Codage de la parole

C.Mailhes

I - Le Contexte
II- Codage temporel : waveform coders
III- Modèles analyse/synthèse sinusoïdaux
IV- Vocodeurs
V- Codeurs prédictifs linéaires :
Analyse par Synthèse

Bibliographie :

A.Spanias, « Speech Coding, a tutorial review », Portions published in Proceedings of the IEEE, Oct. 1994
N.Moreau, « Techniques de compression des signaux », Ed Masson, collec. CNET / ENST.
R.Boite, H.Bourlard, T.Dutoit, J.Hancq, H.Leich, « Traitement du la parole », Presses Polytechniques et Universitaires Romandes, 2000.
http://www.eas.asu.edu/~speech/sp_cod.html
http://www.data-compression.com/
http://www-mobile.ecs.soton.ac.uk/speech_codecs/
http://www.eas.asu.edu/~speech/research/stc/presf/sld001.htm

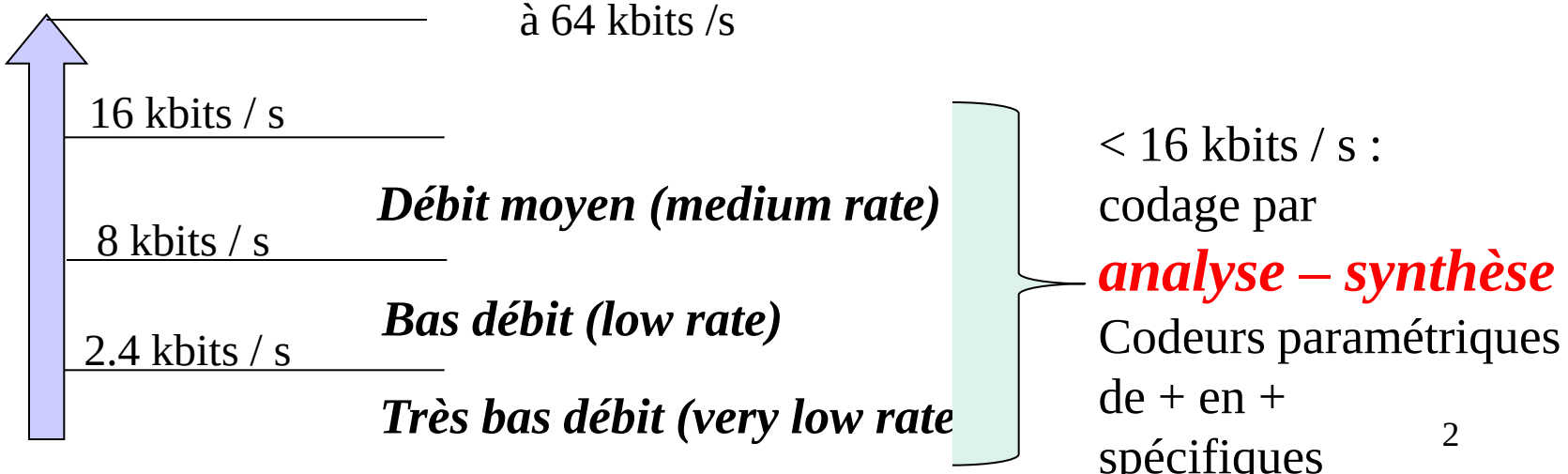
TERME en ANGLAIS	PAGE
ABS(Analyse By Synthesis)	46
ACELP(Adaptative CELP Coder)	13
ACELP(Algebraic CELP Coder)	63;70
AMDF(Average Magnitude Difference Function)	40
AMR-NB(Adaptative Multirate Narrow Band)	71
AMR-WB(Adaptative Multirate Wideband)	72
ATRAC(Adaptative Transform Acoustic Coding: MINIDISC)	77
CELP(code excited linear prediction,codebook excitation)	12;51;61
CNG(Comfort noise generation)	71
DAB(Digital Audion Broadcasting)	77;97
DAM(Diagnostic Acceptability Measure)	17
DPCM(Differential Pulse Code Modulation)	18
DRT(Diagnostic Rythm Test)	17
DTX(Discontinuous Transmission)	71
EFR(Enhanced Full Rate)	70
EPL(Erreur de Prédiction Linéaire)	18
ETSI(European Telecommunication Standard Intsitute)	67
FR(Full Rate)	68;69
GSM(Groupe Special Mobile)	13;68
Hifi(high fidelity)	75
HR(Half Rate)	70;71
IMDCT	95
ISDN	93;97
JSRU(Joint Speech Research Unit)	27
KBD	96
LAR(log area ratio)	
LD-CELP(Low Delay CELP Coder)	13
LP(Linear Prediction)	12
LPC(Linear Prediction Coder)	12,16,31
LSP(line spectrum pairs) ou LSF(LS Frequencies)	
LTP(Long Term Predictor)	47
MBE(Multiband Excitation Coder)	25
MDCT(Modified Discrete Cosine Tranform)	93;94;95
MELP(Mixte Excitation Linear Predictor)	41
MOS(Mean Opinion Score)	17
MPEG(Moving Pictures Expert Group)	77
MPEG-2 AAC(Advanced Audio Coder)	78;99
MPEG-2 BC(backward compatible)	99
MPEG-2 LSF(Low Sampling frequency)	99
MPLPC(Multi Pulse LPC)	51;67
MS(Middle/Side)	93
NICAM(Near-Instantaneously Companded Audio Multiplex-BBC)	77
PARCOR	33
PASC(Precision Adaptative Subband Coding: Digital Compact Cassette (DCC	77;97
PCM (Pulse Code modulation)	2
QMF(Quadrature Mirror Filter)	20
RELP(Residiual Excitation Linear Predictor)	42
RPELPC(Regular Pulse Excited LPC)	51;60
SELP(Self Excitation/ adaptative codebook)	51
SMR(Signal to Mask Ratio)	91
STC(Sinusoidal Transform Coder)	24
STFT(Short Time Fourier Transform)	22
STP(Short Term Predictor)	47
TDAC(Time Domain Aliasing Cancellation)	95
TNS(Temporal Noise Shaping)	100;101
VAD(Voice Activity Detection)	71
Vocoder(Voice Coder)	26
VSELP(Vector Sum Excited Linear Prediction)	64;65;70

I- Le Contexte - Introduction

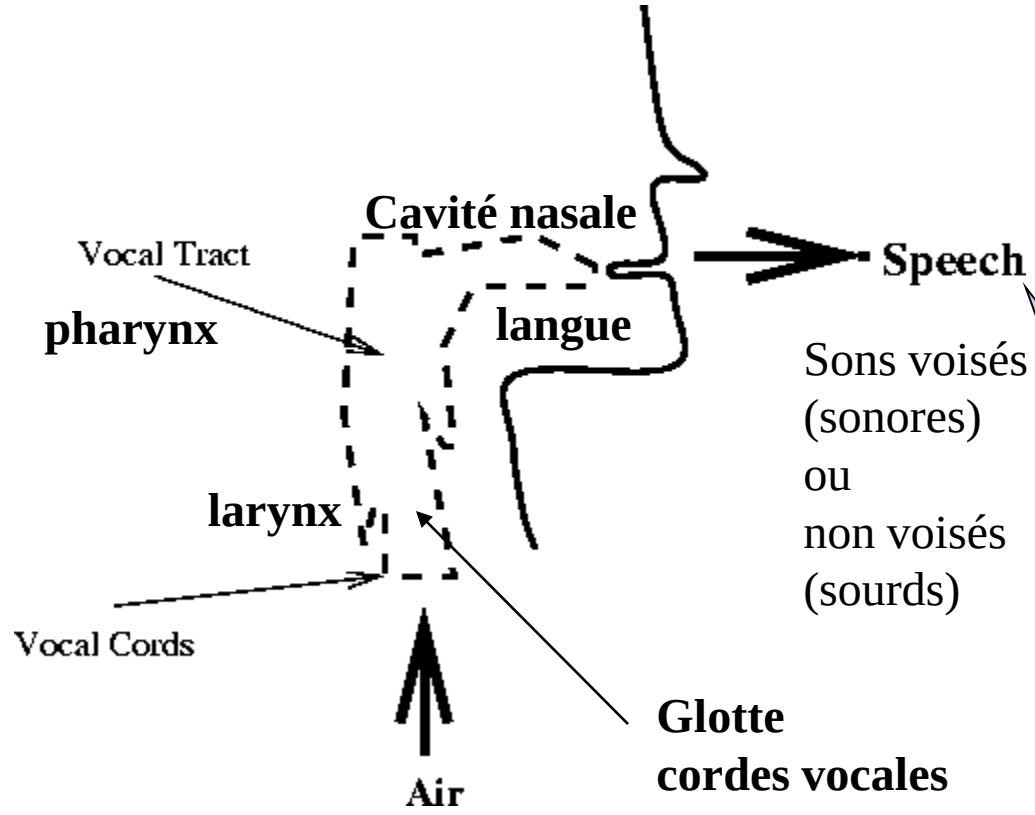
✉ Malgré fibres optiques, besoin croissant en largeur de bande avec sécurité accrue en communications satellitaires et cellulaires sans fil.
✉ Grand Intérêt dans intégration application parole sur PC personnels dans contexte multimédia (voicemail, voice Over IP)

⇒ ***Speech Coding or Speech Compression***

Communications numériques : largeur spectrale de la parole < 4 kHz (3.2 kHz)
Fe = 8 kHz
Technique de codage le plus simple non paramétrique : PCM ou MIC = référence à 64 kbits /s



I- Le contexte : Signal de Parole



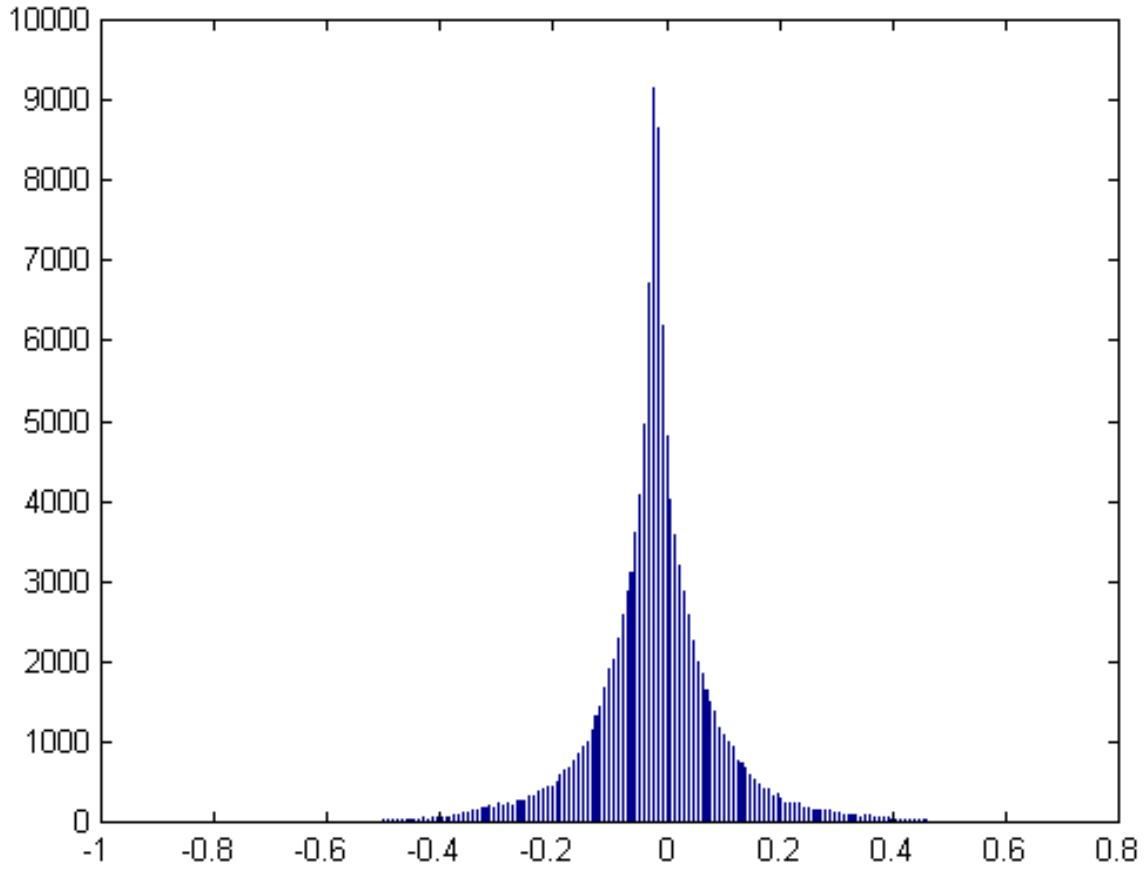
Sons voisés
(sonores)
ou
non voisés
(sourds)

Air poussé à travers la trachée artère
Au sommet, le larynx où pression de l'air modulée avant le conduit vocal.
Larynx : ensemble de muscles et cartilages mobiles
Cordes vocales : 2 lèvres symétriques placées à travers le larynx qui peuvent le fermer et en s'ouvrant forment une ouverture triangulaire : la glotte

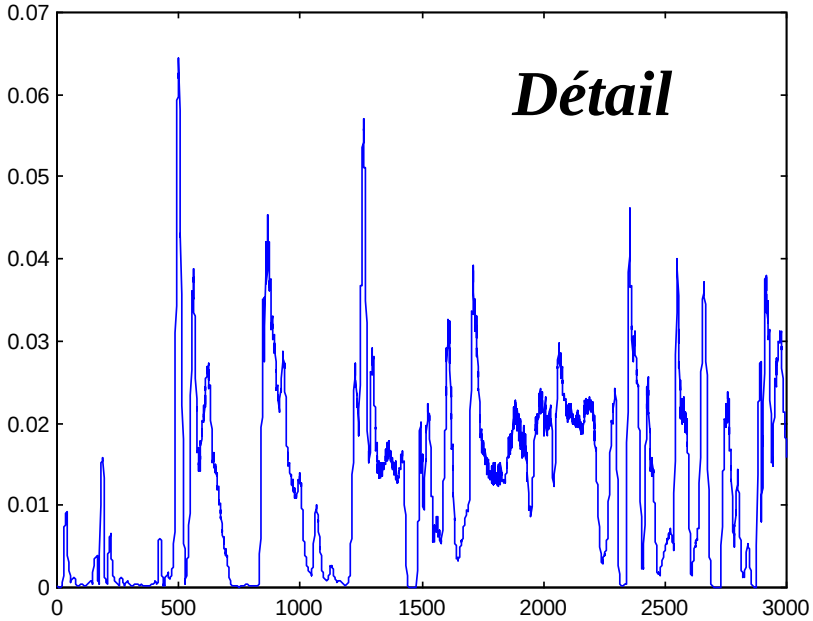
Spectre s'étend jusqu'à 12 kHz
Téléphonie : max à 3400 Hz
Analyse, Synthèse,
Reconnaissance de parole :
entre 6000 et 16 000 Hz
Audio (parole et musique)
jusqu'à 20 kHz (Fe=44.1 kHz)

Statistique de la parole ?

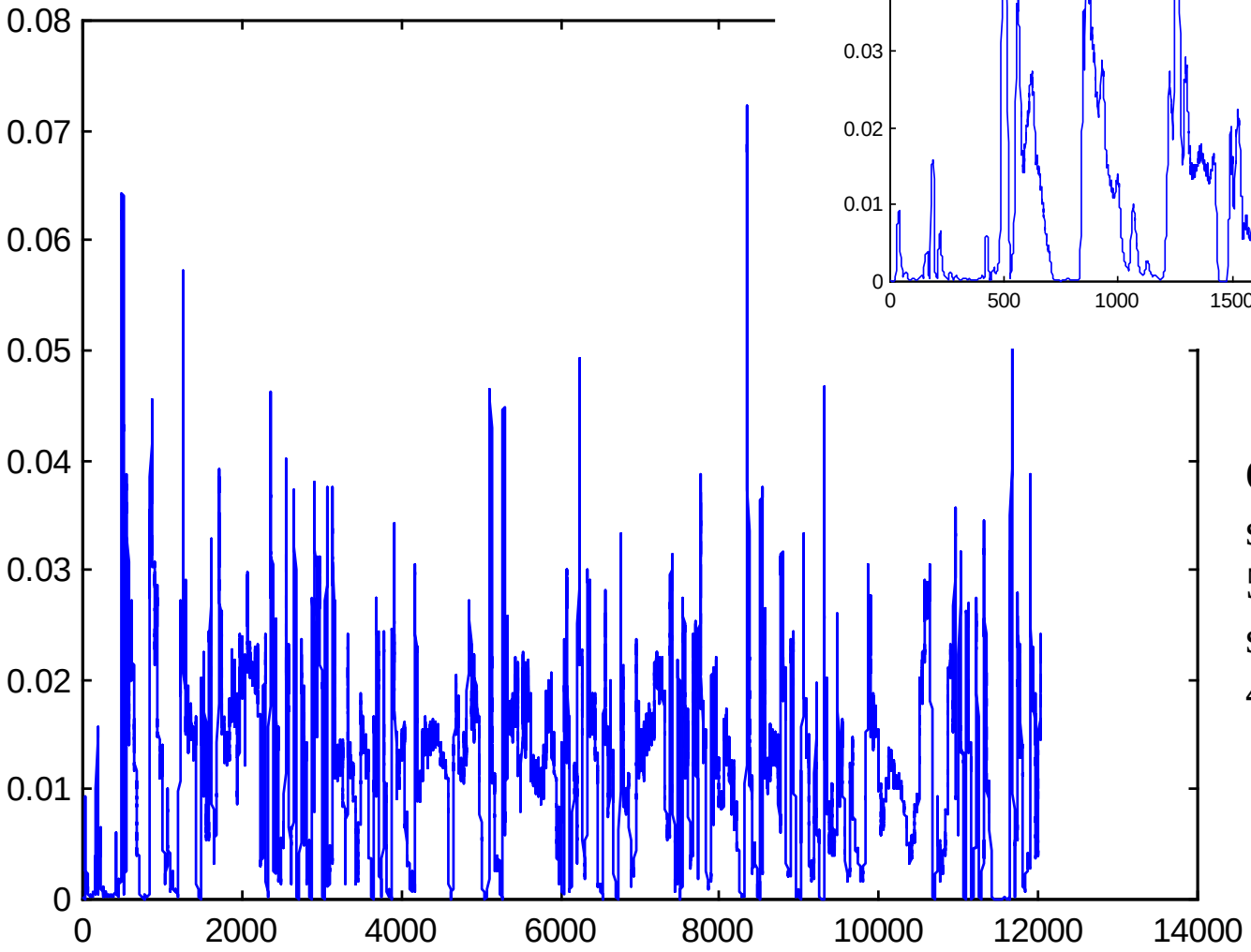
Densité de Probabilité (histogramme)



Stationnaire ?



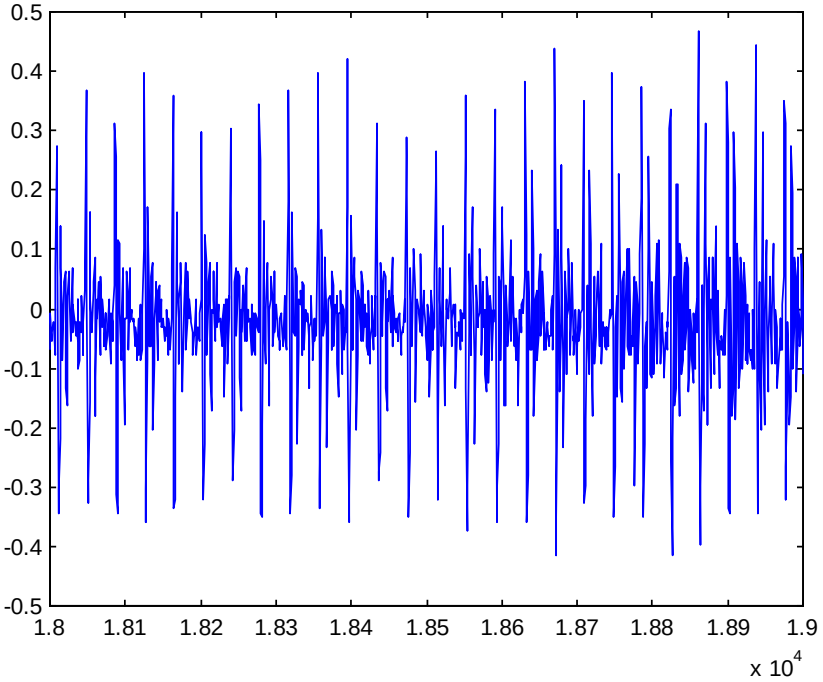
Evolution de la variance



Quasi -stationnaire
sur segments courts :
5 à 20 ms
soit
40 à 160 échantillons

Analyse d'un morceau de parole

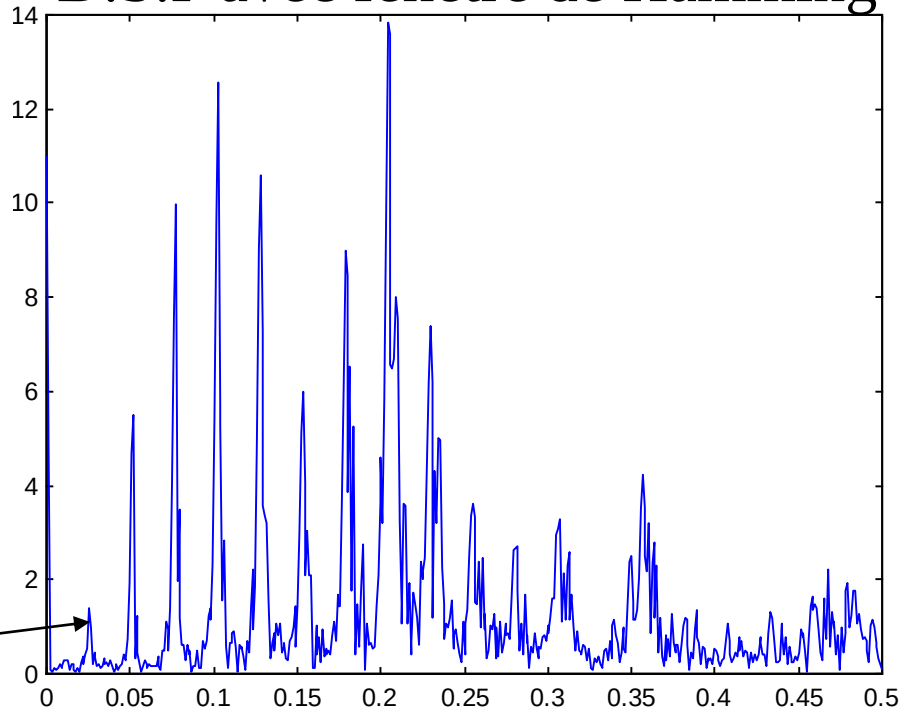
« … *ne* »



Son voisé (voiced)

Signal temporel, Fe=8 kHz
tranche de 1000 échantillons
soit 125 ms

D.S.P avec fenêtre de Hamming



Fréquence fondamentale
Pitch

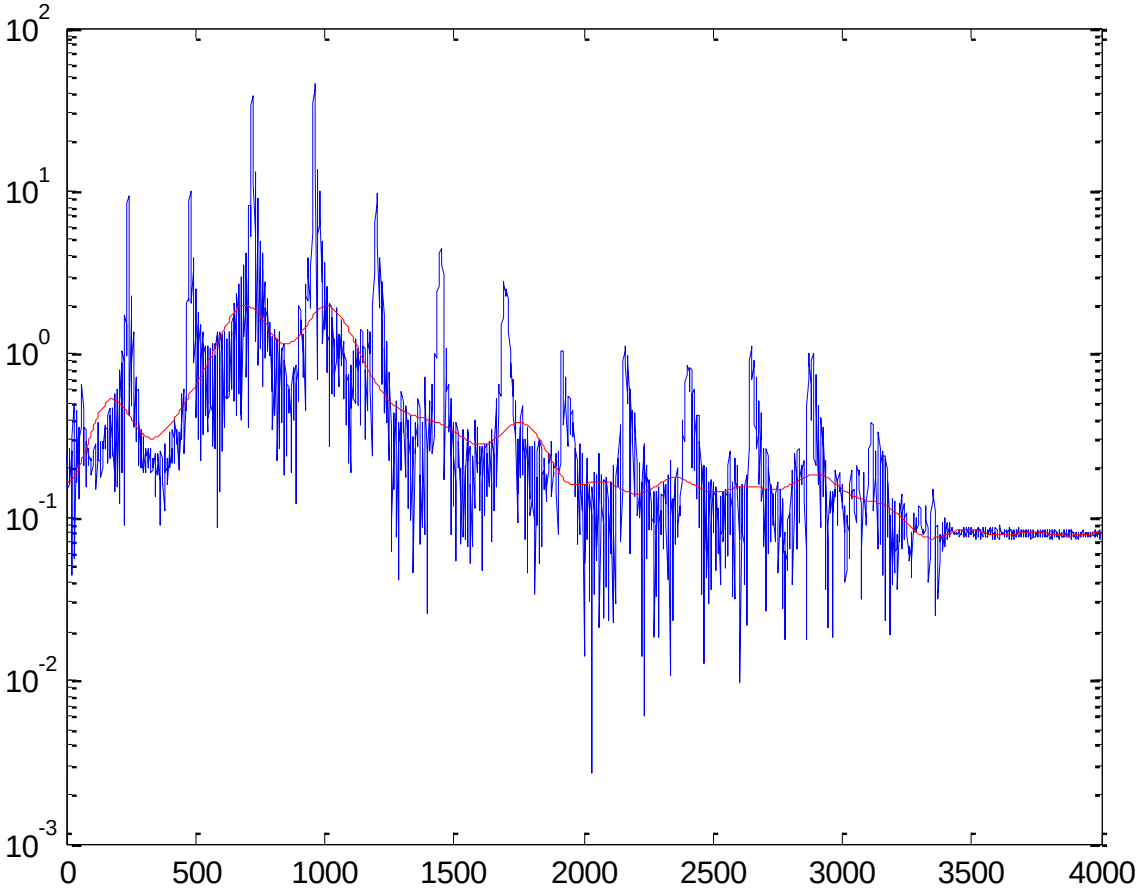
Densité Spectrale de Puissance en échelle logarithmique d'un son voisé

« … fi »

In english

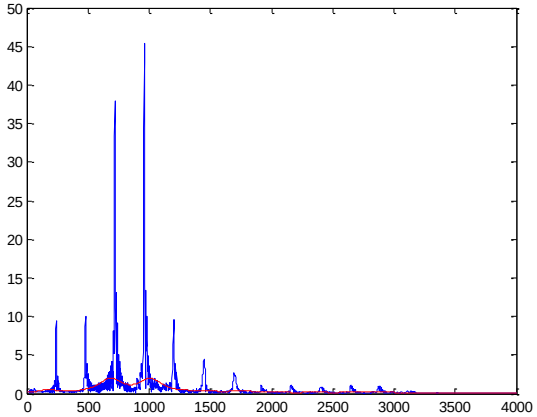
dB

Enveloppe spectrale : formants

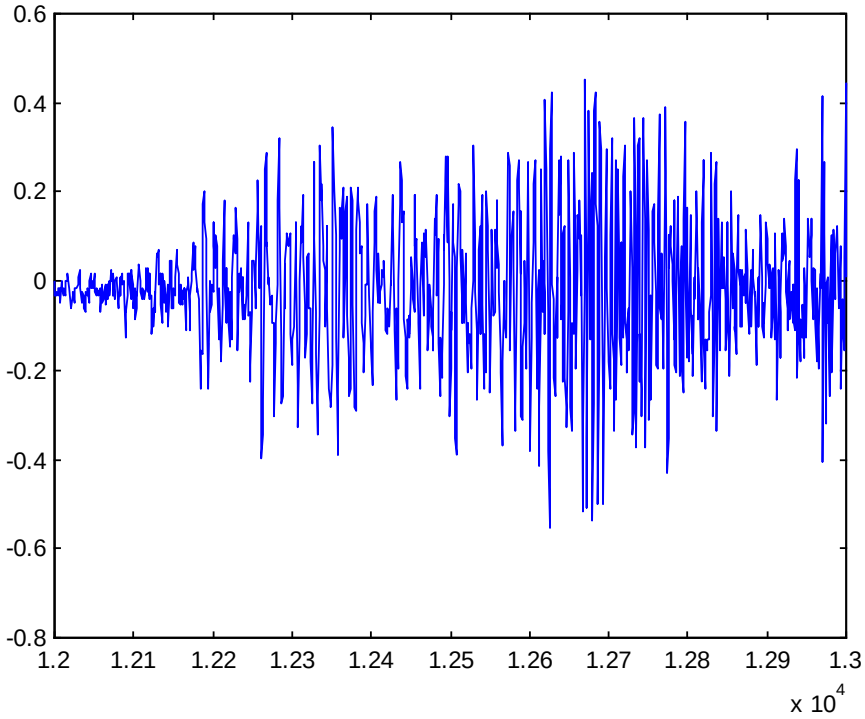


Fréquences (Hz)

(en linéaire)



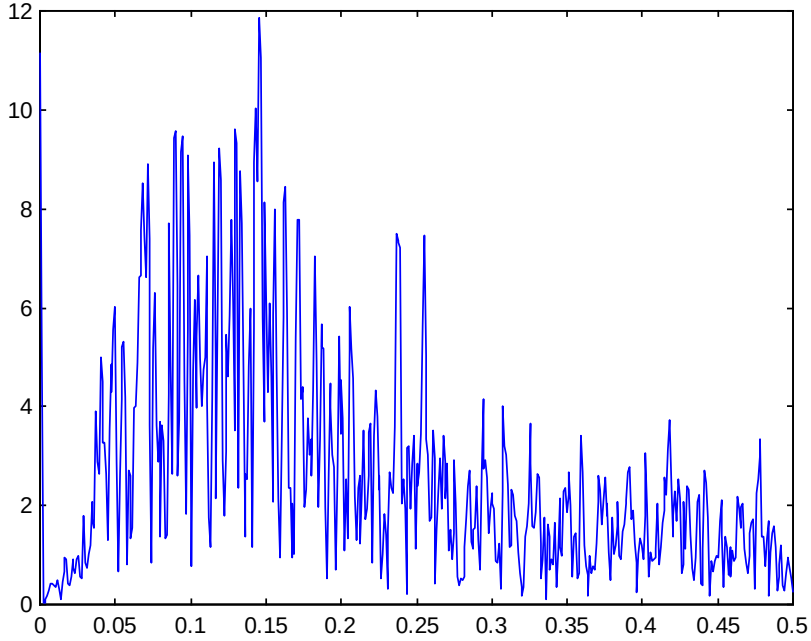
« … c'est »



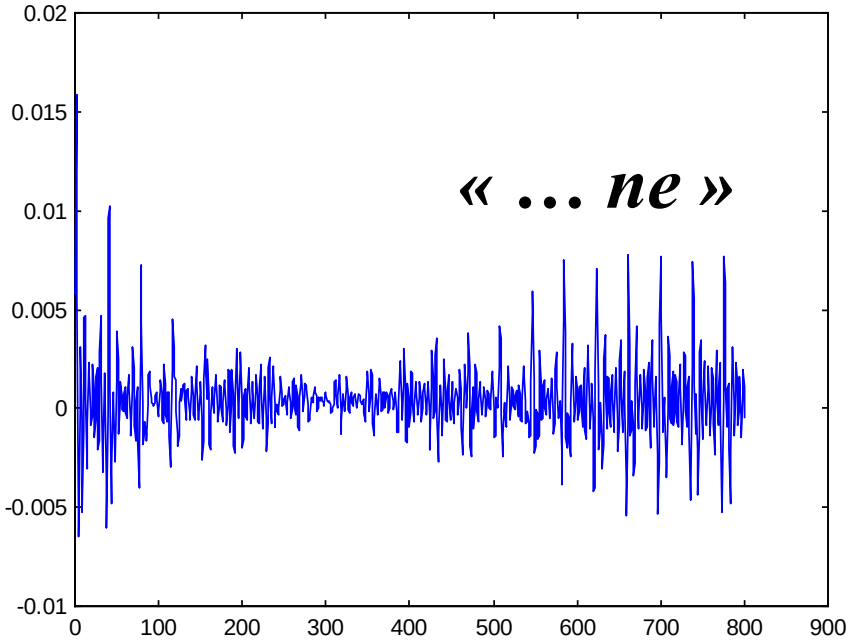
Son non voisé (unvoiced)

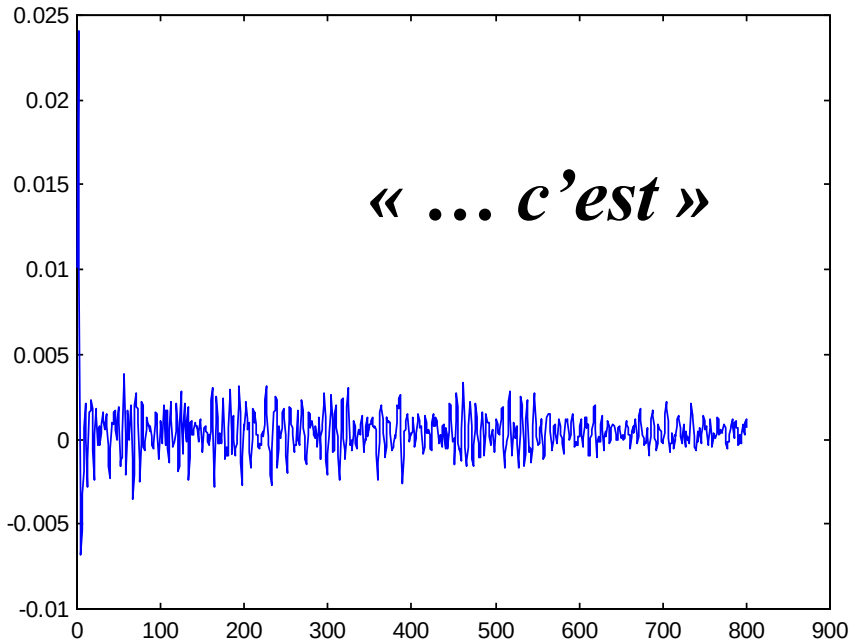
Signal temporel, Fe=8 kHz
tranche de 1000 échantillons
soit 125 ms

D.S.P avec fenêtre de Hamming

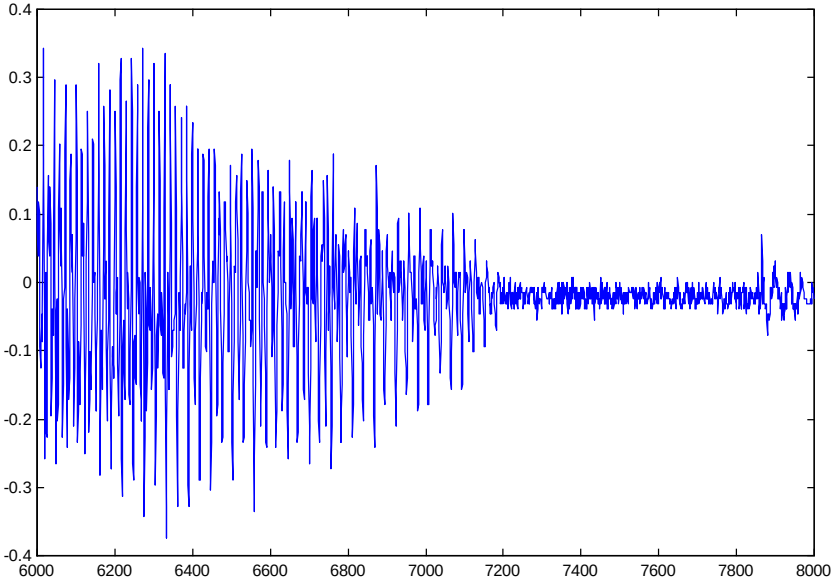


Autocorrélation des sons voisés et non voisés





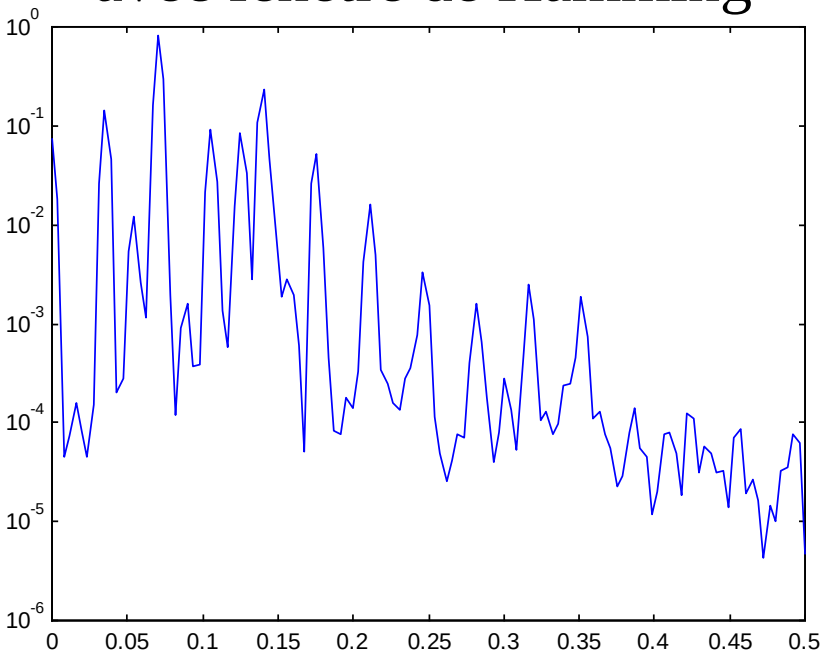
« … *lo* »



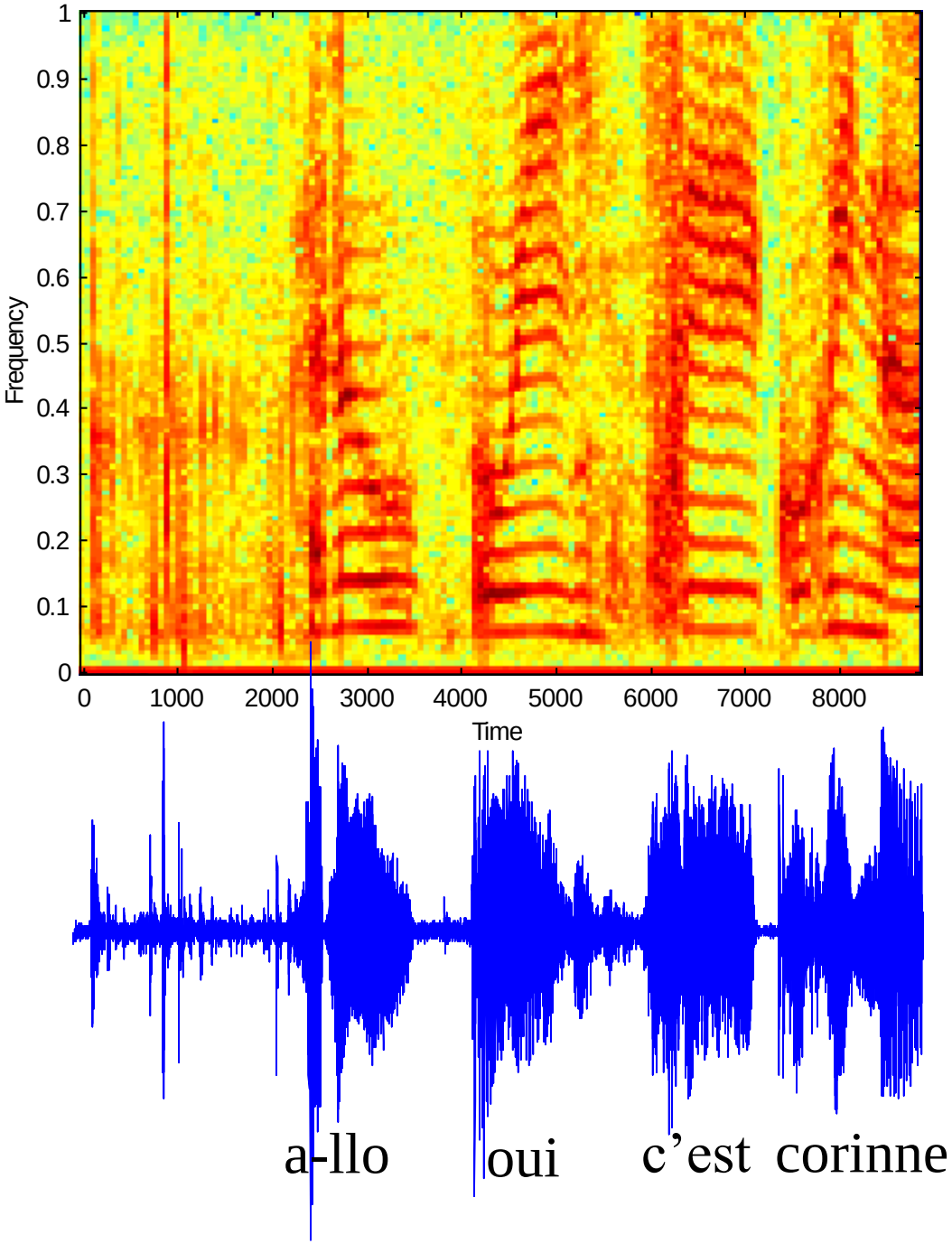
Signal temporel, Fe=8 kHz
tranche de 2000 échantillons
soit 2*125 ms

Son mixte

Périodogramme moyenné
tranches de 256 points
avec fenêtre de Hamming



Emargo rtcepss e

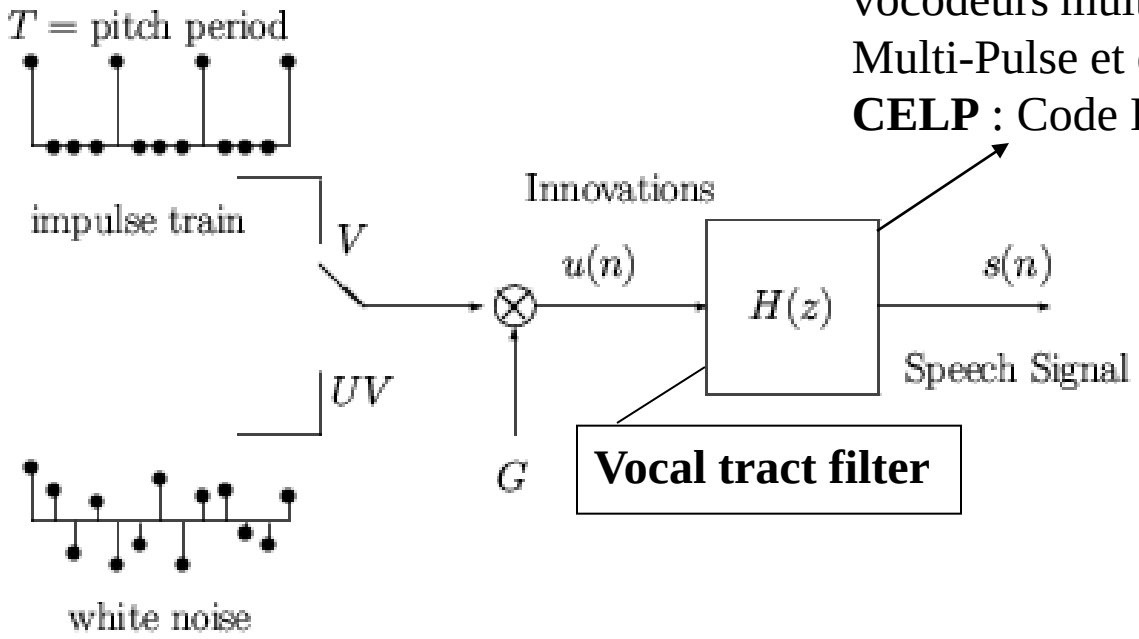


I- Le contexte : Historique

Dès les années 1940, PCM, DPCM, DM; ADPCM…
Fin des années 1950, concentré vers le modèle linéaire de production de la parole

1958 : début du codage de parole : **Dudley** du Bell Telephone Lab
1er système de codage d ’analyse-synthèse analogique :
banc de 10 filtres passe-bande analogiques (représentant le conduit vocal)

Fin des années **60** : Prédiction Linéaire (LP) pour le filtre
Années **60 -70** : apparition du codage en sous-bande
Années **80-90** : codeurs robustes à bas débit pour parole haute qualité:
analyse-synthèse sinusoïdale,
vocodeurs multibandes,
Multi-Pulse et excitation vectorielle dans le LPC
CELP : Code Excited Linear Prediction (1984)



I- Le contexte : Historique

Réseau Téléphonique public : normes
UIT-T : Union Internationale des Télécommunications
(ancien CCITT : Comité Consultatif International Télégraphique et Téléphonique)

1972 : G.711 : PCM à 64 kbits/s :
quantification sur 8 bits après compression non linéaire

1984 : G721 : ADPCM à 32 kbits /s
filtrage de type adaptatif

1991 : G.728 : LD-CELP à 16 kbits /s
Low Delay Code Excited Linear Predictive Coder
tecniques de modélisation et de Q.Vectorielle
faible délai de reconstruction

1994 : ACELP à 8 kbits /s
Adaptive Code Excited Linear Predictive Coder

Comparaison de codeurs

Communications avec les mobiles : normes

1989 : GSM : Groupe Special Mobile
à base de CELP

I- Le contexte : Mesure des performances

Evaluation d'un algorithme de codage de parole : comparaison :

- ***taux de bits***
- ***qualité de parole reconstruite***
- ***complexité de l'algorithme***
- ***retard introduit***
- ***robustesse de l'algorithme aux erreurs de canal et à l'interférence acoustique.***

Codage parole haute qualité à bas débit : algorithmes très complexes
Implantation d'un algo temps réel bas débit ☑ DSP d'au moins 12 MIPS
Retard introduit (codage + décodage) entre 50 et 60 ms

Classification de la qualité de parole :

- Haute qualité (radiodiffusion) > 64 kbits / s
- Réseau : qualité comparable à parole analogique > 16 kbits / s
- Communications : parole plus très naturelle, hautement intelligible, pour télécoms > 4.8 kbits/s
- Synthétique : intelligible mais non naturelle, perte de reconnaissance du locuteur

I- Le contexte : Mesure des performances

Critères Objectifs :

- SNR inadéquat (parole non stationnaire)

➔ SNR segmental (SNR moyenné sur *m* tranches temporelles)

- Index d'articulation : SNR moyenné sur des bandes fréquentielles

- Distance log-spectrale,
- Log Likelihood Ratio
- Distance cepstrale

Mesures spectrales

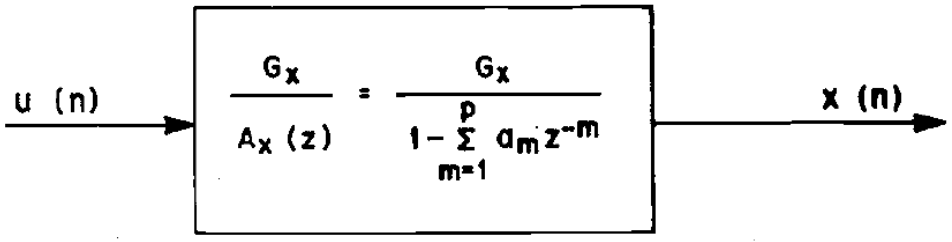
Estimation de la DSP
T.F. non adaptée

$$d_S = \int_{-\pi}^{+\pi} \left| \log\left(S\left(\omega\right)\right) - \log\left(\widehat{S}\left(\omega\right)\right) \right| d\omega$$

Une alternative à l'analyse spectrale à base de T.F. :

La modélisation paramétrique

Modélisation AR (auto-régressive)



$$x(n) = \sum_{m=1}^{p} a_m x(n-m) + G_x u(n)$$

Coefficients auto-régressifs (ou LPC)

Soient a_x le vecteur LPC de la parole originale et a_y celui de la parole codée.
« log-likelihood ratio » :

$$l = \log\left[\frac{a_x R_y a_x^t}{a_y R_y a_y^t}\right]$$

Energie de l'erreur LPC

Distance cepstrale

$$d_{cep}(n) = \sum_{i=-N}^{N} \left(c_{n,i} - c'_{n,i}\right)^2 \qquad \text{avec} \qquad c_i = -a_i - \sum_{k=1}^{i-1}\left(1 - \frac{k}{i}\right) c_{i-k} a_k; \quad i > 0$$

I- Le contexte : Mesure des performances

Critères Subjectifs :

Pour tenir compte propriétés perceptuelles de l'oreille : ***tests d'écoute en double aveugle avec référence cachée*** : on présente aux auditeurs 3 versions A,B,C ;
A est toujours l'originale,
B et C : une est l'original, l'autre la codée-décodée
Echelle **MOS (Mean Opinion Score)** des dégradations perçues :

5 : imperceptibles
4 : perceptible mais non gênant
3 : légèrement gênant
2 : gênant
1 : très gênant

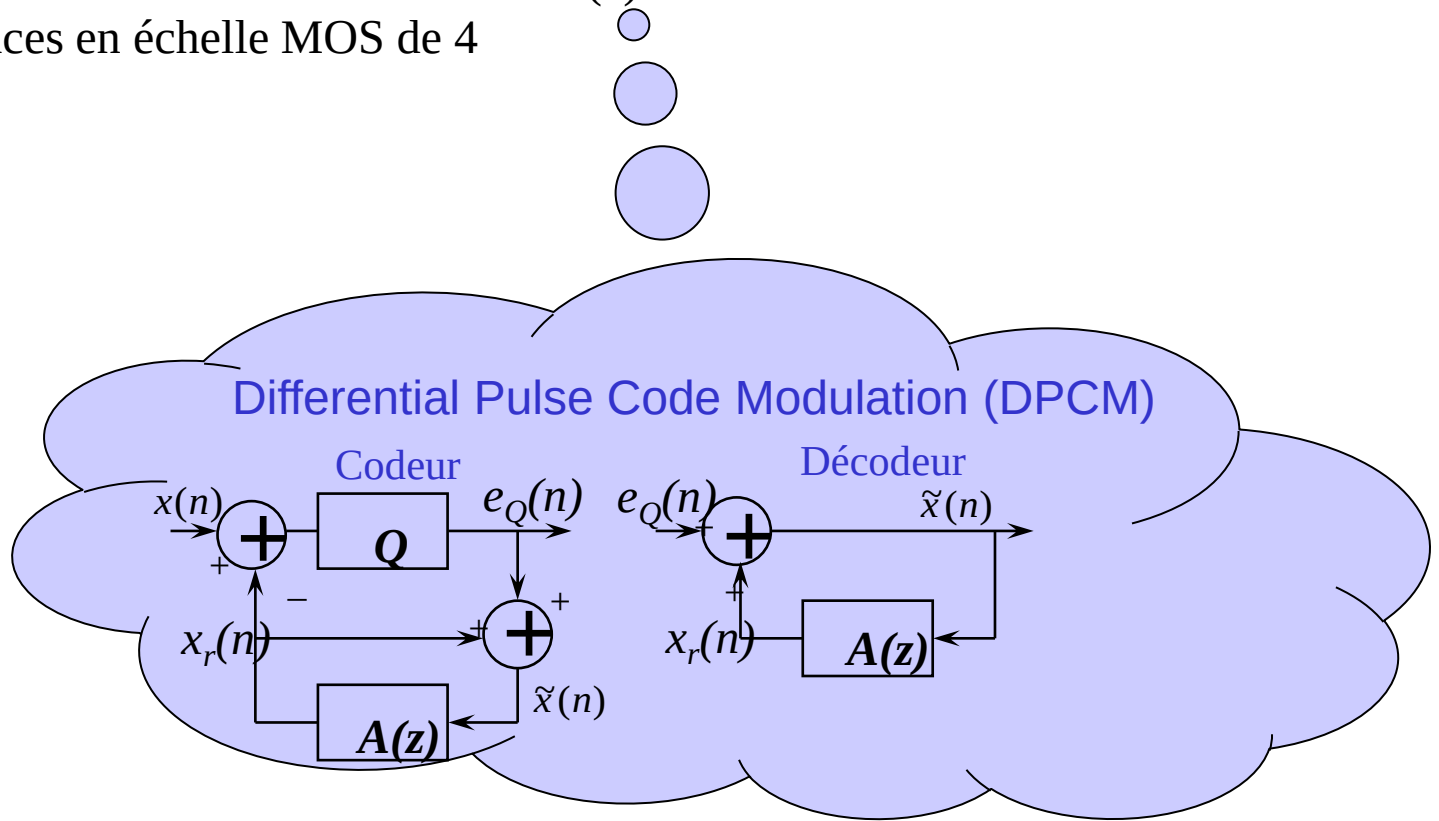
Qualité d'un codeur : moyenne des notes reçues avec intervalle de confiance à 95%
(12 à 24 auditeurs entraînés voire 32 à 64 pour normalisation)
MOS de 4 - 4.5 : qualité réseau, 3.5 - 4 : qualité coms, 2.5 - 3.5 : synthétique
DRT : Diagnostic Rythm Test : mesure d 'intelligibilité visant à reconnaître un de 2 mots d 'une paire qui rime (« meat - heat »)
DAM : Diagnostic Acceptability Measure :
acceptation de la parole à l 'aide d 'auditeurs experts

II- Codage temporel - Waveform coders

Systèmes de codage à débit élevé (>16 kbits / s)
Codeurs génériques (cf cours Compression de données)

❶ Codage Prédictif

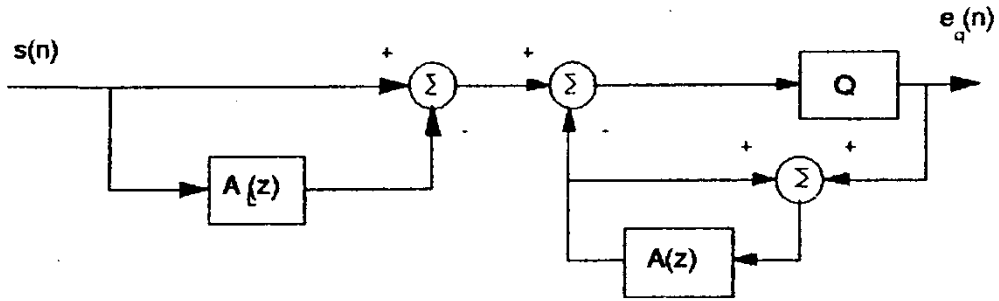
ADPCM à 32 kbits/s : standard G721 : prédicteur adaptatif (2 pôles, 6 zéros) + Q adaptatif
EPL codée sur 4 bits, algorithme du gradient pour prédicteur adaptatif
stabilité contrôlée en testant racines de A(z)
Performances en échelle MOS de 4



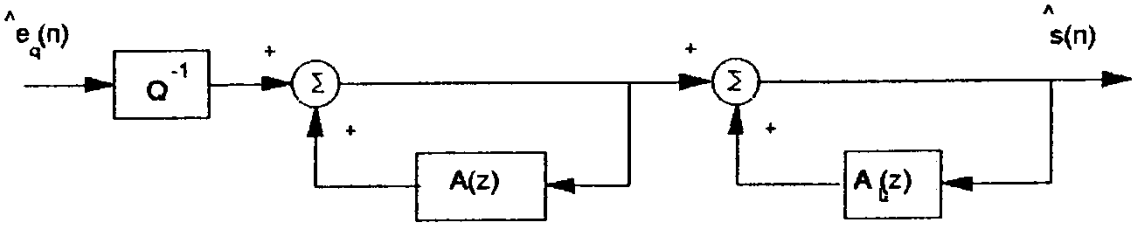
Introduction d'un prédicteur long terme

Utilisation de la **prédiction long terme adaptative** permet haute qualité à 16 kbits /s

Prédicteur long terme fournit le pitch, filtre enlève la périodicité donc la redondance



(a) Transmitter



(b) Receiver

Pitch ? Homme : 100-150 Hz, Femme : 200-300 Hz, Enfant : 300-450 Hz

Retour vers CELP

II- Codage temporel - Waveform coders

❷ Codage en sous-bande

réduction du débit par propriétés spectrales de la parole et propriétés perceptuelles de l'oreille.
Codage en sous-bandes : alloue plus de bits en basses fréquences pour préserver le pitch et les formants
Crochière : 1er sous-bande : 4 bandes (200-700, 700-1310, 1310-2020, 2020-3200 Hz)
opère à 16, 9.6 et 7.2 kbits /s
à 16kbits /s se compare favorablement à ADPCM à 16 kbits /s
Conception des filtres importante ; largeurs de bandes identiques ou non
En parole, bandes basses-fréquences plus étroites pour mieux représenter pitch et formants

Standard : **AT&T SBC voice store à 16 et 24 kbits / s**
5 filtres QMF non uniformes associés à codeurs ADPCM
bandes : 0-500, 500-1000, 1000-2000, 2000-3000, 3000-4000 Hz
allocations de bits : 4 / 4 / 2 / 2 / 0 pour 16 kbits /s et 5 / 5 / 4 / 3 / 0 pour 24 kbits /s
retard de l 'ordre de 18 ms


Comparaison codeurs

Autre standard : **G722 pour l 'audio 7 kHz à 64 kbits /s** pour téléconférences
codeur à 2 bandes + ADPCM
sous bande BF à 48 kbits / s, l'autre à 16 kbits /s, allocation de bits adaptative
retard des QMF < 3ms MOS >4 pour parole et à peine <4 pour musique

Codage de la parole

I - Le Contexte

II- Codage temporel : waveform coders

III- Modèles analyse/synthèse sinusoïdaux

autre classe de codeurs de parole
repose sur modèle sinusoïdal de la parole
1ères méthodes basées sur la TF court-terme (**STFT**)
puis années 80 : **STC** et **MBE**
reposent sur propriétés de parole mais ***plus robustes*** que traditionnels vocodeurs LP à 2 états (voisés ou non) :
fonctionnent pour une plus large classe de signaux

IV- Vocodeurs

V- Codeurs prédictifs linéaires : Analyse par Synthèse

III- Modèles analyse/synthèse sinusoïdaux

1. Analyse-Synthèse à partir de la TF court-terme

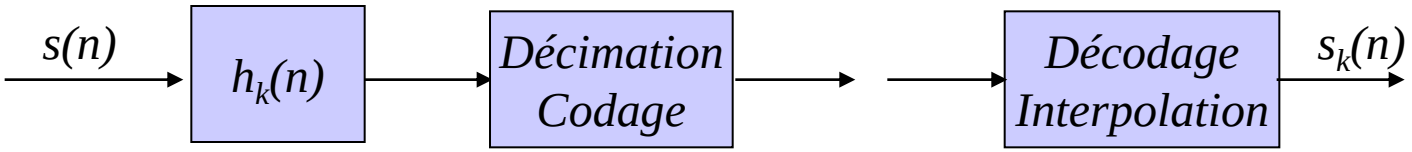
Parole : contenu spectral évolue lentement dans le temps : modélisable par sa PSD court-terme
Analyse temps-fréquence : **STFT** : Short Time Fourier Transform :

$$S(n,f) = \Sigma\, s(m)\, h(n-m) \exp(-j2\pi\, f\, T m) = h(n) * s(n) \exp(-j2\pi\, f\, T n)$$
$$= h(n) \exp(-j2\pi\, f\, T n) * s(n)$$

STFT

$h(n)$: fenêtre d'analyse glissante
Largeur de la fenêtre entre 5 et 20 ms (40 à 160 échantillons) : résolution spectrale pauvre

➢➢ **interprétation de la STFT comme un banc de filtres** : $h_k(n) = h(n) \exp(-j2\pi\, f_k T n)$
avec $f_k = k\,\Delta f$, k=0,…,N-1 de façon à couvrir toute la bande utile



$$\hat{s}_{STFT}(n) = \Sigma_k\, S(n, f_k) \exp(j2\pi\, f_k T n)$$

Flanagan, Golden, 1966 : 1er vocodeur de phase : 30 canaux uniformes de 50 à 3050 Hz
chaque canal : filtre de Bessel d'ordre 6

III- Modèles analyse/synthèse sinusoïdaux

2. Analyse-Synthèse sinusoïdale

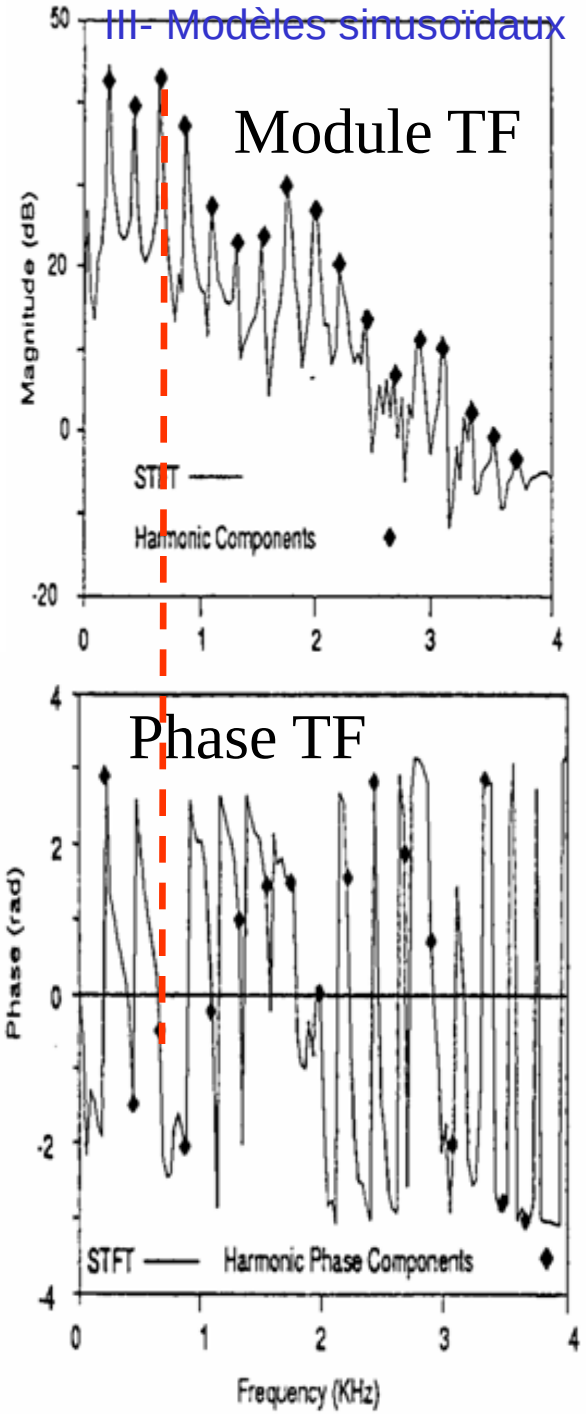
Mc Aulay et Quatieri (1988-1992) :
modèle = combinaison linéaire de L sinusoïdes
dont les **fréquences, amplitudes, phases et le nombre L varient dans le temps** :

$$s_{SR}(n) = \Sigma_{k=1,L} A_k \cos(2 \pi f_k n + \phi_k)$$

Contribution de Mc Aulay et Quatieri :
développement d'algos pour le « tracking » des paramètres des sinusoïdes d'une fenêtre à l'autre,
concept de naissance et de mort de modes
En pratique, fenêtre de Hamming adaptative,
largeur = 2.5 pitch moyen, FFT de 1024 points
mise à jour toutes les 10 ms : L peut aller jusqu'à 80 !!!
Fonctionne sur une large classe de signaux : multiples locuteurs, musique, sons biologiques…

Voisé

Non voisé



III- Modèles analyse/synthèse sinusoïdaux

Applications bas-débit :

fréquences contraintes à être des harmoniques du pitch

$s_{HR}(n) = \Sigma_{k=1,L(f_0)} A_k \cos(2 \pi k f_0 n + \phi_k)$

Parole voisée : pitch constant sur la fenêtre d'analyse

Parole non voisée : jeu de sinus équidistants (très proches <100Hz) en fréquence

3.6 kbits/s :Amplitudes codées par DM, allocations de bits adaptatives au pitch (+ de bits pour pitch ↗)

4.4 kbits/s pour fréquence fondamentale et phases (codées sur 4-5 bits)

Modèle sinusoïdal performant mais sensible aux erreurs (quantification, canal)

→ modèle plus robuste pour le codage de la parole bas débit haute qualité : **STC**

Sinusoïdal Transform Coder : codage de **l'enveloppe d'amplitude A(f)**

interpolation linéaire entre pics de la STFT aux fréquences f_1, f_2,… (pics de la STFT)

On code les **coefficients cepstraux** correspondants : TF $^{-1}$ (log(A(f)))

MOS de 3.52 à 4.8kbits/s et 2.9 à 2.4 kbits/s, 13 MIPS (TMS320c30)

III- Modèles analyse/synthèse sinusoïdaux

Voisé

Non voisé

2 états

3. **MultiBand Excitation Coder (MBE)**

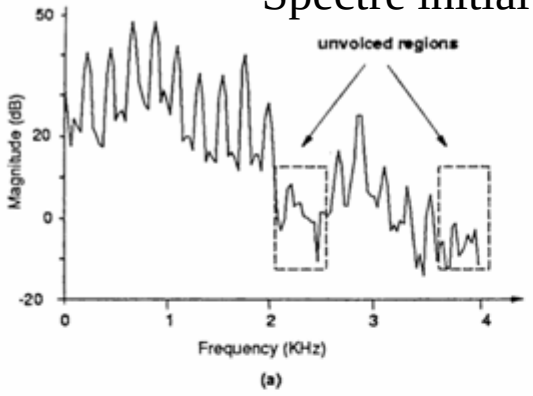
Griffin et Lim (1987)

Modèle : ST PSD = spectre d'excitation X enveloppe du conduit vocal

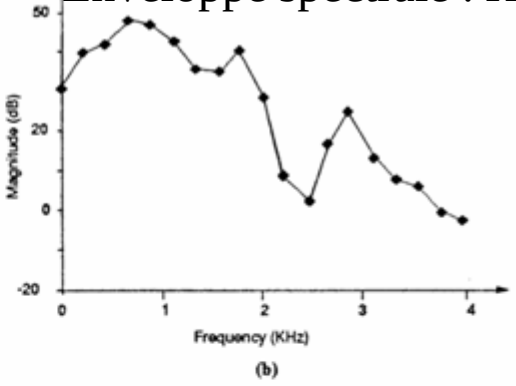
$$S(f) = H(f) \mid X(f) \mid$$

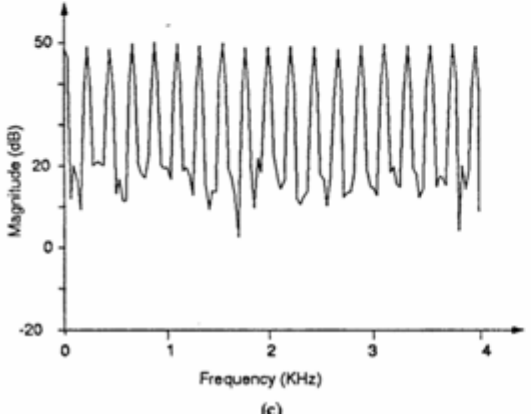
Différence avec le modèle classique à 2 états : le spectre d'excitation = combinaison de contributions harmoniques et aléatoires (voisement dépendant de la fréquence)

Spectre initial

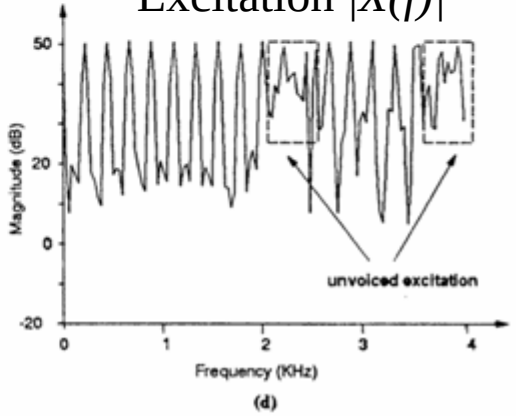


Enveloppe spectrale : *H(f)*





Excitation $|X(f)|$



Paramètres du MBE :
pitch,
enveloppe spectrale,
info de voisement pour chaque harmonique,
info de phase uniquement pour harmoniques voisées.

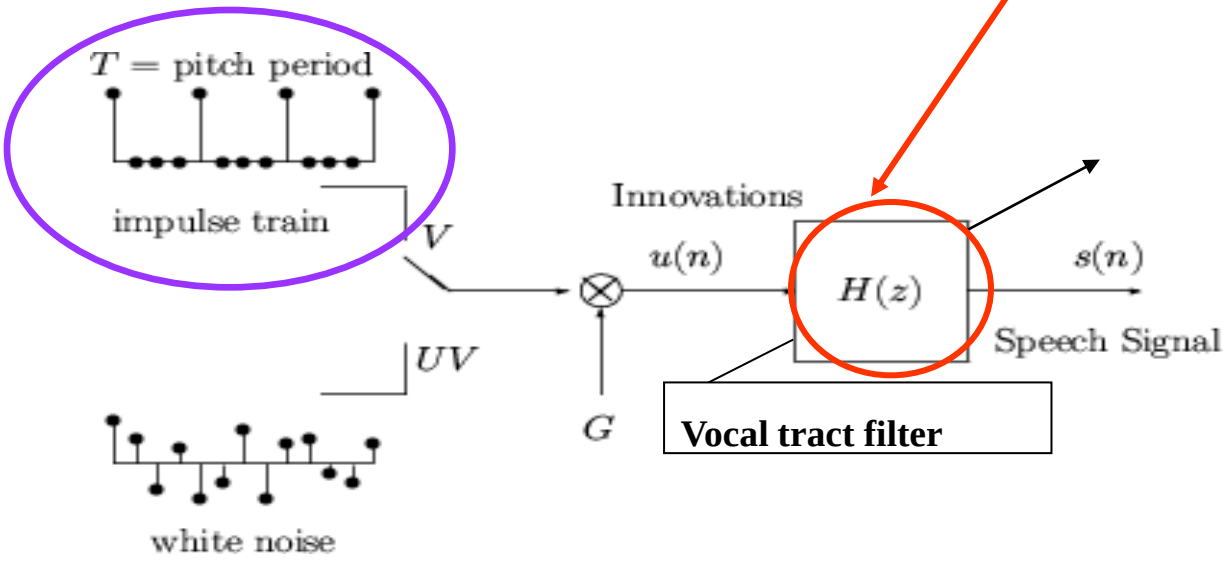
IV- Vocodeurs

Codeurs spécifiques de la parole ou vocodeurs
Performances se dégradent si utilisés sur autres signaux que parole
Basés essentiellement sur modèle de système de source : date des travaux de Dudley
4 vocodeurs décrits :

- ***vocodeur canal***
- ***vocodeur formant***
- ***vocodeur homomorphique***
- ***vocodeur à base de prédiction linéaire***

Plusieurs façons de modéliser le filtre

Comment estimer le pitch ?



Vocodeur Canal

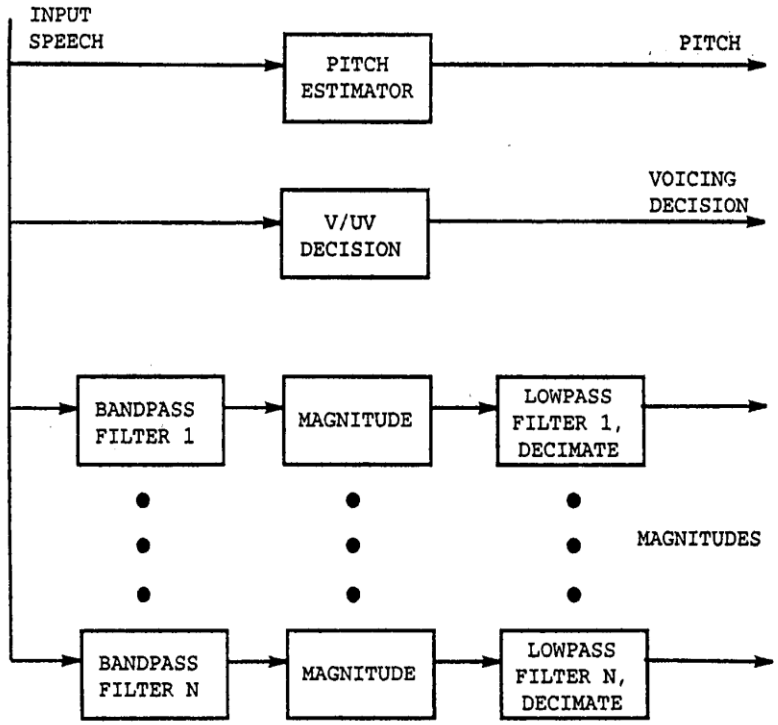


FIGURE 9.1
Channel vocoder analysis of input speech [137].

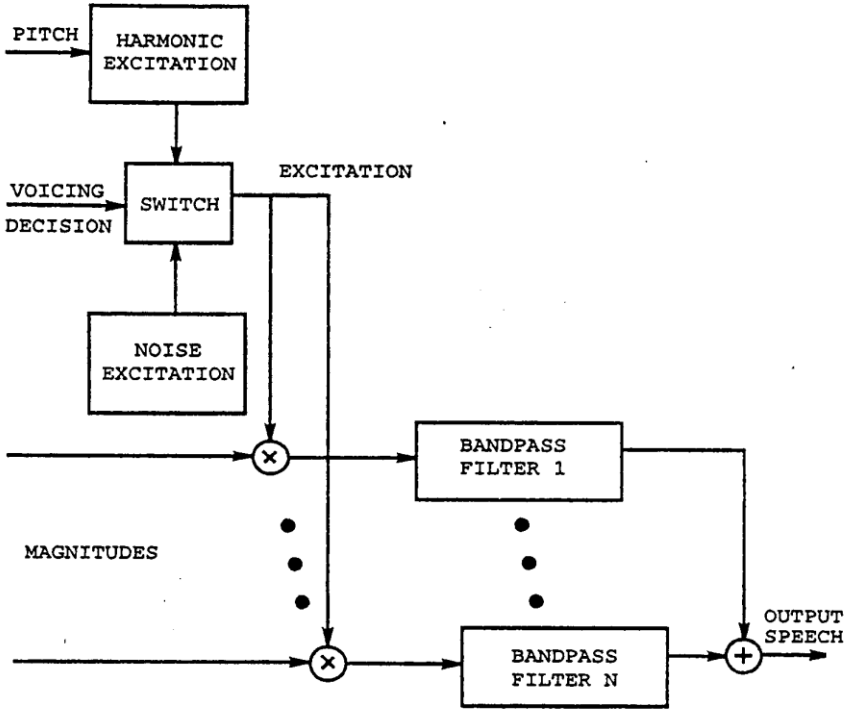


FIGURE 9.2
Channel vocoder synthesis of decoded output speech [137].

Enveloppe du conduit vocal obtenue par un banc de filtres passe-bande (16 à 19), la largeur des canaux augmente avec la fréquence..

Joint Speech Research Unit (JSRU) de U.K. : vocodeur canal à 2.4 kbits/s

→ DRT de 87 et 81 en présence d'erreurs de transmission de 5% (robuste !)

Vocodeur Formant

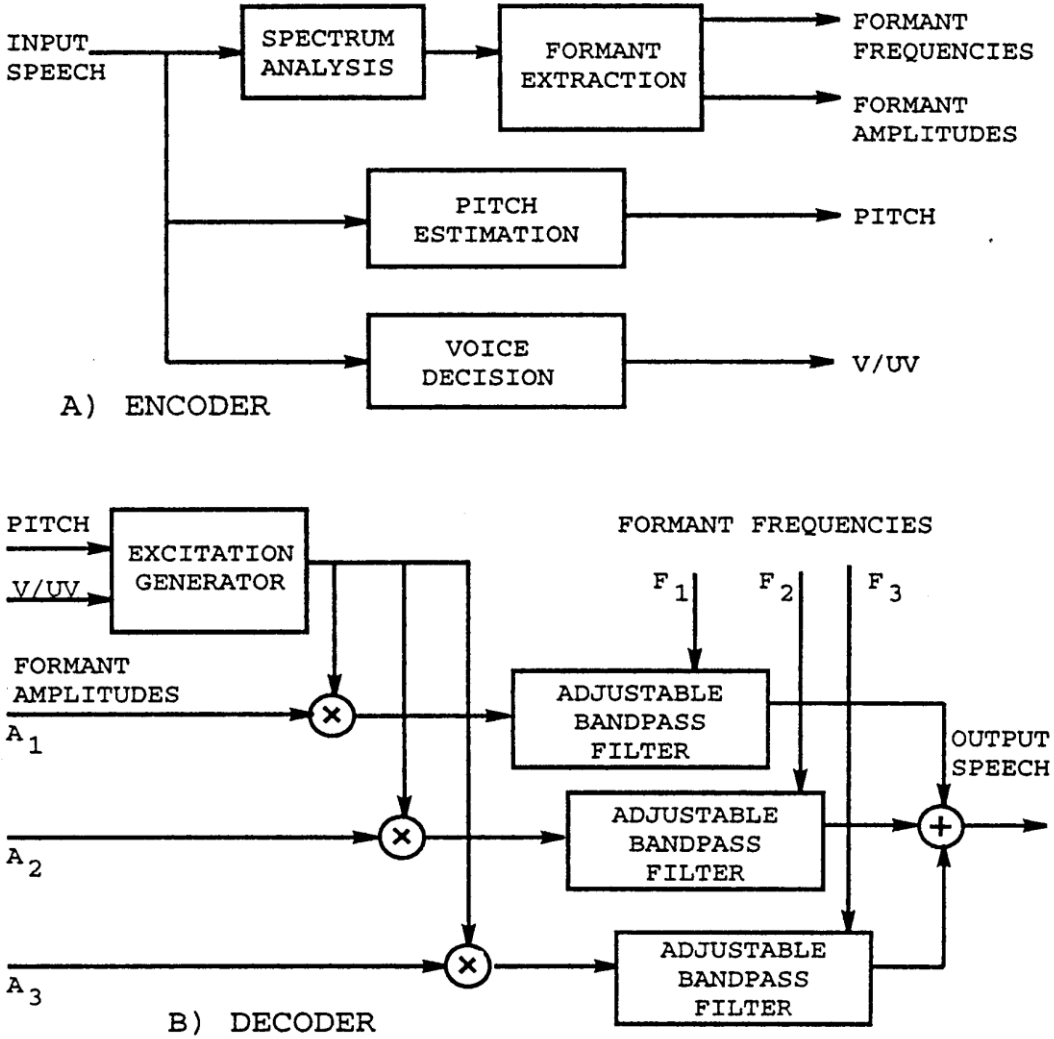


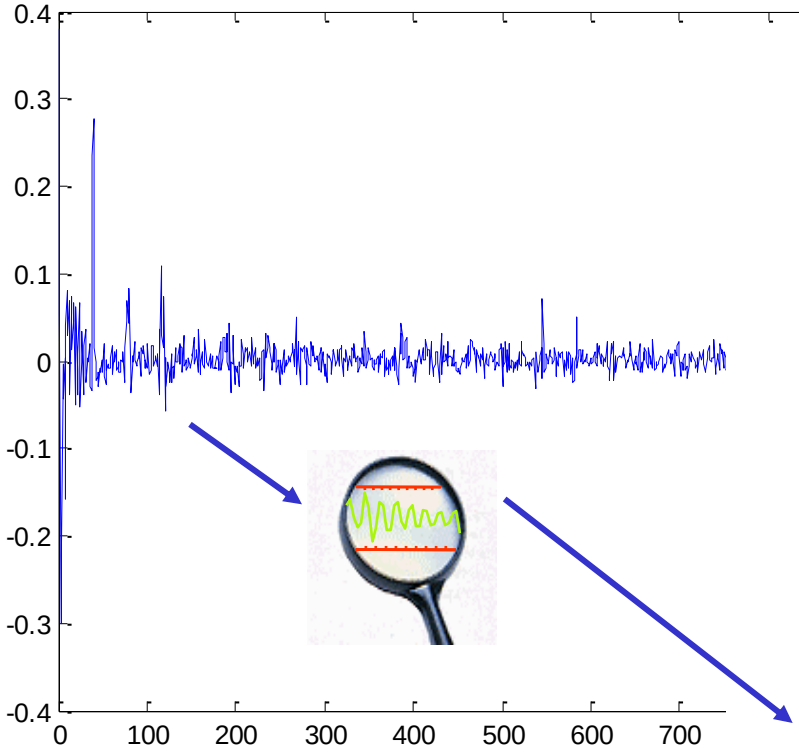
FIGURE 9.3
Formant vocoder analysis and synthesis [143].

Voisé

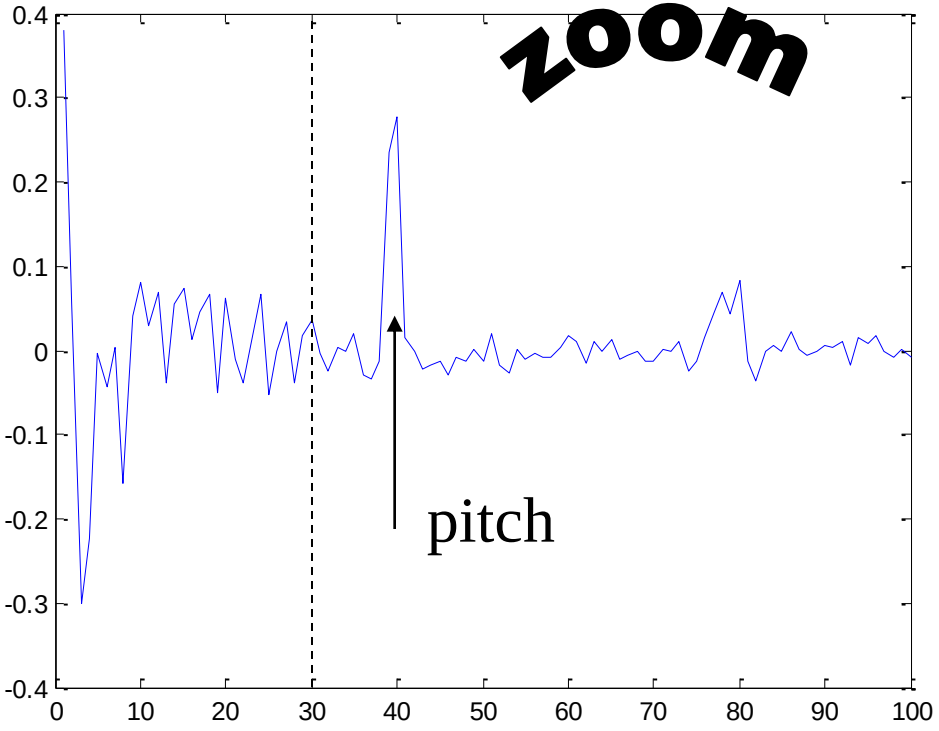
Vocodeur formant :
principale différence :
les caractéristiques de résonance des bancs de filtre
suivent les formants

Vocodeur Homomorphique

Idée de base : utilisation du cepstre : TF $^{-1}$ (log(A(f))), A(f) spectre d'amplitude
les « quéfrences » (échantillons du cepstre) proches de l'origine sont associées au conduit vocal

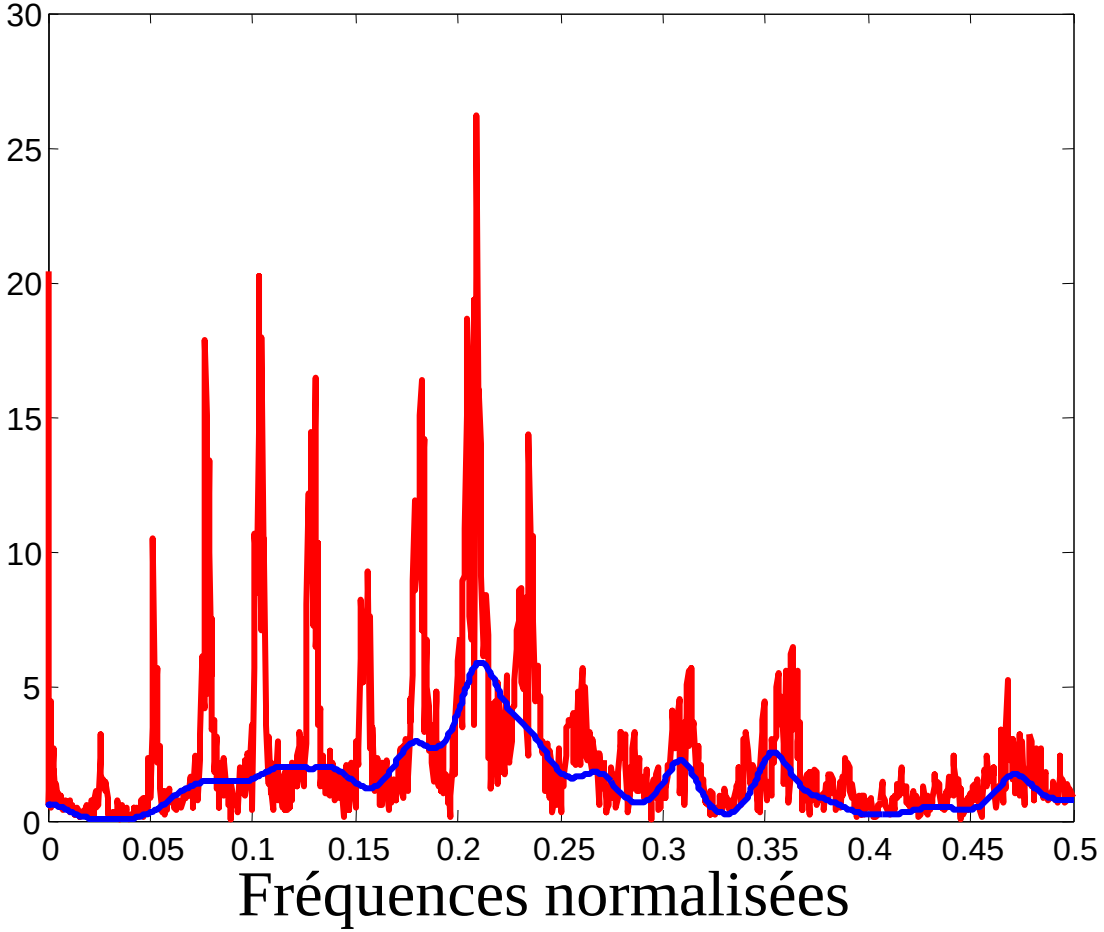


Cepstre de « *ne* »



« *quéfrences* »

Densité Spectrale de « *ne* » en rouge
Filtre homomorphique en bleu



Vocodeurs à base de prédiction linéaire

vers LPC10

Conduit vocal = filtre tout pôle H(z) de la forme ***1 / (1-A(z))***
Estimation des coefficients de ***A(z)*** par prédiction linéaire et minimisation de la MSE : équations de Toeplitz, de Yule-Walker
Beaucoup d'algorithmes existent : Levinson-Durbin récursif en ordre par ex.

On cherche les coefficients AR (ou LPC) tels qu'ils minimisent la puissance de l'erreur de prédiction

$$s(n) = \sum_{k=1,\ldots,p} a_k \, s(n-k) + e(n)$$

e(n) : EPL (LPE) ou LP residual

Remarque : si on transmet les coefficients AR et la LPE, synthèse de la parole ***exacte*** *(principe du DPCM)*
ICI, on ne transmet pas e(n), uniquement le filtre +modèle à 2 états : au décodeur e(n) est soit un train d'impulsions, soit un bruit blanc

LPC : fenêtre d 'analyse de 20-30 ms, paramètres recalculés toutes les 10-30 ms.
+fenêtre sous-divisée en 5ms : paramètres obtenus par interpolation linéaire des paramètres des fenêtres précédentes.

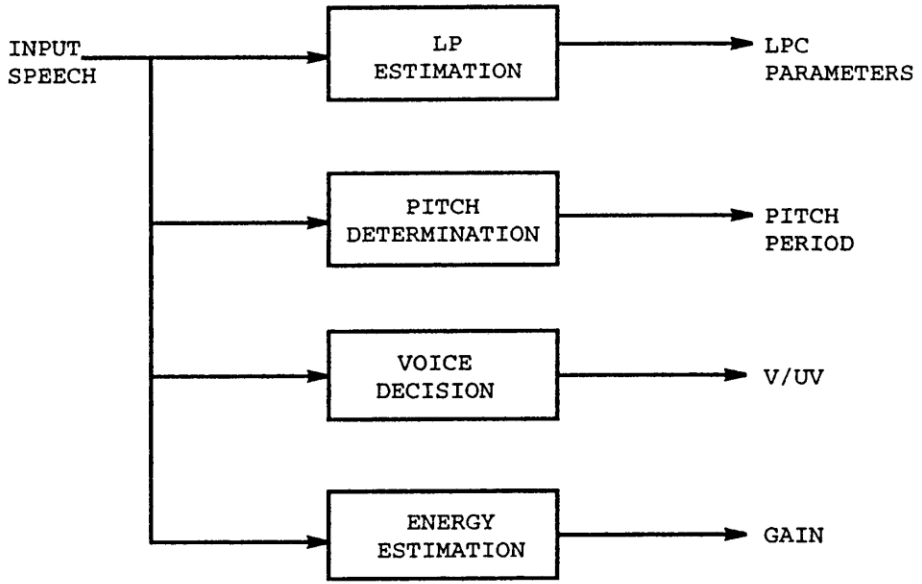


FIGURE 9.7
Linear predictive coding (LPC) encoder.

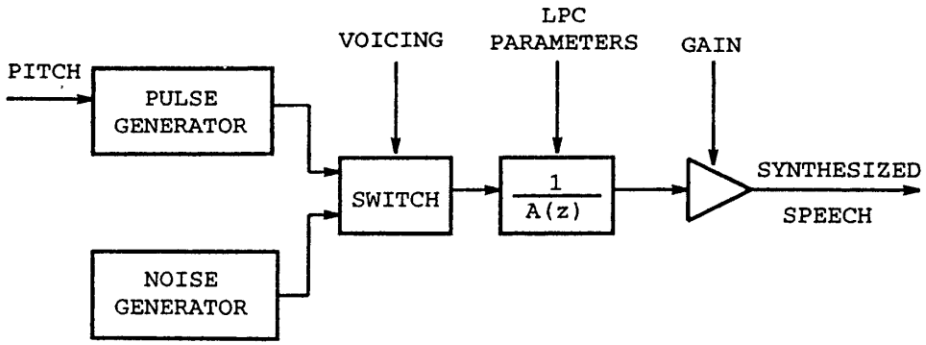


FIGURE 9.8
LPC decoder.

Aspect très étudié du LPC : *la quantification des coefficients*

En général, p = 8 à 14 (10 le plus souvent)
• Quantification directe des coefficients de prédiction déconseillée
• zéros de 1-A(z) ? Difficiles à calculer, non ordonnés, schémas statistiques de Q. difficiles

• **PARCOR** : coefficients partiels de corrélation (ou de réflexion - interprétés comme paramètres physiques du modèle de tube acoustique de la parole) : ordonnés, à dynamique fixe

• **Log Area Ratio** (LAR) : transformation des PARCOR moins sensible à la Q. souvent utilisés

$$LAR(m) = \log \{ (1+ k_m) / (1- k_m) \}$$

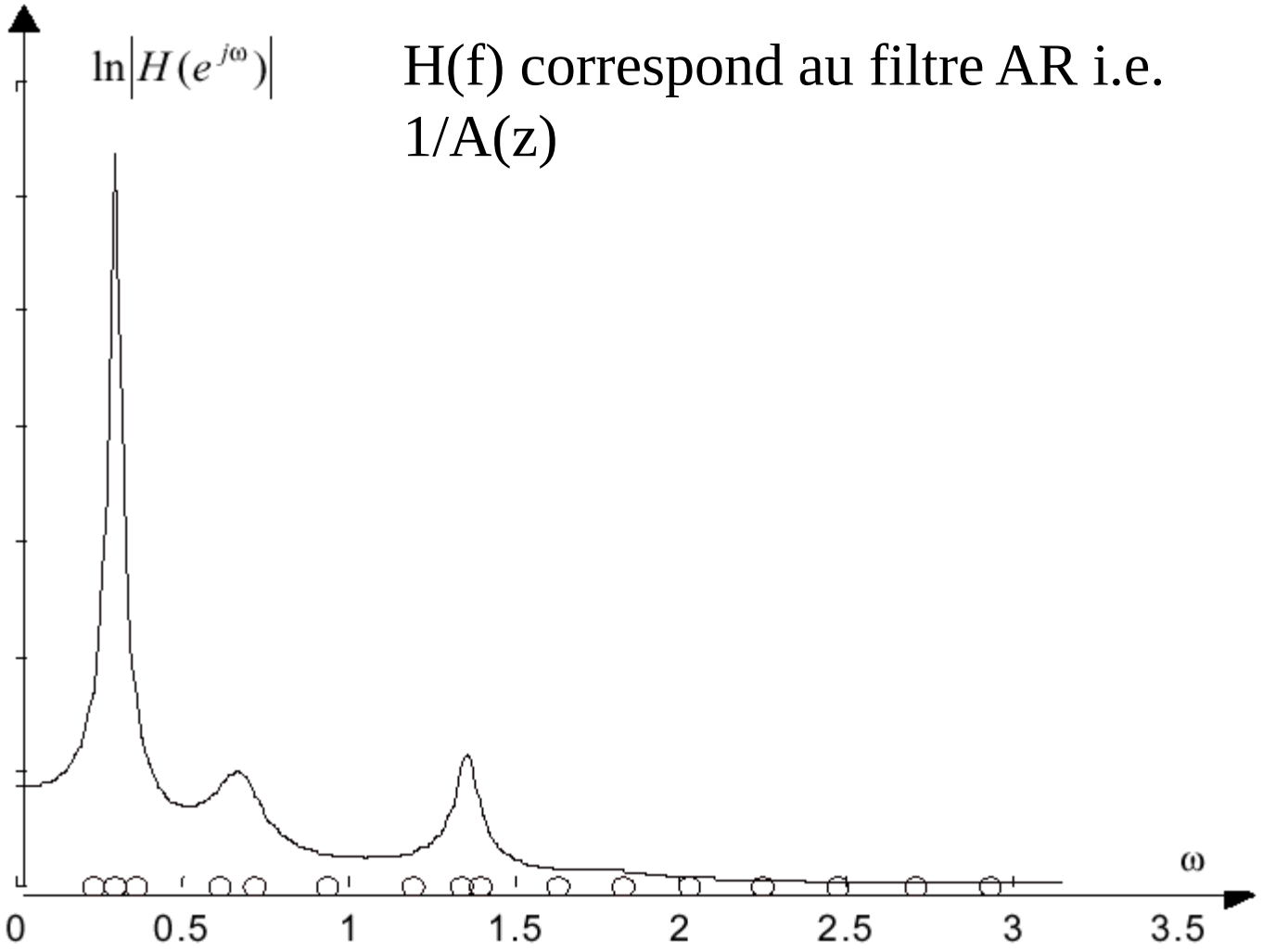
• Transformation en sinus inverse : $\arcsin(k_m)$ (inverse sine transformation)

• **LSP : Line Spectrum Pairs** largement utilisés (ou **LSF Line Spectrum Frequencies**)
exemple : ordre p : $A(z) = 1+ a_1 z^{-1} +\ldots+a_p z^{-p}$
on définit 2 polynômes : $A_1 (z) = A(z) + z^{-p-1} A(z^{-1})$ et $A_2 (z) = A(z) - z^{-p-1} A(z^{-1})$
chaque polynôme représenté par p/2 fréquences de zéros : LSP

$$A(z) = (A_1 (z) + A_2 (z)) / 2$$

intéressant car lien entre fréquence LSP et formants mais coût calculatoire.

Interprétation physique des LSF (Line Spectrum Frequencies)



H(f) correspond au filtre AR i.e. 1/A(z)

O : Line Spectrum Frequencies

Estimation du pitch

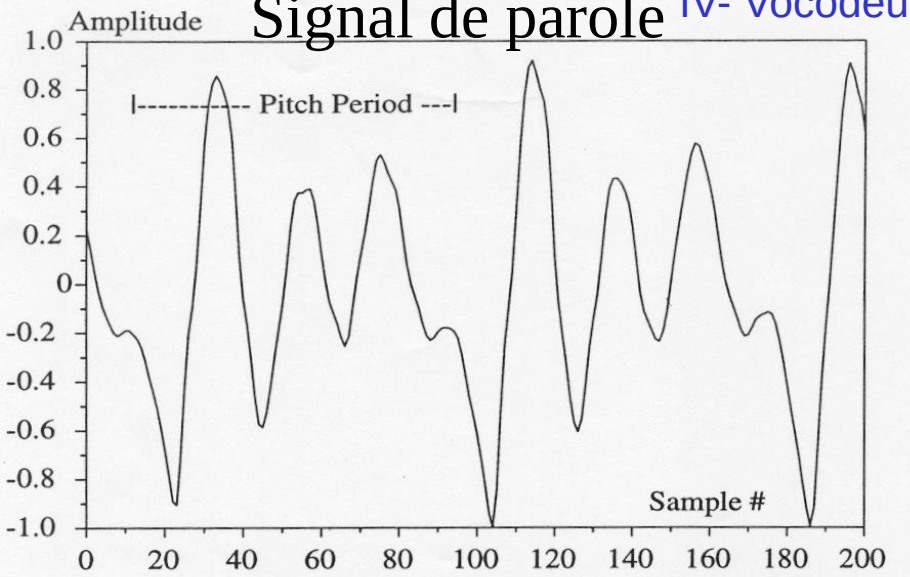
• pitch = propriété fondamentale de la parole voisée,
• se retrouve aussi en musique.

Ouverture et fermeture périodique de la glotte, donnant un caractère périodique à l'excitation
De l'ordre de 50 à 300 Hz (bas pour homme, haut pour femmes)
Information de prosodie portée par montée et descente du pitch

Estimer le pitch ? Plus difficile qu'il ne semble…
seulement quasi-périodique
Réf : W.Hess *Pitch determination of Speech signals*, Springer-Verlag, NY, 1983.
Approches les plus populaires : par autocorrélation.
Méthodes temporelles, fréquentielles

estimation du pitch : par corrélation

Signal de parole



Autocorrélation court-terme :

$$r(k) = \sum_{m=0}^{N-1-k} s(m)\, s(m+k)$$

ex : Fe=8000 Hz

-> pitch = 98.8 Hz

(peak-picking)

Autres formules de corrélation (« cross-correlation », Corrélation normalisée…)

Autocorrélation

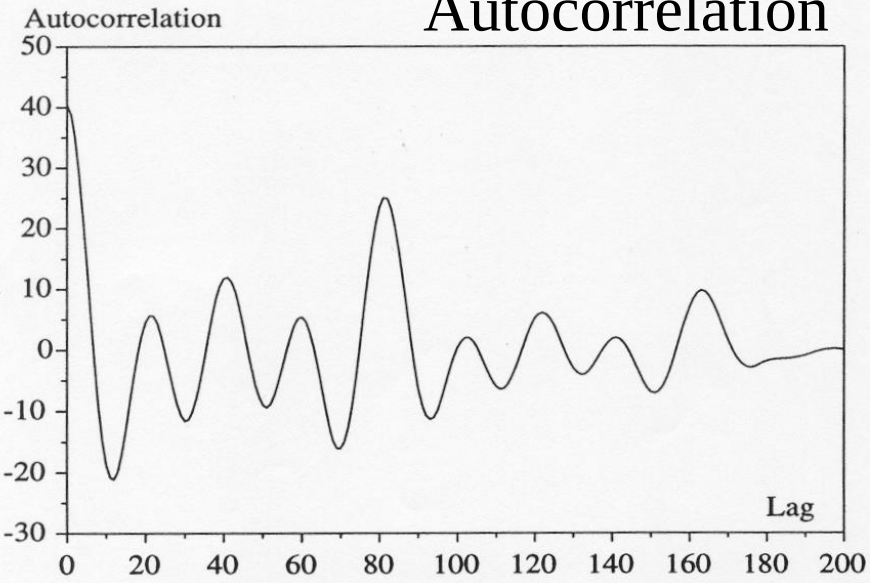


FIGURE 5.1
Time-domain waveform and autocorrelation of a short segment of voiced speech.

estimation du pitch : par corrélation

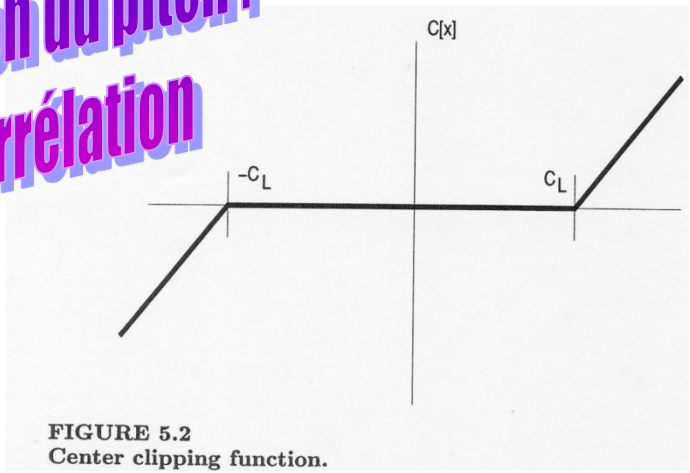
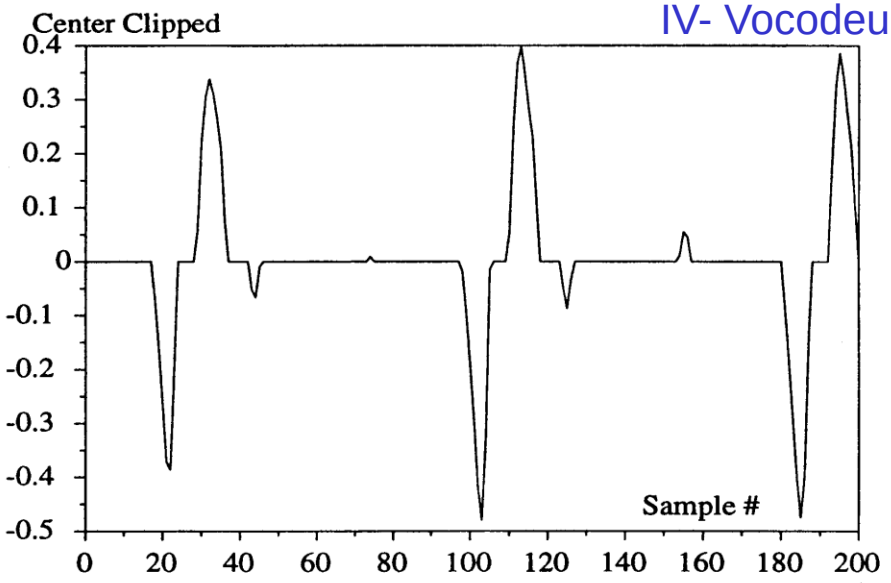


FIGURE 5.2
Center clipping function.



Autocorrélation de la parole « center-clipped »

pour éviter les maxima multiples dans l'autocorrélation (parole non purement périodique)

>> « center clipping the speech » avant de calculer l'autocorrélation

C_L : % fixé de l'amplitude max du signal de parole (30% par ex)

- à éviter en contexte bruité
- pour segments dont l'énergie varie rapidement, difficile de fixer le niveau de « clipping »

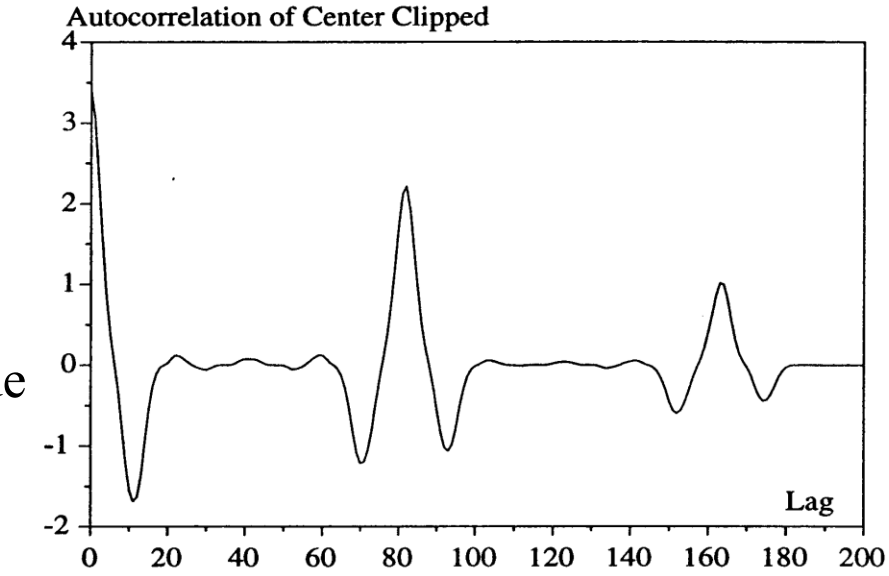
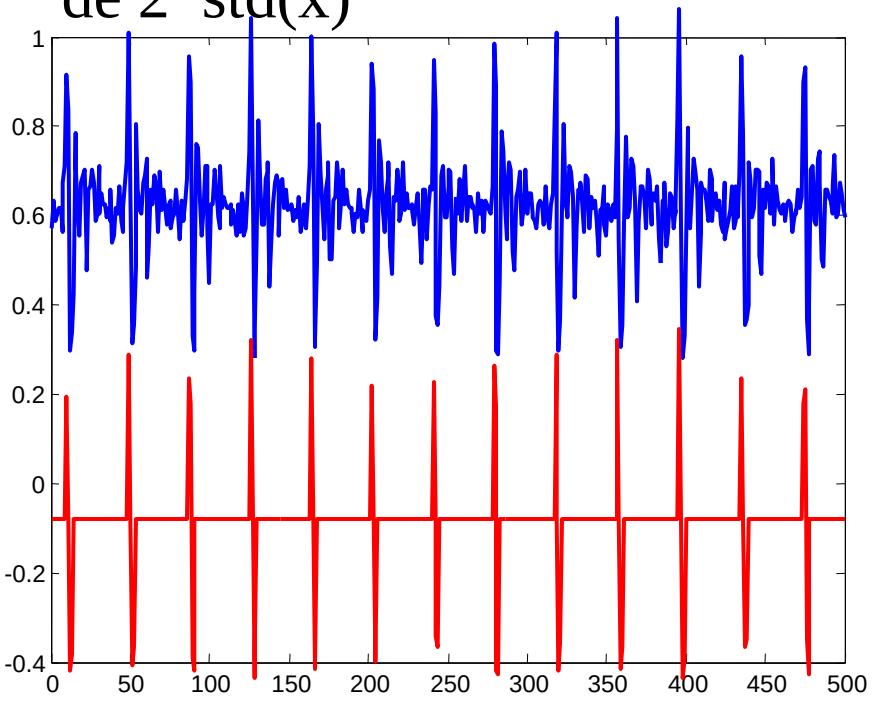
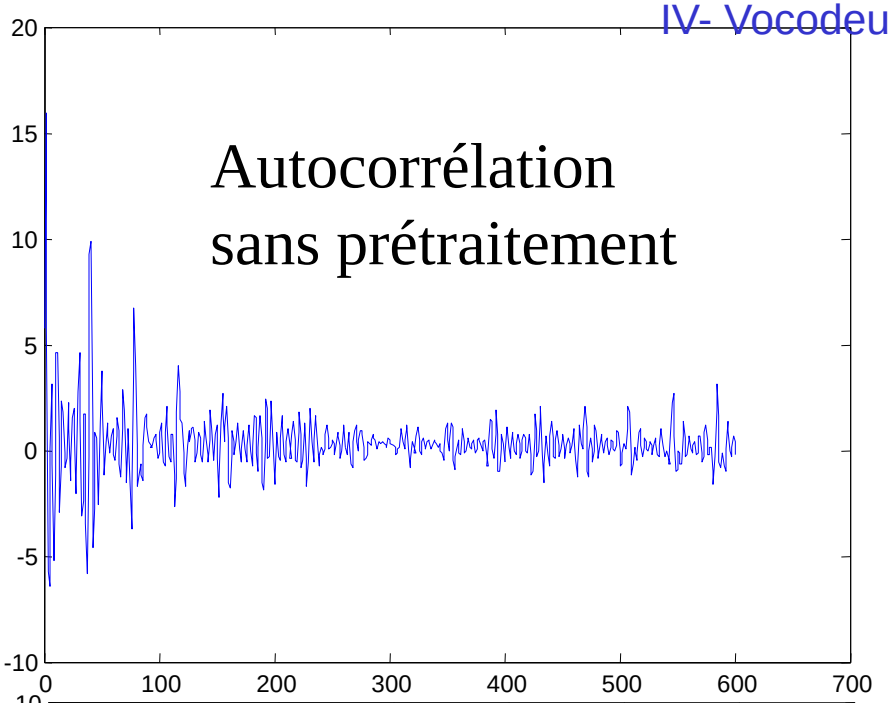


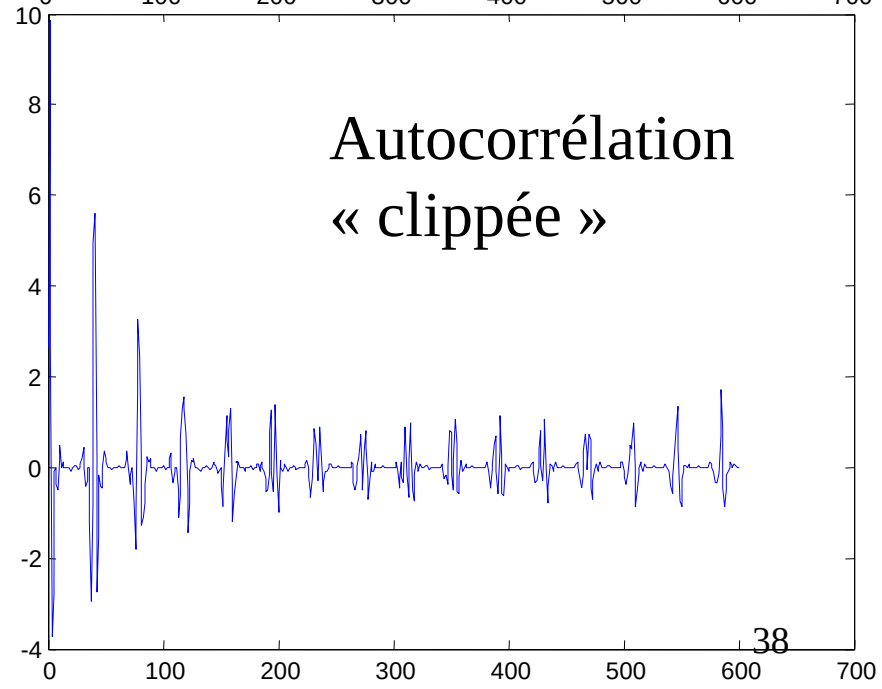
FIGURE 5.3
Center-clipped waveform and autocorrelation of a short segment of voiced speech.

Intérêt du Clipping dans l'estimation du pitch : exemple

Clipping avec un seuil de 2*std(x)





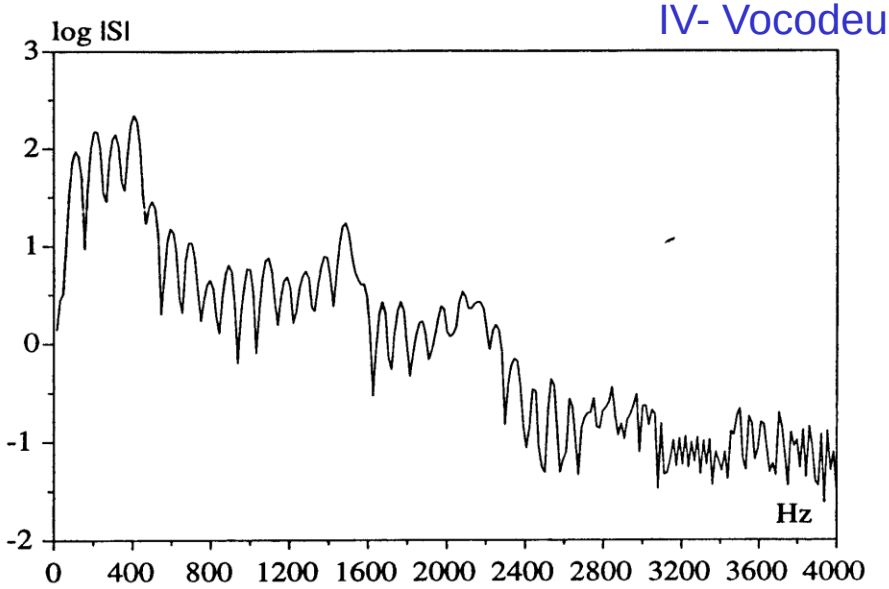


estimation du pitch : par le cepstre

Signal pseudo-périodique
-> spectre TF court terme :
ondulations dues à la structure harmonique
visible sur spectres de sons voisés
visible sur le cepstre :
pic à la quéfrence d = 1/fréq.
Fondamentale
cepstre = IFFT (log10 |FFT(s(n))|)

pour parole non voisée,
pas de pic dominant

peut servir
à la décision voisée / non voisée



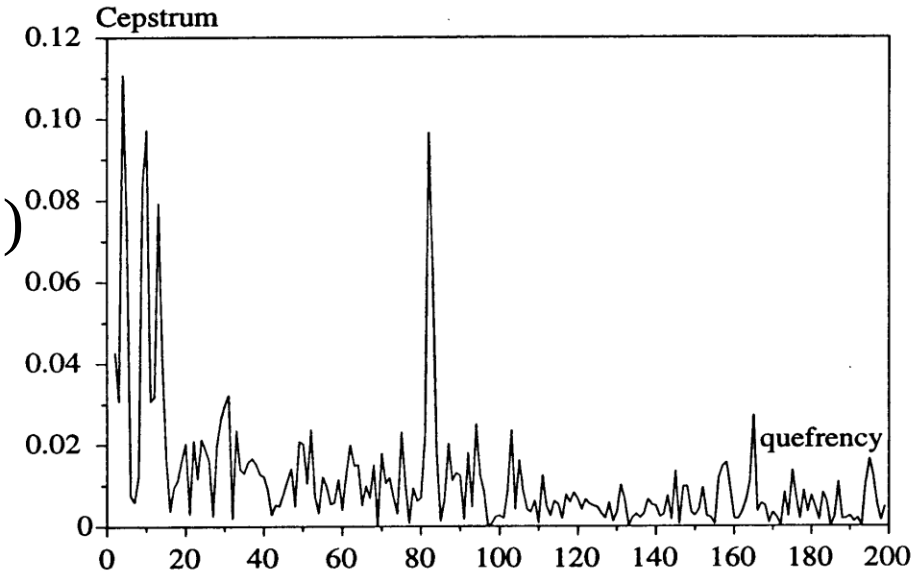


FIGURE 5.6
Log magnitude of DFT and cepstrum of speech segment of Figure 5.1.

IV- Vocodeurs (boucle ouverte)

Exemple 1 - LPC10 (cf LPC)

LPC 10 : standard à 2.4 kbits/s Federal Standard FS-1015 (1976)

utilise un prédicteur d ’ordre 10.
Parole voisée : analyses faites tous les multiples du pitch
Parole non voisée : analyses faites toutes les 22.5 ms
Voisement et pitch estimés sur parole filtrée passe-bas à 800 Hz
Estimation du pitch basée sur AMDF (Average Magnitude Difference Function)

$$AMDF(\tau) = \Sigma \mid s(n) - s(n+\tau) \mid \text{ (autocor simplifiée)}$$

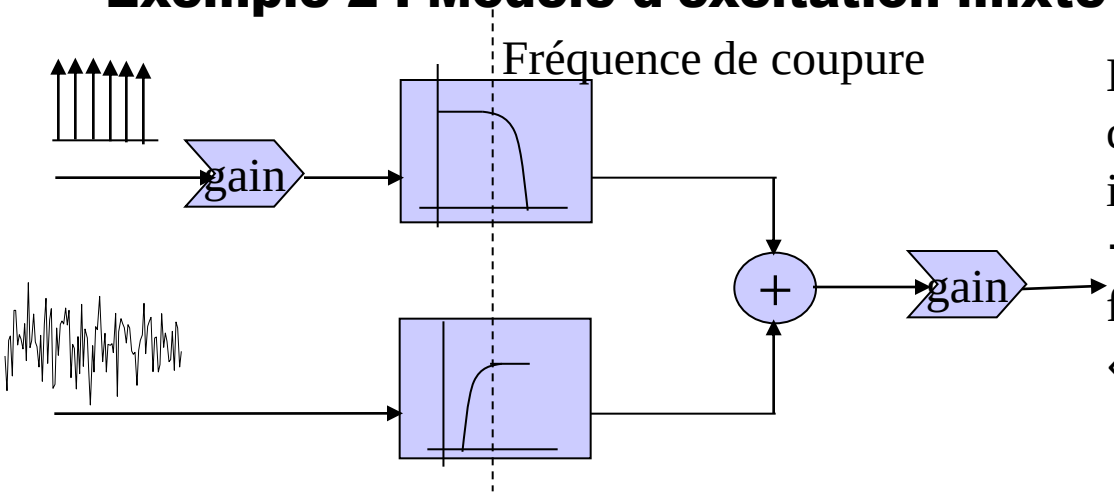
pitch et voisement codés sur 7 bits
gain : 5 bits / tranche d’analyse
segments voisés / non : 10 / 4 coefs de réflexion codés
(k_1 et k_2 par LAR sur 5 bits chacun) DRT de 90

originaux

FS-1015

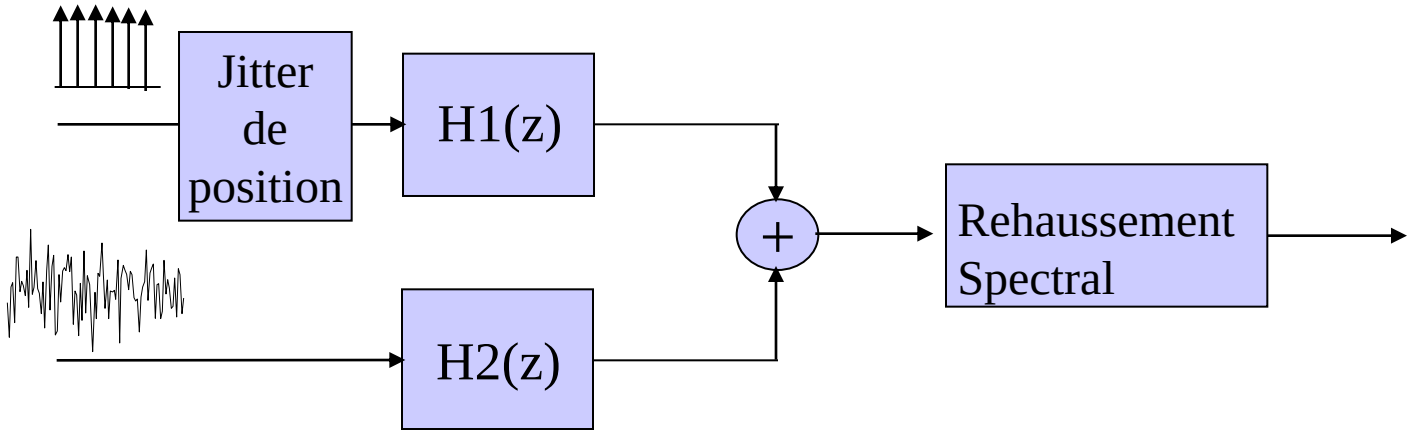
IV- Vocodeurs (boucle ouverte)

Exemple 2 : Modèle d'excitation mixte - MELP



Erreurs de voisement dans le LPC classique à 2 états dégradent qualité et intelligibilité parole
→modèle d'excitation mixte (Makhoul)
fréquence de coupure estimée par algo « peak picking »

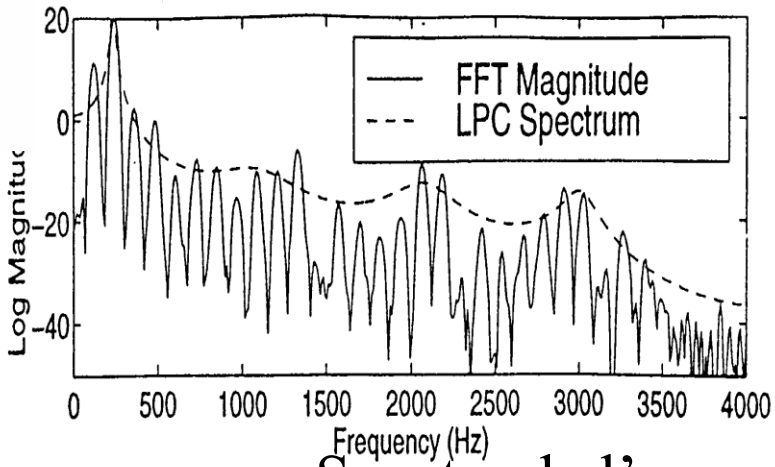
Modèle plus élaboré de McCree et Barnwell

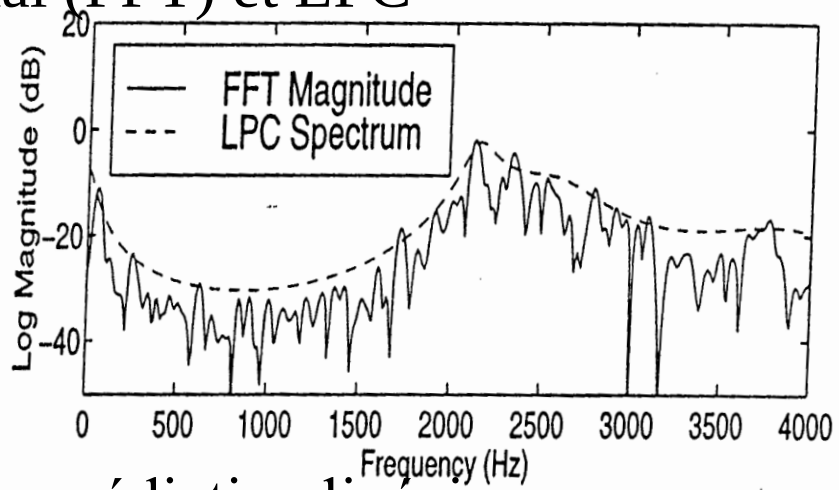


TMS320c30 : 2.4 kbits /s DAM de 58.9 parole sans bruit, 41 parole bruitée

Exemple 3 : Residual Excited Linear Prediction (RELP)

Spectre du signal (FFT) et LPC





Spectre de l'erreur de prédiction linéaire

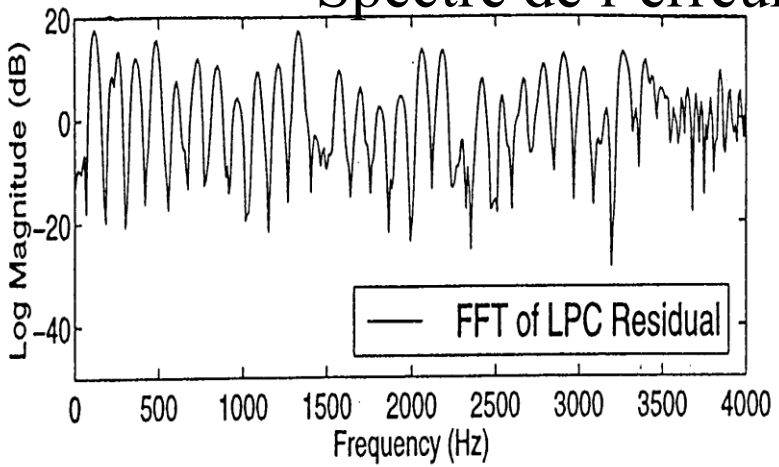


FIGURE 9.5
LP spectrum and residual spectrum for voiced speech frame.

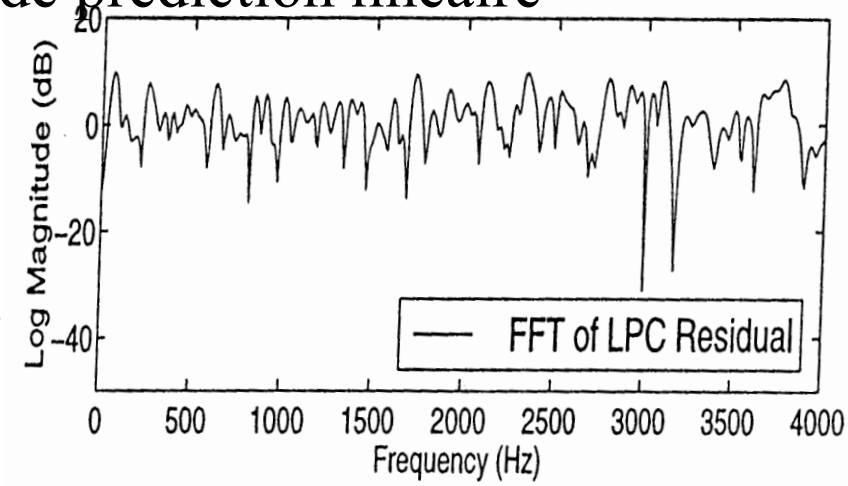


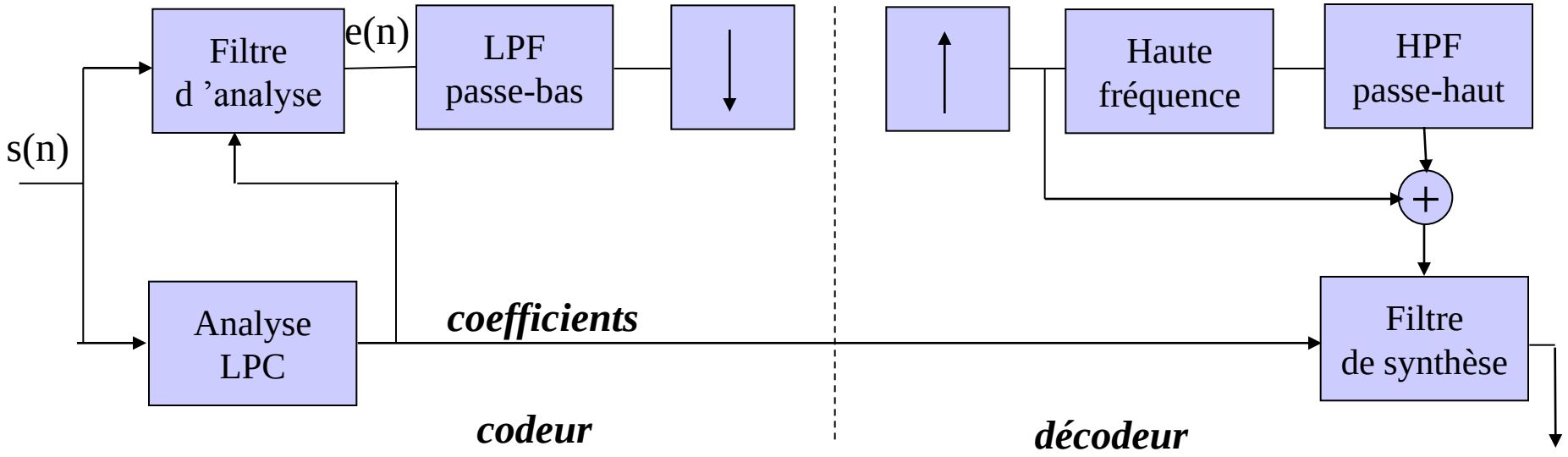
FIGURE 9.6
LP spectrum and residual spectrum for unvoiced speech frame.

Idéalement, EPL blanche mais…

IV- Vocodeurs (boucle ouverte)

Residual Excited Linear Prediction (RELP)

ADPCM : on transmet tout ≠ LPC : on ne transmet que le voisement, le pitch et le gain.
↳ Transmettre le résidu améliore bien sûr !
mais info la plus importante < 1000 Hz : pour son naturel, énergie des sons voisés
RELP : LPC fournit PARCOR codés qui sert par filtrage inverse à produire le résidu qui est filtré (1000 ou 800 Hz), décimé (4 ou 5) et codé (ADM)
Inutile de transmettre pitch et gain car contenus dans résidu : évite calcul du pitch et voisement



Reconstituer la partie HF par repliement spectral (interpolation le provoque)
ou translation spectrale (« copy up »)
codeur RELP 4800 bits/s qualité comparable au PCM à 24-32 kbits/s
RELP entre 4.8 et 13 kbits/s

Codage de la parole

I - Le Contexte
II- Codage temporel : waveform coders
III- Modèles analyse/synthèse sinusoïdaux
IV- Vocodeurs
V- Codeurs prédictifs linéaires :
Analyse par Synthèse

Intérêt de l'Analyse par Synthèse
MPLPC
RPEC
CELP

Analysis-and-synthesis

- Coded speech is not analysed.
- Errors accumulated from previous frames are not considered.

Analysis-by-synthesis

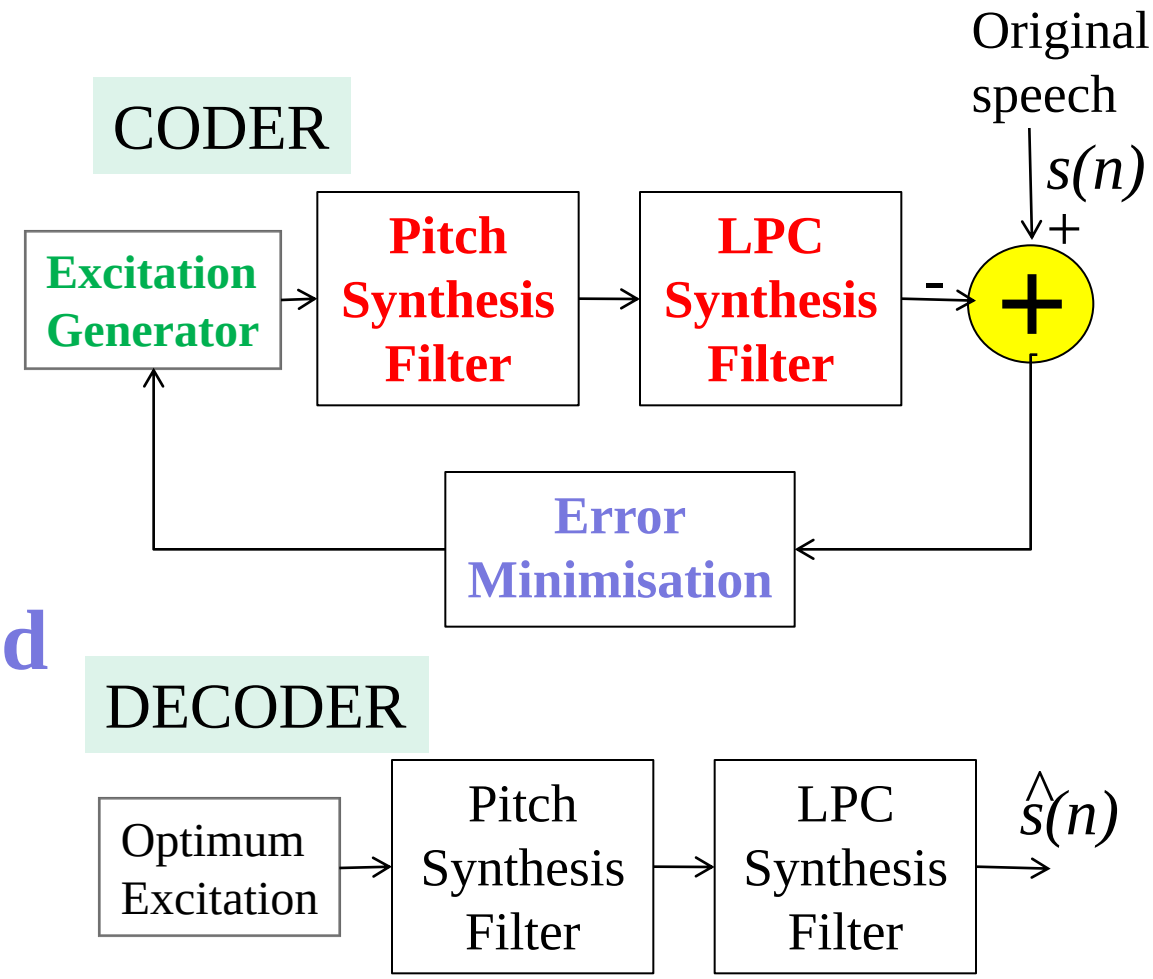
+ Far more succesful at 4.8-9.6 kb/s.

Analysis-by-Synthesis (AbS)

Time-varying filter

Excitation signal

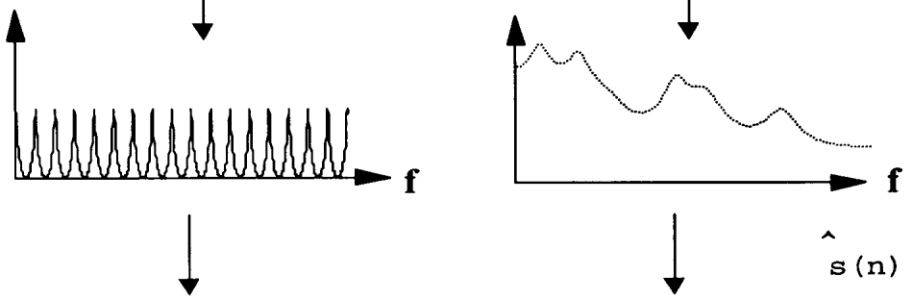
Perceptually based minimisation procedure



Long-term and Short-term predictors

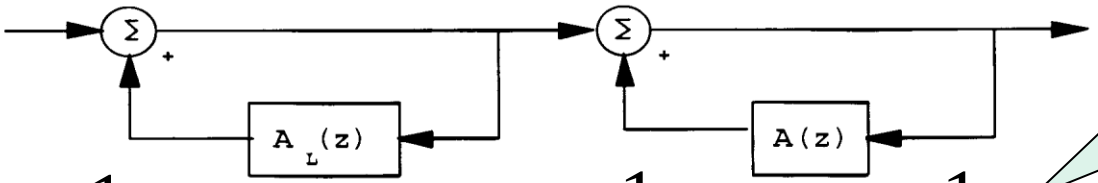
Déjà utilisé dans les waveform coders, ADPCM

$$\frac{1}{1 + a_p z^{-p}} \quad x \quad \frac{1}{1 - \sum_{i=1}^{10} a_i z^{-i}}$$



Estimated every 5-15 ms

Estimated every 10-30 ms



$$\frac{1}{P(z)} = \frac{1}{1 - \sum_{i=-I}^{I} b_i z^{-(D+i)}} \qquad \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

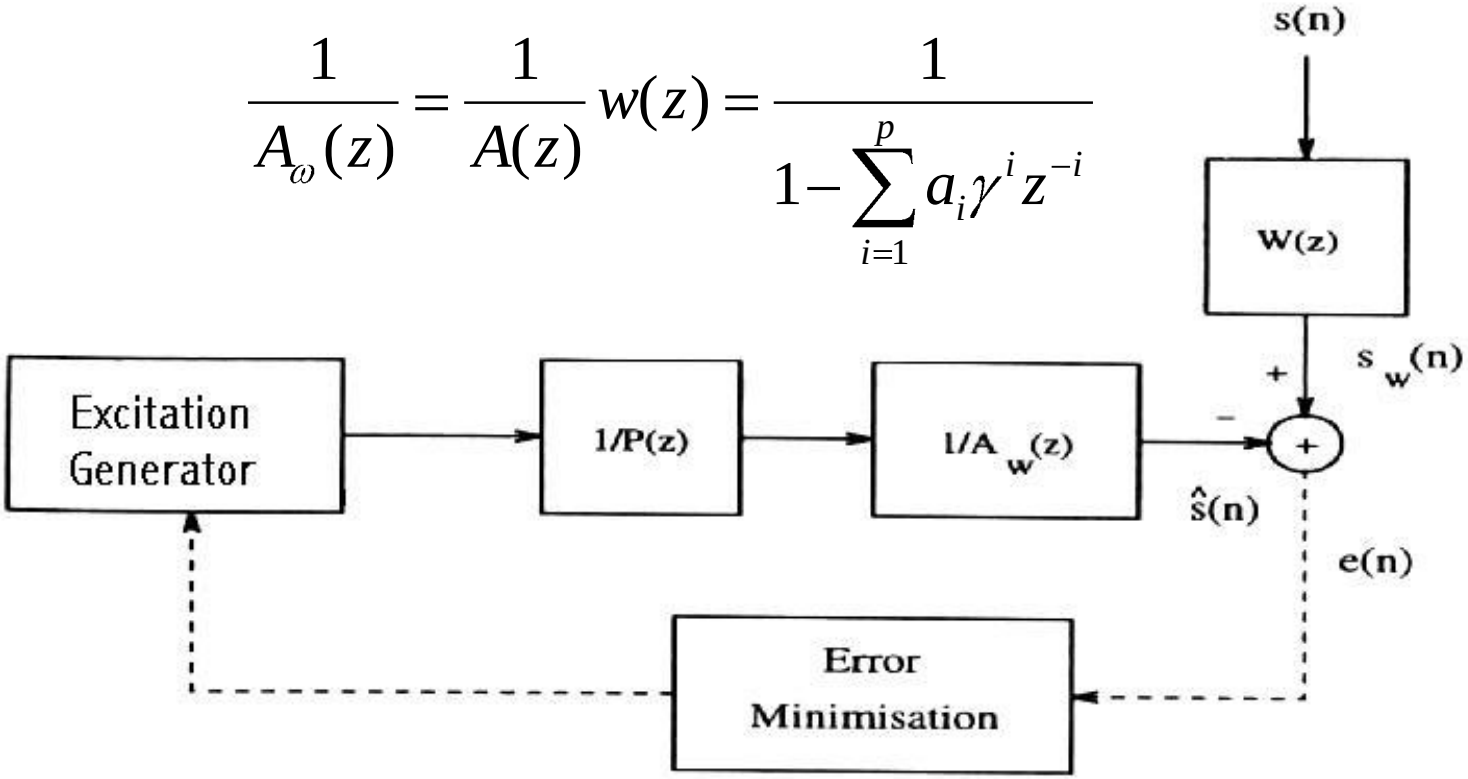
Perceptually based minimisation procedure

- MSE less meaningfull for low bit rates.
- Need a error criterium which is more in sympathy with human perception criterium.
- Use weighting filter.

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}} \qquad 0 \le \gamma \le 1$$

Weighting filter

$$\frac{1}{A_\omega(z)} = \frac{1}{A(z)} w(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}$$



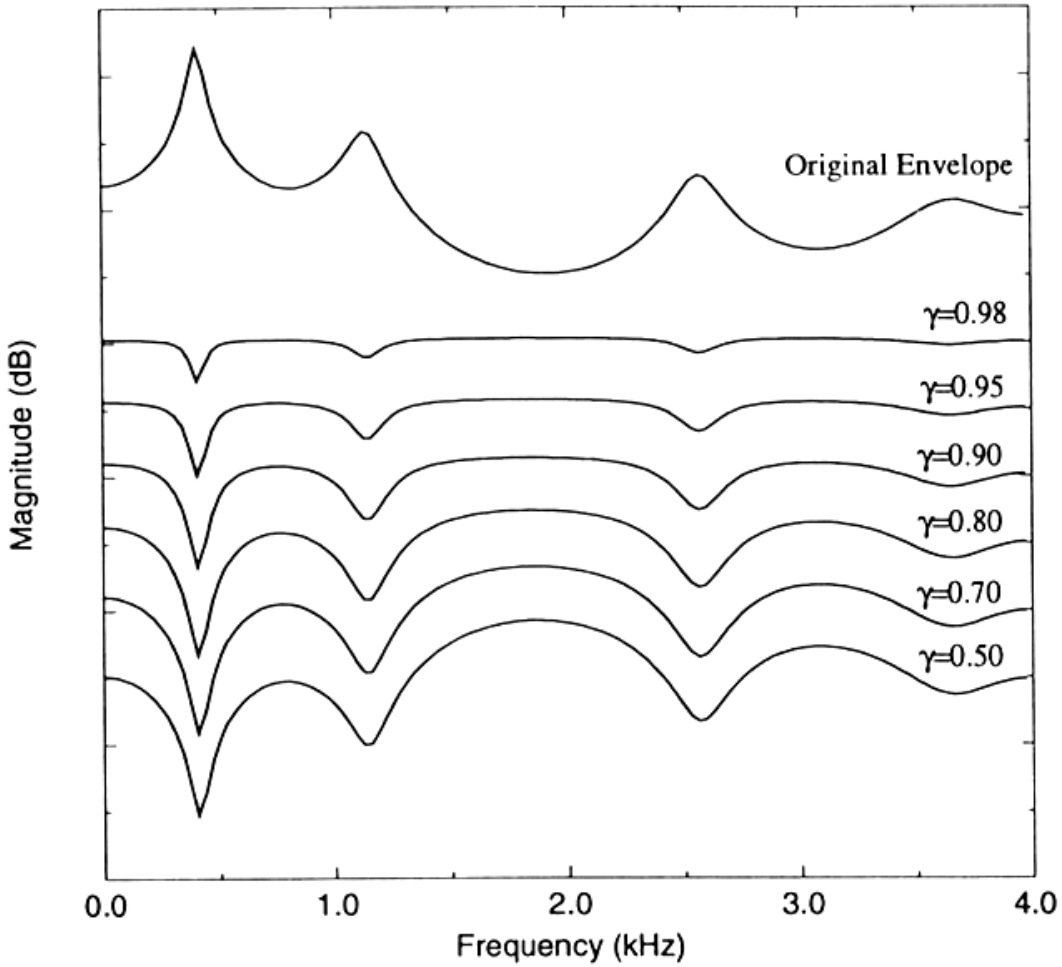


Figure 6.3 Typical plots of weighting filter spectra compared with the original speech envelope

Excitation signal

- Codebook excitation (CELP)
- Self-excitation (*adaptive codebook*) (SELP)
- Multi-pulse LPC (MPLPC)
- Regular pulse excited LPC (RPELPC)

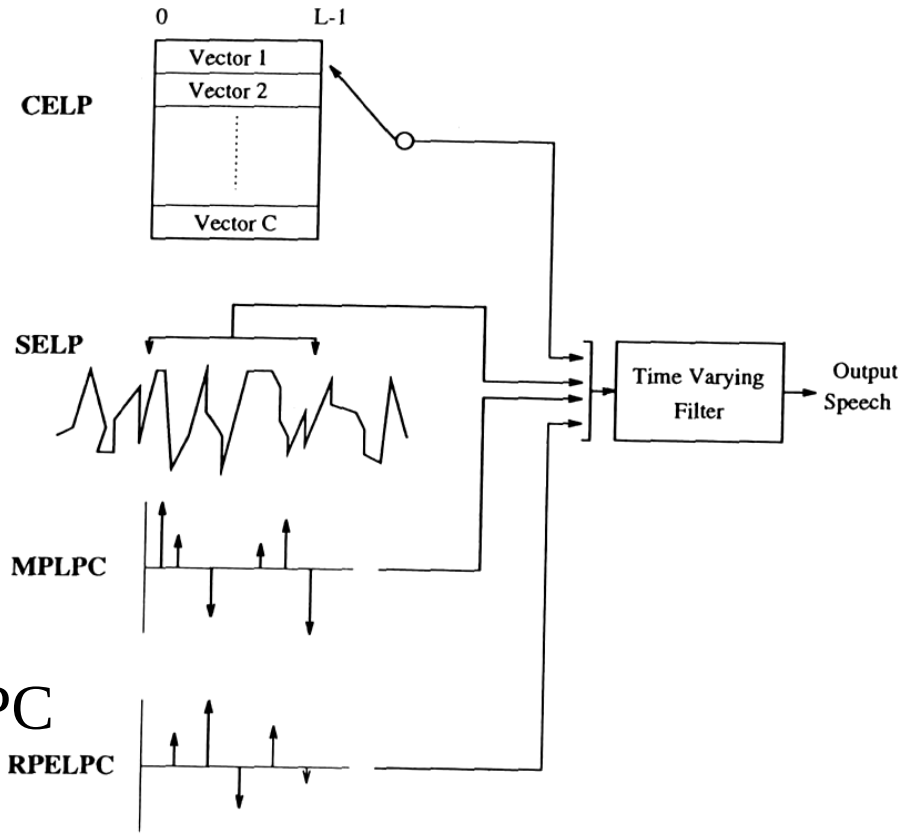
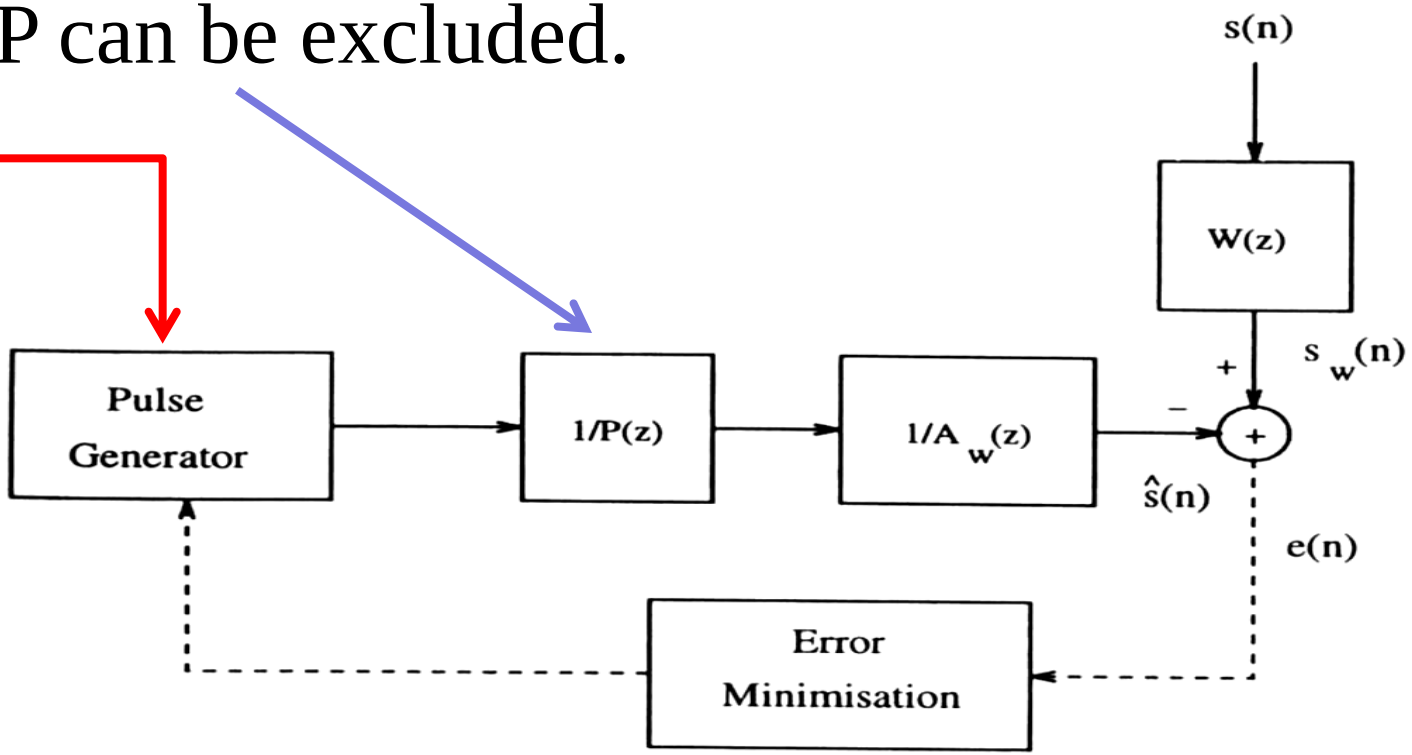


Figure 6.5 Generalised block diagram of AbS-LPC coder with different excitation types

Multi Pulse LPC (MPLPC)

- No distinction between voiced & unvoiced regions.
- Determine pulse locations and amplitudes.
- Minimization of the error signal.
- LTP can be excluded.



Search Methods

- Pulses are optimized one by one(c)
- Improvements:
  - Reoptimize the amplitudes when last pulse is determined(b)
  - Reoptimize the amplitudes after each pulse determination(Line a)

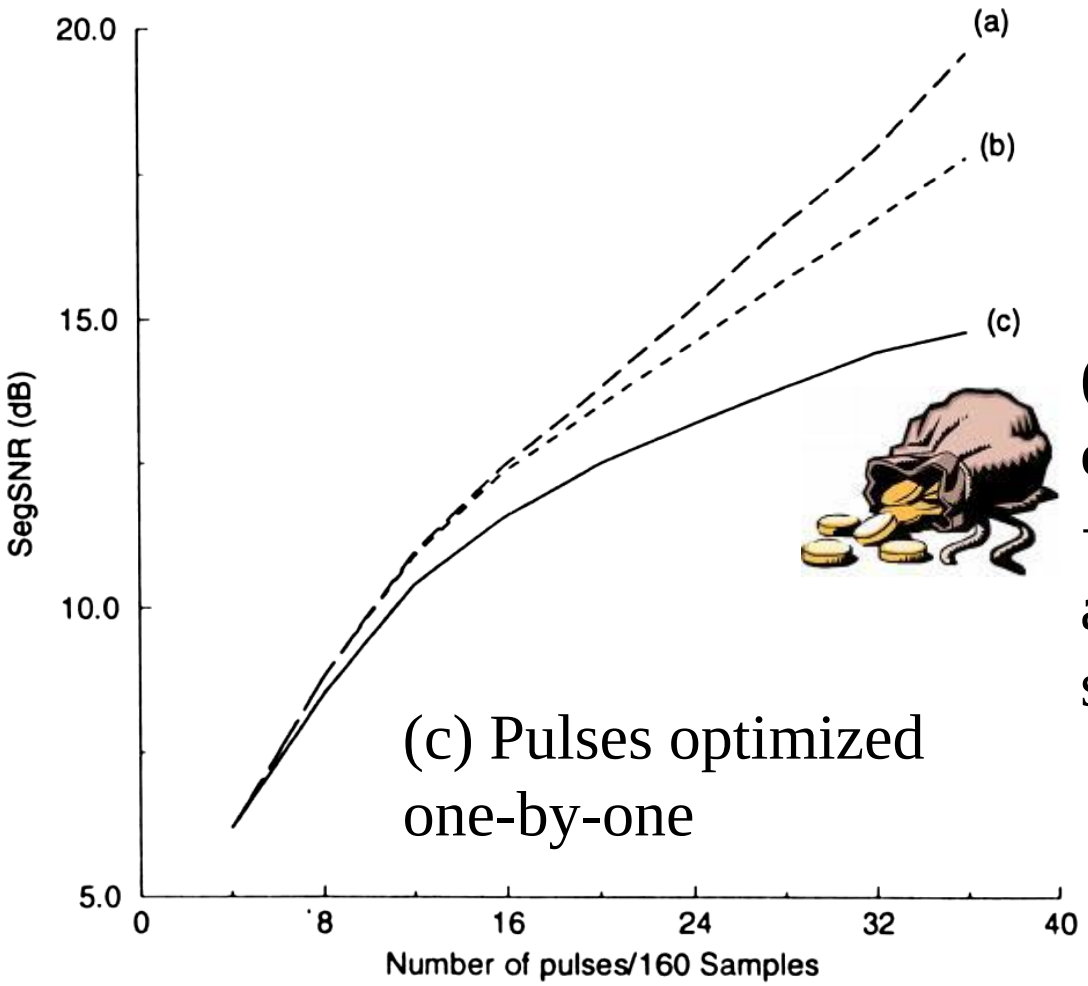
Sequential Search

$$E_w = \sum_{n=0}^{L-1} \left[\tilde{s}(n) - g_i h_w(n - m_i) \right]^2$$

$$\tilde{s}_{i+1}(n+1) = \tilde{s}(n) - g_i h_w(n - m_i)$$

Reoptimization or not

(a) Pulses optimized one-by-one
+ re-optimize amplitudes at EACH step



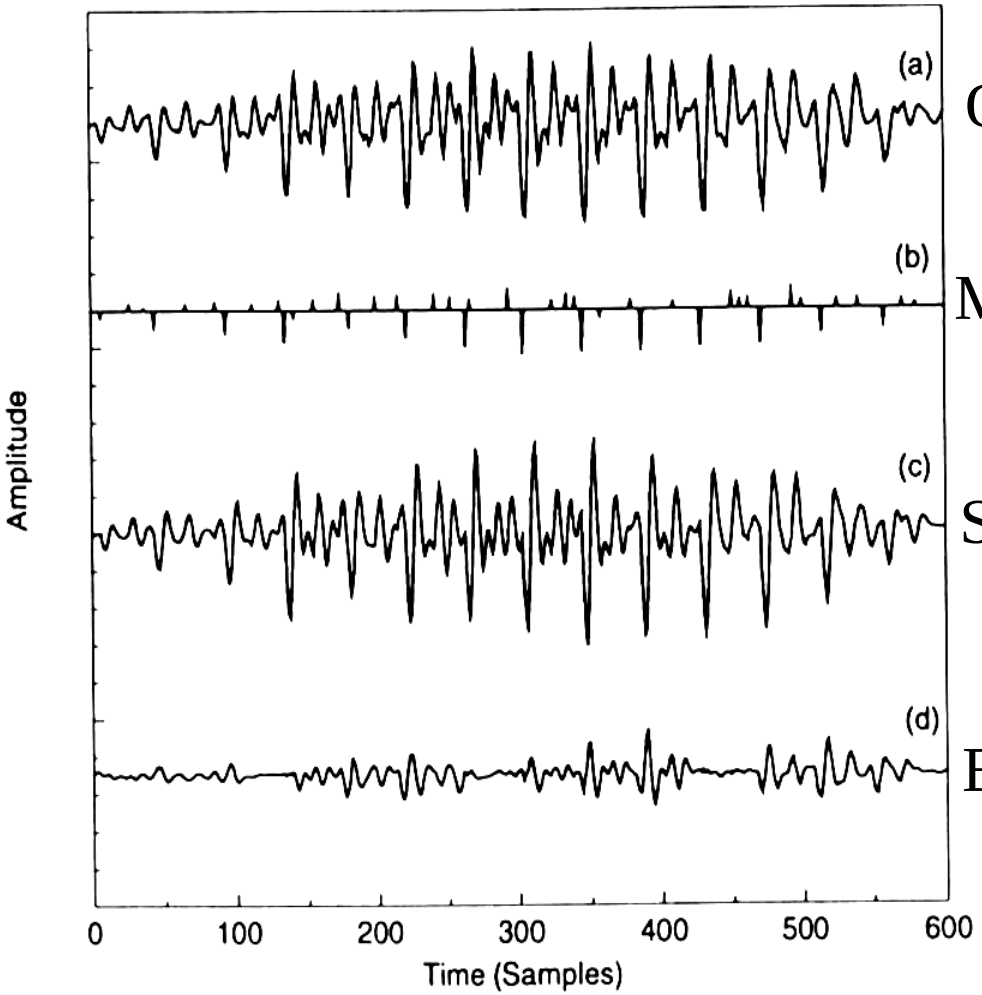
(b) Pulses optimized one-by-one
+ re-optimize amplitudes at the last step

(c) Pulses optimized one-by-one

Frame Size and Number of Pulses

- Large frames and more pulses for better performance.
- Small frames and less pulses for less computation.
- Around 40 samples per frame.
- 5 pulses per 4-5 ms.

MPLPC



Original Speech

Multipulse Excitation

Synthesized Speech

Error Signal

MPLPC with and without LTP

- LTP increases performance at:
  - Low bit rates.
  - In voiced regions with high pitch.

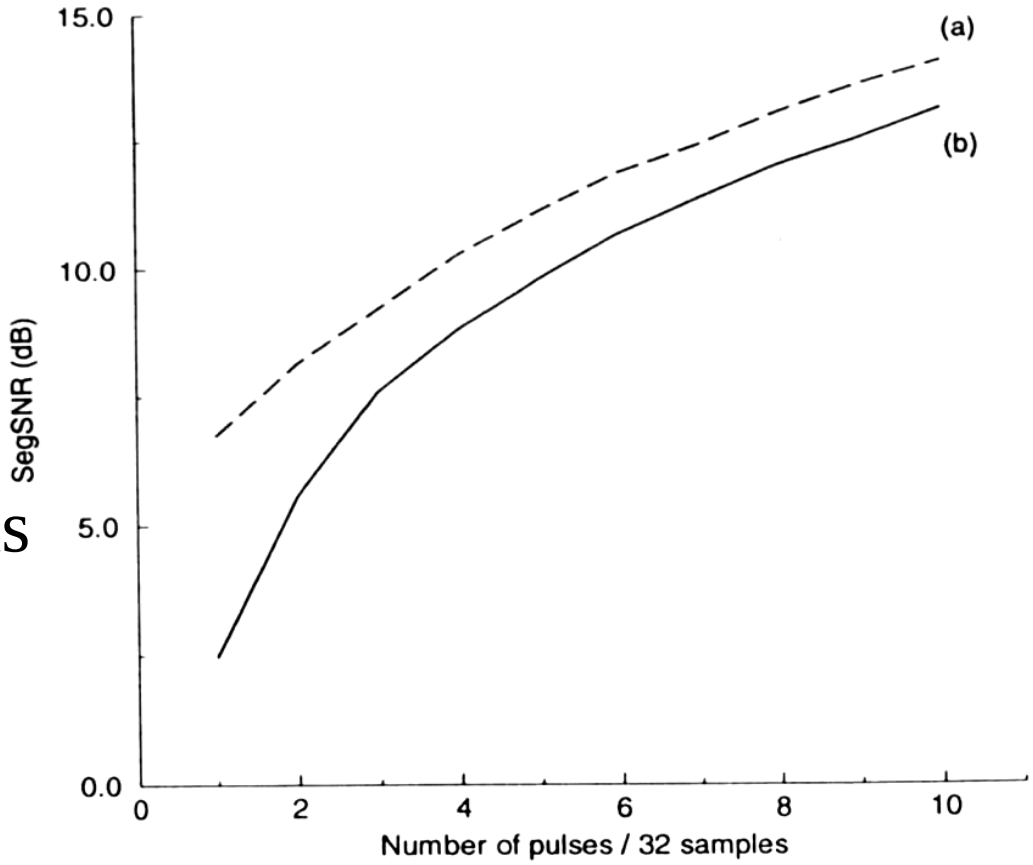
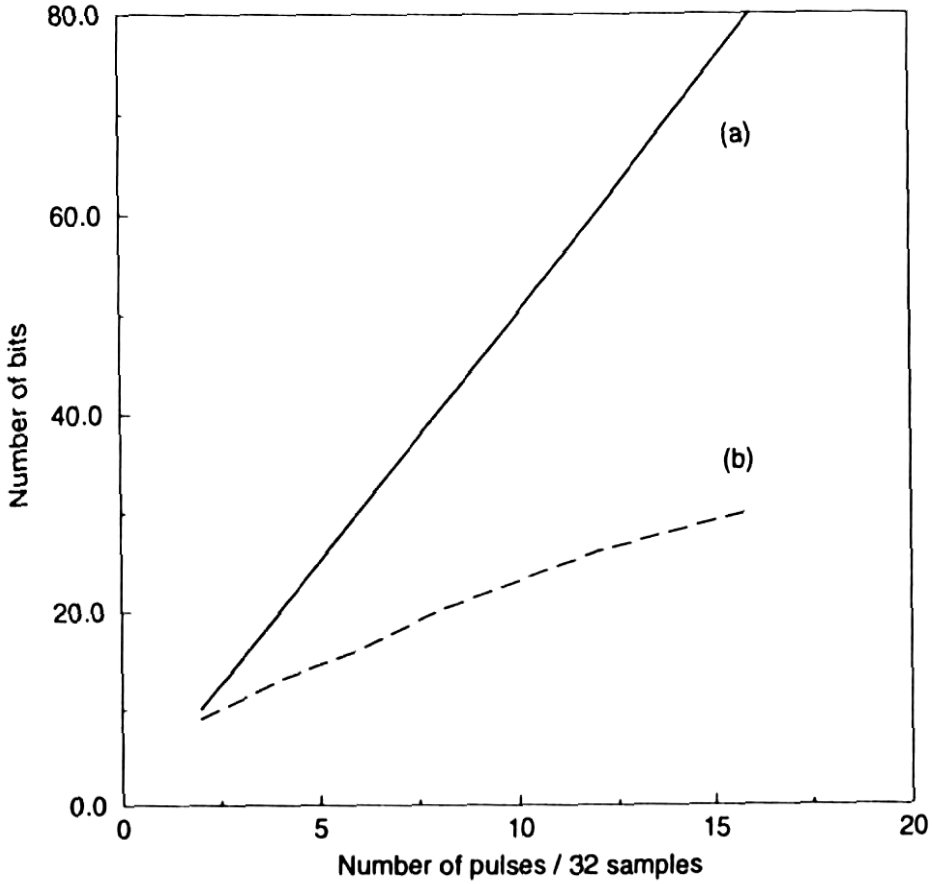


Figure 6.17 Performance of MPLPC (a) with and (b) without LTP

Pulse Position Coding

- Independent
  Every pulse position is coded individual.
- Combinational
  All different possibilities to place M pulses



Ex : C(32,10)=64 512 240 soit 26 bits

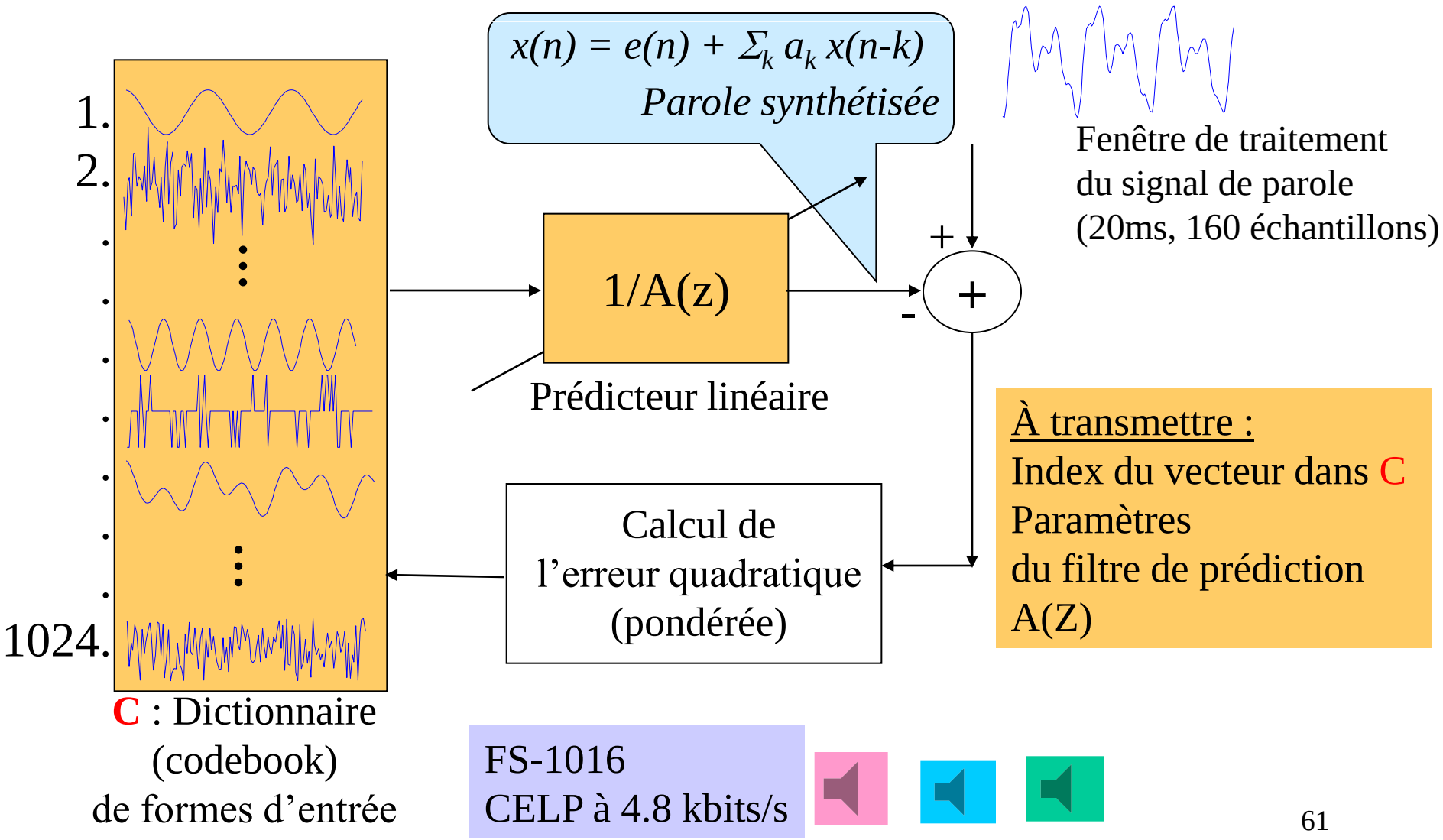
Pulse Amplitude Coding

- First amplitude quantized using a non-uniform quantiser.
- The rest of the amplitudes are normalised with the first amplitude and coded using fewer bits.
  - First pulse magnitude adaptive
  - Previous pulse adaptive
  - Previous subframe energy adaptive
  - LTP filter adaptive

Regular Pulse Excitation LPC (RPE LPC)

- The pulse position are predefined in a structured manner.
- Less computation extensive.
- A performance loss.

CELP : Code Excited Linear Predictor



CELP
avec
dictionnaire
adaptatif
(2 dictionnaires,
1 fixe,
1 adaptatif)
ACELP

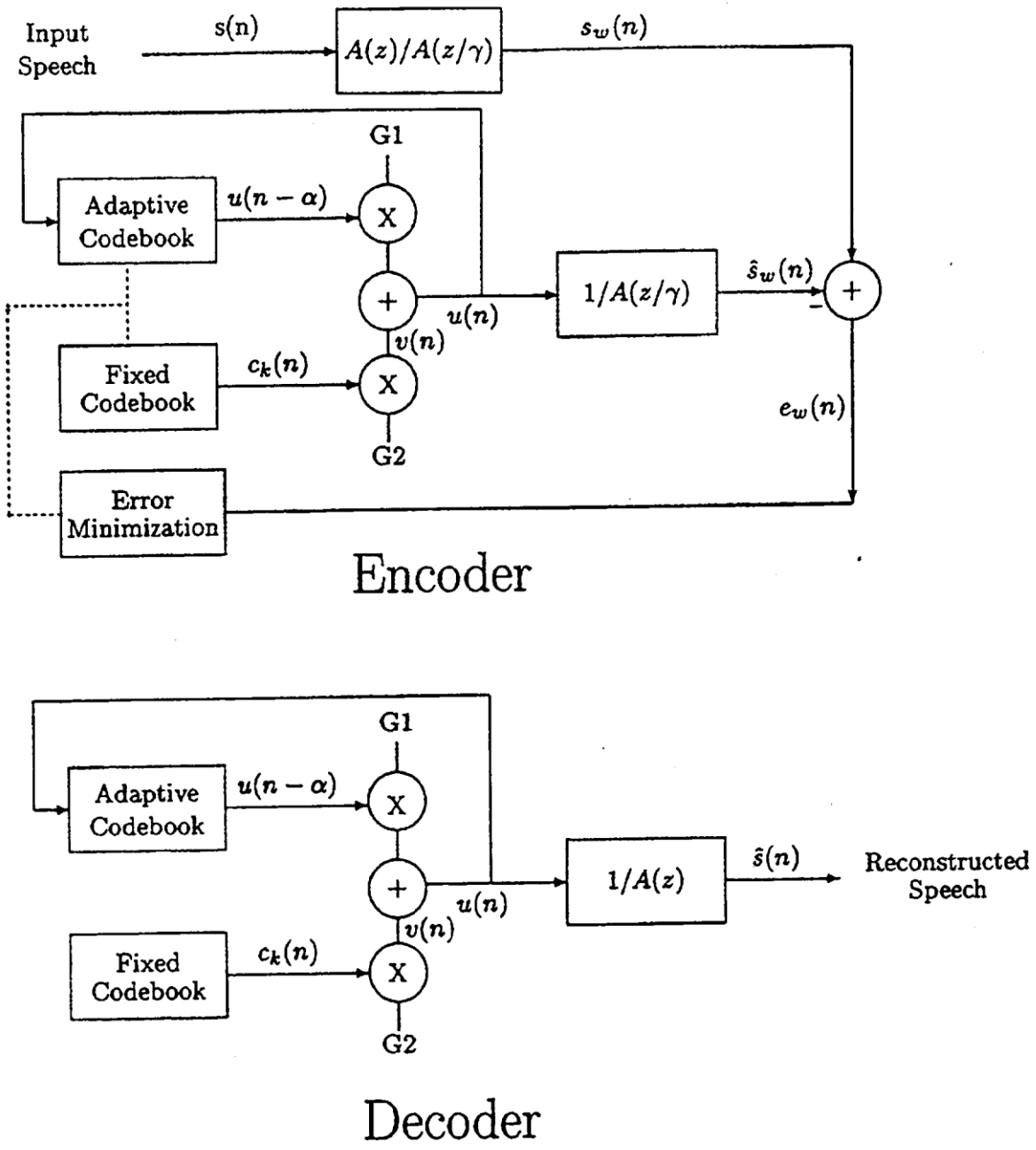


Figure 6.2 Adaptive codebook assisted CELP codec structure.

Algebraic CELP

The name Algebraic CELP implies the structure of the codebook used to select the excitation codebook vector. The codebook vector consists of a set of interleaved permutation codes containing few nonzero elements [3, 4]. The fixed codebook structure is given by,

Table-1: Fixed codebook structure used in ITU –T G.729

Track (k)	Signs	Pulse positions (p_k)
$i_0 = 0$	s_0: ±1	p_0: 0, 5, 10, 15, 20, 25, 30, 35
$i_1 = 1$	s_1: ±1	p_1: 1, 6, 11, 16, 21, 26, 31, 36
$i_2 = 2$	s_2: ±1	p_2: 2, 7, 12, 17, 22, 27, 32, 37
$i_3 = 3$	s_3: ±1	p_3: 3, 8, 13,18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

where, 'p_k' is the pulse position, 'k' is the pulse number, 'L' is the interleaving depth (= 5) and 'j' ranges from 0 to $2^M - 1$, 'M' is the number of bits (=3) describing the pulse positions.

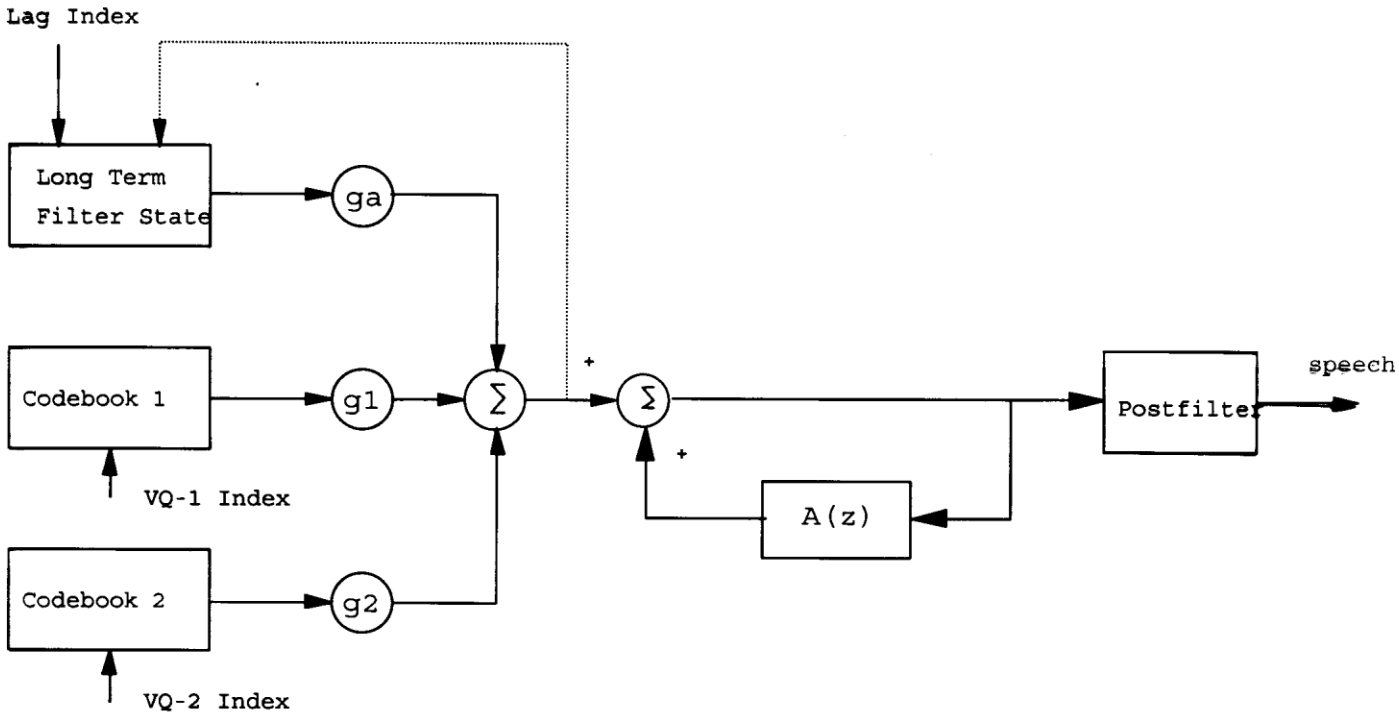
The codebook vector, c(n), is determined by placing the 4 unit pulses at the found locations (p_k) multiplied with their signs (±1) as follows:

$$c(n) = s_0\delta(n-p_0) + s_1\delta(n-p_1) + s_2\delta(n-p_2) + s_3\delta(n-p_3), \quad n = 0, \ldots, 39 \qquad (9)$$

where, δ(n) is the unit impulse function.

ACELP !
Mais
Toujours
2 dictionnaires
1 fixe
et
1 adaptatif

The 8 kbits/s IS-54 VSELP



- developed by Motorola - part of IS-54 cellular standard
- speech sampled at 8 kHz - segmented in 20ms frames - sub-frames of 5 ms
- gain-shape VQ
- LTP lag search over 127 integer (20 to 146)

Copyright 2001 (c) Andreas Spanias

The 8 kbits/s IS-54 VSELP (2)

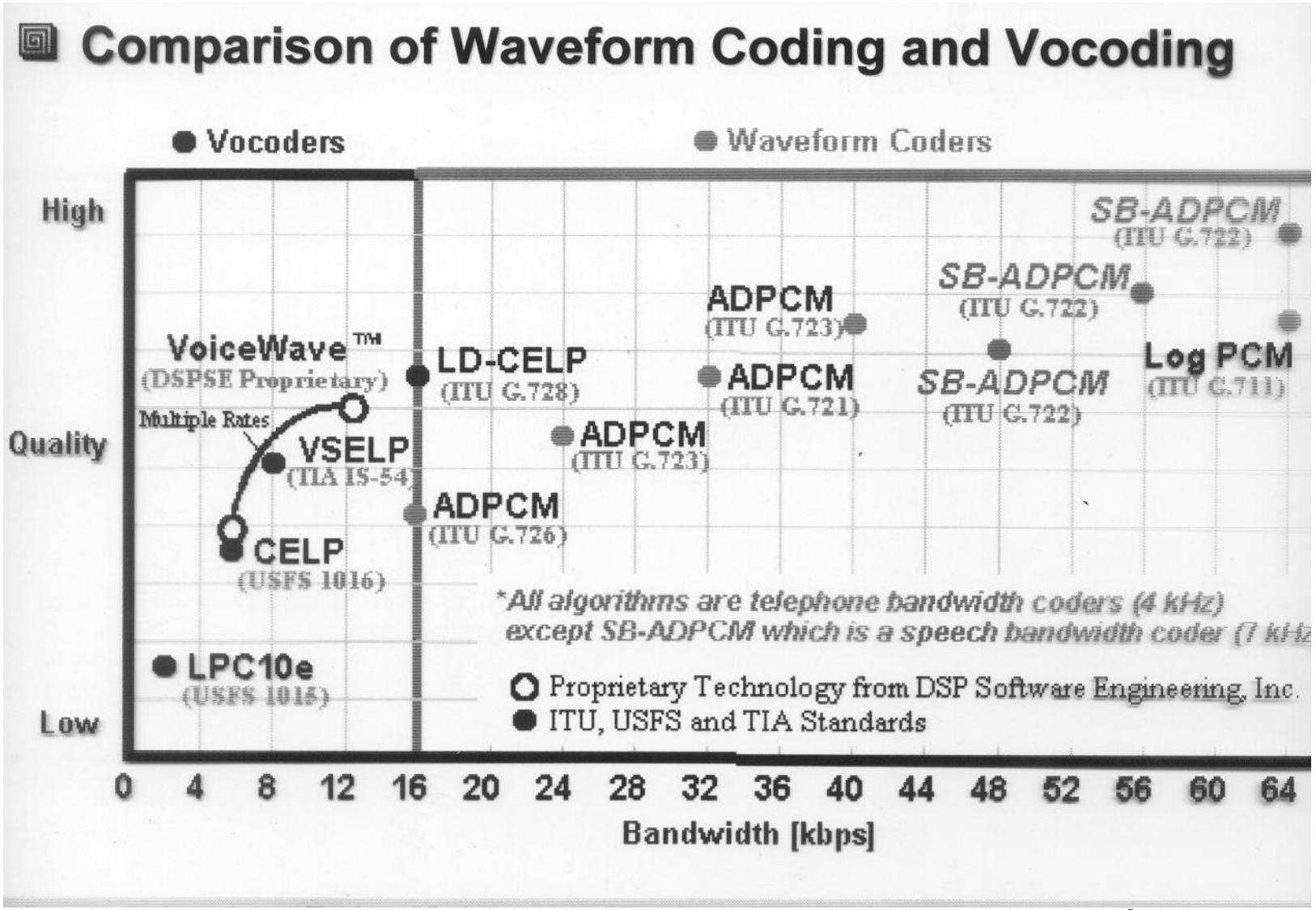
- two stochastic codebooks containing 128 codevectors each
- codevectors formed by combining linearly 7 basis vectors

$$c_{[b_1 b_2 b_3 b_4 b_5 b_6 b_7]} = \sum_{m=1}^{7} u_m \beta_m$$

codevector index consists of 7 bits

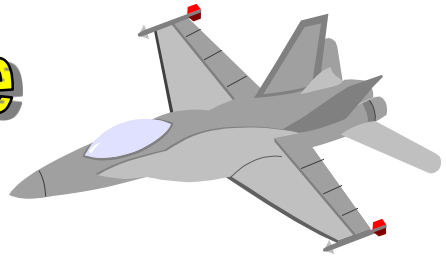
Weights, u_m, assume value of -1 or 1. If address bit b_m is 0 then corresponding coefficient u_m,=-1; if b_m=1 then u_m,=1
This provides robustness to channel errors.

Copyright 2001 (c) Andreas Spanias



LD-CELP : Low Delay CELP, vecteurs de 5 échantillons

Exemple MPLP à 9.6kbits/s pour Skyphone



British Telecom : téléphone dans les avions (Skyphone)
à base de MPLP avec :

- Prédicteur court terme (STP) d'ordre 10,
  - Algorithme de calcul des coefficients de prédiction
  - Levinson-Durbin avec fenêtre d'analyse de Hamming, 32 ms
  - Coefficients recalculés toutes les 20 ms
  - Codés sous forme *d'Inverse Sine Coefficients (Arcsin(k_i))*
- Prédicteur long terme (LTP) d'ordre 1
  - Retard D (pitch) calculé par max de l'autocorrélation
  - et recherché parmi 64 valeurs possibles (6 bits)
  - Gain b codé par Q non linéaire

Implanté sur AT&T DSP 32C : 75% du traitement
MOS de 3.4
Retard Codage-décodage <40ms

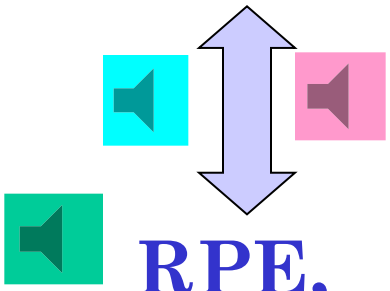
Exemple RPE : Full Rate GSM à 13 kbits/s

GSM : Groupe Special Mobile

ETSI : European Telecommunication Standard Institute

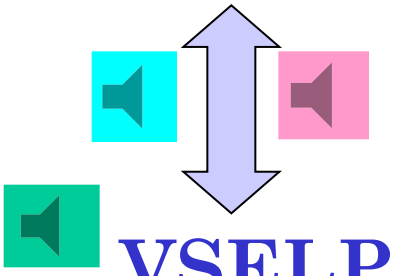
3 différents Codeurs GSM

Full Rate GSM06.10



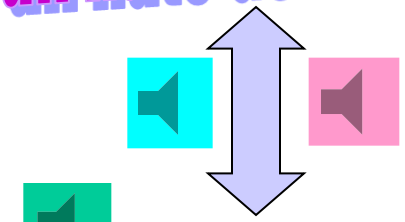
RPE, 1987

Half Rate GSM06.20



VSELP, 1993

Enhanced Full Rate GSM06.60



ACELP, 1997

Entrée : parole codée PCM uniforme 13 bits, Fe=8kHz

Traitement bloc par bloc (frame-by-frame)

taille de bloc = 20ms (160 éch)

Full Rate GSM : GSM RPE-LTP à 13 kbits/s

1987

Pour chaque trame de 20ms (160 éch) :

prédicteur linéaire court terme d'ordre 8

codage des coefficients : LAR sur 36 bits (6/6/5/5/4/4/3/3)

Chaque trame divisée en sous-trame (subframe) de 5ms :

prédicteur long terme d'ordre 1 : retard sur 7 bits, gain sur 2 bits

amplitudes et grille de position des pulses estimés

codage des amplitudes :

normalisées par la plus grande

la plus grande codée logarithmiquement sur 6 bits,

les autres : Q uniforme 3 bits

Total pour trame de 20 ms (160 éch) : 260 bits soit 13 kbits/s

Canal GSM full rate : 22.8 kbits/s *(le reste pour protection contre erreurs)*

MOS de 3.47 5 à 6 MIPS

code disponible (C) dans le domaine public

Surpassé par Enhanced Full Rate (EFR) à base de CELP :

meilleure qualité de parole à un taux de bits plus faible

Half Rate GSM à 5.6 kbits/s

Mis en place pour prendre en compte le nombre croissant d'utilisateurs
à base de Vector Sum Excited Linear Prediction (VSELP de Motorola)
Pour doubler la capacité du système cellulaire GSM, le canal half rate supporte 11.4 kbps *(5.8 kbps utilisés pour protection contre erreurs)*.
Qualité de parole comparable au full rate
sauf en présence de bruit de fond et plusieurs locuteurs
La plupart des mobiles récents le supportent, mais seul le réseau SFR l'utilise, comme secours afin d'éviter la saturation des cellules à Paris aux heures de pointe.

Enhanced Full Rate GSM

EFR : pour canal full rate : grande amélioration par rapport au FR
12.2 kbps pour codage parole, 10.6 kbps pour protection erreurs
Basé sur Algebraic Code Excited Linear Prediction (ACELP)

http://www.data-compression.com/speech.html#8000bps ici

3GPP, AMR : AMR-NB (narrow-band)

Adaptive Multi-Rate (AMR) **audio** data compression scheme optimized for **speech** coding. AMR was adopted as the standard speech codec by 3GPP (oct 1998)
Now widely used in GSM and UMTS,
Uses link adaptation to select from one of eight different bit rates based on link conditions, with bit-rates of 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 and 4.75 kbit/s.
Bad radio conditions -> source coding reduced and channel coding increased.
Improves the quality and robustness of the network connection while sacrificing some voice clarity.
ACELP + DTX (discontinuous transmission)
+VAD (voice activity detection) + CNG (comfort noise generation)
Available in FR (full rate) and HR (half rate) channel

GSM6.90 http://www.vocal.com/speech_coders/gsmamr.html

ici

3GPP, AMR : AMR-WB (wide-band)

Adaptive Multirate Wideband (AMR-WB) codeur standardisé (G722.2),
utilise ACELP (Algebraic CELP) : 6.6 à 23.85 kbits/s

Parole large bande : différence majeure dans 3G
200-3400 Hz → 50-7000 Hz
Apporte beaucoup en qualité (sons non voisés):

http://www.voiceage.com/amrwb.php ici

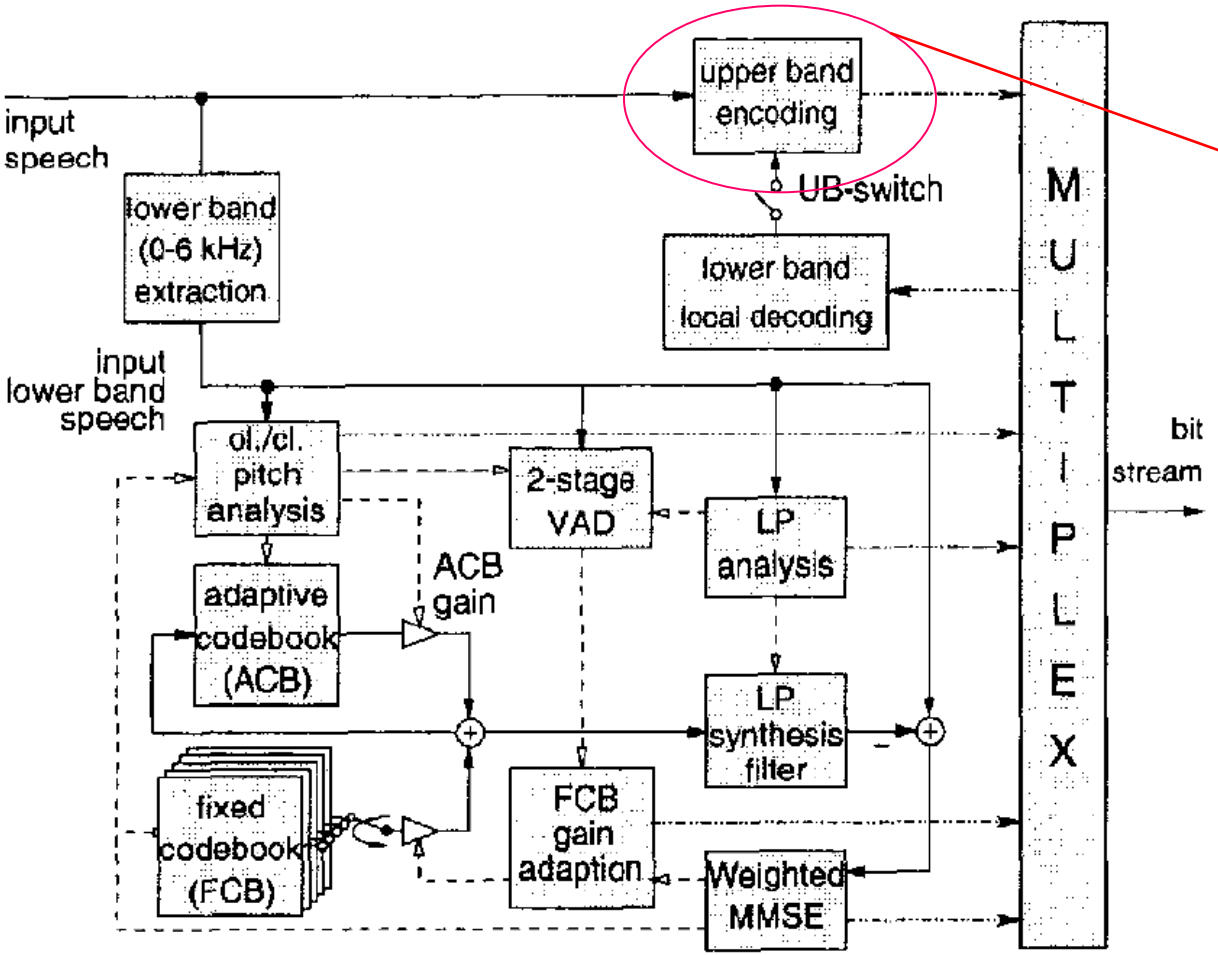
Bande classique   Bande élargie

http://speechcodecs.wordpress.com/ ici

http://www.voiceage.com/media/WidebandSpeech.pdf ici

3GPP, AMR : AMR-WB (wide-band)

Exemple



Codage de
la partie haute :
Par extension
de la bande spectrale
ou
ADPCM

Compression de sources sonores

C.Mailhes

Quelques références :

« A tutorial on MPEG/Audio compression », Davis Pan,
Trans on IEEE Multimedia, 1995, pp. 60-74 (dispo web)

« MPEG Audio and Video Coding »,
IEEE Signal Processing Magazine, Sept 97, vol 14, n5

« Le codeur MPEG-2 AAC expliqué aux traiteurs de signaux »,
O.Derrien, S.Larbi, M.P.Guimares, N.Moreau, Annales des Télécoms, Sept-Oct 2000.

Et de nombreux sites web …

http://www.tnt.uni-hannover.de/project/mpeg/audio/general/
http://www.sericyb.com.au/audio.html/
http://www.cs.sfu.ca/CourseCentral/365/li/material/notes/Chap4/Chap4.4/Chap4.4.html
http://www.iis.fraunhofer.de/amm/techinf/layer3/index.html
http://www.iis.fraunhofer.de/amm/techinf/
http://www.mpeg.org/MPEG/audio.html
...

Historique

- **Référence = CD**
  - Audio Large Bande (WideBand audio) : **20 - 20 kHz**
  - 16 bits * 44.1 kHz = 705 kbit/s (mono)
  - Stéréo : 1.412 Mbps
  - Dynamique et SNR de l'ordre de 90 dB
  - Rajout code correcteur d 'erreurs + codage par plage
  - Débit CD : 2 * 0.705 + 2.91 = **4.32 Mbit/s !!!**
- But : débit moindre pour qualité équivalente

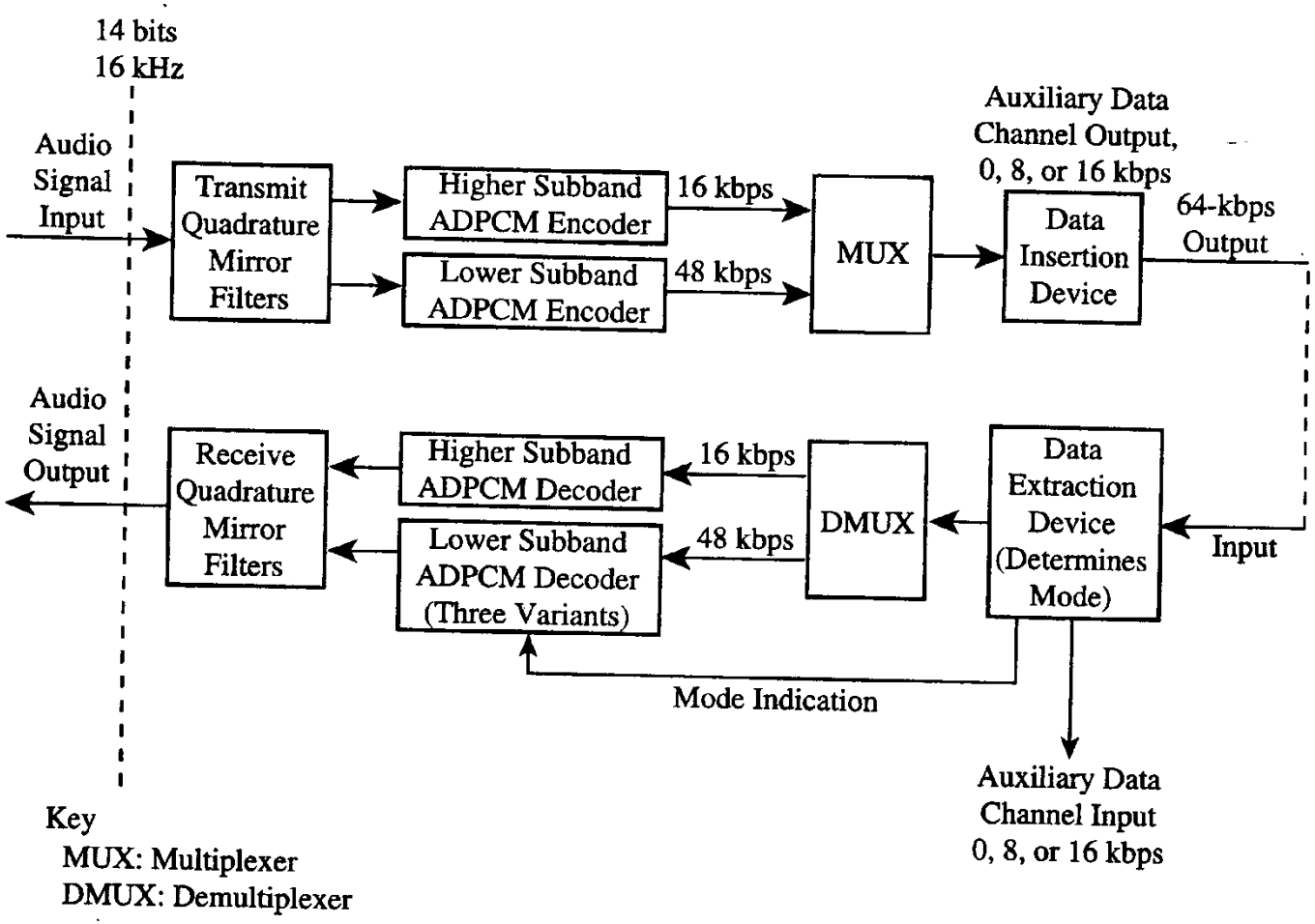
Qualité « transparente » (*transparent quality*)

Qualité Haute Fidélité (Hifi)

Codage fréquentiel très utilisé : variantes du codage en sous-bandes et par transformées

Ex : 1988 : IUT-T G722 : 2 sous-bandes + ADPCM (G721)

G.722 : codeur à Fe=16 kHz



Two-Band Subband Coder for 64-kbps Coding of 7-kHz Audio

Historique

MPEG : Moving Pictures Expert Group : place prépondérante
normes internationales mais codeurs non entièrement normalisés
Seul format du flux de bits fixé : impose structure du décodeur
et une partie du codeur => degré de liberté

- **Entre 1985 et 1992**

  Objectif : Digital Audio Broadcasting

  Normalisation ISO/MPEG1-Audio en **1992**

  En « finale » : MUSICAM et ASPEC

  705 kbits/s => 192, 128 et 96 kbits/s (3 couches)

  Qualité équivalente mais complexité croissante

  MP3 = MPEG1 Couche 3

  Autres codeurs : ATRAC(Adaptive Transform Acoustic Coding :Minidisc),

  NICAM (Near-Instantaneously Companded Audio Multiplex - BBC)

  PASC (Precision Adaptive Subband Coding : Digital Compact Cassette).

Historique

- **Entre 1989 et 1997**
  - Partie son TVHD, cinéma, DVD, Internet
  - MPEG2 : extension multi-voies (5.1) en 1994
  - Dolby AC3 en 1995
  - MPEG2-AAC (Advanced Audio Coder) en 1997
- Après 1997
  - MPEG-3 : prévu pour la HDTV, n'a pas vu le jour
  - MPEG-4 Audio : décembre 1998 : Codage orienté objet

    Very Low Bitrate Audio-Visual Coding

    Inclut MPEG-2 AAC pour codage haute qualité, permet aussi compression musique à bas débit, codage de parole et synthèse musicale, sons réalité virtuelle 3D… "media objects"… TwinVQ : chut ! Secret !
  - MPEG-7 : Novembre 2000 : "Content Description"

    standard pour la recherche multimédia (*je cherche une musique comme celle de Bashung*)
  - MPEG-21 : « End to end e-framework »

Parole et Audio

Pb du Codage Audio postérieur au Codage de Parole

Différences parole - audio : codage audio :

- taux de bits plus élevés,
- meilleure résolution temporelle,
- dynamique plus grande,
- grandes variations dans l'évolution de la DSP,
- représentations stéréophoniques et multi-canaux
- attente d'une plus grande qualité.

Pourtant similaires car basés sur propriétés du système auditif humain

Mais, modèle de production de parole existe

en musique, en audio ???

Compression

4 clefs technologiques

Codage perceptuel

Codage dans le domaine fréquentiel

Passage d'une fenêtre à une autre

Allocation de bits dynamique

Bandes critiques

Système auditif humain : intègre la puissance du signal sonore sur des bandes fréquentielles, comme banc de filtres

résolution limitée, dépendant de la fréquence

Mesure uniforme de perception : largeur de **bandes critiques**

Moins de 100Hz aux fréquences les plus basses audibles,

Plus de 4 kHz dans les fréquences les plus hautes

Système auditif humain

= banc de filtres, constitués de filtres passe-bande

à recouvrement spectral : 24 à 26 bandes critiques définies

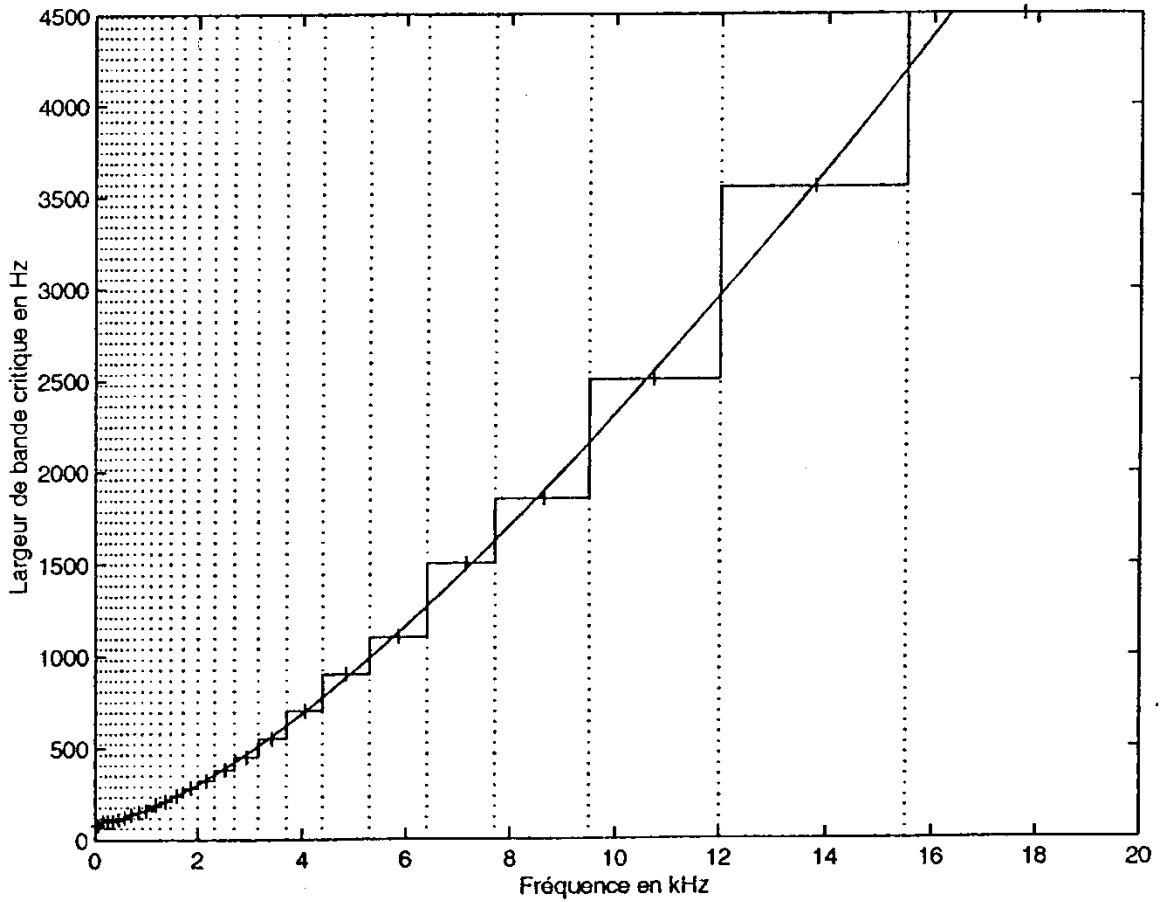
Définition du **Bark** (Barkhausen) :

1 Bark = largeur d'une bande spectrale

$f < 500$ Hz, $f \rightarrow f / 100$ Bark

$f > 500$ Hz, $f \rightarrow 9 + 4 * \log_2(f / 100)$ Bark

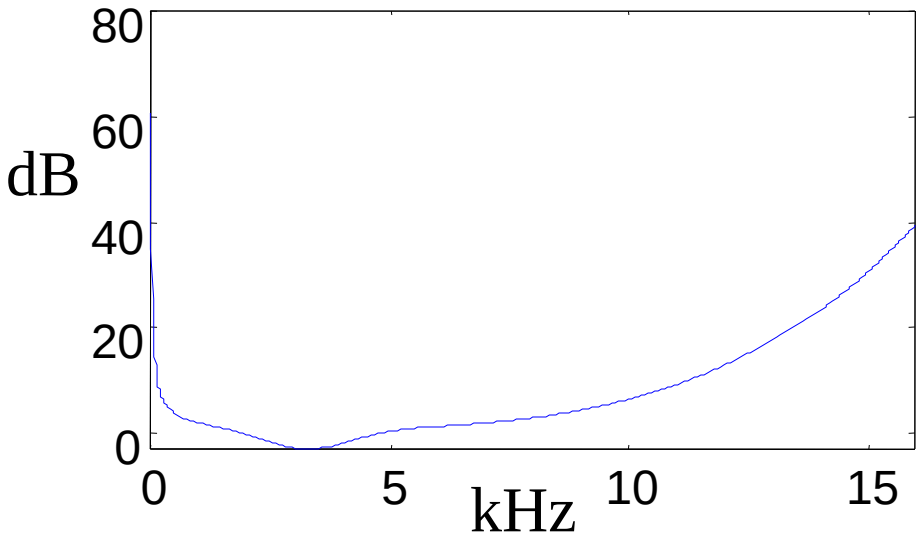
Largeur des bandes critiques en fonction de la fréquence
Partition de l'axe des fréquences en 24 bandes critiques



Fréquence médiane de la b-ième bande critique = b Barks

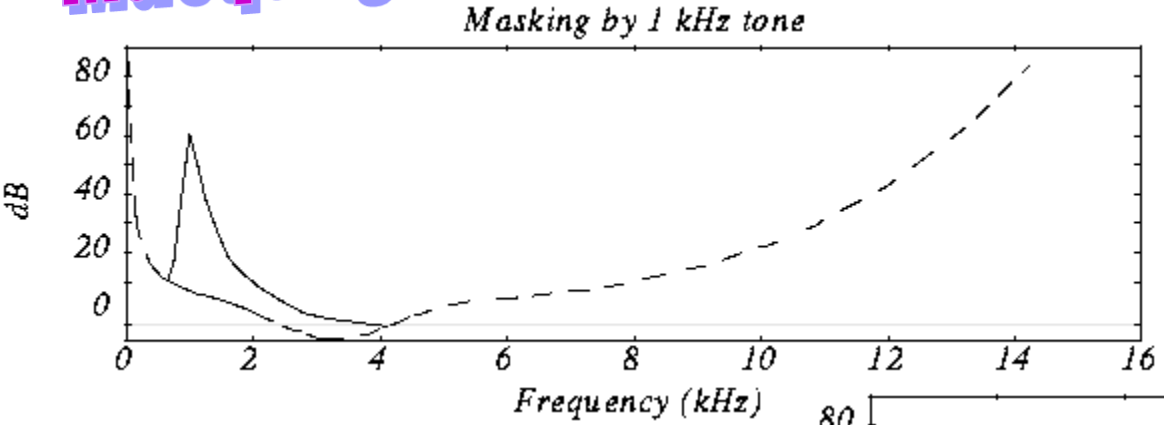
Sensibilité de l'oreille humaine

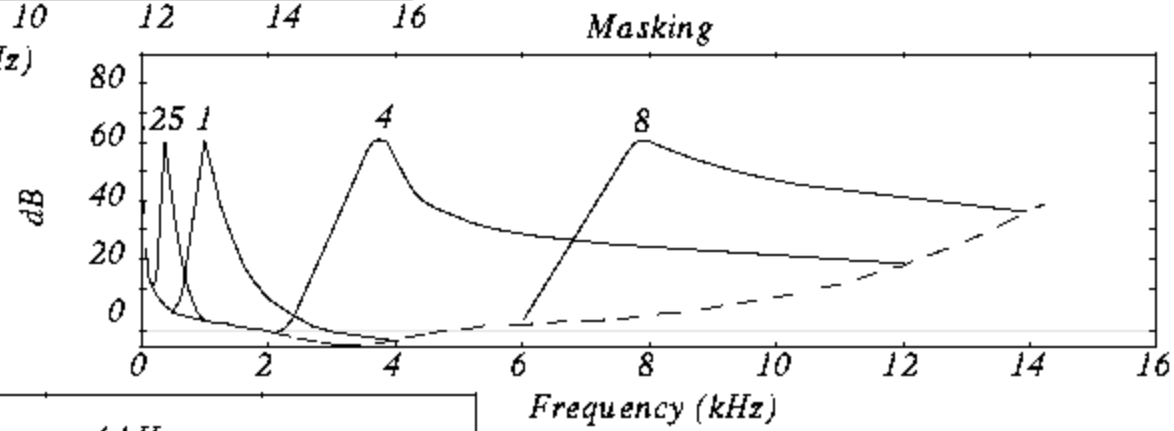
Oreille humaine : 20 Hz à 20 kHz, plus sensible entre 2 - 4 kHz
dynamique autour de 96dB
Seuil d'audition en environnement calme :« Threshold in Quiet »
Expérience : personne dans pièce sans bruit, monter le niveau jusqu'à entendre, fréquence par fréquence

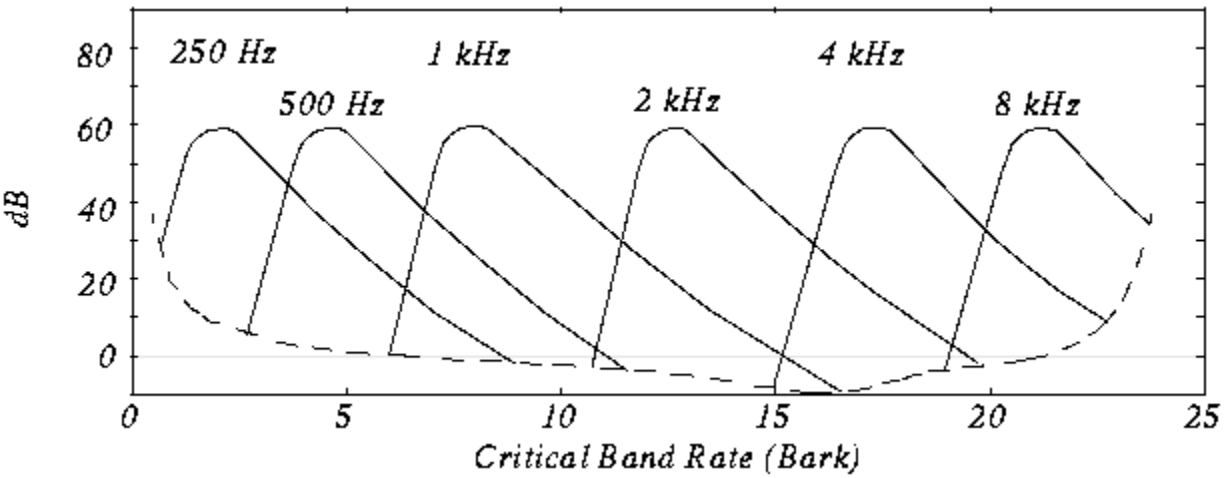


Masquage fréquentiel

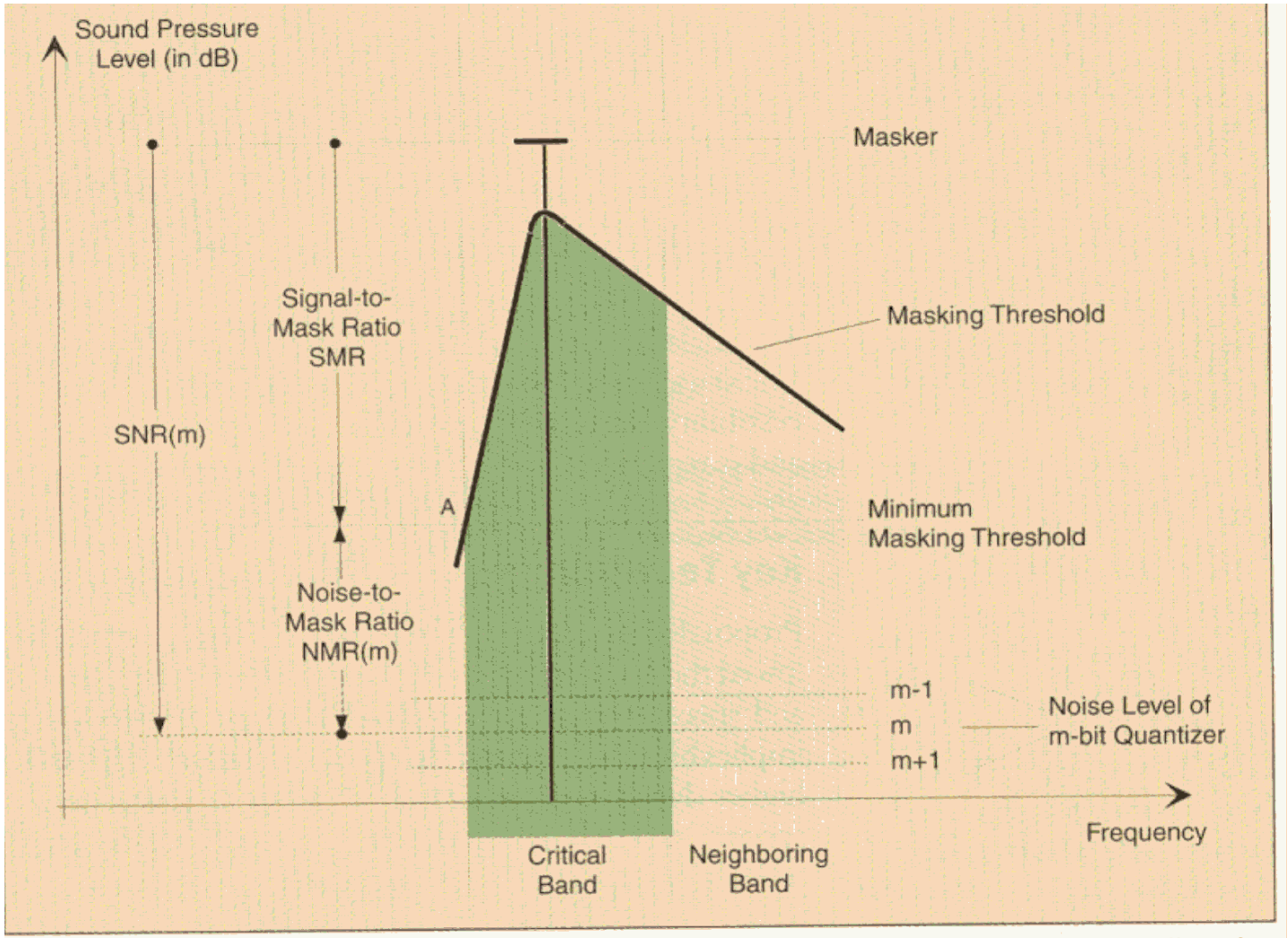
Résultats de psycho-acoustique







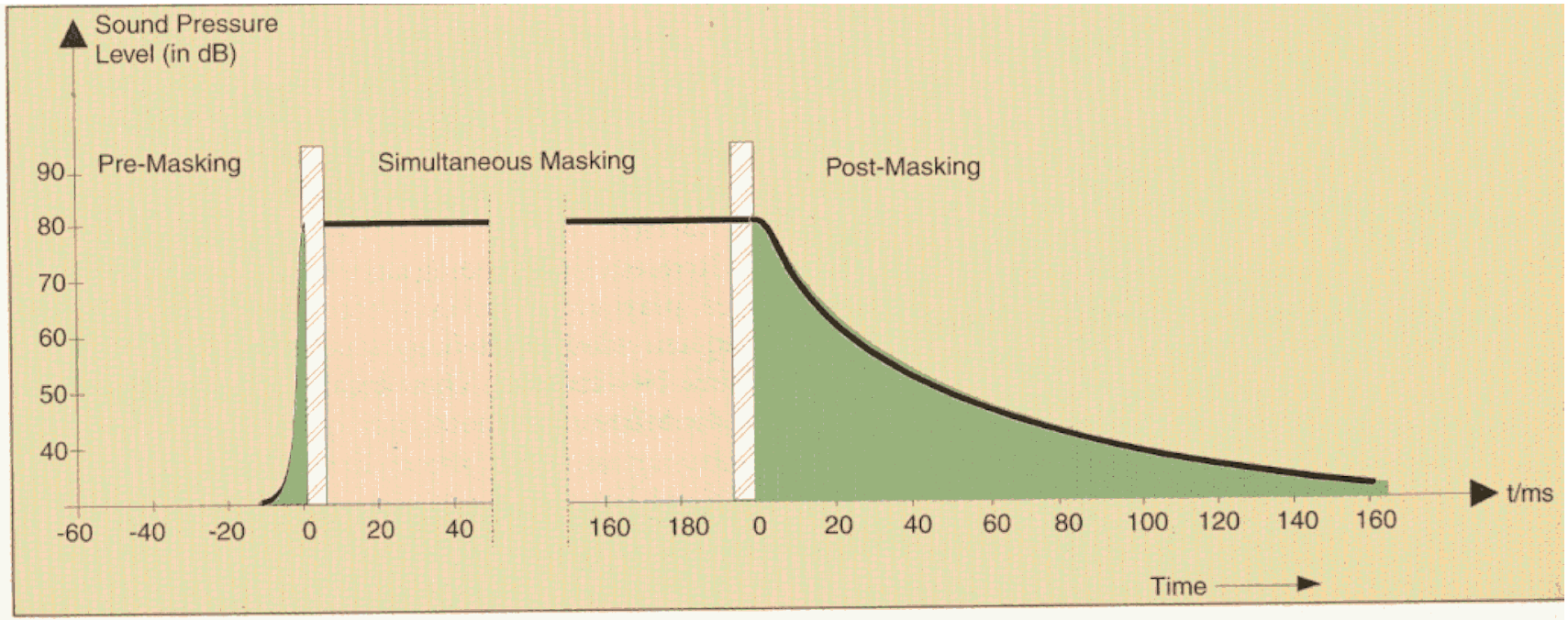
Masquage fréquentiel



▲ *2. Masking threshold and signal-to-mask ratio (SMR) (acoustical events in the gray areas will not be audible).*

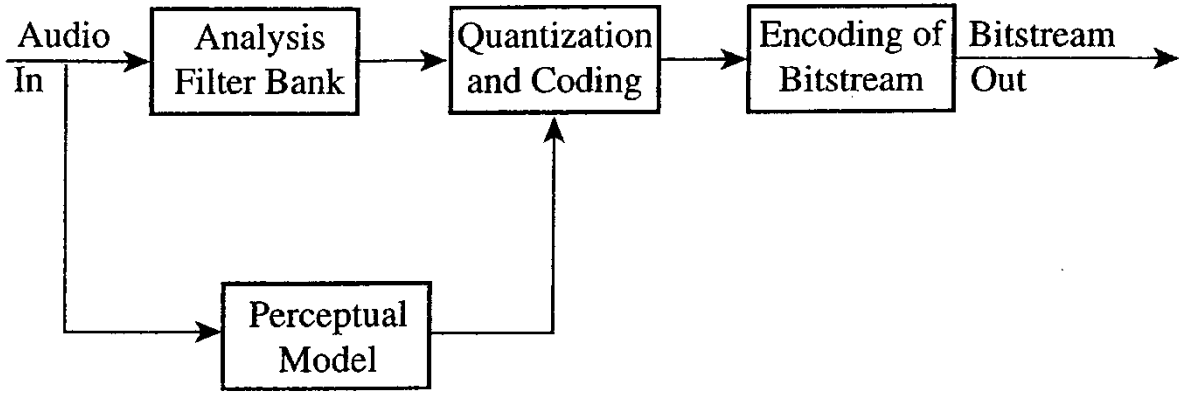
Masquage temporel

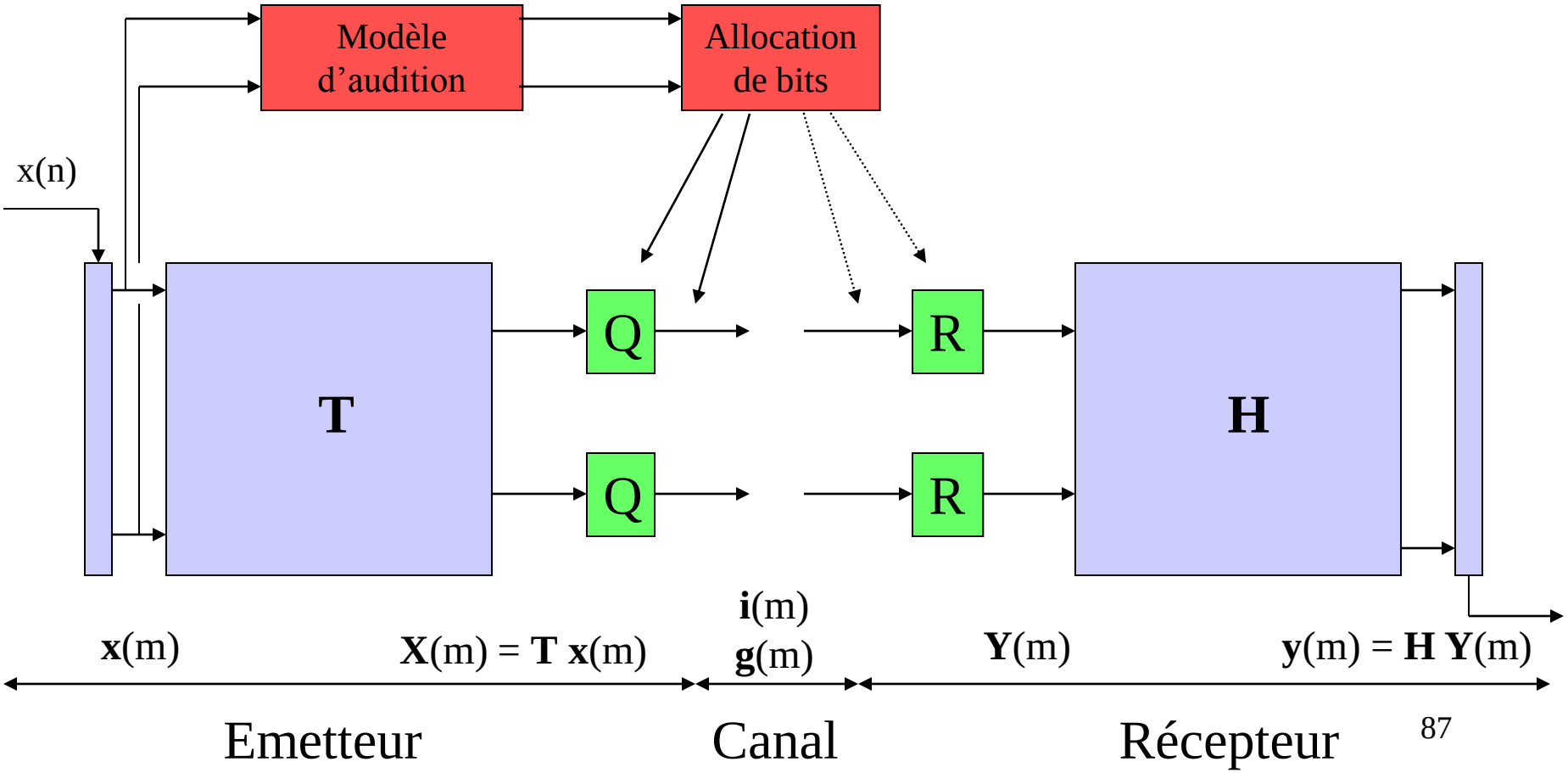
Signal peut masquer d'autres signaux de puissance plus faible arrivant juste avant (Premasking) ou juste après (Postmasking)
Premasking : plus court que Postmasking



▲ 3. Temporal masking (acoustic events in the gray areas will not be audible).

Codeur « perceptuel »





MPEG-1

1.5 Mbits / sec pour audio et vidéo :

environ 1.2 Mbps pour vidéo, 0.3 Mbps pour audio

Facteurs de compression allant de 2.7 à 24

⬇ Pour taux de compression de 6:1, en conditions normales d'écoute, auditeurs expérimentés **ne discernent pas** clips audio original et codé.

MPEG audio supporte différentes Fe : 32 kHz, 44.1 kHz et 48 kHz.

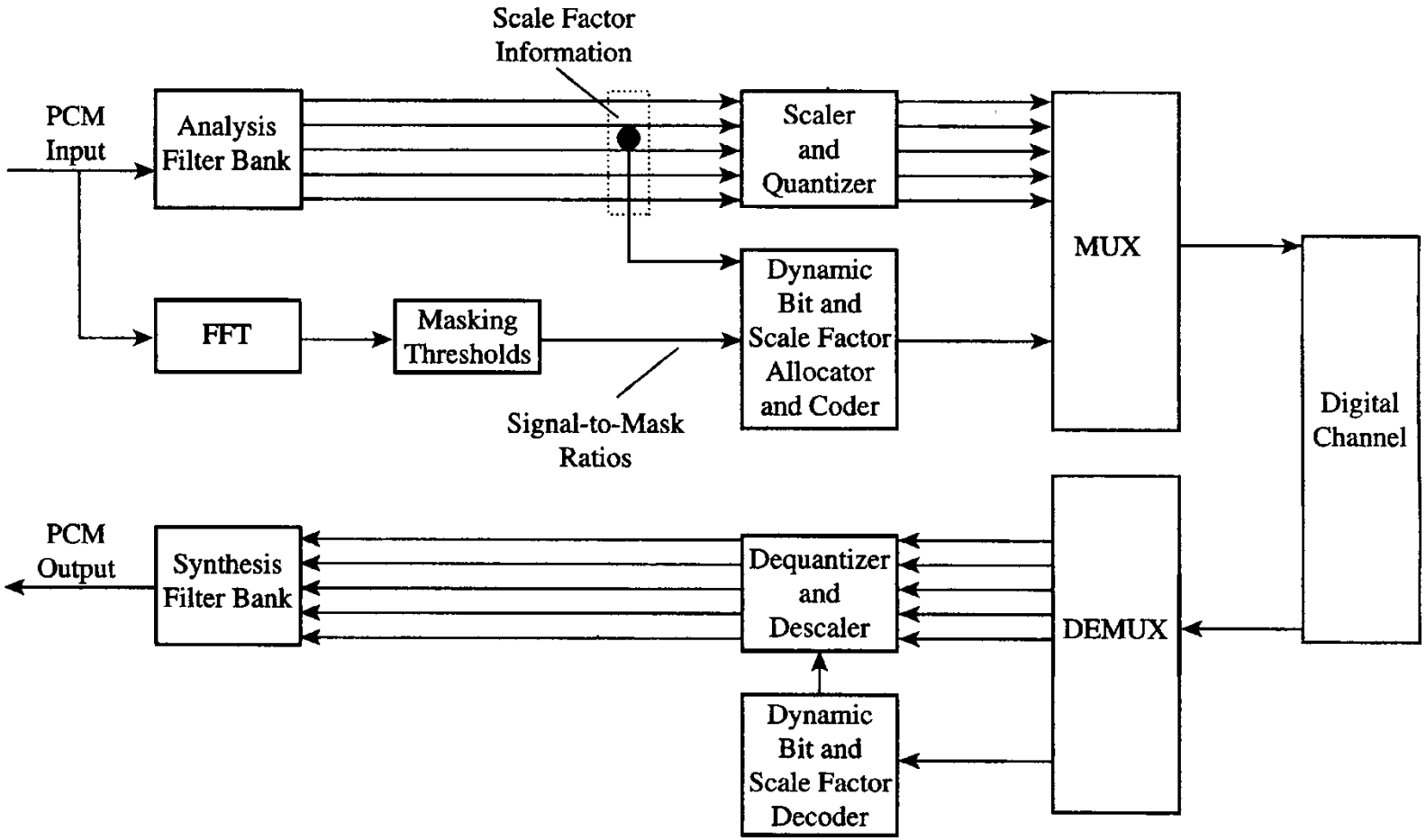
Permet 1 ou 2 canaux audio dans un des 4 modes :

- monophonique (un seul canal audio),
- dual-monophonique (2 canaux indépendants, par ex anglais et français),
- stéréo (canaux stéréo partageant bits mais sans codage conjoint stéréo,
- stéréo conjoints (utilise corrélation entre canaux stéréo).

MPEG-1 Audio : étapes de l'algo

• Banc de filtres pour diviser en 32 bandes spectrales
(filtrage en sous-bandes) et ré-échantillonnage à fréquence min.
• Calcul du seuil de masquage pour chaque bande
dû aux bandes voisines en utilisant modèle psycho-acoustique
• Si la puissance d'une bande < seuil de masquage, on ne la code pas.
• Sinon, calcul du nombre de bits nécessaires pour représenter signal
filtré tel que le bruit de quantification < seuil de masquage
(1 bit supplémentaire = gain de 6 dB)
• Formatage du flux de bits.

MPEG-1 Audio : schéma



Block Structure of ISO/MPEG Audio Encoder and Decoder, Layers 1 and 2

MPEG-1 Audio : les couches (layers)

3 couches, modèles du codeur identiques mais complexité du codec augmente à chaque couche (ainsi que les performances !)

MPEG-1 Layer 1 :

Applications grand public : Digital Compact Disk
32 kbps (mono) à 448 kbps (stéréo)
Faible complexité du décodeur (codeur 1.5 à 3 fois plus complexe)
Proche qualité CD à 256-384 kbps pour stéréo

Filtres de type DCT (polyphase, ordre = 511, RIF=5.33ms)
sur une trame (384 éch soit 12 éch à la sortie de chaque bande)
bandes égales : 750 Hz pour Fe=48 kHz
modèle psychoacoustique : uniquement masquage fréquentiel
SMR (Signal to Mask Ratio) calculé par FFT sur 512 points
Allocation de bits dynamique : choix d'un quantificateur parmi 15 : compromis entre SMR et taux de bits du mieux possible.

MPEG-1 Audio : les couches (layers)

MPEG-1 Layer 2 :
Applications audio grand public et professionnelles
(broadcasting, TV, multimedia, telecommunications)
32-192 kbps (mono) et 64-384 kbps (stéréo)
Proche qualité CD à 192-256 kbps pour stéréo
Complexité du décodeur 25% plus grande que layer I
Complexité du codeur 2 à 4 fois plus grande

très proche de Layer 1,
travaille sur 3 trames (avant, pendant, après) = 1152 éch
utilisation partielle du masquage temporel
SMR (Signal to Mask Ratio) calculé par FFT sur 1024 points,
tables de quantification plus fines

MPEG-1 Audio : les couches (layers)

MPEG-1 Layer 3 (MP3) :

Pour audio professionnel et ISDN bande étroite (64 kbps)
32 filtres de largeur non-égales,
chaque sous-bande décomposée en utilisant la MDCT
produit 6 ou 18 composantes fréquentielles supplémentaires
pour augmenter la résolution fréquentielle
permet d'éviter artéfacts temporels et pré-échos
modèle psychoacoustique utilise masquage temporel,
Quantificateurs non uniformes.

prend en compte la redondance stéréo

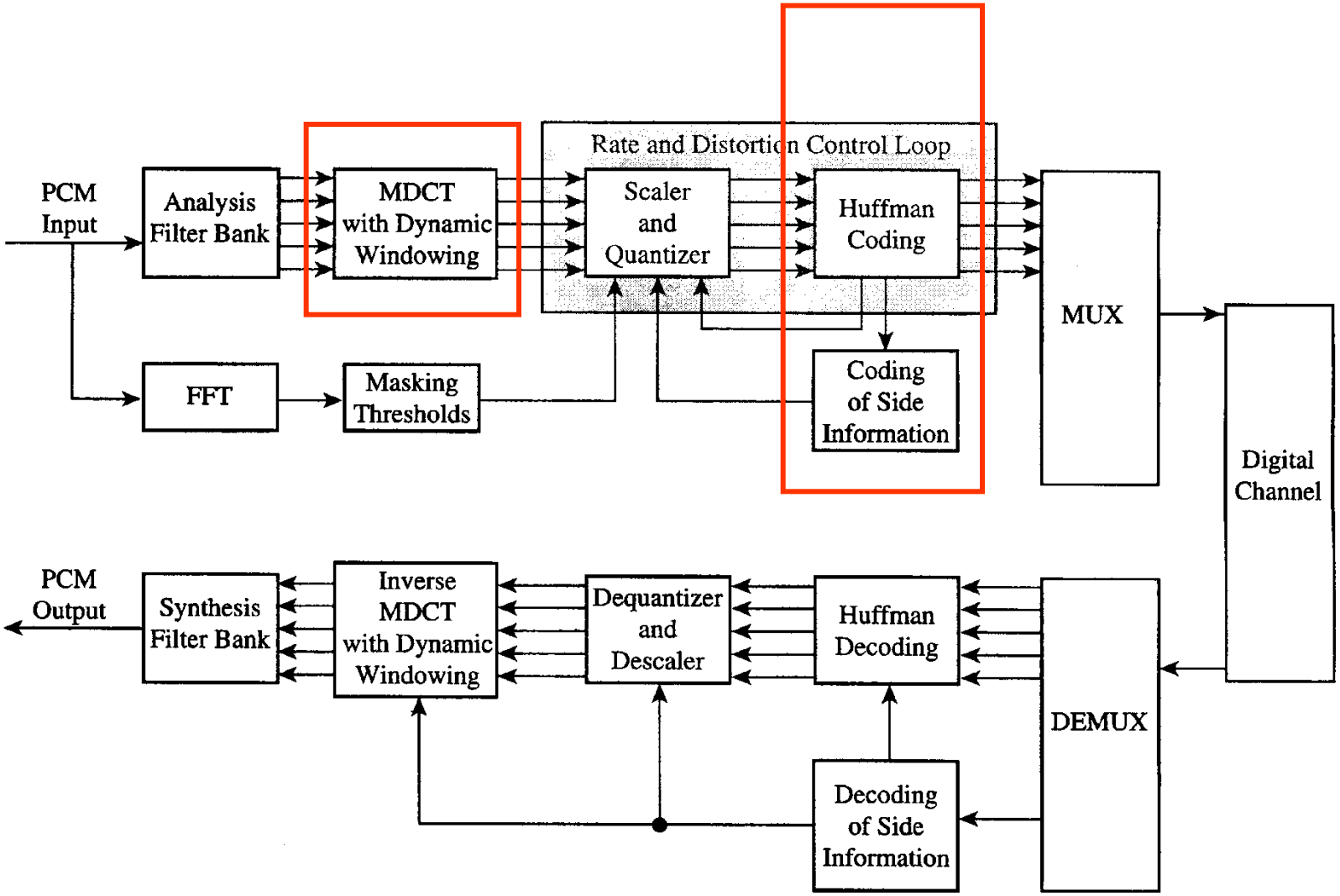
Utilisation d'un codage d'Huffman (18 tables possibles)

Codage de la redondance stéréo (possible dans Layers 1 et 2)

Intensity Stereo Coding :
aux sous-bandes hautes, codage de D+G

Middle/Side (MS) Stereo Coding :
Codage de D+G et D-G

MPEG-1 Layer 3 (MP3) : schéma



Block Structure of ISO/MPEG Audio Encoder and Decoder, Layer 3

Utilisation de la MDCT : Modified Discrete Cosine Transform dans MPEG1 layer 3 et MPEG2

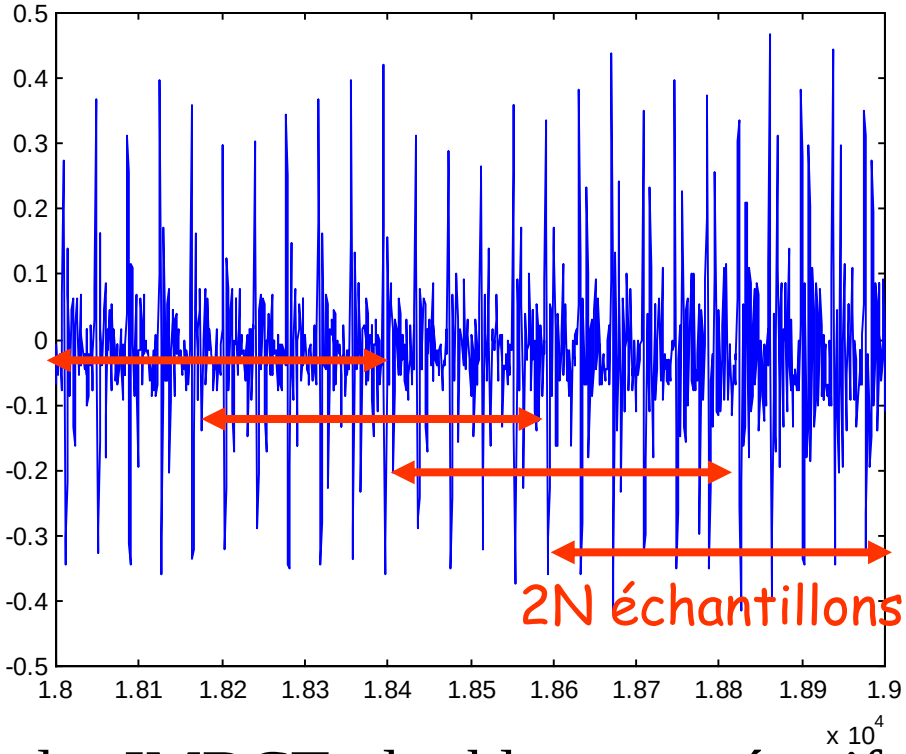
Une « lapped transform » :

En entrée 2N échantillons,

En sortie N points calculés

$$X_k = \sum_{n=0}^{2N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right]$$

Pour k=0,…,N-1



Transformée inverse :

IMDCT

Parfaitement inversible en ajoutant les IMDCTs des blocs consécutifs qui se recouvrent :

Time Domain Aliasing Cancellation : TDAC

$$y_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}+\frac{N}{2}\right)\left(k+\frac{1}{2}\right)\right]$$

Pour n=0,…,2N-1

Utilisation de fenêtres dans la MDCT

Pour éviter discontinuités aux bornes des trames traitées :
Fenêtre de pondération sur x_n et y_n (avant MDCT et après IMDCT).
Différentes fenêtres possibles mais condition de Princen-Bradley

$$w_n^2 + w_{n+N}^2 = 1$$

➡ $w_n = \sin\left[\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right]$ pour MP3 et MPEG2-AAC

➡ $w_n = \sin\left(\frac{\pi}{2}\sin^2\left[\frac{\pi}{2N}\left(n+\frac{1}{2}\right)\right]\right)$ pour Ogg Vorbis (open source, pas de brevet)

➡ une KBD (dérivée de Kaiser-Bessel) pour AC-3 et MPEG4-AAC

MPEG-1 Audio : performances

Layer	Taux de Bits	TC	MOS à 64 kbps	MOS à 128 kbps	Retard min (théor.)
Layer 1	192 kbps	4:1	---	---	19 ms
Layer 2	128 kbps	6:1	2.1 à 2.6	4+	35 ms
Layer 3	64 kbps	12:1	3.6 à 3.8	4+	59 ms

Retard réel =
3 fois le théorique.

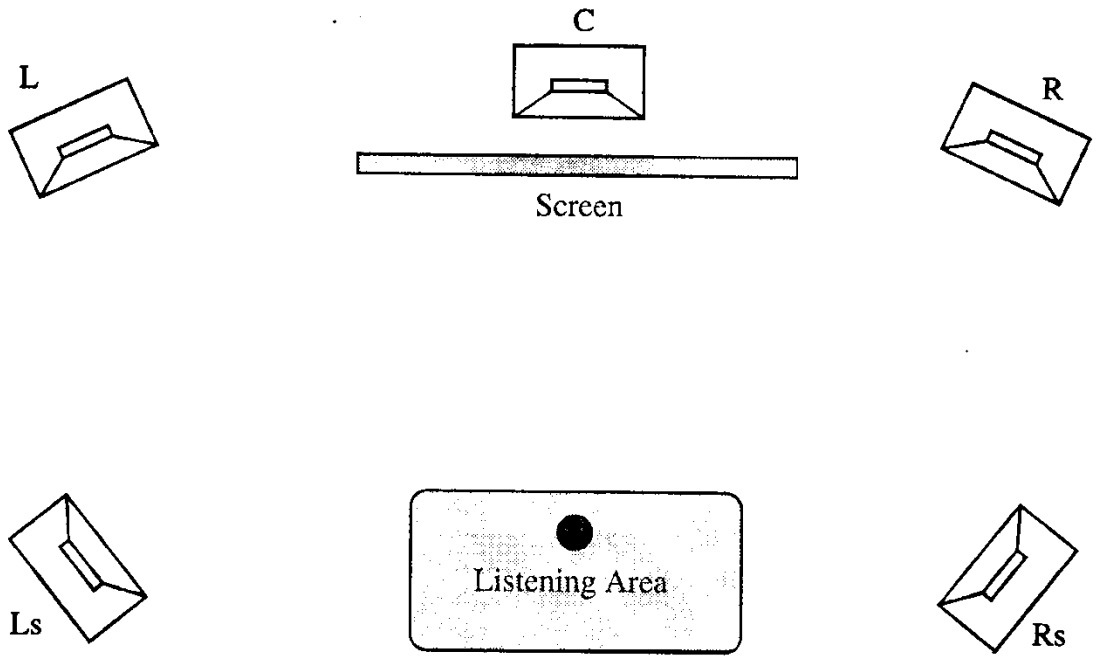
Philips' Digital Compact Cassette (DCC)

audio pour DAB, CD-ROM, Vidéo CD

transmission audio sur ISDN

1994 : MPEG-2

But :
créer système sonore capable de
restituer bandes sons cinéma multicanaux
essentiellement le 5.1 :Avant G et D, Centre, Arrière G et D
+ canal d'extrême grave, passe-bas.



Speaker Configuration for Multichannel MPEG-2 Audio

MPEG-2

MPEG-2 BC (backward compatible - flux binaire décodable par MPEG-1)
MPEG-2 LSF (Low Sampling Frequency) 16, 22.05 et 24 kHz

MPEG-2 NBC : codeur multicanal de très haute qualité, plus efficace
devient MPEG-2 AAC (Advanced Audio Coding)
semble descendant direct de MPEG-1 layer 3
codeur modulaire, autour d'un noyau
permet plusieurs Fe de 8 kHz à 96 kHz (*MP3 max : 48 kHz*)
nombre max de canaux codables : 48 !!!

MPEG formal listening tests have demonstrate that for 2 channels MPEG 2 AAC is able to provide slightly better audio quality at 96 kb/s than layer-3 at 128 kb/s or layer-2 at 192 kb/s.

Differences between MPEG-2 AAC and MP3

- **Filter bank**: in contrast to the hybrid filter bank of MP3, MPEG-2 AAC uses a plain Modified Discrete Cosine Transform (MDCT).
Increased window length (1024 instead of 576 spectral lines / transform), the MDCT outperforms the filter banks of previous coding methods.
- **Temporal Noise Shaping (TNS):** A true novelty.
Shapes the distribution of quantization noise in time by prediction in the frequency domain. In particular voice signals experience considerable improvement through TNS.
- **Prediction:** A technique commonly established in the area of speech coding systems. Certain types of audio signals are easy to predict.
- **Quantization:** by allowing finer control of quantization resolution, the given bit rate can be used more efficiently.
- **Bit-stream format**

TNS

Pour transitoires et pseudo-périodiques
(*transient and pitched signals*)
Problème de masquage temporel

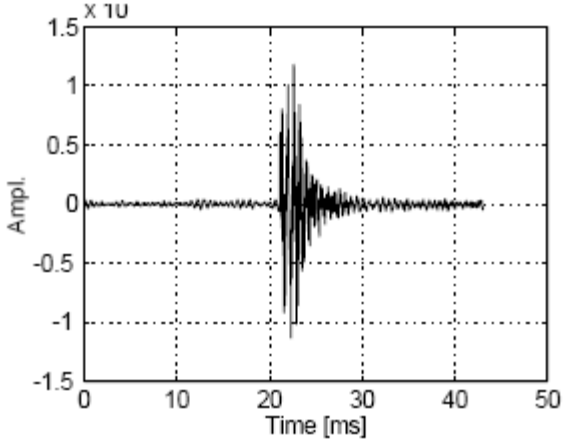
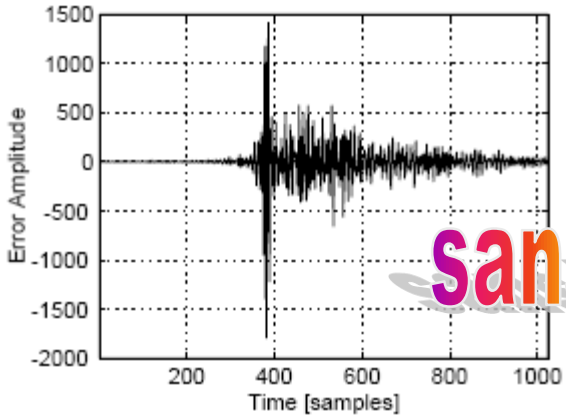


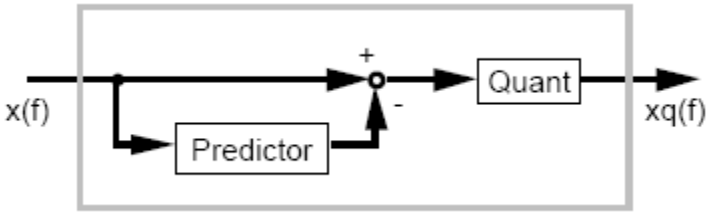
Figure 3: Transient signal (castanets, uncoded).

Utilisation d'une décomposition spectrale
=> Bruit étalé dans le temps
amène pb de pré-écho



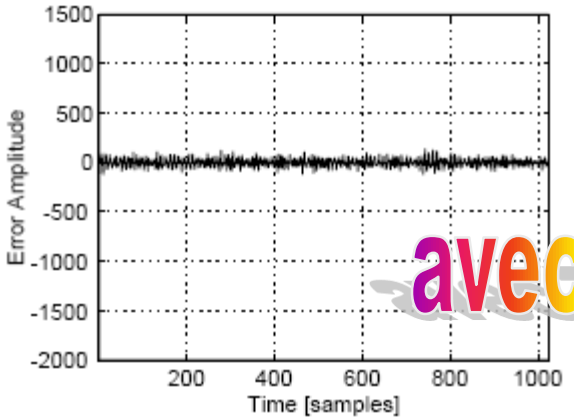
sans TNS

Spectre non-plat : codé par LP sur x(t)
Idem, signal « non-plat » : codé par LP sur X(f)



Codage prédictif en boucle ouverte

D*PCM (en B.O.) sur x(t) :
spectre de l'erreur adapté à Sx(f)
Idem en renversant les rôles ici (temps et fréquence)



avec TNS

Juin 2000 : comparaison de plusieurs codeurs (MP3 et MPEG 2 AAC compris)

http://www.ebu.ch/trev_dolby_frm.html

1998 : MPEG-4

Codeur orienté objet – Multimédia

« Coding of audio-visual objects »

Convergence de tous les codeurs audio :

codage audio haute-fidélité,

codage de la parole, jusqu'à

audio synthétique et parole synthétique,

de 2 à 64 kbps monophonique : 3 types de codeurs :

- codage parole bas débit (2 à 10 kbps) : codage paramétique
- codage parole débit moyen (6 à 16 kbps) : codage par AbS
- codage en sous-bandes / par transformées pour < 64 kbps

merci de votre attention