Compression de signaux parole et musique

G. Richard

- ***Merci à Nicolas Moreau pour les transparents***

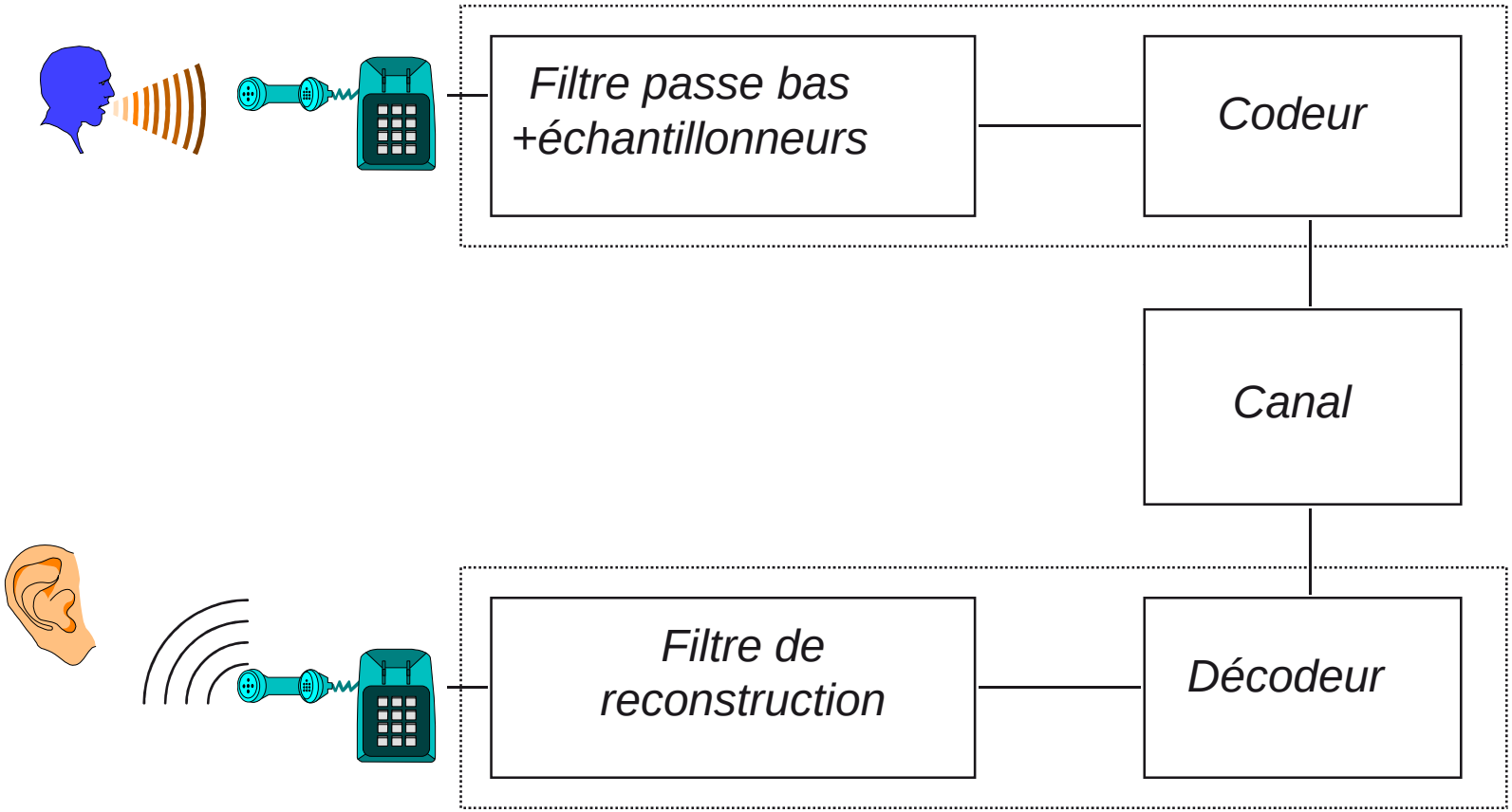
Contenu

- **Codage de la parole**
  - Généralités
  - Organismes de standard, normalisation
  - Codeurs de parole
    - MIC, MICDA, ACELP
  - Quelques extensions
    - Extension de bande
    - Codeur universel parole + musique (USAC)

Le codage audio: généralités

- **Techniques permettant la réduction des coûts de transmission ou de stockage des signaux numériques**
  - par exploitation des **caractéristiques statistiques** des signaux
  - Par exploitation des phénomènes de **masquage** des systèmes perceptuels humains.
  - Éventuellement par exploitation d'un **modèle de production** (ex. codage de la parole)

Système de compression

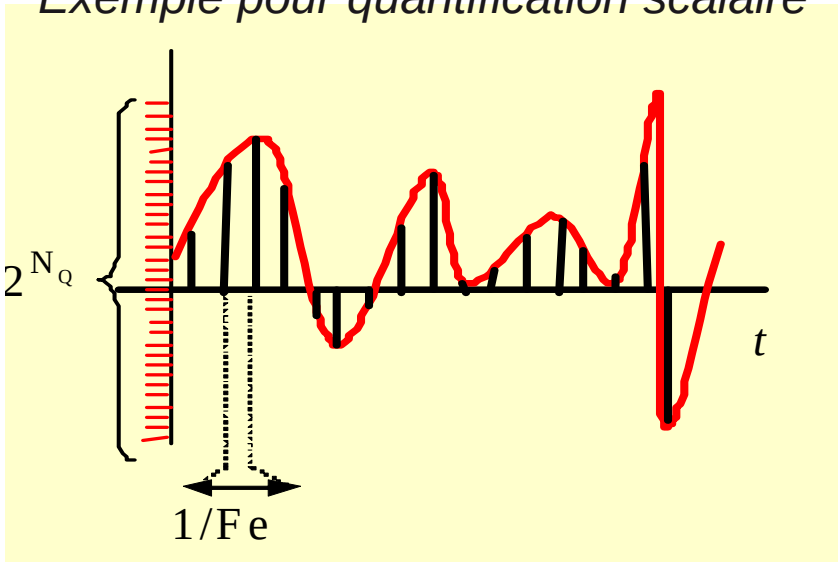


Notion de débit

- **Le débit d'un codeur est:**

$$Débit = F_e \times N_Q \qquad N_Q = Nbre\ de\ bits$$

Exemple pour quantification scalaire



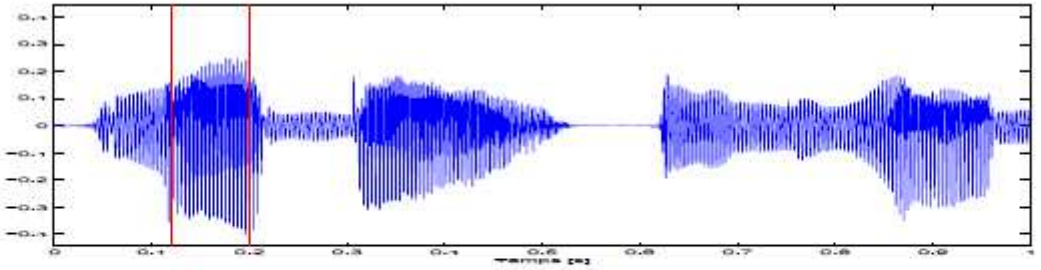
Codage de la parole: Généralités

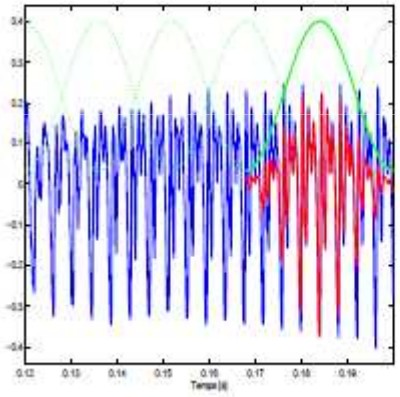
- **Parole: Bande téléphonique :**
  - B = 300 Hz - 3.3 kHz, Fe = 8 kHz
  - Débit : 13 bits @ 8 kHz → 104 kbit/s à 6 kbit/s
- **Parole: Bandeélargie :**
  - B = 50 Hz - 7 kHz, Fe = 16 kHz
  - Débit : 14 bits @16 kHz → 224 à 24 kbit/s
- **Musique**
  - Bande Hi-Fi, Qualité “CD” :
  - B = 20 Hz - 20 kHz, Fe = 44.1 kHz
  - Débit : 16 bits @ 44.1 kHz → 705 à 64 kbit/s
  - *(1.4 Mbit/s → 96 kbit/s en stéréo)*

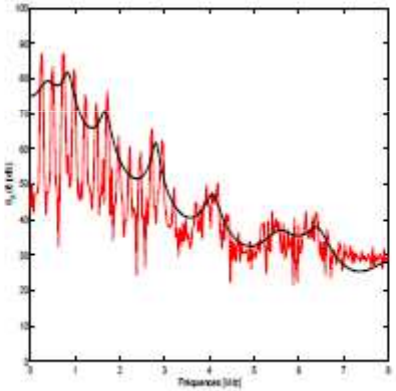
Un “bon” codeur : résultat d’un compromis

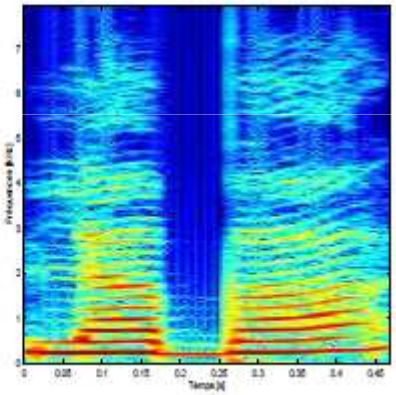
- **Débit**
  - Codeur monodébit, multidébit, hiérarchique (“scalable”, train binaire en plusieurs “couches”)
- **Qualité**
  - Tests d’intelligibilité (Absolute Category Rating,Degradation Category Rating, etc.)
- **Complexité**
  - (exprimée en MIPS sur un DSP, en % sur un PC)
  - Implantation en virgule fixe, en virgule flottante
- **Retard de reconstruction**
  - (critique en “full duplex”)
  - Retard algorithmique : dépend de la durée d’une “fenêtre d’analyse”. Retard total 3 retard algorithmique < 150 ms
- **Tenue aux erreurs de transmission**
  - Communications avec les mobiles : codes correcteurs d’erreurs
  - Communications sur IP : problème de pertes de paquets

Exemple d'un signal de parole : "Là-bas il y a ..."









- **"Fenêtres d'analyse" recouvrantes 20 ms**
  - N = 160 (ou 320) si Fe = 8 (ou 16) kHz

Organismes de standardisation/normalisation

- **UIT-T (SC16/WP3/Q7-10)**
  - Parole en bande téléphonique sur réseaux fixes + paquets
  - Téléphonie, visiophonie, VoIP
- **ISO/IEC (JTC1/SC29/WG11)**
  - Musique en bande Hi-Fi
  - Baladeurs, DVD, TNT, streaming, broadcasting, Digital Radio Mondiale
- **ETSI, TIA, 3GPP**
  - Parole en bande téléphonique ouélargie/Musique sur réseaux mobiles
  - Téléphonie, streaming, ...
- **INMARSAT, OTAN, DoD**
  - Parole sur des réseaux particuliers
  - Transmissions par satellites, communications militaires

Parole en bande téléphonique

- **Réseau téléphonique commuté**
  - Recommandations UIT-T
    - **64 kbit/s** : G.711 en 1972 (PCM : Pulse Code Modulation)
    - **32 kbit/s** : G.721 en 1984 (ADPCM : Adaptive Differential PCM)
    - **16/24/32/40 kbit/s** : G.726 (version multidébit) ; G.727 (version à codes imbriqués)
    - **16 kbit/s** : G.728 en 1991 (LD-CELP : Low Delay Code Excited Linear Prediction)
    - **8 kbit/s** : G.729 en 1995 (CS-ACELP : Conjugate Structure Algebraic CELP)
    - **5.3/6.3 kbit/s** : G.723.1 en 1995

Parole en bande téléphonique: Applications

- **Réseau téléphonique commuté**
  - **64 kbit/s** G.711 : Réseau RNIS
  - **32 kbit/s** G.721 G.726 G.727 :
    - Transmissions par cables sous-marins, satellites
    - Norme européenne téléphone sans cordon *DECT(Digital Enhanced Cordless Telecommunications*)
    - Protocoles de transmission par paquets (Packetized Voiced Protocol)
  - **16 kbit/s** G.728 : Transmissions réseaux par paquets ATM ou IP
  - **8 kbit/s** G.729 : Liaisons satellites, MPEG-4
  - **5.3/6.3 kbit/s** G.723.1 : Visiophonie RTC

Parole en bande téléphonique (suite)

- **Communication avec les mobiles : ETSI en Europe (European Telecommunications Standards Institute)**
  - **1ere génération**
    - 13 (22.8) kbit/s **GSM 06.10** en 1988 (RPE-LTP : Regular Pulse Excitation Long Term Predictor)
  - **2eme génération**
    - 5.6 (11.4) kbit/s **GSM 06.20** “Half-Rate” en 1994
    - 12.2 (22.8) kbit/s **GSM 06.60** “Enhanced Full-Rate” en 1996 (ACELP)
    - 4.75 ⇨ 12.2 (11.4/22.8) kbit/s **GSM 06.90** en 1999 (ACELP-AMR : Adaptive Multi Rate)
  - **3eme génération** : 3GPP (*3rd Generation Partnership Project*)
    - AMR-WB (wide band) : collaboration UIT/ETSI G.722.2

Parole en bande téléphonique (suite)

- **Autres applications**
  - Communications sécurisées :
    - 2.4 kbit/s US DoD (LPC10) en 1976 (puis en 1997)
    - 4.8 kbit/s : US DoD en 1991
  - Communications avec les mobiles par satellite : INMARSAT 6 kbit/s : IMBE en 1990
  - Applications militaires (normes OTAN) 4800 ... 800, 600, 400 bits/s

Parole en bandeélargie

- **Introduction 50-200 Hz** ⇨ voix plus naturelle, amélioration de l'effet de présence
- **Extension 3.4-7 kHz** ⇨ plus grande intelligibilité
- **Communications de groupe : téléconférence, téléphonie sur haut-parleurs**
  - 48/56/64 kbit/s : **G.722** en 1986 (SB-MICDA)
  - 24/32 kbit/s : **G.722.1** en 1999 (LCTC : Low Complexity Transform Coder)
- **Téléphonie mobile**
  - 6.6/ ... /23.85 kbit/s : **AMR-WB, G.722.2** en 2002

”MPEG Unified Speech and Audio Codec”

- **ISO/IEC (JTC1/SC29/WG11) : ”Call for Proposals” en 2007**
  - Statut actuel: Working draft (Version 7): “MPEG-D”
  - Débits : 12 ⇨ 24 kbit/s en mono, 16 ⇨ 64 kbit/s en stéréo
- **Etat de l’art :**
  - Parole : 3GPP AMR-WB
  - Musique : MPEG-4 HE AAC
- **Codeur hybride : 2 modes distincts + un classifieur**
- **Principales difficultés :**
  - assurer des transitions rapides et douces entre parole, musique et signaux mixtes
  - exploiter différents types de fenêtres, de différentes dimensions suivant le type des signaux

Codeurs de parole : plan de l'exposé

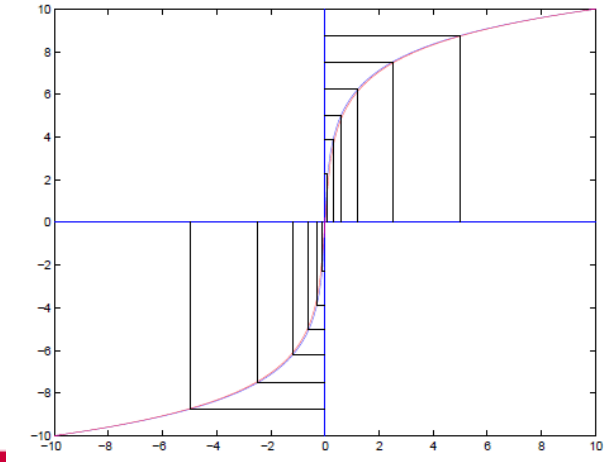
- **Codeur PCM à 64 kbit/s : QS non uniforme**
- **Codeur ADPCM à 32 kbit/s : QS prédictive**
- **Introduction d'un modèle de production de type "source/filtre"**
  - Détermination des coefficients du filtre
    "Analyse LPC" toutes les 20 ms (N = 160)
- **Détermination de l'entrée du filtre**
  - Codeur LPC10 à 2.4 kbit/s (Bruit blanc ou train d'implusions)
  - Codeur CELP à 8 kbit/s : QV + schéma en "boucle fermée"
- **Nouveaux codeurs en bandeélargie : le codeur AMR-WB**

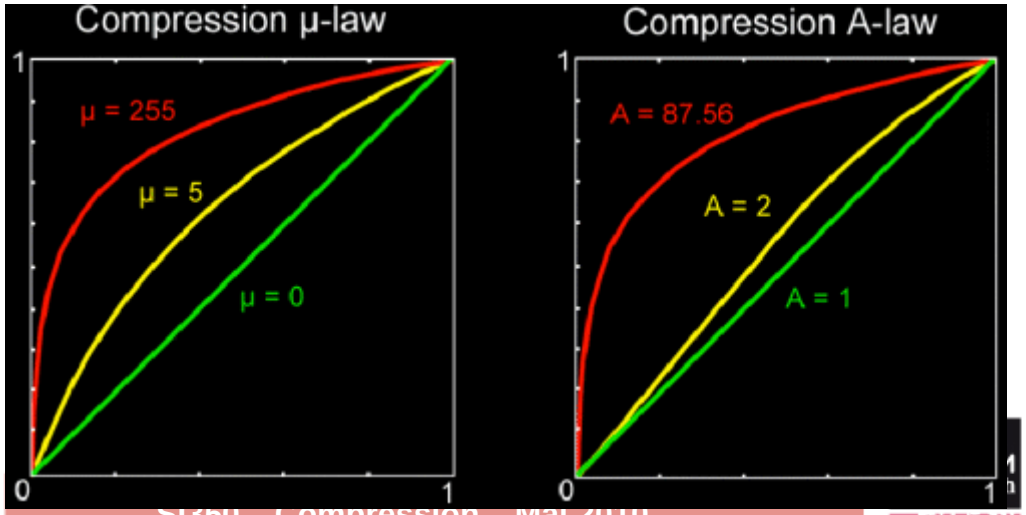
Codeur MIC (64 kbit/s)

- **QS mal adaptée à des signaux présentant des variations de puissance instantanée importantes**
- **Maintenir un RSB ≈ cst**
- **Utilisation d'une caractéristique non-linéaire**

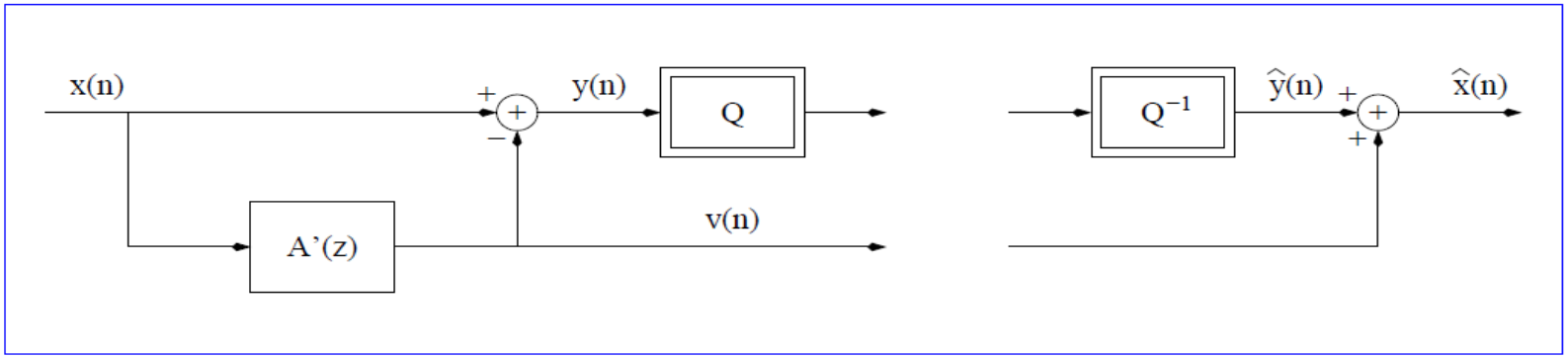
$$y(n) = f[x(n)],\ \hat{y}(n) = QS_{unif}[y(n)],\ \hat{x}(n) = f^{-1}[\hat{y}(n)]$$

  - "Loi A" en Europe, "Loi μ" aux US
- **"COMPANDING" = "COMPressing" + "expANDING"**





Codeur MICDA : QS prédictive



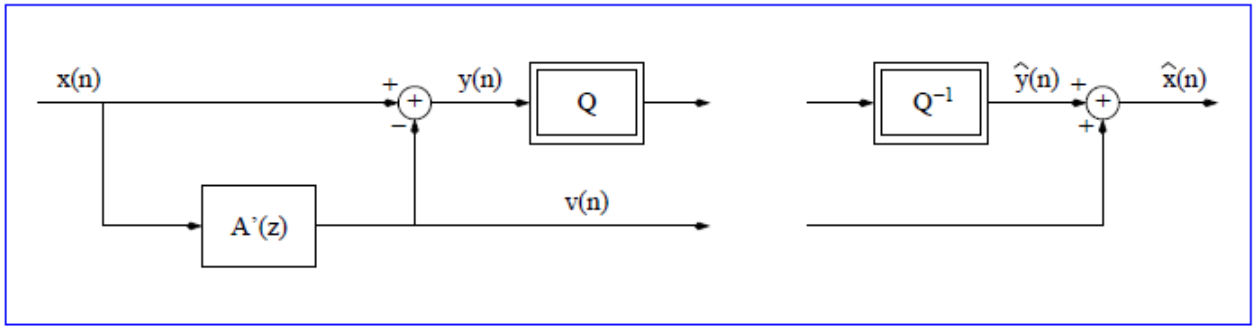
- Erreur de quantification : $q(n) = y(n) - \hat{y}(n)$
- Erreur de reconstruction : $\bar{q}(n) = x(n) - \hat{x}(n) = q(n)$
- Erreur de prédiction : $y(n) = x(n) - v(n) = x(n) + \sum_{i=1}^{P} a_i \; x(n-i)$
- Chaque fenêtre d'analyse $[x(0) \cdots x(N-1)]$

$$\underline{a}^{opt} = \arg \min_{a_1 \cdots a_P} \sum_{n=0}^{N-1} y^2(n) = \arg \min_{a_1 \cdots a_P} \sum_{n=0}^{N-1} [x(n) + \sum_{i=1}^{P} a_i \; x(n-i)]^2$$

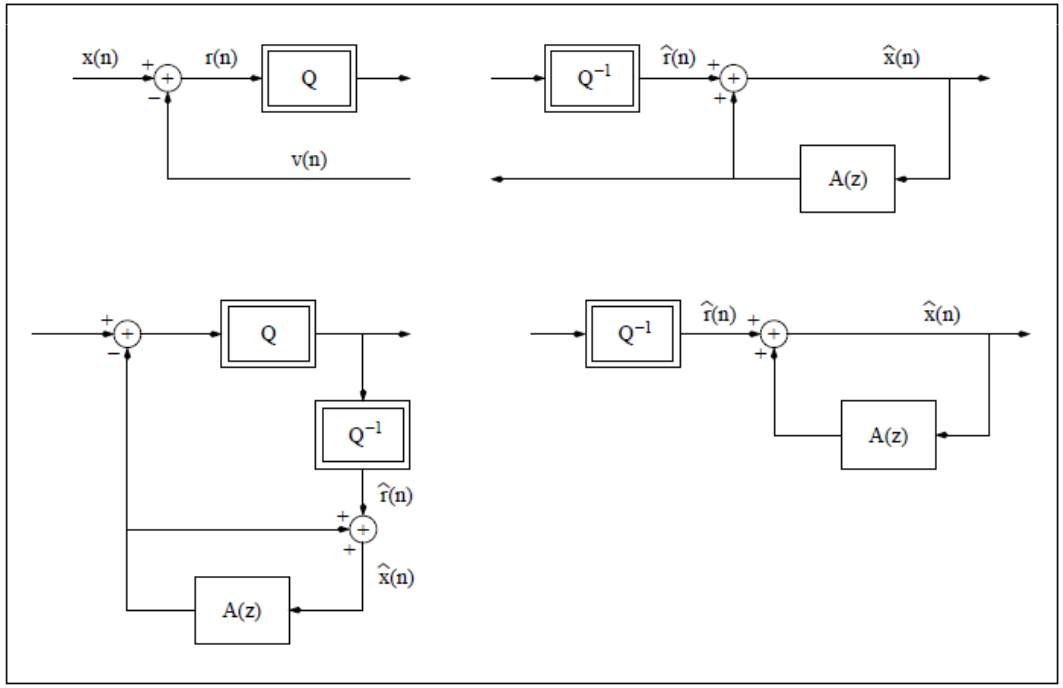
- Codeur à 32 kbit/s ⇨ quantifier $y(n)$ sur 4 bits

Codeur MICDA : QS prédictive en boucle fermée

- **Boucle ouverte**
  - Non réaliste



- **Boucle fermée**
  - Plus réaliste
  - Décodeur « local » / Décodeur distant



Théorie de la prédiction linéaire

- Connaissant le p.a. $X(n)$ on cherche $a_1 \cdots a_P$ minimisant

$$E\{Y^2(n)\} = E\{[X(n) + a_1 X(n-1) + \cdots + a_P X(n-P)]^2\}$$

- Minimisation de :

$$\sigma_Y^2 = \sigma_X^2 + 2\,\underline{r}^t \underline{a} + \underline{a}^t R \underline{a}$$

$$\underline{r} = \begin{bmatrix} r_X(1) \\ \vdots \\ r_X(P) \end{bmatrix} \quad R = \begin{bmatrix} r_X(0) & \cdots & r_X(P-1) \\ \vdots & \ddots & \vdots \\ r_X(P-1) & \cdots & r_X(0) \end{bmatrix}$$

- Solution : équations normales (de Yule-Walker)

$$\underline{a}^{opt} = -R^{-1}\underline{r}$$
$$\sigma_Y^2 = \sigma_X^2 + \underline{a}^t \underline{r}$$

Théorie de la prédiction linéaire, propriétés

- Le filtre $A(z) = 1 + a_1 z^{-1} + \cdots + a_P z^{-P}$ est le filtre `blanchissant"

- Si $X(n)$ est un p.a. AR d'ordre P_0, alors il existe un filtre totalement blanchissant dès que $P \geq P_0$

- Dans ce cas

$$S_Y(f) = |A(f)|^2 S_X(f) = \sigma_Y^2 \quad \Rightarrow \quad S_X(f) = \frac{\sigma_Y^2}{|A(f)|^2}$$

  - Toutes les racines de $A(z)$ à l'intérieur du cercle unité (filtre à ``déphasage minimal") ⇨ filtre $1/A(z)$ toujours stable

Faire une “Analyse LPC”

- A partir de $x(0) \cdots x(N-1)$ calculer

$$\hat{r}(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x(n)x(n+k) \quad \text{pour} \quad k = 0 \cdots P$$

- Construire $\hat{R}$ et $\underline{\hat{r}}$ à partir de $\hat{r}(k)$
- Déterminer les coefficients du filtre : $\underline{a} = -\hat{R}^{-1}\underline{\hat{r}}$
  et la puissance du signal ``résiduel'' : $\sigma_Y^2 = \hat{r}(0) + \underline{\hat{r}}^t \underline{a}$
- Filtrer $x(n)$ ⇨le signal résiduel $y(n)$

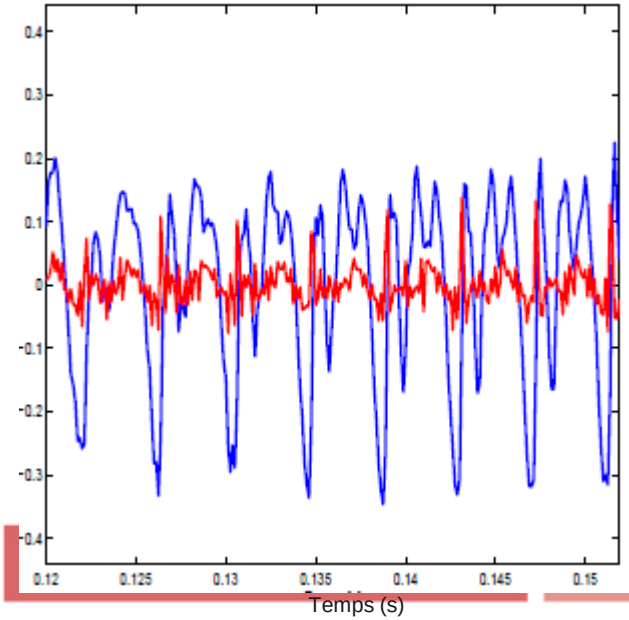
- Calculer le ``spectre LPC'' : $\hat{S}(f) = \sigma_Y^2 / |A(f)|^2$

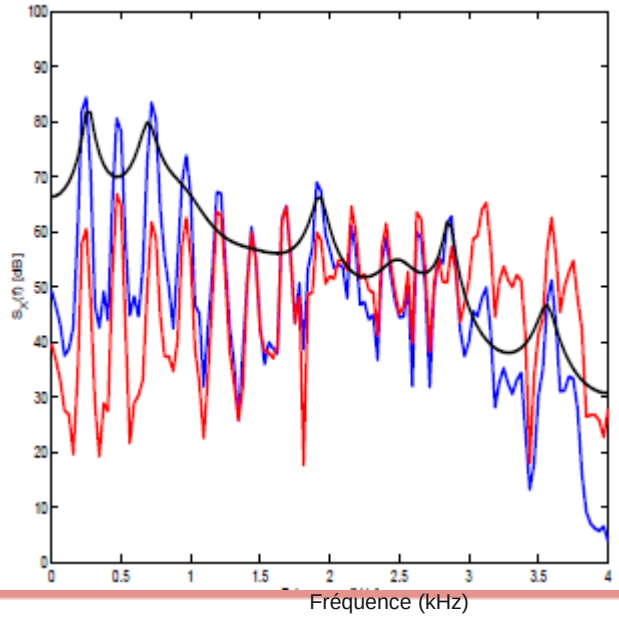
Résultat d'une ``analyse LPC''

- Signal de parole en bande élargie Fe= 16 kHz
  - En bleu : signal original $x(n)$
  - en rouge : signal résiduel $y(n)$

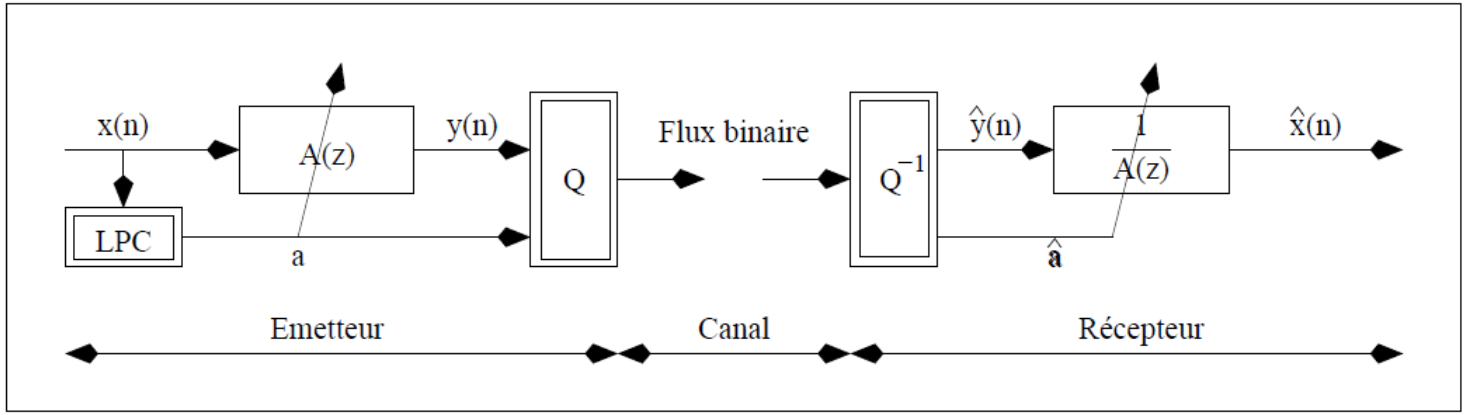
- Dans le domaine fréquentiel
  - En bleu : périodogramme de $x(n)$ = module au carré de la TFD
  - En noir : spectre LPC de $x(n)$ = ``enveloppe spectrale''
  - En rouge : périodogramme de $y(n)$

- Signal blanchi en partie, possibilité d'augmenter P mais $P \lesssim N/10$





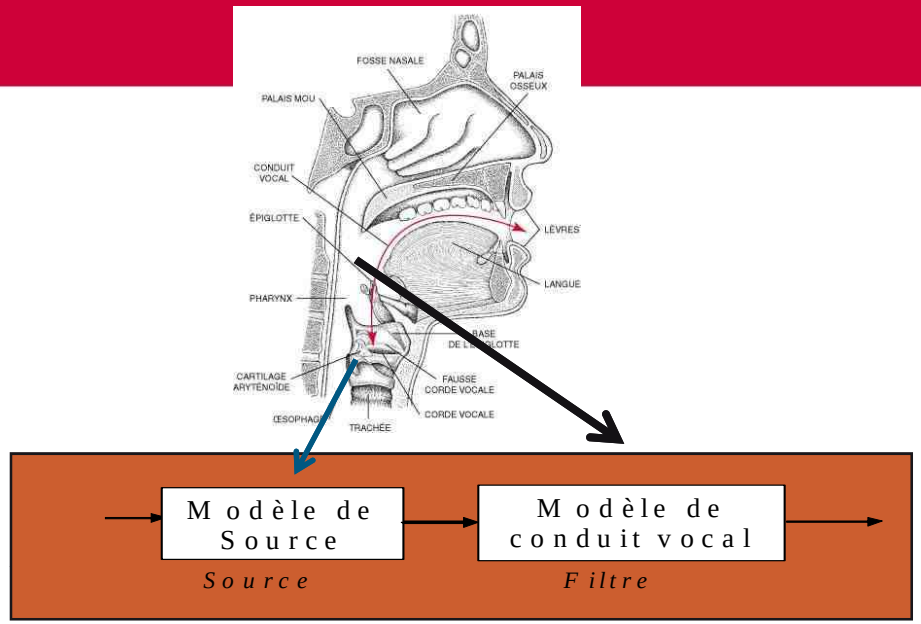
Codeurs de parole : recherche d'un modèle



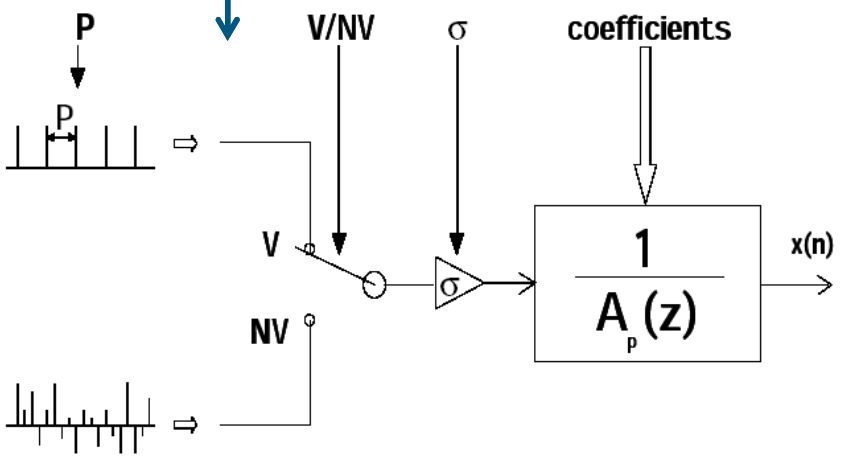
- **1ère étape** : modélisation de $x(n)$ par la réalisation d'un processus AR(P) ⇨ détermination des coefficients du filtre de synthèse $1/A(z)$

- **2ème étape** : modélisation de l'entrée du filtre de synthèse
  - *1ère tentative :* rendre $S_{\hat{X}}(f) \approx S_X(f)$ ⇨ « vocodeur »
  - Autres tentatives : codeurs ``hybrides'', codeurs ``analyse-par-synthèse''

LPC10: Utilisation d'un modèle source-filtre



- Exemple de modèle



Codeur LPC10 (2.4 kbit/s)

- Sons non voisés : $\hat{y}(n)$ = réalisation bruit blanc puissance σ_Y^2

$$S_{\hat{X}}(f) = \frac{\sigma_Y^2}{|A(f)|^2}$$

- Sons voisés : $\hat{y}(n) = \alpha \sum_k \lambda(n - kT_0 + \phi)$

$$S_{\hat{X}}(f) = \alpha^2 \sum_k \frac{\delta(f - kf_0)}{|A(kf_0)|^2}$$

- Débit (actualisation et transmission des paramètres du modèle
- toutes les 20 ms)
- $A(z) \approx$ 50x10x(3 ou 4) $\approx$ 1.8 kbit/s

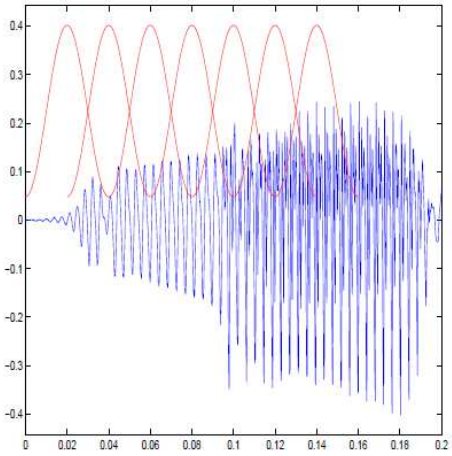
  QV des ``Line Spectrum Pairs''
- $\sigma_Y^2 \approx$ 50 x 6 $\approx$ 300 bit/s (couvrir 50dB par pas de 1 dB)
- Distinction voisé/non voisé = 50 bits/s
- Pitch : $50 \times \log_2(T_0^{max} - T_0^{min})$ = 350 bit/s

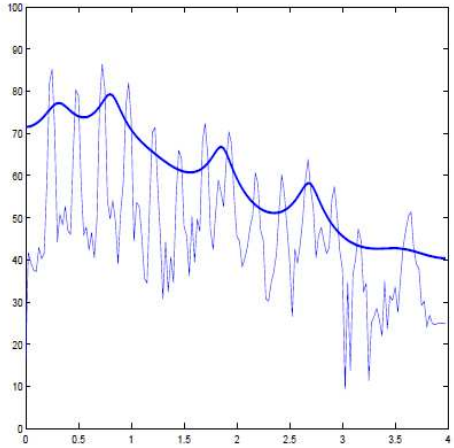
Line Spectrum Pairs (LSP)

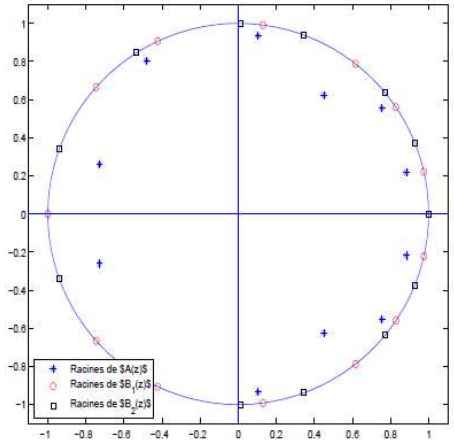
- Pb : codage des coefficients du polynôme $A(z) = 1 + \sum_{i=1}^{P} a_i z^{-i}$
  Mauvaises propriétés ⇨ recherche de nouvelles représentations

- Construction de deux polynômes
  $$B_1(z) = A(z) + z^{-P-1}A(z^{-1}) \text{ et } B_2(z) = A(z) - z^{-P-1}A(z^{-1})$$

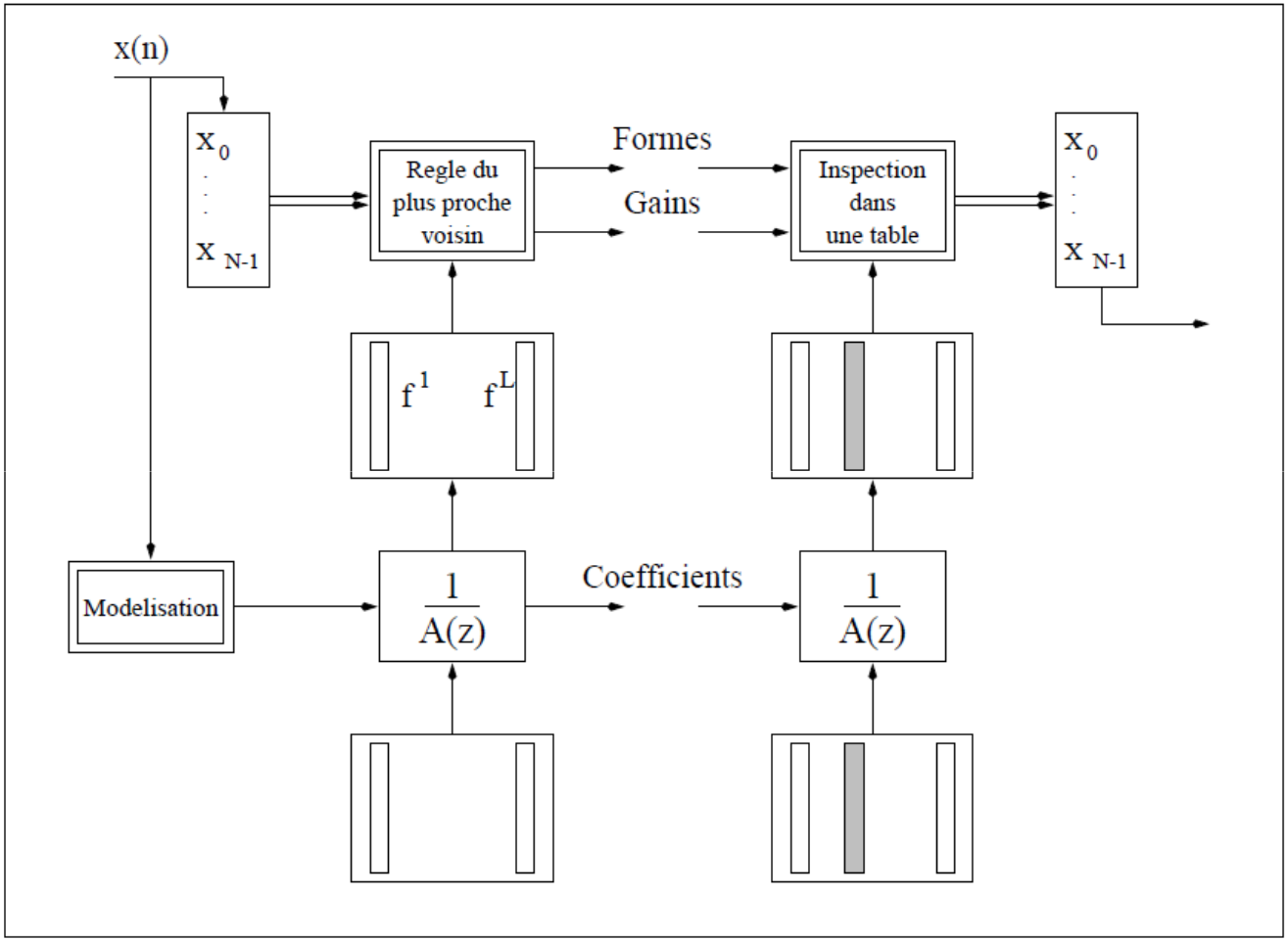
- ⇨ P angles Φ_i ⇨ codage $\Delta\Phi_i = \Phi_i - \Phi_{i-1}$





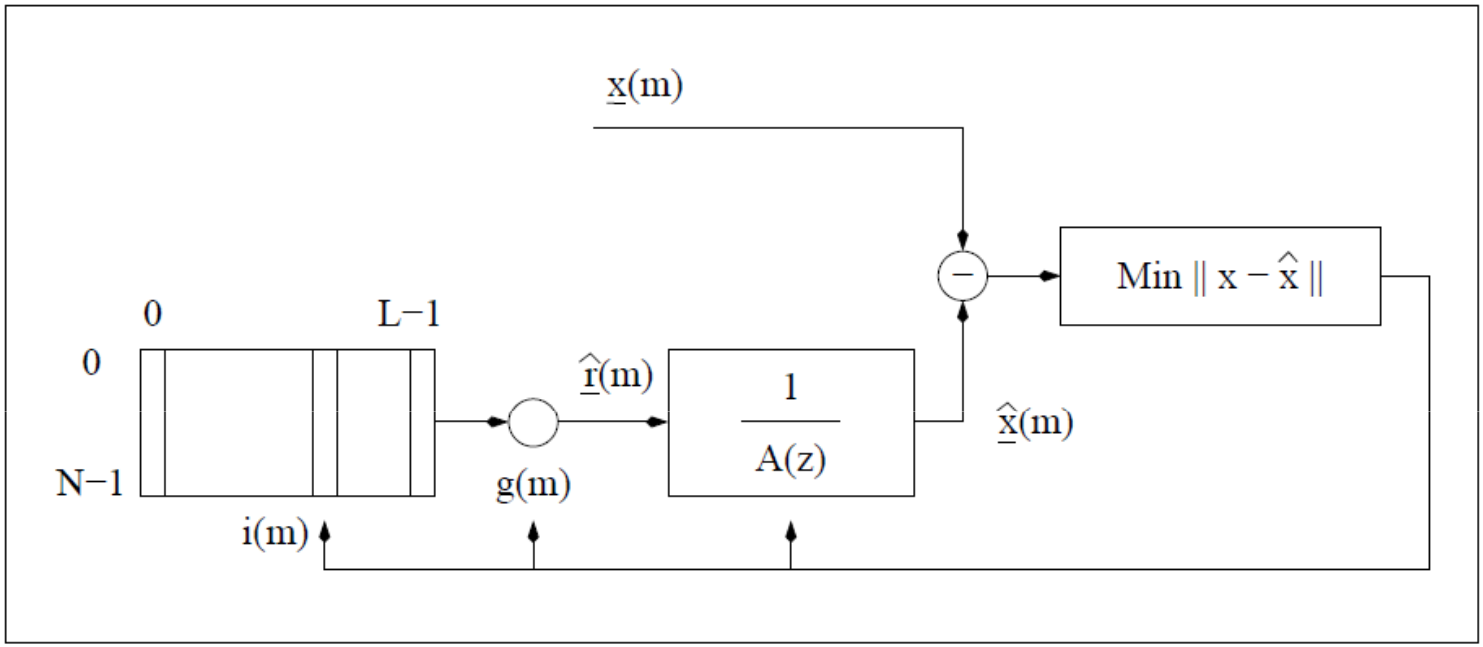


Codeur CELP : Approche “QV”



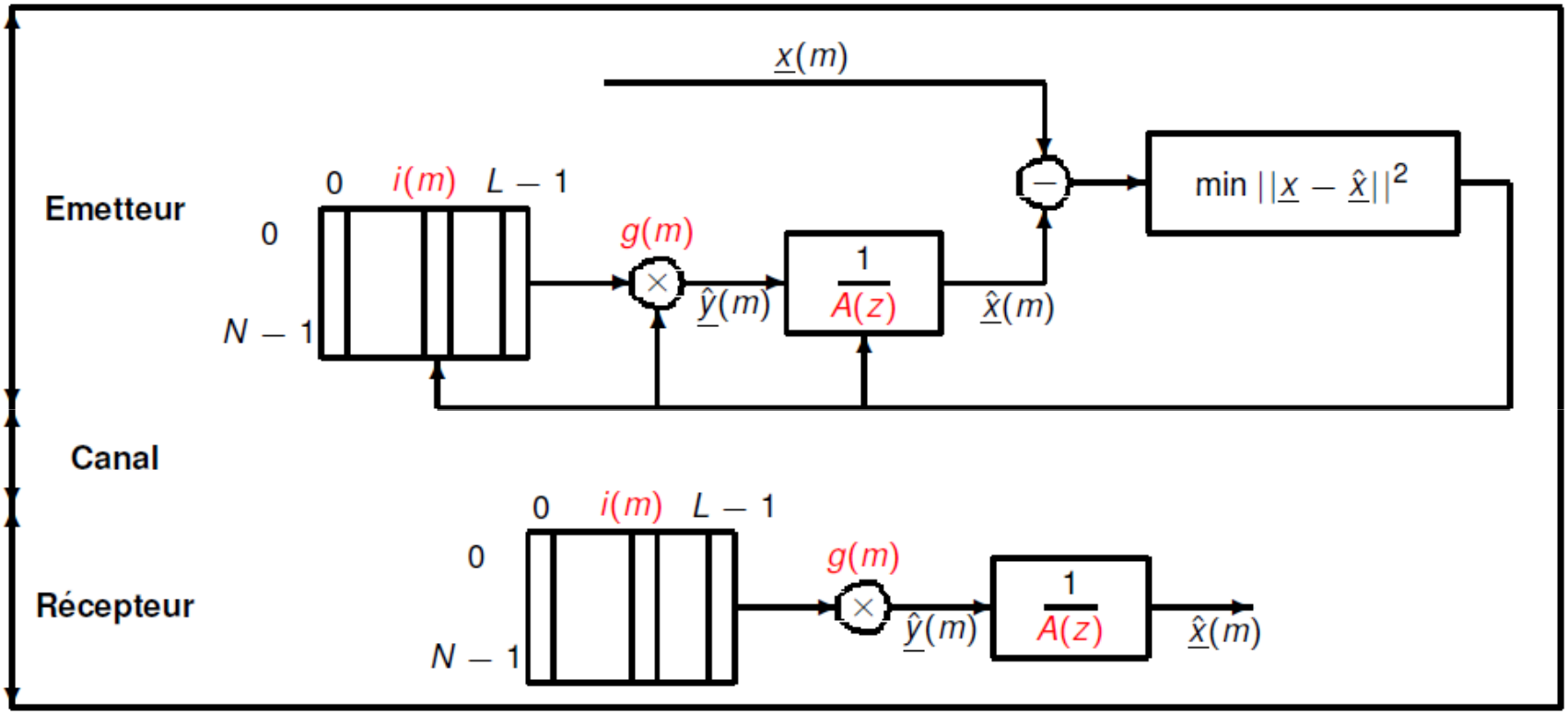
- Dictionnaire de quantification vectorielle adapté par filtrage

Codeur CELP : Approche ``Modélisation paramétrique''



- Modèle de production ``source-filtre''
- Schéma de codage ``analyse-par-synthèse''
- QV ``gain-forme'' et ``multi-étages''

Codeur CELP : Informations transmises



- Chaque fenêtre d'analyse : P coefficients, un n°d'indice, un gain

Choix d'un vecteur et d'un gain

- Modèle pour l'entrée : $\hat{\underline{y}} = g\ \underline{c}^j$ avec $\underline{c}^j \in$ dictionnaire $[\underline{c}^0 \cdots \underline{c}^{L-1}]$

- Critère : $\min ||\underline{x} - \hat{\underline{x}}||^2 = \min ||\underline{x} - g\ \underline{f}^j||^2$
- Choix du vecteur $\underline{f}^j$ le plus colinéaire à $\underline{x}$

$$j^{opt} = \arg\max_j |\cos(\underline{x}, \underline{f}^j)| = \arg\max_j | < \frac{\underline{x}}{||\underline{x}||}, \frac{\underline{f}^j}{||\underline{f}^j||} > |$$

$$j^{opt} = \arg\max_j \frac{< \underline{x}, \underline{f}^j >^2}{||\underline{f}^j||^2}$$

- Détermination du gain : $< \underline{x} - g^{opt} \underline{f}^{j^{opt}}, \underline{f}^{j^{opt}} >= 0$

$$g^{opt} = \frac{< \underline{x}, \underline{f}^{j^{opt}} >}{||\underline{f}^{j^{opt}}||^2}$$

Choix de K vecteurs et de K gains

- Algorithme ``itératif standard'' *(``matching pursuit'')*
- A la *k*ème itération, retirer la contribution des *k-1*premiers vecteurs sélectionnés

$$\underline{e}^k = \underline{x} - \sum_{i=1}^{k-1} g_i \underline{f}^{j(i)}$$

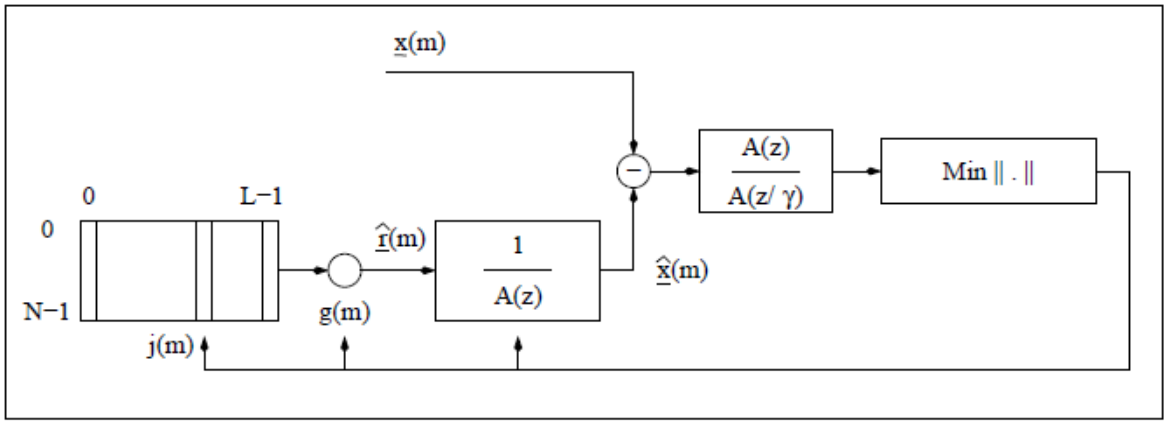
- Choix du *k*ème vecteur et détermination du *k*ème gain

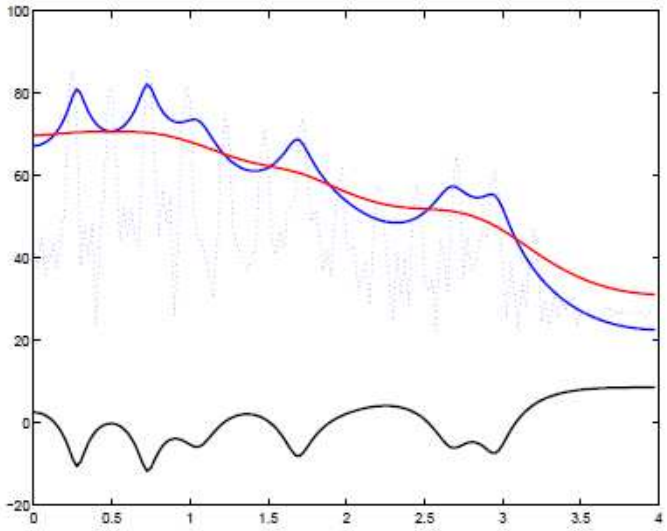
$$j(k) = \arg\max_j \frac{< \underline{e}^k, \underline{f}^j >^2}{||\underline{f}^j||^2} \quad \text{et} \quad g_k = \frac{< \underline{e}^k, \underline{f}^{j(k)} >}{||\underline{f}^{j(k)}||^2}$$

- Nouveau modèle

$$\underline{x}^k = \sum_{i=1}^{k} g_i \underline{f}^{j(i)}$$

Introduction d'un facteur "perceptuel"





Codeur UIT-T G.729 à 8 kbit/s

- Coefficients du filtre de synthèse $1/A(z)$ d'ordre *P=10*
  - Actualisation toutes les 10 ms
  - Codage des ``Line Spectrum Pairs'' sur 18 bits
- Entrée du filtre de synthèse: $\hat{y}(n) = \hat{y}^1(n) + \hat{y}^2(n)$
  - Actualisation toutes les 5 ms
  - Prédicteur à long terme : $\hat{y}^1(n) = g_1 \ \hat{y}(n-Q)$

    Q : caractéristique de la période fondamentale (``pitch'')
  - $\hat{y}^2(n) = g_2 \ c^k(n)$ où $c^k(n) \in [\underline{c}^0 \cdots \underline{c}^{L-1}]$ = dictionnaire de QV
  - Codage (en première approximation)
    - g_1 codé sur 3 bits, *Q* codé sur 7 bits
    - g_2 codé sur 4 bits, *k* codé sur 13+4 bits
- Toutes les 10 ms : 18 + 2 x (3+7+4+17) = 80 ⇨ 8 kbit/s

Codeur UIT-T G.729 à 8 kbit/s (suite)

- **Structure du dictionnaire fixe de QV** $[\underline{c}^0 \cdots \underline{c}^{L-1}]$

Impulsions	Amplitude	Positions	Bits
0	± 1	0, 5, 10, 15, 20, 25, 30, 35	1 + 3
1	± 1	1, 6, 11, 16, 21, 26, 31, 36	1 + 3
2	± 1	2, 7, 12, 17, 22, 27, 32, 37	1 + 3
3	± 1	3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39	1 + 4

$$\log_2 L = 17 \text{ bits}$$

Codeur 3GPP AMR-WB (UIT-T G.722.2) 6.6 …23.85 kbit/s

- **Introduction 50-200 Hz ⇨voix plus naturelle, amélioration de l'effet de présence**

  Extension 3.4-7 kHz ⇨ plus grande intelligibilité

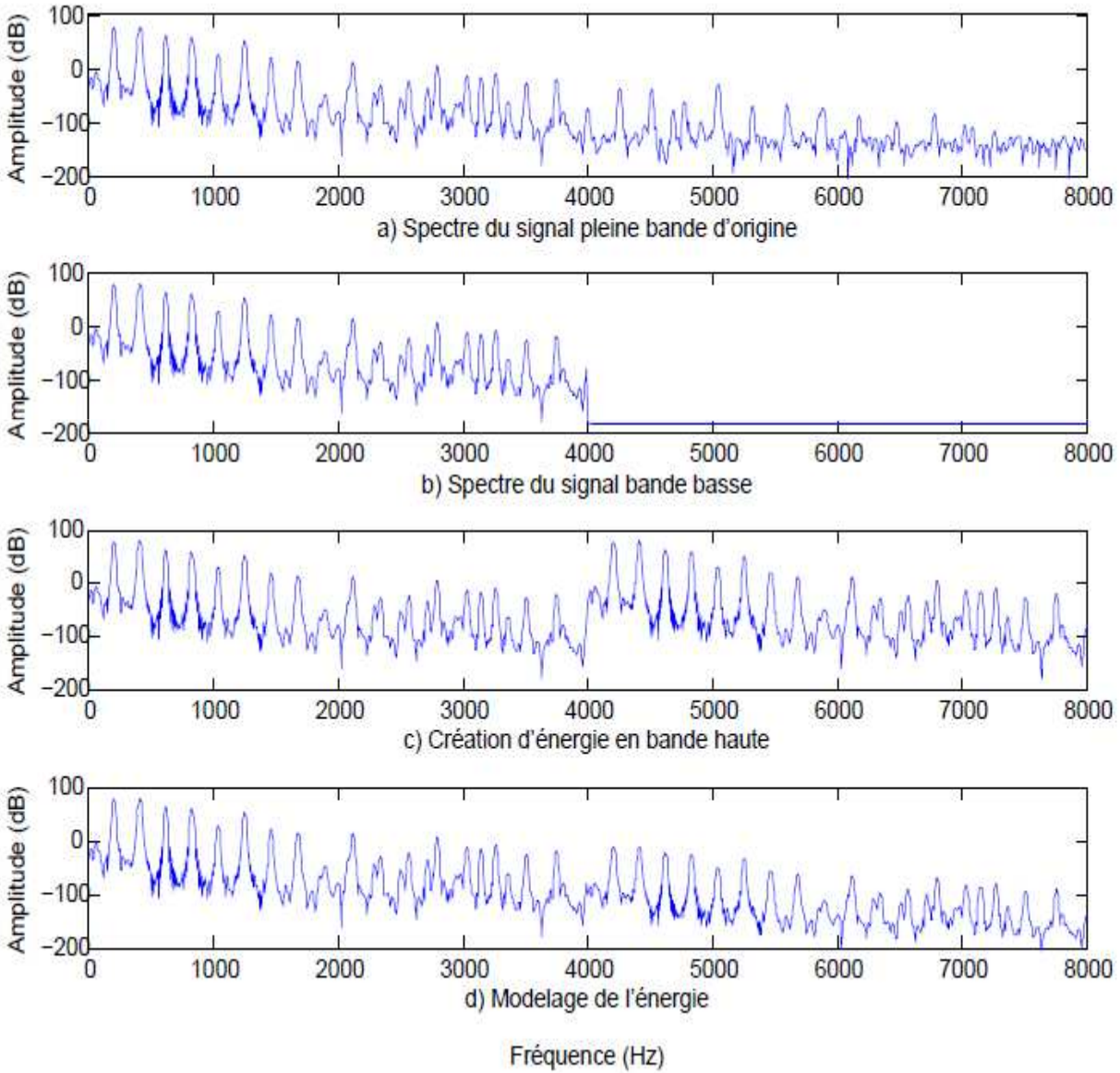
- **Premier codeur adopté pour les réseaux fixes ou mobiles (suppression des transcodages)**
- **Codeur de type ACELP très comparable au G.729 mais**
  - modification du filtrage perceptuel (extension à la bande élargie)
  - modification de l'exploitation de l'information de pitch (pas de structure harmonique sur toute la bande)
  - introduction d'un très grand dictionnaire d'excitation $(\log_2 L = 88 \text{ bits})$

Quelques extensions…..

- **Extension de bande**
- **Quelques éléments de codage Audio**
- **Vers le codeur Parole-Audio universel (*USAC*)**

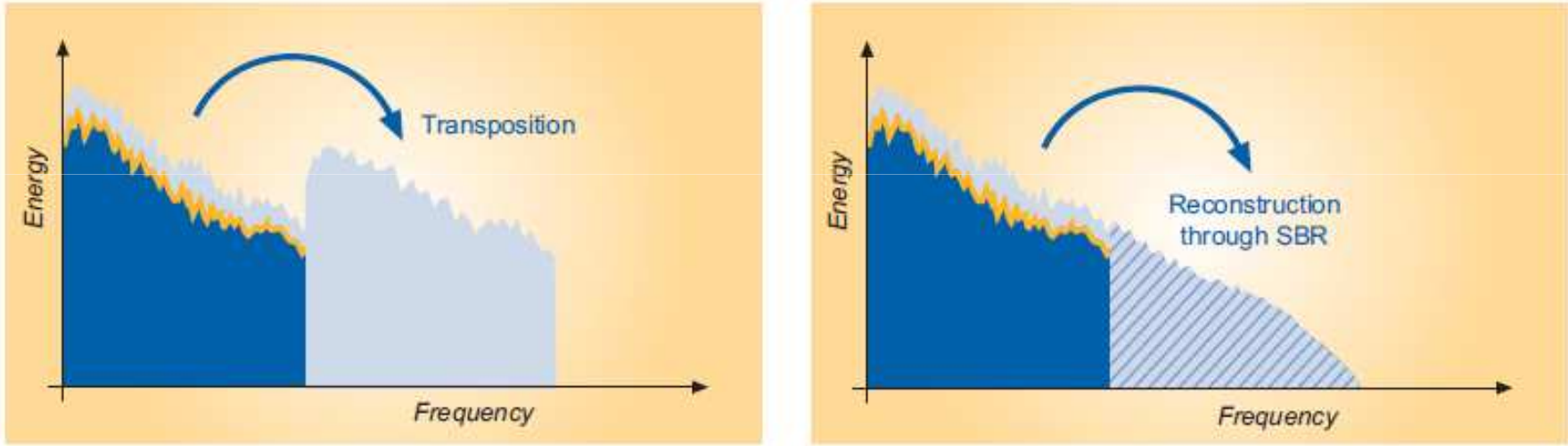
Extension de bande

- Extension en basses fréquences (< 340 Hz) et en hautes fréquences (> 4000 Hz)
- Principe général:
  - Génération d'énergie dans la zone fréquentielle considérée
  - Modulation en amplitude de cette énergie (utilisation de l'enveloppe spectrale)

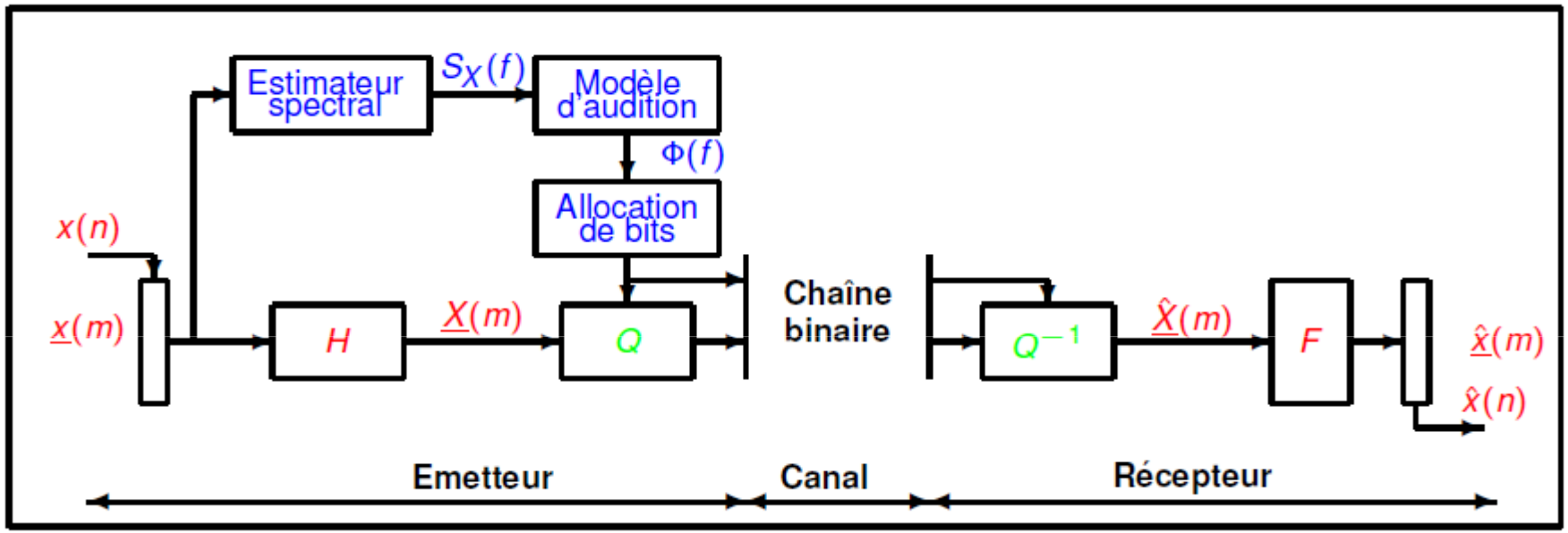


Extension de bande (SBR dans MPEG)

- D'après: S. Meltzer and G. Moser« HE-AAC v2 MPEG-4 - audio coding for today's digital media world- » http://www.ebu.ch/fr/technical/trev/trev_305-moser.pdf



Codeur audio perceptuel



- Transformation temps-fréquences : X(m) = H x(m)
- Allocation de bits sous le contrôle d'un modèle d'audition
- QS/QV des composantes X(m) + codage entropique
- Au récepteur, reconstruction du signal : $\hat{\underline{x}}(m) = F\ \hat{\underline{X}}(m)$

Vers USAC « *Unified Speech-Audio Coding* »

- D'après: Neuendorf & al. « UNIFIED SPEECH AND AUDIO CODING SCHEME FOR HIGH QUALITY AT LOWBITRATES, in Proc. Of ICASSP 2009.