

ENSEA 3ème SyM (2017-2018)

Traitement du signal audio musical

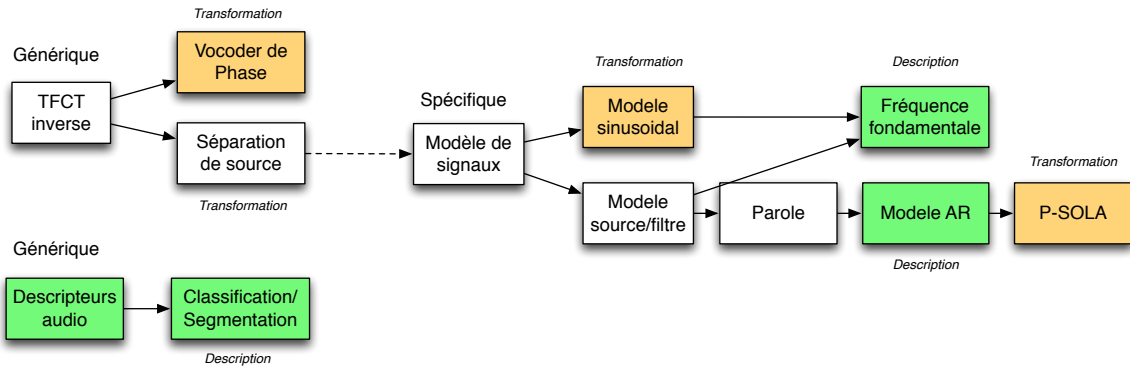
Partie 2: modèles de signaux, estimateurs, descripteurs audio et
apprentissage machine

Geoffroy.Peeters@ircam.fr

UMR SMTS 9912 (IRCAM CNRS UPMC)

1. Introduction
2. Les modèles de signaux
 - 2.1 Le modèle sinusoïdal
 - 2.2 Le modèle sinusoïdal harmonique
 - 2.3 Le modèle source/ filtre
 - 2.4 Production du signal vocal
 - 2.5 La prédiction linéaire, modèle auto-régressif (AR)
 - 2.6 Transformation du signal par la méthode P-SOLA
3. Utilisation de modèles de signal
 - 3.1 Modèle de signal (son quasi-périodique)
 - 3.2 Méthodes temporelles
 - 3.3 Méthodes fréquentielles
 - 3.4 Méthodes combinées
 - 3.5 Transformée à Q-Constant (CQT)
4. Applications du traitement audio pour la description musicale
 - 4.1 Identification audio
5. Descripteurs audio
 - 5.1 Introduction
 - 5.2 Taux de passage par zéro
 - 5.3 Enveloppe ADSR
 - 5.4 Description du spectre (barycentre, étendue spectral)
 - 5.5 Mel Frequency Cepstral Coefficients (MFCCs)
 - 5.6 Chroma - Pitch Class Profile (PCP)
 - 5.7 Spectral Flatness Measure (SFM)
 - 5.8 Intégration temporelle
6. Classification Audio
 - 6.1 Extraction des descripteurs
 - 6.2 Apprentissage

1- Introduction



Les modèles de signaux

Pourquoi des modèles de signaux ?

- On suppose que le signal a été produit par un certain **modèle**
- Permet de **réduire** le nombre de paramètres observés du signal
 - la TFCT contient beaucoup trop d'information
- Permet d'obtenir des paramètres plus facilement **interprétables** (indexation) et manipulables (transformation, synthèse)
 - la TFCT fournit des paramètres non directement exploitables
- Quels modèles ?
 - Modèle sinusoidal harmonique
 - Modèle source/ filtre
 - Modèle autorégressif

Le modèle sinusoidal

2- Les modèles de signaux

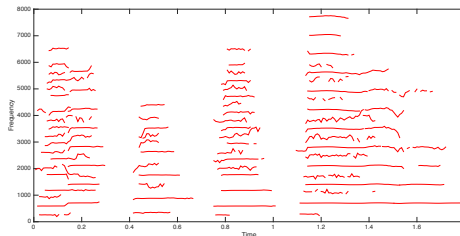
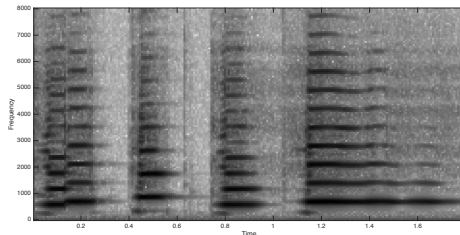
Le modèle sinusoïdal

Le modèle sinusoïdal

- Le signal $x(t)$ est représenté par une somme de sinusoides plus un bruit

$$x(t) = \sum_{h=1}^{H(t)} A_h(t) \cos(\phi_h(t)) + b(t)$$

- les sinusoides sont en nombre limité
- les paramètres des sinusoides sont lentement variables en temps



2- Les modèles de signaux

Le modèle sinusoidal

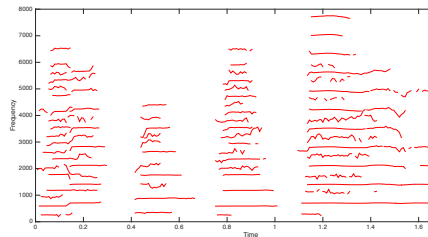
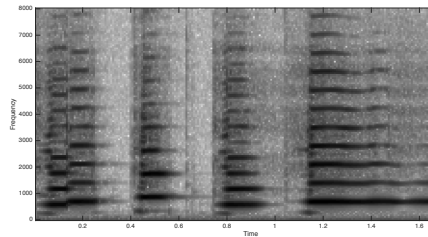
Le modèle sinusoidal

- Lentement variables en temps?
 - → stationnarité locale
 - on peut approximer localement en temps $A_h(t)$ par sa valeur à une trame donnée $A_h(t_m)$
 - on peut approximer localement en temps l'évolution de la phase par $\phi(t) = \omega_h(t_m)(t - t_m) + \phi_h(t_m)$

$$x(t) = \sum_{h=1}^{H(t)} A_h(t) \cos(\phi_h(t)) + b(t)$$

$$\hat{x}_m(t_m) = \sum_{h=1}^{H(t)} A_h(t_m) \cos(\omega_h(t_m)(t - t_m) + \phi_h(t_m))$$

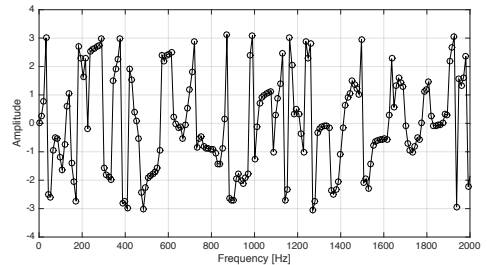
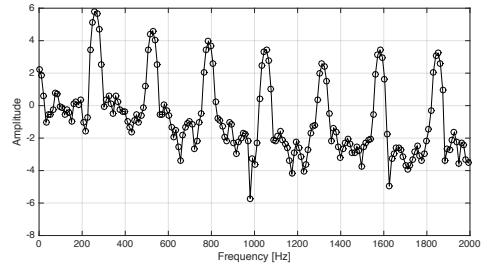
$$\hat{x}(t) = \sum_m \hat{x}_m(t) + b(t)$$



2- Les modèles de signaux

Le modèle sinusoidal

Comment trouver les $\omega_h(t_m)$ et $A_h(t_m)$?

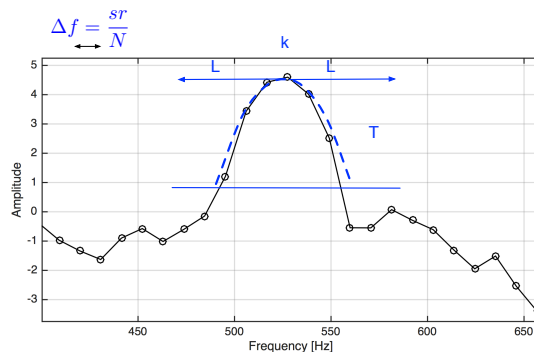


2- Les modèles de signaux

Le modèle sinusoidal

Comment trouver les $\omega_h(t_m)$ et $A_h(t_m)$?

- 1. Détection des maxima locaux/ des pics de $|X(k, m)|$ (peak-picking)
 - k est un maximum local si
 - si il est supérieur à la valeur de ses voisins sur une longueur L à gauche et à droite
 - si il est supérieur à T fois la valeur de ces voisins
 - ...
- Amélioration de la **précision** fréquentielle :
 - utilisation du zero-padding



2- Les modèles de signaux

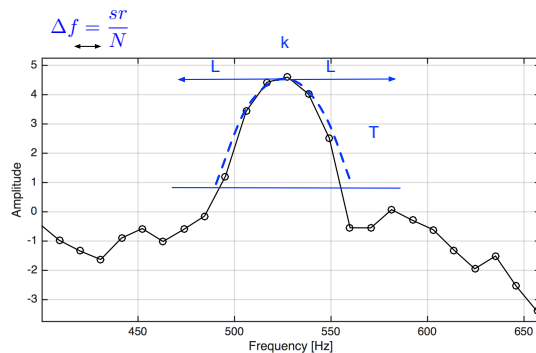
Le modèle sinusoidal

Comment trouver les $\omega_h(t_m)$ et $A_h(t_m)$?

- 2. **Interpolation** du spectre de puissance sur trois points :

$$\omega_h = \omega_k + \frac{\Delta}{2} \frac{P_{k-1} - P_{k+1}}{P_{k-1} - 2P_k + P_{k+1}} \quad (1)$$
$$P_h = P_k - \frac{1}{8} \frac{(P_{k-1} - P_{k+1})^2}{P_{k-1} - 2P_k + P_{k+1}}$$

- avec $P_k = |X(k, m)|^2$
- avec $\Delta = \omega_k - \omega_{k-1}$



2- Les modèles de signaux

Le modèle sinusoidal

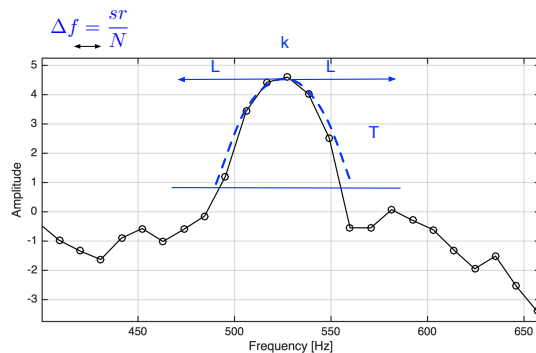
Comment trouver les $\omega_h(t_m)$ et $A_h(t_m)$?

- 3. **Régression** parabolique
 - on peut montrer que la forme du spectre (dans le cas d'une fenêtre d'analyse gaussienne de variance σ^2) est
 - $\log(|X(\omega)|) = \log(A_h) - \frac{\sigma^2}{2}(\omega - \omega_h)$

$$\omega_h = \omega_k + \frac{s}{\Delta}(\log(A_{k+1}) - \log(A_{k-1}))$$

$$\log(A_h) = \frac{\Delta^2}{6s} + \frac{\omega_h^2}{4s} + \frac{\sum \log(A_i)}{3} \quad (2)$$

- avec $\log(A_k) = \log(|X(k, m)|)$
- avec $s = \frac{1}{2(2\pi\sigma)^2}$

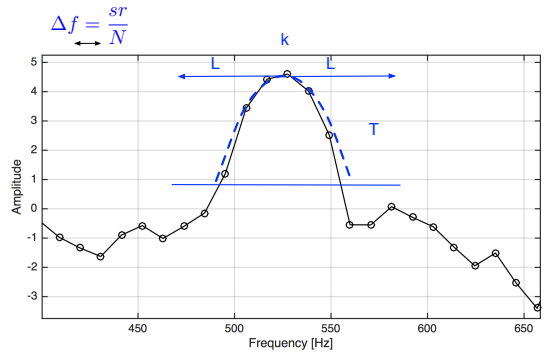


2- Les modèles de signaux

Le modèle sinusoïdal

Comment trouver les $\omega_h(t_m)$ et $A_h(t_m)$?

- 4. Fréquence instantannée pour trouver plus précisément ω_h

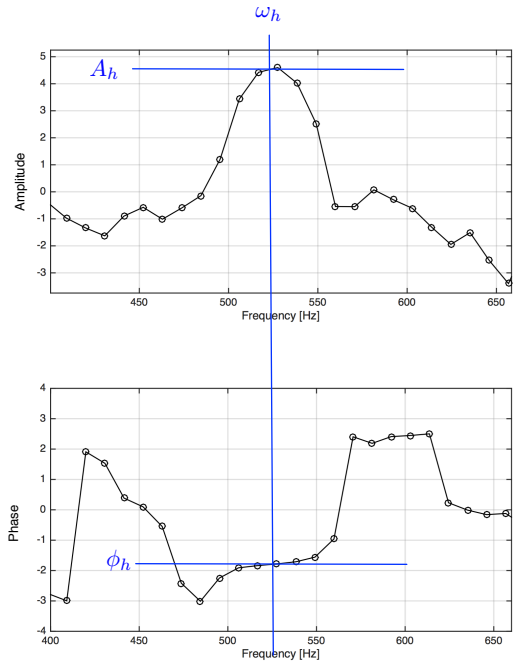


2- Les modèles de signaux

Le modèle sinusoidal

Comment trouver les $\phi_h(m)$?

- Prendre la phase correspondant à ω_h
- Le spectre de phase correspondant à une sinusoïde est un plateau
- Attention :
 - la phase retournée par l'algorithme FFT est donnée par rapport au début de la fenêtre d'analyse
 - mais l'énergie de la fenêtre d'analyse est au milieu
 - il faut corriger le spectre de phase :
 - $x(t + t_0) \Leftrightarrow X(f) \exp(j2\pi f t_0)$
 - avec $t_0 = L/2$

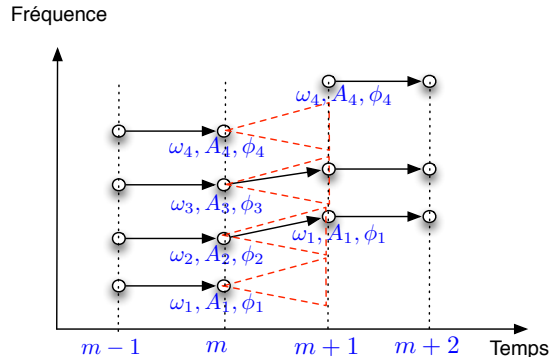


2- Les modèles de signaux

Le modèle sinusoïdal

Création de trajectoire temporelle de sinusoïde (partial tracking)

- A chaque trame d'analyse, on a estimé
 - un ensemble de H triplets de paramètres $\{\omega_h, A_h, \phi_h\}$
- Nous cherchons à créer des trajets continus (à travers le temps) de sinusoïdes
 - Nous devons connecter les $\{\omega_h, A_h, \phi_h\}$ à un instant m donné aux $\{\omega_h, A_h, \phi_h\}$ à l'instant suivant $m + 1$
- Méthode du cône fréquentielle
 - on cherche à connecter le peak $\omega_h(m)$ à un peak $\omega_\gamma(m + 1)$
 - $\omega_\gamma(m + 1) \in [\omega_h - \Delta, \omega_h + \Delta]$
 - Δ représente la variation de fréquence acceptée entre deux trames
 - $A_\gamma(m + 1) \in [A_h - \Delta, A_h + \Delta]$
 - Δ représente la variation d'amplitude acceptée entre deux trames

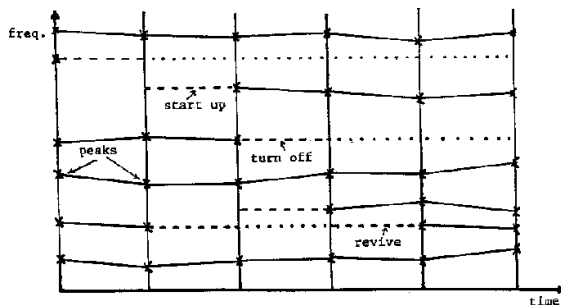


2- Les modèles de signaux

Le modèle sinusoïdal

Création de trajectoire temporelle de sinusoïde (partial tracking)

- Il faut également gérer au cours du temps
 - les naissances de sinusoïdes
 - les morts de sinusoïdes



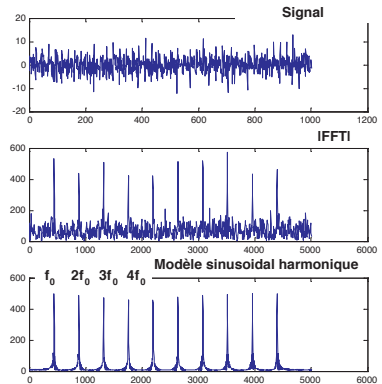
Le modèle sinusoidal harmonique

2- Les modèles de signaux

Le modèle sinusoidal harmonique

Le modèle sinusoidal harmonique

- Même chose que le modèle sinusoidal mais
 - $x(t)$ = source mono-phonique et harmonique
 - note de musique, parties voisées parole
- Conséquences :
 - les fréquences $f_h(t_m) = hf_0(t_m)$
 - Si on connaît $f_0(t_m)$, on en déduit la position des sinusoides : $f_h(t_m) = hf_0(t_m)$,
 - Il ne reste plus qu'à affiner $f_h(t_m)$ et déterminer le $A_h(t_m)$ et le $\phi_h(t_m)$ correspondant
 - Nombre de sinusoides H = nombre d'harmoniques
 - $H(t)$ est constant au cours du temps
 - Plus besoin de créer les trajectoires
 - le ω_1 à m se connecte par définition au ω_1 à $m + 1$



2- Les modèles de signaux Le modèle sinusoidal harmonique

Le modèle sinusoidal harmonique

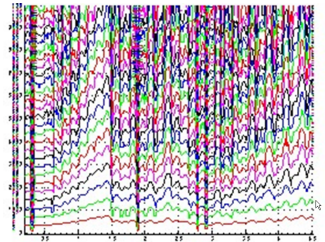
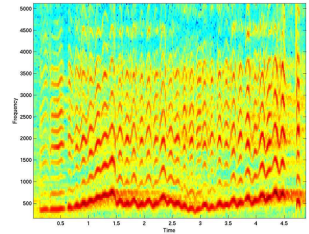
$$\hat{x}_m(t_m) = \sum_{h=1}^{H(t)} A_h(t_m) \cos(\omega_h(t_m)(t - t_m) + \phi_h(t_m))$$

devient

$$\hat{x}_m(t) = \sum_{h=1}^H A_h(t_m) \cos(h\omega_0(t_m)(t - t_m) + \phi_h(t_m)) \quad (3)$$

$$\hat{x}(t) = \sum_m \hat{x}_m(t) + b(t)$$

- Méthode :
 - on commence par détecter la fréquence fondamentale au cours du temps $f_0(t_m)$
 - autour de chaque $h\omega_0(t_m)$ on cherche les valeurs précises de $\{\omega_h, A_h, \phi_h\}$
 - pas de création de trajets



2- Les modèles de signaux

Le modèle sinusoidal harmonique

Utilisation du modèle sinusoidal, sinusoidal harmonique ?

- Transformations de haute qualité
 - traitement séparé de la partie sinusoidale, de la partie bruitée (contrairement au vocodeur de phase)
- Traitements plus poussés
 - changer l'harmonicité du signal, modifier certaines harmoniques
- Compression
 - transmission uniquement de $f_0(t)$ et de l'enveloppe spectrale
 - valeurs de A_h sous forme compressé
- Indexation audio
 - extraire des descripteurs audio plus précis du signal

Différentes méthodes d'estimation

- de la fréquence fondamentale
 - auto-corrélation, AMDF, Yin, Cepstre, maximum de vraisemblance, ...
- de l'enveloppe spectrale
 - LPC, Cepstre, MFCC, ...

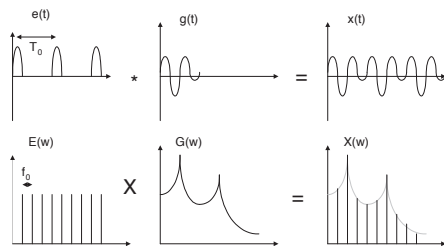
Le modèle source/ filtre

2- Les modèles de signaux

Le modèle source/ filtre

Le modèle source/ filtre

- Hypothèse :
 - le signal $x(t)$ est le résultat du passage d'une excitation (un pulse, une série de pulse) dans un filtre (résonnant)
 - Exemples : le signal de parole, certains instruments de musique (trompette)
- Modélisation temporelle :
 - un signal d'excitation $e(t)$ passe (convolution) à travers un filtre $g(t)$:
 - $x(t) = e(t) \otimes g(t)$
- Modélisation fréquentielle
 - la multiplication de la TF du signal d'excitation (source) par la TF du filtre.
 - $X(\omega) = E(\omega) \cdot G(\omega)$

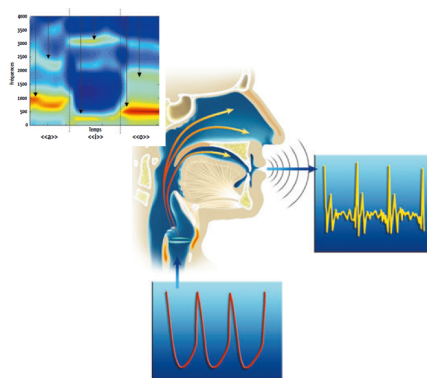


Production du signal vocal

2- Les modèles de signaux

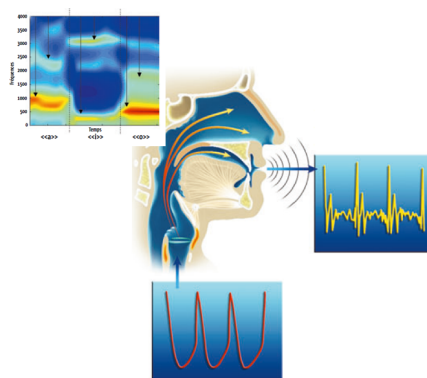
Production du signal vocal

- Le signal de parole (pour sa partie voisée) est créé par
 - les cordes vocales
 - une excitation régulière, périodique
 - le conduit bucco-nasal (bouches et nez)
 - filtrage résonant et anti-résonant



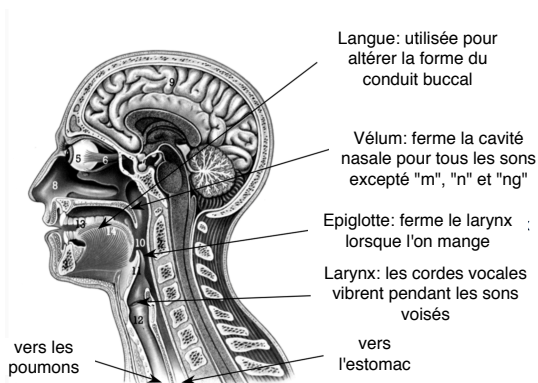
2- Les modèles de signaux Production du signal vocal

- Ouverture/fermeture périodique des cordes vocales
 - détermine la hauteur
 - hauteur de 100Hz?
 - pulses d'air espacés de
$$T_0 = \frac{1}{f_0} = \frac{1}{100} = 10ms.$$
 - appelé **signal d'excitation** (ou **source**), $e(t)$.
- Conduit bucco-nasal
 - créer les différentes voyelles pour une hauteur donnée
 - renforce (résonance) et retire (anti-résonances) certains fréquences.
 - filtre résonant (AR : Auto-Regressif)
 - filtre anti-résonant (MA : Moving Average)
 - Total = filtre dit "ARMA".

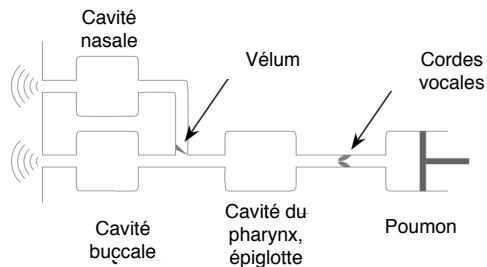


2- Les modèles de signaux

Production du signal vocal

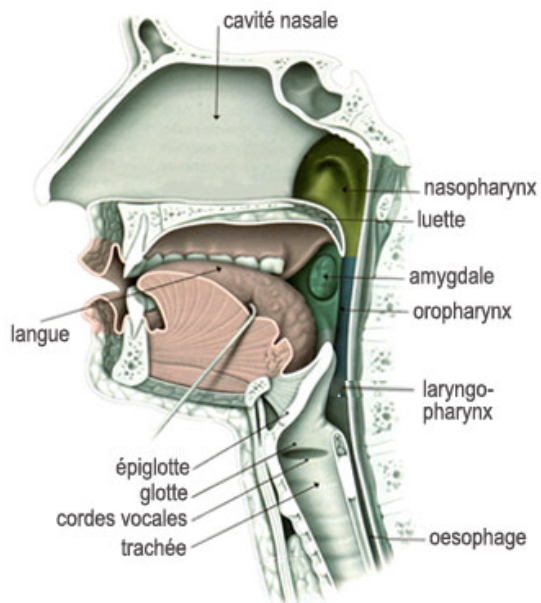


source : Mike Brookes

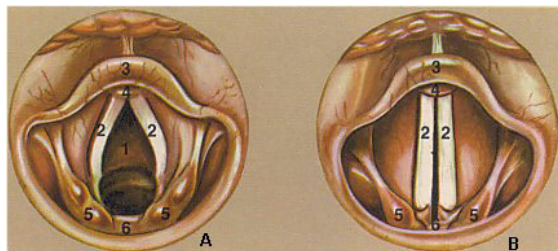


source : Mike Brookes

2- Les modèles de signaux Production du signal vocal

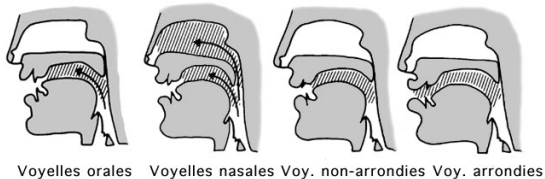


source : outilsrecherche.over-blog.com

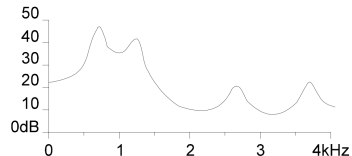
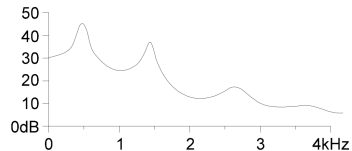
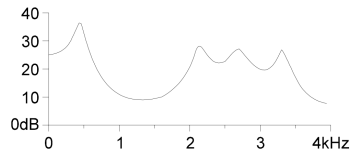


source : outilsrecherche.over-blog.com

2- Les modèles de signaux Production du signal vocal



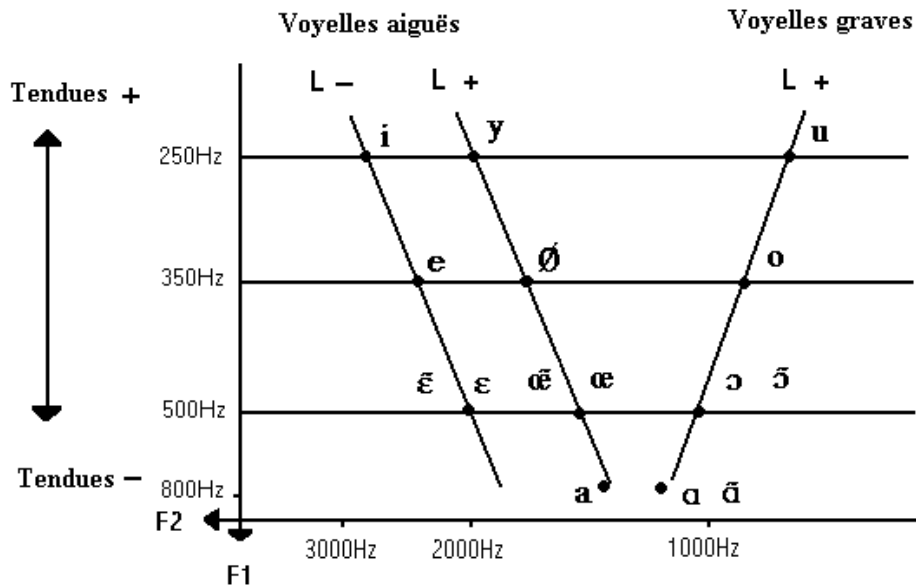
source : outilsrecherche.over-blog.com



source : Mike Brookes

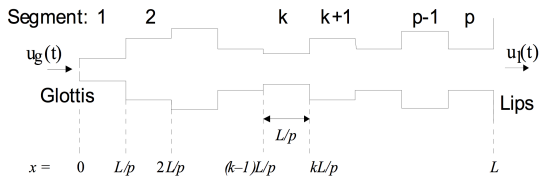
2- Les modèles de signaux Production du signal vocal

Fréquences des formants pour les différentes voyelles



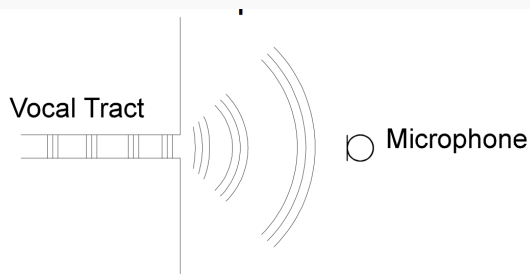
2- Les modèles de signaux Production du signal vocal

Représentation sous-forme de tube



source : Mike Brookes

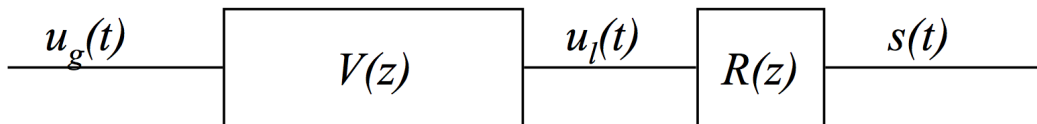
Radiation des lèvres



source : Mike Brookes

- Filtre passe-haut $R(z) = 1 - z^{-1}$

Système équivalent



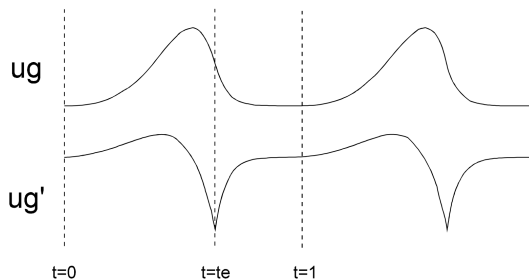
2- Les modèles de signaux Production du signal vocal

Modèle de forme-d'onde pour la glotte

- Modèle "LF" (Liljencrants et Fant)

$$\begin{aligned}u'_g(t) &= e^{at} \sin(bt) \text{ pour } 0 \leq t < t_e \\ &= c + de^{-ft} e^{at} \sin(bt) \text{ pour } t_e \leq t < 1\end{aligned}$$

- avec $u_g(0) = u_g(1) = 0$ et $u'_g(t)$ continu en t_e



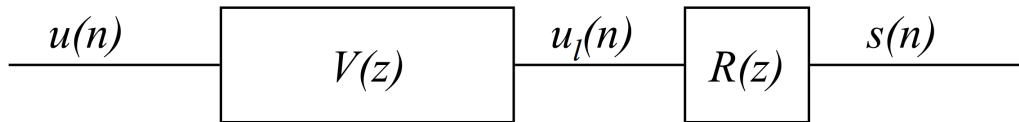
source : Mike Brookes

La prédiction linéaire, modèle auto-régressif (AR)

2- Les modèles de signaux

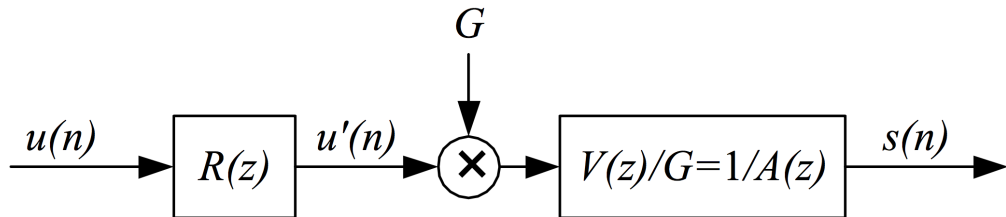
La prédiction linéaire, modèle auto-régressif (AR)

- Système équivalent de production vocale



source : Mike Brookes

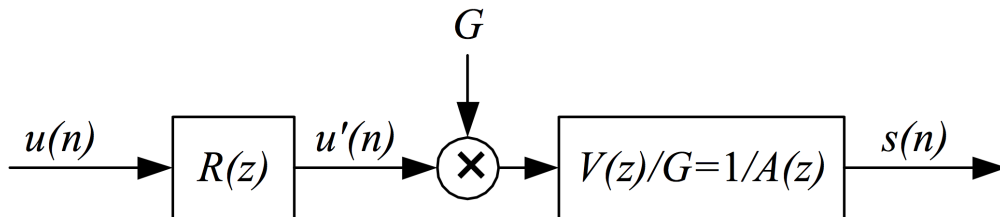
- Inversion de l'ordre de $V(z)$ et $R(z)$
 - puisque linéaire et
 - puisque $V(z)$ ne change pas significativement durant la réponse impulsionnelle de $R(z)$ et inversement



source : Mike Brookes

2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)



source : Mike Brookes

- $x(n) = Gu'(n) + \sum_{j=1}^P a_j x(n-j)$
 - Si le gain des résonances du conduit vocal est important, le second terme va dominer
- $x(n) \simeq \sum_{j=1}^P a_j x(n-j)$
 - La partie de droite est la prédiction de $x(n)$ comme combinaison linéaire des échantillons passés de la voix

2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)

Modèle auto-régressif : $x(n) \simeq \sum_{j=1}^P a_j x(n-j)$

- On définit l'erreur de prédiction à l'échantillon n comme

$$\begin{aligned} e(n) &= x(n) - \sum_{j=1}^P a_j x(n-j) \\ &= x(n) - a_1 x(n-1) - a_2 x(n-2) \dots - a_P x(n-P) \end{aligned}$$

- En transformée en Z
 - $E(z) = X(z)A(z)$
- Etant donné une trame de signal de parole $\{F\}$, on cherche les valeurs a_i qui minimize
 - $Q_E = \sum_{n \in \{F\}} e^2(n)$
- Minimisation par rapport aux $a_i \rightarrow$ différenciation de Q_E par rapport aux a_i

$$\frac{\partial Q_E}{\partial a_i} = \sum_{n \in \{F\}} \frac{\partial (e^2(n))}{\partial a_i} = \sum_{n \in \{F\}} 2e(n) \frac{\partial e(n)}{\partial a_i} = - \sum_{n \in \{F\}} 2e(n)x(n-i)$$

2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)

- Les valeurs optimales de a_i doivent satisfaire les P équations :

$$\text{pour } i = 1 \dots P \quad \sum_{n \in \{F\}} e(n)x(n-i) = 0$$

$$\text{puisque } e(n) = x(n) - \sum_{j=1}^P a_j x(n-j)$$

$$\text{on a } \sum_{n \in \{F\}} \left(x(n)x(n-i) - \sum_{j=1}^P a_j x(n-j)x(n-i) \right) = 0$$

$$\sum_{j=1}^P a_j \sum_{n \in \{F\}} x(n-j)x(n-i) = \sum_{n \in \{F\}} x(n)x(n-i)$$

- Système de i équations à résoudre

$$\sum_{j=1}^P \phi_{ij} a_j = \phi_{i0}$$

- avec $\phi_{ij} = \sum_{n \in \{F\}} x(n-i)x(n-j)$

2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)

- Système de i équations à résoudre

$$\sum_{j=1}^P \phi_{ij} a_j = \phi_{i0} \quad (4)$$

- avec $\phi_{ij} = \sum_{n \in \{F\}} s(n-i)s(n-j)$
- Sous forme matricielle

$$\begin{aligned} \underline{\underline{\phi}} \underline{\underline{a}} &= \underline{\underline{c}} \\ \underline{\underline{a}} &= \underline{\underline{\phi}}^{-1} \underline{\underline{c}} \end{aligned} \quad (5)$$

- avec $\underline{\underline{\phi}}$ matrice **symétrique** et définie **semi-positive**

- Rappel

- Matrice **symétrique** :
 - $\phi_{ji} = \phi_{ij} \Leftrightarrow \underline{\underline{\phi}}^T = \underline{\underline{\phi}}$
- Matrice **définie semi-positive** :
 - $\sum_{i,j} x_i \phi_{ij} x_j \geq 0 \Leftrightarrow \underline{\underline{x}}^T \underline{\underline{\phi}} \underline{\underline{x}} \geq 0$ pour tout $\underline{\underline{x}}$
- Matrice de **Toeplitz** :
 - les diagonales sont constantes :
 - $\phi_{i+1,j+1} = \phi_{ij} = f(i-j)$

2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)

Solution 1 : prédiction linéaire par auto-corrélation

- on choisit $\{F\}$ comme l'intervall infini
 - $\phi_{ij} = \sum_{n=-\infty}^{+\infty} x(n-i)x(n-j)$
- à cause de la symétrie et de l'intervall infini on a
 - $\phi_{ij} = \phi_{|i-j|,0} = R_{|i-j|}$
 - avec R_k la séquence d'auto-corrélation du signal de parole
- Dans ce cas ϕ_{ij} est une matrice de Toeplitz (les diagonales sont constantes),
 - inversion en $O(p^2)$ au lieu de $O(p^3)$
- Les équations $\underline{\phi}\underline{a} = \underline{c}$ sont appelées **équations de Yule-Walker**.
 - l'algorithme d'inversion correspondant est appelé **algorithme de Durbin-Levinson**
- Rappel
 - Matrice **symétrique** :
 - $\phi_{ij} = \phi_{ji} \Leftrightarrow \underline{\phi}^T = \underline{\phi}$
 - Matrice **définie semi-positive** :
 - $\sum_{i,j} x_i \phi_{ij} x_j \geq 0 \Leftrightarrow \underline{x}^T \underline{\phi} \underline{x} \geq 0$ pour tout \underline{x}
 - Matrice de **Toeplitz** :
 - les diagonales sont constantes :
 - $\phi_{i+1,j+1} = \phi_{ij} = f(i-j)$

2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)

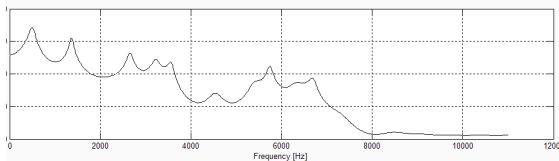
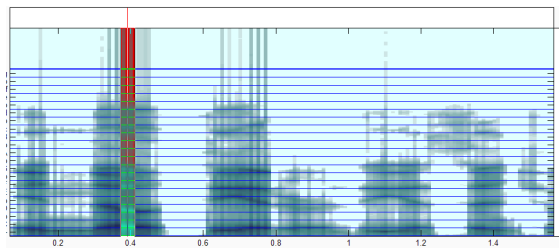
Solution 2 : prédiction linéaire par covariance

- on choisit $\{F\}$ comme un interval fini de parole $\{F\} = s(n) \quad 0 \leq n \leq (N-1)$
 - $\phi_{ij} = \sum_{n=0}^{N-1} x(n-i)x(n-j)$
- la matrice ϕ_{ij} est symétrique mais plus de Toeplitz
 - calcul plus lourd $O(p^3)$
- Rappel
 - Matrice **symétrique** :
 - $\phi_{ji} = \phi_{ij} \Leftrightarrow \underline{\underline{\phi}}^T = \underline{\underline{\phi}}$
 - Matrice **définie semi-positive** :
 - $\sum_{i,j} x_i \phi_{ij} x_j \geq 0 \Leftrightarrow \underline{\underline{x}}^T \underline{\underline{\phi}} \underline{\underline{x}} \geq 0$ pour tout x
 - Matrice de **Toeplitz** :
 - les diagonales sont constantes :
 - $\phi_{i+1,j+1} = \phi_{ij} = f(i-j)$

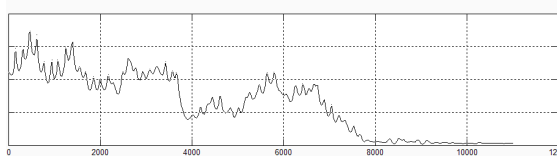
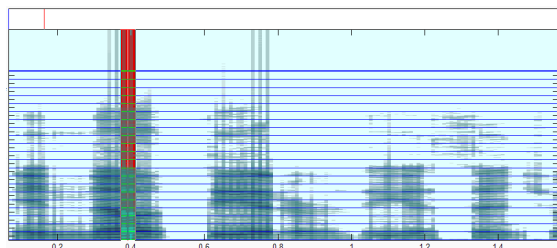
2- Les modèles de signaux

La prédiction linéaire, modèle auto-régressif (AR)

Choix du nombre de pôle P



$P = 40$



$P = 200$

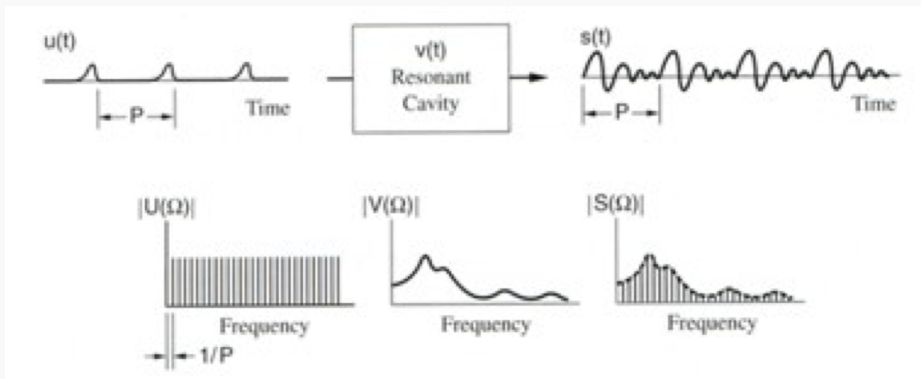
Transformation du signal par la méthode P-SOLA

2- Les modèles de signaux

Transformation du signal par la méthode P-SOLA

Transformation du signal par la méthode P-SOLA

- Modèle source/filtre
 - un signal d'excitation $e(t)$ passe (convolution) à travers un filtre $g(t)$:
 - $x(t) = e(t) \otimes g(t)$
 - $e(t) = \sum_m \delta(t - mT_0)$
 - Modélisation fréquentielle
 - la multiplication de la TF du signal d'excitation (source) par la TF du filtre.
 - $X(\omega) = E(\omega) \cdot G(\omega)$
 - $X(\omega) = \frac{1}{T_0} G(h\omega_0)$

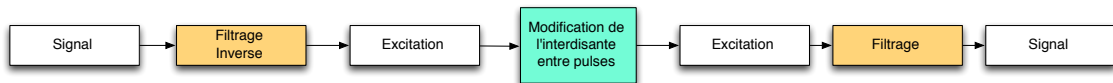
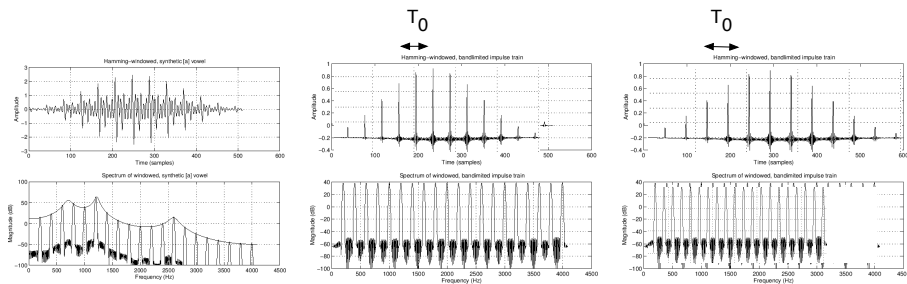


2- Les modèles de signaux

Transformation du signal par la méthode P-SOLA

Transformation du signal par la méthode P-SOLA

- LP-P-SOLA (Linear Predictive P-SOLA)
 - déconvolution du signal de parole $x(t)$ par le filtre de prédiction linéaire $g(t)$ estimé
 - modification de la position des pulses glottiques dans $e(t)$
 - convolution du signal résultant par le filtre de prédiction linéaire $g(t)$

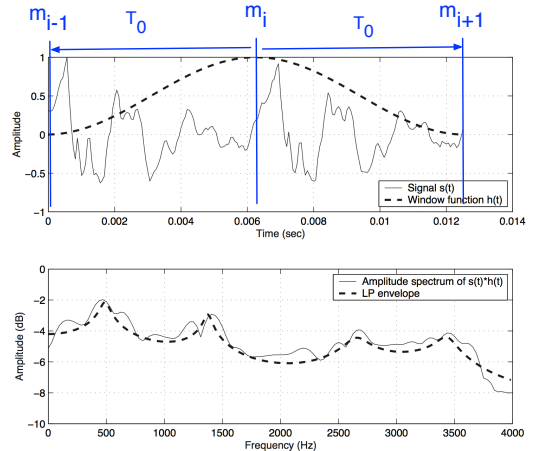


2- Les modèles de signaux

Transformation du signal par la méthode P-SOLA

Transformation du signal par la méthode P-SOLA

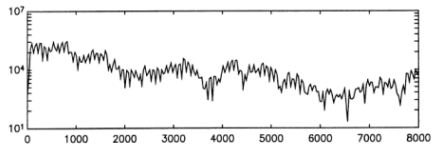
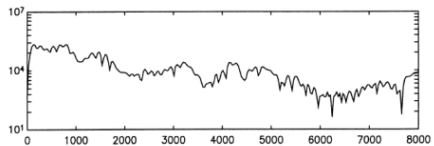
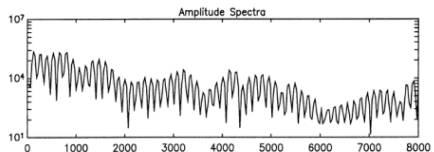
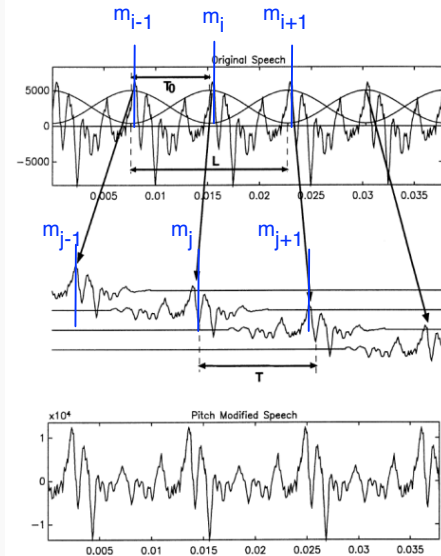
- **Pitch-Synchronous** Overlap Add
- P-SOLA **Analyse**
 - **forme d'onde élémentaire**
 - fenêtrage du signal autour des instants de fermeture de glotte sur une longueur $[-T_0, T_0]$
 - $x_i(t) = x(t - m_i)w_i(t)$
 - m_i : instants d'analyse, position des instants de fermeture de glotte
 - tel que $m_{i+1} - m_i = T_0$
 - tel que $w_i(t) = w\left(\frac{t}{2T_0(m_i)}\right)$
fenêtre de taille égale à $2T_0$
 - on fait l'hypothèse que $x_i(t)$ approxime bien la R.I. du filtre de prédiction linéaire



2- Les modèles de signaux

Transformation du signal par la méthode P-SOLA

P-SOLA Synthèse (abaissement de la hauteur : $T > T_0$)



2- Les modèles de signaux

Transformation du signal par la méthode P-SOLA

Transformation du signal par la méthode P-SOLA

- P-SOLA **Synthèse**
 - pour modifier la hauteur :
 - on modifie directement l'interdistance entre formes d'ondes élémentaires
 - $m_{j+1} - m_j = T$
 - pour allonger la durée :
 - on recopie plusieurs fois formes d'ondes élémentaires
 - pour raccourci la durée :
 - on retire certaines formes d'ondes élémentaires
- re-synthèse par addition/recouvrement dans le domaine temporelle
 - $\hat{x}(t) = \sum_j x_j(t + m_j)$
 - m_j instants de synthèse

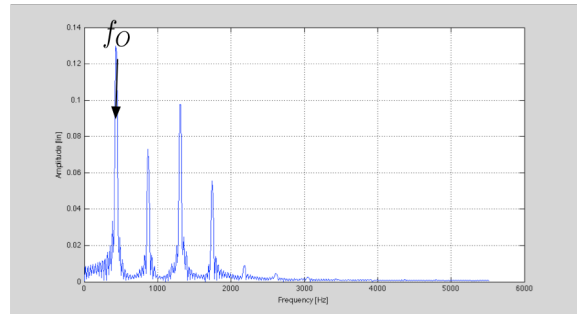
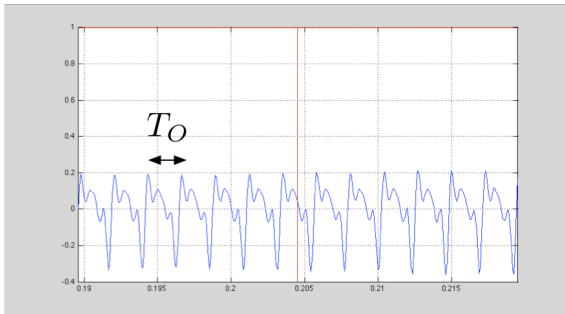
- P-SOLA Pro :
 - synthèse extrêmement rapide, de très grandes qualité pour la parole
- P-SOLA Con :
 - limiter au traitement de signaux répondant au modèle source/filtre
 - approximation difficile pour les pitchs élevés
 - nécessite l'estimation de $f_0(t)$
 - nécessite une estimation des instants de fermeture de glotte t_a^i

3- Utilisation de modèles de signal

3- Utilisation de modèles de signal

Période fondamentale T_0 ou fréquence fondamentale f_0

- f_0 : fréquence fondamentale en Hz
 - exemple La3/A4 = 440Hz
- $T_0 = \frac{1}{f_0}$: période fondamentale en secondes
 - exemple La3/A4 = 0.0023s.



3- Utilisation de modèles de signal

Méthodes temporelles

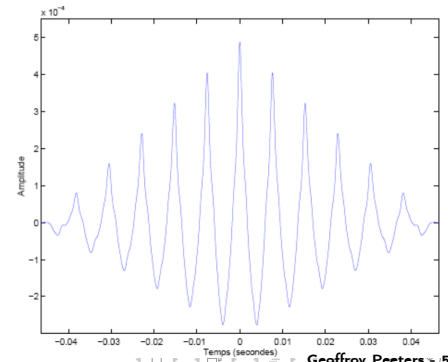
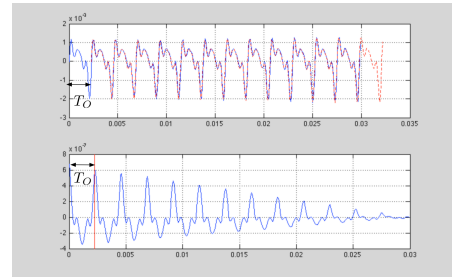
Auto-corrélation biaisée

$$\hat{r}_x(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n)x(n+m) \text{ si } m \geq 0$$

- Propriétés :

$$E[\hat{r}_x(m)] = \frac{N - |m|}{N} r_x(m)$$

$$|\hat{r}_x(m)| \leq \hat{r}_x(0)$$



3- Utilisation de modèles de signal

Méthodes temporelles

Auto-corrélation non-biaisée

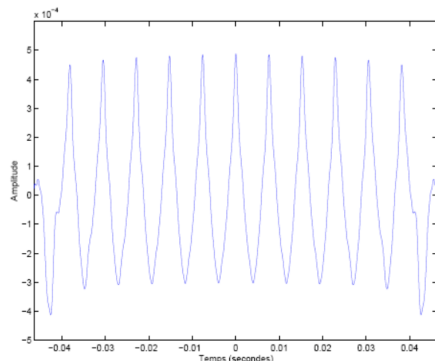
$$\tilde{r}_x(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} x(n)x(n+m) \text{ si } m \geq 0$$

- Propriétés :

$$E[\tilde{r}_x(m)] = r_x(m)$$

$$\text{Var}[\tilde{r}_x(m)] = \left(\frac{N}{N-m}\right)^2 \text{Var}[\hat{r}_x(m)]$$

$$|\tilde{r}_x(m)| \leq \tilde{r}_x(0)$$



source : Richard, 2012

3- Utilisation de modèles de signal

Méthodes temporelles

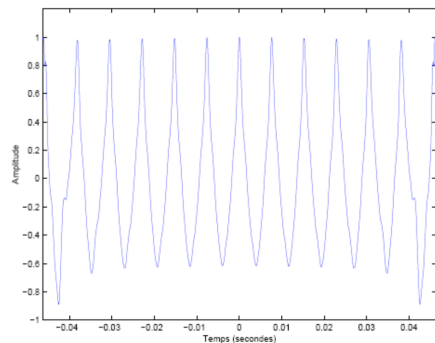
Auto-corrélation normalisée

$$\bar{r}_x(m) = \frac{\sum_{n=0}^{N-1-m} x(n)x(n+m)}{\sqrt{\sum_{n=0}^{N-1-m} x(n)^2} \sqrt{\sum_{n=0}^{N-1-m} x(n+m)^2}}$$

- Propriétés :

$$|\bar{r}_x(m)| \leq \bar{r}_x(0) = 1$$

$$|\bar{r}_x(m)| = 1 \text{ ssi les vecteurs sont colinéaires}$$



source : Richard, 2012

3- Utilisation de modèles de signal

Méthodes temporelles

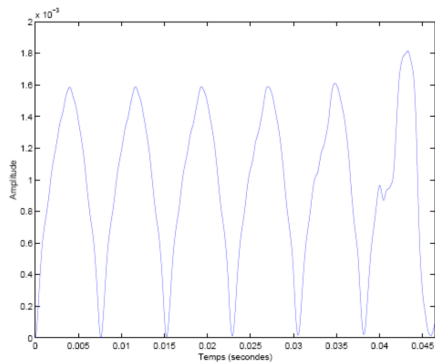
Average Square Difference Function (ASDF)

$$ASDF(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} (x(n) - x(n+m))^2$$

- La période T_0 peut être estimée en recherchant le minimum de l'écart quadratique entre les signaux $x(n)$ et $x(n+m)$
- Propriétés :

$$ASDF(m) = 0 \text{ ssi } x \text{ est de période } T_0 = m$$

$$E[ASDF(m)] = 2(r_x(0) - r_x(m))$$



source : Richard, 2012

3- Utilisation de modèles de signal

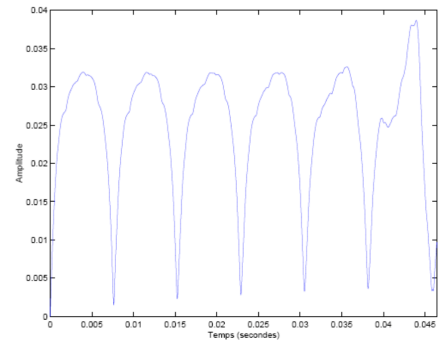
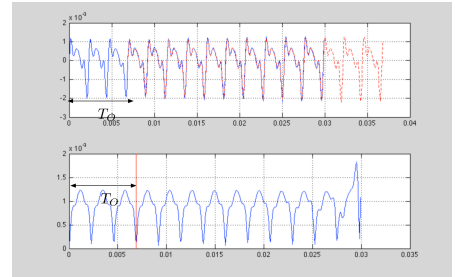
Méthodes temporelles

Average Magnitude Difference Function (AMDF)

$$AMDF(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x(n) - x(n+m)|$$

- Propriétés :

$$AMDF(m) = 0 \text{ ssi } x \text{ est de période } T_0 = m$$



3- Utilisation de modèles de signal

Méthodes temporelles

Algorithme Yin

[A. de Cheveigné, H. Kawahara, YIN, a fundamental frequency estimator for speech and music, JASA, 2002]

- Point de départ : méthode de l'auto-corrélation
- Améliorations
 - 1) Utilisation de l'ASDF
 - 2) Normalisation
 - 3) Seuillage
 - 4) Interpolation
 - 5) Minimisation locale en temps

Version	Gross error (%)
Step 1	10.0
Step 2	1.95
Step 3	1.69
Step 4	0.78
Step 5	0.77
Step 6	0.50

source : Richard, 2012

3- Utilisation de modèles de signal

Méthodes temporelles

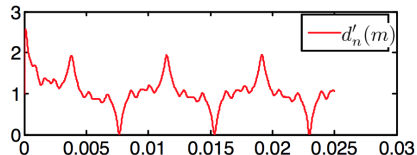
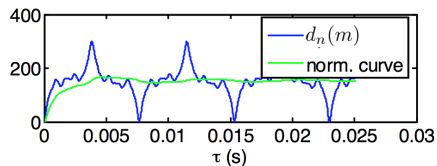
Algorithme Yin

- 1) Utilisation de l'ASDF
 - $d_t(\tau) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2$
 - lien avec l'auto-corrélation
 $d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau)$
 - Gain ?
 - l'ASDF est beaucoup moins sensible aux variations des amplitudes relatives que l'ACF (qui est sensible, par exemple, à l'accentuation des partiels d'ordre pair)

- 2) Normalisation
 - Normalisation par la "moyenne cumulée"

$$d'_t(\tau) = \begin{cases} 1 & \text{si } \tau = 0 \\ \frac{d_t(\tau)}{\frac{1}{\tau} \sum_{j=1}^{\tau} d_t(j)} & \text{sinon} \end{cases}$$

- Gain ?
 - permet d'éviter les erreurs pour les F0 élevées (suppression du lobe en 0)



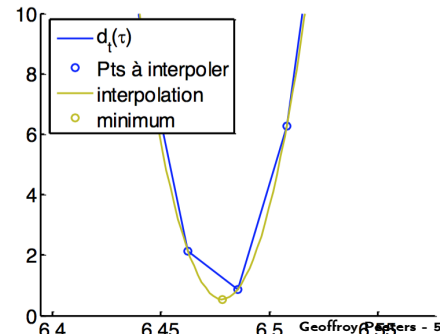
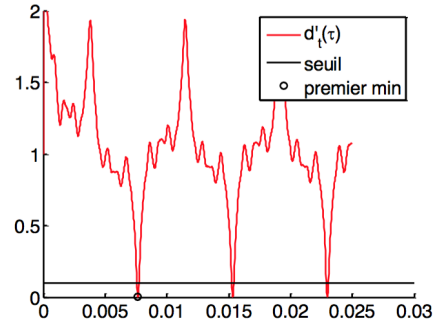
source : Richard, 2012

3- Utilisation de modèles de signal

Méthodes temporelles

Algorithme Yin

- 3) Seuillage absolu
 - La plus petite période inférieure au seuil est choisie
 - Si aucune période n'est inférieure au seuil, alors le minimum global est choisi
- 4) Interpolation parabolique autour du minimum
 - Réalisée sur $d_t(\tau)$ (i.e avant normalisation)
 - Gain : meilleure précision sur la valeur de F_0
- 5) Minimisation locale en temps
 - si on note T_t la période estimée au temps t
 - pour un temps t ,
 - on cherche pour $\theta \in [t - T_{\max}/2, t + T_{\max}/2]$ (T_{\max} est la période la plus grande considérée, 25 ms)
 - le minima de $d_\theta(T_\theta)$
 - on réitère avec cette nouvelle estimation et intervalle de recherche de $\pm 20\%$
 - Gain : effet de lissage en cas de fluctuations de l'estimation
- Autres méthodes possibles pour le lissage : filtre médian, programmation dynamique



3- Utilisation de modèles de signal

Méthodes temporelles

Algorithme Yin

Evaluation sur quatre bases de données de parole

- annotées automatiquement (par YIN, à partir du laryngographe) puis vérifiées et triées à la main

Method	Gross error (%)					
	DB1	DB2	DB3	DB4	Average	(low/high)
pda	10.3	19.0	17.3	27.0	16.8	(14.2/2.6)
fxac	13.3	16.8	17.1	16.3	15.2	(14.2/1.0)
fxcep	4.6	15.8	5.4	6.8	6.0	(5.0/1.0)
ac	2.7	9.2	3.0	10.3	5.1	(4.1/1.0)
cc	3.4	6.8	2.9	7.5	4.5	(3.4/1.1)
shs	7.8	12.8	8.2	10.2	8.7	(8.6/0.18)
acf	0.45	1.9	7.1	11.7	5.0	(0.23/4.8)
nacf	0.43	1.7	6.7	11.4	4.8	(0.16/4.7)
additive	2.4	3.6	3.9	3.4	3.1	(2.5/0.55)
TEMPO	1.0	3.2	8.7	2.6	3.4	(0.53/2.9)
YIN	0.30	1.4	2.0	1.3	1.03	(0.37/0.66)

source : Richard, 2012

3- Utilisation de modèles de signal

Méthodes temporelles

Cepstre réel

- **Auto-correlation** du signal temporel $\hat{r}(\tau)$:

$$\hat{r}(\tau) = \int_t x^*(t)x(t+\tau)dt$$

- Sa Transformée de Fourier $\Gamma(\omega)$:

$$\Gamma(\omega) = \int_{\tau} \left(\int_t x^*(t)x(t+\tau)dt \right) e^{-j\omega\tau} d\tau$$

$$\Gamma(\omega) = |X(j\omega)|^2$$

- Donc **Auto-correlation** du signal temporel :

$$\hat{r}(l) = \frac{1}{N-l} \sum_k |X(k)|^2 \cos\left(2\pi k \frac{l}{N}\right)$$

- **Cepstre réel** du signal temporel :

$$\hat{c}(l) = \frac{1}{N-l} \sum_k \log(|X(k)|) \cos\left(2\pi k \frac{l}{N}\right)$$

- Relation avec le modèle source/filtre :

$$x(t) = e(t) \otimes g(t)$$

$$X(\omega) = E(\omega) \cdot G(\omega)$$

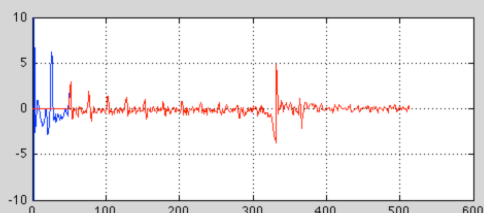
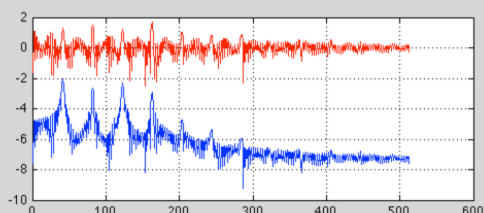
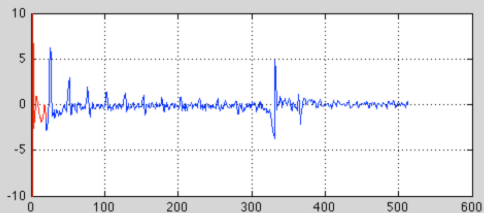
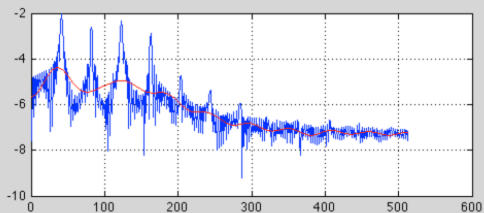
$$\log(X(\omega)) = \log(E(\omega)) + \log(G(\omega))$$

3- Utilisation de modèles de signal

Méthodes temporelles

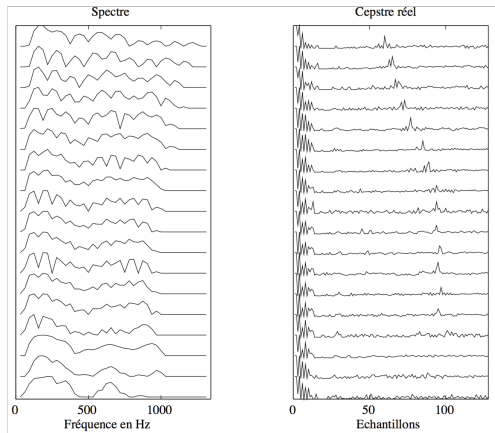
Cepstre réel

- Le cepstre permet de séparer
 - l'enveloppe spectrale
 - ce qui varie lentement
 - basse fréquence de la TF^{-1}
 - la fréquence fondamentale
 - ce qui varie rapidement
 - haute fréquence de la TF^{-1}

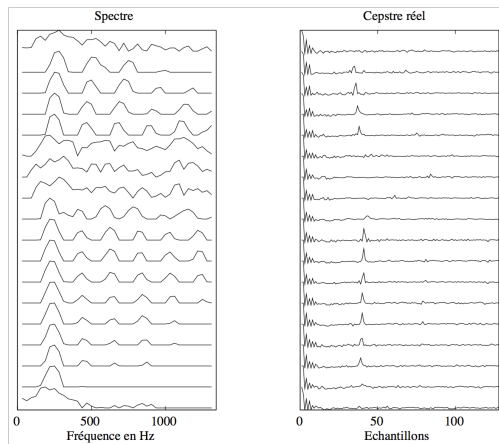


3- Utilisation de modèles de signal Méthodes temporelles

Cepstre réel



source : voix d'homme, Laroche, 1995



source : voix de femme, Laroche, 1995

3- Utilisation de modèles de signal

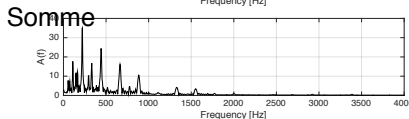
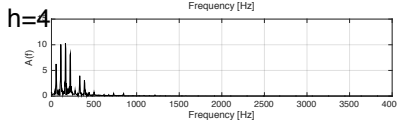
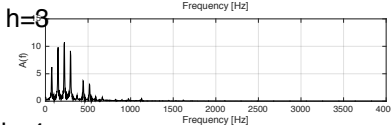
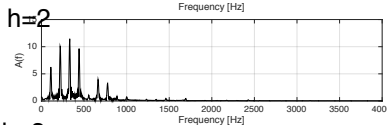
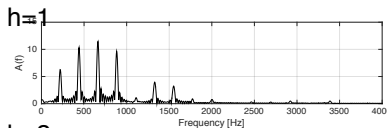
Méthodes fréquentielles

Somme spectrale

- On peut montrer que la maximisation de L par rapport à $F_0 = \frac{m}{N}$ revient à maximiser la somme spectrale

$$S(e^{j2\pi \frac{m}{N}}) = \sum_{h=1}^H \hat{R}_x(e^{j2\pi \frac{m}{N} \cdot h})$$

$$S(\omega) = \sum_{h=1}^H |X(e^{j\omega \cdot h})|^2 \text{ pour } \omega < \frac{\pi}{H}$$



3- Utilisation de modèles de signal

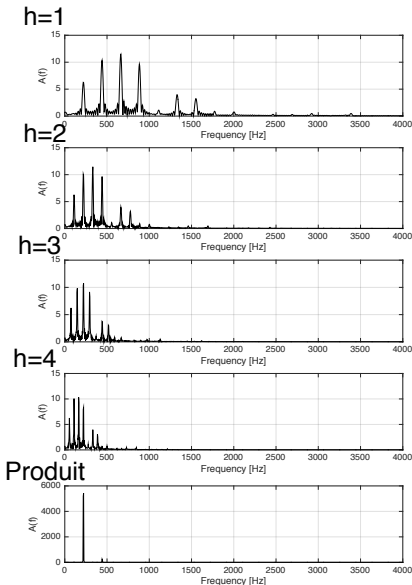
Méthodes fréquentielles

Produit spectral

- Par similitude avec la somme spectrale on peut définir le produit spectral (souvent plus robuste)

$$P(e^{j2\pi \frac{m}{N}}) = \prod_{h=1}^H \hat{R}_x(e^{j2\pi \frac{m}{N} \cdot h})$$

$$P(\omega) = \prod_{h=1}^H |X(e^{j\omega \cdot h})|^2 \text{ pour } \omega < \frac{\pi}{H}$$



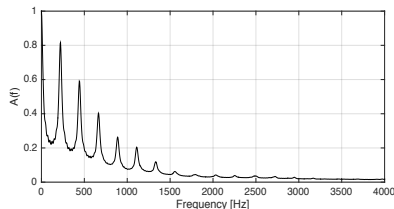
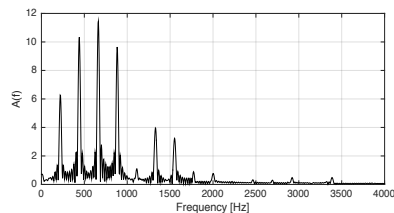
3- Utilisation de modèles de signal

Méthodes fréquentielles

Auto-corrélation du spectre d'amplitude

- Mesure de la périodicité de l'espace entre les harmoniques
 - ne fait pas l'hypothèse qu'il existe de l'énergie à la fréquence f_0

$$\hat{R}(k) = \frac{1}{N-k} \sum_{\kappa=0}^{N-k-1} |X(\kappa)| |X(\kappa+k)| \quad (11)$$



Applications du traitement audio pour la description musicale

4- Applications du traitement audio pour la description musicale



Enter a keyword, record a query or drag an example clip.



Search Audio

[Audio Preferences](#)
[Audio Help](#)



[Steve Jobs interview](#)

7 min 14 sec
Speech



[Metric - Raw Sugar](#)

3 min 47 sec
Music - Indie Pop



[Grenade explosion](#)

23 sec
Sound effect

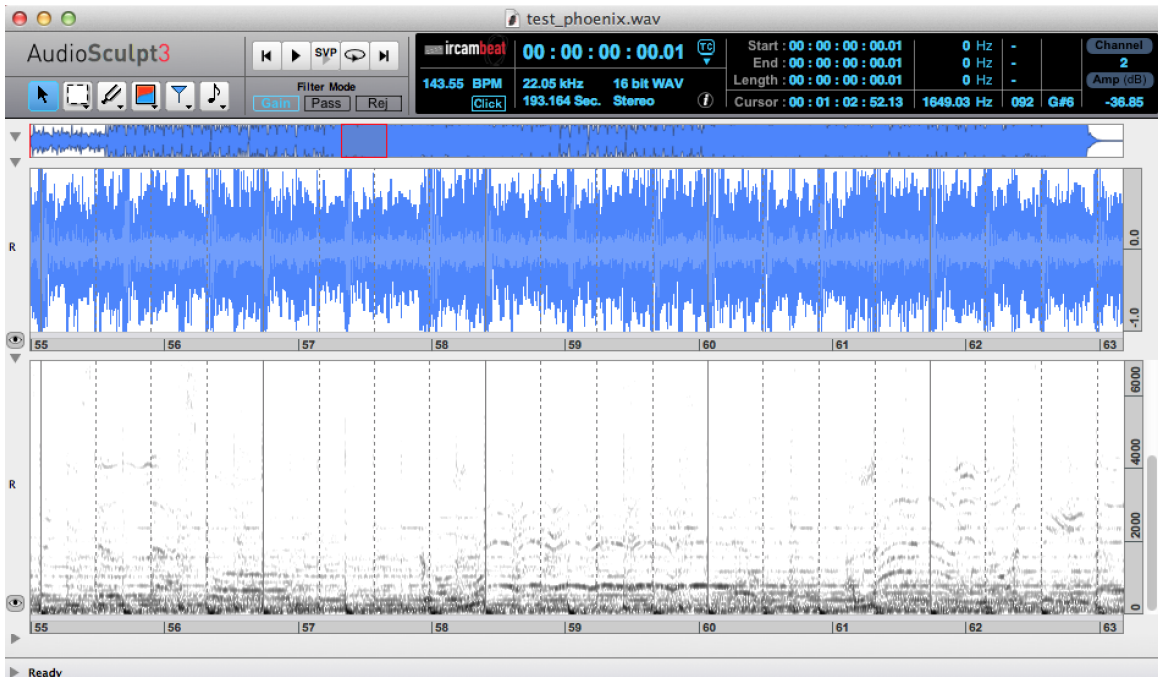
4- Applications du traitement audio pour la description musicale

- Identification audio
 - recherche de doublons, gestion de copyright, attacher des méta données à une instance d'un morceau



4- Applications du traitement audio pour la description musicale

- Estimation du tempo, de la position des temps/ premier-temps
 - DJing, manipulation du contenu (add swing ...)



4- Applications du traitement audio pour la description musicale

- Nouveaux modes de recherche :
 - par chantonement/ sifflement

Query by Humming v0.51b

IIS

Query by Humming

1.7 3.3 5.0 6.6 8.3 sec

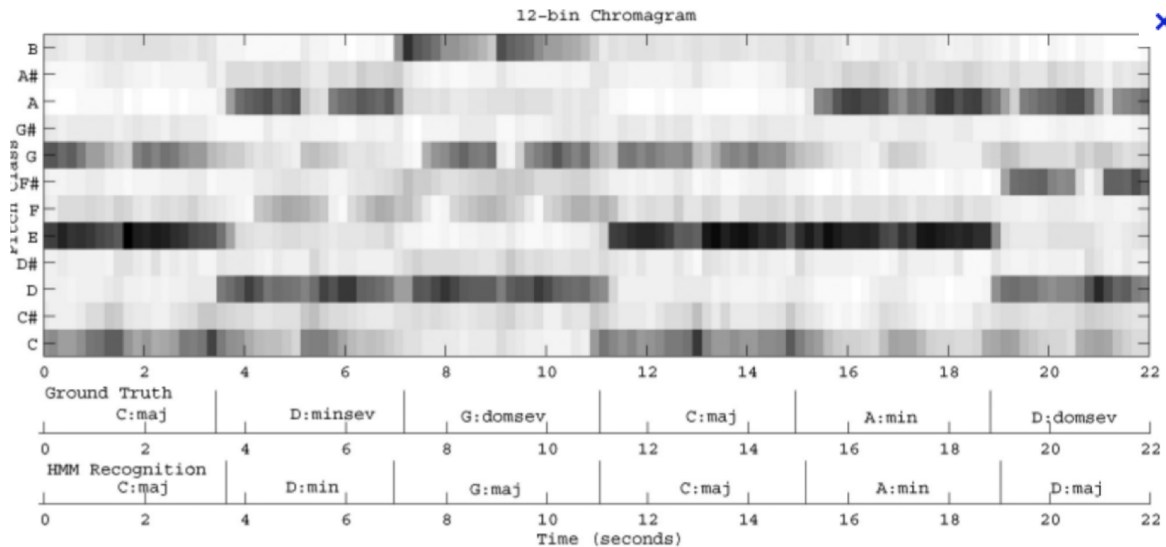
Mic-Gain

good Find ster Top10

Geoffroy Peeters - 80

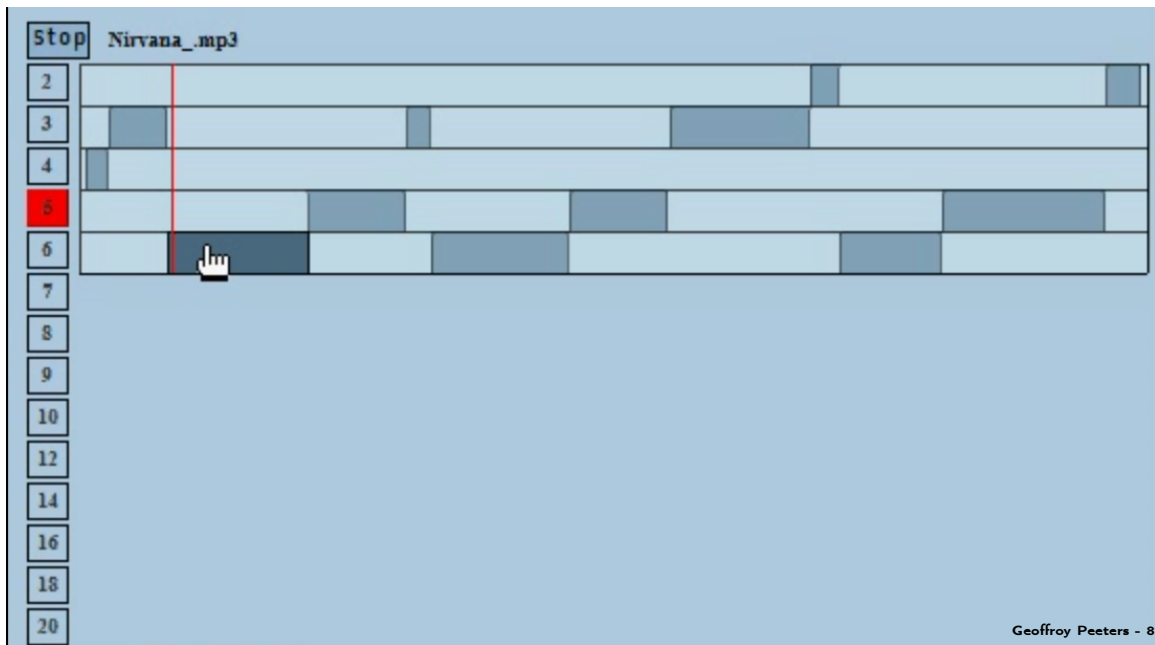
4- Applications du traitement audio pour la description musicale

- Estimation des accords
 - obtenir des guitar-tab automatiquement



4- Applications du traitement audio pour la description musicale

- Navigation à l'intérieur d'un morceau de musique par couplet/refrain
 - Génération automatique de résumé audio
- Dé-linéarisation d'un flux audio :
 - segmentation de flux radio, télé et étiquetage des parties



4- Applications du traitement audio pour la description musicale

- Détection des cover, reprises ou ... des plagias

Titre	Artiste	Album	D.	Pop.	
Let It Be	∨ The Beatles Recovered Band	30 Beatles Top Hits	03:50		<input type="checkbox"/>
Let It Be	∨ The Hit Co., The Tribute Co.	A Tribute to the Beatles: The Lat...	03:42		<input type="checkbox"/>
Let It Be	∨ Labrinth	Let It Be	03:05		<input type="checkbox"/>
Let It Be Me	∨ Ray LaMontagne	Gossip in The Grain	04:41		<input type="checkbox"/>
Let It Be – The Beatles Tribute	Let It Be	Let It Be – The Beatles Tribute	03:49		<input type="checkbox"/>
Let It Be	Lois	Let It Be – The Voice 2	03:15		<input type="checkbox"/>
Let It Be	The Yesteryears	A Tribute to #1 Beatles Hits – T...	03:48		<input type="checkbox"/>
Let It Be	∨ Aretha Franklin	This Girl's In Love With You	03:33		<input type="checkbox"/>
Let It Be Sung	∨ Jack Johnson, Matt Costa, Zach Gill,...	If I Had Eyes	04:09		<input type="checkbox"/>
Let It Be	Vox Angeli	Gloria	03:26		<input type="checkbox"/>
Let It Be	∨ Paul McCartney	Good Evening New York City	03:54		<input type="checkbox"/>
Hey Jude	Let It Be	Hey Jude	03:55		<input type="checkbox"/>
Let It Be	Joan Baez	Greatest Hits And Others	03:51		<input type="checkbox"/>

4- Applications du traitement audio pour la description musicale

- Recherche d'un contenu audio dans une base de données
 - autrement que par "artistes", "titres" (Google musical)

The screenshot shows a music search interface. At the top, there is a search bar with the text 'maceo parker' and a 'Rechercher' button. Below the search bar, there are search suggestions: 'maceo_parker (8)' and 'all_the_king_s_men/maceo_parker (3)'. The main content area features a large card for the song 'Got to get you' by Maceo Parker, with a play button and a progress bar. Below this, there is a 'RÉSULTATS (8)' section with a table of search results. The table has columns for 'Titre', 'Artiste', 'Album', and 'Durée'. The first result is 'Got to get you' by Maceo Parker, with a duration of 07:10. To the right of the table, there are filters for 'HUMEURS' (JOYEUX, CALME (1), DYNAMIQUE (5), ROMANTIQUE, TRISTE) and 'GENRES' (POP/ROCK, BLUES (2), ELECTRONIQUE, METAL/PUNK, REGGAE, CLASSIQUE, JAZZ, RAP, SOUL/FUNK (7), LATIN, RNB). Below these are filters for 'INSTRUMENTATIONS' (GUITARE ELECTRIQUE (6), GUITARE ACOUSTIQUE, ELECTRONIQUE, BATTERIE (8), CUIVRES (2), ORCHESTRE A COIRES, PIANO, ACOUSTIQUE) and 'ENREGISTREMENTS' (STUDIO (1), LIVE (7)). At the bottom right, there is a 'MES PLAYLISTS' section with a '+ nouvelle playlist' button.

MSSE PROJECT

maceo parker tous les champs Rechercher

Recherches proches : maceo_parker (8) all_the_king_s_men/maceo_parker (3)

Got to get you *Dynamique - Soul/Funk - Batterie, Guitare électrique*
Maceo Parker
Life on Planet Groove

ajouter à une playlist
chercher des musiques similaires

4:53 - 7:10 écouter le résumé | écouter l'intégral afficher les passages chantés

RÉSULTATS (8)

enregistrement dans une playlist

Modos de lecture : **automatique** | manuel

Titre	Artiste	Album	Durée
▶ Got to get you Dynamique - Soul/Funk - Batterie, Guitare électrique - En studio	Maceo Parker	Life on Planet Groove	07:10
▶ Pass the pees Dynamique - Soul/Funk - Guitare électrique, Batterie - En public	Maceo Parker	Life on Planet Groove	11:28
▶ Addictive Love - Soul/Funk - Cuivres, Batterie - En public	Maceo Parker	Life on Planet Groove	09:00
▶ Shake everything you've got Dynamique - - Batterie, Cuivres - En public	Maceo Parker	Life on Planet Groove	16:41
▶ Soul Power #2 Dynamique - Soul/Funk - Guitare électrique, Batterie - En public	Maceo Parker	Life on Planet Groove	14:13
▶ Georgia on my mind - Soul/Funk, Blues - Guitare électrique, Batterie - En public	Maceo Parker	Life on Planet Groove	07:25
▶ I got you (I Feel Good) Dynamique - Soul/Funk, Blues - Guitare électrique, Batterie - En public	Maceo Parker	Life on Planet Groove	03:47
▶ Children's World Calme - Soul/Funk - Batterie, Guitare électrique - En public	Maceo Parker	Life on Planet Groove	06:23

<< 1/1 >>

HUMEURS
JOYEUX CALME (1) DYNAMIQUE (5)
ROMANTIQUE TRISTE

GENRES
POP/ROCK BLUES (2) ELECTRONIQUE
METAL/PUNK REGGAE CLASSIQUE JAZZ
RAP SOUL/FUNK (7) LATIN RNB

INSTRUMENTATIONS
GUITARE ELECTRIQUE (6) GUITARE ACOUSTIQUE
ELECTRONIQUE BATTERIE (8) CUIVRES (2)
ORCHESTRE A COIRES PIANO ACOUSTIQUE

ENREGISTREMENTS
STUDIO (1) LIVE (7)

MES PLAYLISTS
1 - ismir (2)

+ nouvelle playlist

4- Applications du traitement audio pour la description musicale

Identification audio

Identification audio

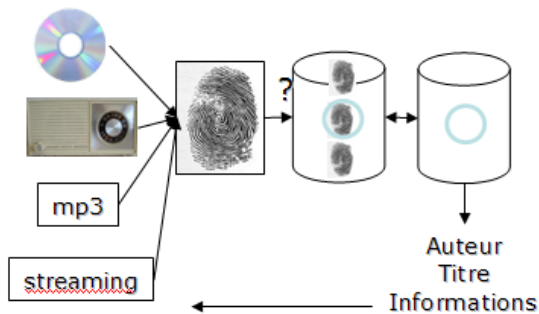
- Objectif :
 - Reconnaître un morceau diffusées sur radio, télé, Internet, bar, discothèque, ...
 - Identifier l'enregistrement (ISRC), pas l'oeuvre (ISWC)

4- Applications du traitement audio pour la description musicale

Identification audio

Méthode du Fingerprint

- Shazam, Midomi, Philips, ...
- Codage :
 - prise d'empreinte du signal, stockage dans une base de données
- Décodage :
 - pour un nouveau signal, prise d'empreinte, comparaison avec les empreintes de la base de données
- Challenge :
 - déterminer un ensemble réduit de descripteurs audio extraits du signal sonore permettant d'identifier de manière unique un extrait musical



4- Applications du traitement audio pour la description musicale

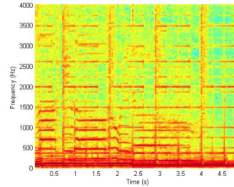
Identification audio

Algorithme de Fingerprint de Shazam

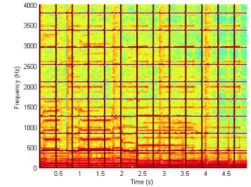
[A. L.-C. Wang. An industrial strength audio search algorithm. In Proc. of ISMIR, 2003.]

[S. Fenet. Empreintes Audio et Stratégies d'Indexation Associées pour l'Identification Audio à Grande Echelle. PhD thesis, Télécom Paris-Tech, 2013.]

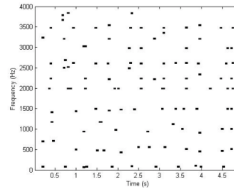
- Extraction de points saillants dans le plan temps/fréquence
 - Calcul du spectrogramme
 - fenêtre de Hamming, $L=64$ ms,
 $S=32$ ms
 - Dans chaque pavé du spectrogramme ($\Delta t=0.4$ s, Δf) :
 - détection du maximum \rightarrow valeur = 1
 - = "constellation points"



(a)



(b)



(c)

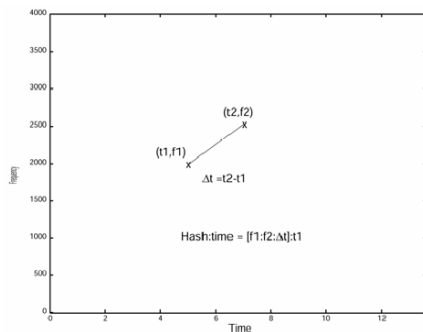
source : Sébastien Fenet

4- Applications du traitement audio pour la description musicale

Identification audio

A) Partie stockage de signature

- Représentation des "constellation points" :
 - chaque point est pris comme un "anchor point" ayant une "target zone"
 - $[f_1, f_2, t_2 - t_1]$
 - + le temps de l'anchor t_1
- Méthode de "pruning" des points
 - on ne garde que les pairs de points pour lesquels
 - $f_2 - f_1 < \Delta f_{\max} = 350\text{Hz}$
 - $t_2 - t_1 < \Delta T_{\max} = 3\text{s}$
- Stockage des triplets
 - $[f_1, f_2, t_2 - t_1]$ stocké sur 32 bits

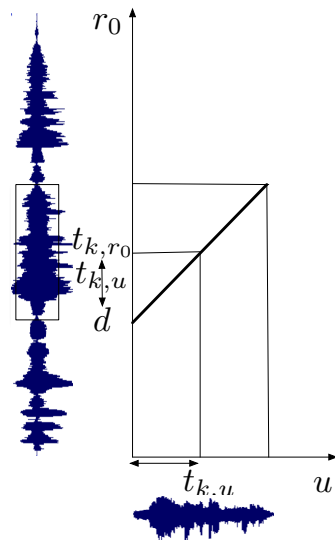


4- Applications du traitement audio pour la description musicale

Identification audio

B) Partie matching de signature

- si le signal inconnu u est un extrait de r_0 démarrant au temps d
 - alors toutes les clefs apparaissant dans u doivent être trouvées dans r_0
 - une clef k de u au temps $t_{k,u}$ doit être trouvé dans r_0 au temps $t_{k,r_0} = d + t_{k,u}$
 - si on étudie l'ensemble des valeurs $\{t_{k,r_0} - t_{k,u}\}$ pour toutes les clefs k de u , on doit avoir un maximum d'accumulation en d

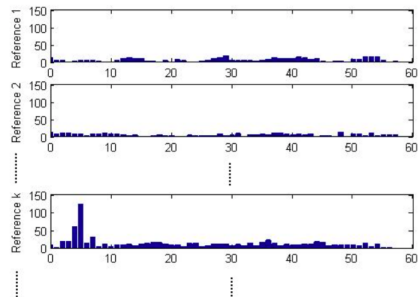


4- Applications du traitement audio pour la description musicale

Identification audio

B) Partie matching de signature

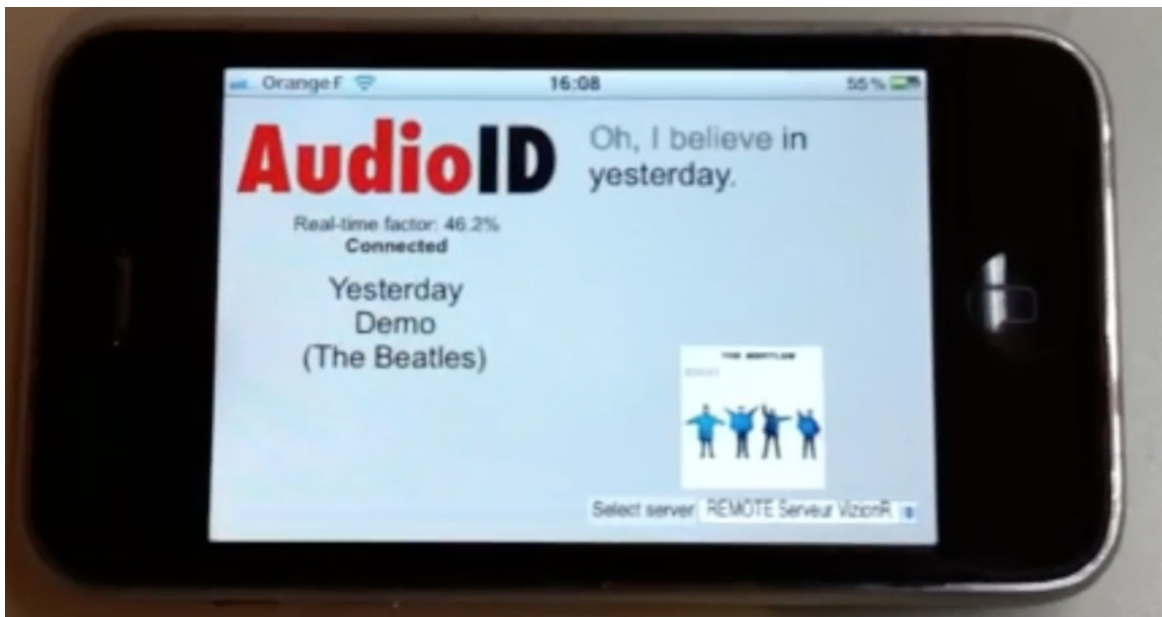
- Méthode :
 - pour toutes les clefs k de u , pour chaque référence r_i , on stocke toutes les valeurs $\{t_{k,r_i} - t_{k,u}\}$ dans un histogramme
 - l'histogramme avec le plus grand maximum donne la référence du signal inconnu
 - la position du maximum dans cet histogramme donne le point de démarrage d dans le signal de référence



source : Sebastien Fenet

4- Applications du traitement audio pour la description musicale

Identification audio



5- Descripteurs audio

Les descripteurs audio

[G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.]

- Valeurs numériques extraites du signal audio dont le but est de représenter une propriété particulière de son contenu
 - Tout est dans la forme d'onde, dans la TFCT, difficile à lire, trop grande dimension
- Contrainte :
 - on veut le même nombre de dimensions pour toutes les données
- Extraction ?
 - Algorithme d'estimation
 - Opérateurs mathématique

5- Descripteurs audio

Introduction

Les descripteurs audio

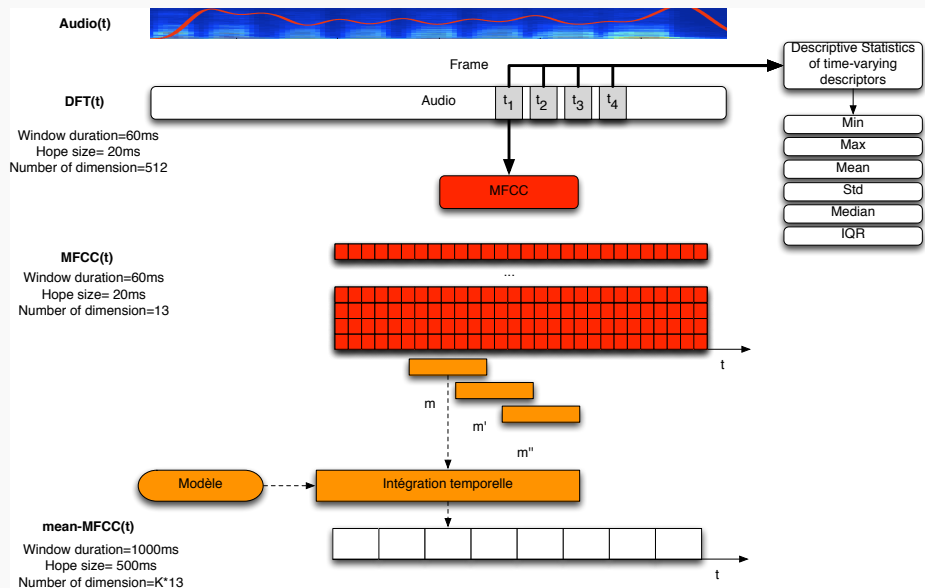
[G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.]

- Différentes **formes** :
 - **scalaire** : Centroïde spectral, étendue spectrale, fréquence fondamentale, spectral roll-off, spectral flux, zero-crossing rate, RMS, ...
 - **vecteur** : Mel Frequency Cepstral Coefficients, coefficients LPC, coefficients PLP ...
- Différentes **temporalité** :
 - représente une **trame** du signal audio → descripteurs "instantanés"
 - représente le résumé du contenu d'un **ensemble local de trame** → texture windows
 - représente **globalement** le signal audio
- Mise en évidence de différents **contenus** (, harmonique, bruité, ...)
 - contenu **timbral** : Mel Frequency Cepstral Coefficients, coefficients LPC, coefficients PLP ...
 - contenu **harmonique** : Pitch Class Profiles/ Chroma ...
 - contenu **bruité** : Spectral Flatness Measure
 - contenu **rythmique** : ...

5- Descripteurs audio

Introduction

Les descripteurs audio

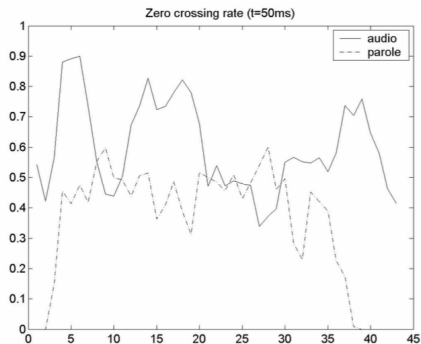
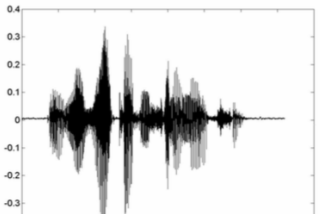
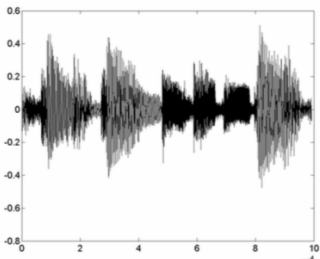


5- Descripteurs audio

Taux de passage par zéro

Taux de passage par zéro / zero-crossing rate (zcr)

- Mesure le nombre de fois que la forme d'onde croise l'axe zéro
 - $zcr = 0.5 \sum_{n=1}^N |sign(x(n)) - sign(x(n-1))|$
- Utilisation :
 - permet de distinguer les signaux bruités → zcr élevé
 - permet de distinguer les signaux harmoniques → zcr bas



source : Gaël Richard

5- Descripteurs audio

Mel Frequency Cepstral Coefficients (MFCCs)

Objectif

- décrire la forme du spectre (du timbre) d'un signal à l'aide d'un nombre réduit de coefficients

Cepstre complexe

- Cepstre complexe** $c(\tau)$:

$$\begin{aligned}c(\tau) &= TF^{-1}[\log(X(\omega))] \\ &= \frac{1}{2\pi} \int_{\omega} \log(X(\omega)) e^{j\omega\tau} d\omega\end{aligned}\tag{12}$$

- τ est appelé "céfrence"
- $x(t) \xrightarrow{TF} X(\omega) \xrightarrow{\log} \log(X(\omega)) \xrightarrow{TF^{-1}} c(\tau)$

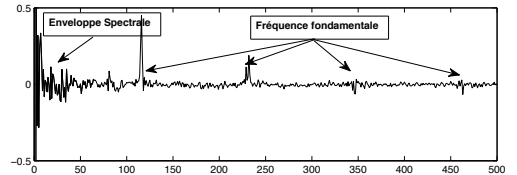
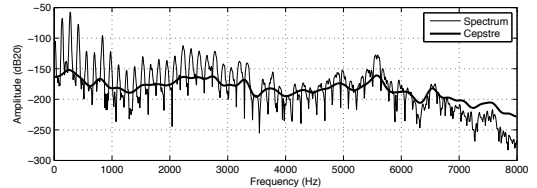
5- Descripteurs audio

Mel Frequency Cepstral Coefficients (MFCCs)

Cepstre complexe

- Modèle source/ filtre :
 - Source : signal périodique
 - Filtre : résonant/ anti-résonant

$$\begin{aligned}
 x(t) &= e(t) \circledast g(t) \\
 \xrightarrow{TF} X(\omega) &= E(\omega) \cdot G(\omega)
 \end{aligned}
 \tag{13}$$



$$\begin{aligned}
 \xrightarrow{\log} \log(X(\omega)) &= \underbrace{\log(E(\omega))}_{\text{variation rapide à travers } \omega} + \underbrace{\log(G(\omega))}_{\text{variation lente à travers } \omega} \\
 \xrightarrow{TF^{-1}} TF^{-1} [\log(X(\omega))] &= \underbrace{TF^{-1} [\log(E(\omega))]}_{\text{énergie aux céfrenes } \tau \gg} + \underbrace{TF^{-1} [\log(G(\omega))]}_{\text{énergie aux céfrenes } \tau \ll}
 \end{aligned}
 \tag{14}$$

5- Descripteurs audio

Mel Frequency Cepstral Coefficients (MFCCs)

Cepstre réel

- **Cepstre réel** :
 - Cepstre calculé sur la partie réelle du log-spectrum

$$X(\omega) = A(\omega) \cdot e^{j\phi(\omega)}$$

$$\log(X(\omega)) = \log(A(\omega)) + j\phi(\omega) \quad (15)$$

$$\Re(\log(X(\omega))) = \log(A(\omega))$$

$$\begin{aligned} \text{cepstre réel} &= TF^{-1} [\Re(\log(X(\omega)))] \\ &= TF^{-1} [\log(A(\omega))] \end{aligned} \quad (16)$$

$$c(\tau) = \frac{1}{2\pi} \int_{\omega} \log(A(\omega)) e^{j\omega\tau} d\omega$$

- Le spectre d'amplitude étant réel et symétrique
 - sa TF se réduit à sa partie réelle
 - donc à la projection de $\log(A(\omega))$ sur un ensemble de cosinus \rightarrow DCT

5- Descripteurs audio

Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs)

- **Mel Frequency Cepstral Coefficient :**
 - Cepstre réel calculé sur un spectre d'énergie exprimé en convertissant l'énergie $|X(\omega)|^2$ en échelle perceptive (échelle de Mel)
- Pourquoi ?
 - La transformée de Fourier :
 - décomposition sur une série de sinusoides linéairement espacées (10Hz, 20Hz, 30Hz, ... Hz)
 - L'oreille :
 - décomposition sur une série de filtres de fréquences logarithmiquement espacé (10, 20, 40, 80, ... Hz).
 - meilleure résolution en basses fréquences que en hautes fréquences.
 - résonances de l'enveloppe spectrale sont plus rapprochées en basse fréquence.
 - MFCCs permet une représentation plus compacte que le cepstre réel
- Comment ?
 - On utilise des échelles dites perceptives : échelles de Mel, de Bark, filtres ERB, Gamma tone
- Utilisation ?
 - Les coefficients les plus utilisés dans le monde de la reconnaissance audio : parole, musique, sons environnementaux, ...

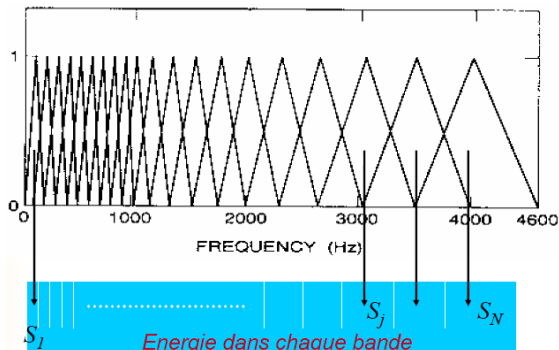
5- Descripteurs audio Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs)

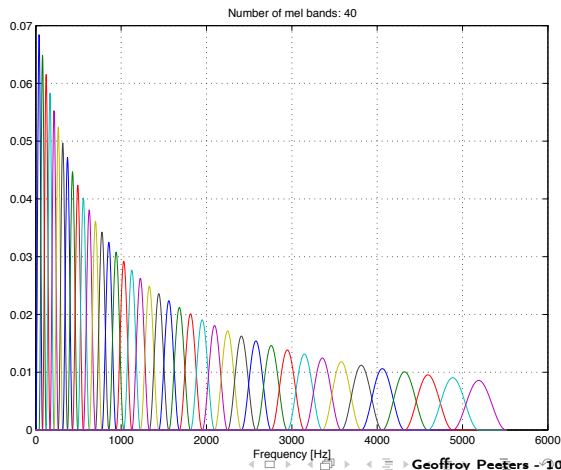
- Echelle de Mel :

$$M = f \text{ pour } f < 1000\text{Hz}$$

$$M = f_c \left(1 + \log_{10} \left(\frac{f}{f_c} \right) \right) \text{ pour } f \geq 1000\text{Hz} \quad (17)$$



source : Gaël Richard

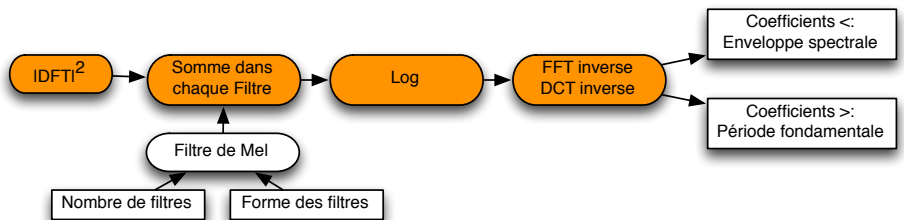


5- Descripteurs audio

Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs)

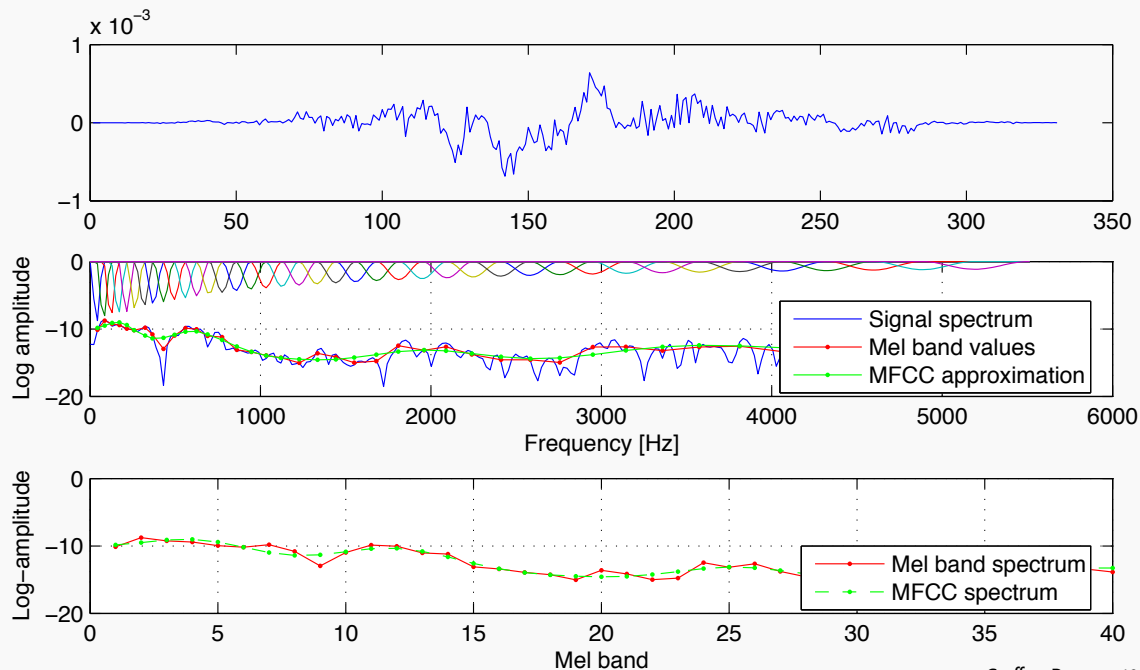
- Calcul du spectre de puissance : $|X(\omega)|^2$
- Calcul des filtres de Mel : $H_b(\omega)$ avec $b \in [1, B]$
 - choix du nombre de filtres B : 40
 - choix de la forme des filtres : triangulaire, hanning, tanh, ...
- Conversion du spectre de puissance en bandes de Mel : $S(b) = \sum_{\omega} |X(\omega)|^2 \cdot H_b(\omega)$
- Passage en échelle logarithmique : $\log(S(b))$
- Calcul de la IFFT (ou de la IDCT) :
- Sélection des coefficients de la IDCT proches de zéro (jusqu'à 13)
 - les coefficients proches de zéro représentent la décomposition du spectre en échelle de Mel sur un ensemble de cosinus à variation lente



5- Descripteurs audio

Mel Frequency Cepstral Coefficients (MFCCs)

Exemple de calcul de MFCCs

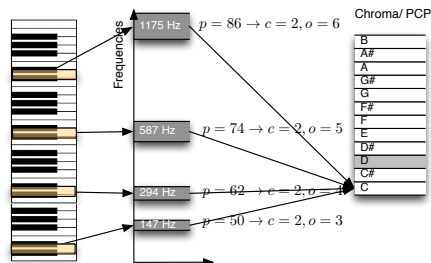
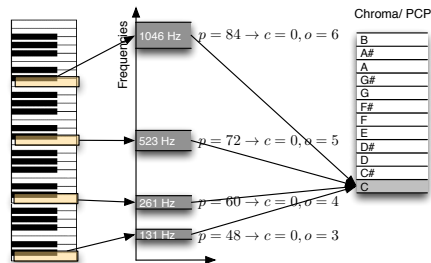


5- Descripteurs audio

Chroma - Pitch Class Profile (PCP)

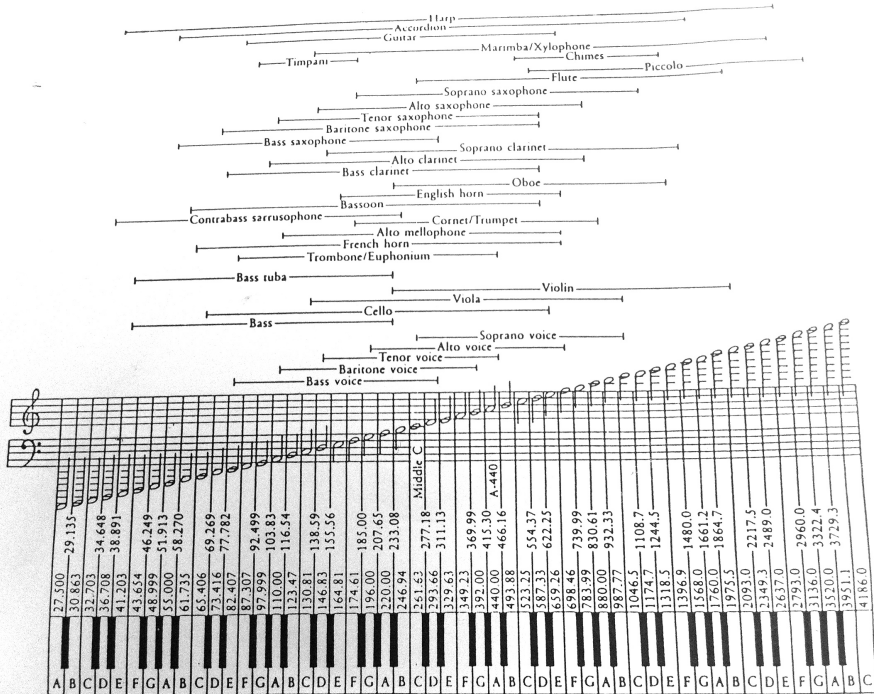
Définition des Chroma - Pitch Class Profile (PCP)

- **Objectif :**
 - le spectre à l'instant n : $X(k, n)$
 - représenter son contenu harmonique sous forme d'un vecteur : $C(c, n)$ $c \in [0, 12[$
- Utilisations :
 - reconnaissance de tonalité,
 - reconnaissance de suite d'accords,
 - détection de "cover versions"
- Shepard-1964 :
 - représenter la hauteur d'une note p comme une structure bi-dimensionnelles :
 - $p = c + o \cdot 12$
 - le chroma c (classe de hauteur).
 - la hauteur tonale o (numéro d'octave),



5- Descripteurs audio

Chroma - Pitch Class Profile (PCP)

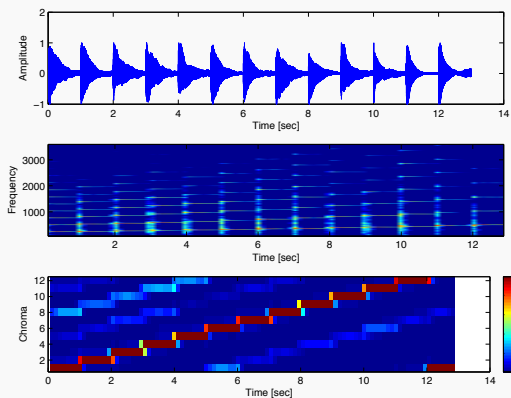


5- Descripteurs audio

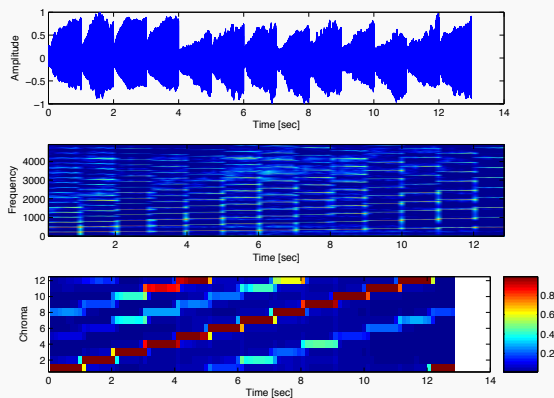
Chroma - Pitch Class Profile (PCP)

Limitations des Chromas - Pitch Class Profile (PCP)

Exemple piano



Exemple violon

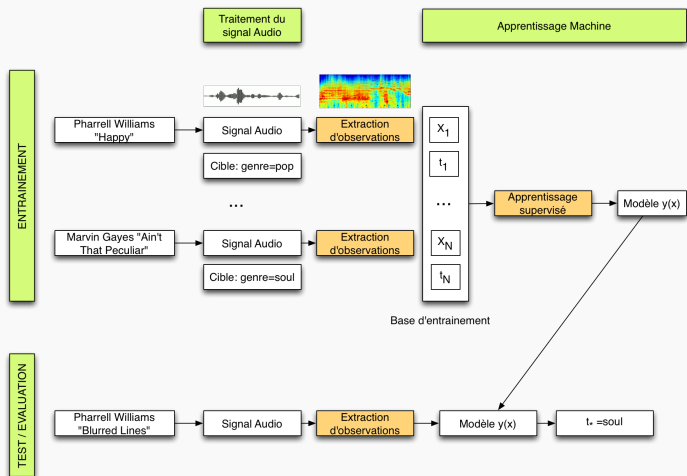


Classification Audio

6- Classification Audio

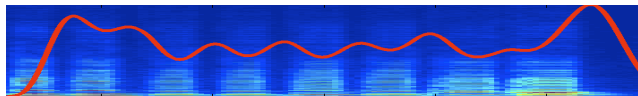
Classification Audio

- Utilisation des descripteurs audio en entrée d'un algorithme d'**apprentissage supervisé**
- Exemples d'utilisation :
 - auto-tagging en genre, en mood (humeur), en instrumentation
 - segmentation d'un flux tempos en catégories paroles/musiques, musique instrumentale/chantée

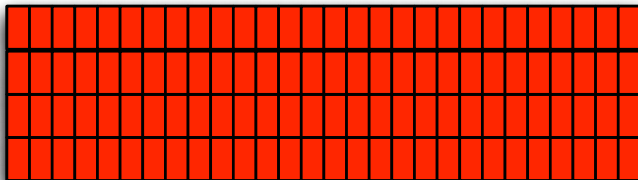


Extraction des descripteurs sur un fichier audio

Morceau 1



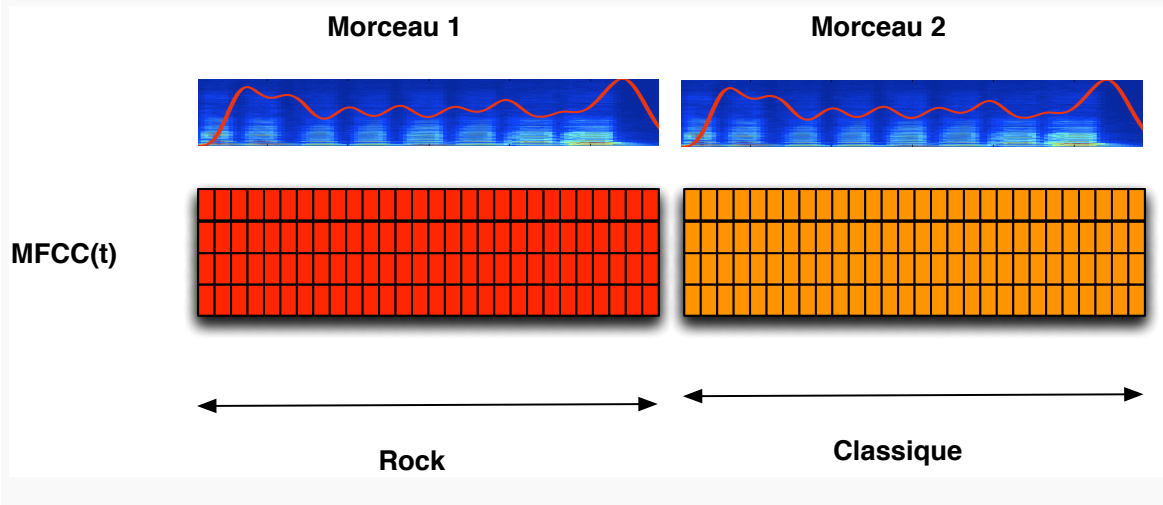
MFCC(t)



6- Classification Audio

Extraction des descripteurs

Extraction des descripteurs sur plusieurs fichier audio + assignation des labels de classes aux données

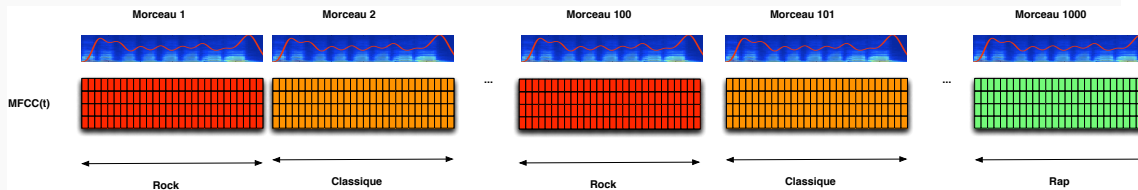


6- Classification Audio

Extraction des descripteurs

Extraction des descripteurs sur une collection de fichiers + assignation des labels de classes aux données

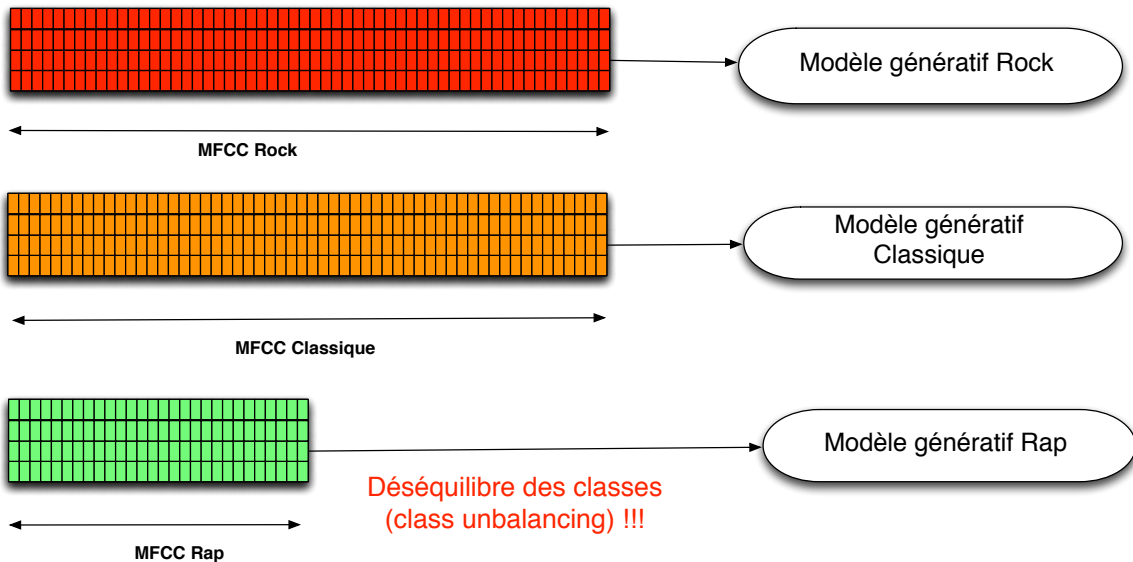
- La collection peut contenir plusieurs millions de fichiers audio
- Le nombre de labels de classes peut être très important (99 genre musicaux)



6- Classification Audio

Apprentissage

Apprentissage



6- Classification Audio

Apprentissage

Algorithmes d'apprentissage supervisé

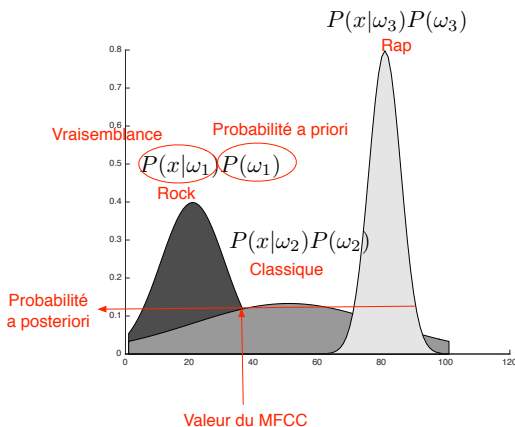
- Modèles génératifs : Gaussien, GMM
- Modèles discriminants : LDA, SVM
- Approche par exemplification : KNN

Modèle génératif gaussien

- On modélise chaque classe ω_i par une densité de probabilité gaussienne
 - $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\}$
- On applique la règle de décision Bayésienne

$$p(\omega = \omega_i | \mathbf{x}) = p(\omega = \omega_j) \cdot \frac{p(\mathbf{x} | \omega = \omega_j)}{p(\mathbf{x})}$$

$$\text{posterior} = \text{prior} \cdot \frac{\text{vraisemblance}}{\text{evidence}}$$



6- Classification Audio

Apprentissage

Estimation du label de classe d'un fichier audio inconnu

