

Master de sciences et technologie
Mention : Informatique
Cours : traitement du signal



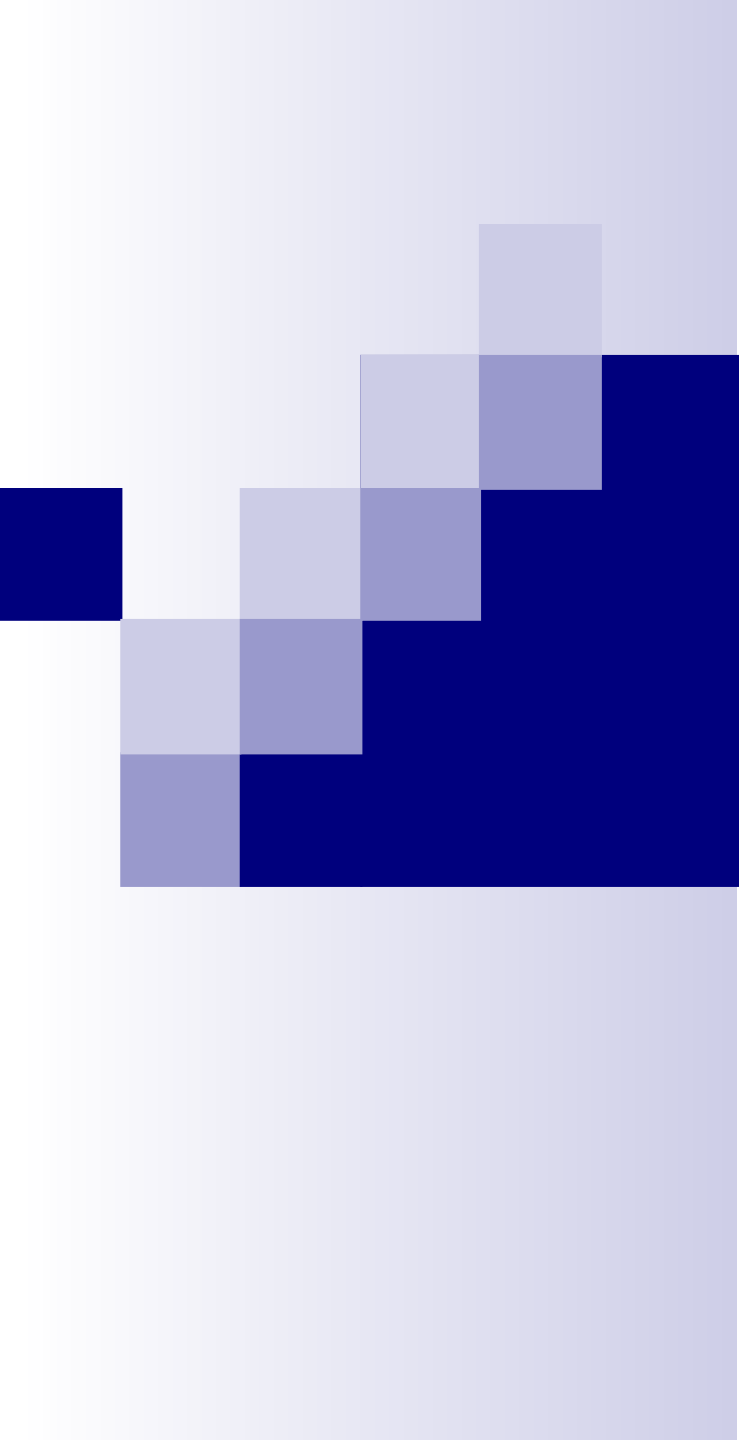
Introduction à la compression Audio

Yvan BONNASSIEUX
Email : yvan.bonnassieux@polytechnique.edu

2008-2009

Sommaire

- *Introduction*
- *Numérisation d'un signal*
- *Compression Différentiel*
- *Appareil phonatoire humain*
- *Compression LPC*
- *Appareil auditif humain*
- *Compression MPEG II layer 3*



Introduction

LE SON, qu'est ce que c'est ?

Exemple de l'onde sonore d'un bruit



Son musical = fréquence fondamentale
+ harmoniques
Bruit = + Oscillations aléatoires

Définition

- Objectivement :

phénomène physique d'origine mécanique,
fluctuations rapides de la pression de l'air au niveau
des oreilles (ondes acoustiques)

- Subjectivement :

sensation traduisant la perception par le cerveau
d'une information extérieure

- Le non audible :

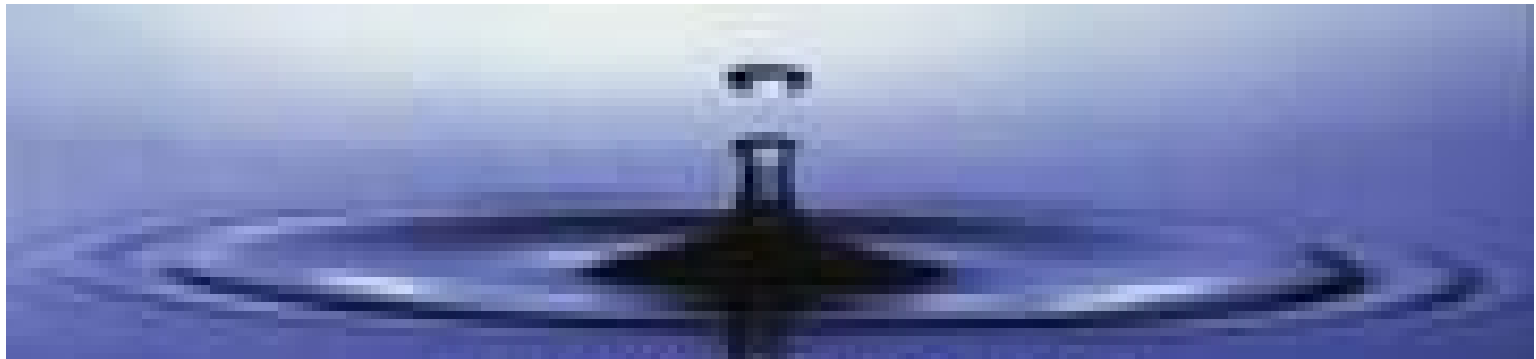
infrasons (< 20 ou 25 Hz), et ultrasons (> 15 ou 20 kHz).

Emission - Propagation

Propagation sous forme d'ondes :

- pression
- vitesse vibratoire

⇒ intensité sonore = flux d'énergie par unité de surface



Emission - Propagation

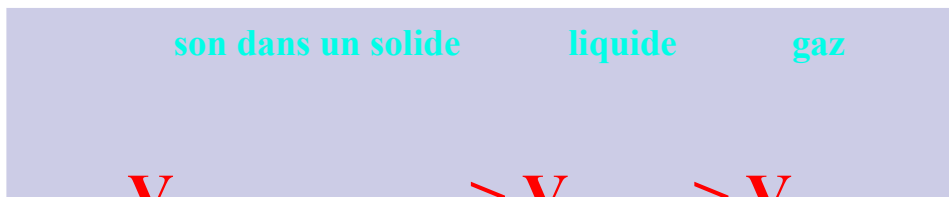
Son → Onde sonore → molécules du milieu vibrent autour d'une position moyenne

La vitesse varie suivant le milieu de propagation

Facteurs : densité (masse volumique), pression, température, dilatation...

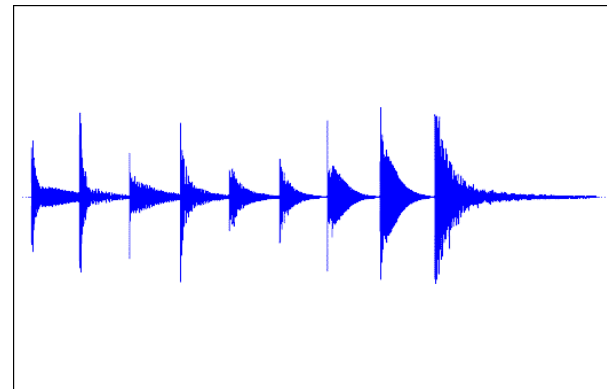
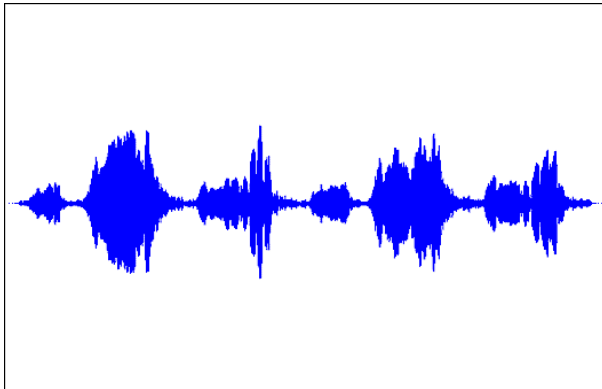
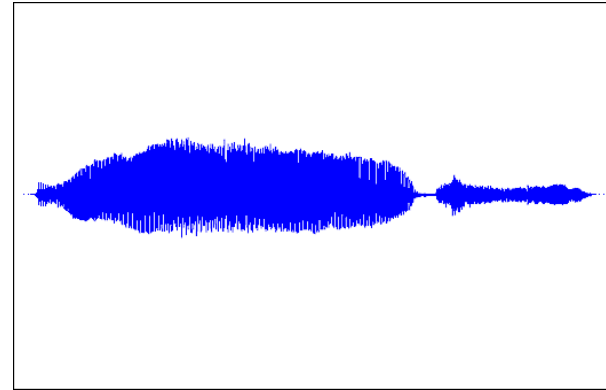
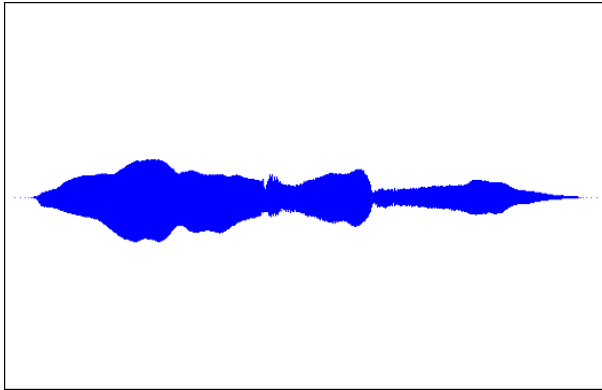
Vitesses (m/s) : ordre de grandeur (à 0°C):

Dans l'air	341
Eau douce	1435
Eau de mer	1512
Acier	5000
Aluminium	6400



Représentation du son

- Représentation temporelle: amplitude-temps

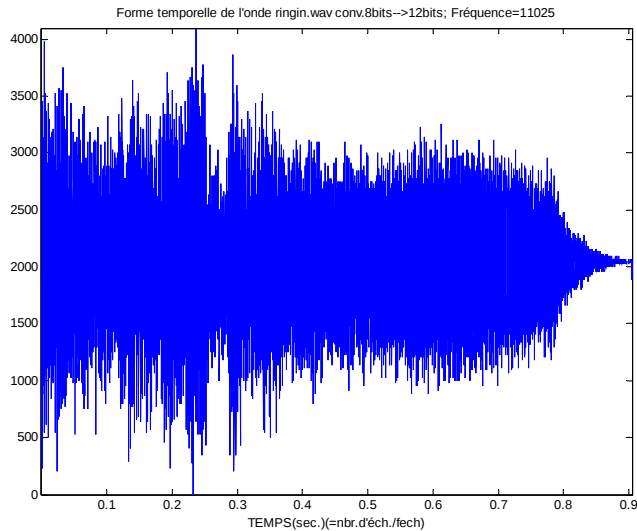


Evolution temporelle de l'enveloppe → peu d'informations sémantiques/caractéristiques

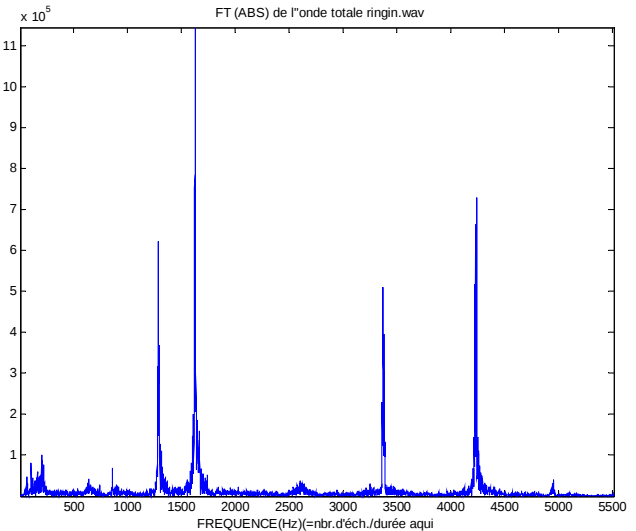
Regarder le son n'informe pas sur son contenu fréquentiel: le son n'est pas une image!

Représentation du son

- Représentation fréquentielle: amplitude-fréquence



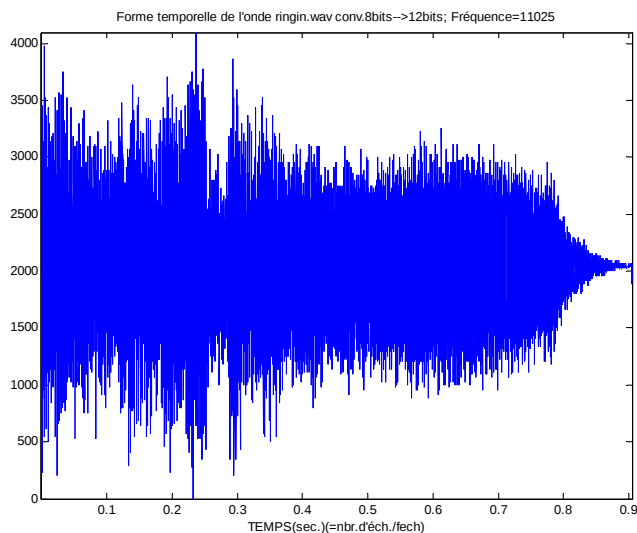
FT



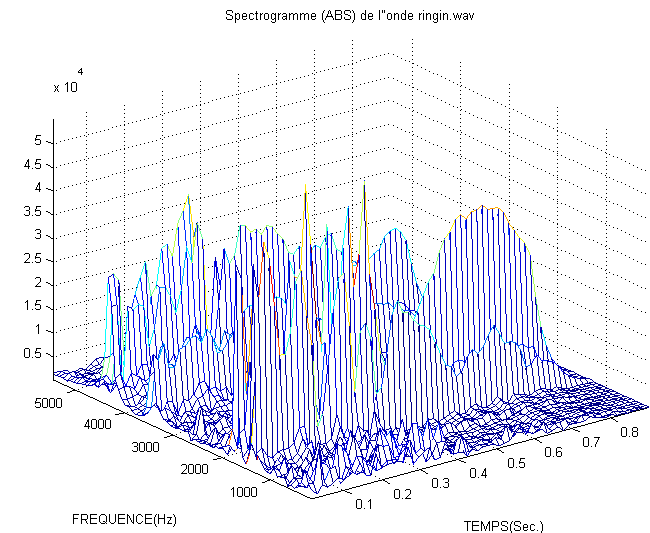
Spectre → perte de l'information temps

Représentation du son

■ Représentation Temps-fréquence: amplitude-temps-fréquence



**Découpage-
FT à court
terme**



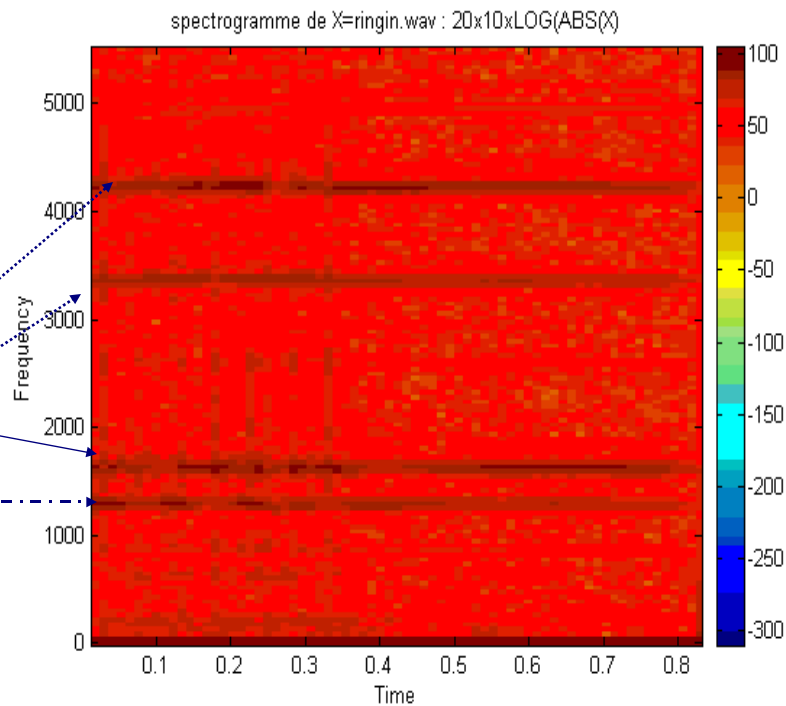
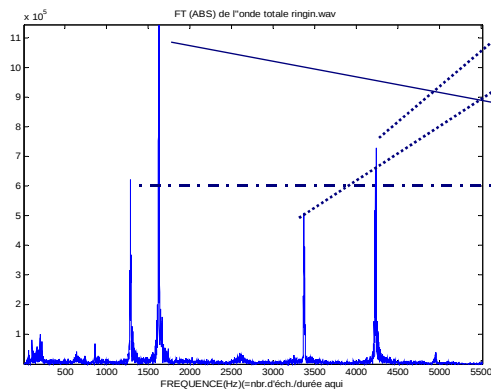
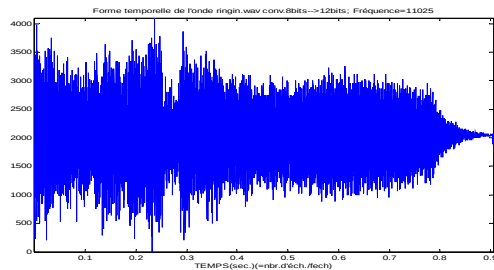
signal quasi stationnaire → courte durée.

On segmente le signal en unité temporel de 20 à 30ms

FT des segments -> caractéristiques du signal plus significatives

Représentation du son

- Représentation Temps-fréquentielle: niveau de couleur-fréquence-temps (Spectrogramme, sonagramme)

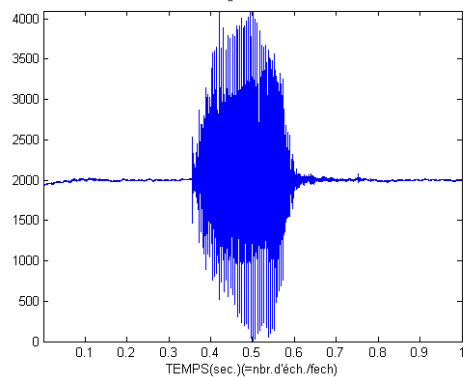


Représentation du son

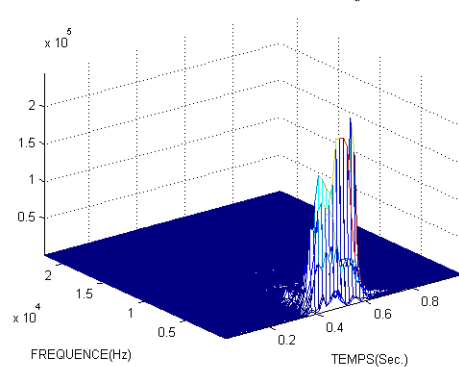
■ Représentations du SON 'a'



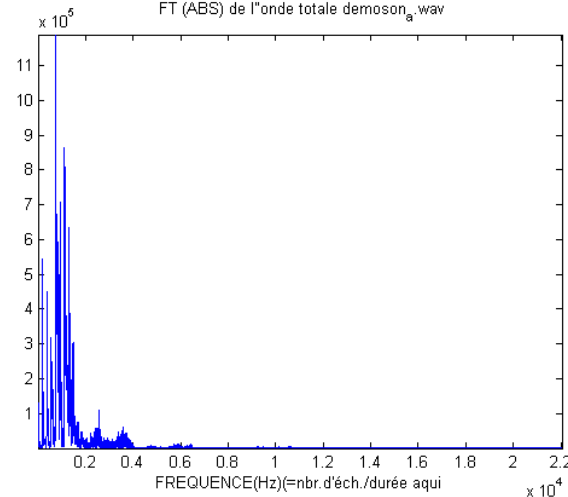
Forme temporelle de l'onde demoson_a.wav conv.16bits-->12bits; Fréquence=44100



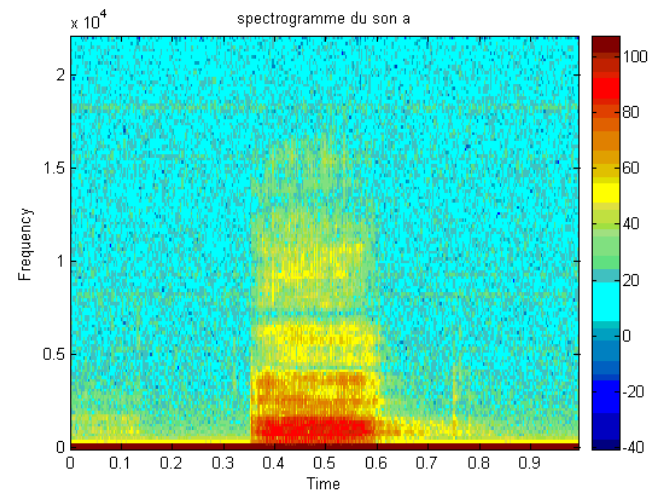
Spectrogramme (ABS) de l'onde demoson_a.wav



FT (ABS) de l'onde totale demoson_a.wav



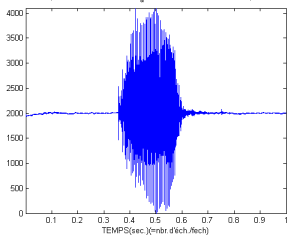
spectrogramme du son a



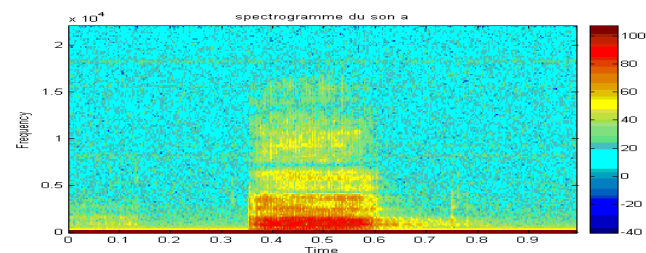
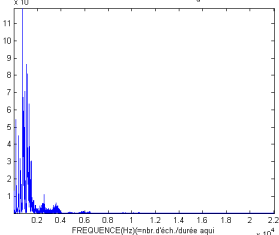
Représentation du son

Représentations des SONS 'a b i aaa'

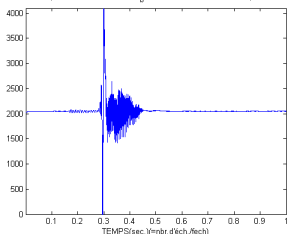
Forme temporelle de l'onde demon_01.wav con: 16bits->12bits; Fréquence=44100



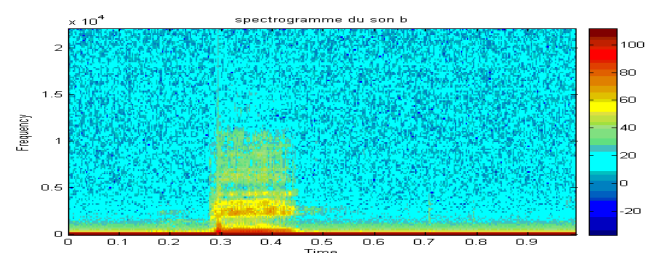
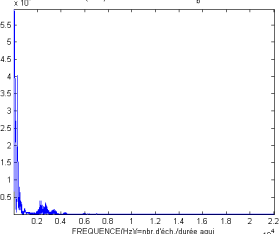
FT (ABS) de l'onde totale demon_01.wav



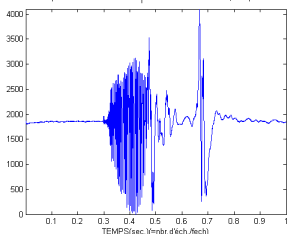
Forme temporelle de l'onde demon_02.wav con: 16bits->12bits; Fréquence=44100



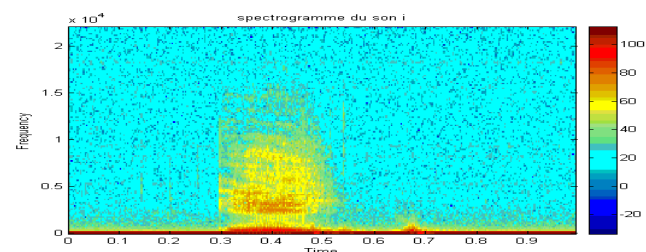
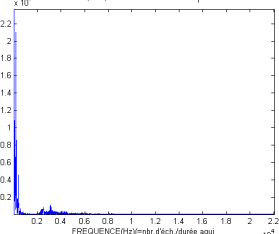
FT (ABS) de l'onde totale demon_02.wav



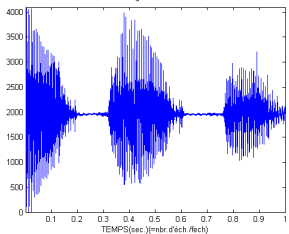
Forme temporelle de l'onde demon_03.wav con: 16bits->12bits; Fréquence=44100



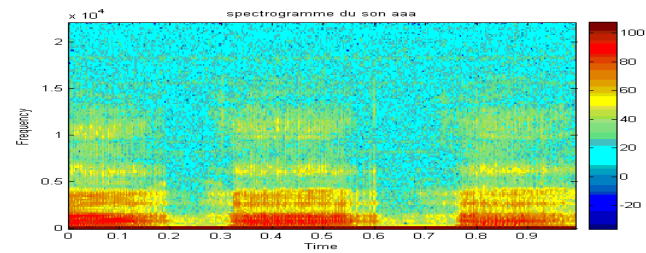
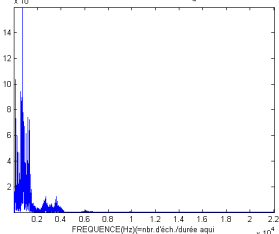
FT (ABS) de l'onde totale demon_03.wav



Forme temporelle de l'onde demon_04.wav con: 16bits->12bits; Fréquence=44100



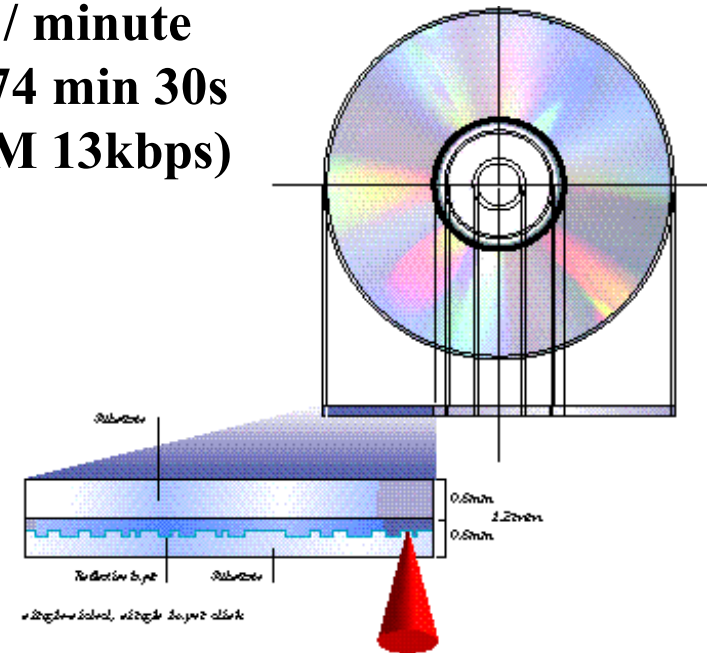
FT (ABS) de l'onde totale demon_04.wav



Pourquoi la compression ?

Beaucoup de bits pour peu d'espace ou de temps

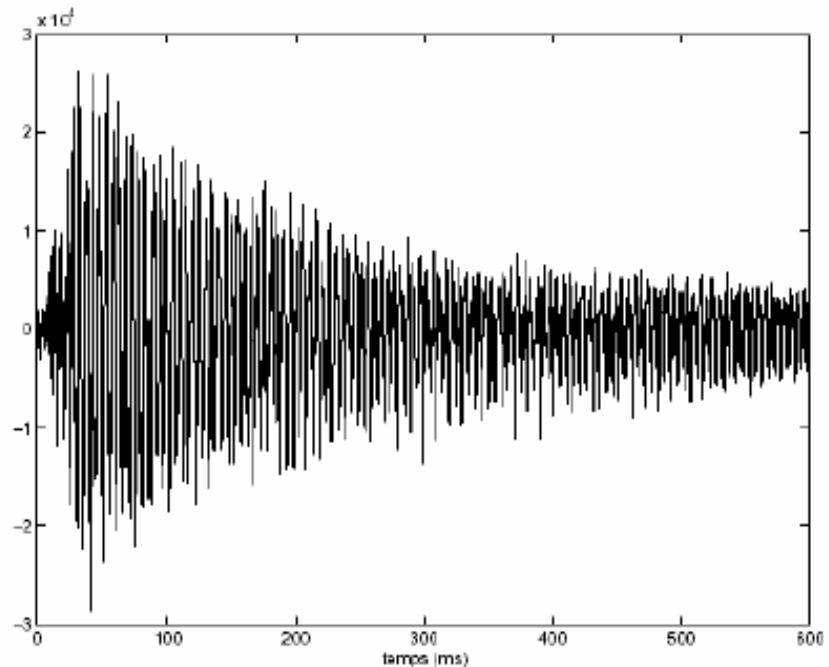
- Débit binaire brut CD : $2 * 16 * 44100 = 1.411.200$ bps
- CD audio : 10,584,000 octets / minute
- CD : 680 Mo soit seulement 74 min 30s
- Téléphone RTC 64kbps (GSM 13kbps)
- Modem ADSL 1Mbps



Robustesse : permettre la correction d'erreur par redondance

Codage parole , Codage musique

- Signal de parole
 - ✚ Modèle de production simple
 - ✚ Modèle source/filtre
- Signal de musique (audio)
 - ✚ Difficile à caractériser
 - ✚ Pas de modèle de production simple
 - ✚ Structure harmonique forte
 - ✚ Dynamique en puissance plus importante



Note de guitare échantillonnée à 32 kHz



Des applications/contextes variés

Stockage, téléchargement:

- μ Equipements audio, CD, DVD, cartes mémoires ...
- Messageries, répondeurs

Diffusion Audio

- télévisions, radios numériques
- diffusion sur Internet: Musique à la demande, streaming, radios

Communication interpersonnelle & de groupe

- téléphonie (fixe, mobiles, IP)
- audio/visioconférences, chats, forums
- communications militaires
- communications par satellites, flottes embarquées, ..

Gamme de qualité

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

Caractéristique du signal audio

- Débit
 - ☞ Reflète le degré de compression
 - ☞ Varie selon la qualité de restitution demandée (384 ... 2 kbit/s)
- Complexité
 - ☞ Impact sur coût et puissance consommée
 - ☞ MIPS, RAM, ROM, ...
- Retard
 - ☞ Paramètre critique pour applications conversationnelles
 - ☞ < 150 ms, perte d'interactivité au-dessus de 400 ms
- Qualité
 - ☞ Fonction du type de signal transmis (parole, bruit, musique, modems...)
 - ☞ Déterminée par des tests subjectifs

Nécessité de la normalisation

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

Principaux organisme de normalisation

- ➔ **ITU-T SG16/WP3/Q7-10**: Codage Audio/Parole pour les **services multimedia** sur réseaux fixes & paquets
 - ✓ Téléphonie, CME, VoIP, Frame Relay, visiophonie, communications de groupe, ...
- ➔ **ISO/IEC SC29/WG11 MPEG Audio**: **Représentation codée de l'information audio**, contenu multimedia, diffusion, streaming de musique sur internet, radios en ligne ,...
- ➔ **3GPP**: Organisme de standardisation pour les systèmes mobiles de **3^{ème} génération** basé W-CDMA & TD-CDMA (ETSI, T1, TTC, ARIB, TTA, CWTS)
 - ✓ **SA4** : téléphonie NB & WB, téléphonie multimedia bas débits, services de streaming/ mode paquet, services conversationnels/ mode paquet
- ➔ **3GPP2** : autre organisme de standardisation pour les systèmes mobiles de **3^{ème} génération** basé CDMA (TIA, TTC, ARIB, TTA, CWTS)
- ➔ **+ INMARSAT, NATO, DoD, standards "régionaux", ...**
 - ✓ Mobiles, transmissions par satellites, communications militaires ...

Codage : les technologies I

➡ Codeurs par Transformées :

- ✓ Principalement pour l'audio (*MP3, MPEG2-AAC, MPEG4-AAC, TVQ*) + *G.722.1*
 - ⇒ Utilisés pour différentes bandes passantes (≈ 3 à 20 kHz)
 - ⇒ Qualité : d'excellente à bonne pour la musique mais la parole demande du débit
 - ⇒ Délais : potentiellement élevés
 - ⇒ Complexité : moyenne à élevée /codeur, basse à moyenne/ décodeur

➡ Codeurs temporels :

- ✓ Parole NB haute qualité : (*G.711, G.726*), technologies anciennes
 - ⇒ Généralement peu complexes, pas de délai

➡ Codeurs en Sous-bandes :

- ✓ Pour les bandes passantes/qualités élevées (*MPEG1 LI,II, MPEG2, G.722*)
 - ⇒ Délais moyens, débits élevés pour Haute Qualité

Codage : les technologies II

➡ Codeurs AbS/CELP :

- ✓ Parole NB ou WB : *GSM-FR, GSM-HR, GSM-EFR, NB-AMR, WB-AMR/G.722.2, G.728, G.723.1, G.729, MPEG-CELP*
- ✓ CELP/sous bandes en WB (*CELP-MPEG, WB-AMR/G722.2*)
- ✓ RCELP (*3GPP2/EVRC, SMV*)
- ✓ CELP/RCELP (*3GPP2/VMR-WB*)
 - ⇒ **Qualité** : Bonne/parole, Moyenne/Bruits, Musique: généralement mauvaise
 - ⇒ **Complexité** : relativement élevée (codeur)
 - ⇒ **Délais** : de l'ordre de $\approx 35-100$ ms

➡ Codeurs Paramétriques :

- ✓ Parole : *HVXC* (paramétrique), *MELP* (NATO, DoD), *IMBE* (INMARSAT)
- ✓ Musique : *SSC* (norm. MPEG4 en cours), *HILN* (sinusoïdal)
 - ⇒ **Modélisation** du signal d'entrée
 - ⇒ **Qualité** : moyenne, tendance à dépendre du type de signal
 - ⇒ **Complexités et délais potentiellement plus élevés**

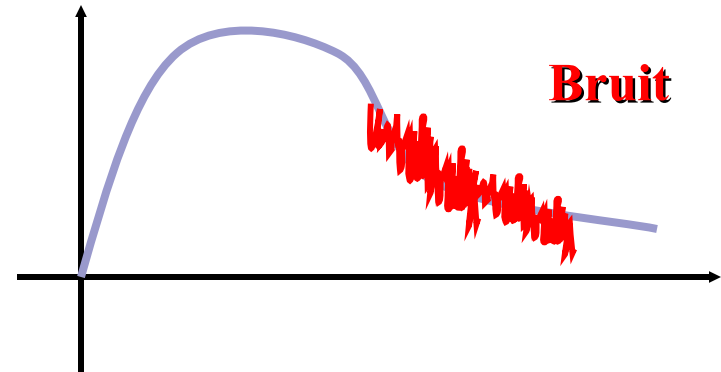


Numérisation d'un signal

Signal Analogique ou Numérique

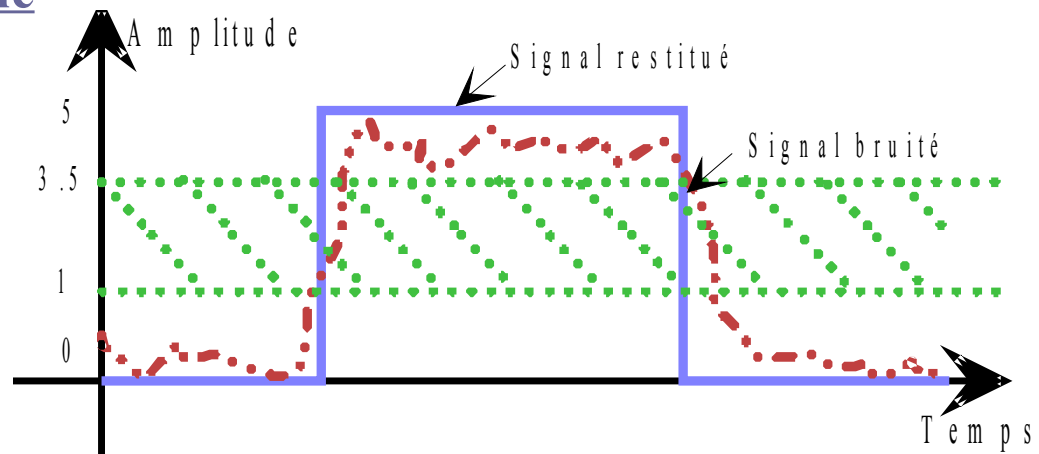
Signal Analogique

Pour traduire un signal quelconque, on associe à chaque instant l'amplitude du dit signal à l'amplitude d'un signal électrique image.



Signal numérique

Insensible au bruit



Codage binaire

2 niveaux électriques : un pour coder « **1** » ; l'autre pour coder « **0** »

Échantillonnage des signaux

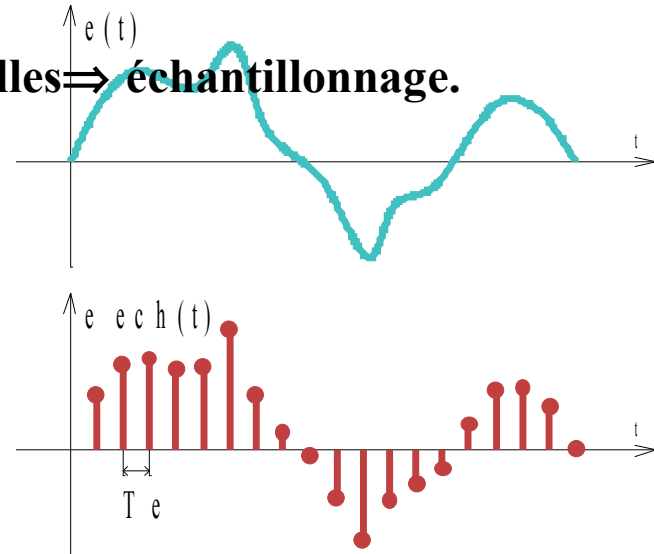
Définition

Système numérique $\Rightarrow$ nombre finit de données.

Décomposer en une suite de valeurs ponctuelles $\Rightarrow$ échantillonnage.

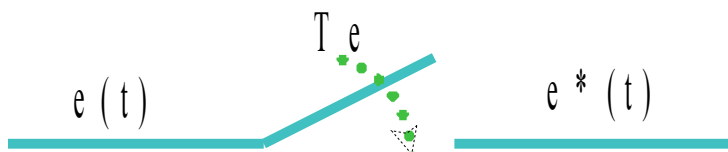
Période d'échantillonnage T_e .

Nombre fini de données



Échantillonneur
idéal

$$e^*(t) = \sum_{n=-\infty}^{+\infty} e(t) \cdot \delta(t - nT_e) = e(t) \cdot p_{nT_e}(t)$$



$e(t)$ un signal temporel analogique
 T_e la période d'échantillonnage

Réversibilité de l'échantillonnage

Comment choisir la fréquence d'échantillonnage, de façon à permettre la reconstruction de $e(t)$ à partir de $e^*(t)$?

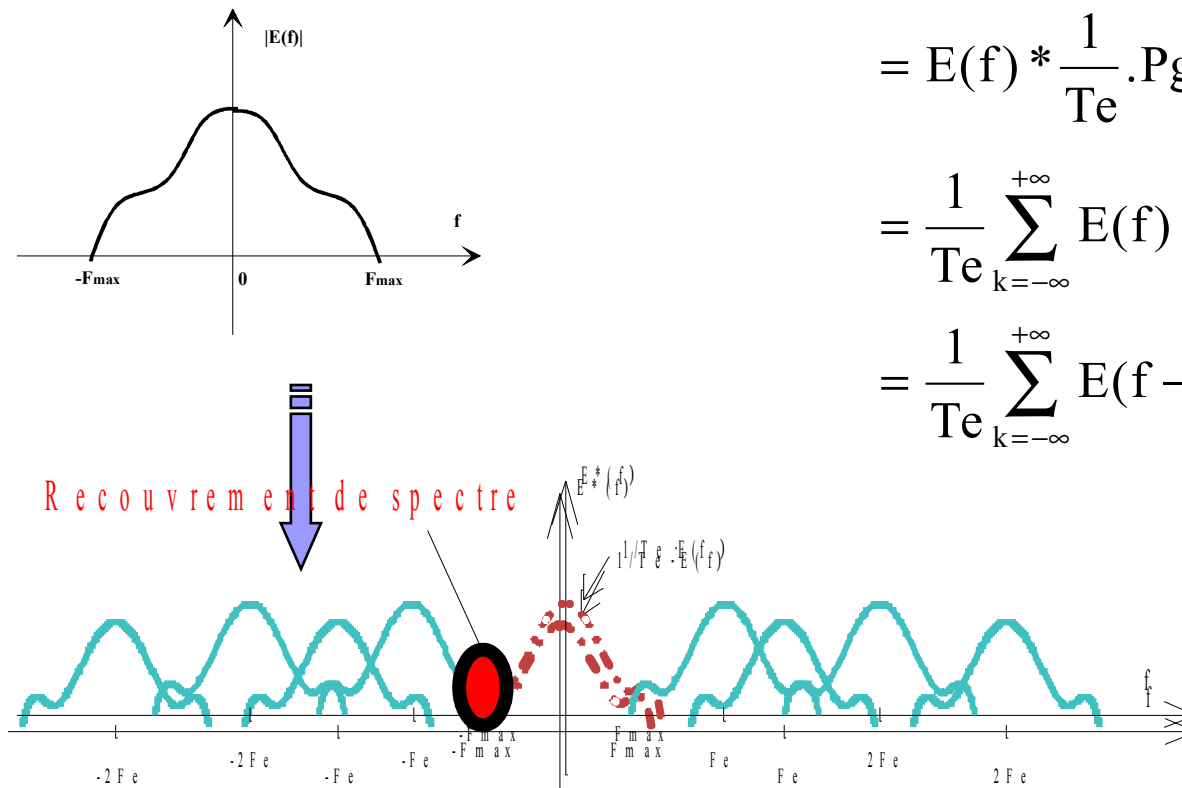
Spectre du signal échantillonné

$$E^*(f) = F(e^*(t)) = F(e(t) \cdot p_{n_{T_e}}(t))$$

$$= E(f) * \frac{1}{T_e} \cdot p_{n_{\frac{1}{T_e}}}(f)$$

$$= \frac{1}{T_e} \sum_{k=-\infty}^{+\infty} E(f) * \delta(f - \frac{n}{T_e})$$

$$= \frac{1}{T_e} \sum_{k=-\infty}^{+\infty} E(f - \frac{n}{T_e})$$

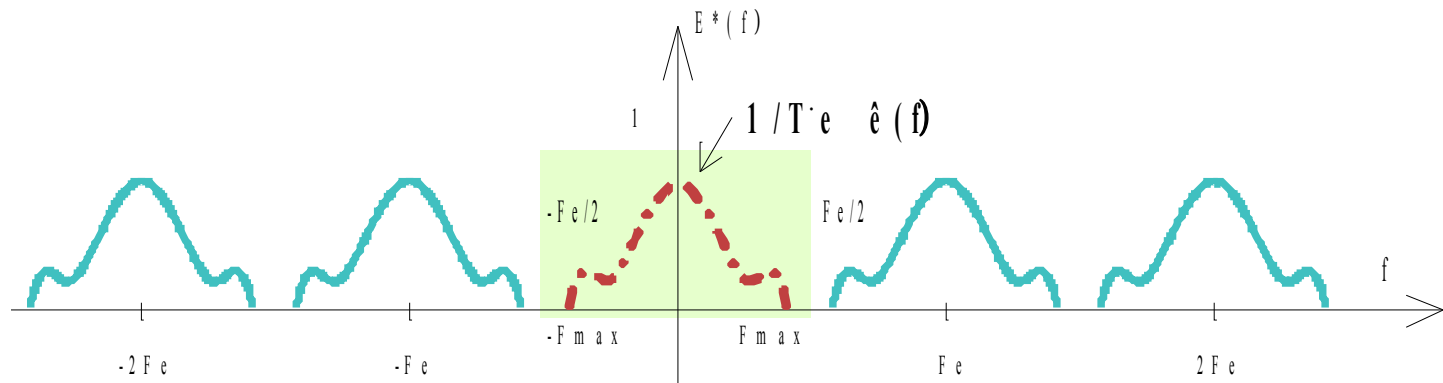


Théorème de Shannon

Définition

Pour pouvoir envisager la reconstruction du signal $e(t)$ à partir du signal $e^*(t)$, il faut donc respecter l'inégalité suivante :

$$\mathbf{F_e > 2.F_{max.}} \text{ avec } \begin{cases} F_{\max} & \text{borne supérieure de } E(f) \\ F_e & \text{Fréquence d'échantillonnage} \end{cases}$$



Autre vision

Minimum 2 échantillons par période pour définir une sinusoïde $e(t) = \sin(2\pi F_0 t)$.

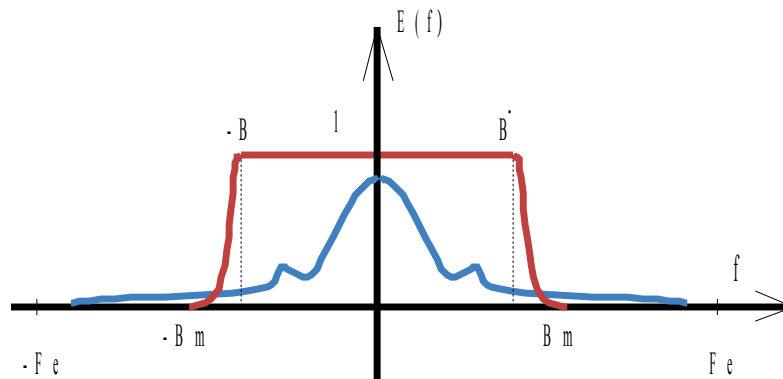
0

Donc, si la fréquence de $e(t)$ est F_0 , on doit échantillonner à $F/2$

Filtre anti-repliement

atténuation du spectre du signal d'origine au delà de $F_e/2$

⇒ filtre anti-repliement.



Dans la réalité, tout filtre anti-repliement possède une bande de transition qui reporte la bande passante limite B_m bien au-delà de la bande passante B . Dans ce

m

cas, le théorème de Shannon devient: **$F_e > 2.B$** .

Exemples

le CD Audio

$F_e = 44.1$ Khz

Ligne MIC

$F_e = 8$ Khz

Échantillonneur bloqueur

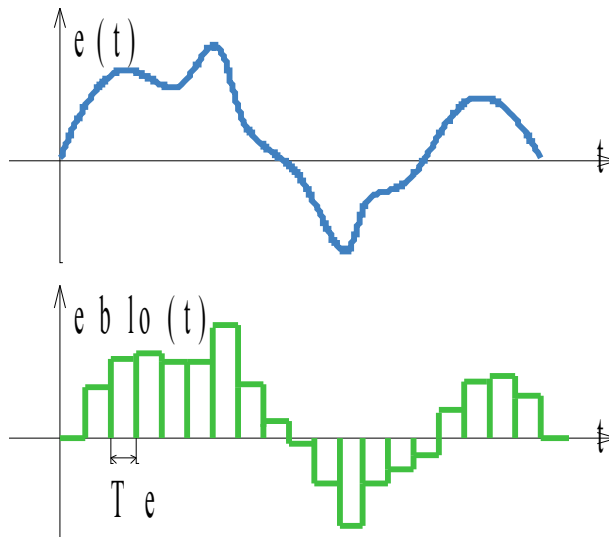
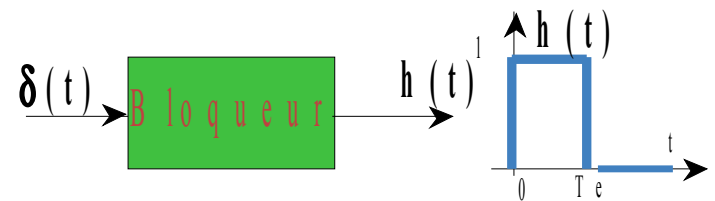
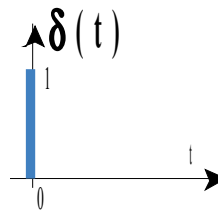
Nécessité d'un bloqueur

convertisseur analogique numérique à un temps de conversion non nul
Signaux échantillonnés bloqués.

Le blocage est d'une durée d'une période d'échantillonnage T_e .

Bloqueur d'ordre Zéro

Fonction de transfert

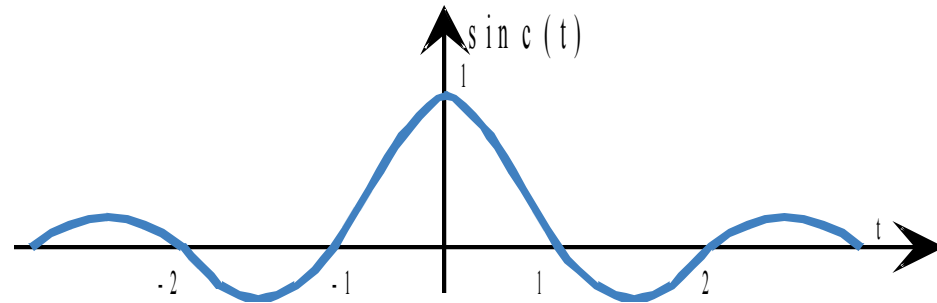


Forme temporelle d'un signal échantillonné bloqué

Échantillonneur bloqueur

TF d'un bloqueur

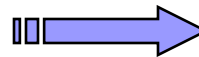
$$B_o(t) = T_e \frac{\sin(\pi T_e f)}{\pi T_e f}$$



TF d'un Signal échantillonné bloqué

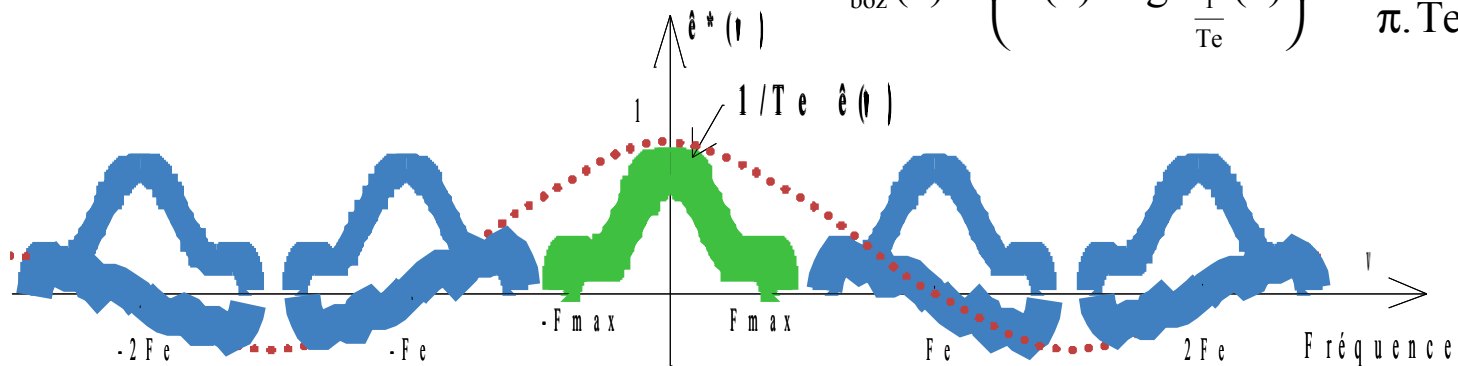
$$e_{BOZ}(t) = e * (t) * b_o(t) = (e(t) \cdot \text{pgn}_{T_e}(t)) * b_o(t)$$

$$E_{boz}(f) = \frac{1}{T_e} (E(f) * \text{Pgn}_{1/T_e}(f)) \cdot B_{boz}(f)$$



TF

$$E_{boz}(f) = \left(E(f) * \text{Pgn}_{\frac{1}{T_e}}(f) \right) \cdot \frac{\sin(\pi \cdot T_e \cdot f)}{\pi \cdot T_e \cdot f}$$



er

Transformée de Fourier discrète

Définition

Soit N échantillons $x(i)$ avec des échantillons du signal $x(t)$. La Transformée de Fourier Discrète “ T.F.D. ” notée $X^*(f)$ est définie par les N coefficients $X(k)$

$$X(k) = \sum_{i=0}^{N-1} x(i) e^{-j \frac{2\pi}{N} ik} \quad k \in \{0 \dots N-1\}$$

Remarques importantes

- Le calcul ne prend apparemment pas en compte la fréquence
- La précision c’est à dire l’écart entre deux raies contiguës est donnée par:

$$\Delta f = X(k+1) - X(k) = \frac{F_e}{N}$$

F_e fréquence d’échantillonnage et N le nombre d’échantillons



Pour être précis : N ou F

Différences TFD & TF

Échantillonnage temporel

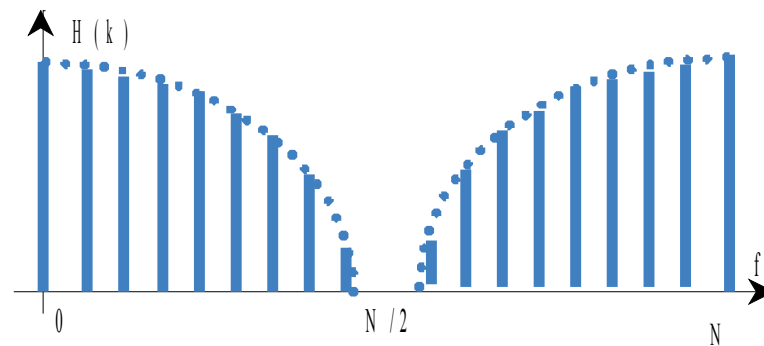
Problème : Le signal de départ est discret $\Rightarrow$ repliement de spectre.

Solution : **filtre passe-bas anti-repliement.**

Échantillonnage fréquentiel

Le spectre est échantillonné.

Il comporte autant de points que le nombre d'échantillons temporels.



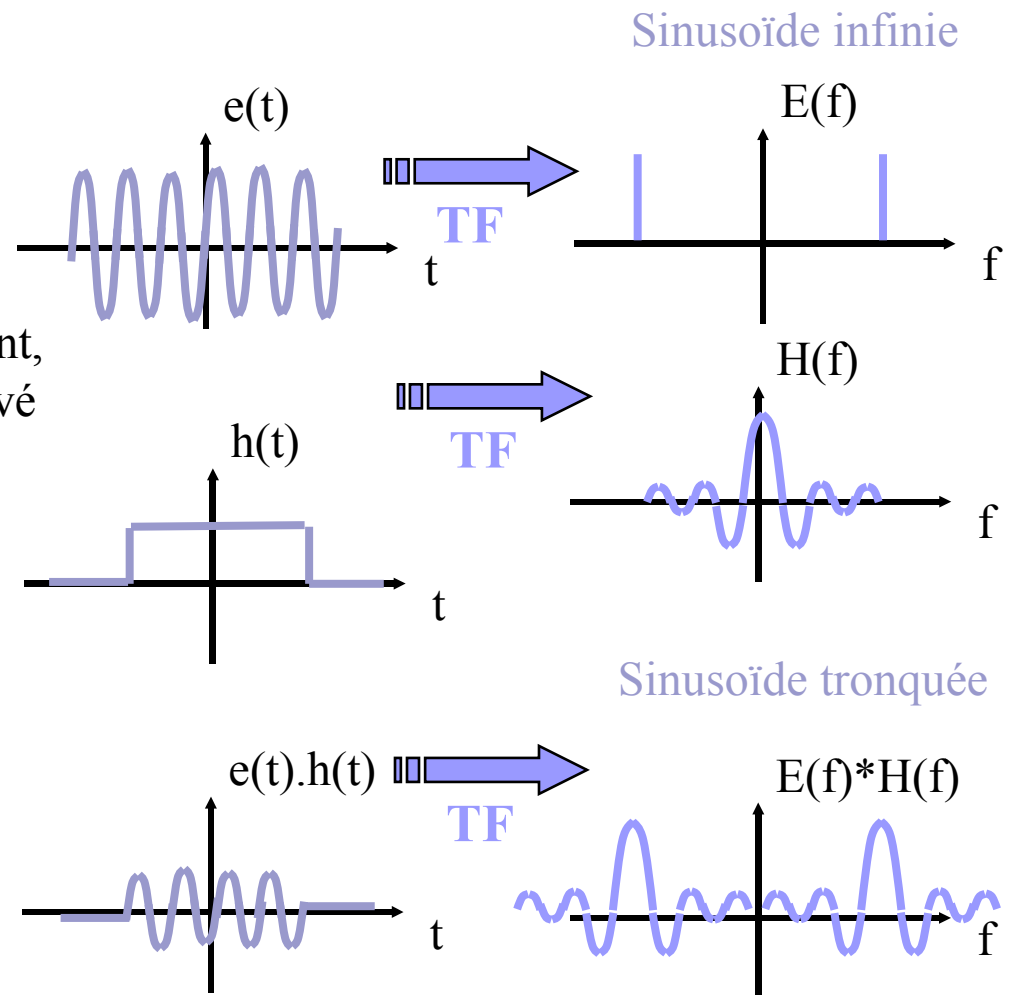
Le spectre est répétitif : Si le nombre total des $X(k)$ est N , le spectre discret formé par les $X(k)$ est symétrique par rapport à $N/2$.

On ne résonne donc que entre 0 et $F/2$

Fenêtrage :Principes

Durée de la fenêtre temporelle

Pour être traité numériquement,
le signal analogique est prélevé
pendant un temps limité
(N échantillons)



Types de fenêtre

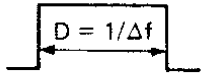
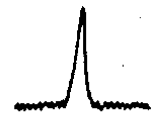
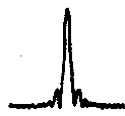
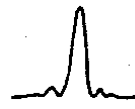
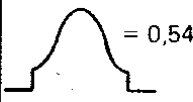
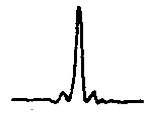
Type de fenêtre	Forme de la fenêtre temporelle et équation correspondante	Allure du spectre en fréquence	Amplitude du 1er lobe secondaire par rapport au lobe central (dB)	Largeur du lobe central (à -3dB)	Largeur équivalente au bruit	Pente d'atténuation des lobes secondaires (dB/oct)
Rectangle (fenêtre naturelle)	 $D = 1/\Delta f$		-13,2	$0,88 \Delta f$	Δf	6
Demi-sinusoïde	 $\sin \frac{\pi t}{D}$		-22,4	$1,15 \Delta f$	$1,26 \Delta f$	12
Triangle (Bartlett)	 $\frac{2t}{D}$, de 0 à $\frac{D}{2}$ $2 - \frac{2t}{D}$, de $\frac{D}{2}$ à D		-26,6	$1,28 \Delta f$	$1,33 \Delta f$	12
Cosinus carré (Hann)	 $\frac{1}{2} (1 + \cos \frac{\pi t}{D})$ $= \cos^2 (\frac{\pi t}{2D})$		-31,6	$1,39 \Delta f$	$1,5 \Delta f$	18
Cosinus carré et décalage 8 % (Hamming)	 $0,08 + 0,46 (1 + \cos \frac{\pi t}{D})$ $= 0,54 + 0,46 \cos \frac{\pi t}{D}$		-43,9	$1,26 \Delta f$	$1,36 \Delta f$	6 (à partir de $5 \Delta f$)

Fig. 6.26

Fast Fourier Transform

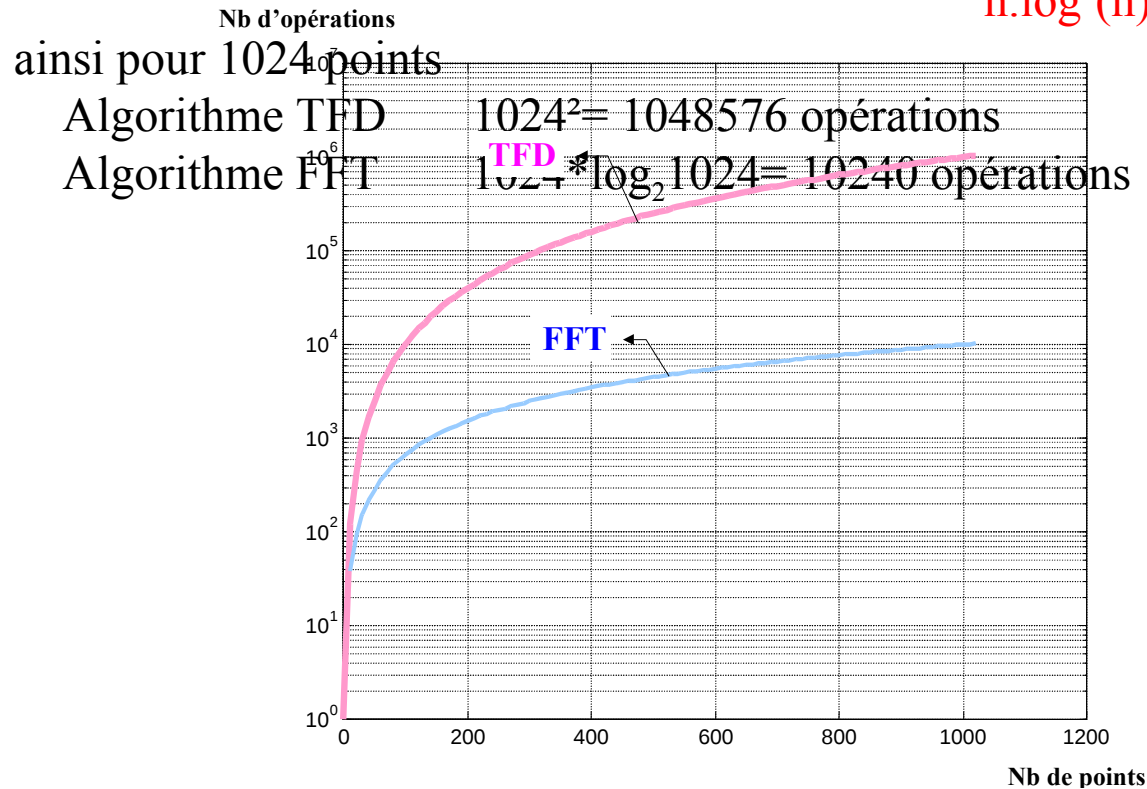
- TFD

n^2 additions et multiplication.

-

FFT “ Fast Fourier Transform ” qui pour $n=2^2$

$n \cdot \log(n)$ opérations.



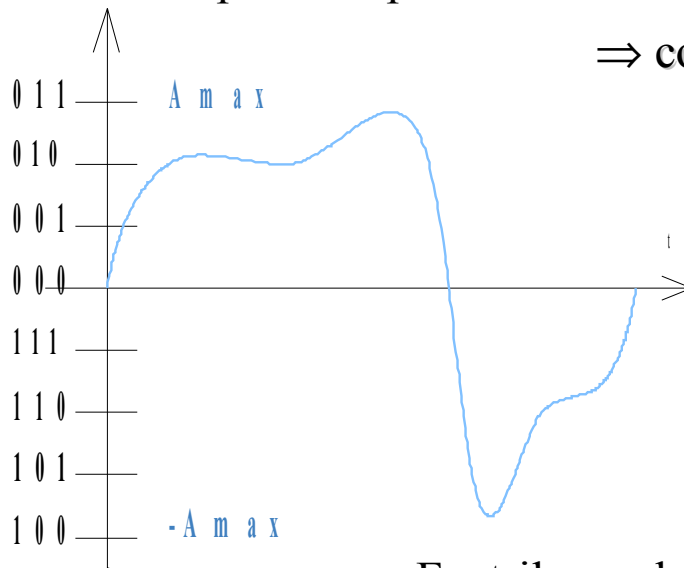
Quantification du signal

Définition

Impossible d'enregistrer toutes les valeurs des échantillons numériquement
⇒ codage avec des mots infinis.

N valeurs possible pour les échantillons

⇒ codage avec des mots de N bits



Notion de pas de quantification :
“ q ”

Problèmes de la quantification

- Faut-il prendre un pas de quantification constant quelque soit le niveau ?
- Comment choisir le pas de quantification pour que l'erreur de codage correspondante soit acceptable ?

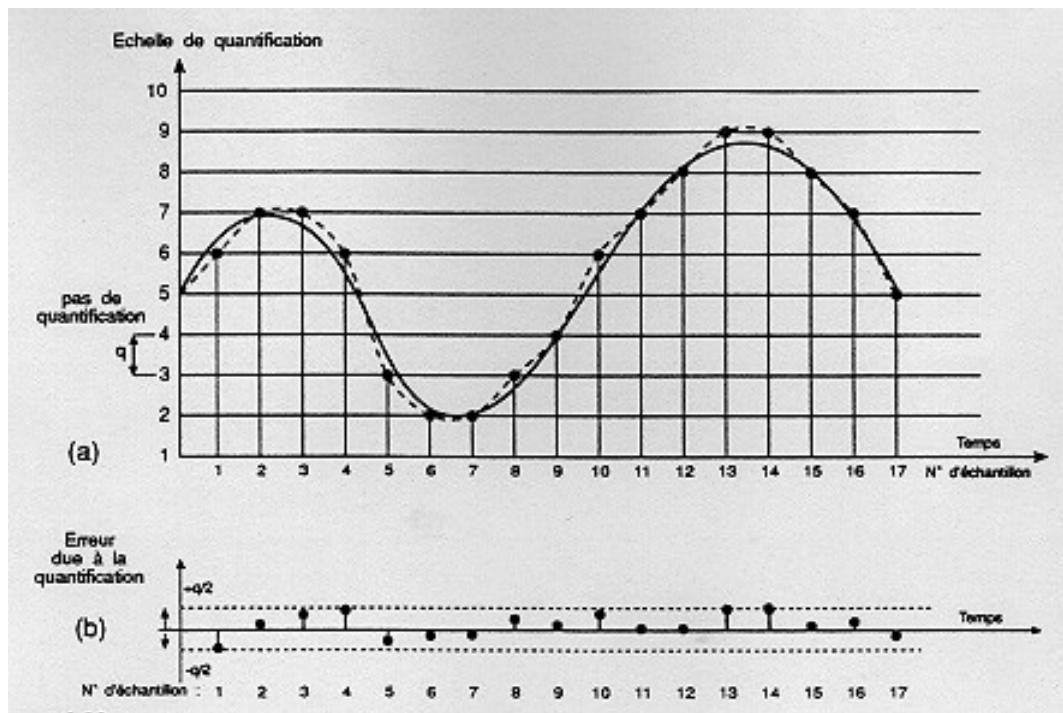
bruit de quantification

Pas de quantification linéaire

Le codage est effectué en binaire sur n bits, ceci autorise 2^n niveaux différents.

En téléphonie n=8 (256 niveaux) le Disque Compact 16 bits (65535 niveaux)

Comment définir le nombre de bits nécessaire ?



⇒ **bruit de quantification.**

Rapport S/B

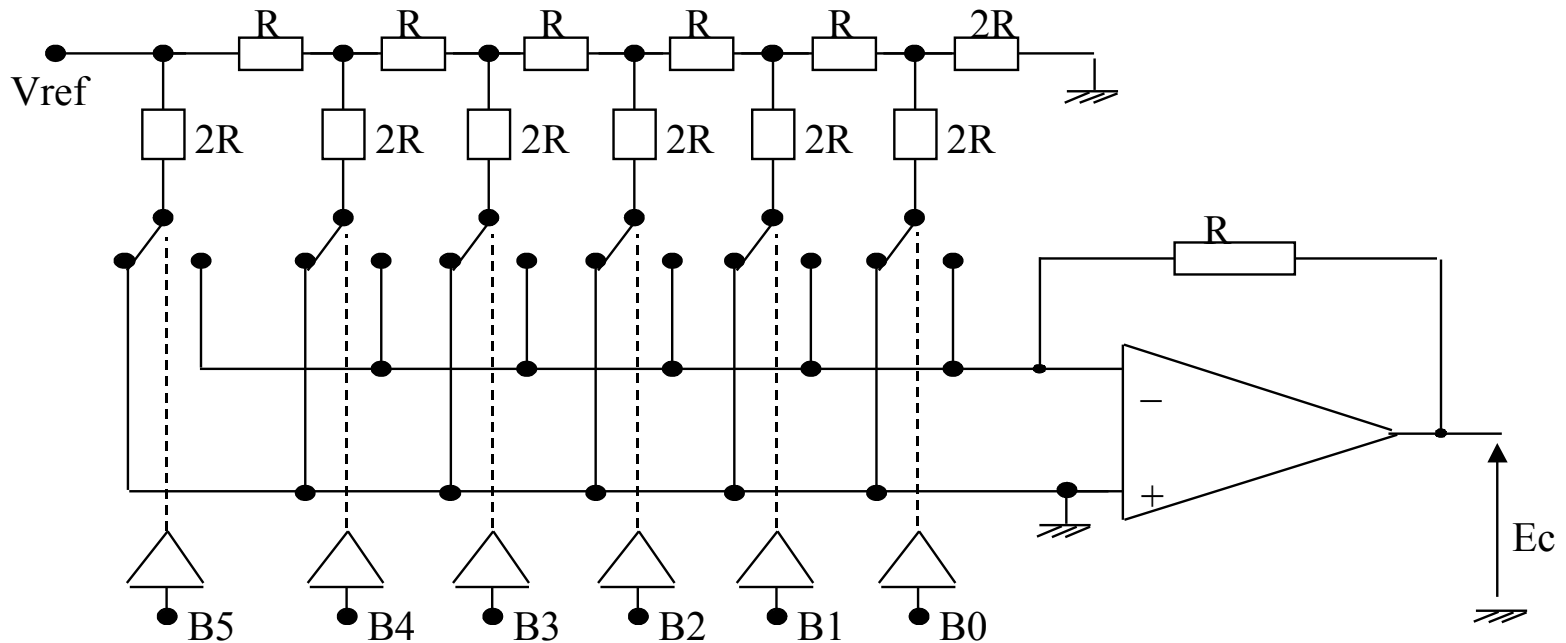
$$S = \frac{q \cdot 2^N}{q} = 2^N$$

$$S_{dB} = 20 \cdot \log(S) \approx 6 \cdot N$$

Attention à la validité de ce critère

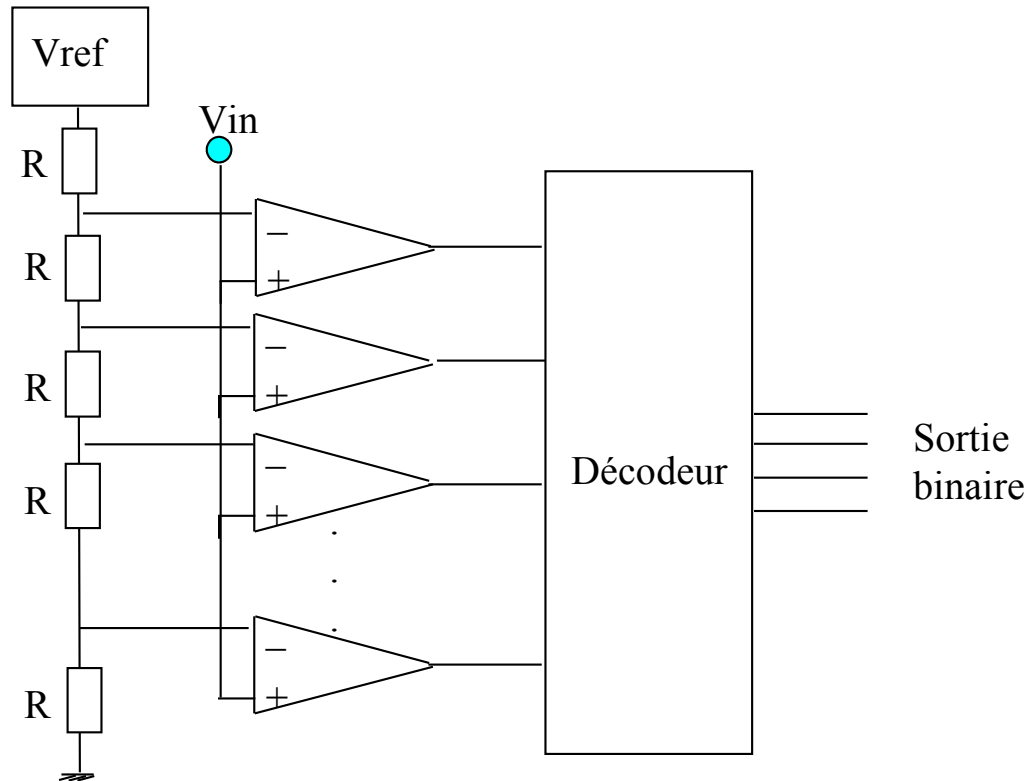
Convertisseurs Numériques Analogiques

Convertisseur CNA à réseau $R-2R$



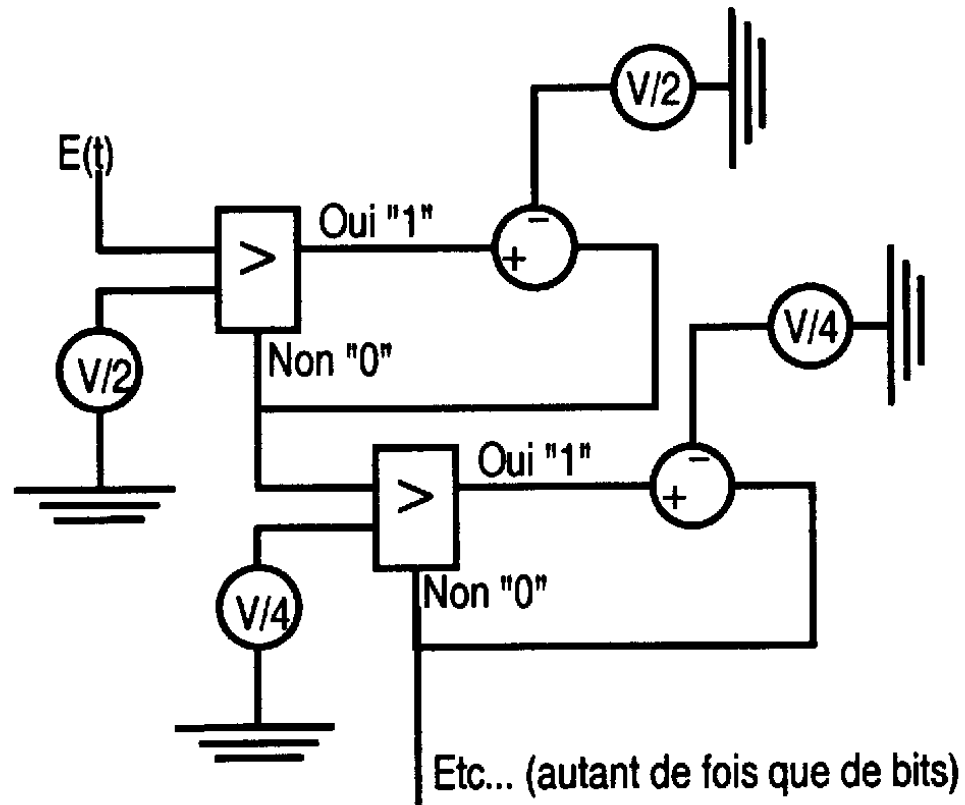
Les Convertisseurs Analogiques Numériques

Convertisseur C.A.N Parallèle ou Flash



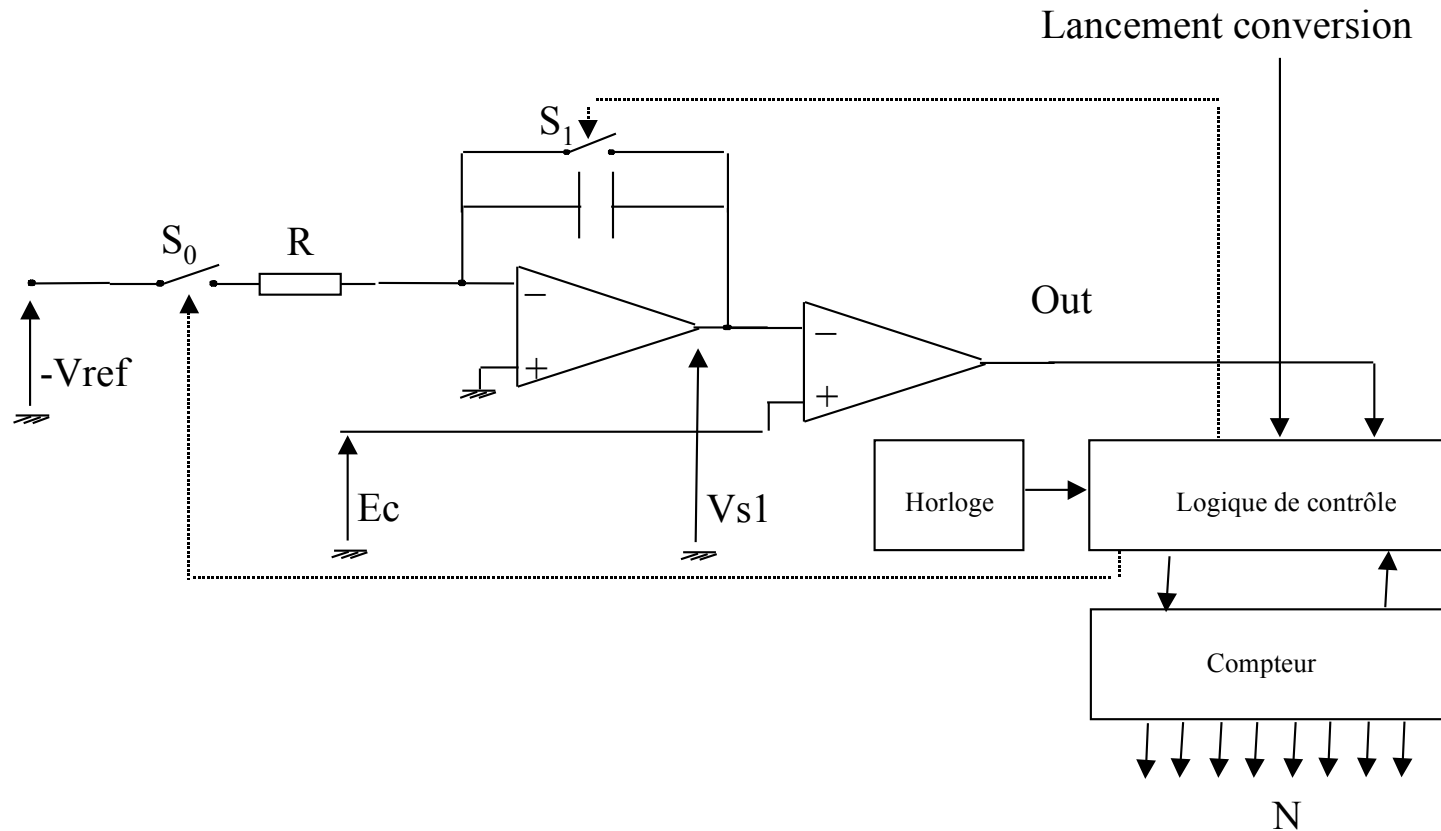
Convertisseurs Numériques Analogiques

Convertisseur par approximations successives



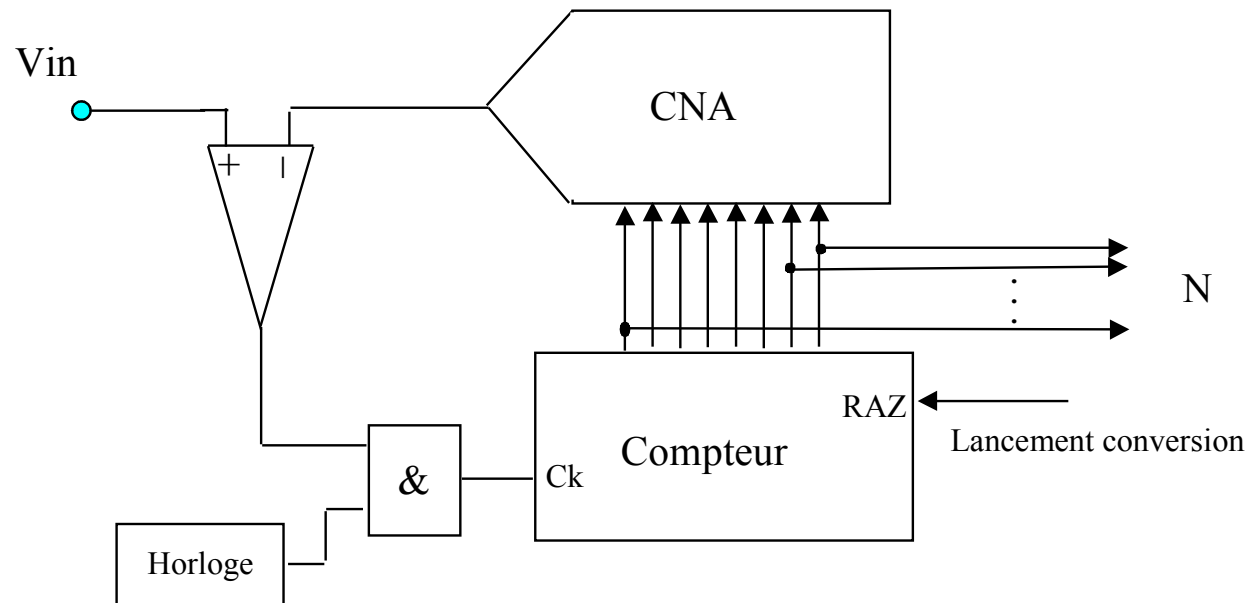
Les Convertisseurs Analogiques Numériques

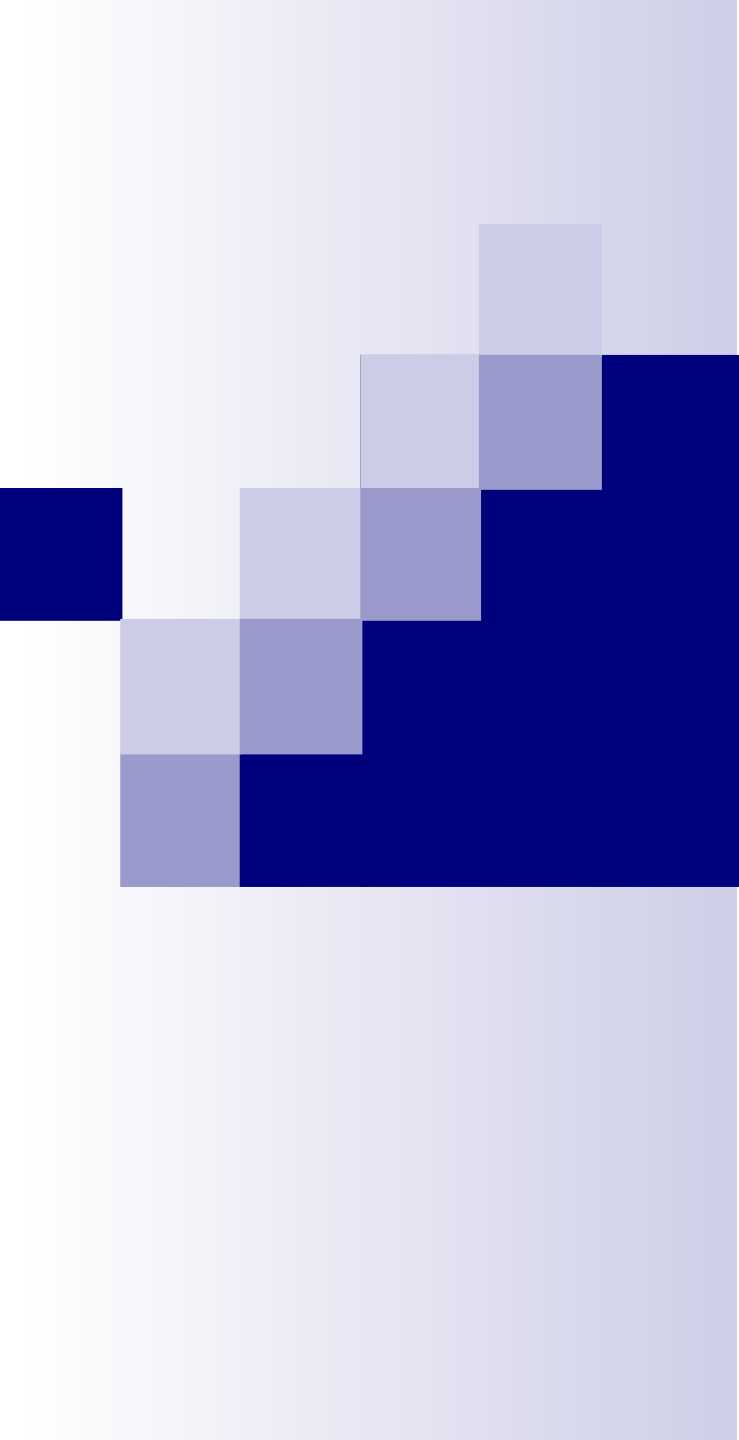
Convertisseur CAN à Intégration



Les Convertisseurs Analogiques Numériques

Convertisseur CAN utilisant un CNA





Compression Différentielle



PCM (Pulse Code Modulation)

C'est la quantification brute

PCM (ou MIC, Modulation par Impulsions et Codage) utilisé par le réseau numérique à intégration de services (RNIS ou ISDN, Integrated Services Digital Network).

Un échantillonnage préalable

Une quantification non uniforme privilégiant les amplitudes faibles

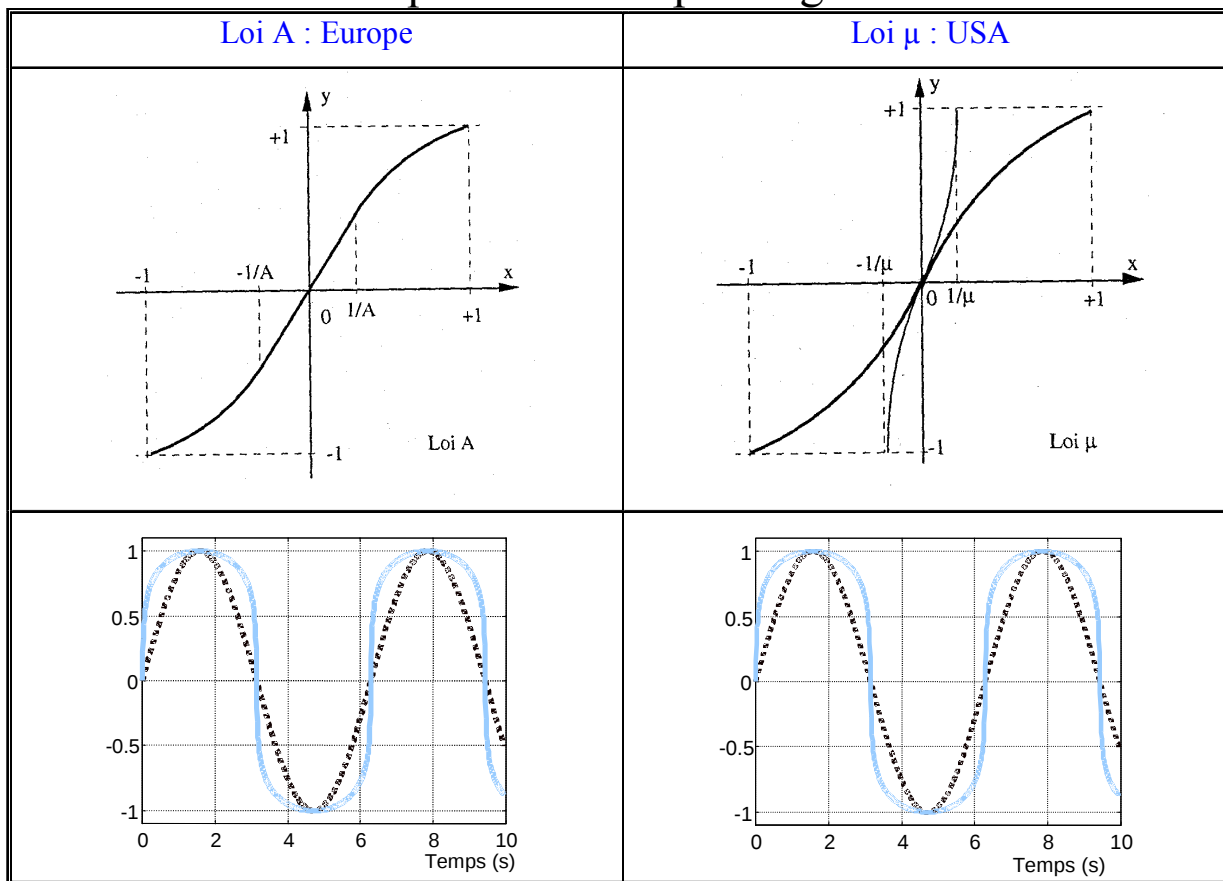
Permet d'avoir un signal téléphonique sur 8 bits avec un S/N équivalent à une quantification sur 12 bits

Norme internationale **G.711** (dépassée format *.au de Sun)

Pas de quantification non linéaire

On construit un codage qui assurera :

- Une quantification plus fine des échelons de faible niveau
- Une quantification plus “grossière” des échelons de fort niveau

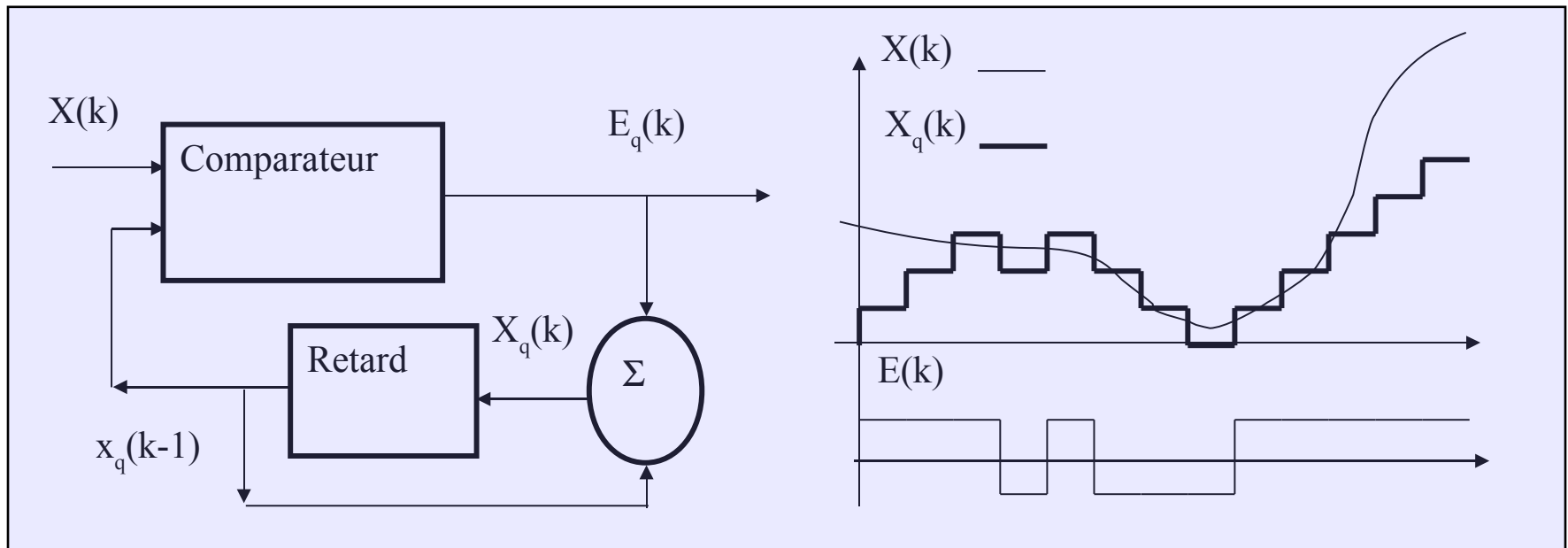


One bit PCM » MIC Delta (Delta Modulation)

Reconstituer un signal analogique quantifié $X_q(k)$ soit en ajoutant soit en retranchant une quantité fixe Δ à la valeur précédente $X_q(k-1)$, qui soit le plus près possible du signal $X(k)$ à transmettre.

Le signal transmis $E(k)$ est binaire : "one bit PCM".

Exemple de transmetteur :



Si Δ faible (bonne résolution), la fréquence F_s doit être très importante
(pb dans notre cas).

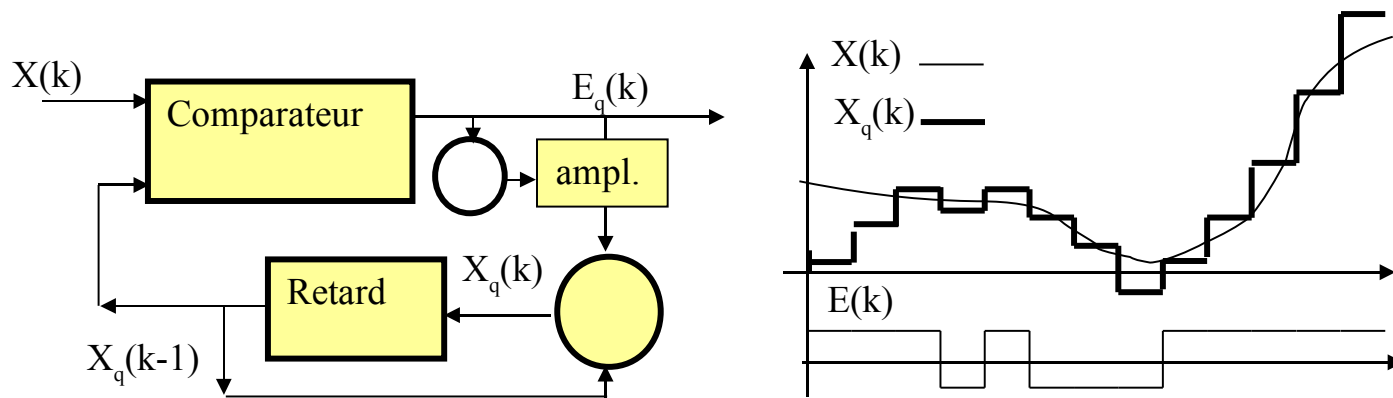
Modulation Delta adaptative

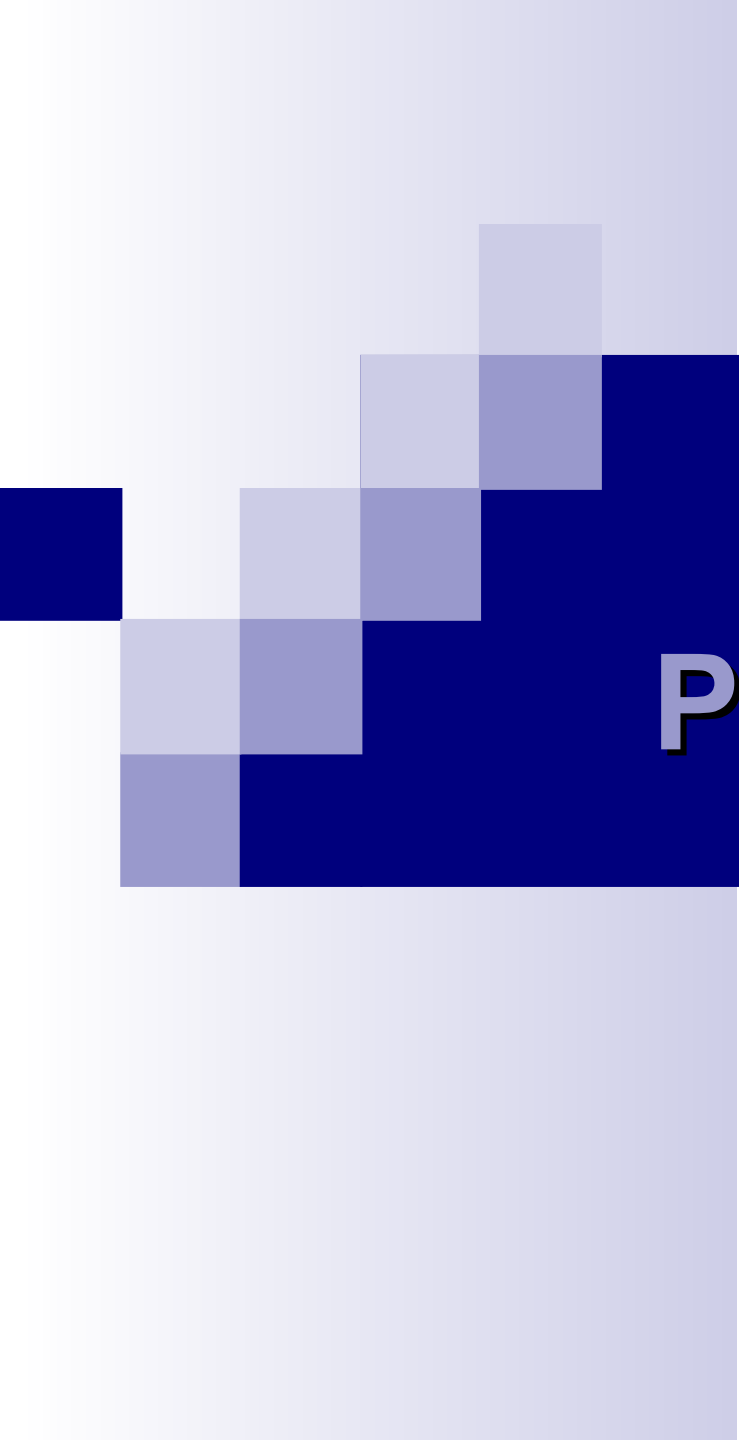
Objectif : réduit les effets de dépassement de pente sans augmenter le bruit de quantification.

La correction à ajouter ou retrancher à la valeur précédente est multipliée ou divisée par un coefficient selon que la correction change de sens ou non.

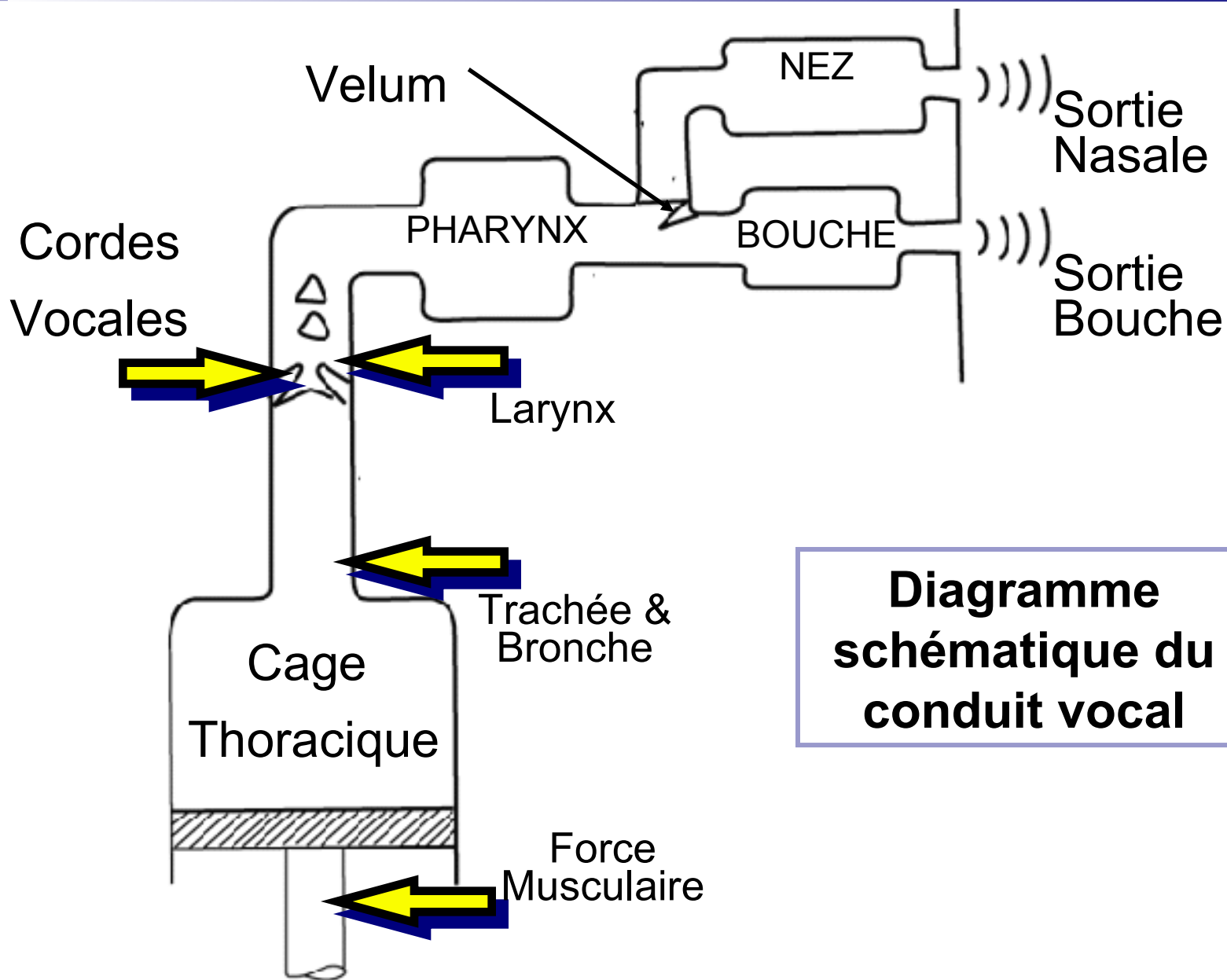
Le rapport signal à bruit du codage ADM est typiquement amélioré de 8 à 14 dB, et l'on obtient une meilleure dynamique (écart entre signaux faibles et forts).

La transmission de la voix peut utiliser un échantillonnage 6 à 8 fois seulement supérieur à la fréquence de shanon, est donc utiliser un canal de largeur 24-32 kHz.





Appareil Phonatoire humain

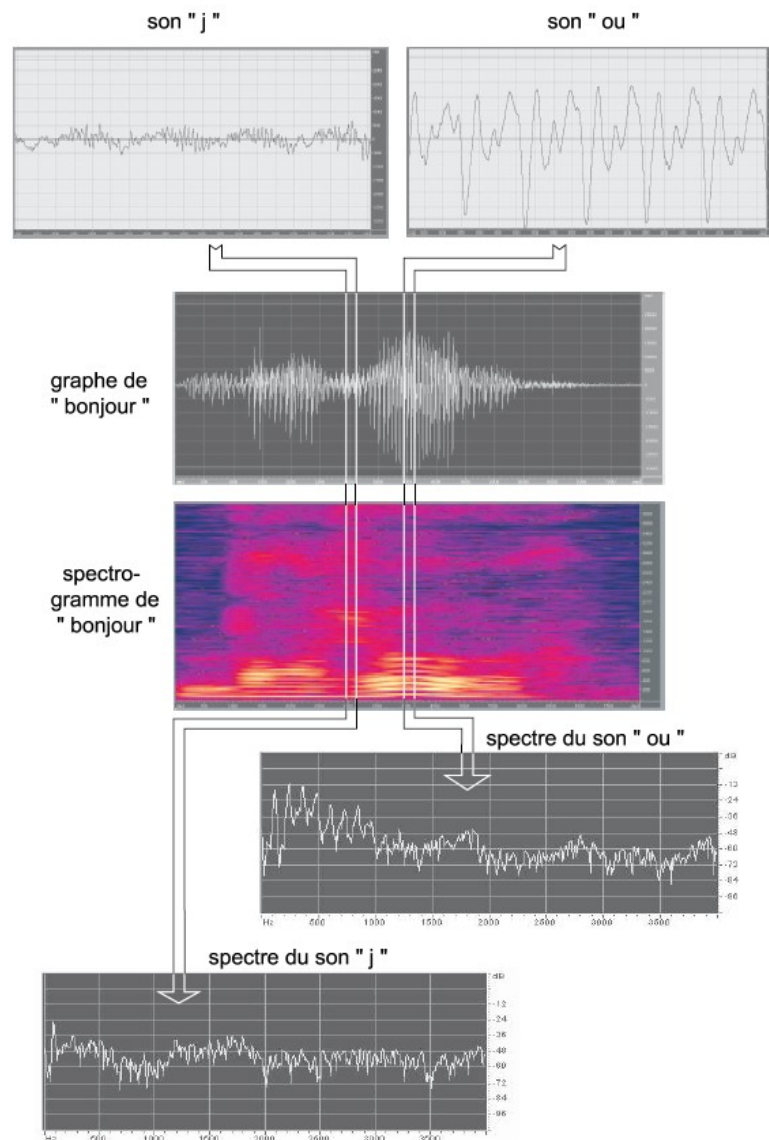


Sons Voisés et non Voisés

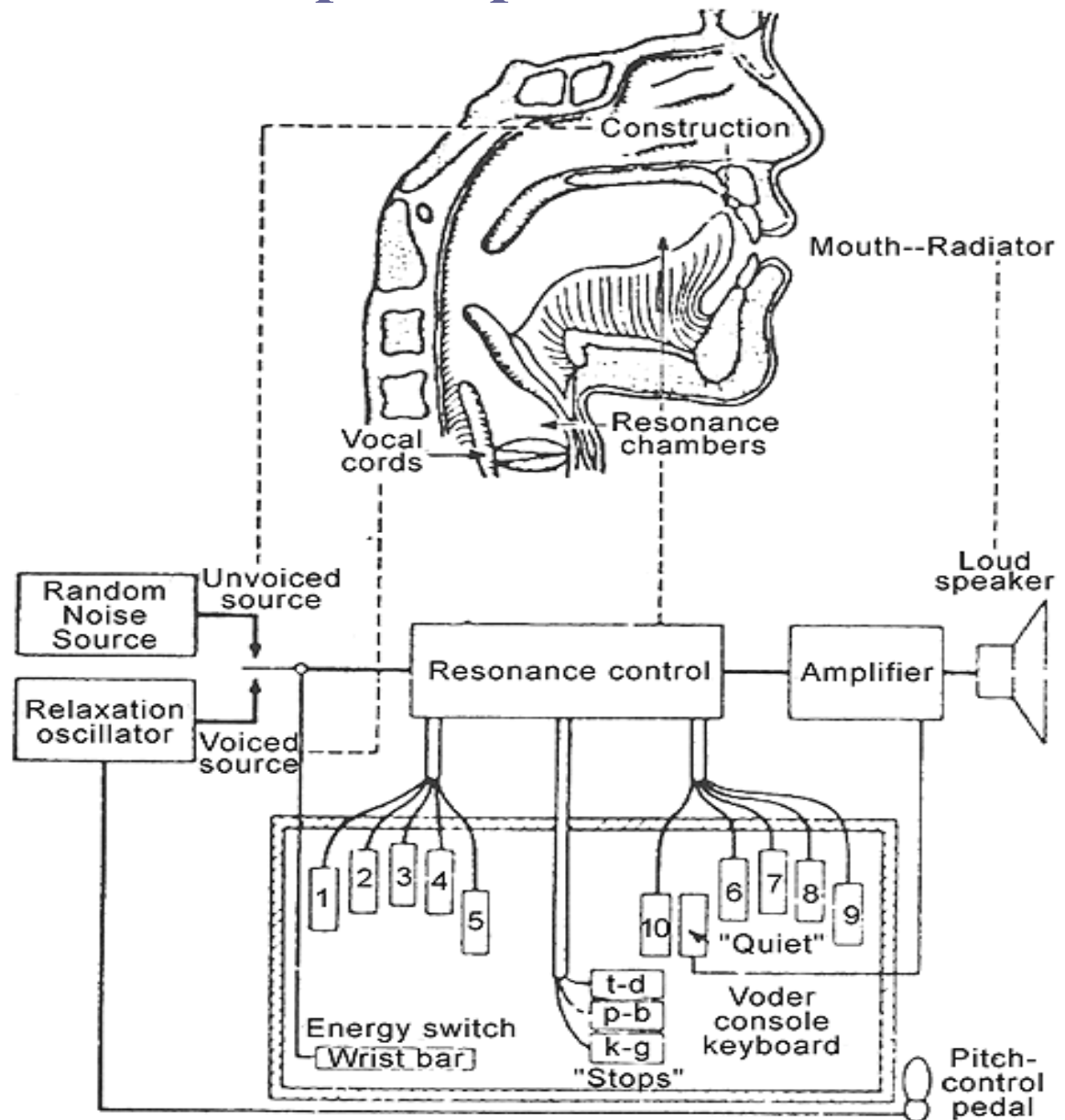
Voisés : contenu périodique marquée
fréquence fondamentale : pitch

Homme : 40Hz à 250Hz

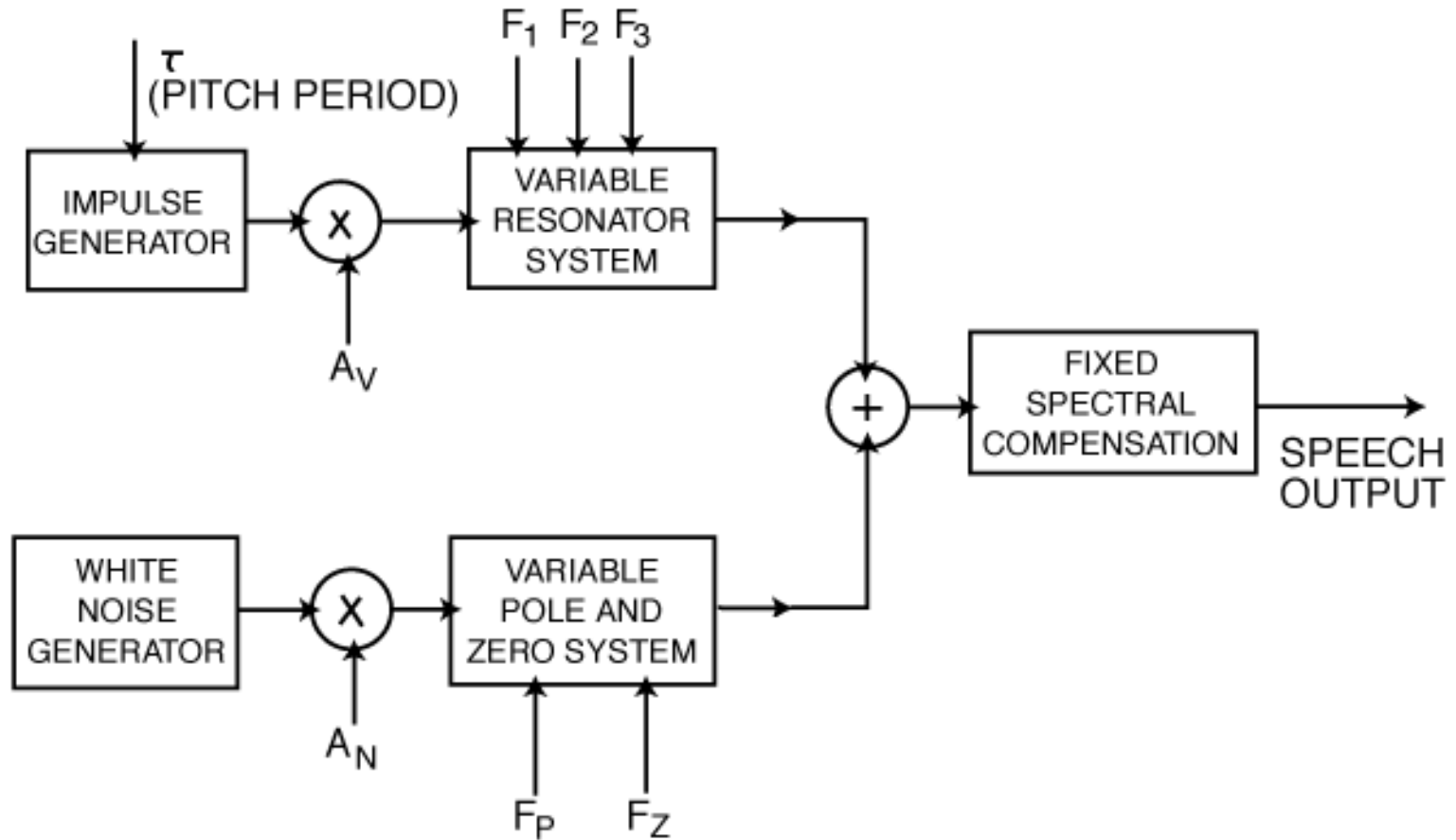
Femme : 150Hz à 750Hz

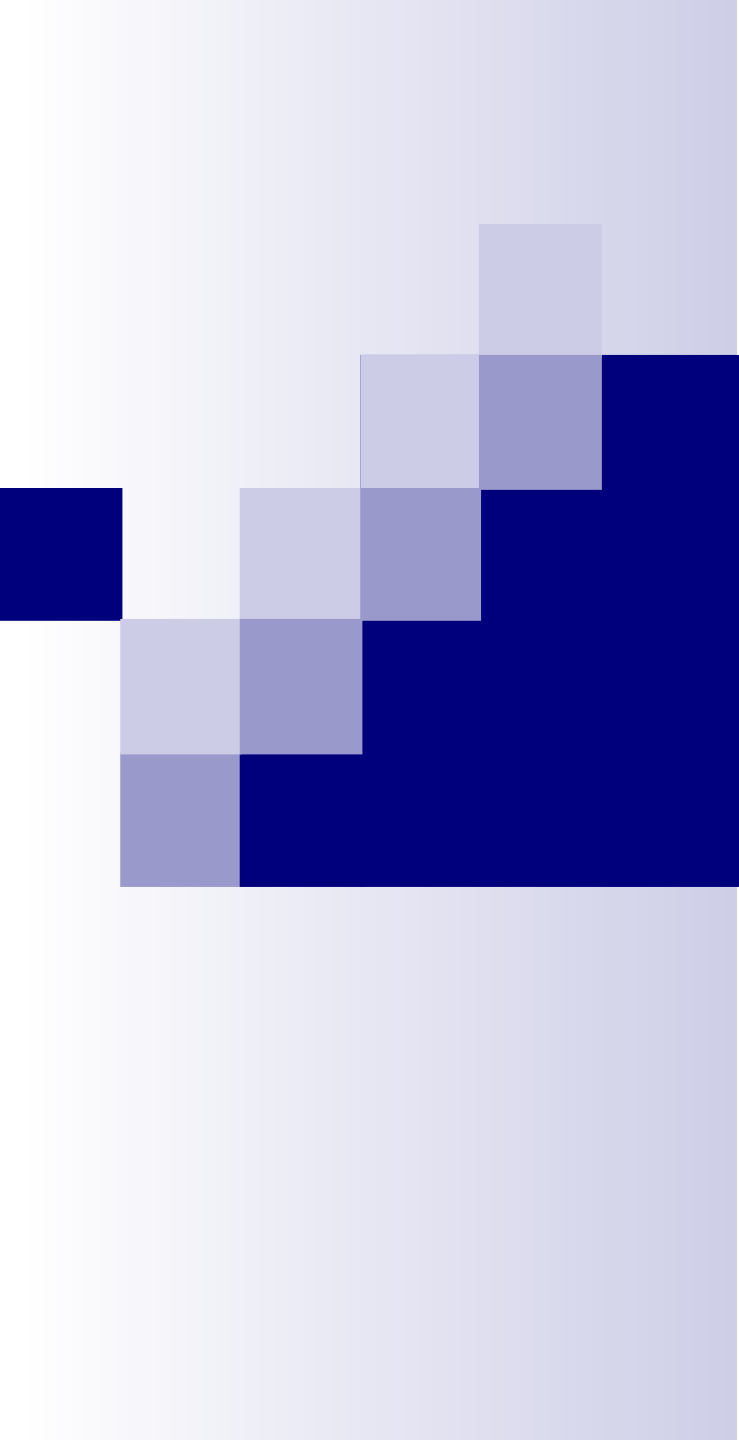


Génération de la Parole : principe



Génération de la Parole : Modèle

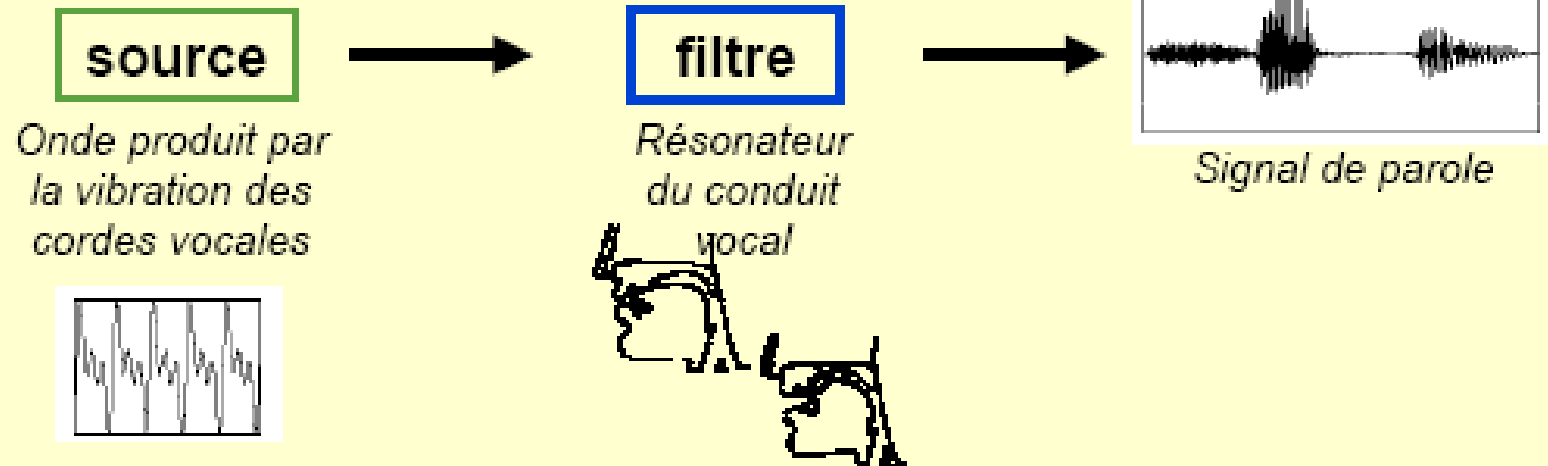




Compression LPC

Codage linéaire prédictif pour la parole (LPC)

Modèle « source-filtre » :



Notion de formants :

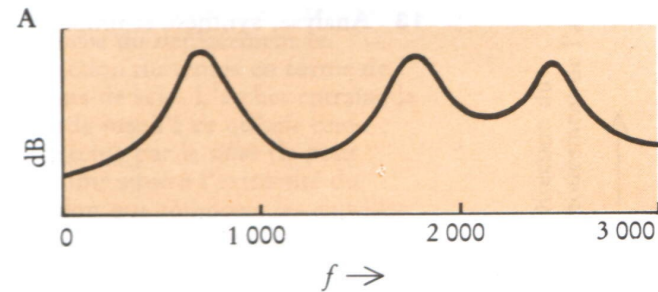
Reconnaissance de la parole:
reconnaissance « des formants »

Exemple :

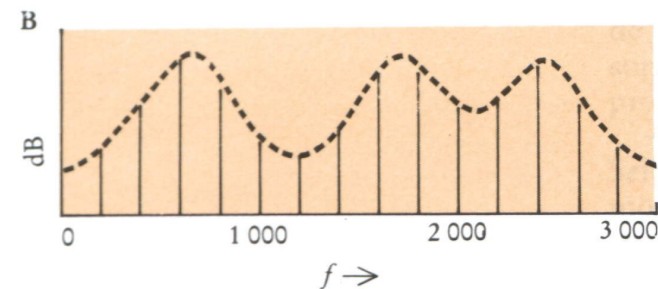
Chuchoté : beaucoup de bruit : spectre continu

Voix grave : le spectre de raies
(harmoniques du son fondamental)

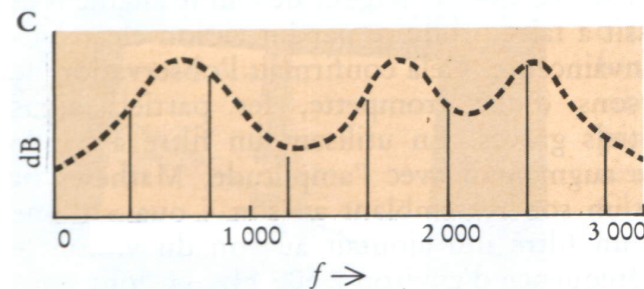
Voix aigüe : harmoniques sont plus
écartés



Chuchoté



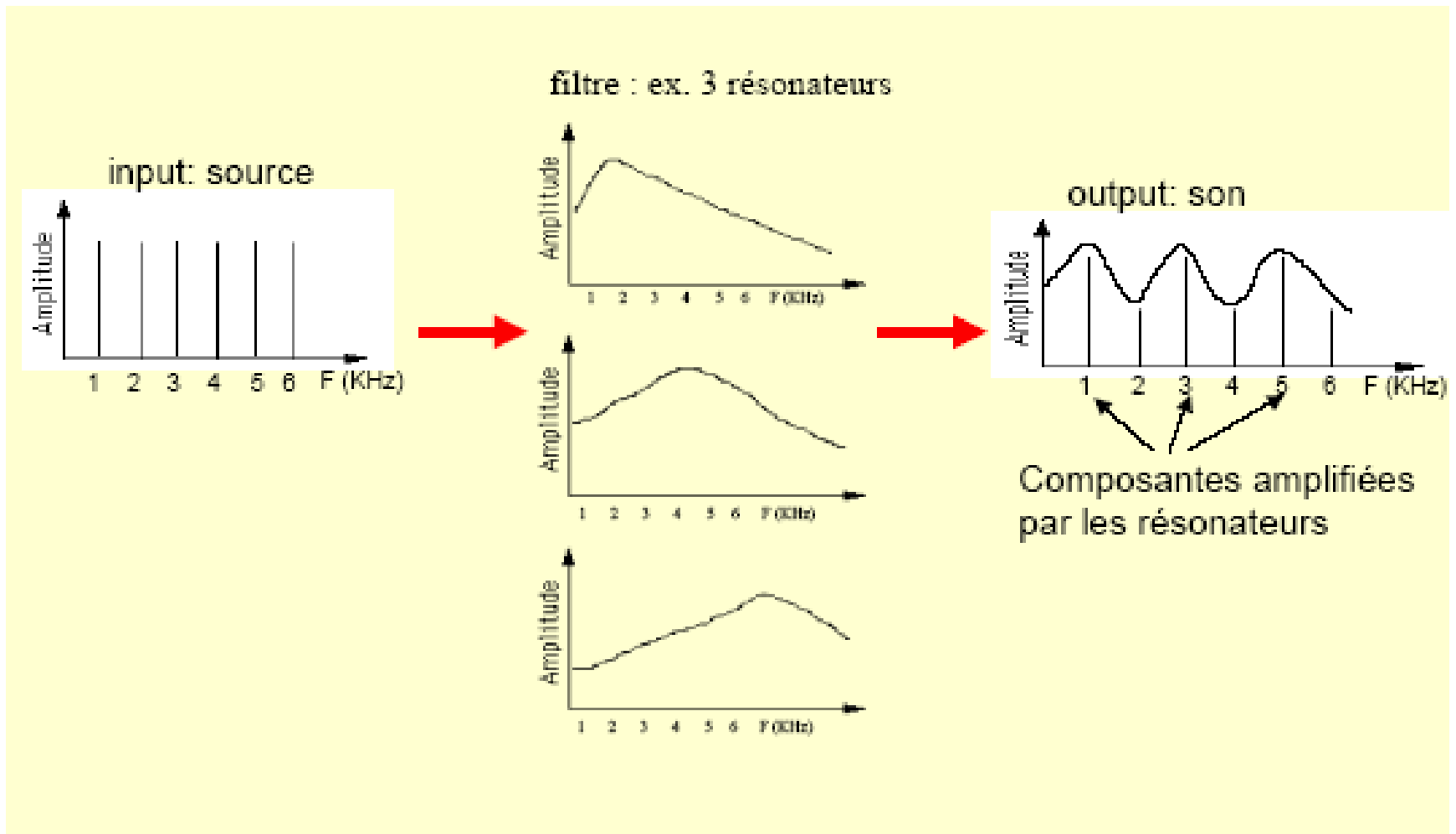
200 Hz



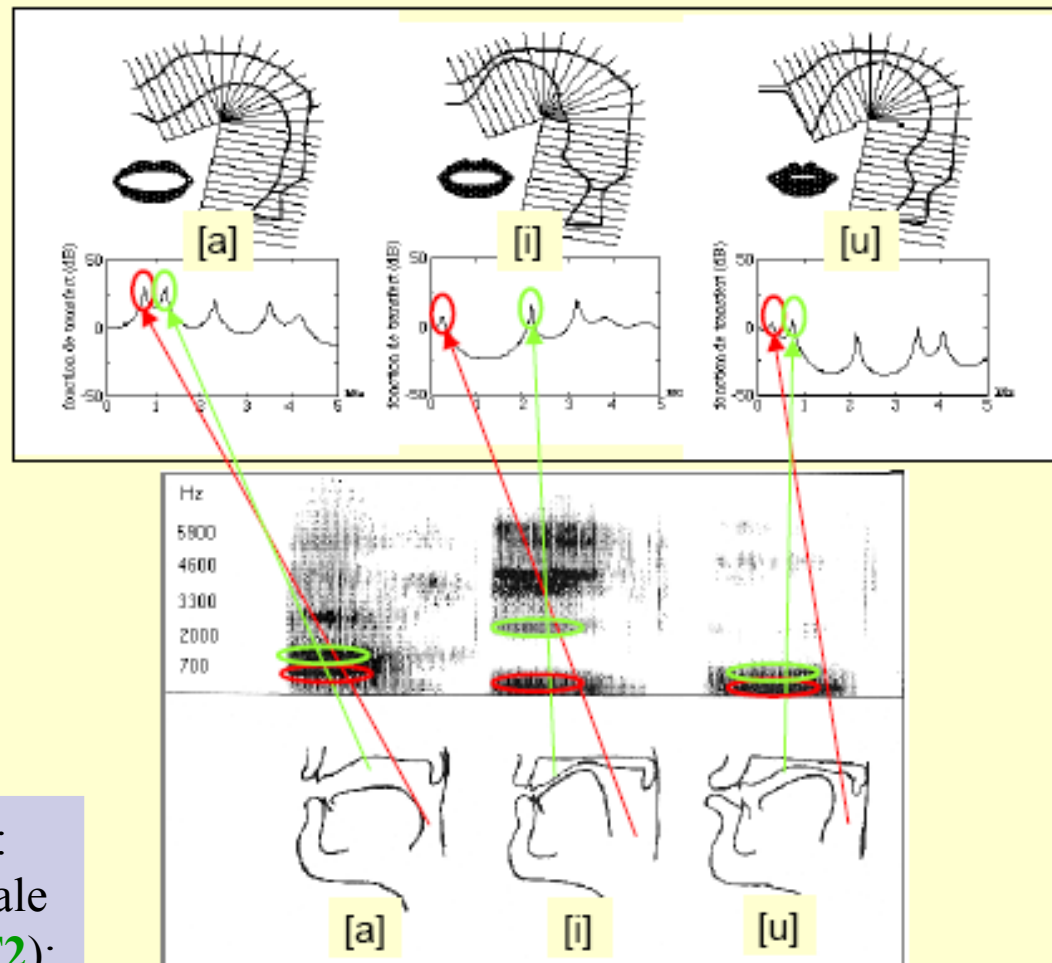
400 Hz

Chaque voyelle a entre trois et cinq formants pour se distinguer.

Notion de formants (II):



Notion de formants (III) :



/a/ /i/ /u/

Fréquence du 1^{er} formant (**F1**):

– dépend de la cavité pharyngale

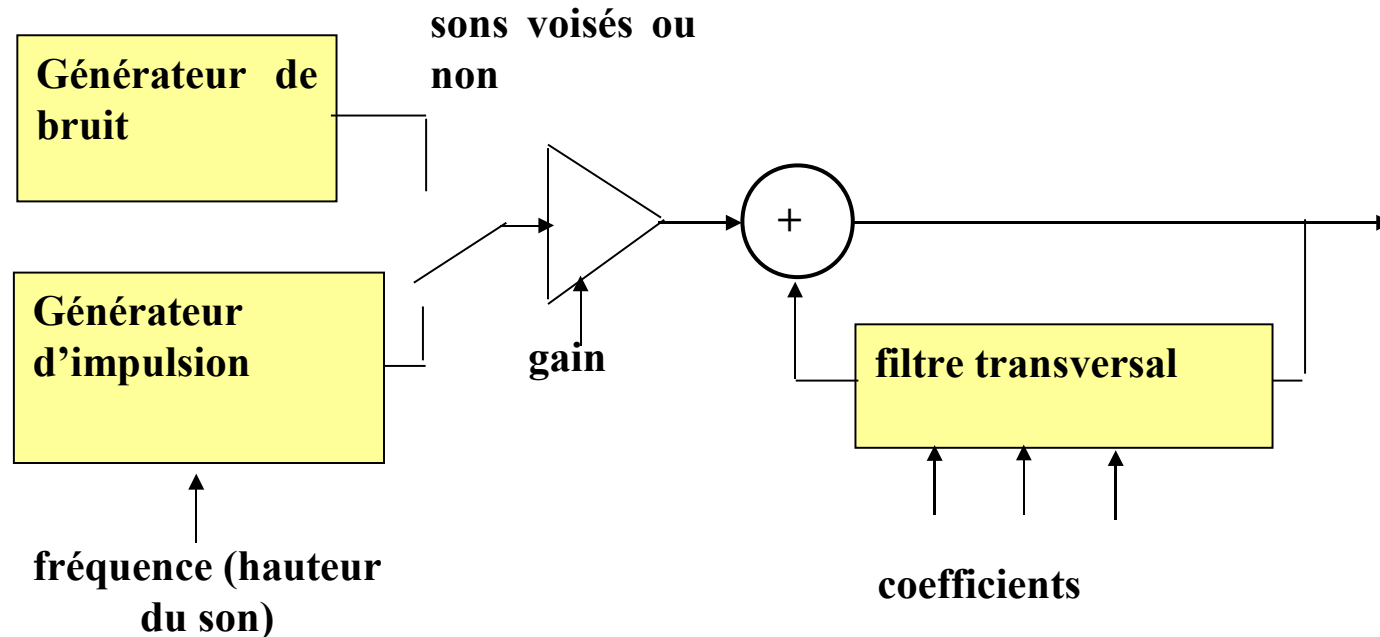
• Fréquence du 2^{ème} formant (**F2**):

– dépend de la cavité buccale

• Fréquence du 3^{ème} formant (**F3**):

– dépend de la position des lèvres

Codage linéaire prédictif pour la parole (LPC)



Adapté à la compression de la voie

Basé sur la modélisation de l'appareil phonatoire

**on arrive à reconstituer la voie avec une dizaine de coefficients du filtre,
et un échantillonnage toutes les 20 msec**

LPC principe (I)

Modèle source Filtre, La parole $S(z)$ modélisée par :

$$\begin{cases} S(z) = P(z).H(z) & \text{parole voisée } P(z) \text{ train périodique d'impulsions} \\ S(z) = N(z).H(z) & \text{parole non voisée } N(z) \text{ bruit blanc} \end{cases}$$

Soit $S(z) = G.E(z).H(z)$ avec $\begin{cases} G & \text{gain} \\ E(z) & \text{source à spectre plat} \\ H(z) & \text{filtre de synthèse} \end{cases}$

Ou $S(z) = G.E(z)/A(z)$ avec $A(z) = \sum_{i=1}^M a_i z^{-i}$ Filtre d'analyse

Prédiction linéaire $\Rightarrow$ Corrélation entre échantillons adjacents de la parole
 $\Rightarrow$ connaissance de p échantillons jusqu'à l'instant $n-1$ permet de prédire l'échantillon suivant :

$$s_n \approx \hat{s}_n = \alpha_1 s_{n-1} + \dots + \alpha_p s_{n-p} = \sum_{i=1}^p \alpha_i s_{n-i}$$

$$\hat{S}(z) = S(z).(\alpha_1 z^{-1} + \dots + \alpha_p z^{-p}) = S(z). \sum_{i=1}^p \alpha_i z^{-i} = S(z).F(z)$$

LPC principe (II)

Donc erreur de prédiction ϵ_n entre prédiction et signal véritable

$$\epsilon_n = s_n - \hat{s}_n = s_n - \left(\sum_{i=1}^p \alpha_i s_{n-i} \right) \quad \text{ou} \quad E(z) = S(z) - \hat{S}(z) = S(z) \cdot \left(1 - \sum_{i=1}^p \alpha_i z^{-i} \right)$$

Prédiction linéaire $\approx$ modèle acoustique linéaire de production $\Rightarrow$ Identification

$\Rightarrow$ erreur résiduelle ϵ_n = source d'excitation

▮ filtre inverse $A(z)$ associé au filtre prédicteur (en prenant $M=p$)

$$\epsilon_n + \sum_{i=1}^p \alpha_i s_{n-i} = Ge(n) - \sum_{i=1}^p a_i s_{n-i}$$

Identification de $A \Rightarrow$ résiduel à spectre plat donc excitation = ▮ bruit blanc
▮ une seule impulsion

Modélisation source en LPC soit ▮ générateur impulsion $\Rightarrow$ **voisée**
▮ bruit blanc $\Rightarrow$ **non voisée**

LPC principe (III)

Le $n^{\text{ième}}$ échantillon est défini par $x(n)$:

- une combinaison linéaire de p échantillons précédents.
- un résidu correspondant à l'erreur de prédiction $\varepsilon(n)$

$$\text{Détermination de } s(n) = \alpha s(n-1) + \alpha s(n-2) + \dots + \alpha s(n-p) + \varepsilon(n) \text{ iction}$$

Soit sur la plage temporelle n_0 n_1 (trame):

$$\varepsilon_n^2 = \left[s_n - \sum_{i=1}^p \alpha_i s_{n-i} \right]^2 \quad \text{Erreur quadratique} \quad E = \sum_{n=n_0}^{n_1} \varepsilon_n^2 \quad \text{Erreur totale}$$

k

$$\frac{\partial E}{\partial \alpha_k} = 0$$

Minimisation = on cherche les α tels que :

$$\text{Soit } 2 \sum_{n=n_0}^{n_1} s_{n-k} \left[s_n - \sum_{i=1}^p \alpha_i s_{n-i} \right] = 0$$

LPC principe (IV)

Détermination des coefficients de prédiction (2)

Donc système à résoudre :
$$\sum_{n=n_0}^{n_1} s_{n-k} s_n = \sum_{i=1}^p (\alpha_i \sum_{n=n_0}^{n_1} s_{n-k} s_{n-i}) \quad 1 \leq k \leq p$$

Qui donne par changement de variable

$$c_{k_0} = \sum_{i=1}^p \alpha_i c_{k_i} \quad 1 \leq k \leq p \quad \text{avec} \quad c_{k_i} = \sum_{n=n_0}^{n_1} s_{n-k} s_{n-i}$$

Plusieurs méthodes de résolution possibles classiquement : **Autocorrelation**

Car si on prends une plage infini pour l'erreur total

$$c_{k_i} = \sum_{n=-\infty}^{+\infty} s_{n-k} s_{n-i}$$

Plusieurs algorithmes

Par exemple approche récursive : N. Levinson (1947) modifié par J. Durbin (1959)

LPC principe (V)

Détermination des coefficients de prédiction (3)

Équation aux différences $\Rightarrow$ un échantillon ,reconstitué d'après les échantillons précédents.

Ses coefficients (**formants**) ajustés pour minimiser l'écart (quadratique moyen) entre le signal prédit et le signal réel.

$\Rightarrow$ Résolution d'un système d'équations linéaires : plusieurs méthodes possibles.

Remarques :

➤ Pour les sons nasaux, Modèle plus un simple tube

le nez = branche latérale $\Rightarrow$ des zéros, mathématiquement parlant, et rend

algorithmes plus complexes les algorithmes.

problème souvent négligé, est délégué au niveau du résidu.

➤ Certaines positions de la langue conduisent aussi à une prise en compte par le seul résidu,

$\Rightarrow$ un nombre important de bits ! !

LPC principe (VI)

filtre transversal + amplificateur ajustés pour imiter les filtres des cordes vocales.

Le filtre est mathématiquement une combinaison linéaire des échantillons successifs
(Comme le prédicteur linéaire précédent)

Un codeur transmet :

- la fréquence (6bits, 0=bruit)
- le gain de l'amplificateur (6 bits)
- les valeurs des coefficients (6 bits x 10)
- 8 bits pour la correction à apporter à la synthèse.



80 bits toutes les 20 msec.

Un canal de transmission d'un débit entre 3 kbps et 8 kbps peut être suffisant.

Voix « robot » à 2.4 kbps

Codage prédictif de la norme G.S.M.06.10 (I)

Algorithme **RPE-LTP (regular pulse excitation- long term prediction)**

- construit des trames de 260 bits à partir de 160 échantillons PCM à 13 bits, à 8 kHz.
- Une seconde ne nécessite que 1625 octets, et un 1Mo suffit pour 10 mn.
- Une trame couvre donc 20 ms (160 éch.) : une période pour une voix très grave, et à 10 pour une voix très aiguë.

Deux filtres sont utilisés :

- L'un fonctionne à court terme (« short term prediction »), reconstitue le rôle des cordes vocales et autres cavités résonnantes humaines.
- L'autre filtre excite le précédent et reconstitue un mélange d'ondes et de bruit par « prédiction à long terme ».

Codage prédictif de la norme G.S.M.06.10 (II)

Comparaison des codages pour la voix

Méthode	taux d'éch.(kHz)	bits/ech.	Débit (kbps)	
DM		64-128	1	64-128
PCM		8	7-8	56-64
ADM		48-64	1	48-64
DPCM		8	4-6	32-48
ADPCM		8	3-4	24-32
LPC	0.04-0.1	~80	2-8	
CELP				4.8 (conference)
GSM	0.05	~260	13 (mobile)	

Exemple de son GSM



Son (OLE2)

8kbps



Musique

44kbps



Parole

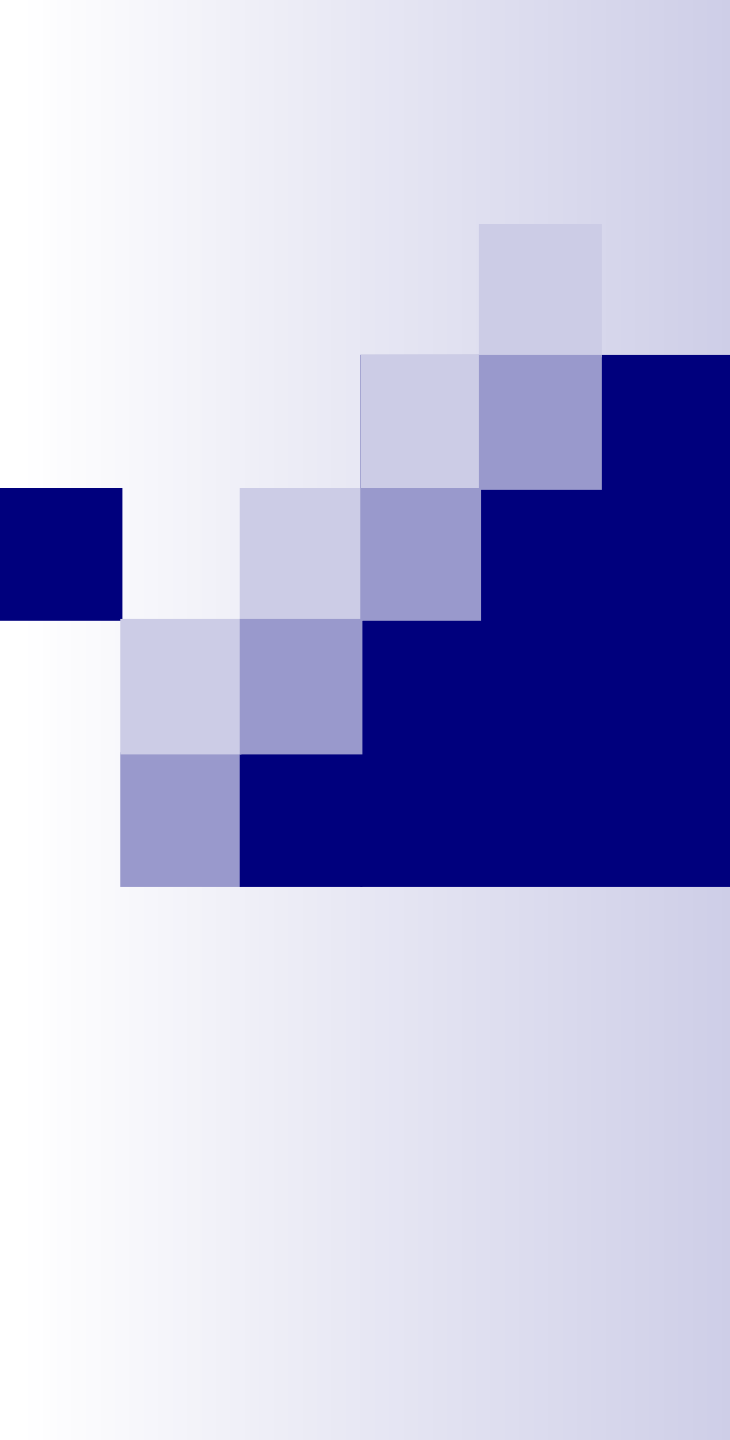
Codage CELP

Réduire encore le débit

⇒ pour coder le résidu : codage **CELP (Code Excited Linear Prediction)**, n

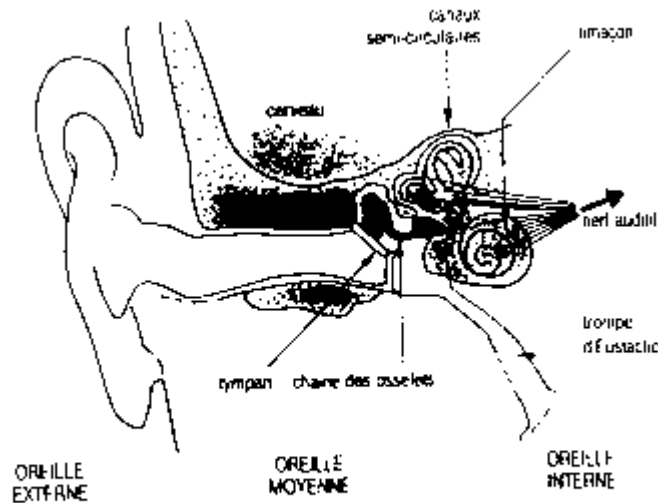
utilise un « livre de codes » plutôt qu'un générateur d'impulsions.

- L'analyse du résidu essaie de trouver à chaque instant, le résidu type le plus proche parmi ceux proposés par le « livre de codes ».
- Le synthétiseur utilise son code pour exciter le filtre à formants.
- Le problème est que le nombre de codes doit être très important si l'on veut une qualité et intelligibilité correcte, et considérer toutes les hauteurs de voix.
- Les concepteurs ne fixent que quelques codes pour une seule hauteur de voix et un utilise un autre « livre de codes », adaptatif, vide au départ, qui se remplit durant le fonctionnement du système

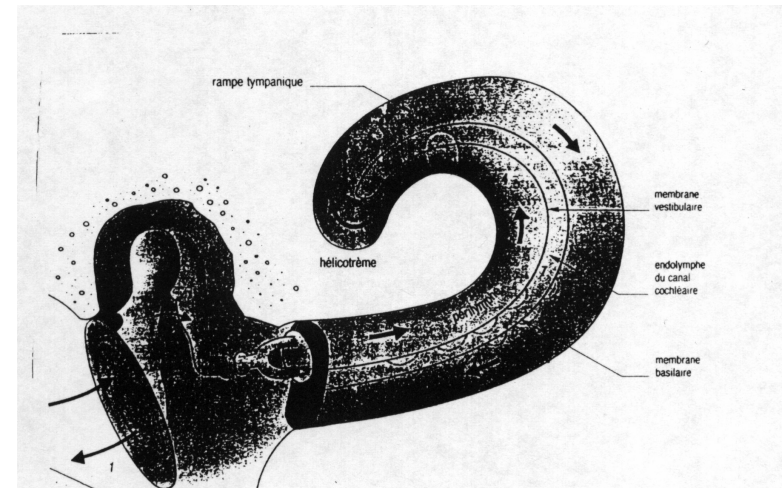


Appareil auditif humain

Description de l'appareil auditif humain



Mécanisme de la transmission



entre 700 Hz et 1.4 kHz pour les osselets
environs 3 kHz pour le conduit auditif.

Heureusement ces résonances sont peu marquées.

la transmission des sons à partir de la cochlée est excellente entre **600 et 6 kHz**
mais mauvaise en dessous et au dessus de ces limites.

Notions de perception auditive

Notion de sonie subjective

la sonie, c'est-à-dire la sensation (subjective) d'intensité sonore, est proportionnelle au logarithme de l'excitation

$$S = K \cdot \log(I) \quad \text{loi dite de WEBER-FECHNER}$$

Échelle des dB acoustiques

référence la pression acoustique p_0
correspondant au seuil de perception d'un son pur à 1000 Hz

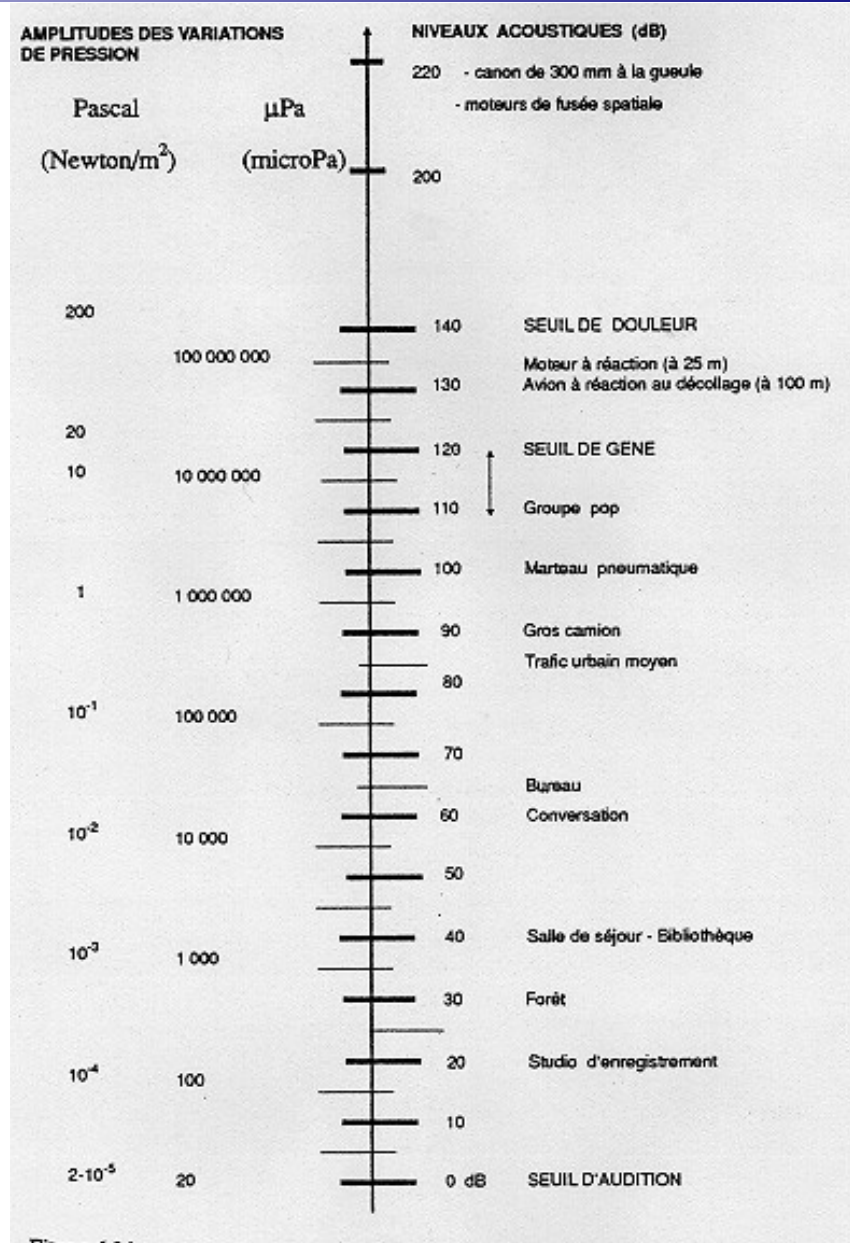
$p = 2 \cdot 10^{-5} \text{ Pa}$ et $I = 10^{-12} \text{ W/m}^2$

Échelle de mesure des niveaux de pressions ou d'intensité acoustique :

$$L_{\text{dB}} = 20 \cdot \log\left(\frac{p}{p_0}\right) = 10 \cdot \log\left(\frac{I}{I_0}\right)$$

La sensibilité différentielle d'intensité est de l'ordre de 0.5 dB.

Niveaux acoustique



Données pratiques sur l'oreille

Perception de la hauteur (Pitch)

La hauteur tonale H (grandeur de "sensation", subjective)
est proportionnelle au logarithme de la fréquence

$$H_1 - H_2 = k \cdot \log\left(\frac{f_1}{f_2}\right)$$

Échelle des octaves

Correspond à un doublement de la fréquence

La gamme musicale : une division de cet intervalle en 12 demi-tons égaux.
le rapport des fréquences correspondant étant de

$$\sqrt[12]{2} = 1.059$$

Quelques chiffres

- Le rapport extrême des énergies normalement audibles ("dynamique" de l'oreille)
- Le rapport extrême des fréquences audibles

13

10 OCTAVES, ou encore 10 300 HZ

- Seuil absolu de perception sonore

-12

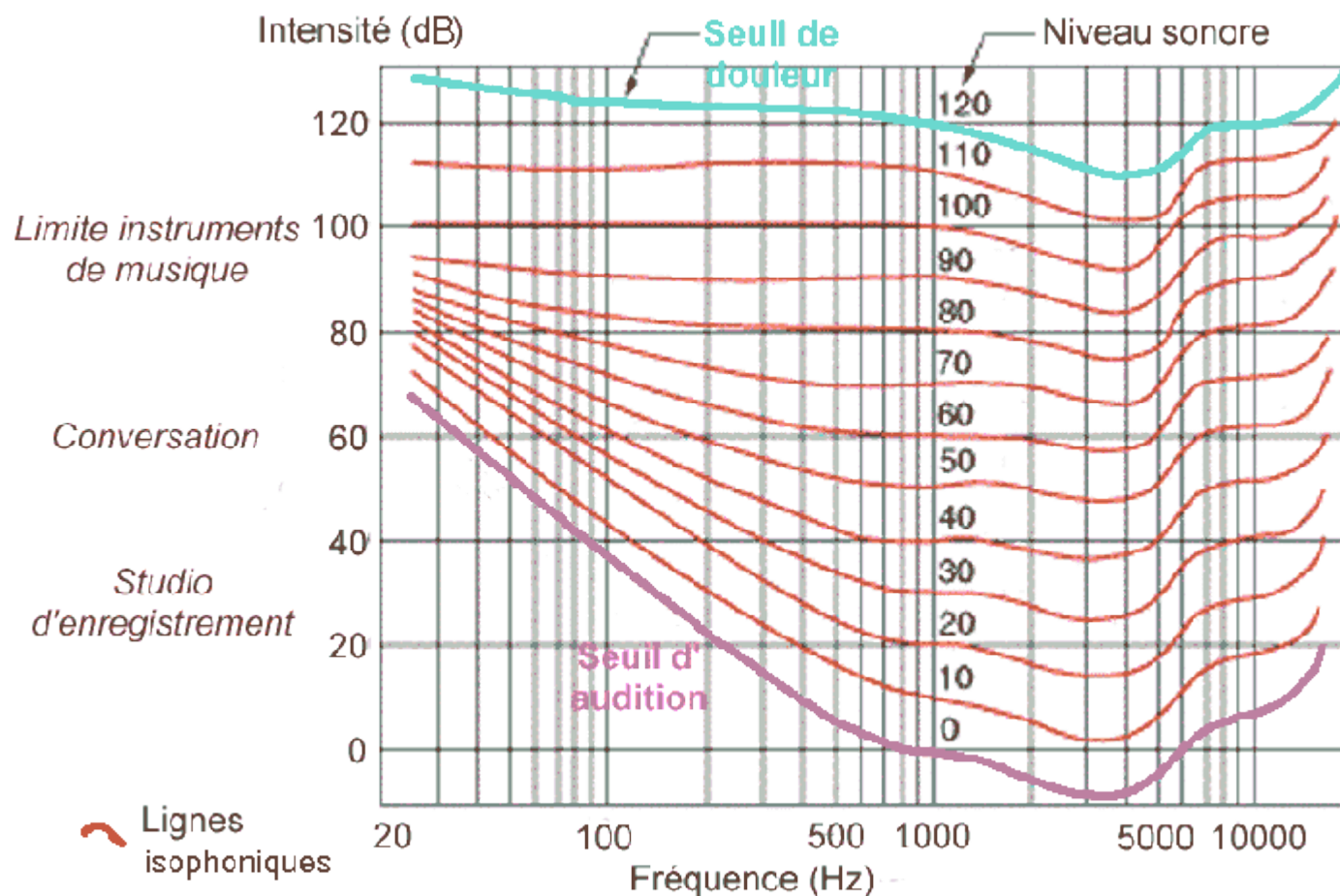
2

bruit d'agitation thermique des molécules d'air est de l'ordre de 10^{-14} W/m^2

10⁻¹² W/m²

Données pratiques sur l'oreille

Notion de sonie subjective





Compression

MPEG II Layer 3

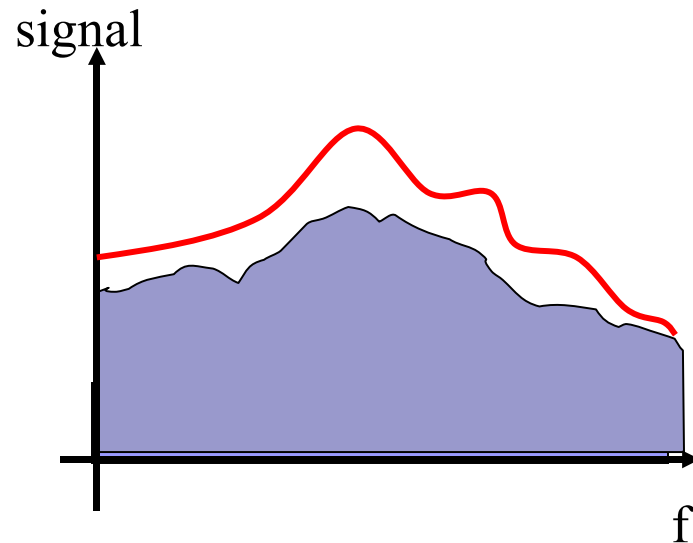
Objectifs

La numérisation d'un signal audio conduit à des débits trop important
16 bits à $F_e=48$ kHz donne un débit de 1,5 Mbits/s en stéréo

- Assurer une qualité sonore qui soit jugée transparente
256 kbits/s en stéréo pour une qualité CD
- Ne pas faire de présupposé sur le signal audio à compresser
- Le décodeur doit être le plus simple possible

Allocation binaire dynamique par sous bande fréquentielle

Idée de base



CD: bruit de quantification constant
Quelque soit les fréquences

$$S/N_{dB} = 6,02n \text{ soit } 16 \text{ bits pour } 96 \text{ dB}$$

Mise en Forme du bruit de quantification
Selon les fréquences en fonction d'un
modèle Psycho-acoustique de l'oreille

**Idée injecter le maximum de bruit de quantification possible
mais qui reste inaudible :** définir par bande de fréquences le
nombre de bits de quantification strictement nécessaire

Idée de base

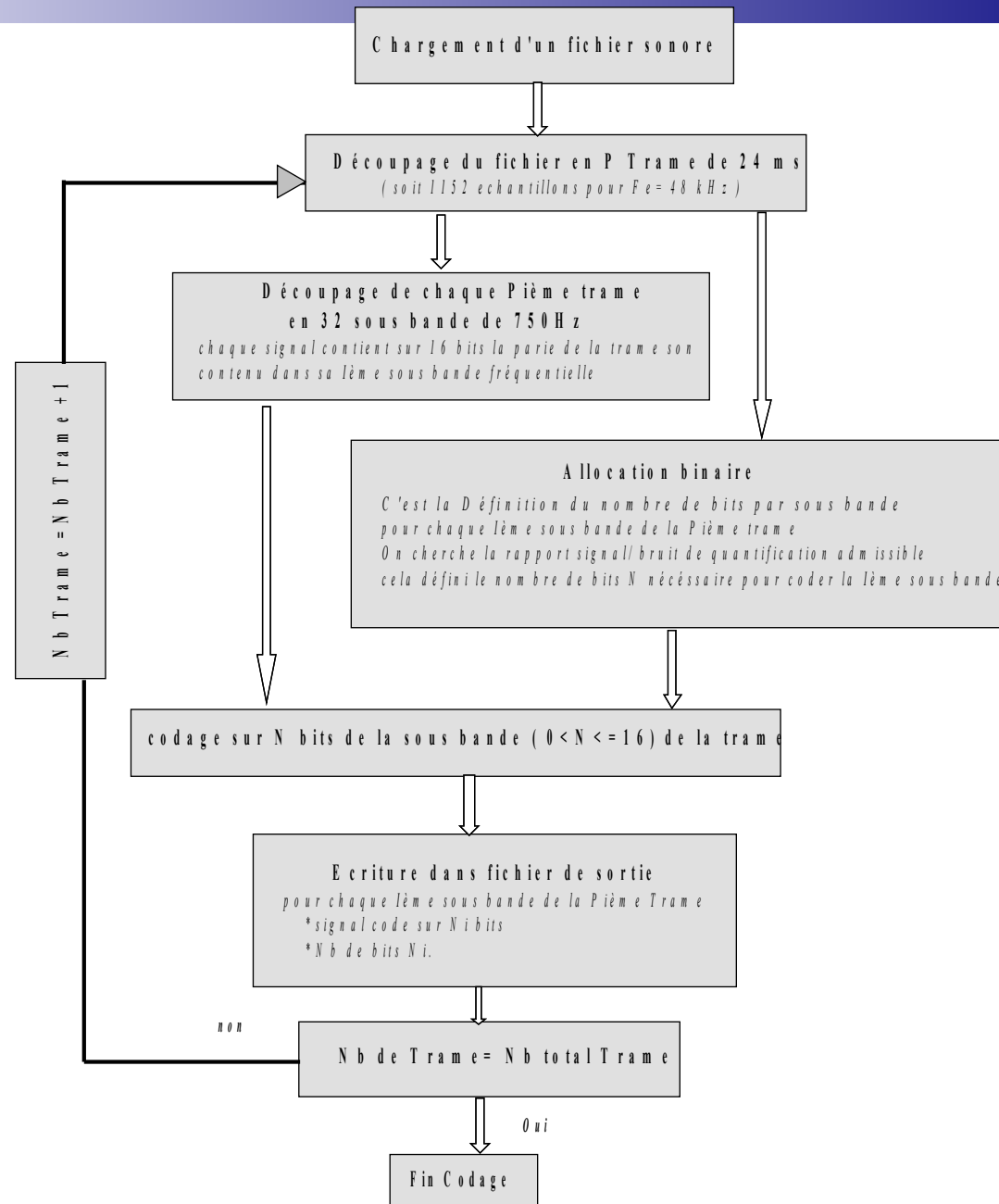
les Principes.

- ☐ Travailler par trame de son quasi-statique temporellement.
- ☐ Découper le signal en sous bandes fréquentielle.
- ☐ Allouer à chaque sous bande le nombre de bits nécessaire et suffisant
- ☐ Reconstruire simplement le signal.

Les problèmes.

- ☐ **Définir la courbe de masquage dynamique de l'oreille**
- ☐ **Découpage en sous-bandes parfait sans augmenter le nombre d'info**
- ☐ **Découpage en sous-bandes réversible**

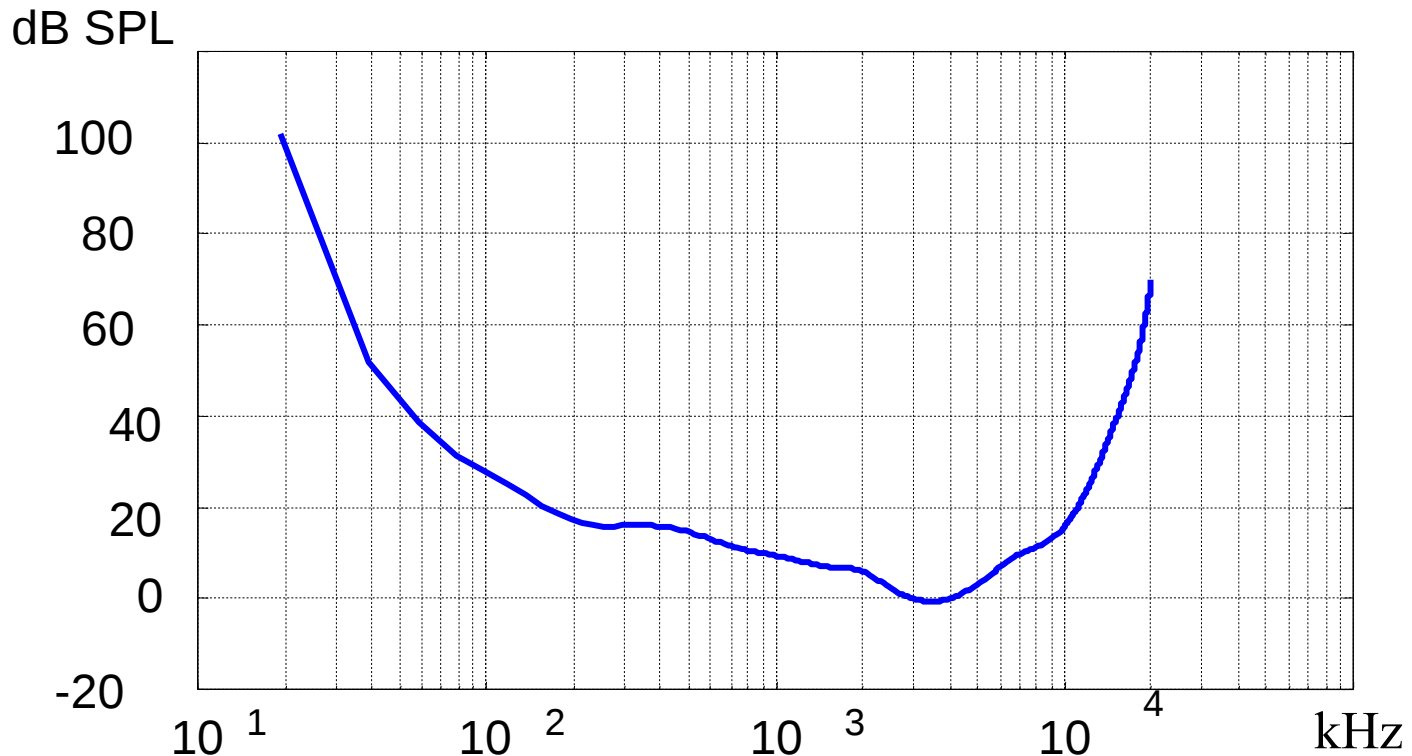
codeur





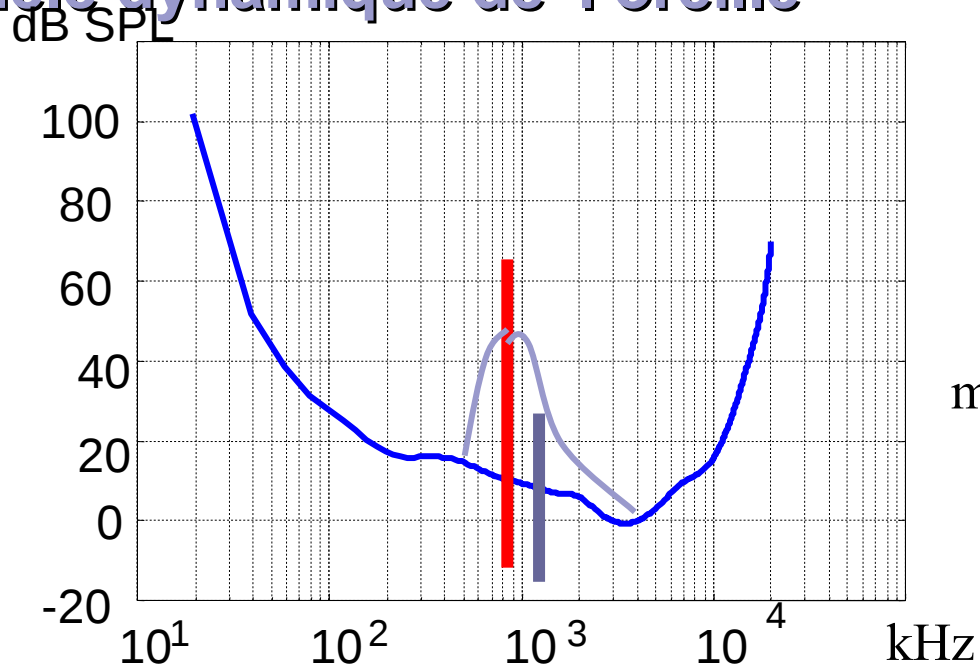
Allocation binaire

Modèle Psycho-acoustique de l'oreille



La courbe de « **Seuil absolu au repos** » est liée au bruit interne de l'oreille. Un signal présenté à l'oreille dont la puissance acoustique se situe en dessous de cette courbe n'est pas perçu

Modèle dynamique de l'oreille



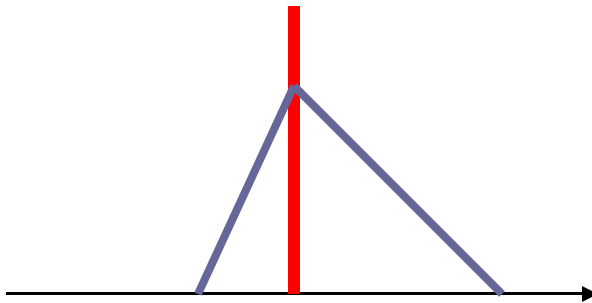
En présence d'un son pur
(**Tonale**) on obtient une
nouvelle courbe de
masquage qui masque le son
faible -----

Ainsi la courbe de masque dynamique calculée tout les 24ms
est la est le max entre les courbes de masquages des tonales et
celle du modèle psycho-acoustique de l'oreille

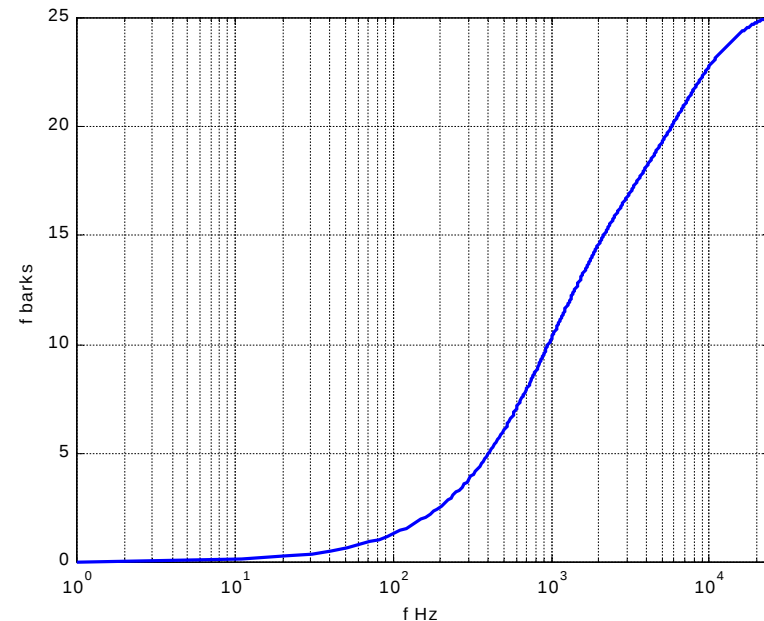
Modèle dynamique de l'oreille

La courbe de masque d'une tonale est difficilement modélisable :
Faire un changement d'échelle de fréquence pour tenir compte de la non-linéarité de l'oreille (**fréquence en Barks**)

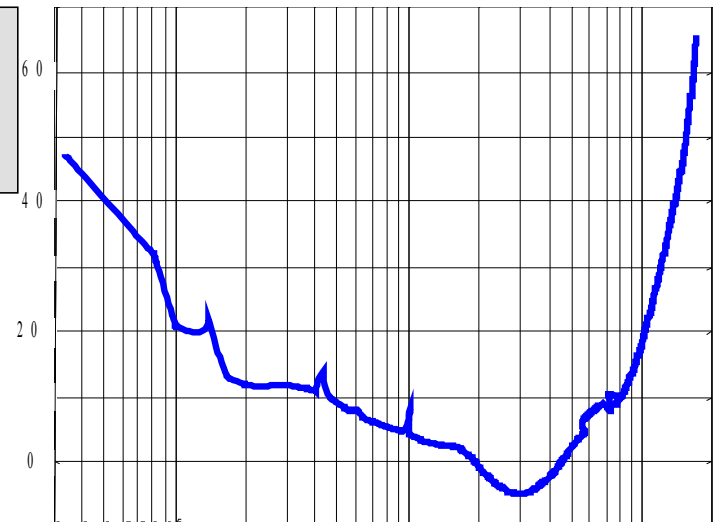
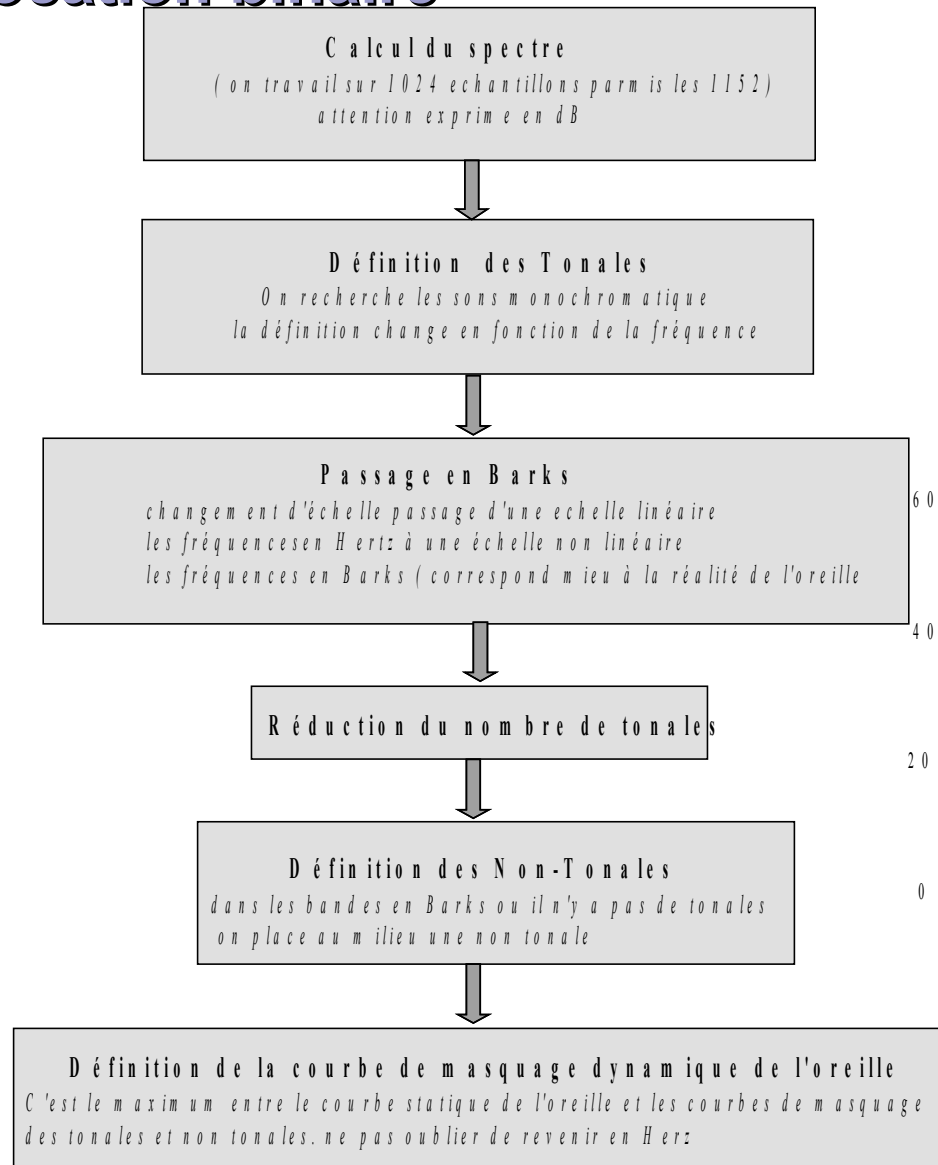
$$f_{\text{barks}} = 13 \cdot \arctan\left(\frac{f_{\text{Hz}}}{1000}\right) + 3,5 \cdot \arctan\left(\frac{f_{\text{Hz}}}{7500}\right)^2$$



On obtient une courbe de masquage affine



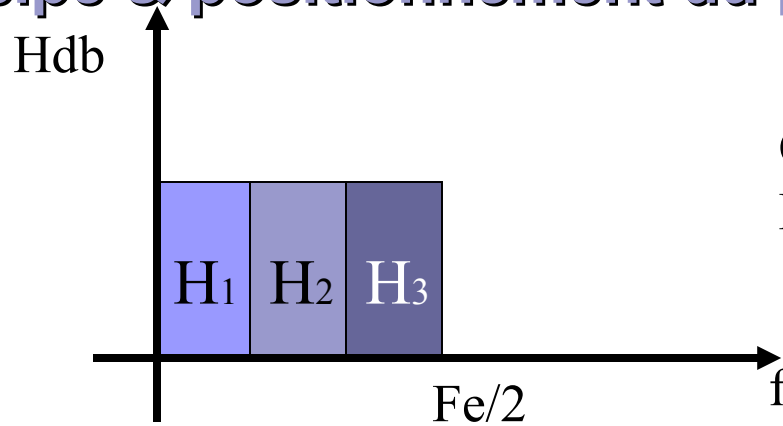
Allocation binaire



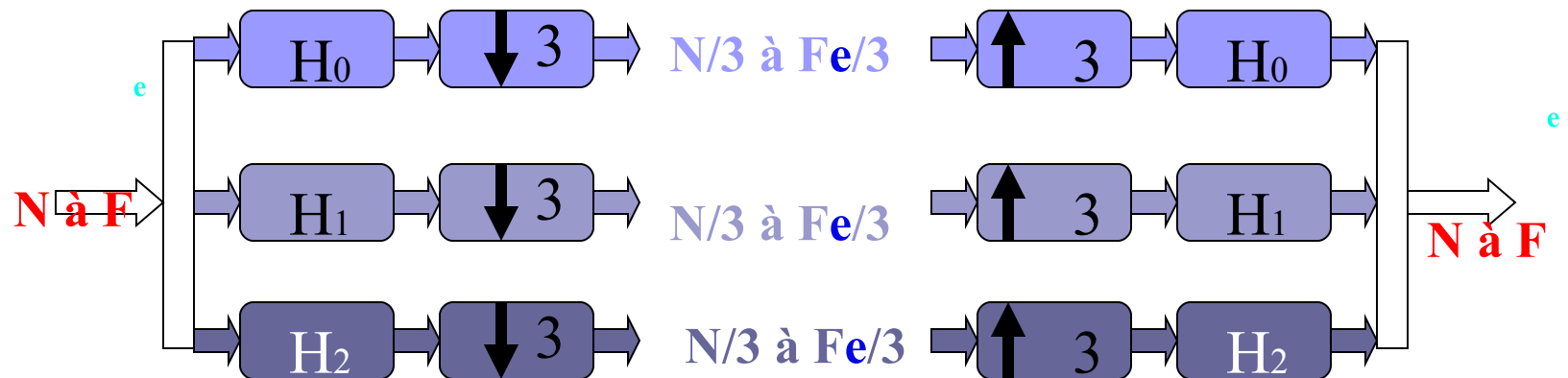


Filtrage en sous bandes

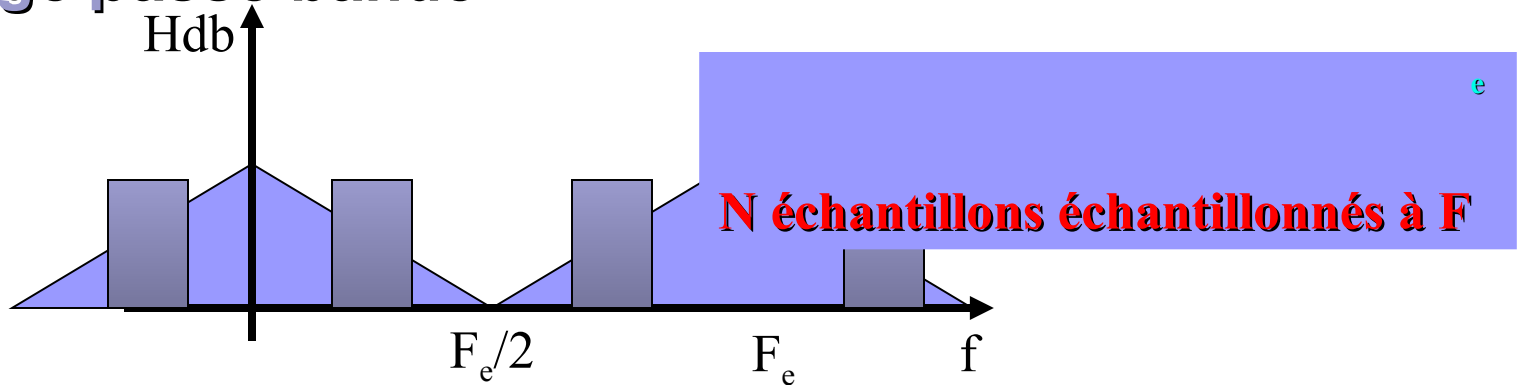
Principe & positionnement du problème



On découpe $[0, Fe/2]$ par 3 filtres FIR
 Passe Bande strictement identique

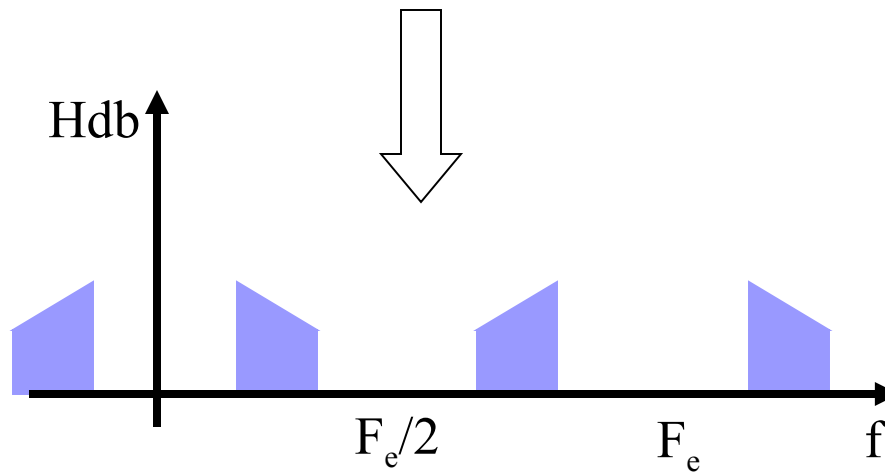


Filtrage passe bande



➤ Signal échantillonnée

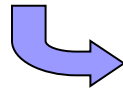
➤ Passe Bande H_1
Largeur $F_e/6$



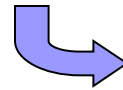
N échantillons échantillonnés à F

Décimation

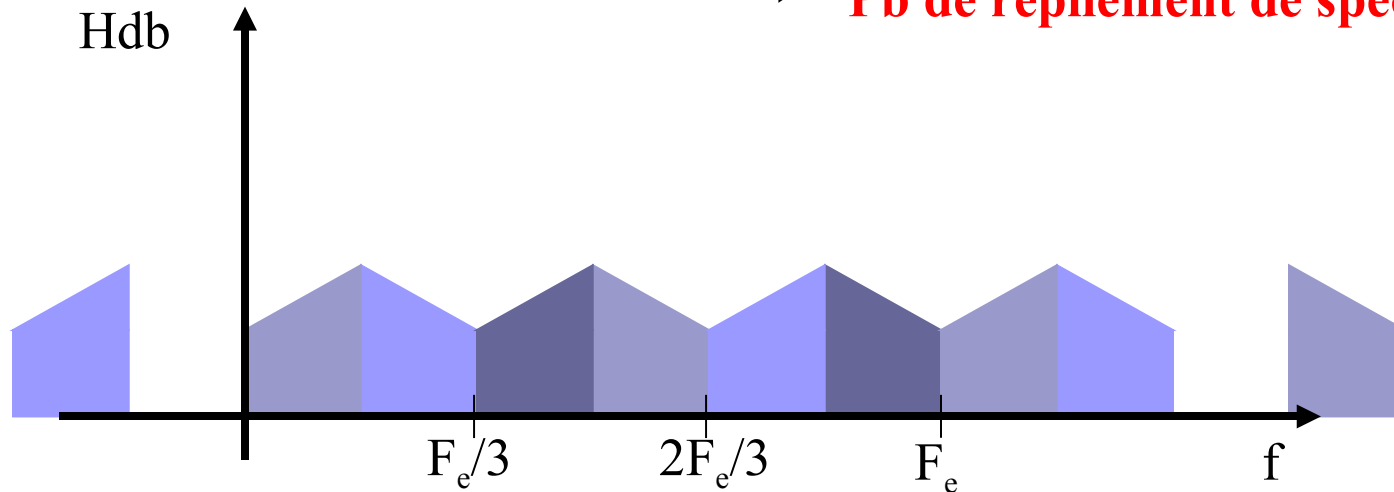
On prend 1 échantillon sur 3



Signal de départ échantillonné à $F_e/3$

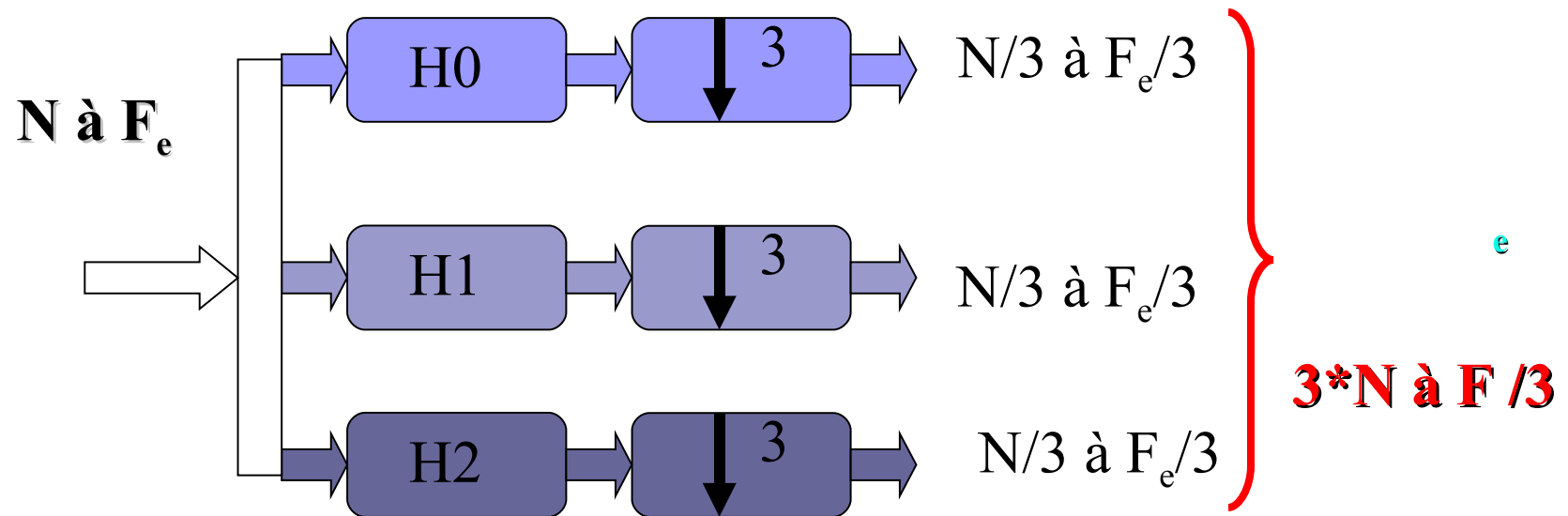


Pb de repliement de spectre



Impératif : Découpage en N sous Bande = Décimation par N

Récapitulatif du codage en sous bande



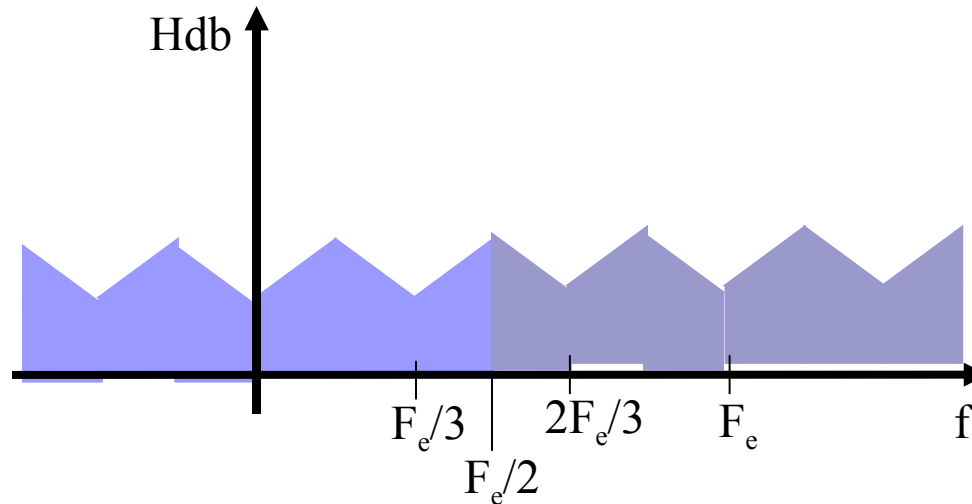
On ne change pas le nombre global d'échantillons

On peut maintenant coder séparément chaque sous bande :

- Nombre de bits différent

Sur échantillonnage

Opération inverse de la décimation



Réplication de $[-F_e/2, F_e/2]$ autour de F_e



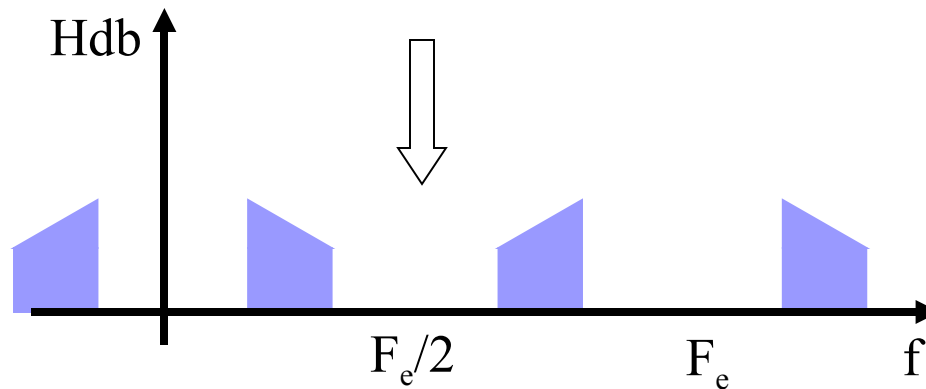
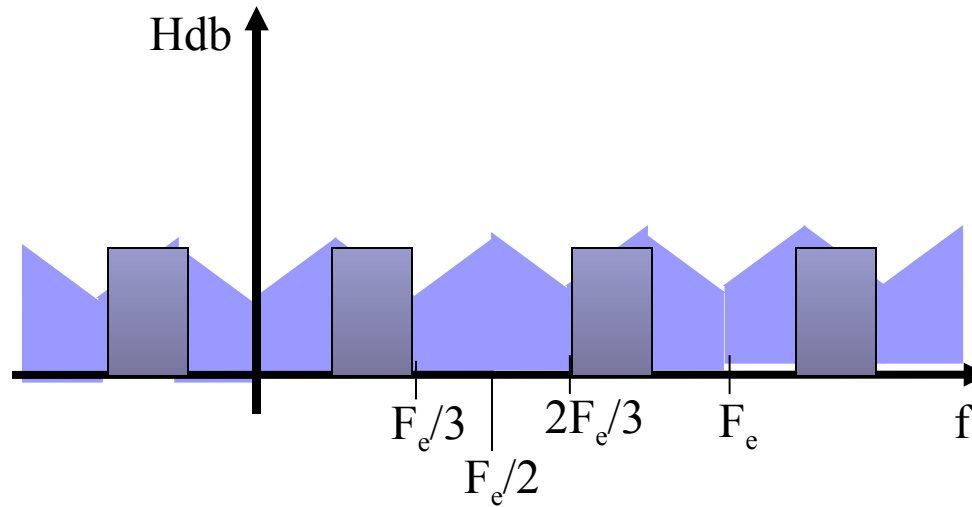
Superposition parfaite



On ne change que le gain

Retour à N signaux échantillonnés F_e

Filtrage passe bande

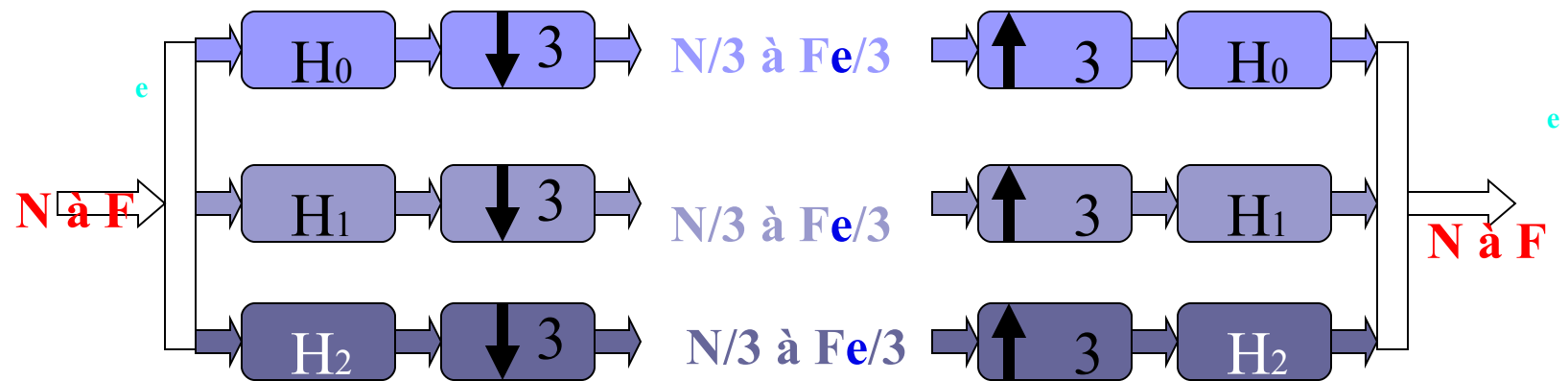


On passe par le même

1

Filtre Passe Bande H

Définition des filtres

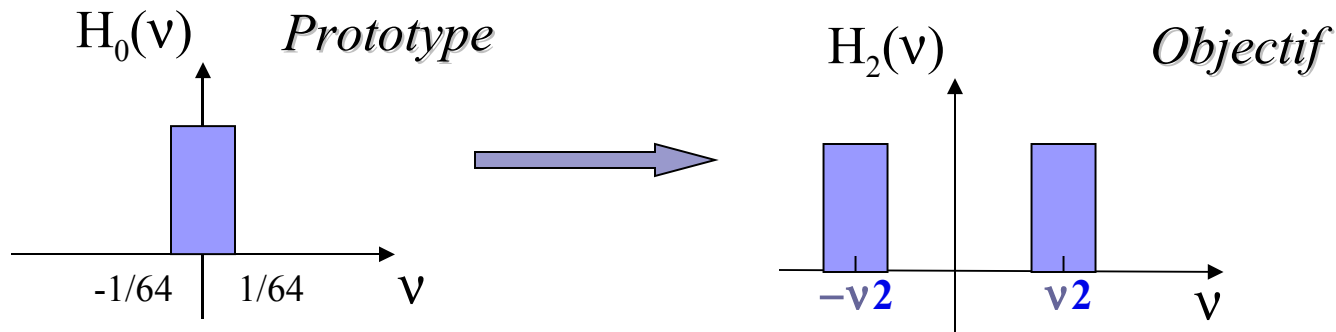


Construction des filtres (32 dans le cas du MP3)

Réalisation pratique des filtres

FIR passe-bas prototype sur 512 Coefficients.

Faire une translation dans chaque sous bande par une modulation cosinus



$$H_2(v) = H_0(v) * \delta(v - v_2) + H_0(v) * \delta(v + v_2)$$

$$H_2(v) = 2 \cdot H_0(v) * \left[\frac{1}{2} (\delta(v - v_2) + \delta(v + v_2)) \right]$$

2 e

TF(cos(2πv F .t))

$$\hookrightarrow [h_k(n)] = \left[2 \cdot h_0(n) \cdot \cos\left(\frac{n(2k+1)\pi}{64}\right) \right]$$

Exemples de Compression Mp3



fichier Initial : 32,1 Mo 16 bits 44kHz **1,5Mo/s** Stéréo



Compression 4,37 Mo 44 kHz **192kbits/s** Stéréo

1/7,3



Compression 1,45 Mo 44 kHz **64kbits/s** Mono

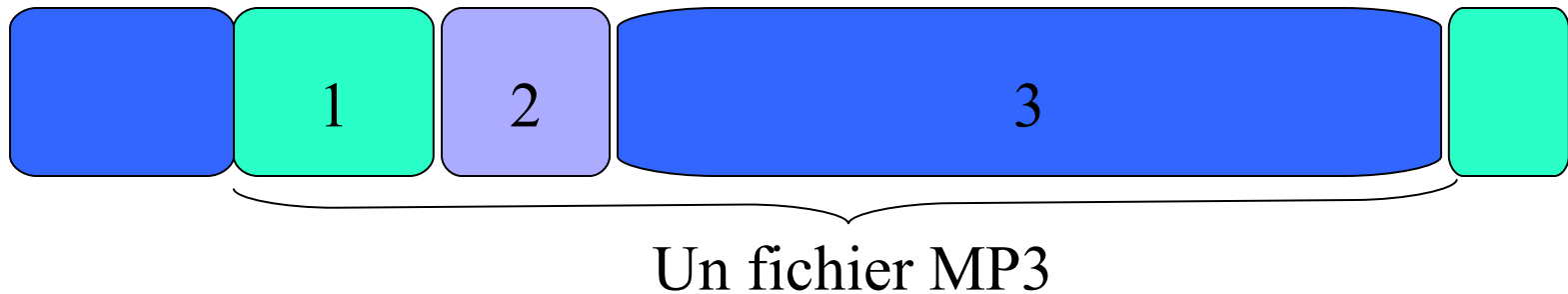
1/22



Compression 373 ko 16 kHz **16kbits/s** Mono

1/88

Structure de données fichier MP3



1 : synchronisation

2 : ID (renseignements sur la compression)

3 : données musicales

Parfois un 4ème wagon (ID3 ou Lyrics 3)