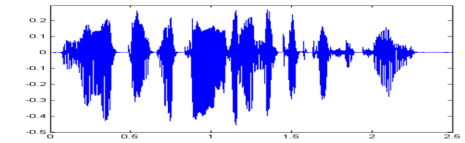
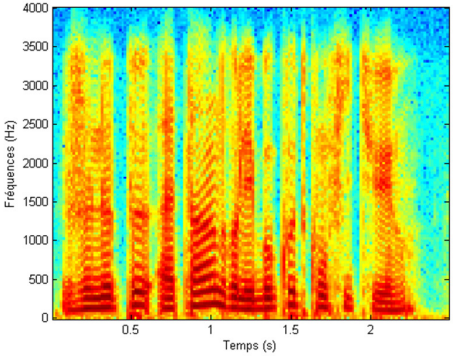
Modifications de hauteur, d'échelle temporelle et de timbre

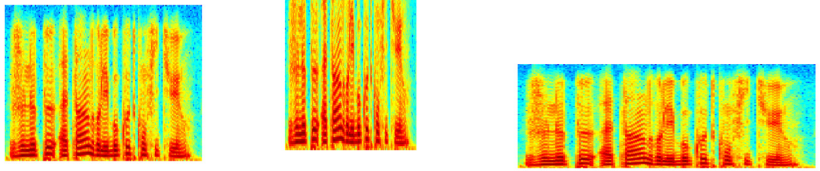
M2 Mathématiques / Vision / Apprentissage
Analyse des signaux audiofréquences

Introduction



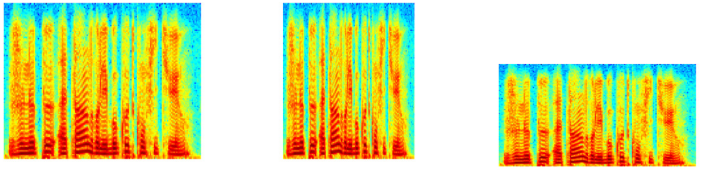


Modification de la vitesse de lecture



Origine du problème : $y(t) = x(\alpha t) \Leftrightarrow Y(f) = \frac{1}{|\alpha|} X(\frac{f}{\alpha})$

Modifications de durée et de hauteur



But : contrôler séparément les échelles temps et fréquence

Plan

- Contrôle séparé des échelles temporelle et fréquentielle
  - Synthèse par échantillonnage d'une table d'onde
  - Post-synchronisation du son et de la vidéo
  - Post-production musicale
- Trois catégories de méthodes
  - Méthodes fréquentielles : vocodeur de phase
  - Méthodes temporelles : TD-PSOLA
  - Méthodes paramétriques : LPC, sinus + bruit

Partie I

Définitions

Modèle de production vocale

- Modèle source / filtre linéaire variant dans le temps :
  $x(t) = \int_{-\infty}^{+\infty} g(t,\tau)\, e(t-\tau)\, \mathrm{d}\tau$
- Réponse en fréquence du filtre :
  $G(t,f) = \int_{-\infty}^{+\infty} g(t,\tau)\, e^{-j2\pi f\tau} \mathrm{d}\tau = M(t,f)\, e^{j\varphi(t,f)}$
- Source harmonique : $e(t) = \sum_{k=1}^{L} e^{j\xi_k(t)}$, où $\frac{\mathrm{d}\xi_k}{\mathrm{d}t} = 2\pi f_k(t)$
- Hypothèse de quasi-stationnarité : $\xi_k(t-\tau) \simeq \xi_k(t) - 2\pi f_k(t)\tau$
- Signal filtré : $x(t) = \sum_{k=1}^{L} M(t, f_k(t))\, e^{j(\xi_k(t)+\varphi(t,f_k(t)))}$

Modèles de signaux

Modèle de McAulay et Quatieri (codage de la parole)
$x(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Psi_k(t)}$ où $\frac{\mathrm{d}\Psi_k}{\mathrm{d}t} = 2\pi f_k(t)$
et $A_k(t)$ et $f_k(t)$ sont à variation lente devant $e^{j\Psi_k(t)}$

Modèle de Serra et Smith (synthèse de signaux de musique)
$x(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Psi_k(t)} + b(t)$
où $b(t)$ est un bruit blanc filtré par un filtre variant dans le temps

Système complet d'analyse / modification / synthèse

- estimation des composantes déterministes
- interpolation linéaire des amplitudes et cubique des phases
- soustraction de la partie déterministe pour obtenir $b(t)$
- transformation de chacune des deux composantes
- resynthèse

Modifications d'échelles

Distortion temporelle

- Fonction de distortion temporelle : $\tau = T(t)$
- Signal modifié : $y(\tau) = \sum_{k=1}^{L} A_k(T^{-1}(\tau))\, e^{j\phi_k(\tau)}$
- Conservation des fréquences : $\phi_k(\tau) = 2\pi \int_0^{\tau} f_k(T^{-1}(u))du$

Modification de hauteur

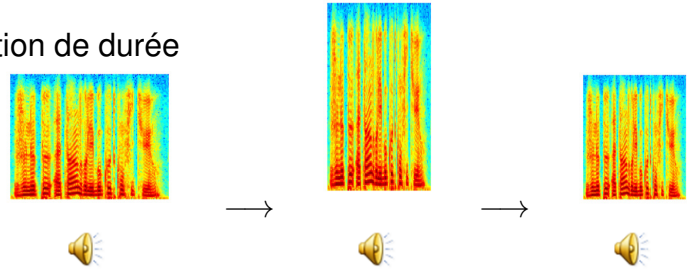
- Taux de compression fréquentiel : $\alpha(t)$
- Signal modifié : $y(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Phi_k(t)}$
- Altération des fréquences : $\Phi_k(t) = 2\pi \int_0^{t} \alpha(u) f_k(u)du$

Réciprocité

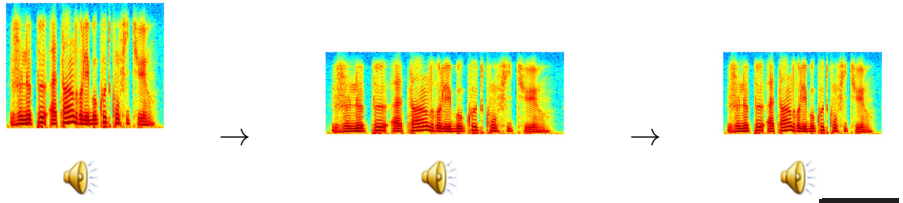
- distortion temporelle T plus ré-échelonnement T^{-1}
  $\Leftrightarrow$ modification de hauteur de taux $\alpha(t) = T'(t)$

Équivalence des deux modifications

- Modification de durée



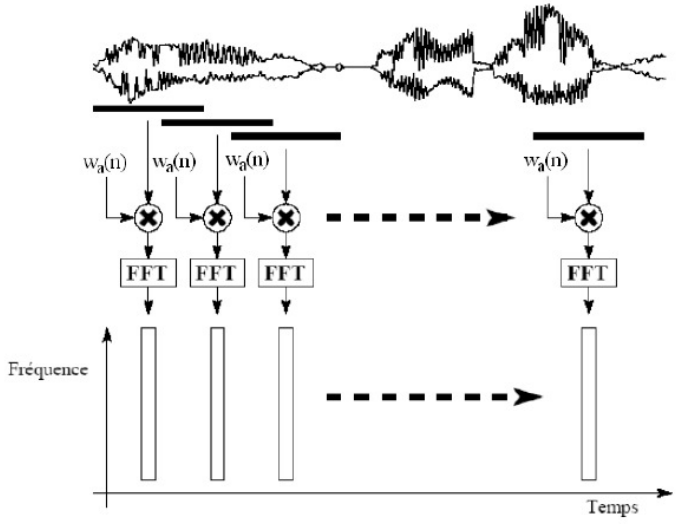
- Modification de hauteur



Partie II

Transformée de Fourier à Court Terme

Schéma de principe



Rappels théoriques

Définition : $\widetilde{X}(t_a, \nu) = \sum_{n\in\mathbb{Z}} x(n+t_a)\, w_a(n)\, e^{-j2\pi\nu n}$, où

- la fenêtre d'analyse $w_a(n)$ est finie, réelle et symétrique
- les instants d'analyse t_a sont indexés par un entier u

Interprétation : convention passe-bande

- $\widetilde{X}(t_a, \nu_p) = [x \star h](t_a)$ où $h(n) = w_a(-n)\, e^{j2\pi\nu_p n}$
- la TF de $h(n)$ est $H(e^{j2\pi\nu}) = W_a\left(e^{j2\pi(\nu_p-\nu)}\right)$

Version discrète de la TFCT : on pose $\nu_p = \frac{p}{N}$

- $\widetilde{X}(t_a, \nu_p) = \sum_{n=0}^{N-1} x(n+t_a)\, w_a(n)\, e^{-j2\pi\frac{pn}{N}}$
- la longueur des fenêtres d'analyse doit être $\leq N$

Filtrage passe-bande équivalent

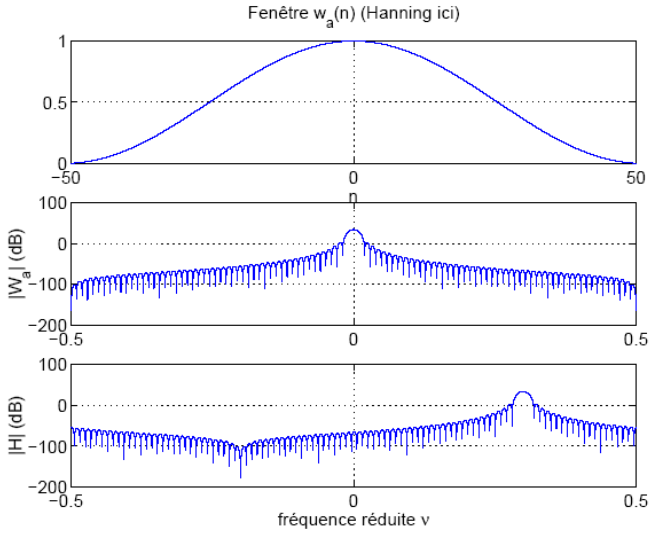
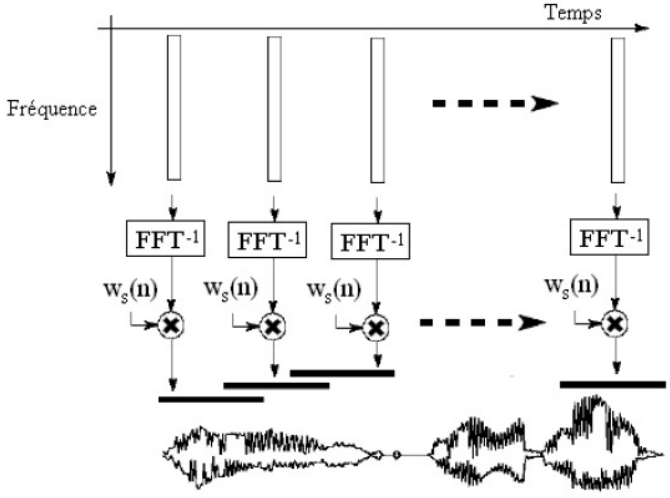


Schéma de synthèse



Reconstruction du signal

Condition de reconstruction parfaite ($t_s = t_a$ et $Y = \widetilde{X}$)

- synthèse par addition-recouvrement (overlapp-add ou OLA)
  $$y(n) = \sum_u w_s(n - t_s(u))\, y_w\left(n - t_s(u), t_s(u)\right)$$
  $\mathrm{supp}(w_s) \subset [0, N-1]$, $y_w(n, t_s(u)) = \frac{1}{N}\sum_{p=0}^{N-1} Y(t_s(u), \nu_p)\, e^{j2\pi\nu_p n}$
- condition suffisante : $\sum_u w_a(n - t_a(u))\, w_s(n - t_a(u)) \equiv 1$

Modifications et problèmes posés :

- Modification des amplitudes et phases de la TFCT
- $t_a \longrightarrow t_s$, $\widetilde{X}(t_a(u), \nu_p) \longrightarrow Y(t_s(u), \nu_p)$
- Difficulté : Y n'est généralement pas la TFCT d'un signal
- Resynthèse à partir d'un modèle sinusoïdal

Partie III

Vocodeur de phase

Fréquence instantanée

- Modèle de McAulay et Quatieri : $x(t) = \sum_{k=1}^{L} A_k(t)\, e^{j\Psi_k(t)}$
- **Hypothèses de quasi-stationnarité** : $\forall n \in \{0 \dots N-1\}$
$$\begin{cases} A_k(n+t_a) & \simeq & A_k(t_a) \\ \Psi_k(n+t_a) & \simeq & \Psi_k(t_a) + 2\pi f_k(t_a) n \end{cases}$$
- Alors $\widetilde{X}(t_a(u), \nu_p) = \sum_{k=1}^{L} A_k(t_a)\, e^{j\Psi_k(t_a)}\, W_a\left(e^{j2\pi(\nu_p - f_k(t_a))}\right)$
- Soit f_c la fréquence de coupure du filtre passe-bas $w_a(n)$
- **Condition de bande étroite** : $\exists!\ l$ tel que $|\nu_p - f_l(t_a)| \leq f_c$
Interprétation (spectre harmonique) : $N \geq \frac{4}{f_0}$
- Alors $\widetilde{X}(t_a(u), \nu_p) = A_l(t_a)\, e^{j\Psi_l(t_a)}\, W_a\left(e^{j2\pi(\nu_p - f_l(t_a))}\right)$
$\Rightarrow$ la TFCT donne accès aux phases $\Psi_l(t_a)$ modulo 2π

Condition de recouvrement

Levée de l'indétermination de la phase modulo 2π :

- Différence de phases entre deux instants successifs :
$\Delta\Phi_p = 2\pi(f_l(t_a) - \nu_p)\Delta t_a(u) + 2\pi\nu_p\Delta t_a(u) + 2n\pi$
- Condition de recouvrement minimal : $f_c\, \Delta t_a(u) < \frac{1}{2}$
Interprétation (fenêtre de Hanning) : $f_c = \frac{2}{N} \Rightarrow \Delta t_a < \frac{N}{4}$
- $\exists!\ n$ tel que $|\Delta\Phi_p - 2\pi\nu_p\Delta t_a(u) - 2n\pi| < \pi$

Estimation de la fréquence instantanée $\forall p \in \{0 \dots N-1\}$

1. calcul de la TFCT à deux instants successifs $\longrightarrow \Delta\Phi_p$
2. calcul de $Q(n_0) = \Delta\Phi_p - 2\pi\nu_p\Delta t_a - 2n_0\pi$ tel que $|Q(n_0)| < \pi$
3. calcul de la fréquence instantanée $f_l(t_a) = \nu_p + \frac{Q(n_0)}{2\pi\Delta t_a}$

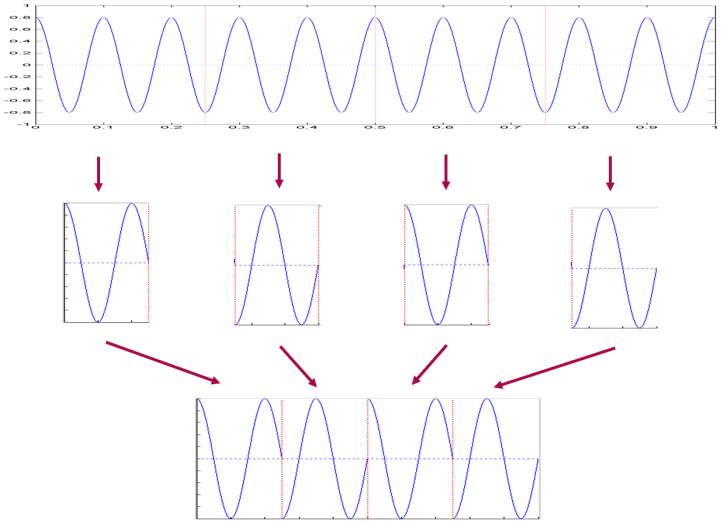
Distortion temporelle

Déroulement des phases instantanées pour une distortion $T(t)$

Algorithme de modification :

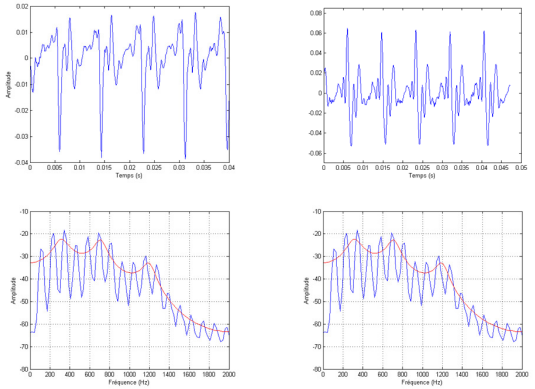
1. calcul de la TFCT et de $f_l(t_a(u))$ dans chaque canal
2. calcul du nouvel instant de synthèse $t_s(u) = T(t_a(u))$
3. calcul de la phase instantanée de synthèse
$\Phi_s(t_s(u+1), \nu_p) = \Phi_s(t_s(u), \nu_p) + 2\pi f_l(t_a(u))\,(t_s(u+1) - t_s(u))$
4. calcul de la TFCT de synthèse en $u+1$
$\widetilde{Y}(t_s(u+1), \nu_p) = A_p(t_a(u+1))\, e^{j\Phi_s(t_s(u+1), \nu_p)}$

Déroulement de la phase



Influence des phases initiales

$$x(t) = \sum_{k=1}^{H} a_k \cos(2\pi k \frac{t}{T} + \psi_k)$$



Modification de hauteur

Méthode de rééchantillonage temporel

1. étirement temporel de distortion $T(t) = \int_0^t \alpha(u)du$
2. ré-échelonnement de taux $T^{-1}(\tau)$

Méthode de rééchantillonage spectral

1. interpolation linéaire de la TFCT d'analyse
   - $\alpha(t_a) > 1$: perte d'information en hautes fréquences
   - $\alpha(t_a) < 1$: complétion du spectre en hautes fréquences
2. resynchronisation des phases en synthèse

Problème en traitement de la parole : effet "Donald Duck"

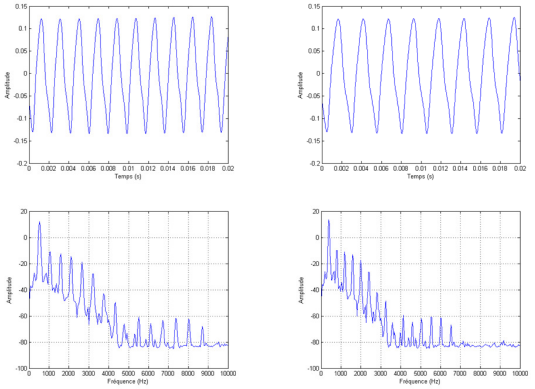
- estimation de l'enveloppe du spectre (LPC) et "blanchiment"
- modification de hauteur, puis filtrage inverse

Partie IV

Traitement spécifique de la parole

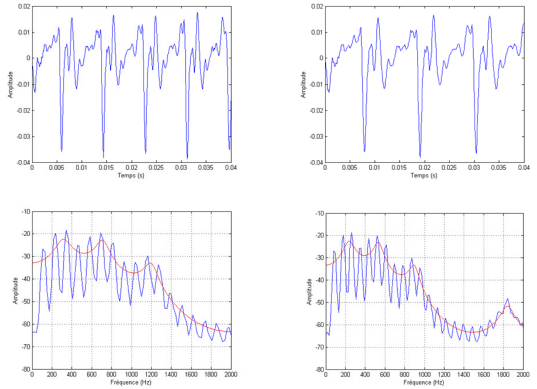
Réciprocité temps-fréquence

$$x(\frac{t}{\alpha}) = \sum_{k=1}^{H} a_k \cos(2\pi \frac{k}{\alpha T} t + \phi_k)$$

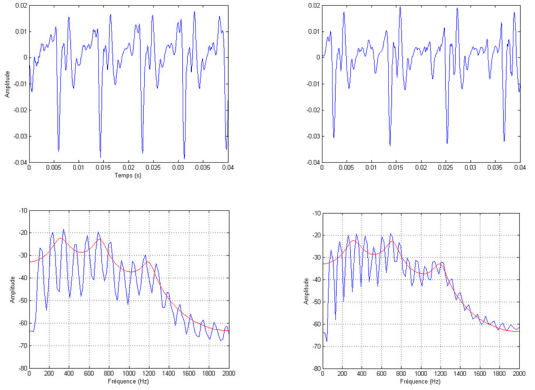


Réciprocité temps-fréquence

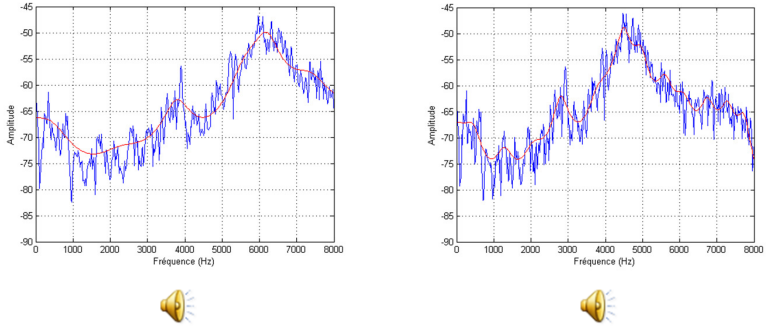
$$x(\frac{t}{\alpha}) = \sum_{k=1}^{H} a_k \cos(2\pi \frac{k}{\alpha T} t + \phi_k)$$



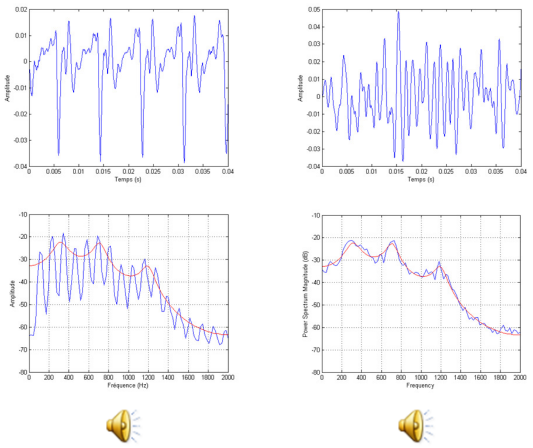
Modification de hauteur pour la parole



Cas des sons non voisés



Timbre et enveloppe spectrale



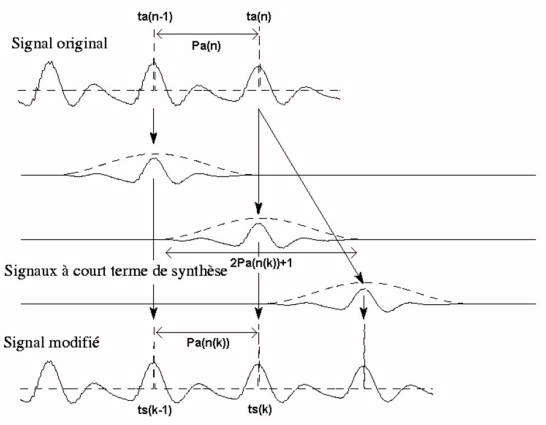
Modification de hauteur

- Sons voisés :
  - modifier la fréquence fondamentale
- Sons voisés/non voisés :
  - laisser l'enveloppe spectrale inchangée
- Utilisation du vocodeur
  1. Blanchiment du signal par filtrage (analyse LPC)
  2. Modification d'échelle fréquentielle
  3. Filtrage inverse
- Méthodes spécifiques aux signaux de parole monophoniques
  - Segmentation voisé / non voisé
  - Estimation de hauteur sur les trames voisées

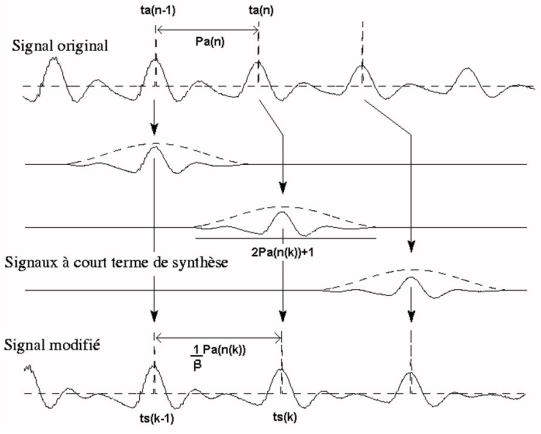
Partie V

TD-PSOLA

Modifications temporelles

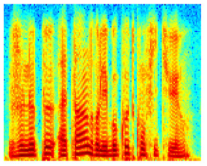
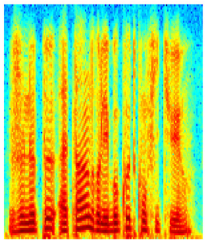
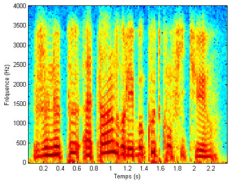


Modifications fréquentielles



Exemple de modification de hauteur

Comparaison vocodeur de phases / PSOLA

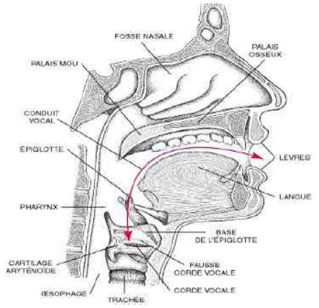




Partie VI

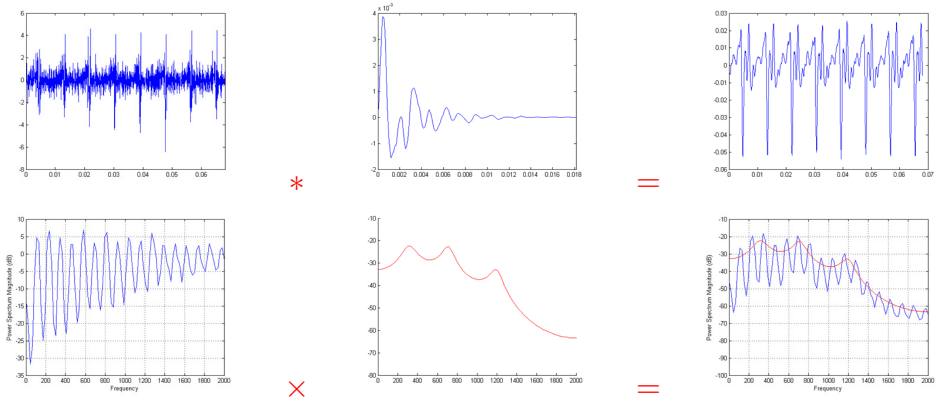
Modèles autorégressifs

Mécanisme de production de la parole

- Sons voisés : vibration des cordes vocales filtrée par le conduit vocal
- Sons non voisés : souffle turbulent filtré par le conduit vocal

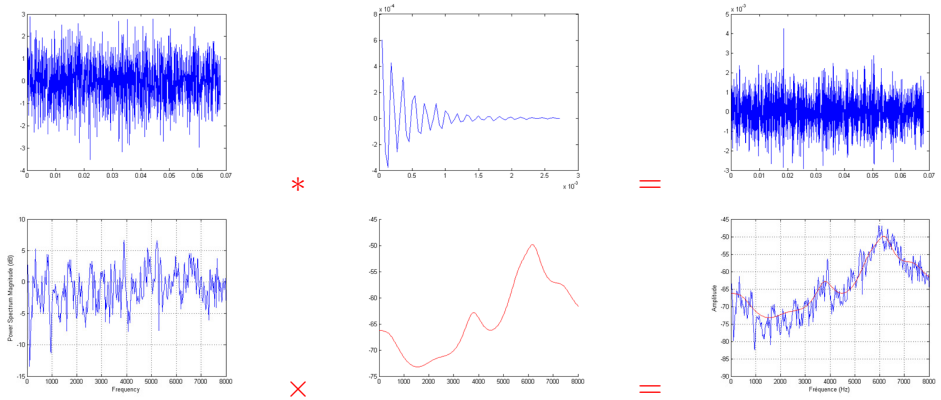


Production des sons voisés



Ondes glottiques × Conduit vocal = Son voisé

Production des sons non voisés



Bruit turbulent × Conduit vocal = Son non voisé

Modélisation du signal

- Le conduit vocal est modélisé par un filtre AR

$$h(z) = \frac{1}{1 + a_1 z^{-1} + \ldots + a_p z^{-p}}$$

  estimé par prédiction linéaire (analyse LPC)
- Modèle de source selon le cas voisé / non voisé
  - Le train d'ondes glottiques est modélisé par un train d'impulsions de période T

$$s(t) = \sum_n \delta(t - nT)$$

  - Le souffle turbulent est modélisé par un bruit blanc

Synthèse par modèles autorégressifs

- Synthèse sans modification
  - par addition / recouvrement des trames
  - convolution de la source par le filtre sur chaque trame
- Synthèse avec modification
  - Modification de durée
    - Synthèse d'une source de durée appropriée
  - Modification de hauteur
    - Trames non voisées : inchangées
    - Trames voisées : on modifie la période des impulsions