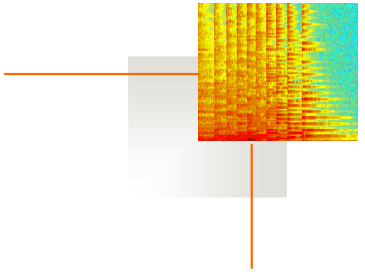
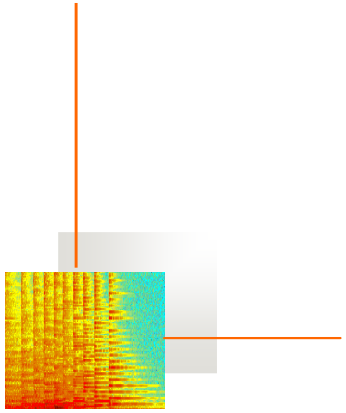




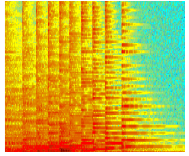
Normes de codage audio

Slim ESSID

INT - Février 2006

e-mail : slim.essid@enst.fr

Page web: http://www.enst.fr/~essid



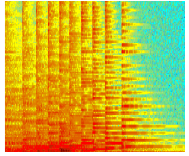
Objectifs, applications

- Codage PCM (Pulse Coded Modulation), qualité CD
  - » Échantillonnage (44.1 kHz)
  - » Quantification (16 bits)
  - » Débit 1.41 Mbit/s ! (stéréo)

- Codage ⇔ Compression
  - » Stockage (CD, MINI DISC, DVD, …)
  - » Services de diffusion
    - TV numérique, radio numérique (DAB)
    - Réseau IP, téléchargement/streaming
  - » Communications interactives (sur réseau RTC, RNIS, IP, GSM)
    - Téléphonie
    - Visiophonie, télé-enseignement, …

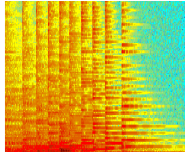
Gammes de qualité

			Fe (kHz)	R (bits)	Débit nominal (kbit/s)	Débit usuel (kbit/s)	Taux de compression
PAROLE	Bande téléphonique		8	13	104	64-...-4-...	1.6-...-26-...
PAROLE	Bande élargie		16	14	224	64-...-16-...	3.5-...-14-...
MUSIQUE	Bande HiFi	Qualité "FM"	32	16	512 monovoie (1024 stéréo)	192-...-64-...	2.6-...-8-...
MUSIQUE	Bande HiFi	Qualité "CD"	44.1	16	705.6 (1411 stéréo)	192-...-56-...	3.6-...-12-...
MUSIQUE	Bande HiFi	Qualité "parfaite"	96	24	13824 en 5.1 canaux	...-1000-...	...-13.8-...



Un "bon" codeur: résultat d'un compromis entre

- Débit
  - » Varie selon la qualité de restitution demandée (384 … 2 kbit/s)
  - » Monodébit, multidébits/hiérarchique (imbriqué, à vol de bits, scalable)
- Complexité
  - » Impact sur coût et puissance consommée
  - » MIPS, RAM, ROM, …
- Retard de reconstruction
  - » Paramètre critique pour applications conversationnelles
  - » < 150 ms, perte d'interactivité au-dessus de 400 ms
- Tenue aux erreurs de reconstruction
  - » Communication avec mobiles: codes correcteurs d'erreurs
  - » Communications sur IP : récupération de trames effacées
- Qualité
  - » Fonction du type de signal transmis (parole, bruit, musique, modems…)
  - » Déterminée par des tests subjectifs



Evaluation de la qualité

- Tests subjectifs (écoutes) formels (protocoles complètement définis)
  - » Codeurs de parole : qualité médiocre ► tests d'intelligibilité ► méthodes de jugement :
    - par catégories absolues ACR (Absolute Category Rating)
    - par catégories de dégradation DCR (Degradation Category Rating), etc.
  - » Codeurs audio débits compris entre 20 et 64kbits/s : qualité "acceptable"
  - ► méthode MUSHRA (MUlti Sitmulus test with Hidden Reference and Anchor)
  - » Codeurs audio de très bonne qualité : la "transparence" >> méthode dite "doublement aveugle à triple stimulus et référence dissimulée" :
- Recommandation UIT-R BS.1116
  - » Enregistrements courts (entre 5 et 10 secondes) répétés 3 fois
  - » Deux possibilités : ABA ou AAB (A= signal original, B=signal codé/reconstruit)
  - » Réponse réclamée
    - B en 2ème ou 3ème position ?
    - Opinion sur B (5: bruit totalement inaudible, 4: très légèrement gênant, 3: un peu gênant, 2: gênant, 1: mauvais, 0: très mauvais)
  - » Traitement statistique ► comparaison objective entre codeurs

Nécessité de la normalisation

- Apparition de produits propriétaires avec nouvelles applications
- Exemples :
  - » Stockage : AC3, Dolby - ATRAC, Mini-Disc Sony
  - » Streaming : Real Audio, Real Networks – Microsoft, windows media
- Incompatibilités, décodeurs propriétaires
- Recours au transcodage
  - » Complexité supplémentaire
  - » Dégradation de qualité
- Normalisation
  - » Interopérabilité
  - » Consensus entre industriels

Normalisation

- Identification d'un nombre suffisant d'applications potentielles
- Cahier des charges par participants au groupe de travail
- Appel à candidatures
- Sélection de la meilleure technologie
- Organismes
  - » Union Internationale des Télécommunications (UIT)
  - » European Telecommunications Standards Institue (ETSI)
  - » Organisation Internationale de normalisation (ISO)
    - Groupe WG 11 de la sous-commission 29 (ISO/IEC/JTC1/SC29/WG11) : MPEG (Moving Picture Expert Group)
    - Développement des normes multimédia (audio et vidéo numériques)

Normes de codage audio de l'ISO : MPEG1

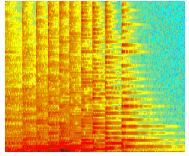
- MPEG1 (ISO/CEI 11172): "Technologie de l'information – Codage de l'image animée et du son associé pour les supports de stockage numérique jusqu'à environ 1.5 Mbit/s"
  - » Partie 1 : Systèmes
  - » Partie 2 : Vidéo
  - » **Partie 3 : Audio**
  - » Partie 4 : Tests de conformité
- Normalisation du décodeur + annexes informatives (1992)
  - » Fréquences d'échantillonnage : 44,1 kHz – 48 kHz – 32 kHz
  - » Modes
    - Stéréophonique, voies droites et gauche dans même train binaire
    - Stéréo combiné, exploitation de la redondance stéréo – corrélation entre voie gauche et droite
    - « Dual monophonique », 2 canaux mono contenant des programmes indépendants (exple. bilingues) dans même train binaire
    - Monophonique, 1 seul canal
  - » 3 couches ou layers offrant des taux de compression différents

MPEG1 audio, couches

- Couche I
  - » Studio, diffusion par satellites,...
  - » Débits : 32, 64, 96, ..., **192**, ..., 384,416,448 kbit/s
  - » Retard min. théorique de codage/décodage : 19 ms
- Couche II
  - » DAB en Europe, ...
  - » Débits : 32, 48, 56, ..., **128**, ..., 256,320,384 kbit/s
  - » Retard min. théorique de codage/décodage : 35 ms
- Couche III
  - » mp3 = MPEG1 Layer 3
  - » Débits : 32,40,48, ...,**96**, ..., 224,256,320 kbit/s
  - » Retard min. théorique de codage/décodage : 59 ms

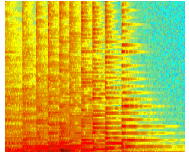
MPEG2 (ISO/IEC 13818 et 13818-7)

- MPEG2 audio (1994)
  - » Extension de MPEG-1
    - à la configuration multi-canaux dont les 5.1 canaux (L,C,R,LS,RS.LFE)
    - au fonctionnement en mono et stéréo à f_e réduites 16, 22.05 et 24 kHz (MPEG-2 LSF, Low Sampling Frequencies)
  - » Diffusion : TVHD – Stockage : DVD, minidisque
- MPEG2 AAC (Advanced Audio Coding) (1997 - 1998)
  - » Codage multi-canaux à des débits raisonnables
  - » Transparence à 384 kbit/s - 5.1)
  - » 1 à 48 canaux, 8 à 96 kHz, 8 à 160 kbit/s
  - » Performances nettement supérieures
  - » Complexité ~ 100 MIPS (codeur), 10 MIPS (décodeur)
  - » 3 profils : "Main profile", "Low complexité profile", "Scalable Sampling Rate Profile"
  - » Etat de l'art en codage musical



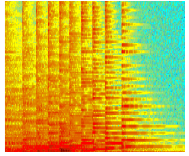
MPEG4

- Version 1, 1998 – version 2, 1999
- Représentation des sons (parole et musique) d'origine naturelle (issu d'un microphone) ou d'origine synthétique (fabriqués par une machine)
- Définition d'objets sonores susceptibles d'être manipulés de façon à former des scènes sonores



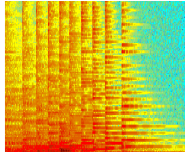
MPEG4 (2/3)

- Sons d'origine naturelle
  - » Définition d'une famille de codeurs hiérarchiques (de 2 à 64 kbit/s)
  - » Existence d'une boite à outils regroupant plusieurs algorithmes de compression
    - de 6 à 24 kbit/s : parole en bande téléphonique ou en bande élargie : **codeur UIT-T G.729**
    - de 16 à 64 kbit/s : musique en bande Hi-Fi
      - Codeurs **MPEG2-AAC, AAC-LD** (Low Delay) pour des communications interactives, **BSAC** (Bit Slice Arithmetic Coding) pour avoir une "granularité" très fine (1 kbit/s)
      - Codeur **TWIN-VQ** (Transform Weighted INterleave-Vector Quantization), meilleures performances à 16 kbit/s
    - de 2 à 24 kbit/s parole/musique (``streaming sur modem") : **codeurs paramétriques**
      - Codeur **HVXC** : Harmonic Vector eXcitation Coding
      - Codeur **HILN** : Harmonic and Individual Line plus Noise coding



MPEG4 (3/3)

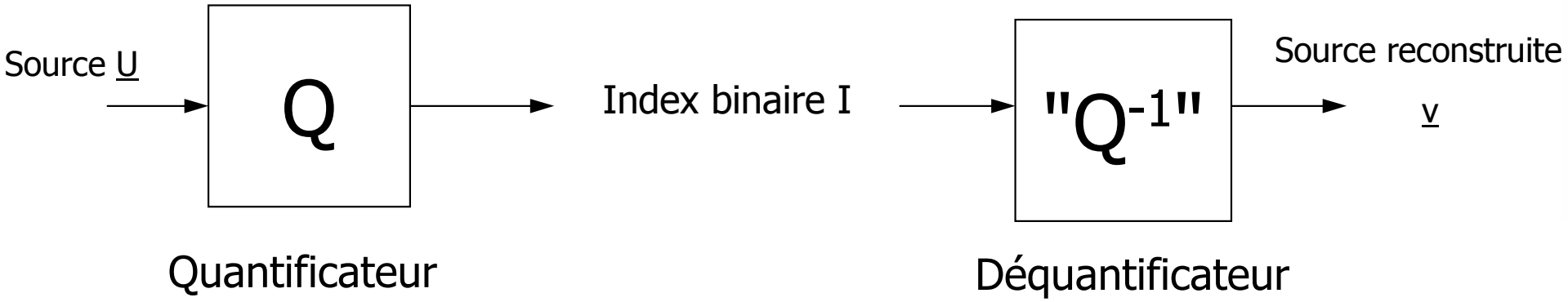
- Sons d'origine synthétique
  - » Algorithme de synthèse de la parole (synthèse "Text-To-Speech")
  - » Langage pour engendrer de la musique (langage SAOL : Structured Audio Orchestra Language)
  - » Exploitation du format MIDI
- Standardization de la façon de décrire une scène
  - » Définition de l'endroit dans un système de coordonnés où se trouve (se déplace) un objet sonore (navigation dans une scène)
  - » Façon dont est modifiée l'apparence de chaque objet
    - Modifications prosodiques pour de la parole
    - Réverbération, spatialisation pour de la musique



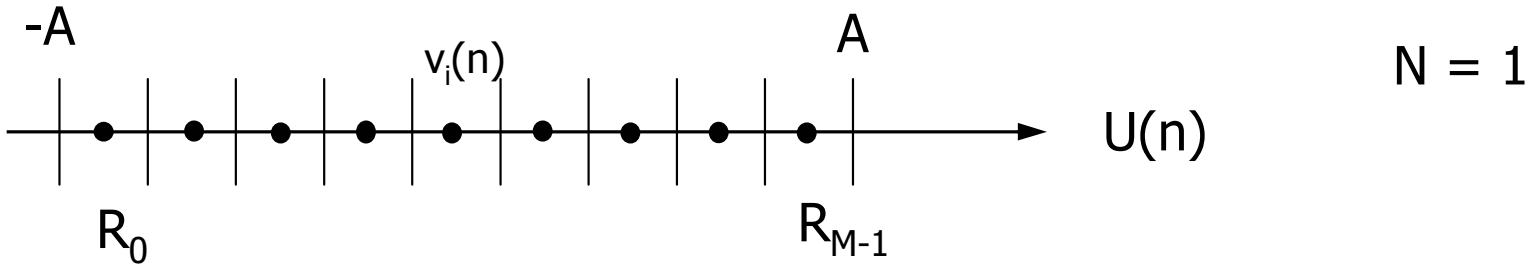
MPEG7 & MPEG21

- MPEG7
  - » Normalisation des descripteurs de contenus multimédia
  - » Faciliter l'accès aux bases de données multimédia
  - » Requête par fredonnement, …
- MPEG21
  - » Sécurité
  - » Tatouage

Quantification, principe



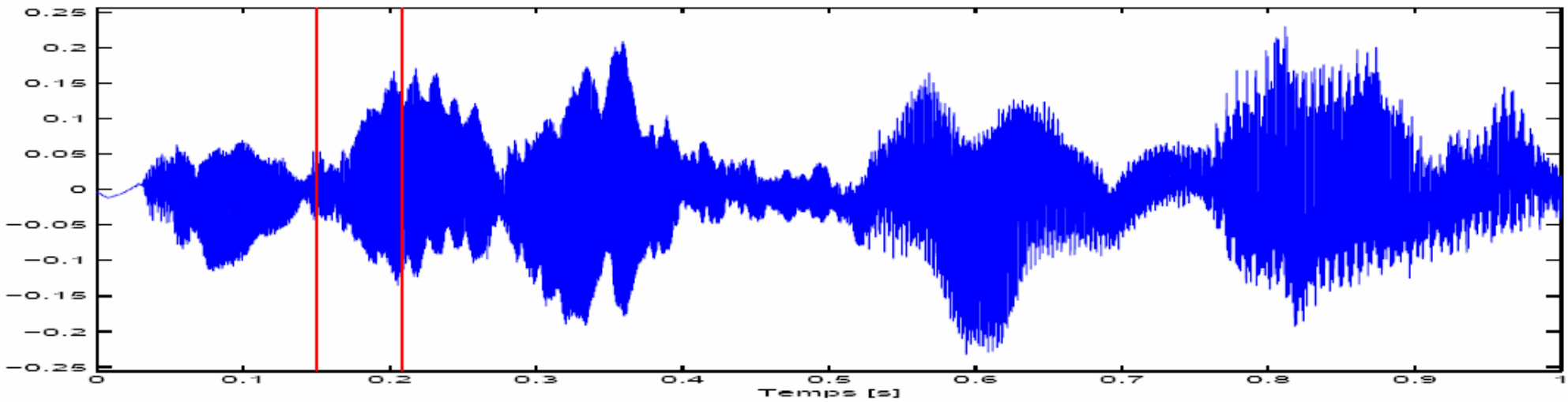
- $\underline{U} = (U_1, U_2, \dots, U_N) \in A^N$, $I \in \{0,1,\dots,M-1\}$, $\underline{v} = (v_1, v_2, \dots, v_N) \in A^N$,
  A ensemble des valeurs prises par U_i ; card(A) = K (dans le cas discret)
- Quantificateur : $Q(\underline{U}) = i \Leftrightarrow \underline{U} \in R_i$; $i \in \{0,1,\dots,M-1\}$
  - » $R_0, R_1, \dots, R_{M-1}$: cellules de quantification, forment une partition
- Déquantificateur : $Q^{-1}(i) = \underline{v}_i$
  - » $\{\underline{v}_0, \underline{v}_1, \dots, \underline{v}_{M-1}\}$: codebook ou dictionnaire

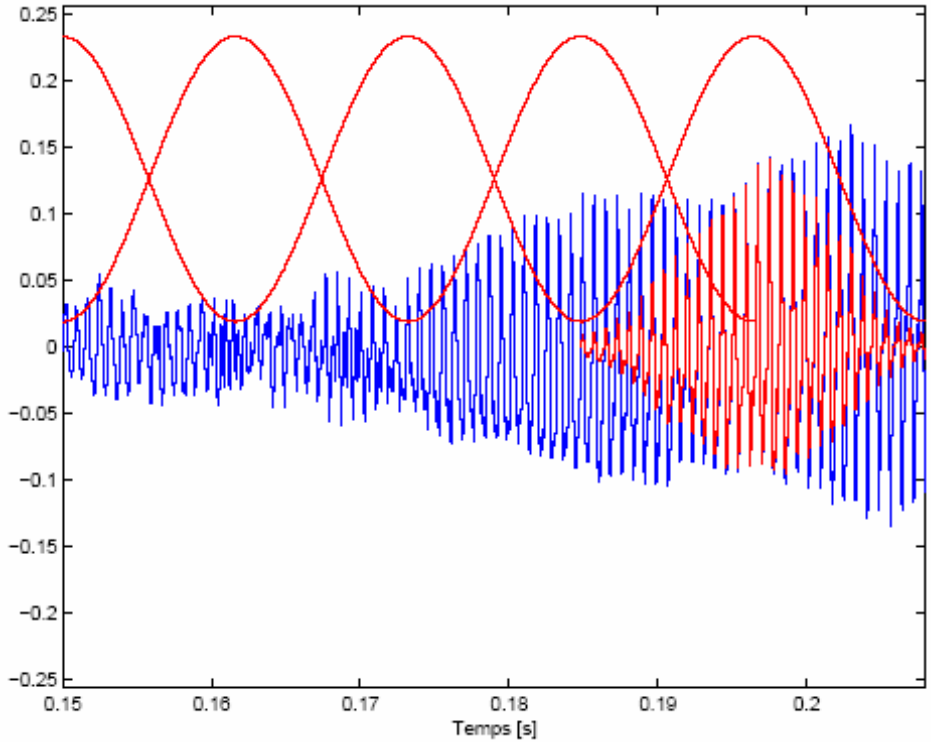


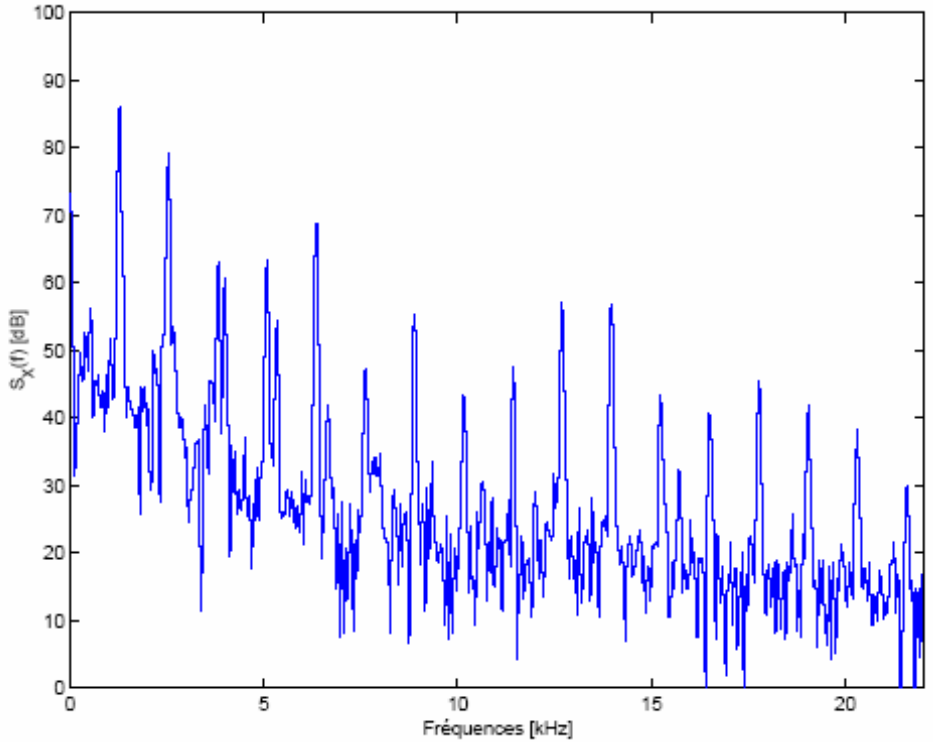
Quantification, caractérisation

- N = 1 : quantification scalaire , N ≠ 1 : quantification vectorielle
- Taux de codage : $R = log_2(M)/N$ (bit/ech)
- Débit : $B = R\,fe$ (bit/s)
- Facteur de compression : $\tau = log_2(K)/R$
- Erreur de quantification : $q(n) = u(n) - v(n)$
- Mesure de distorsion :
  - » $D = \frac{1}{N} E\left\{\left\|\underline{U} - \underline{v}\right\|^2\right\} = \frac{1}{N}\int_{IR^N} p(\underline{u})\left\|\underline{u} - \underline{v}\right\|^2 d\underline{u} = \frac{1}{N}\sum_{i=0}^{M-1}\int_{R_i} p(\underline{u})\left\|\underline{u} - \underline{v}_i\right\|^2 d\underline{u}$
  - » D : puissance moyenne de l'erreur de quantification $q(n)$, $D = \sigma^2_Q$

Un signal de musique : violon

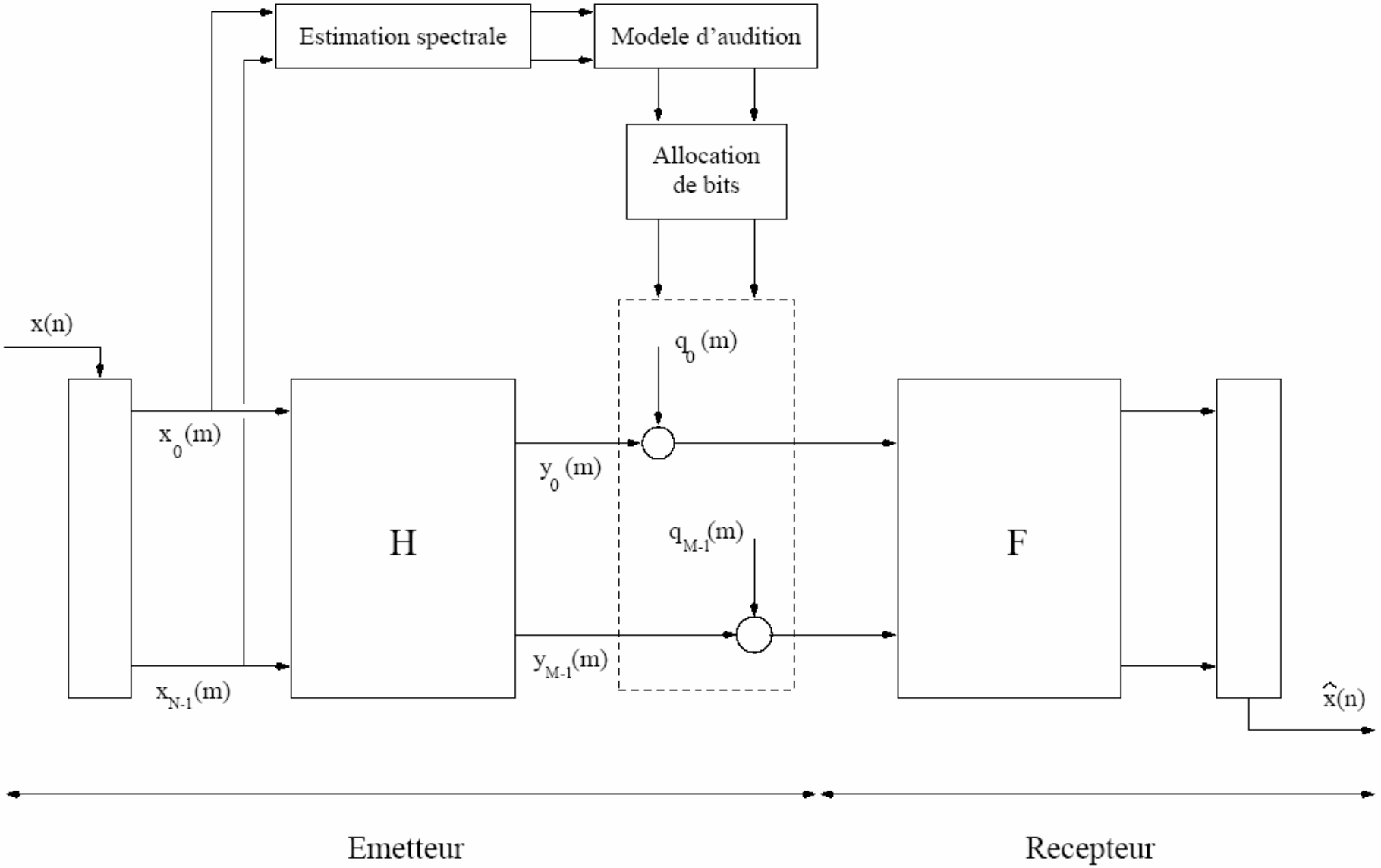




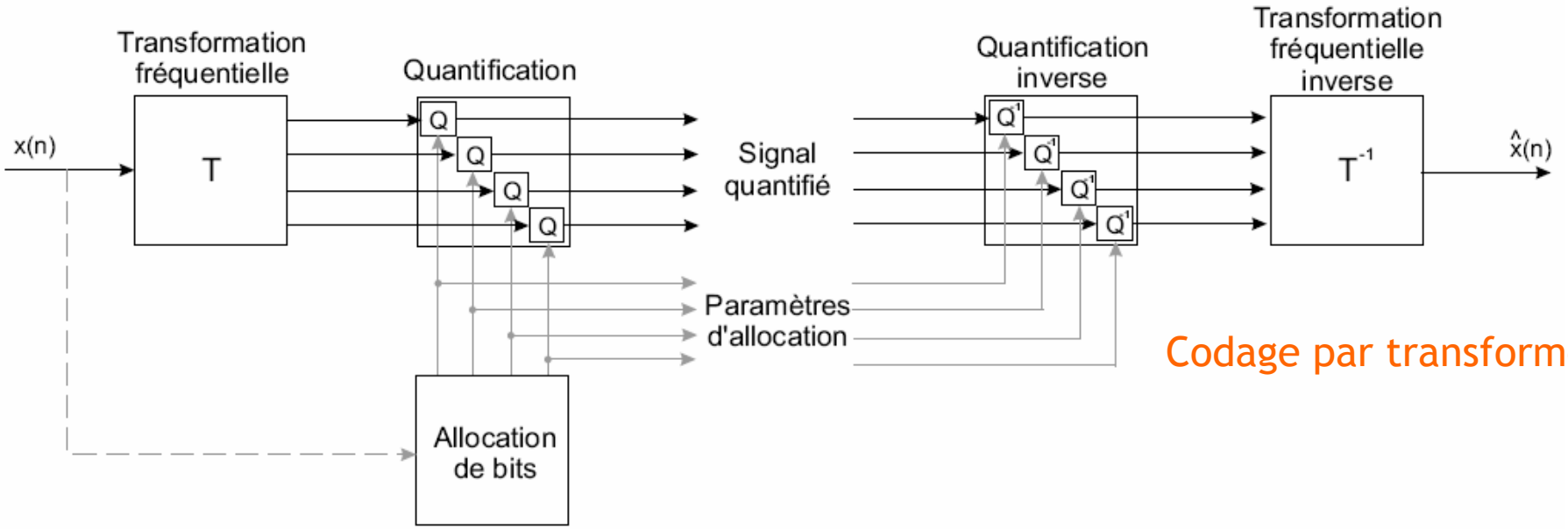


Fenêtres d'analyse recouvrantes de 20ms (fe=44.1kHz, 20ms : N=882 échantillons)

Codage perceptuel : principe

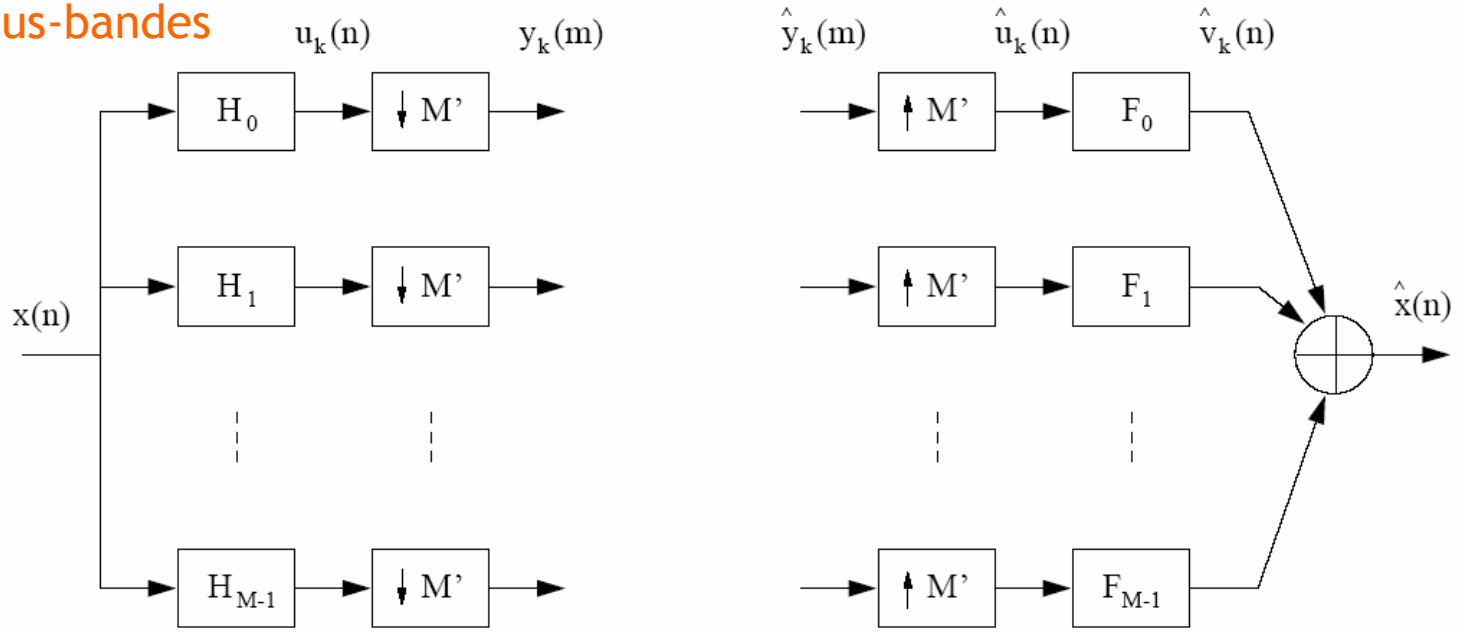


Transformation temps-fréquence



Codage par transformée

Codage en sous-bandes



Banc de filtres d'analyse

Banc de filtres de synthèse

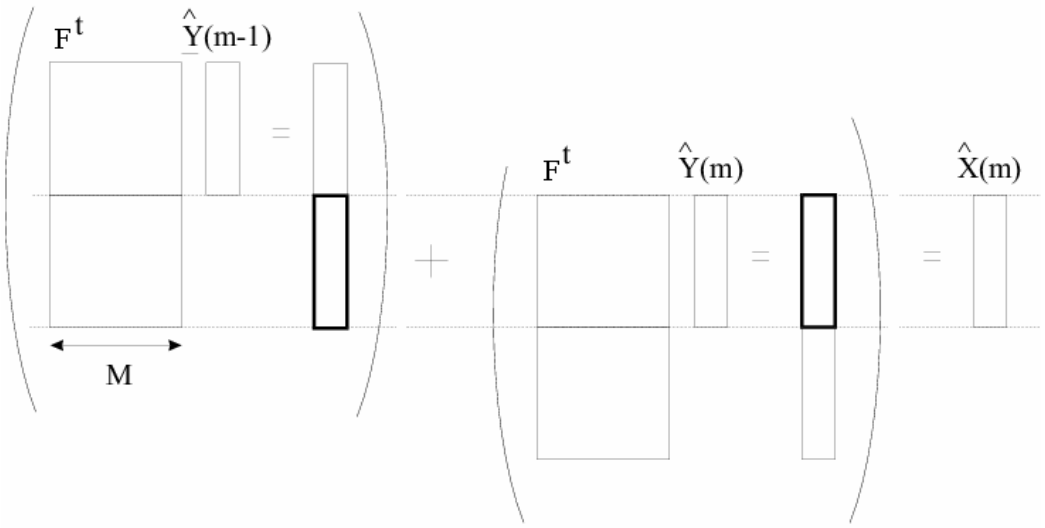
Equivalence banc de filtres et transformée

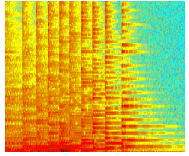
- Analyse $\underline{y}(m) = \mathrm{H}\,\underline{x}(m)$

$$\mathbf{H} = \begin{bmatrix} h_0(N-1) & \cdots & h_0(0) \\ \vdots & & \vdots \\ h_{M-1}(N-1) & \cdots & h_{M-1}(0) \end{bmatrix} \qquad \underline{x}(m) = \begin{bmatrix} x(mM-N-1) \\ \vdots \\ x(mM) \end{bmatrix} \qquad \underline{y}(m) = \begin{bmatrix} y_0(m) \\ y_1(m) \\ \vdots \\ y_{M-1}(m) \end{bmatrix}$$

- Synthèse par addition et recouvrement (Overlap add)

$$\mathbf{F} = \begin{bmatrix} f_0(0) & \cdots & f_0(N-1) \\ \vdots & & \vdots \\ f_{M-1}(0) & \cdots & f_{M-1}(N-1) \end{bmatrix}$$



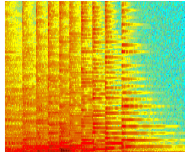


Transformée

- Transformée à reconstruction parfaite
  - » Importance du recouvrement : N > M
  - » Banc de filtres modulés : filtre prototype et modulation

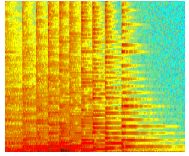
- Dimensionnement (N, M ??)
  - » Résolution fréquentielle
    - ► M grand : grand nbre de coef spectraux
    - ► N grand : filtres sélectifs
  - » Résolution temporelle
  - » Adaptation à diverses fréquences d'échantillonnage

► Réaliser un compromis



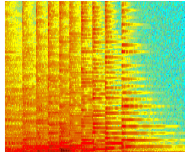
Transformée optimale et allocation de bits

- Transformée optimale "théorique" : Transformée Karhunen-Loeve (KLT)
- En pratique MDCT : Modified Discrete Cosine Transform
- Allocation optimale
  - ► sous le contrôle d'un modèle d'audition



Psychoacoustique

- Caractérisation de la perception auditive humaine
- Analyse temps-fréquence des capacités de l'oreille interne
- Relation entre grandeurs physiques et grandeurs perceptuelles
- Expériences/tests psychoacoustiques
- Jugement de l'individu testé

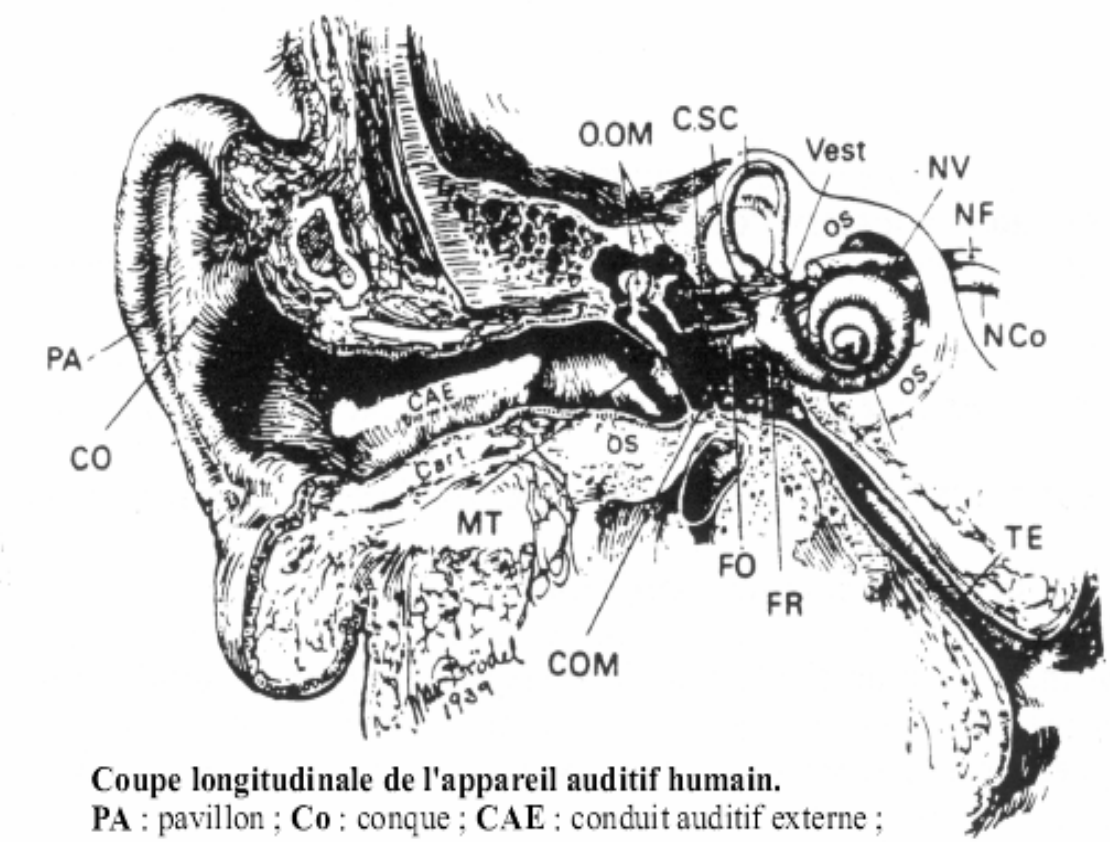


Niveau de pression acoustique

- Sensation sonore : onde acoustique
- Vibrations du tympan
- $P(t) = H_0 + p(t)$
  - » P(t), pression instantanée;
  - » H_0, pression atmosphérique (pression moyenne de l'air, 10^5 Pa (N/m^2));
  - » p(t) : vibrations.
- p/p_r , $p_r = 2\ 10^{-5}$ Pa ; p_r : pression de référence ~ pression min. audible (auditeur moyen, 1kHz)
- Niveau de pression acoustique
  - » $SPL = 20 \log_{10}(p/p_r)$ (Sound Pressure Level dB SPL)
  - » 0 dB SPL ~ silence, > 140 dB SPL ~ douleur

Structure de l'oreille

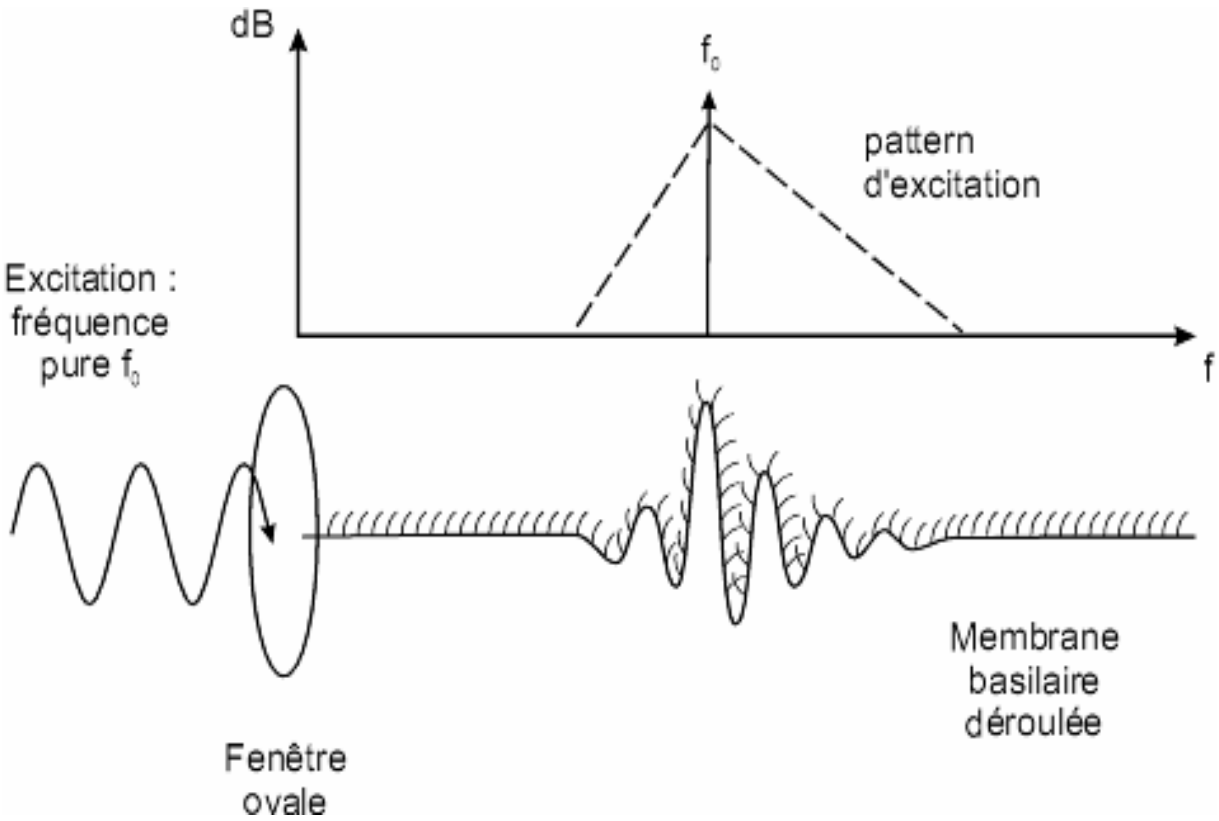
- Oreille externe : pavillon, conduit auditif, tympan
- Oreille moyenne : Osselets (marteau-enclume-étrier), fenêtre ovale
- Oreille interne : cavité en colimaçon, membrane basilaire



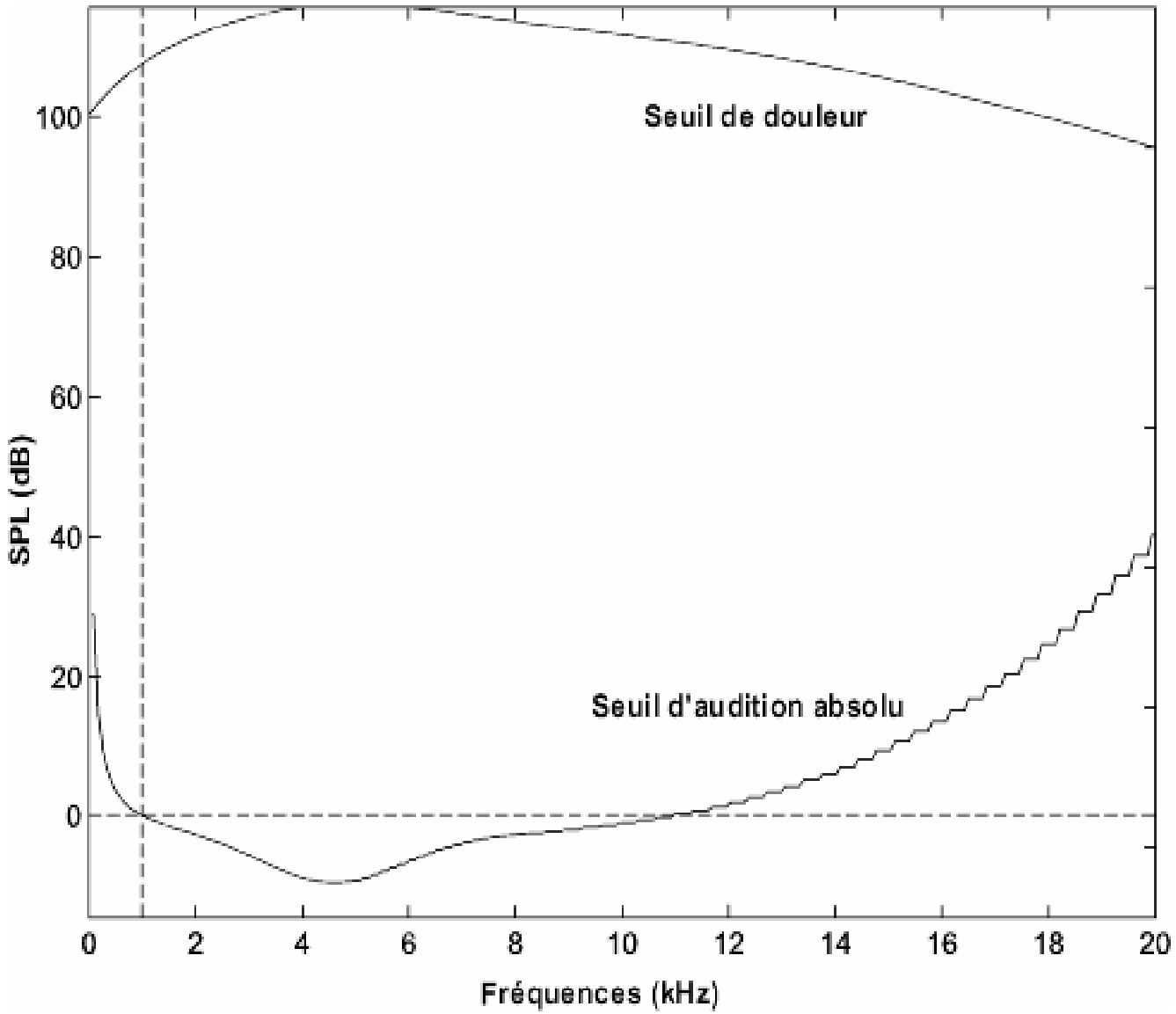
Coupe longitudinale de l'appareil auditif humain.
PA : pavillon ; **Co** : conque ; **CAE** : conduit auditif externe ; **Cart** : cartilage ; **MT** : membrane tympanique ; **O.OM** : osselets de l'oreille moyenne ; **CSC** : canaux semi-circulaires ; **FO** : fenêtre ovale ; **FR** : fenêtre ronde ; **COM** : cavité de l'oreille moyenne ; **Vest** : vestibule ; **NV** : nerf vestibulaire ; **NF** : nerf facial ; **NCo** : nerf cochléaire ; **TE** : trompe d'Eustache

Membrane basilaire, pattern d'excitation

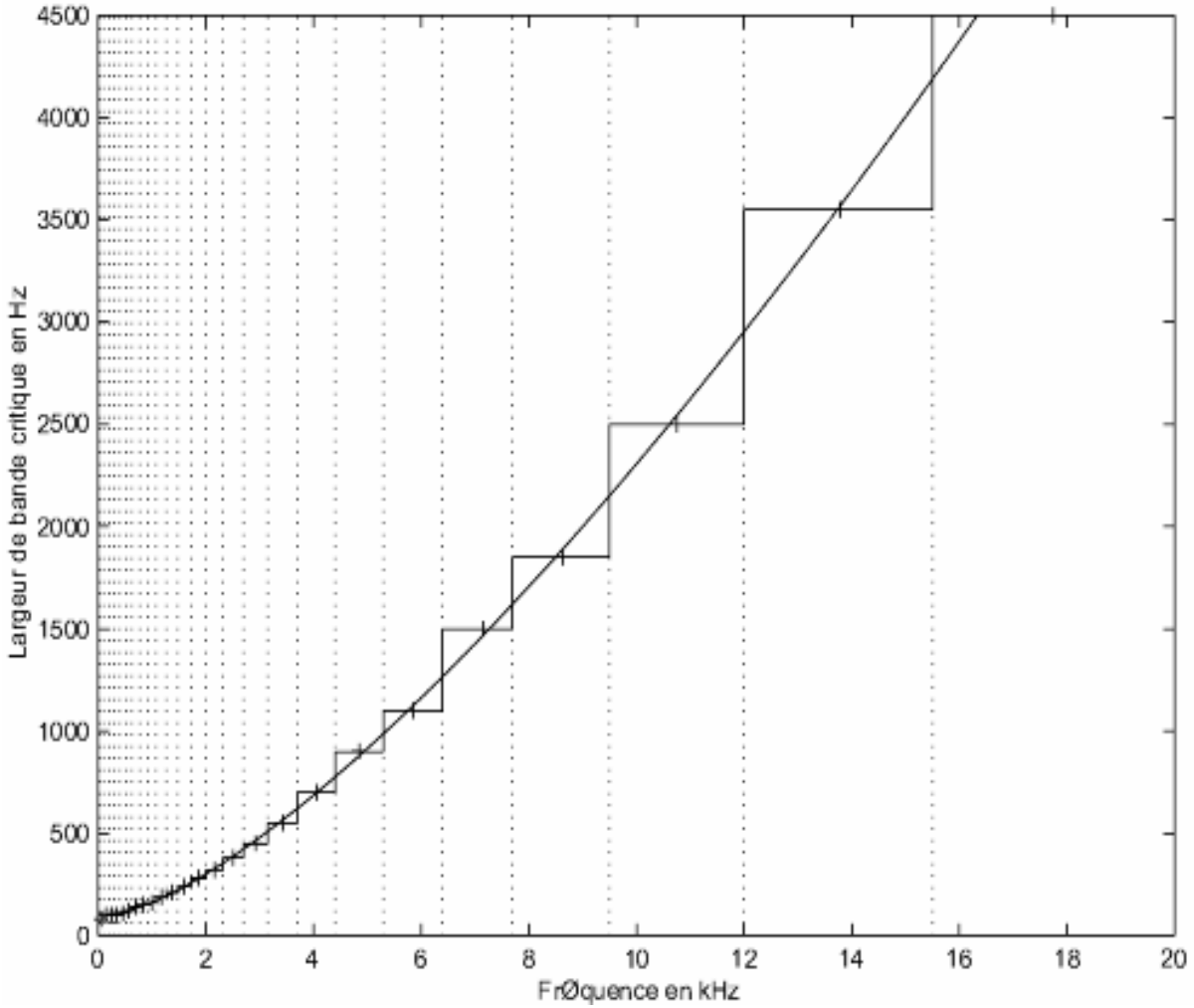
- Vibration le long de la membrane, fonction de la fréquence
- Cellules cilliées, spécialisées en fréquence
- Pattern d'excitation → masquage fréquentiel



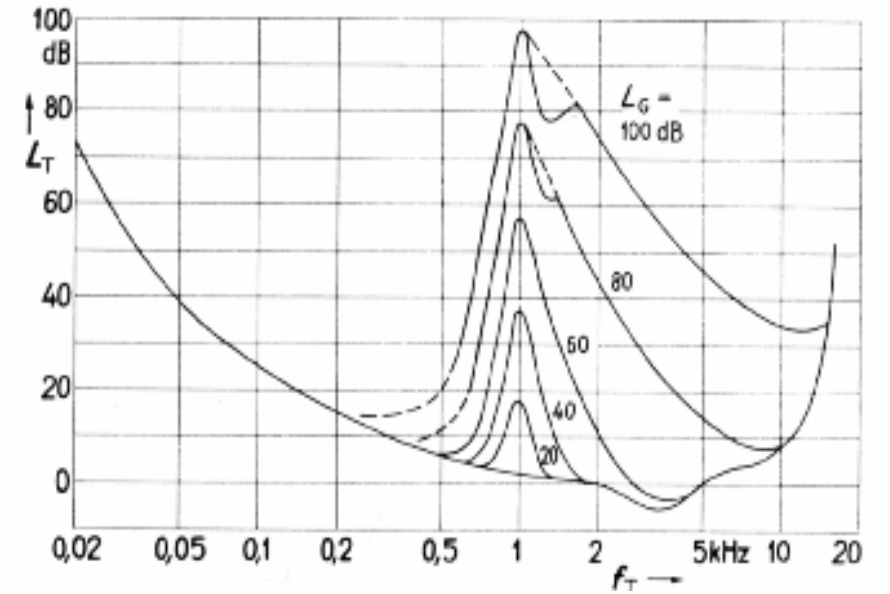
Seuil d'audition absolu

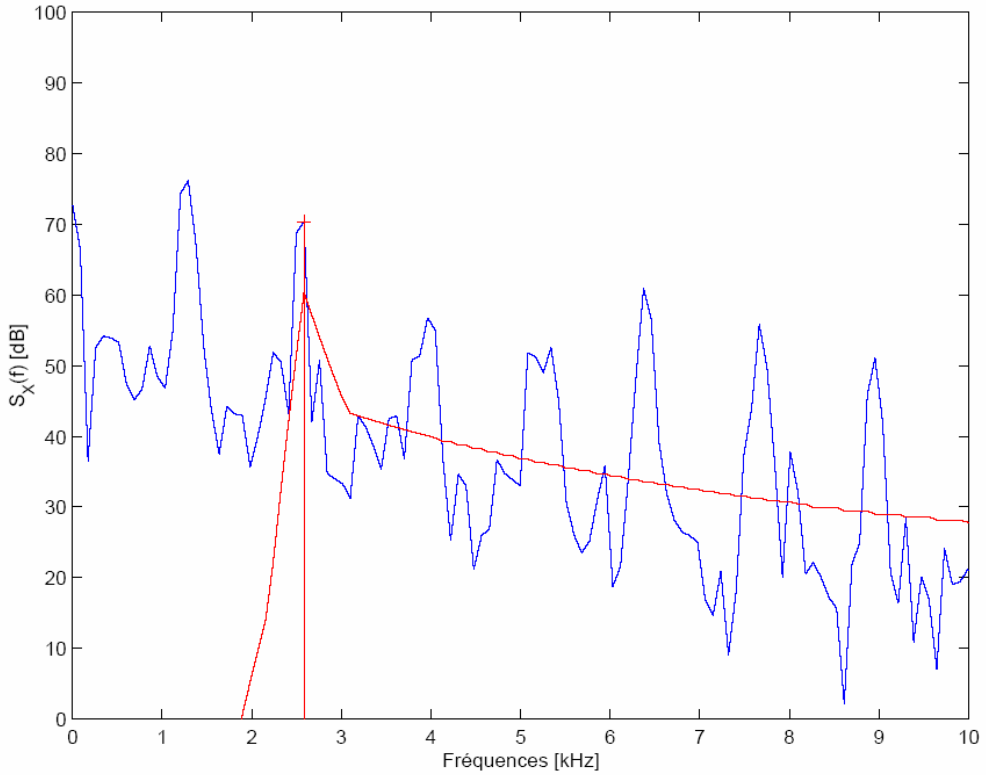


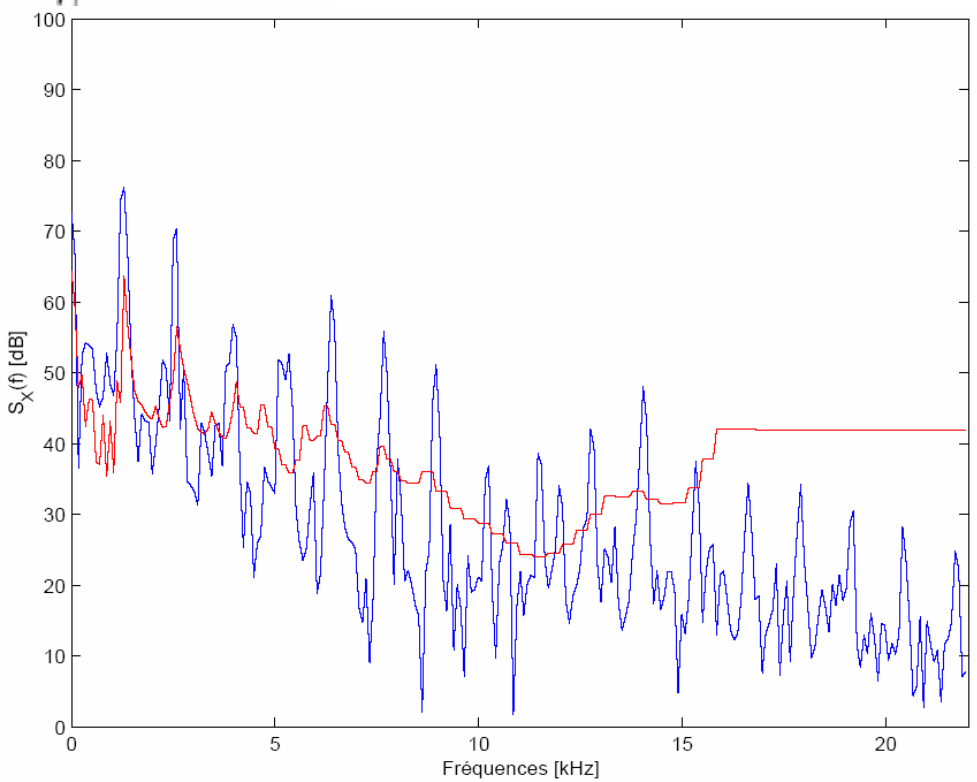
Bandes critiques



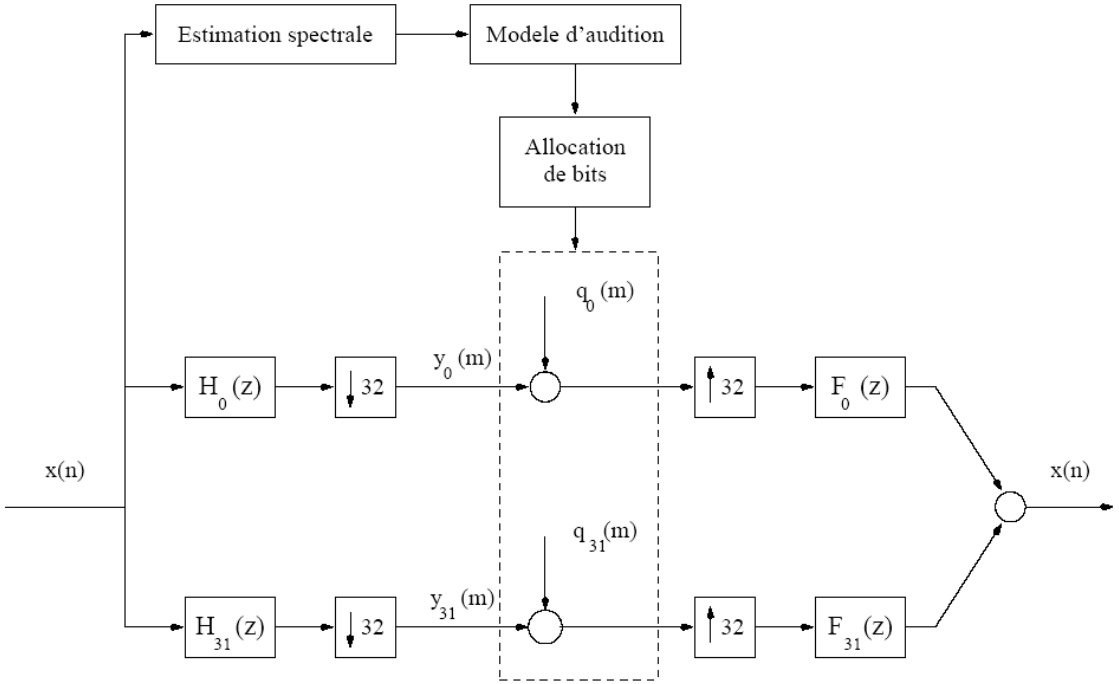
Courbe de masquage, seuil de masquage

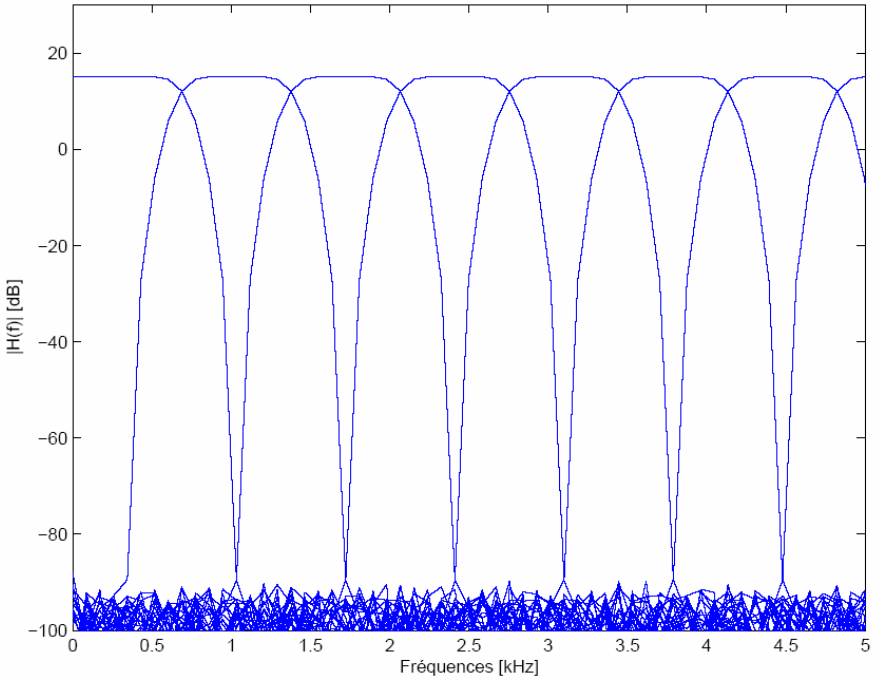






Codeur MPEG1 Layer 1: Transformation temps-fréq.





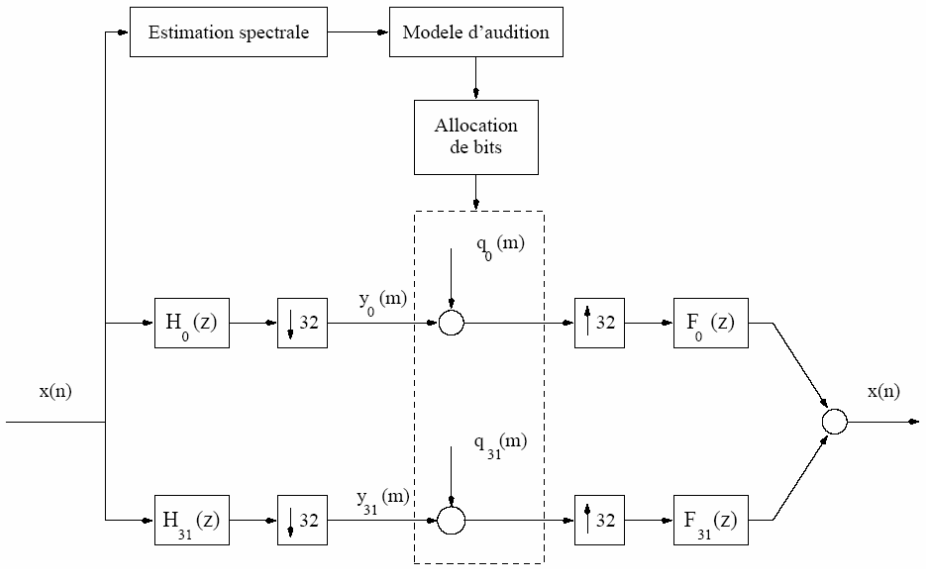
- Banc de M=32 filtres modulés (exploitation d'un filtre prototype) :

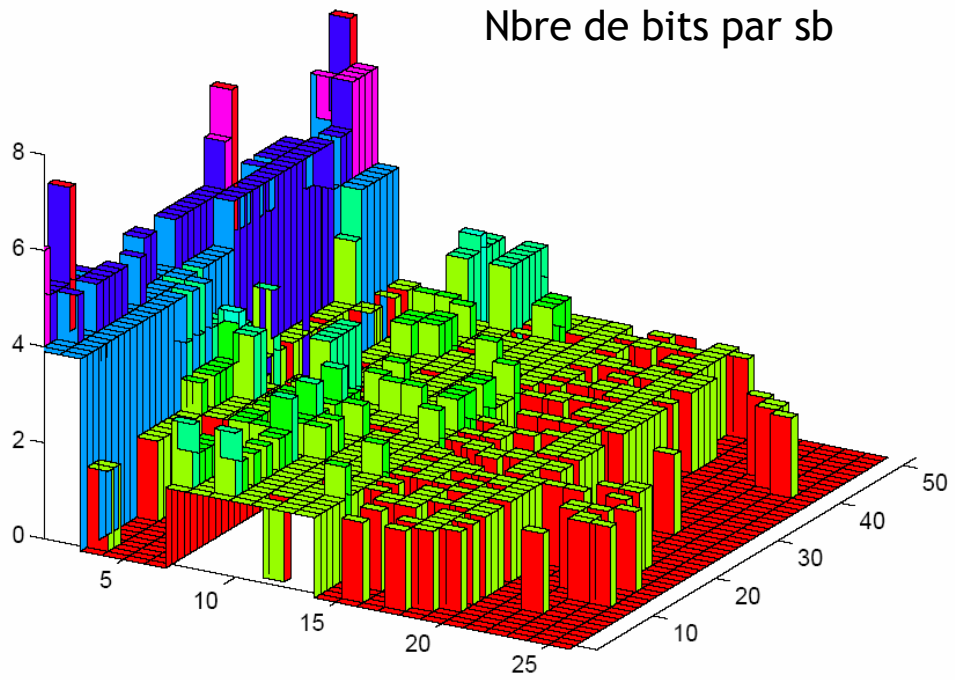
$$H_k(f) = H(f - \frac{2k+1}{4M}) + H(f + \frac{2k+1}{4M})$$

$$h_k(n) = 2\ h(n)\ cos(2\pi \frac{2k+1}{4M} n + \varphi_k) \quad n = 0 \cdots N-1$$

- Longueur de la réponse impulsionnelle du filtre prototype : N = 512
- Sous-échantillonnage "critique" : Facteur de sous-échantillonnage = M
- Pas de reconstruction parfaite mais RSB > 90dB
- $M << N$: bonne résolution temporelle (α M) mais mauvaise résolution fréquentielle (α $1/M$)
- Pas de problème de pré-écho avec ce codeur

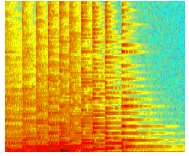
Codeur MPEG1 Layer 1





- Banc de M = 32 filtres Pseudo-QMF
- Pour chaque sous-bande $k \in \{0 \cdots M-1\}$
  - » Construction d'un vecteur : $\underline{y}_k = [y_k(0) \cdots y_k(11)]$ (trames de 384 échantillons)
  - » Détermination d'un "facteur d'échelle" : $g_k = \max[y_k(0) \cdots y_k(11)]$ (6 bits)
  - » Normalisation des composantes $[y_k(0) \cdots y_k(11)]/g_k$
  - » QS uniforme des composantes normalisées sur b_k bits (choix entre 15 quantificateurs)

- Détermination de $b_0 \cdots b_{M-1}$: procédure "d'allocation de bits" sous le contrôle d'un modèle d'audition
- Complexité : essentiellement due au modèle d'audition (uniquement au codeur)

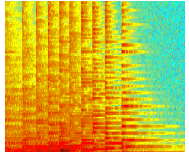


Modèles psychoacoustiques

- 2 propositions de modèle psychoacoustique
  - » Modèle I - faible complexité, bonne précision pour débits élevés
  - » Modèle II – complexité supérieure, plus bas débits

- Fonctionnement
  - » Calcul périodogrammes du signal, domaine de Fourier
  - » Mapping vers bandes critiques
  - » Distinction composantes tonales/bruit
  - » Application des fonctions d'étalement aux composantes
  - » Calcul de la fonction de masquage
  - » Retour au domaine de Fourier en sous-bandes

- Mapping, paramètres de représentation donnés pour différentes fe

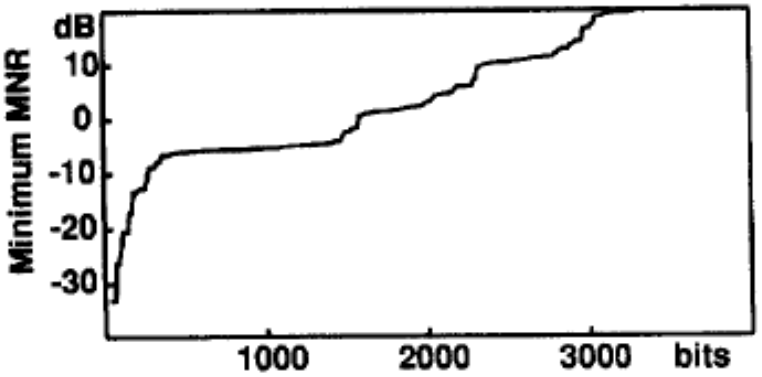


Allocation des bits

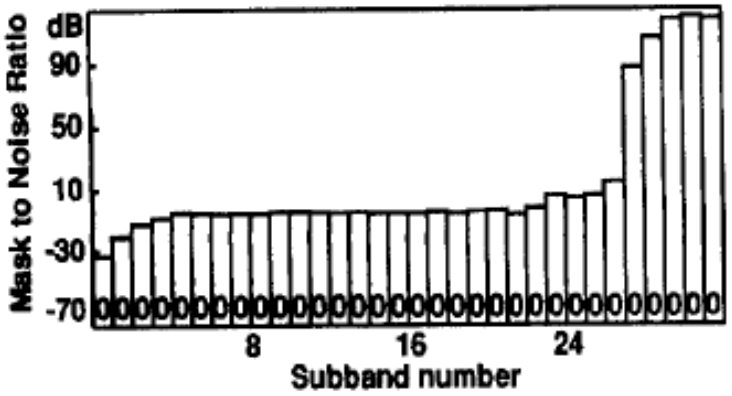
- Modèle psychoacoustique >> SMR par sous-bande
- Algo. Allocation >> MNR (Mask to Noise Ratio)
  - » MNR = SMR – SNR
  - » SNR fonction de l'allocation des bits, donnés par tables du standard
- Nombre de bits alloués à une trame
  - » bits/trame = (bits/seconde)/(trames/seconde)
  - » trames/seconde = (échantillons/seconde)/(échantillons/trame)
- Prendre en compte les bits d'entête et de donnés aux...
- Allouer les bits restant de façon à maximiser le MNR min de sous-bande

Allocation des bits

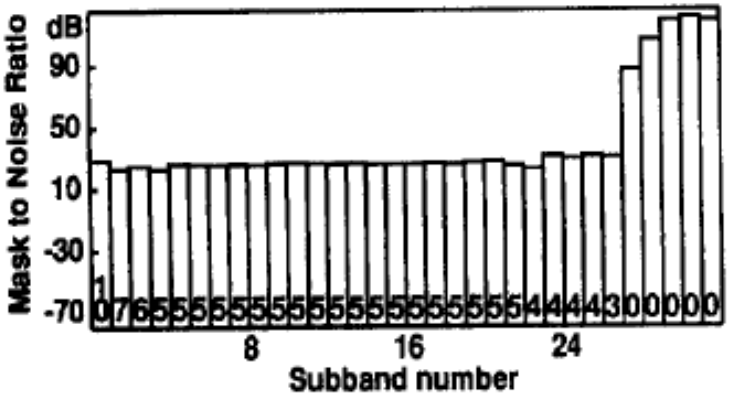
- Init. : 0 bits par sous-bande
- Calculer MNR pour chaque sous-bande
- Trouver sous-bande avec MNR min et nbits < limite max
- Incrémenter de 1 bit ...

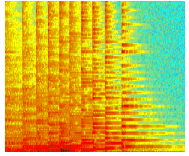


Début



Fin



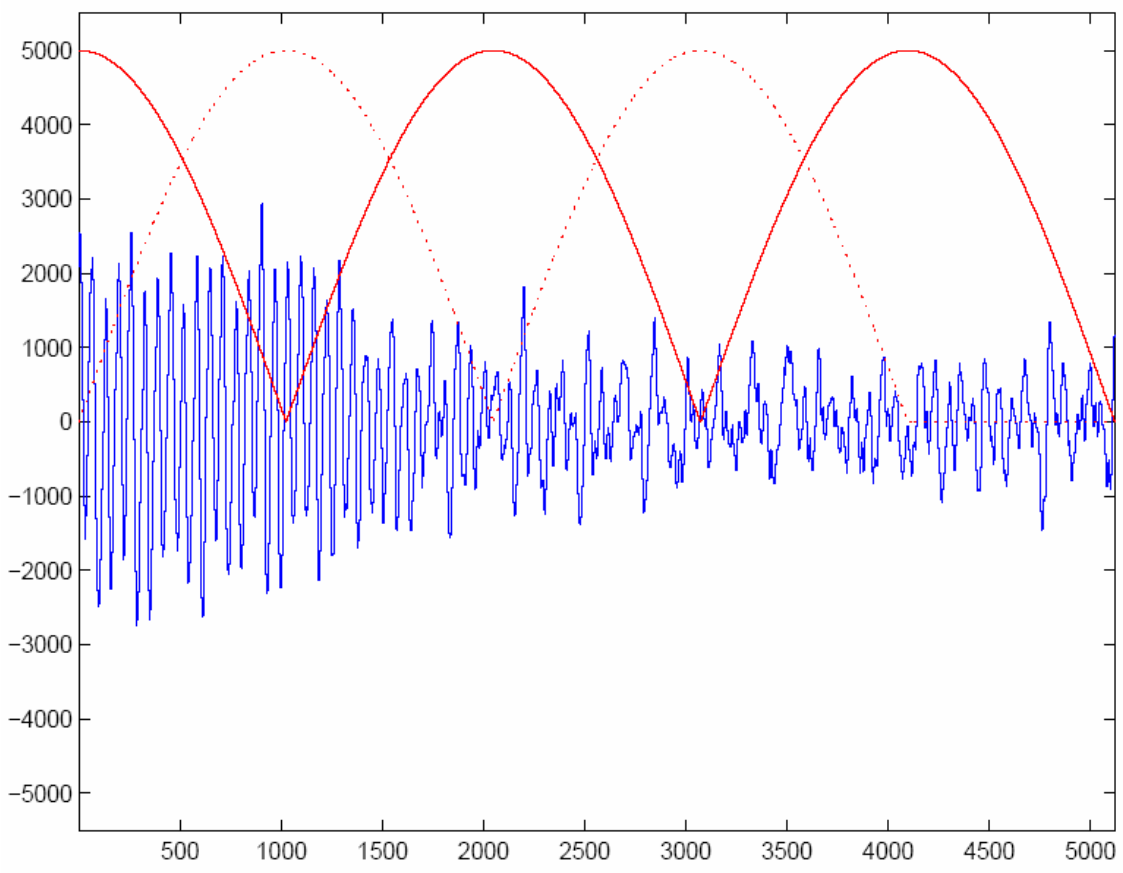


Couche III

- Banc de filtre hybride, MDCT
- Fenêtres courtes / fenêtres longues
- Réduction des effets d'aliasing
- Quantification non-uniforme
- Bandes de facteurs d'échelle
- Codage entropique des données
- Utilisation de « réservoir de bits »

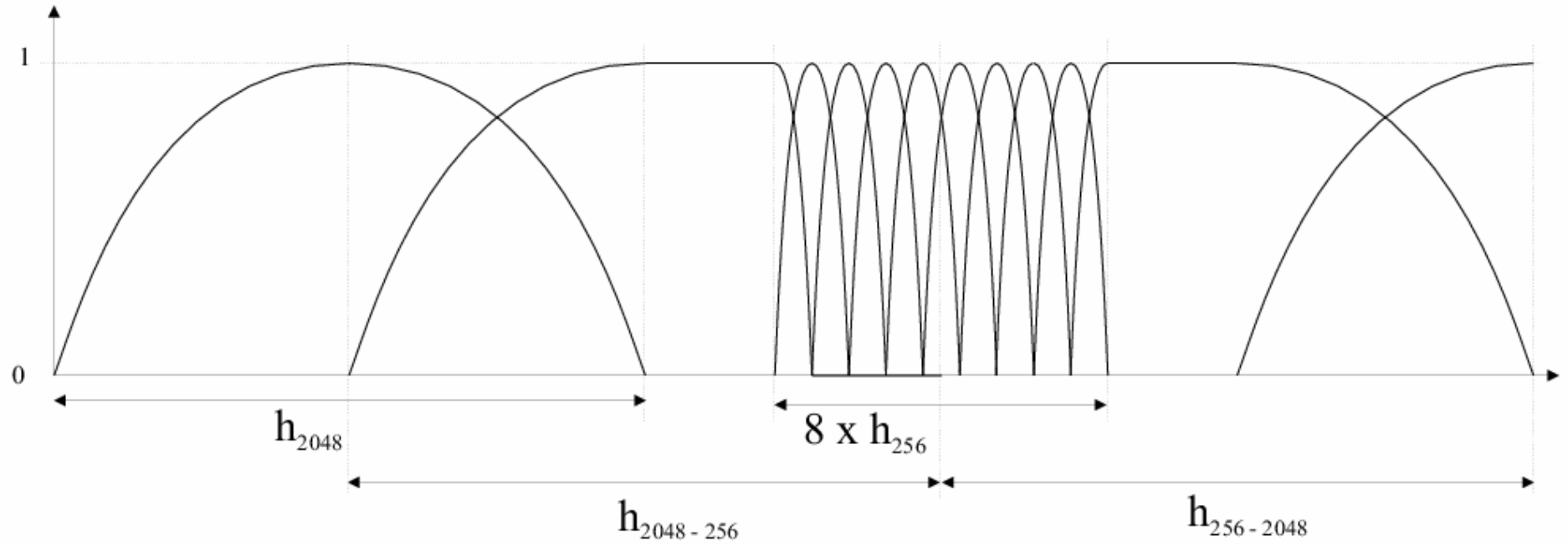
Codeur MPEG2-AAC : Transformation temps-fréquences

- Connaissant le vecteur $\underline{x}(m) = [x(mM) \cdots x(mM+N-1)] \Rightarrow \underline{X}(m) = [X(0) \cdots X(M-1)](m)$
- MDCT : Transformée en cosinus discrète modifiée (N=2048 et M=1024)
- Transformée avec recouvrement



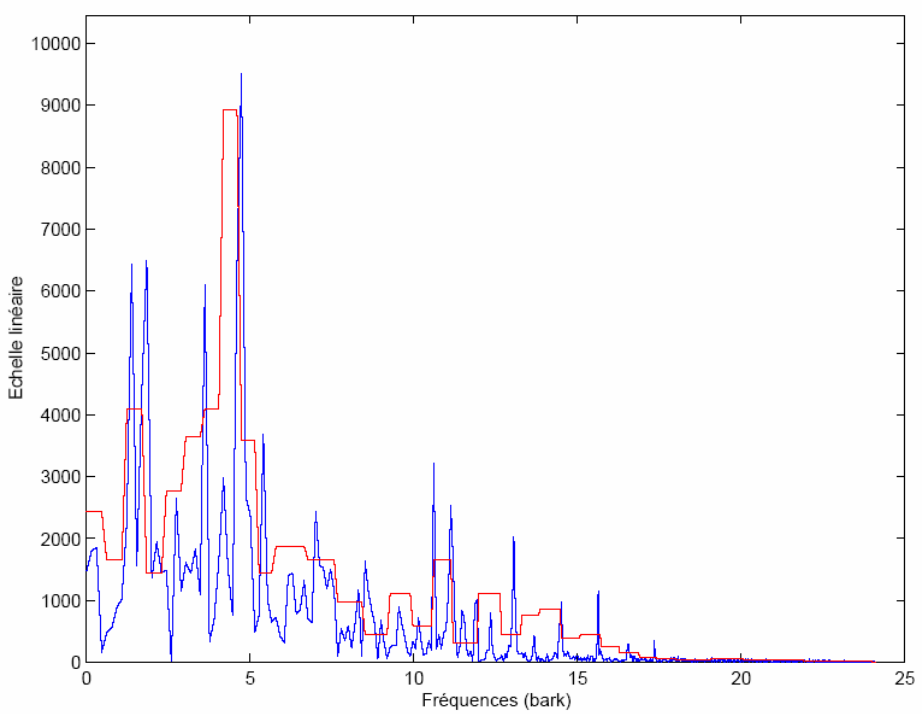
Codeur MPEG2 AAC, fenêtrage dynamique

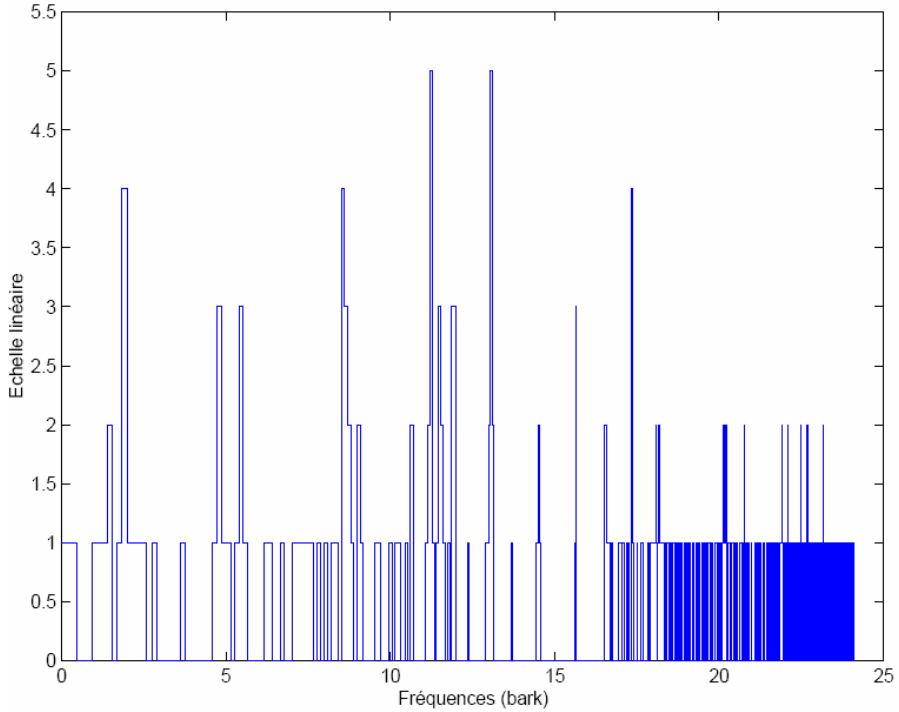
- Fenêtres longues (N=2048 et M=1024)
- Fenêtres courtes (N=256 et M=128)



Quantification des coefficients $\underline{X} = [X(0) \cdots X(M-1)]$

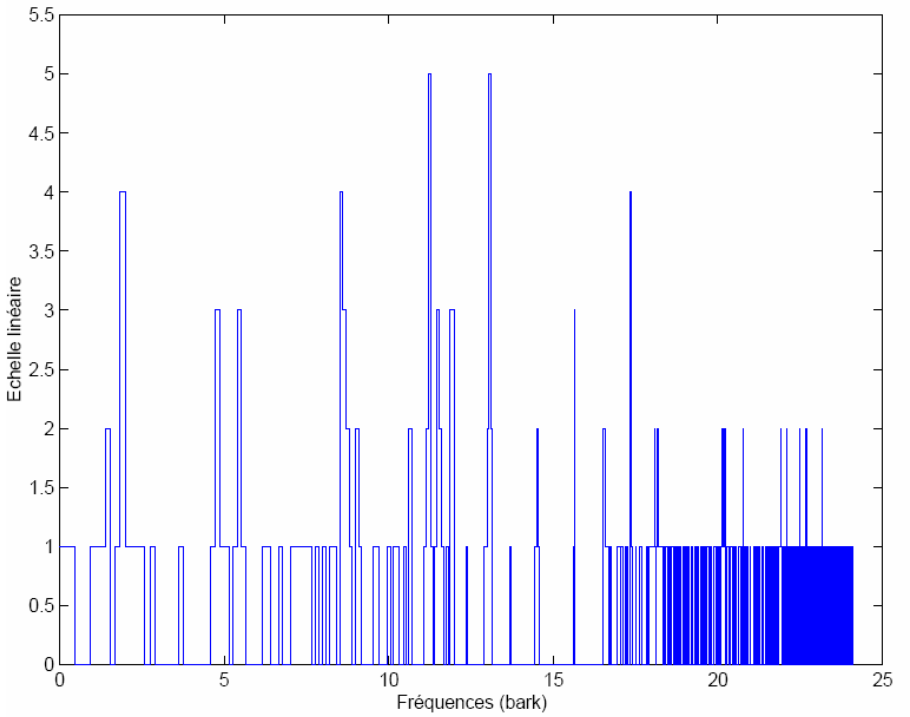
- Si on connaît le vecteur des "facteurs d'échelle" $\underline{g} = [g(0) \cdots g(M-1)]$
- A l'émetteur, calcul du vecteur d'entiers $\underline{i} = round([X(0)/g(0) \cdots X(M-1)/g(M-1)])$
- Au récepteur, reconstruction de $[\hat{X}(0) \cdots \hat{X}(M-1)] = [g(0) \times i(0) \cdots g(M-1) \times i(M-1)]$ puis $[\hat{x}(0) \cdots \hat{x}(M-1)]$
- Erreur de reconstruction $\underline{q} = [x(0) \cdots x(M-1)] - [\hat{x}(0) \cdots \hat{x}(M-1)]$





Codage du vecteur $\underline{i} = [i(0) \cdots i(M-1)]$

- 1ère solution : M = 512 entiers vérifiant $0 \leq i(k) \leq 8$ (dans cet exemple)
  - » nombre de bits nécessaires B = 512x3 > nombre de bits disponibles (~ 750)
- Autre solution plus économique : codage de Huffman
  Les valeurs les plus probables codées sur moins de 3 bits, les valeurs les moins probables sur plus de 3 bits
- En réalité : partition de l'axe des fréquences en 51 "bandes"
  Dans chaque bande : détermination de max $i(k)$ puis codage séparé
- Dans le "bitstream" : mots de code des $i(k)$ + mots de code des max $i(k)$ + mots de code des $g(k)$



Détermination des facteurs d'échelle $\underline{g} = [g(0) \cdots g(M-1)]$

- A priori simple problème d'optimisation : déterminer $\underline{g}$ minimisant la puissance de l'erreur de reconstruction sous la contrainte que le nombre de bits nécessaire soit inférieur au nombre de bits disponibles
- Solution non réaliste car elle n'autorise que des taux de compression faibles (de l'ordre de 2) si on veut que le signal reconstruit soit "transparent"
- Pour obtenir des taux de compression élevé (de l'ordre de 10) : exploitation des résultats de psychoacoustique
- "Mise en forme spectrale" du bruit de reconstruction
- Problème d'optimisation sous 2 contraintes
  - » Contrainte de débit : nombre de bits nécessaire < nombre de bits disponible
  - » Contrainte "psychoacoustique" : $S_Q(f) < \Phi(f) \ \forall f$

Codeur MPEG2 AAC : informations transmises

- Dans chaque fenêtre d'analyse (1024 échantillons i.e. 23 ms lorsque fe = 44.1 kHz)
  - » Partition de l'axe des fréquences en 51 "sous-bandes"
  - » Facteurs d'échelle
    - Codage du 1er directement sur 8 bits
    - Codage de $\Delta(k) = g(k)-g(k-1)$ pour les 50 suivants
    - Utilisation d'une table de Huffman
  - » Coefficients de la MDCT
    - 4 composantes dans la 1ère sous-bande, 32 dans la dernière
    - Codage du signe à part
    - Détermination de la valeur max dans chaque sous-bande, choix d'une table de Huffman parmi 11 (information transmise dans la chaîne binaire)
    - Codage des $i(k)$