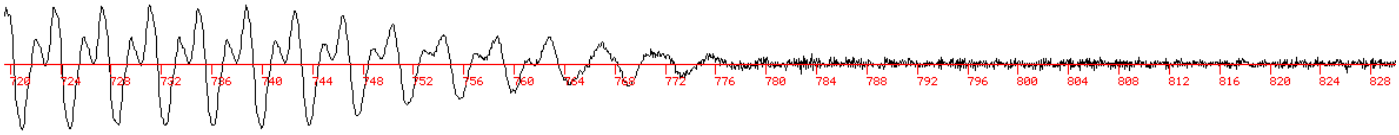
Analyse spectrale de la parole

Yves Laprie

16 octobre 2009

Plan

Pourquoi analyser [19] spectralement le signal de parole ?



1. **Spectrogramme**
2. **Spectrogramme de parole**
3. **Lissages cepstraux**
   (a) Quelle utilisation du lissage cepstral ?
   (b) Coefficients Mel cepstraux
   (c) Enveloppe vraie
   (d) Cepstres discrets
4. **Prédiction linéaire**
   (a) Prédiction linéaire sélective
   (b) Prédiction linéaire perceptive
   (c) Prédiction linéaire perceptive RASTA
5. **Caractérisation spectrographique des sons de la parole**
6. **Deux problèmes d'analyse de la parole**
7. **Détection et modification de la fréquence fondamentale**
   (a) Détection du fondamental (F0)
   (b) Méthode par recouvrement et addition (OLA)
   (c) PSOLA (Pitch Synchronous Overlap and Add)
   (d) Modèles sinusoïdaux et harmoniques

1 Spectrogramme

Objectif : connaître l'évolution temporelle du spectre de parole

- Transformée de Fourier Discrète

$$F(k) = \sum_{n=0}^{N-1} f(n) exp(-j\frac{2\pi}{N}kn)$$

- Pour utiliser la transformée de Fourier discrète il faut un signal périodique $\longrightarrow$ fenêtrage du signal
- Soit T la période d'échantillonnage, f le signal de départ, w la fenêtre appliquée au signal, F_w sa transformée de Fourier, on a :

$$F_w(\omega) = \sum_{n=-\infty}^{\infty} w(nT) f(nT) exp(-j\omega nT)$$

  où $w(nT) = 0$ pour $|n| > N/2$ avec N pair et w(-nT) = w(nT) avec $w(N/2) = 0$)

- F_w est la transformée d'un produit, c'est donc la convolution des transformées.

$$F_w(\omega) = \int_{-\infty}^{\infty} F(x) W(\omega - x) dx/2\pi$$

$$F_w(\omega) = F(\omega) * W(\omega)$$

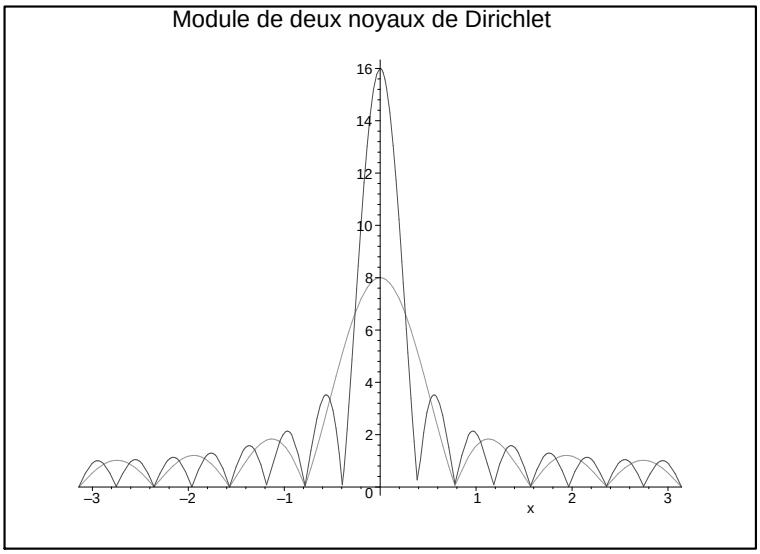
- Exemple très simple : fenêtre rectangulaire $w(nT) = 1$ (voir [6] par exemple)

$$W(\omega) = \sum_{n=-N/2}^{N/2-1} w(nT) exp(-j\omega nT)$$

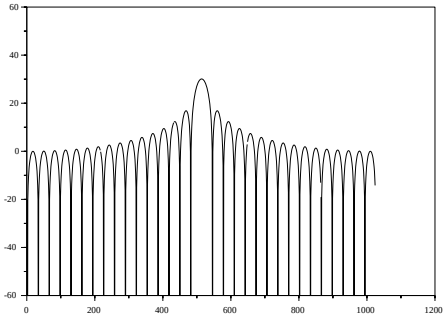
  $W(\omega) = exp(j\frac{\omega T}{2})\frac{sin(\frac{N}{2}\omega T)}{sin(\frac{1}{2}\omega T)}$ (noyau de Dirichlet)

- Attention si on somme de 0 à $N-1$ (cas de la transformée de Fourier discrète) alors $W'(\omega) = exp(-j\omega\frac{N}{2}T) \times W(\omega)$

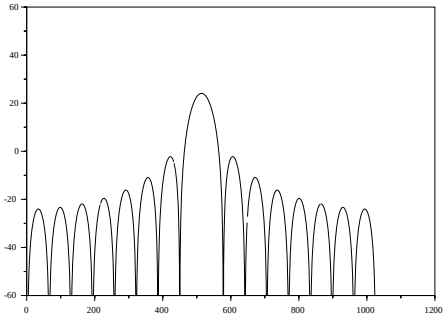
Importance de la fenêtre



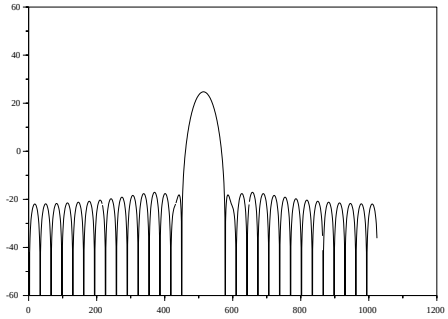
- Enjeu : trouver une bonne fenêtre qui réduit la convolution dont la largeur du lobe central est faible et la proéminence du lobe central forte.
- Difficulté : pour réduire la largeur du lobe central il faut augmenter N, donc amplifier les effets de moyenne.
- Exemples de fenêtres (rectangulaire, triangulaire et Hamming) représentées en dB ($20 * log10(|H(n)|)$)
- Remarque : multiplier un signal par 2 élève son spectre de 6 dB.



rectangulaire



triangulaire

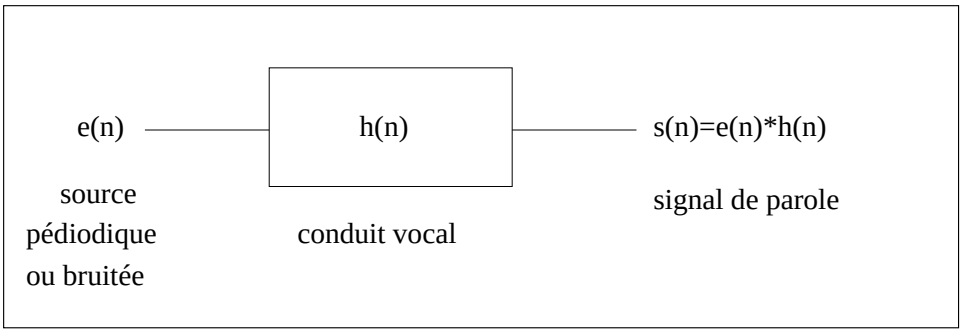


hamming

- Deux fenêtres célèbres :
  - Hamming: $w(n) = 0.54 - 0.46cos(2 * \pi/N * n)$ pour $n = 0, 1, 2, ..., N - 1$
  - Hanning: $w(n) = 0.5 + 0.5cos(2 * \pi/N * n)$ pour $n = -N/2, ..., -1, 0, 1, ..., N/2 - 1$
    $F(k)_{|Hanning} = \frac{1}{2}[F(k) + \frac{1}{2}(F(k-1) + F(k+1))]$
- L'effet de la convolution est donc de lisser le spectre :
  - fenêtre longue $\rightarrow$ lissage faible
  - fenêtre courte $\rightarrow$ lissage fort

2 Spectrogramme de parole [8]

- Modèle simplifié de la production de la parole



  Le spectre utile s'étend de 0 à 8kHz.

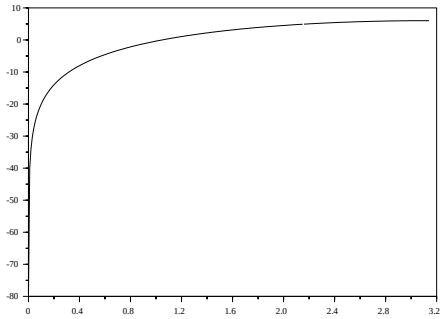
- Caractéristiques de la source :
  - bruit pour les fricatives et les bruits d'explosion
  - vibration des cordes vocales pour les sons vocaliques (voyelles en particulier).
    $\rightarrow$ fréquence fondamentale
    120Hz pour un locuteur, 200Hz pour une locutrice

- Calcul d'un spectrogramme : DFT d'une fenêtre de 4 à 32 ms qui se déplace de la moitié de sa durée. Avec une fréquence d'échantillonnage de 16kHz, cette fenêtre a donc entre 64 et 512 points.
  $\longrightarrow$ DFT entre 64 et 512 points
  $\longrightarrow$ pour lisser le spectre on utilise en fait une DFT avec plus de points (au moins 256) ce qui permet d'interpoler le spectre plus finement.
  $\longrightarrow$ "zero padding" : le signal de départ complété par des zéros. Si on utilise une DFT de 512 points et que la fenêtre a 64 points on complète par (512 - 64) zéros.

- fenêtre plus courte que la période fondamentale (spectre à large bande) $\longrightarrow$ fort lissage fréquentiel et pas d'harmonique

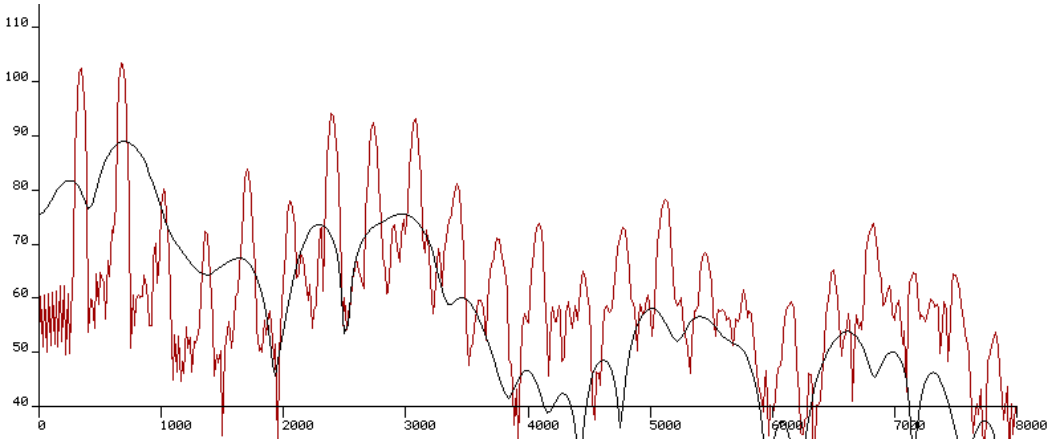
- fenêtre plus longue (spectre à bande étroite) $\longrightarrow$ faible lissage fréquentiel et visualisation des harmoniques

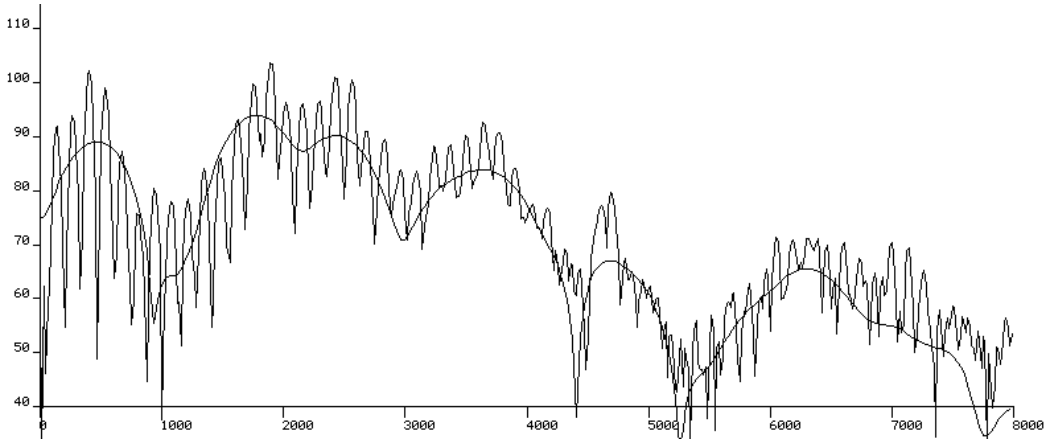
- Pour renforcer la contribution des hautes fréquences on préaccentue le signal : $u(n) = s(n) - s(n-1)$
  En passant à la transformée en z l'effet est donc $1 - z^{-1}$ c'est-à-dire un renforcement de la contribution des hautes fréquences.



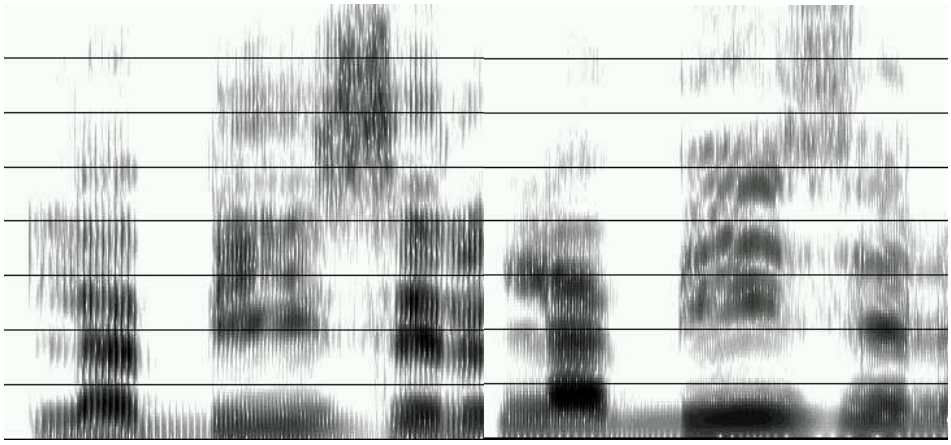
  Effet de la préaccentuation (dB entre 0 et Π)

- Exemples (locuteur en bas, locutrice en haut) :

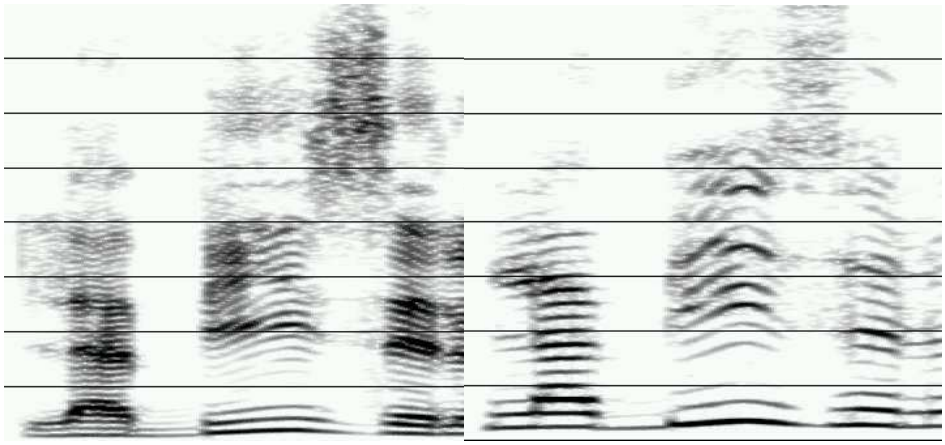




- Exemples de spectrogrammes



Spectrogrammes à bande large (locuteur à gauche, locutrice à droite)



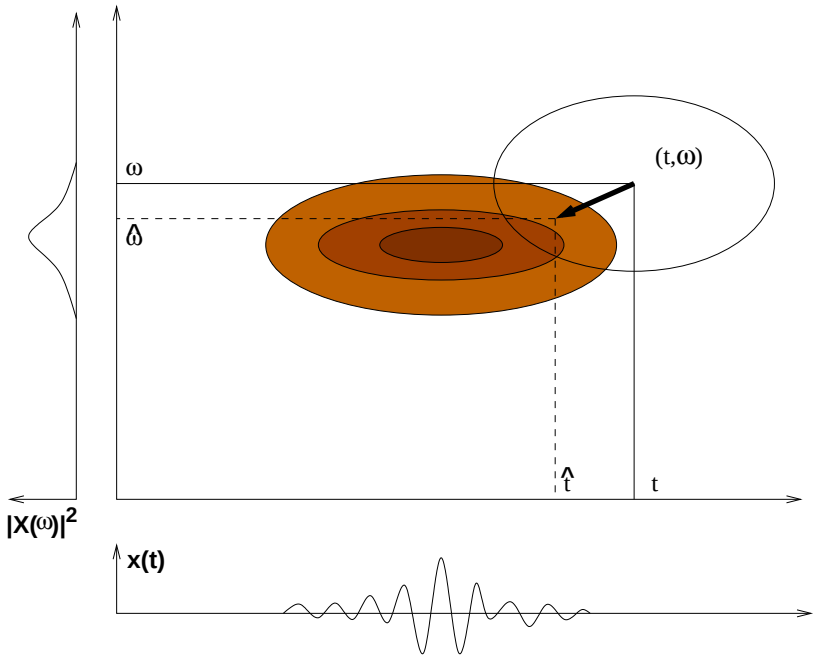
Spectrogrammes à bande étroite (locuteur à gauche, locutrice à droite)

-

Comment lisser le spectre ? Comment séparer la contribution du conduit vocal de celle de l'excitation ?

2.1 Amélioration de la précision du spectrogramme par la méthode de réassignation spectrale

- l'énergie de $X(t,\omega)$ qui est l'énergie calculée sur une fenêtre centrée en (t,ω) est affichée (Figure d'après [2]) à (t,ω) alors que l'énergie n'est pas forcément centrée en ce point[2, 16].

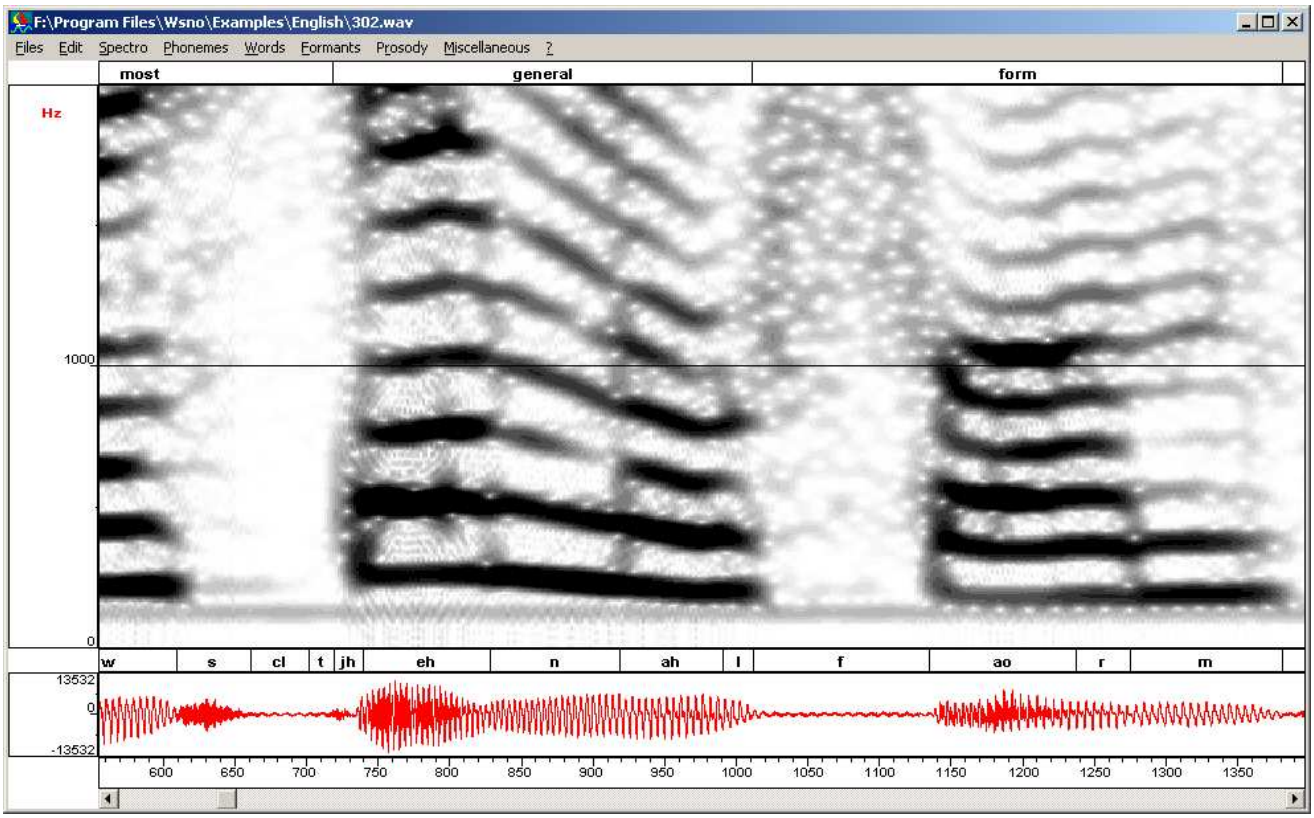


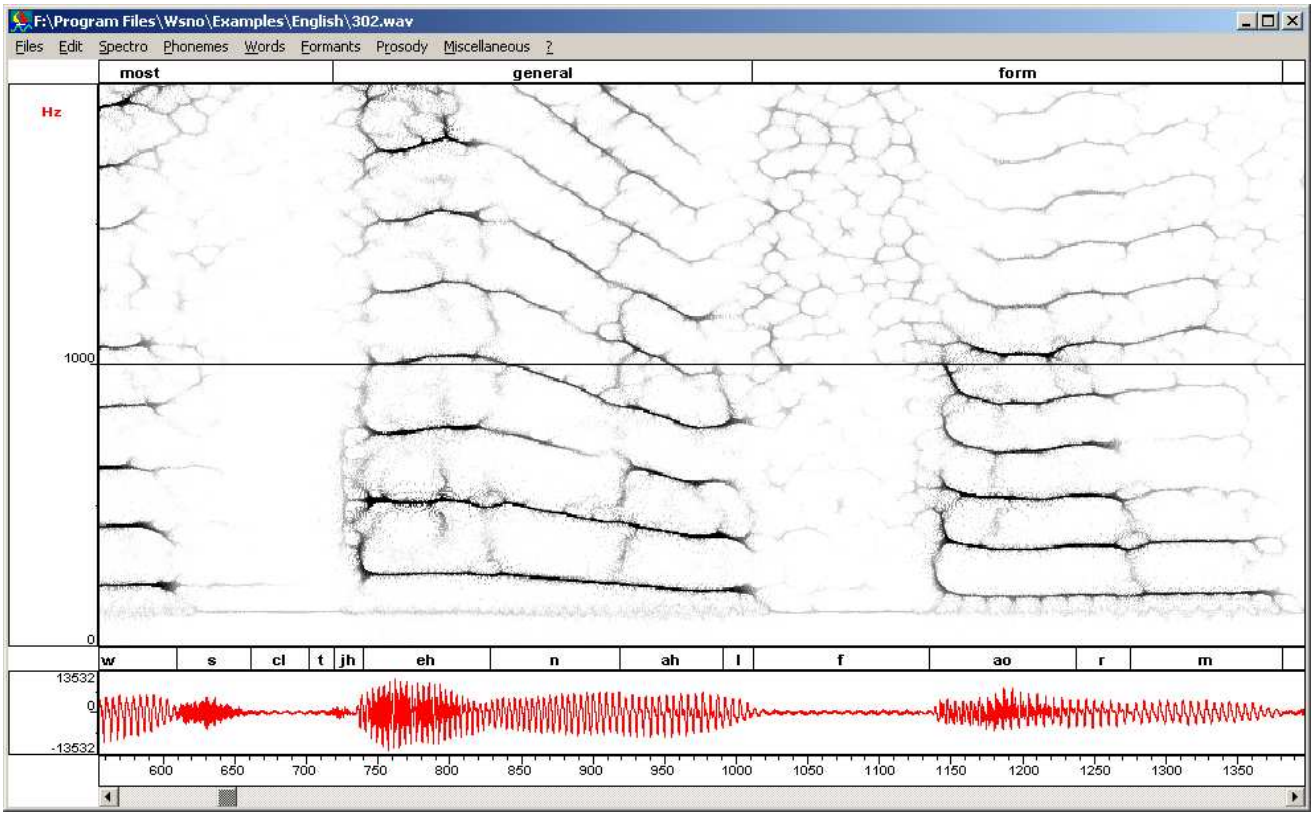
- l'idée est de corriger la position du point où apparaît l'énergie là où l'énergie de la fenêtre est effectivement présente.
  $t_r(x;t,\omega) = t - \mathcal{R}e\frac{STFT_{th}(x;t,\omega)STFT_h^*(x;t,\omega)}{|STFT_h(x;t,\omega)|^2}$
  $\omega_r(x;t,\omega) = \omega + \mathcal{I}m\frac{STFT_{dh}(x;t,\omega)STFT_h^*(x;t,\omega)}{|STFT_h(x;t,\omega)|^2}$ où h est la fenêtre d'analyse, th la fenêtre multipliée par le temps et dh la dérivée de la fenêtre.

- l'énergie affichée en un point est la somme de toutes les énergies déplacées en ce point

$$Reass(x;t',\omega') = \int\int STFT(x;t,\omega)\delta(t'-\hat{t}(x;t,\omega)).\delta(\omega'-\hat{\omega}(x;t,\omega))dt\frac{d\omega}{2\pi}$$

- $\longrightarrow$ meilleure résolution temporelle et fréquentielle (exemple avec fenêtre de 32ms)


F:\Program Files\Wsno\Examples\English\302.wav
Files Edit Spectro Phonemes Words Formants Prosody Miscellaneous ?
most
general
form
Hz
1000
0
w s cl t jh eh n ah l f ao r m
13532
0
-13532
600 650 700 750 800 850 900 950 1000 1050 1100 1150 1200 1250 1300 1350


F:\Program Files\Wsno\Examples\English\302.wav
Files Edit Spectro Phonemes Words Formants Prosody Miscellaneous ?
most
general
form
Hz
1000
0
w s cl t jh eh n ah l f ao r m
13532
0
-13532
600 650 700 750 800 850 900 950 1000 1050 1100 1150 1200 1250 1300 1350

3 Lissages cepstraux

Est-il possible de séparer dans le signal de parole $s(n) = e(n) * h(n)$ les contributions de l'excitation et du conduit vocal ?

- traitement homomorphique [18, 4] :
  - Signal $x(n) = x_1(n) * x_2(n)$
  - Transformée de Fourier (pour passer de la convolution à une multiplication) $X(\omega) = X_1(\omega)X_2(\omega)$
  - Logarithme $\hat{X}(\omega) = ln[X(\omega)] = ln[X_1(\omega)] + ln[X_2(\omega)] = \hat{X}_1(\omega) + \hat{X}_2(\omega)$
  - Transformée de Fourier inverse. Le signal revient dans le domaine temporel mais il reste additif. $\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n)$
  - Traitement linéaire, par exemple le liftrage (ne conserver que les premiers coefficients cepstraux pour éliminer la contribution de la source). $\hat{y}(n) = \hat{x}_1(n)$
  - Transformée de Fourier $\hat{Y}(\omega) = \hat{X}_1(\omega)$
  - Exponentielle. $Y(\omega) = X_1(\omega)$
  - Transformée de Fourier inverse. $y(n) = x_1(n)$
  - Signal déconvolué $x_1(n)$

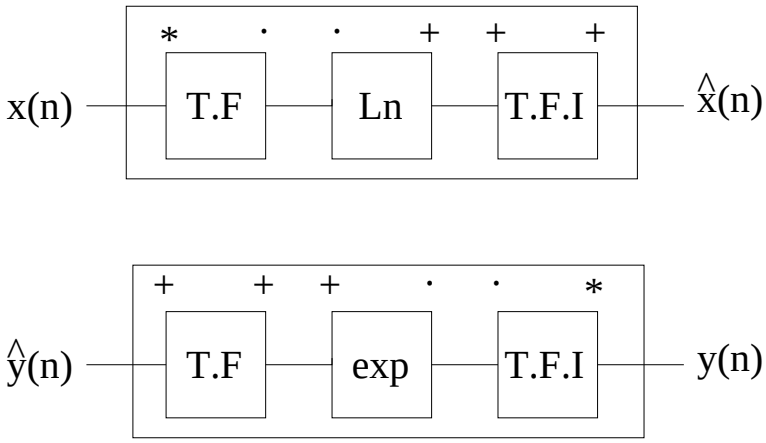
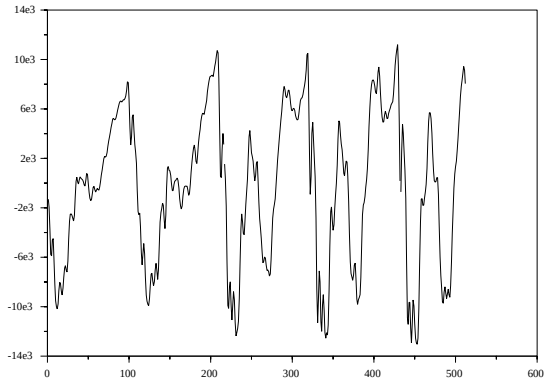
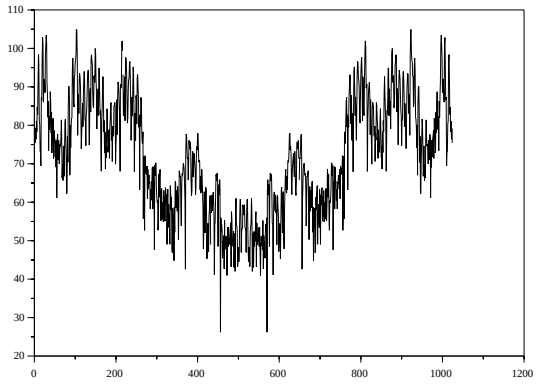


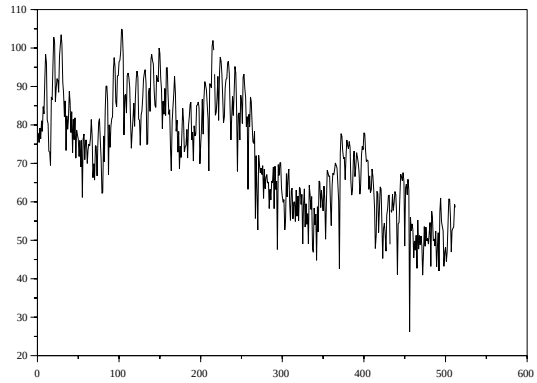
Schéma homomorphique



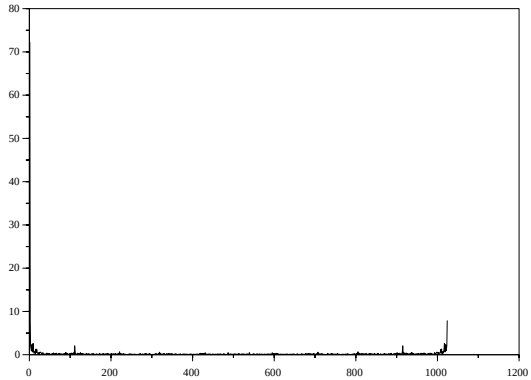
signal temporel avant hamming



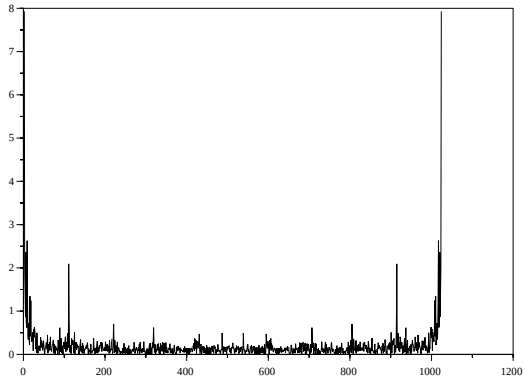
spectre (0 à 2π)



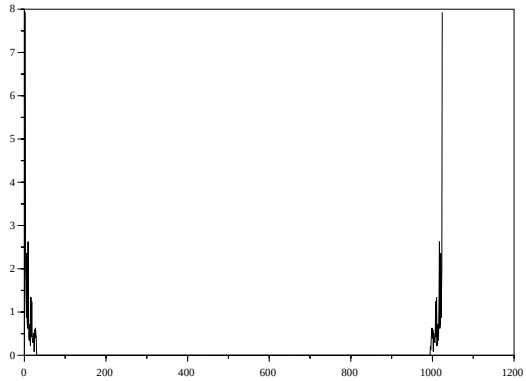
spectre (0 à π)



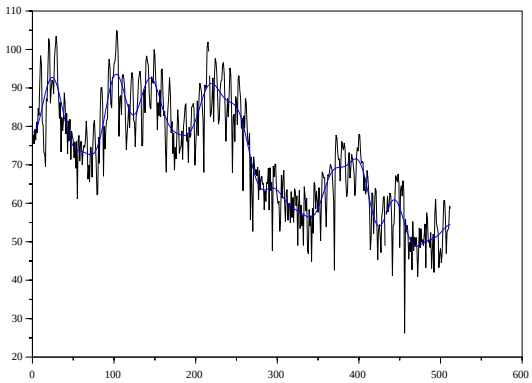
cepstre



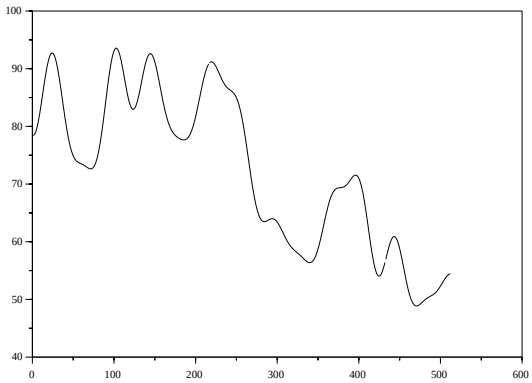
cepstre sans le premier coef.



cepstre liftré sans le premier coef.



spectre lissé(0 à π)



spectre lissé(0 à π)

Remarques :

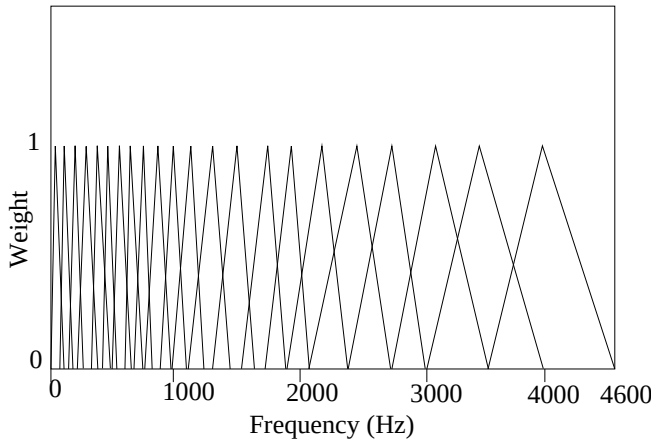
- le lissage cepstral est un lissage, le spectre résultat ne passe donc pas par les harmoniques.
- il faut connaître, ou au moins avoir une idée, de la fréquence fondamentale pour choisir le liftrage.
- le cepstre est, au signe près de la TF, la TF du spectre donc le liftrage correspond à l'élimination des "hautes quéfrences" du spectre.
- considérations d'implantation : la transformée de Fourier inverse est appliquée à un "signal" réel symétrique, on peut donc utiliser une transformée en cosinus.
  $X(k) = \sum_{n=0}^{N-1} s(n) e^{-j\frac{2\pi}{N}kn}$
  Si $s(n)$ est réel et symétrique ($s(N-m) = s(m)$) alors
  $X(k) = s(0) + (-1)^k s(N/2) + 2\sum_{n=1}^{N/2-1} s(n) cos\frac{2\pi}{N}kn$
- généralement on ne revient pas au signal temporel déconvolué.

3.1 Quelles utilisations du lissage cepstral ?

1. fournir un vecteur spectral pour la reconaissance automatique de la parole (coefficients mel cepstre)
2. lisser le spectre de parole pour trouver les formants (fréquences de résonance du conduit vocal)
3. retrouver les signaux d'excitation et le filtre correspondant au conduit vocal,

3.2 Coefficients Mel cepstraux [1]

- Point de départ : la contribution à la perception des sons de la parole des hautes fréquences est plus faible que celle des basses fréquences. $\longrightarrow$ changement d'échelle.
- échelle Mel : linéaire en basse fréquence, logarithmique en haute fréquence.



$$M = \frac{1000}{Log2} log(1 + \frac{f}{1000})$$

- échelle Bark (à peu près identique) :

```
float bark(float f) // f en Hz
{
if (f <= 200.0) return(((9.88e-3 + 6.222e-8 * f) * f));
else return((26.81 * f / (1960.0 + f) - 0.53));
}
```

- Mise en œuvre :
  1. Calcul du spectre sur une fenêtre plus longue que le fondamental (environ 20ms)
  2. Calcul de l'énergie en sortie des filtres traingulaires X_k pour $k = 1, 2, \ldots, 20$
  3. Calcul des cepstres Mel

$$MFCC_i = \sum_{k=1}^{N} X_k cos[i(k - \frac{1}{2})\frac{\pi}{N}]$$

  4. Récupération des premiers coefficients pour la reco (en général $N/2$).

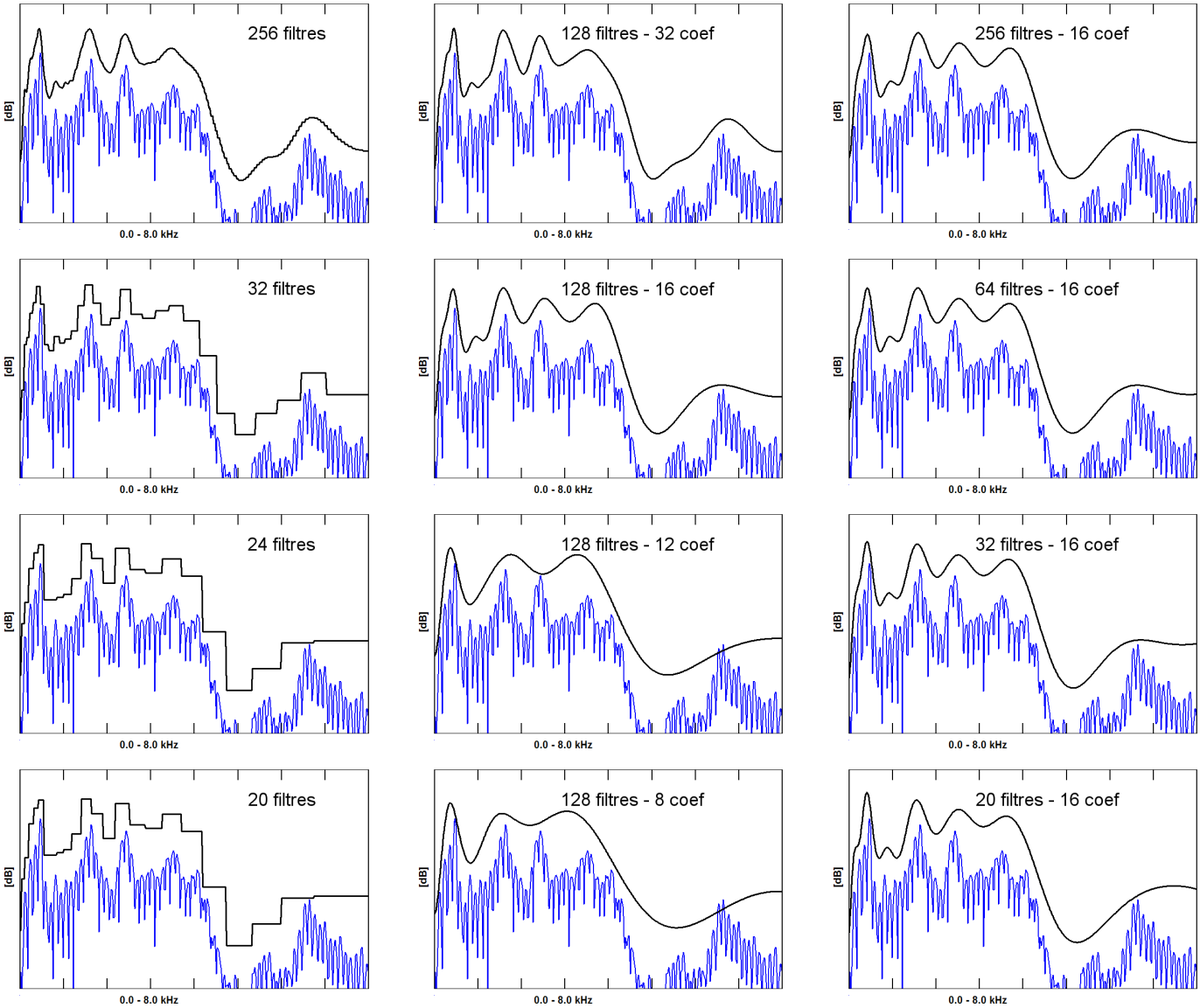
- généralement on prend des fenêtres entre 20 et 30ms, N égal à 24 et un déplacement égal à la moitié de la fenêtre.

- Comment prendre en compte le bruit :

  bruit additif supprimé après le calcul du spectre

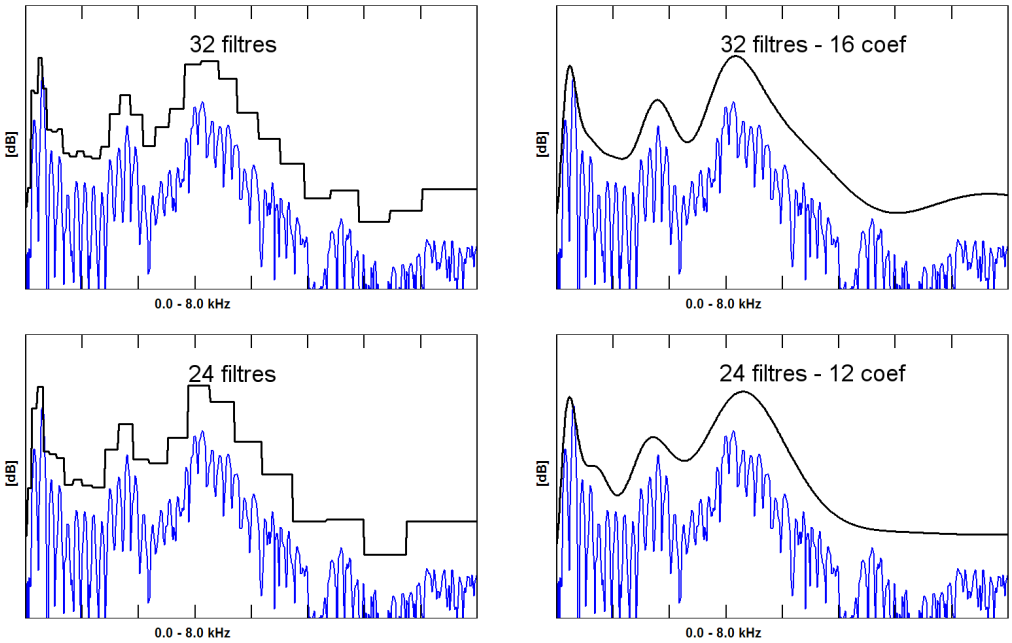
  bruit convolutif supprimé après le calcul des cepstres

Effet comparatif du nombre de filtres Mel

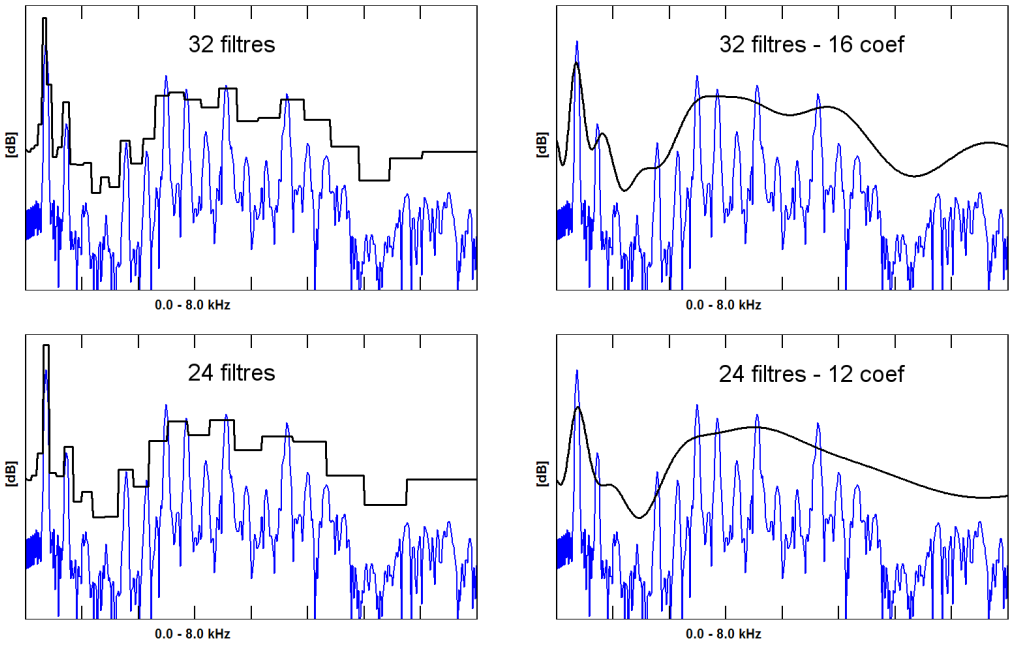


Nombre de filtres Mel (colonne de gauche), nombre de coefficients conservés par le liftrage (colonne du centre) et nombre de filtres en fixant l'ordre du liftrage (nombre de coefficients conservés). La courbe du bas représente la transformée de Fourier à bande étroite sur la même fenêtre temporelle (32 ms).

Effet des filtres et du liftrage Mel sur un voix masculine (voyelle [i])



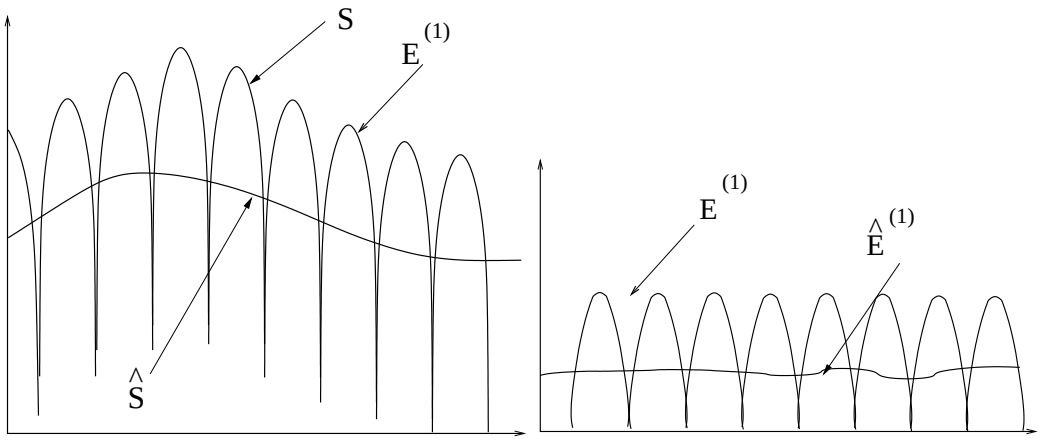
Effet des filtres et du liftrage Mel sur un voix féminine (voyelle [i])



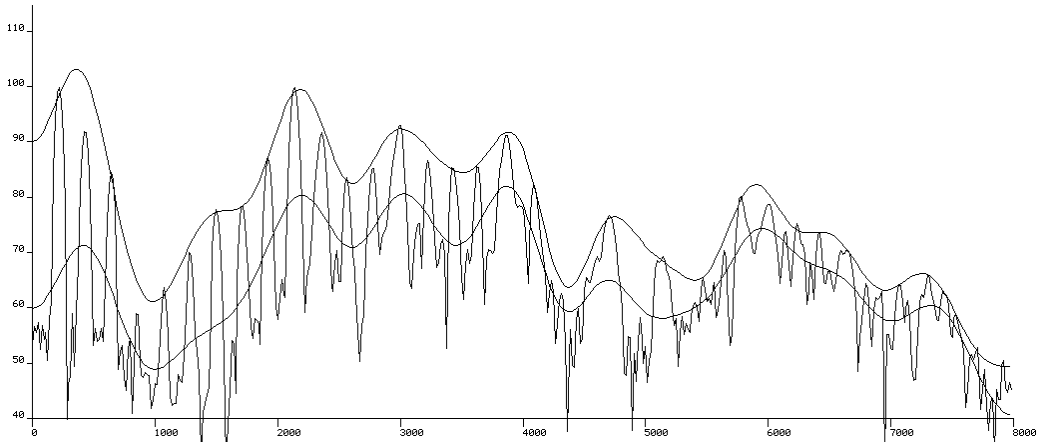
3.3 Enveloppe vraie

- But : obtenir un spectre lissé [5] qui passe par les harmoniques (ce que l'oreille perçoit effectivement).
- Idée : partir du lissage classique et le corriger itérativement en écartant la contribution des valeurs situées au-dessous du spectre lissé.

S spectre
$V^{(1)} = \hat{S}$ (lissage cepstral classique)
$E^{(1)} = g(S - \hat{S})$ où $g(y)$ = si $y > 0$ alors y sinon 0 fsi
$E^{(1)}$ représente les dépassements de S sur $\hat{S}$
$\hat{E}^{(1)}$ est le spectre de dépassement lissé cepstralement :



- Algorithme :
  1. solution initiale
     $\hat{E}^{(1)} = \sum_{m=0}^{N-1} e_m^{(1)} h_m cos(\frac{2}{N} mk)$ où $e^{(1)} = IDFT(E^{(1)})$ et h_m est la fenêtre de liftrage.
  2. itération $i+1$
     soit $V^{(i)}$ l'enveloppe obtenue à l'étape précédente, $E^{(i)}$ et $\hat{E}^{(i)}$ les dépassements et lissage du dépassement correspondants.
     $V^{(i+1)} = V^{(i)} + \hat{E}^{(i)}$
     $E^{(i+1)} = g(E^{(i)} - (1+\alpha)\hat{E}^{(i)})$ où α est un coefficient accélérateur.
     $\hat{E}^{(i+1)} = DFT(h(IDFT(E^{(i+1)})))$
  3. fin ou nouvelle itération



spectre bande étroite, lissé cepstralement et enveloppe vraie

- Avantages :
  - pas de battement d'énergie dû à la place de la fenêtre par rapport à la période,
  - spectre passant par les harmoniques.

3.4 Cepstres discrets [3]

- Idée : n'évaluer le spectre qu'aux fréquences des harmoniques, ou des points du spectre qui doivent être pris en compte,
- Point de départ : il faut disposer des valeurs du spectre pour les harmoniques (une estimation de F0 est nécessaire).

Soit $P(\omega_k)$, $1 \leq k \leq n$ les points du spectre à approcher, et $X(\omega_k)$, $1 \leq k \leq n$ les points fournis par l'interpolation. En choisissant la famille des cosinusoïdes $cos(i\omega)$, $0 \leq i \leq p$ pour interpoler le spectre

$$X(\omega_k) = \sum_{i=0}^{p} c_i cos(i\omega_k)$$

où c_i est le coefficient de la i^{ème} harmonique. Erreur d'approximation:

$$E = \sum_{k=1}^{n} (\sum_{i=0}^{p} c_i cos(i\omega_k) - P(\omega_k))^2$$

En annulant chacune des dérivées de E par rapport à c_i on obtient :

$$\sum_{k=1}^{n} (\sum_{j=0}^{p} c_j cos(j\omega_k) - P(\omega_k)) \times cos(i\omega_k) = 0$$

soit sous forme matricielle $A.C = B$ avec :

$$a_{ij} = \sum_{k=1}^{n} cos(i\omega_k) cos(j\omega_k)$$

$$b_i = \sum_{k=1}^{n} P(\omega_k) cos(i\omega_k)$$

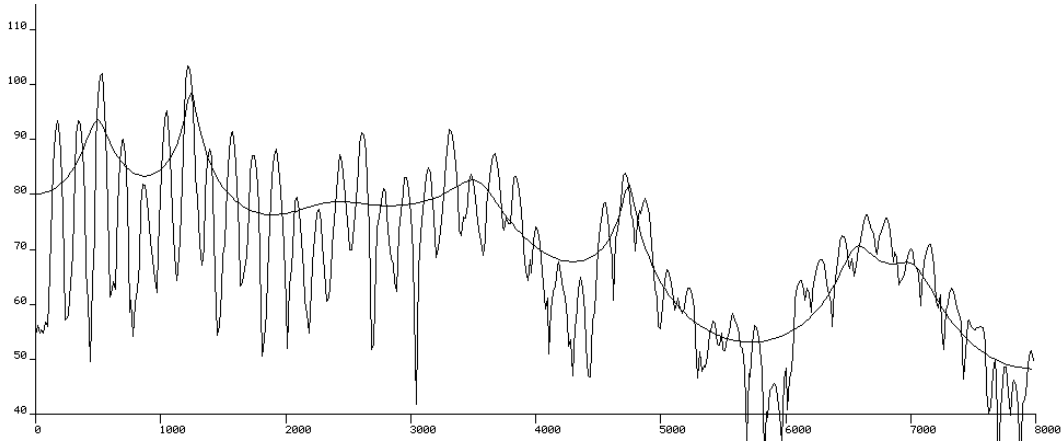
et C le vecteur des p coefficients inconnus c_i.

Voir l'article de Gallas et Rodet [3] qui donne de plus amples détails sur d'autres versions des cepstres itératifs (résolution rapide, prise en compte de l'incertitude de la position des pics spectraux, application aux cepstres Mel...)

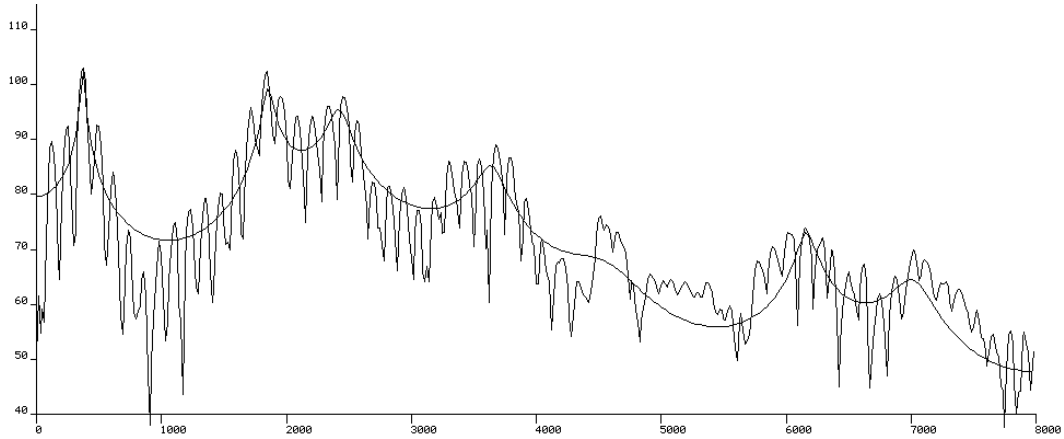
4 Prédiction linéaire [18]

- Origine : le signal de parole n'étant pas complètement aléatoire, les échantillons successifs sont corrélés. Peut-on utiliser cette corrélation pour réduire la quantité de données ?
- Principe :
  - $s(n)$ est représenté par la somme d'une combinaison linéaire des échantillons précédents et d'une erreur,
  - les coefficients de la combinaison linéaire sont trouvés de façon à minimiser l'erreur,
  - cette modélisation correspond à un modèle tout pôle :
    Soient $S(z)$, $H(z)$ et $U(z)$ les transformées en z du signal, du filtre du conduit vocal et de l'excitation. Un modèle tout pôle de $H(z)$ est :
    $$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}$$
- Calcul des coefficients a_k
  La prédiction de $s(n)$ est $\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k)$.
  L'erreur est $e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$.
  L'énergie de l'erreur est $E_n = \sum_m e^2(m) = \sum [s(m) - \hat{s}(m)]^2$
  En minimisant E_n par rapport aux a_k :
  $\frac{\partial}{\partial a_i} E_n = 0$ pour $(i = 1, 2, ..., p)$ on obtient :
  $\sum_m s_n(m-i) s_n(m) = \sum_{k=1}^{p} a_k \sum_m s_n(m-i) s_n(m-k)$ pour $(1 \leq i \leq p)$.
  En posant $\Phi_n(i,k) = \sum_m s_n(m-i) s_n(m-k)$ on obtient :
  $\sum_{k=1}^{p} a_k \Phi_n(i,k) = \Phi_n(i,0)$ pour $(1 \leq i \leq p)$.
- Deux méthodes suivant la sommation :
  - par autocorrélation : signal multiplié par une fenêtre,
  - par covariance : limite de la somme des erreurs
- Solutions de calcul efficaces (Toeplitz ou Cholesky).

- Exemples



Un exemple qui ne marche pas : /ε̃]/



Un exemple qui marche : /ε/

- Evaluation du spectre LPC
  Calculer $H(z)$ pour $z = exp(jk\frac{2\pi}{N})$ avec $0 \leq k \leq N/2 - 1$.
  $H(z) = \frac{G}{\sum_{k=0}^{p} b_k z^{-k}}$ avec
  $b_0 = 1$
  $b_k = -a_k$ pour $1 \leq k \leq p$
  $\longrightarrow$ DFT sur les b_k complétés par des zéros.

4.1 Prédiction linéaire sélective [10]

Idée : Appliquer le calcul des coefficients de prédiction sur une partie du spectre

- Comment obtenir les $\Phi(i,k)$ à partir du spectre ?
  éléments de réponse :

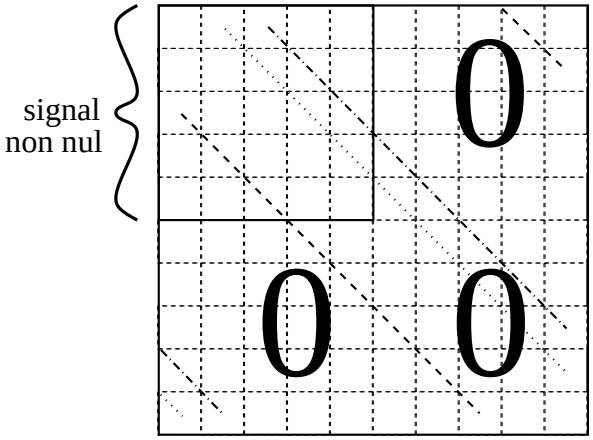
$$|X(k)|^2 = \sum_{n=0}^{N-1} s(n)e^{-j\frac{k2\pi}{N}n} \times \sum_{m=0}^{N-1} s(m)e^{+j\frac{k2\pi}{N}m}$$

- en réorganisant la somme précédente ([15] pages 556 et suivantes) :

$$|X(k)|^2 = \sum_{l=0}^{L-1}(\sum_{m=0}^{L-1} s(m)s_L(m+l))e^{-j\frac{k2\pi}{N}l}$$

  avec $l = n - m$ et $s_L(m+l) = s((m+l) \text{ modulo } L)$

- et en complétant le signal par suffisamment de zéros (M) pour que :
  $\sum_{n=0}^{N-1-l} s(n)s(n+l) = \sum_{m=0}^{L-1} s(m)s_L(m+l)$



Adjonction de zéros pour éviter le repliement

- Calcul des coef. d'autocorrélation par transformée de Fourier inverse
  $R(l) = \frac{1}{L}\sum_{k=0}^{L-1} |X(k)|^2 e^{j\frac{2\pi}{L}kl}$.

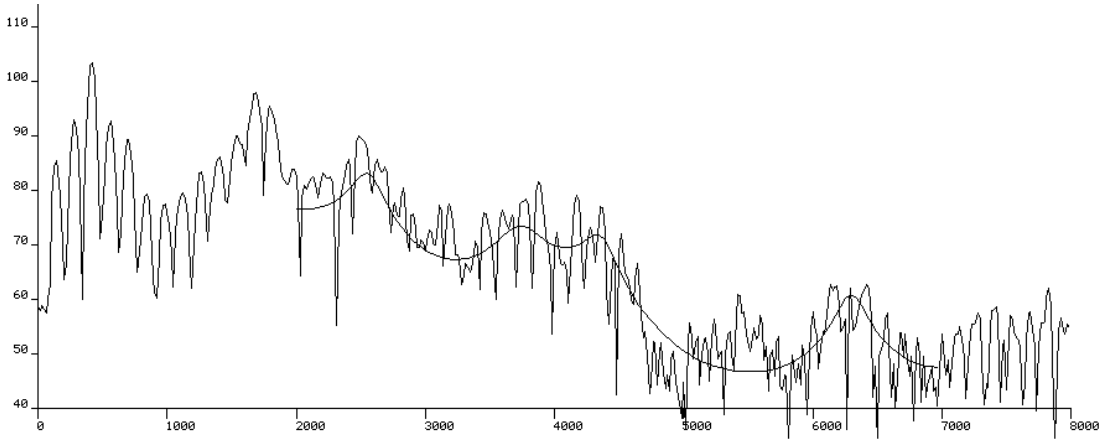
- L doit être la première puissance de 2 telle que $L \geq N + M$.

- Comme $|X(k)|^2$ donne un ”signal” réel et pair la TF se réduit à une somme de cos.
  1. si L = 2l

$$R(m) = \frac{2}{L}\sum_{k=1}^{l-1} S(k)cos\frac{2\pi}{L}km + \frac{1}{L}(S(0) + (-1)^m S(l))$$

  2. si L = 2l+1

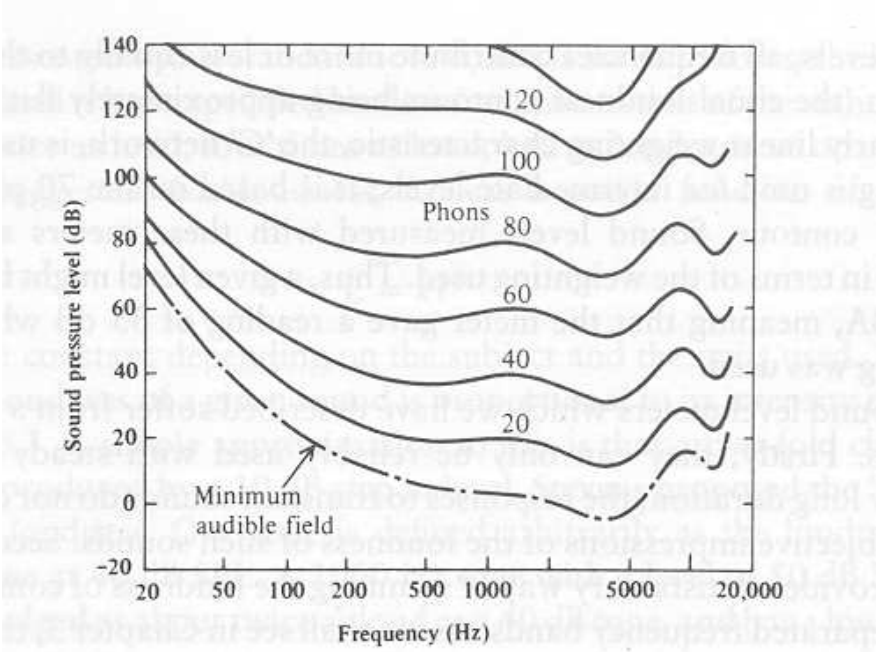
$$R(m) = \frac{1}{L}S(0) + \frac{2}{L}\sum_{k=1}^{l} S(k)cos\frac{2\pi}{L}km$$



Exemple de spectre LPC sélective entre 4000 et 7000 Hz

4.2 Prédiction linéaire perceptuelle [7]

4.2.1 Les origines psychoacoustiques



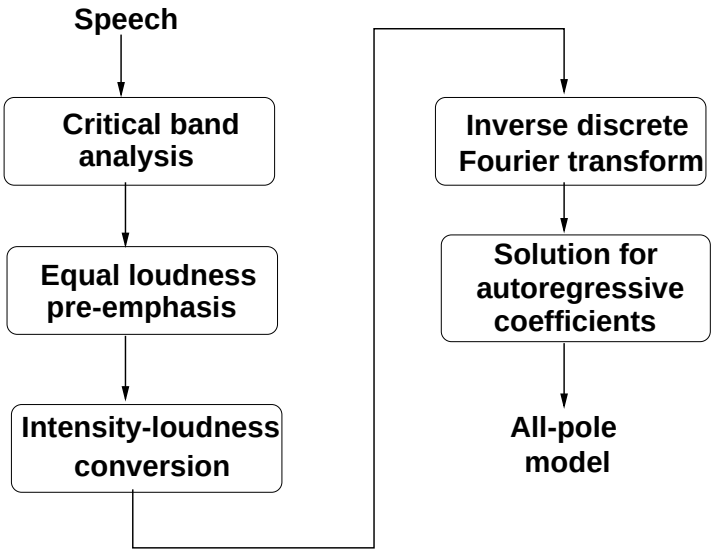
Contours d'égale intensité sonore [12] (Equal loudness contours)
d'après Robinson and Dadson 1956

- Pour passer du seuil d'audition à 100 Phons :
  - à 100 Hz il faut 79 dB,
  - à 1000 Hz il faut 97 dB.
- $\longrightarrow$ élever l'intensité sonore change "l'équilibre tonal" (vers les graves).
- Echelle d'intensité sonore (perception de l'élévation de l'intensité SPL)

$$L = kI^{0.3}$$

  où I est l'intensité physique (dB), L l'intensité sonore et k une constante.
- $\longrightarrow$ une augmentation de 10 dB double l'intensité sonore.

4.2.2 Mise en œuvre



Exemples avec une PLP d'ordre 5 (5 coefficients de prédiction).

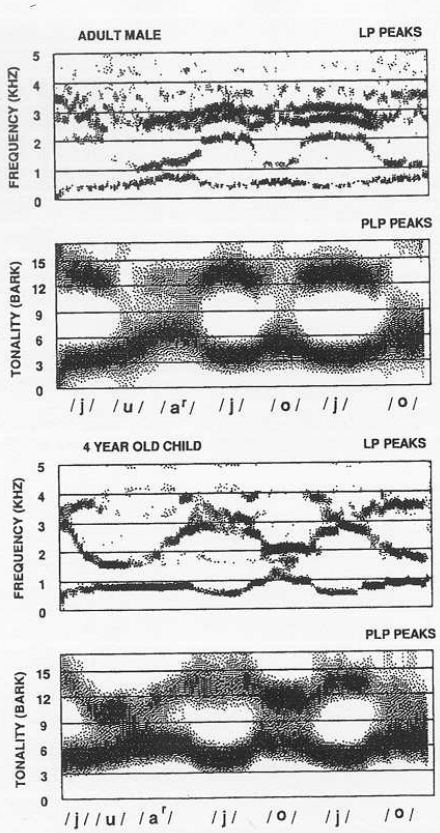
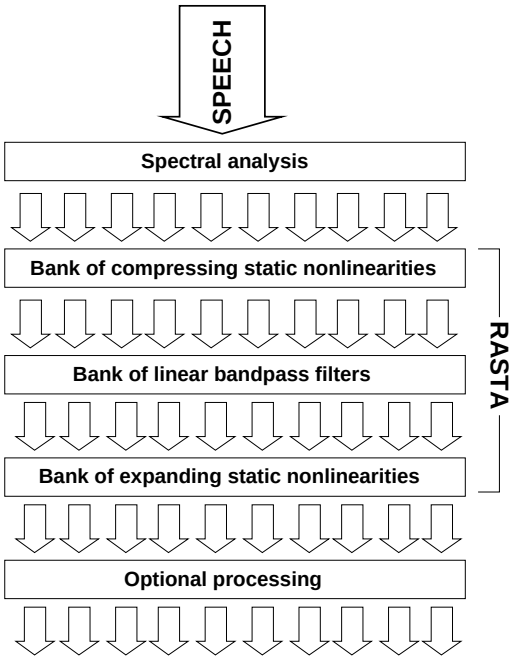


FIG. 13. Peak-enhanced spectrograms from conventional LP and PLP analyses of the utterance "You are yo-yo" uttered by adult male and a 4-year-old child. While for these two speakers LP analysis finds a different number of formants at different positions, PLP analysis finds the same number of peaks at similar positions.

4.3 Prédiction linéaire perceptive RASTA

- Origine : La perception humaine réagit aux valeurs relatives plus qu'aux valeurs absolues.
- Idée : éliminer les variations trop lentes ou trop rapides par filtrage sur le spectre d'amplitude.



Algorithme

1. Calcul le spectre d'amplitude en bandes critiques (comme pour la PLP),
2. Compression de l'amplitude à l'aide d'une transformation non linéaire,
3. Filtrage des trajectoires temporelles de chaque composante spectrale,
4. Expansion de l'amplitude à l'aide d'une transformation non linéaire,
5. Préaccentuation à l'aide du contour d'égale intensité sonore et prise en compte de l'échelle sonore par élévation à la puissance 0.33,
6. Calcul du modèle tout pôle du spectre selon la méthode PLP classique.

Filtre RASTA

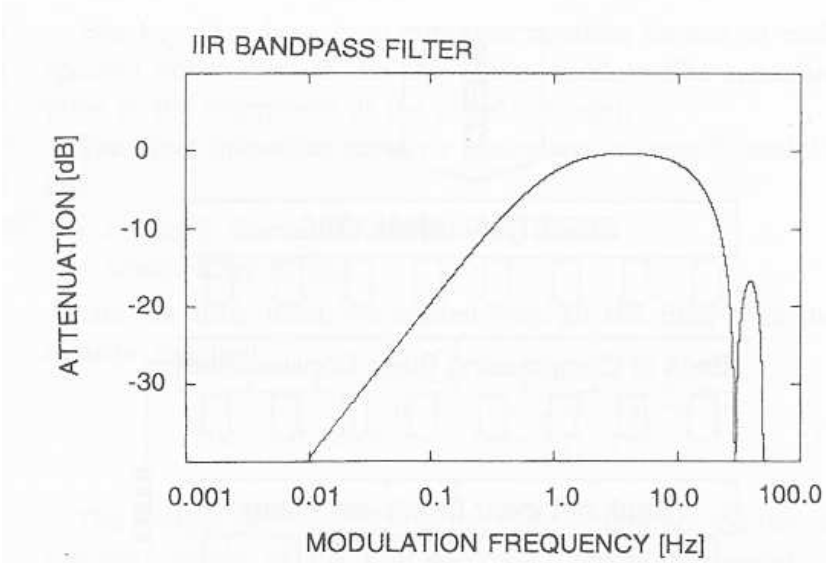
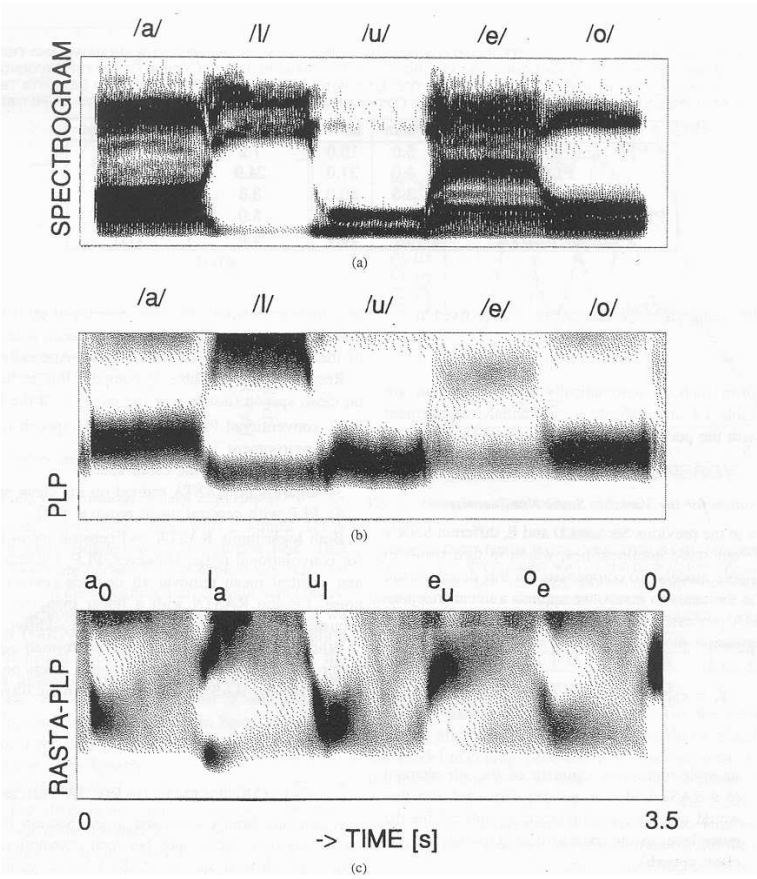


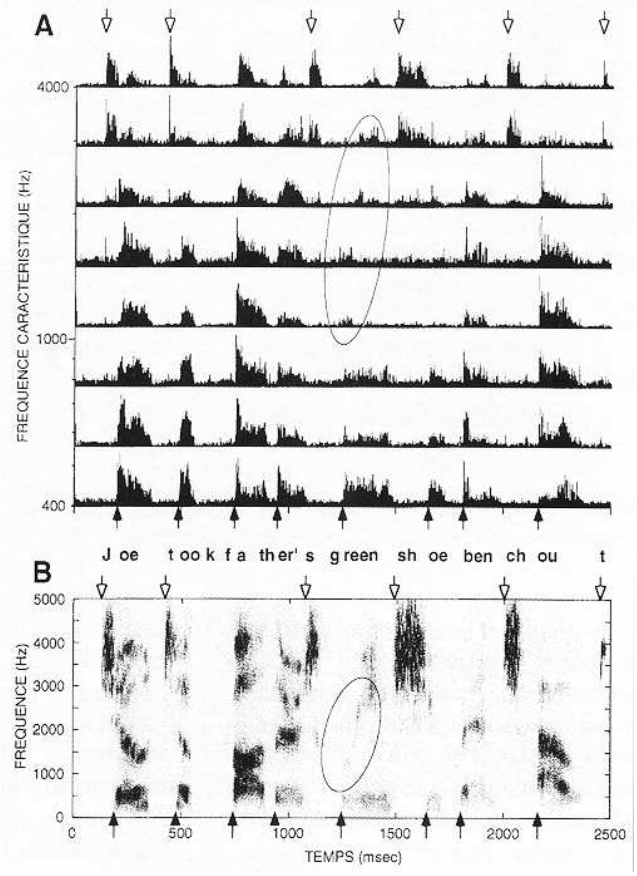
Fig. 3. Frequency response of RASTA band-pass filter.

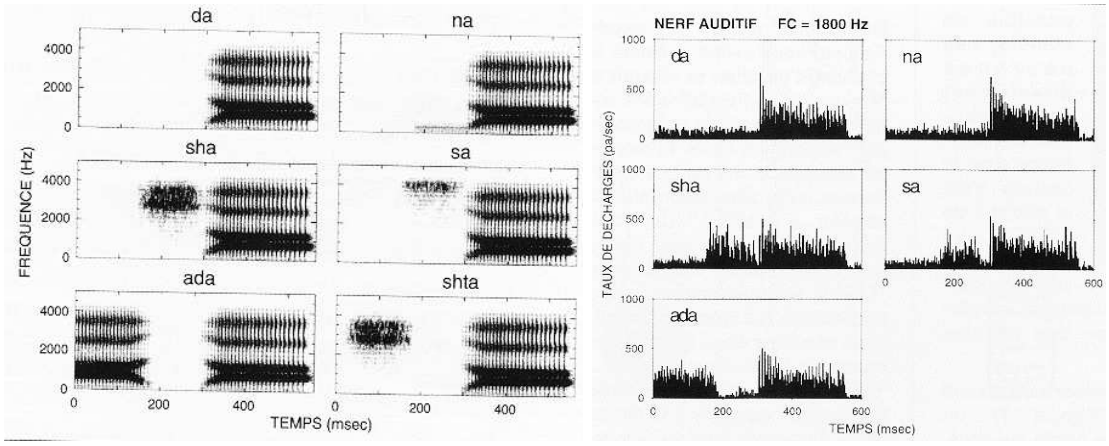


De haut en bas, spectrogramme FFT normal, PLP et RASTA-PLP.

4.4 Retour sur les bases de RASTA-PLP

Réponse du nerf auditif à une phrase (d'après Delgutte et Hammond 97) chez un chat.





Effet du contexte sur la réponse du nerf auditif (à gauche les syllabes synthétiques, à droite la réponse d'un neurone du nerf auditif).

4.5 Remarques sur la mise en œuvre

- l'étape de compression est plus ou moins bien adaptée au bruit éventuel :
  - bruit convolutif $\longrightarrow$ somme du spectre de parole et du bruit après le log
  - bruit additif $\longrightarrow$ le log n'améliore pas les choses du tout
- remplacement de l'étape de compression en log par un opérateur du genre $log(1 + Jx)$ où x est l'énergie dans le canal fréquentiel considéré et J une constante numérique.
- comportement :
  - linéaire si $J \ll 1$
  - logarithmique si $J \gg 1$
- Détermination de J :
  - Détermination pour différents niveaux de rapports signal sur bruit (en maximisant le taux de reconnaissance),
  - Choix de J en fonction du rapport signal sur bruit estimé sur le signal à traiter.

5 Caractérisation spectrographique des sons de la parole

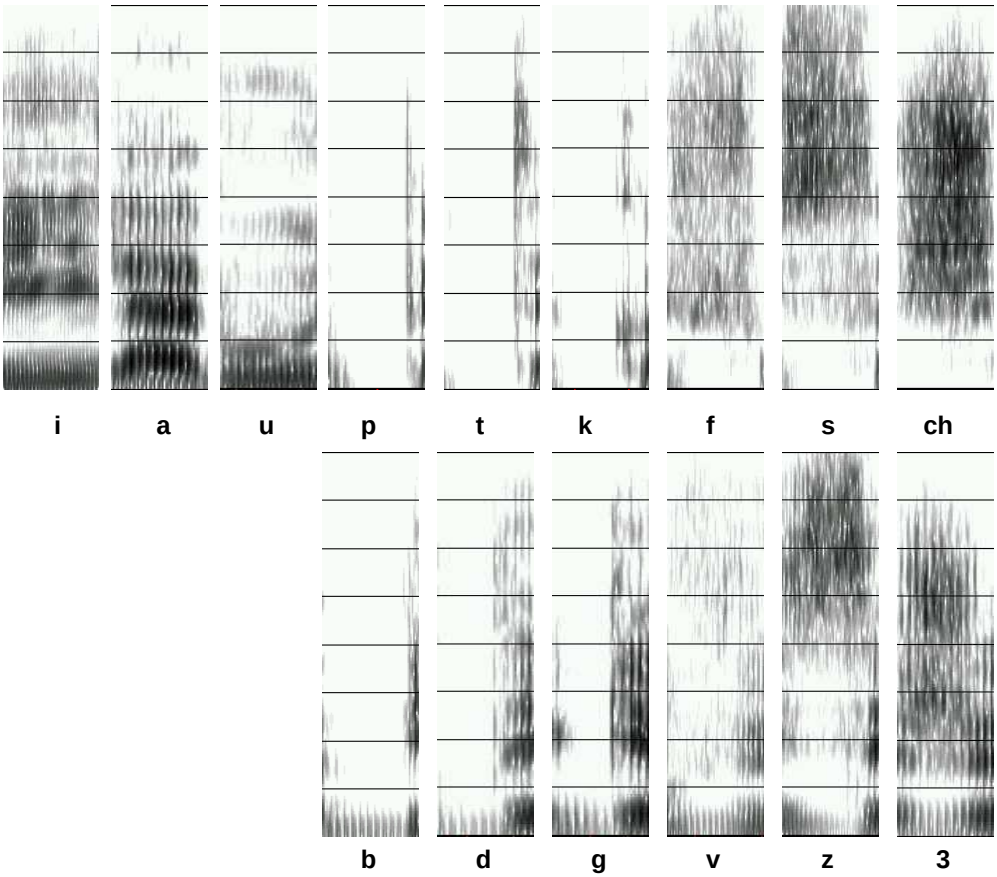
- Mode d'articulation

  vocalique vibration des cordes vocales (voisement) et constriction pas trop forte,

  fricatif fort resserrement du conduit vocal provoquant un bruit de friction,

  occlusif fermeture partielle ou totale du conduit vocal, augmentation de la pression derrière la constriction puis relâchement brutal de l'occlusion qui produit un bruit d'explosion (burst).

- Lieu d'articulation = lieu de la constriction principale du conduit vocal : pharynx, palais /k/, dents /t/, lèvres /p/



- Caractérisation acoustique et articulatoire des voyelles
  - les voyelles sont caractérisées acoustiquement par leurs formants (renforcement spectraux qui correspondent aux fréquences de résonance du conduit vocal)
  - les voyelles sont caractérisées articulatoirement par le lieu de la plus forte constriction du conduit vocal
  - parole continue $\longrightarrow$ **COARTICULATION**

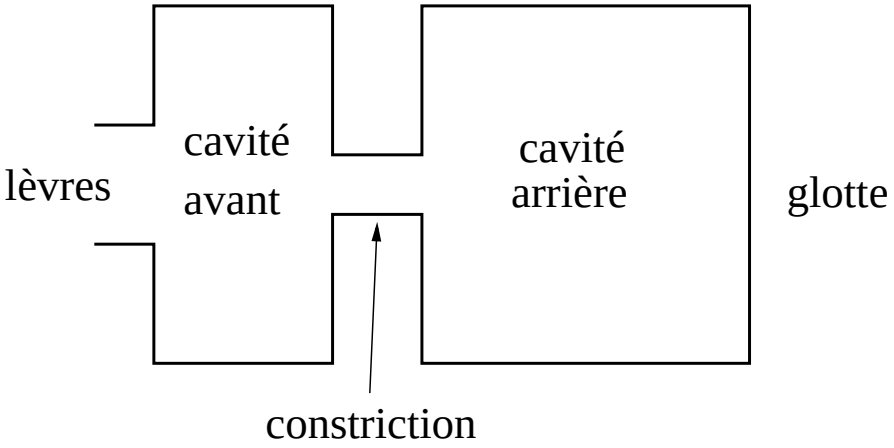
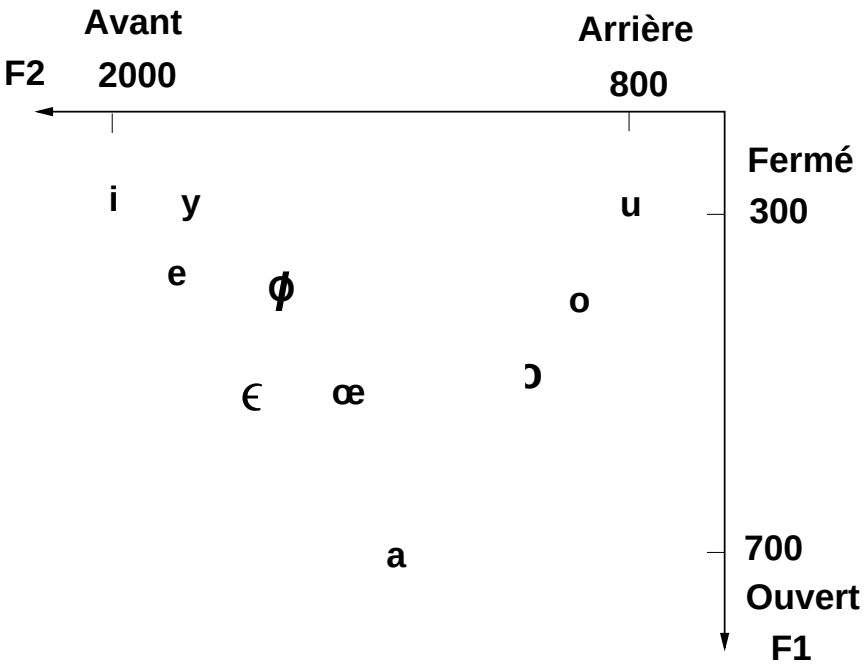
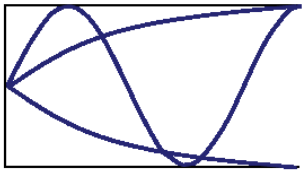


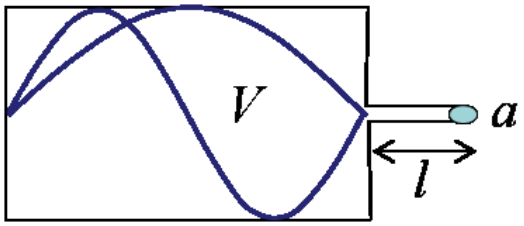
Schéma simplifié du conduit vocal



Triangle vocalique pour le français

5.1 Approximation du conduit vocal sous la forme de tubes

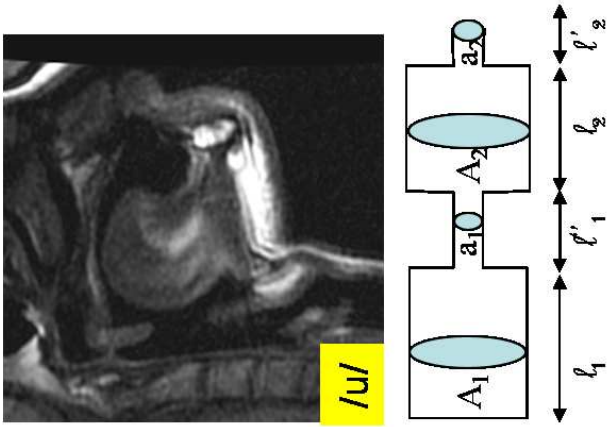


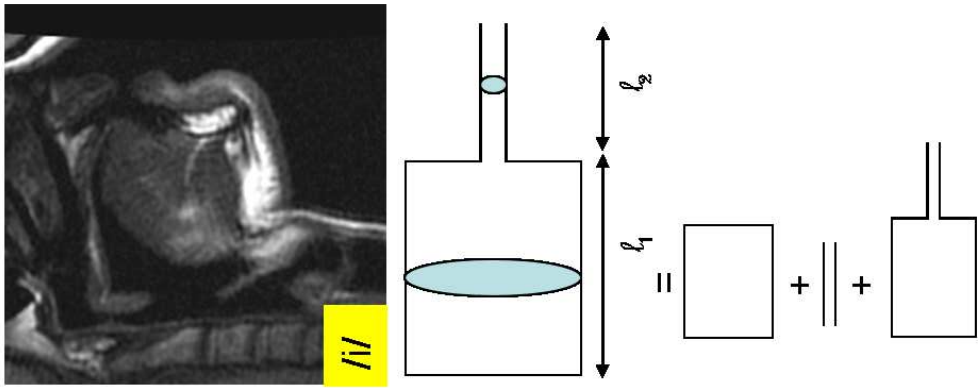
1. ⟷ Tube fermé à une extrémité et ouvert à l'autre :
   - **résonateur en quart d'onde**
   - fréquences de résonance : $(2n-1)\frac{c}{4L}$
   - Exercice: trouver les fréquences de résonance pour L=17 cm and c = 350m/s
2. Tube fermé aux deux extrémités
   - **résonateur en demie longueur d'onde**
   - fréquences de résonance : $n\frac{c}{2L}$
3. 

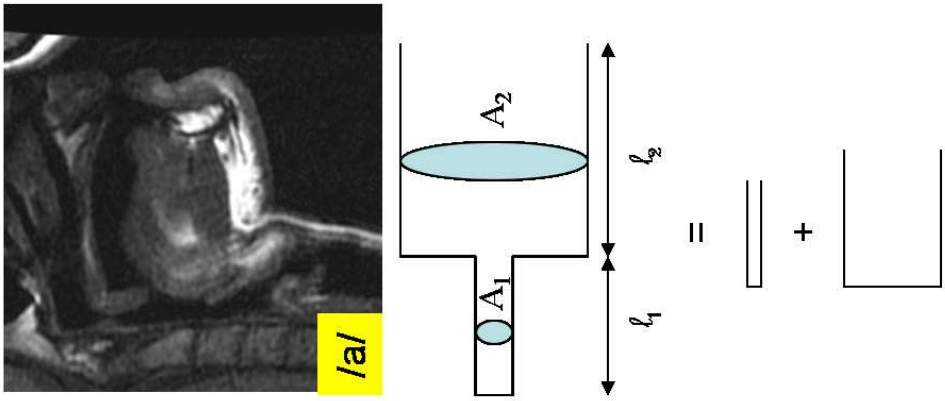
Tube fermé par un petit tuyau à l'une des extrémités et fermé à l'autre
   - **résonateur de Helmholtz**
   - fréquences de résonance :

   Fréquences de résonance : $\frac{c}{2\pi}\sqrt{\frac{a}{lV}}$ où V est le volume du gros tuyau, l la longueur du petit tuyau et a sa section.

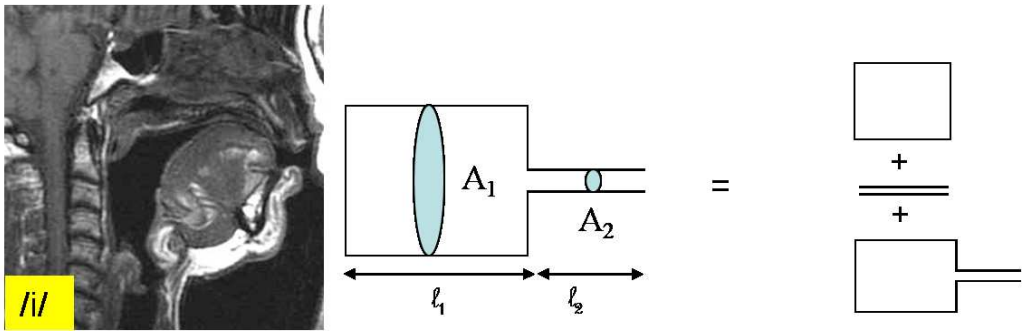
Les voyelles cardinales sous la forme de tubes simples







Exercice: voyelle /i/



l_1 = 9cm, l_2= 6cm, A_1= 8cm2, A_2=1.5cm2

calculer F1, F2 et F3

F1 résonance de Helmholtz de la cavité arrière:

$$F_1 = \frac{c}{2\pi}\sqrt{\frac{a}{lV}} = \frac{340}{2\pi}\sqrt{\frac{0.00015}{0.06 \times 0.0008 \times 0.09}} = 319\,Hz$$

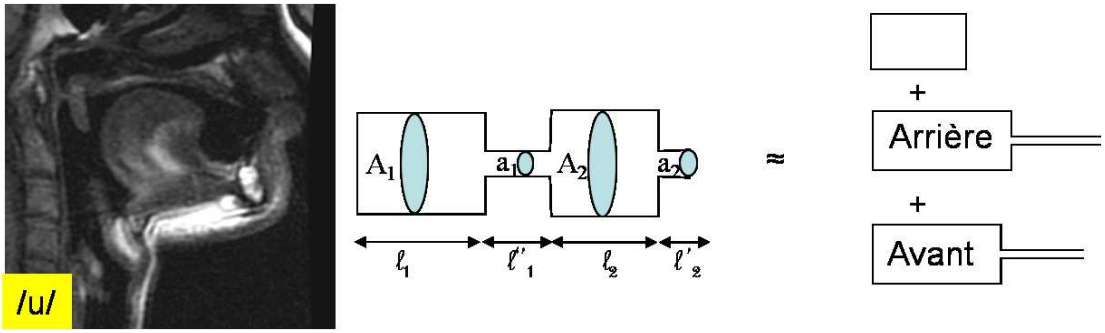
F3 demi longueur d'onde de la cavité avant

$$F_3 = \frac{340}{2 \times 0.06} = 2833\,Hz$$

F2 demi longueur d'onde de la cavité arrière (pharynx)

$$F_2 = \frac{340}{2 \times 0.09} = 1888\,Hz$$

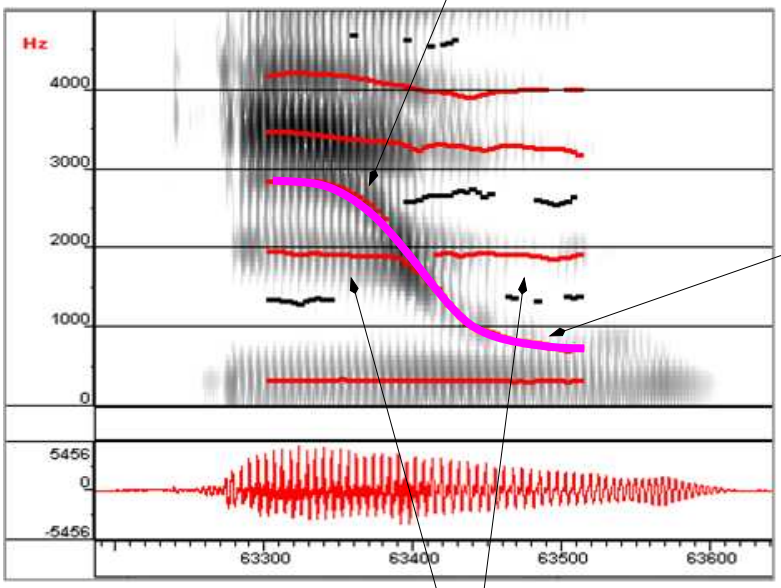
Exercice: voyelle /u/



- Cavité de la bouche
  - résonateur de Helmholtz:
    A_2=7cm2, l_2=5cm, l'_2= 1.5cm, a_2= 1cm2
    La fréquence de résonance est 747 Hz.
- Cavité du pharunx
  - résonateur de Helmholtz:
    A_1=8cm2, l_1=8cm, l'_1= 3cm, a_1= 0.7cm2
    La fréquence de résonance est 326 Hz.
  - résonateur en demi longueur d'onde:
    La fréquence de résonance est 2125 Hz.

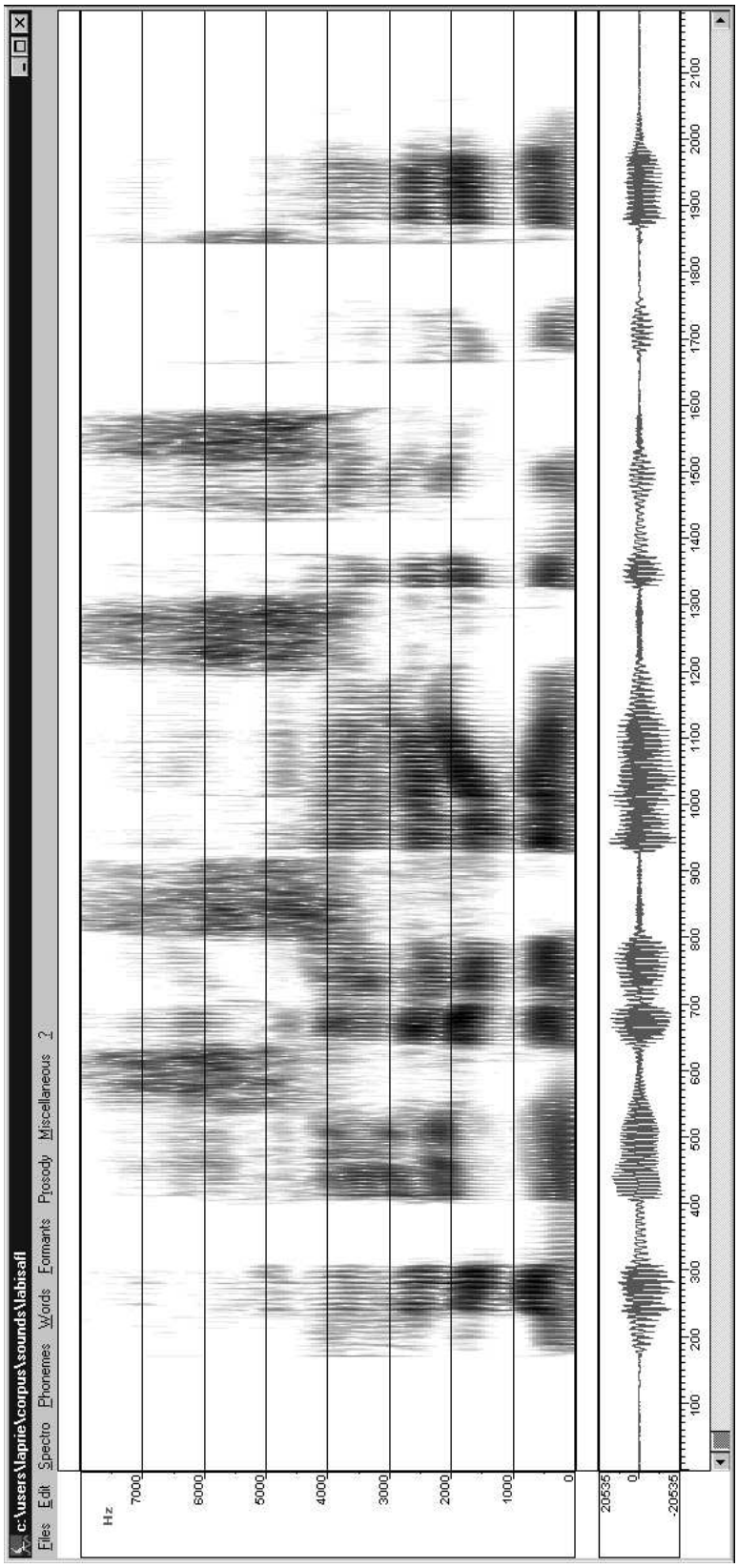
Transition /iu/

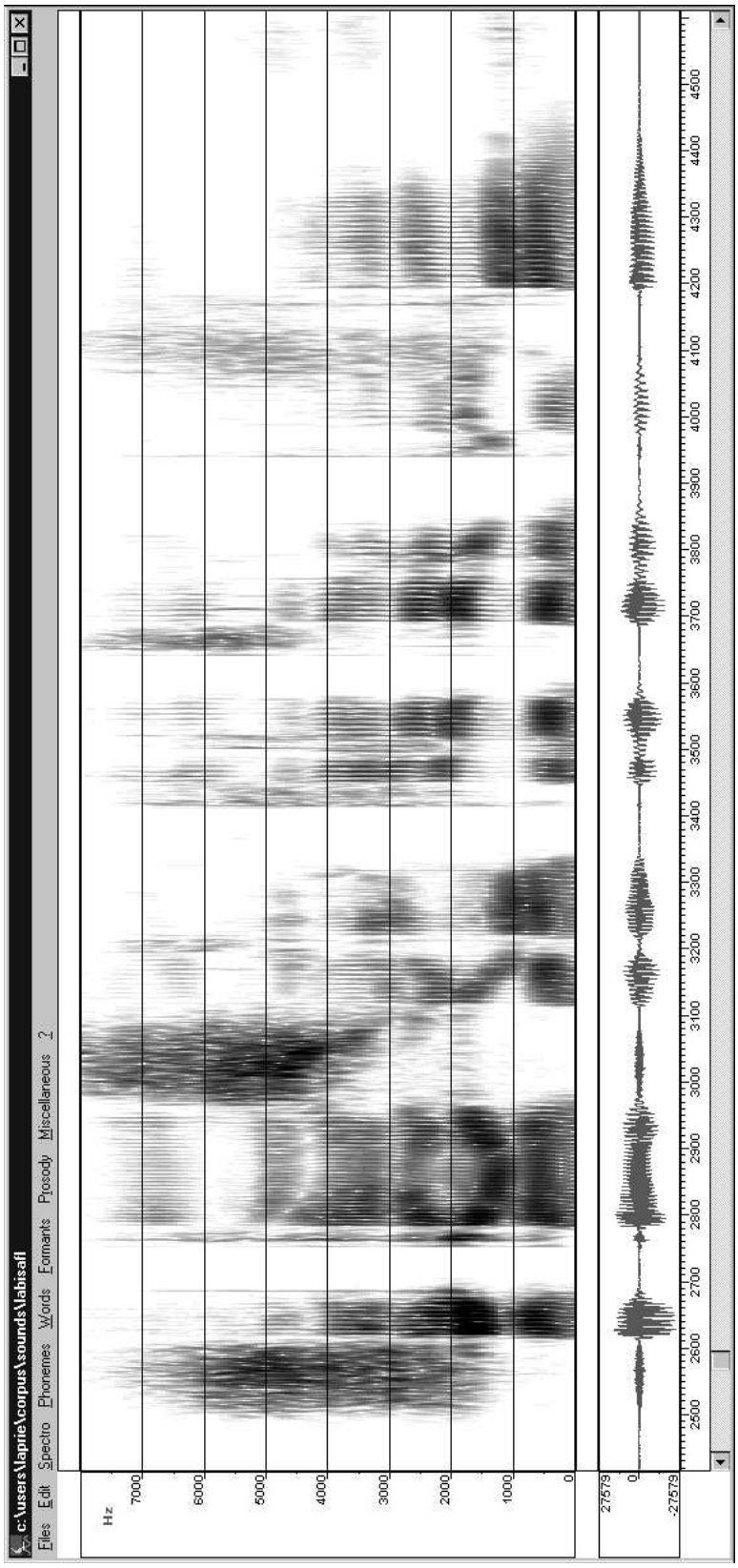
Demi longueur d'onde
de la cavité avant (bouche)



Cavité de la bouche
par continuité mais
résonance de fréquence
faible.
−−> résonance de Helmholtz
de la cavité avant.

Rien à déduire de particulier.


c:\users\lapirie\corpus\sounds\labisafl
Files Edit Spectro Phonemes Words Formants Prosody Miscellaneous ?
Hz
7000
6000
5000
4000
3000
2000
1000
0
20535
0
-20535
100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 1600 1700 1800 1900 2000 2100


c:\users\lapirie\corpus\sounds\labisafl
Files Edit Spectro Phonemes Words Formants Prosody Miscellaneous ?
Hz
7000
6000
5000
4000
3000
2000
1000
0
27579
0
-27579
2500 2600 2700 2800 2900 3000 3100 3200 3300 3400 3500 3600 3700 3800 3900 4000 4100 4200 4300 4400 4500

6 Deux problèmes d'analyse de la parole

- Calcul de la fréquence fondamentale
  - Fréquence de vibration des cordes vocales
  - Essentiel pour le codage, la synthèse acoustique de la parole et l'étude de la prosodie
  - Méthodes opérant dans le domaine temporel (autocorrélation, résidu LPC) ou spectral (cepstre, peigne),
  - Deux sous-problèmes :
    - détection du fondamental $\longrightarrow$ traitement du signal
    - correction ou lissage des résultats bruts $\longrightarrow$ reconnaissance des formes
- Suivi de formants
  - Les formants caractérisent l'évolution de la forme du conduit vocal
  - Problème difficile à cause du couplage avec les cavités nasales
  - Comment interpoler la fonction de transfert du conduit vocal à partir du spectre de parole ? $\longrightarrow$ quelle est l'analyse spectrale la plus appropriée ?
  - Interprétation des pics spectraux en termes de formants ou choix des formants pour représenter au mieux le spectre de parole ?
  - Suivi dans le temps pour créer des trajectoires ou simple détection ?

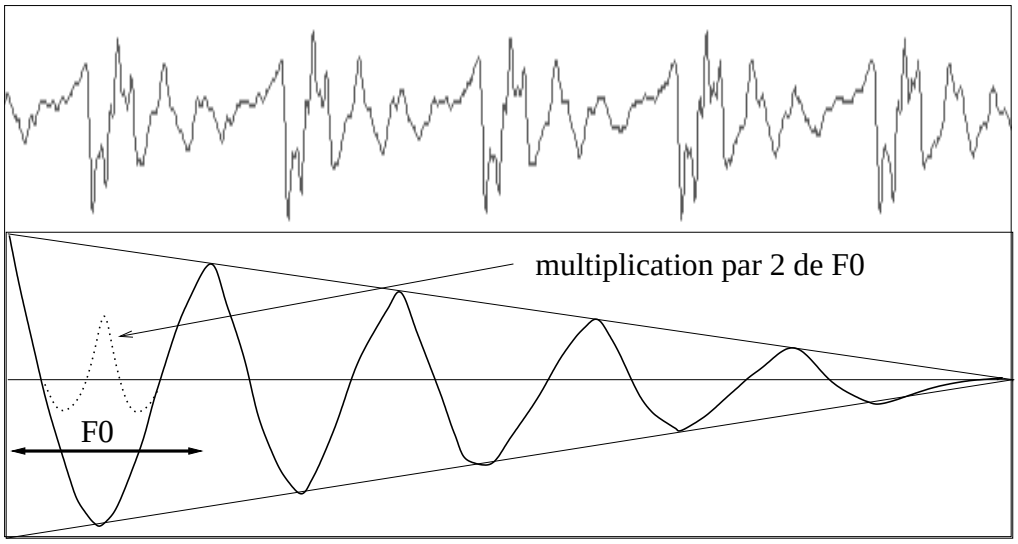
7 Détection et modification de la fréquence fondamentale

7.1 Détection du fondamental (F0)

Quelques méthodes :

autocorrélation (domaine temporel)

- calcul de la fonction d'autocorrélation $\phi(n) = \sum_{i=0}^{N} x(i)x(n+i)$
- $\phi(n)$ est périodique si $x(b)$ est périodique



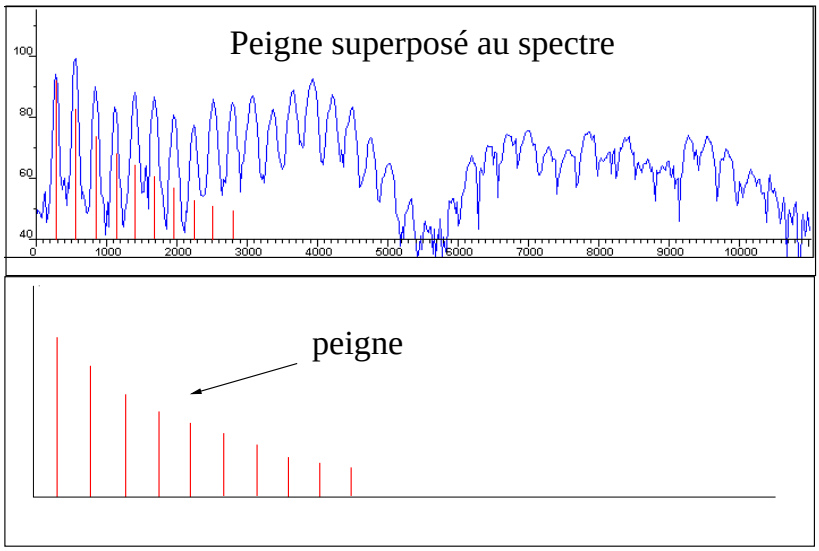
Fonction d'autocorrélation d'un signal pseudopériodique

- problèmes possibles : division ou multiplication par deux de F0
- parades : ne conserver le signal que au-dessous de 1000 Hz (éviter les formants)

peigne (domaine spectral) [?]

- Intercorrélation entre le spectre d'harmonique et un peigne :

$$I(\omega_p) = \int_0^{\infty} P(\omega_p, \omega)|F(\omega)|d\omega$$



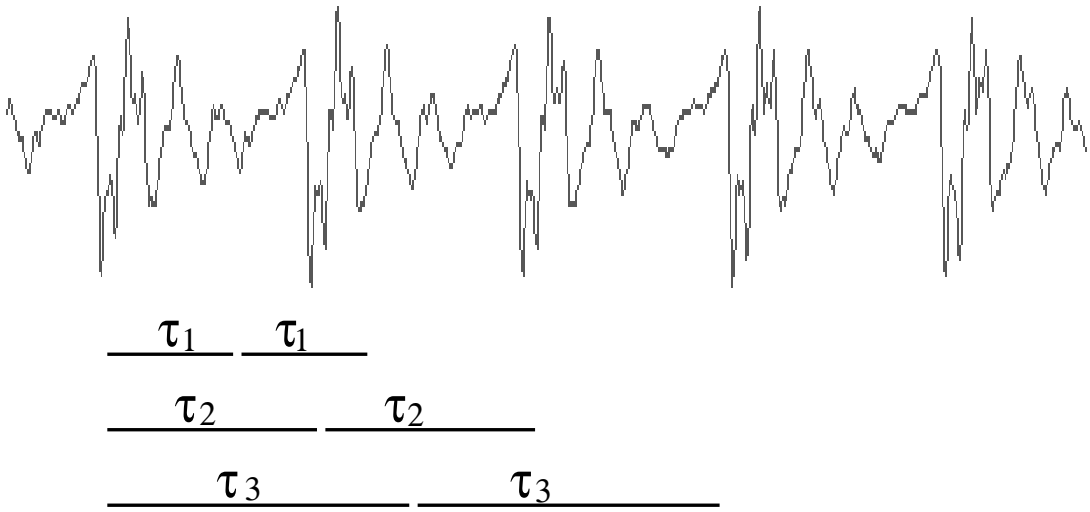
Méthode du peigne

- Si les dents ont une hauteur égale il n'est pas possible de distinguer $F0$ et $F0/k$.
- Peigne de la forme $A_n = n^{-p}$ avec $p > 1$.
- Pour simplifier les calculs et renforcer F0 le spectre est réduit à l'interpolation du spectre à proximité des pics.

Détermination à super résolution (domaine temporel) [11]

- modèle de similarité :
  $x_\tau(t, t_0) = s(t)w_\tau(t - t_0)$
  $y_\tau(t, t_0) = s(t + \tau)w_\tau(t - t_0)$
  $x_{T_0}(t, t_0) = a(t_0)y_{T_0}(t, t_0) + e(t, t_0)$
  $s(t)$ est le signal de parole, w_τ est une fenêtre rectangulaire de durée τ, a et e permettent d'assurer la similarité.

$$T_0 = \underset{\tau, a(t_0)>0}{argmin} \{ \frac{\int_{t_0}^{t_0+\tau} (x_\tau(t,t_0) - a(t_0) y_\tau(t,t_0))^2 dt}{\int_{t_0}^{t_0+\tau} x_\tau(t,t_0)^2 dt} \}$$



Calcul de la similarité

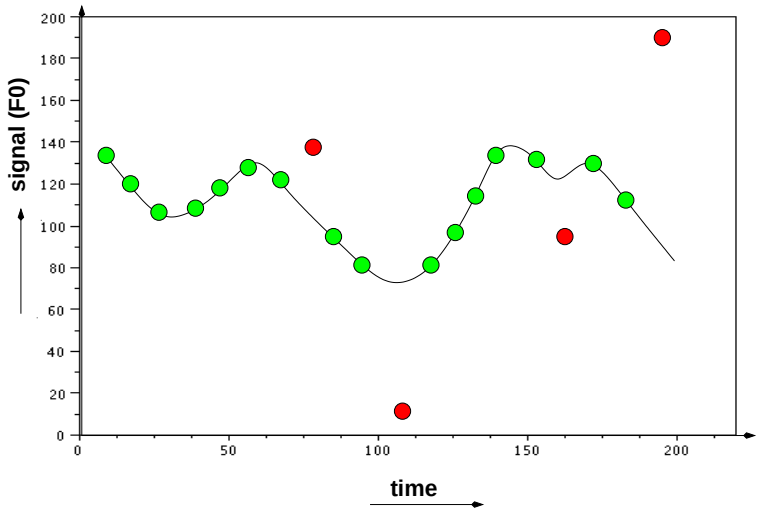
- Une fois T_0 calculé on augmente la précision autour de $T0$.

Correction des valeurs de F0 - Programmation dynamique à bonus

Problèmes :

1. décider si la parole est voisée ou non (si les cordes vocales vibrent)
2. choisir la bonne valeur de F0 (éviter les erreurs, division par 2, multiplication par 2)

En pratique, c'est un problème très important. $\longrightarrow$ optimiser ces choix à l'aide de la programmation dynamique
De nombreux algos s'inspirent de celui proposé par Ney [14] qui consiste à **choisir** les points à conserver



Principe de l'algorithme

$$A = [a(i)] = a(1), ...a(j), ..., a(N)$$

où i est un indice temporel. Sélectionner des points consiste à construire une sélection, cad une séquence d'indices :

$$J = [j(k)] = j(1), ..., j(k), ..., j(K)$$

où K est inférieur ou égal à N, avec $j(k) < N$ et $j(k) < j(k+1)$
La courbe résulat est $\hat{A} = [a(j(k))]$

$$\hat{A} = a(j(1)), ..., a(j(k)), ..., a(j(K)).$$

j est construite en minimisant un critère sur la courbe (par exemple la différence de F0 entre deux points).

$$\min_J \sum_{k=2}^{K} (d(j(k), j(k-1)) - B)$$

où B est un bonus qui représente l'intérêt d'inclure le point dans la courbe finale. B peut prendre en compte l'énergie, la valeur de l'autocorrélation, de l'intercorrélation...

Résolu simplement avec la programmation dynamique:

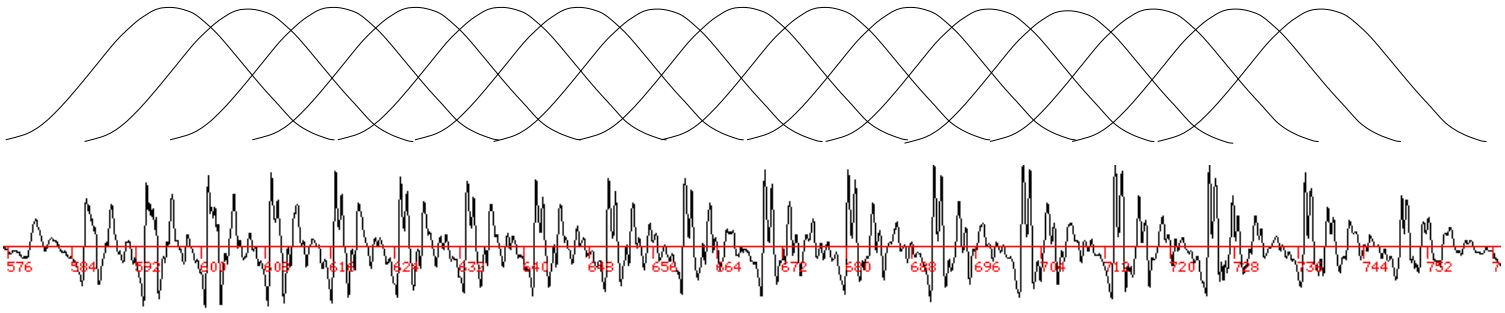
1. initialisation $D(i) = 0, \quad i = 1, ..., N$
2. itération $D(i) = -B + min\{d(i,l) + D(l) : l = 1, ..., i\}$

Bien sûr, il est possible de réduire le balayage pour gagner du temps, ou réduire l'horizon de la recherche.

7.2 Méthode par recouvrement et addition (OLA)

Méthode très générale (voir [18] pages 274 et suivantes) pour transformer un signal

1. Décomposition du signal en fenêtres recouvrantes :
   - Utilisation de la fenêtre de Hamming ou Hanning et zéro padding pour éviter les problèmes de repliement (doubler la taille de la fenêtre)
   - Recouvrement de 75 %
2. Transformée de Fourier
3. Modification du spectre:
   - Généralement seul le spectre d'amplitude est transformé. Faire attention si la phase est modifiée (le zéro padding n'est plus possible).
   - Peut être utilisé pour le filtrage, le débruitage, PSOLA, ...
   - La modification du spectre ne signifie pas que le spectre du signal modifié est celui qui a été modifié.
4. Transformée de Fourier inverse
5. Addition dans le domaine temporel: En dehors des extrémités, la somme des fenêtres est constante pour les fenêtres classiques (Hamming, Hanning). Par exemple pour Hamming $w(n) = 0.54 - 0.46cos(2 * \pi/N * n)$ et donc $w(n + N/2) + w(n) = 2 \times 0.54$.



7.3 PSOLA (Pitch synchronous Overlap and Add) [13]

- Décomposition du signal en fenêtres recouvrantes synchronisées avec les périodes du fondamental
- Les changements de F0 ou de débits reviennent à jouer sur la duplication et/ou l'écart entre les fenêtres.

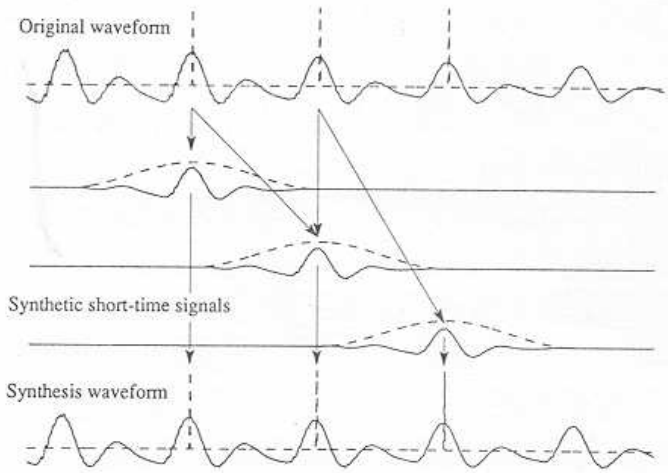


Fig. 8. Time-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch-marks. Middle panel: three short-time analysis signals. The mapping between these analysis signals and the associated analysis pitch-marks are indicated by the arrows. Lower panel: time-scale modified waveform along with the synthetic pitch-marks. Same signal as in Fig. 5. The time-scale modification factor is set to 2; the interpolation mode is used: the intermediate short-time signal is obtained by linearly interpolating the two adjacent waveforms, with a coefficient $\alpha = 0.5$.

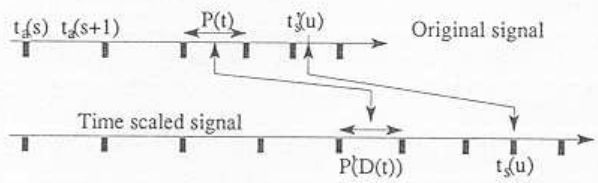


Fig. 7. Calculation of the synthesis pitch-marks for time-scale modifications. We have $P'(D(t)) \approx P(t)$ and $t_s(u) = D(t'_s(u))$.

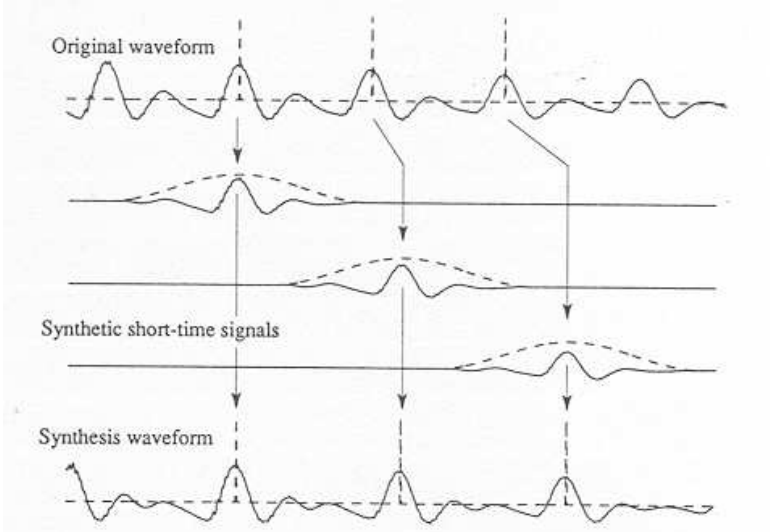


Fig. 5. Pitch-scale modification with the TD-PSOLA method. Upper panel: original signal along with the analysis pitch-marks. Middle panel: three short-time synthetic signals. The mapping between these short-time signals and the analysis pitch-marks are indicated by the arrows. Lower panel: pitch-scale modified waveform along with the synthetic pitch-marks. The signal is a vowel /i/, uttered by a male speaker (pitch frequency around 100 Hz). The pitch-scale modification factor is equal to 0.8.

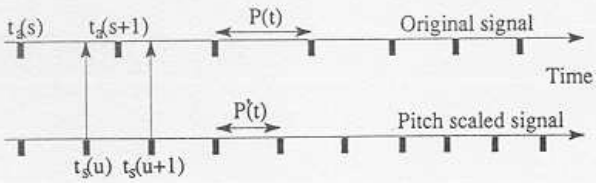


Fig. 4. Calculation of the synthesis pitch-marks for pitch-scale modifications. We have $P'(t) = P(t)/\beta$, with $\beta = 3/2$.

7.4 Modèles sinusoïdaux et harmoniques

Idée : Représenter le signal de parole par une somme de composantes sinusoïdales correspondant aux maxima spectraux
Avantage : pouvoir modifier indépendamment les composantes spectrales du signal de parole

- L'idée vient des recherches en synthèse musicale [20]
- $s(t) = \sum_{r=1}^{R} A_r(t) cos[\theta_r(t)]$ où $A_r(t)$ est l'amplitude instantanée et $\theta_r(t)$ la phase instantanée.
- La mise en œuvre nécessite le développement d'un algorithme pour construire des trajectoires spectrales qui correspondent aux pics spectraux.
- Lors de la synthèse il faut interpoler les amplitudes entre deux trames, et surtout les phases et les fréquences instantanées.
- Les modifications temporelles et de fondamentales sont faciles mais difficiles en pratique si on veut éviter la désynchronisation des phases [17].

$\longrightarrow$ modèles harmoniques + bruit [9]

- On suit directement les harmoniques :
  $s(t) = \sum_{r=-K(t)}^{K(t)} A_r(t) exp(jr\omega_0(t)) + e(t)$ où $\omega_0(t)$ est la fréquence fondamentale.
- réglage des paramètres en mesurant l'erreur sur le signal temporel $E = \sum_{n=t_{i-1}}^{t_{i+1}} \alpha(t)(s(t) - \hat{s}(t))^2$ où $\alpha(t)$ est une fonction de pondération.
- calcul de la partie non déterministe par simple soustraction puis réglage du spectre par un modèle tout pôle.
- synthèse synchronisée avec les périodes du fondamental.

References

[1] S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(4):357–366, August 1980.

[2] Auger F. and Flandrin P. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, May 1995.

[3] T. Gallas and X. Rodet. Generalized functionnal approximation for source-filter system modelling. In *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991.

[4] Scilab group. Scilab. In *http://www-rocq.inria.fr/scilab/*. INRIA.

[5] P. Halle. Techniques cepstrales amÃ©liorÃ©es pour l'extraction d'enveloppe spectrale et la dÃ©tection du pitch. In *Actes du séminaire "Traitement du signal de parole"*, pages 83–93, Paris, 1983.

[6] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978.

[7] H. Hermansky. Perceptual linear predictive (lpl) analysis of speech. *Journal of Acoustical Society of America*, 87:1738–1752, April 1990.

[8] Y. Laprie. WinSnoori. In *http://www.loria.fr/~ laprie/WinSnoori*. LORIA, 2000.

[9] J. Laroche, Y. Stylianou, and E. Moulines. HNS: Speech modification based on a Harmonic + Noise Model. In *Proc. ICASSP*, volume 2, pages 550–553, Minneapolis, April 1993.

[10] J.D. Markel and A.H. Gray. *Linear Prediction of Speech*. Springer-Verlag, Berlin Heidelberg New York, 1976.

[11] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-39(1):40–48, January 1991.

[12] Brian. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press, London, 1989.

[13] E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for a text-to-speech synthesis using diphones. *Speech Communication*, 9(5,6):453–467, 1990.

[14] H. Ney. A dynamic programmation algorithm for nonlinear smoothing. *Signal Processing*, 5(2):163–173, March 1983.

[15] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, Inc, 1975.

[16] F. Plante, G. Meyer, and W.A. Ainsworth. Improvement of speech spectrogram accuracy by the method of reassignment. *IEEE Trans. on Speech, and Audio Processing*, 6(3):282–287, 1998.

[17] T.F. Quatieri and R.J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, March 1992.

[18] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, N.J., 1978.

[19] T. Robinson. Speech Analysis. In *http://svr-www.eng.cam.ac.uk/svr.html*, http://svr-www.eng.cam.ac.uk/˜ ajr/SpeechAnalysis/index.html, 1998. Speech Vision and Robotics Group, University of Cambridge.

[20] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.