

ENSEA AuPaMu

Traitement du signal audio musical, descripteurs et estimation

07/04/2016

Geoffroy.Peeters@ircam.fr

UMR SMTS (IRCAM CNRS UPMC)

1. Examen
 - 1.1 Option Audio-Parole-Musique 2nd semestre
2. Introduction
 - 2.1 Applications des techniques d'indexation audio pour la musique
3. Théorie : Traitement du signal
 - 3.1 Transformée de Fourier (temps et fréquences continus)
 - 3.2 Transformée de Fourier (temps et fréquences discrets)
 - 3.3 Transformée de Fourier (à Court Terme) : TFCT
 - 3.4 Identification audio
 - 3.5 Estimation du tempo
 - 3.6 Estimation de la hauteur d'une note
4. Modèles de signaux
 - 4.1 Modèle source/ filtre
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. Applications
7. Classification Audio
 - 7.1 Extraction des descripteurs
 - 7.2 Apprentissage
 - 7.3 Estimation d'accords
 - 7.4 Détection de cover-version

1- Examen

1.1- Option Audio-Parole-Musique 2nd semestre

Examen Option Audio-Parole-Musique 2nd semestre

- Quand : lundi 2 mai 2016 - 8h à 9h30 (1h30)
- Salle : ???
- sans document

2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique



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

 [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

[similarly random recordings »](#)

[Google Labs](#) - [Discuss](#) - [Terms of use](#) - [About Google Audio](#) - [Submit your recording](#)

source : Gaël Richard

2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

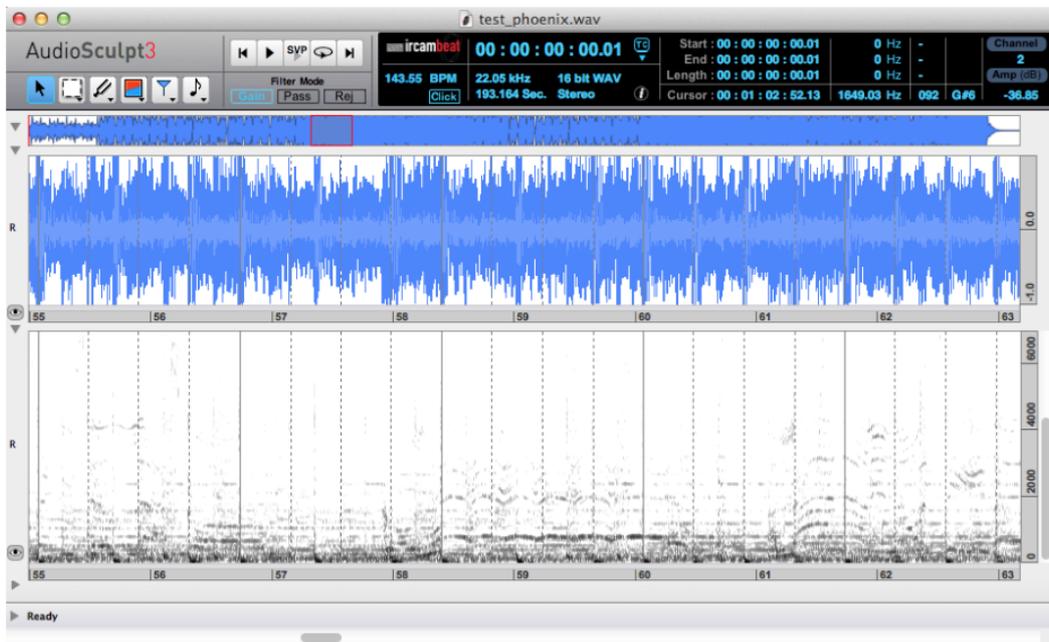
- Identification audio
 - recherche de doublons, gestion de copyright, attacher des méta données à une instance d'un morceau



2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

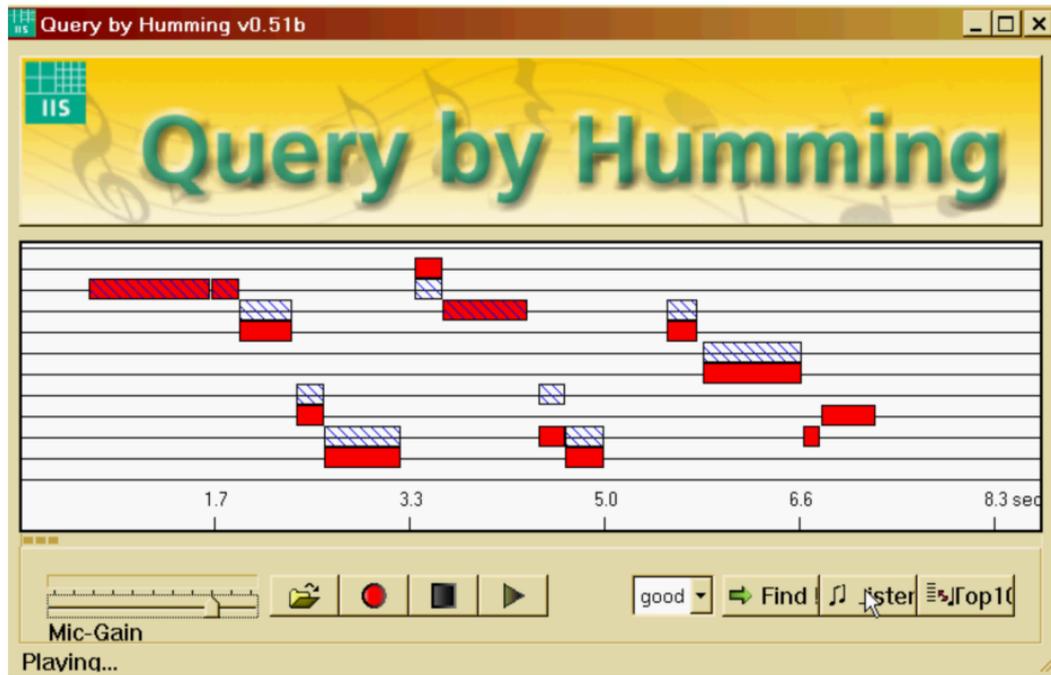
- Estimation du tempo, de la position des temps/ premier-temps
 - DJing, manipulation du contenu (add swing ...)



2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

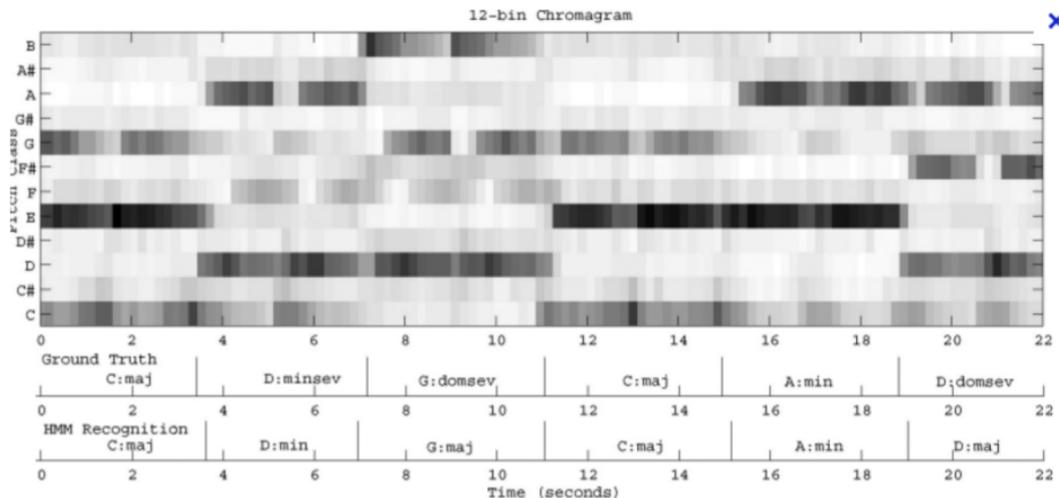
- Nouveaux modes de recherche :
 - par chantonement/ sifflement



2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

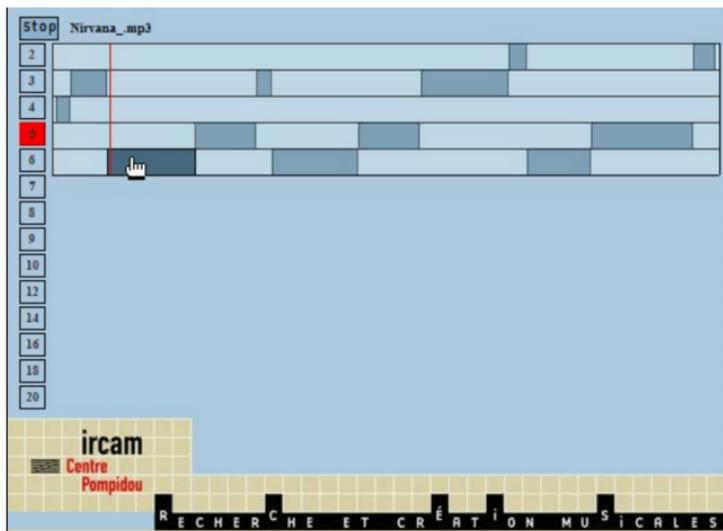
- Estimation des accords
 - obtenir des guitar-tab automatiquement



2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

- 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



2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

- 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 Wlth 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"/>

2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique

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

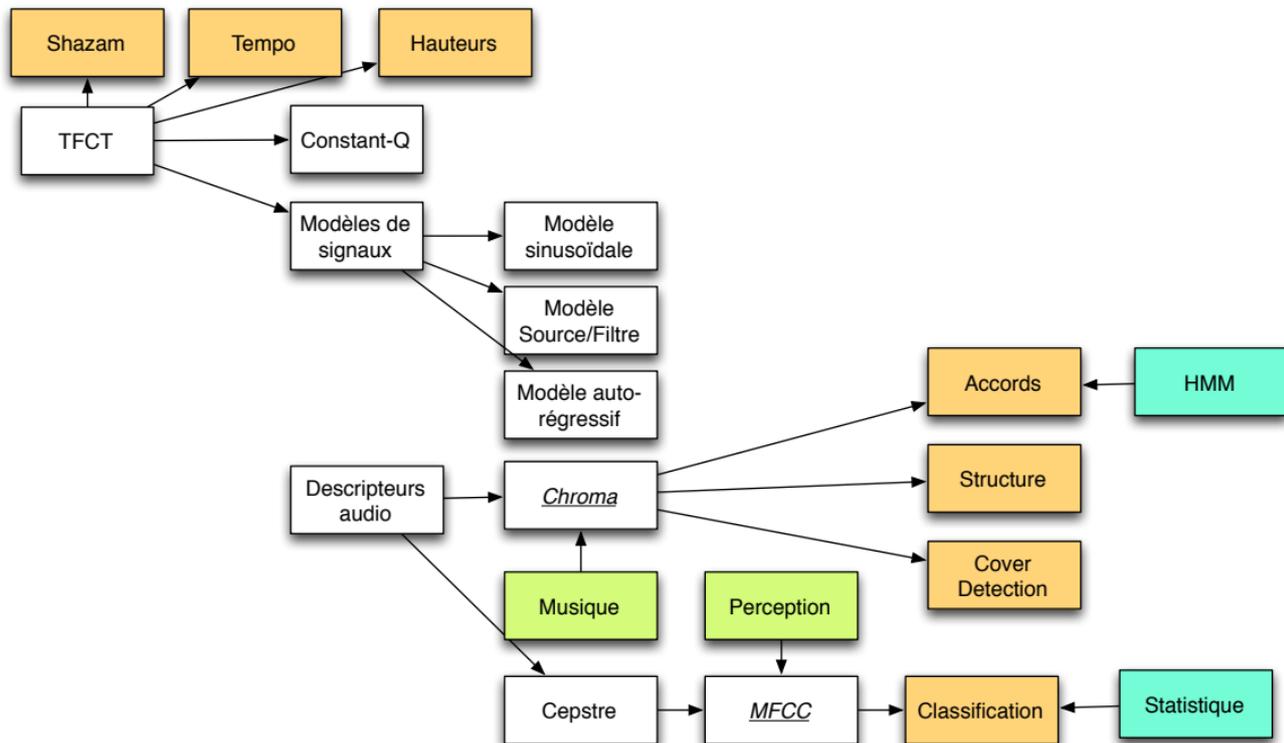
Search results for "maceo parker" on MSSE PROJECT. The interface shows a search bar, a player for the selected track "Got to get you" by Maceo Parker, and a table of search results.

Titre	Artiste	Album	Durée
Got to get you	Maceo Parker	Life on Planet Groove	07:10
Pass the pass	Maceo Parker	Life on Planet Groove	11:28
Addictive Love	Maceo Parker	Life on Planet Groove	09:00
Shake everything you've got	Maceo Parker	Life on Planet Groove	16:41
Soul Power 92	Maceo Parker	Life on Planet Groove	14:13
Georgia on my mind	Maceo Parker	Life on Planet Groove	07:25
I got you (I Feel Good)	Maceo Parker	Life on Planet Groove	03:47
Children's World	Maceo Parker	Life on Planet Groove	06:23

Additional interface elements include a search bar with "maceo parker", a "Rechercher" button, a "Recherches proches" section, a player for "Got to get you" (4:53 / 7:10), and a sidebar with "HUMEURS" (JOYEUX, CALME, DYNAMIQUE), "GENRES" (POP/ROCK, BLUES, ELECTRONIQUE, METAL/PUNK, REGGAE, CLASSIQUE, JAZZ, RAP, SOUL/FUNK, LATIN, RNB), "INSTRUMENTATIONS" (GUITARE ELECTRIQUE, GUITARE ACOUSTIQUE, ELECTRONIQUE, BATTERIE, CUIVRES, ORCHESTRE A CORDES, PIANO, ACOUSTIQUE), "ENREGISTREMENTS" (STUDIO, LIVE), and "MES PLAYLISTS" (1 - ismir).

2- Introduction

2.1- Applications des techniques d'indexation audio pour la musique



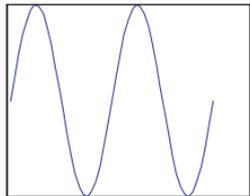
3- Théorie : Traitement du signal

3.1- Transformée de Fourier (temps et fréquences continus)

Transformée de Fourier (temps et fréquences continus)

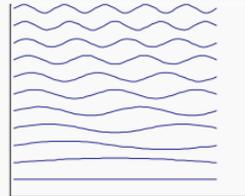
$$X(\omega) = \int_{t=-\text{inf}}^{+\text{inf}} x(t)e^{-j\omega t} dt \quad X(f) = \int_{t=-\text{inf}}^{+\text{inf}} \exp(-j2\pi ft) dt \quad (1)$$

- Variables :
 - t est le **temps**
 - $\omega = 2\pi f$ les **fréquences continues** exprimées en radian,
 - $\exp(j2\pi ft) = \cos(2\pi ft) + j \cdot \sin(2\pi ft)$.
- Pourquoi la Transformée de Fourier ?
 - Difficile d'extraire des observations directement à partir de la forme d'onde $x(t)$
 - Reproduire la décomposition en fréquences de l'oreille humaine

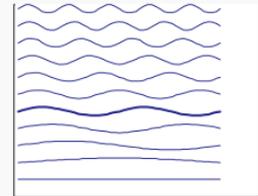


x(t)

x



sin(2 pi f t)



3- Théorie : Traitement du signal

3.1- Transformée de Fourier (temps et fréquences continus)

Propriété de la Transformée de Fourier (temps et fréquences continus)

Propriétés	$x(t)$	$X(f)$
Similitude	$x(at)$	$\frac{1}{ a } X\left(\frac{f}{ a }\right)$
Linéarité	$ax(t) + by(t)$	$aX(f) + bY(f)$
Translation	$x(t - t_0)$	$X(f) \exp(-j2\pi f t_0)$
Modulation	$x(t) \exp(j2\pi f_0 t)$	$X(f - f_0)$
Convolution	$x(t) \otimes y(t)$	$X(f)Y(f)$
Produit	$x(t)y(t)$	$X(f) \otimes Y(f)$
Parité	réelle paire réelle impaire imaginaire paire imaginaire impaire complexe paire complexe impaire réelle $x^*(t)$	réelle paire imaginaire paire imaginaire paire réelle impaire complexe paire complexe impaire $X(f) = X^*(-f)$ $\Re(X(f))$ est paire $\Im(X(f))$ est impaire $X^*(f)$

3- Théorie : Traitement du signal

3.2- Transformée de Fourier (temps et fréquences discrets)

$$X(k) = \sum_{m=0}^{N-1} x(m)e^{-j2\pi \frac{k}{N}m} \quad \forall k \in [0, N]$$

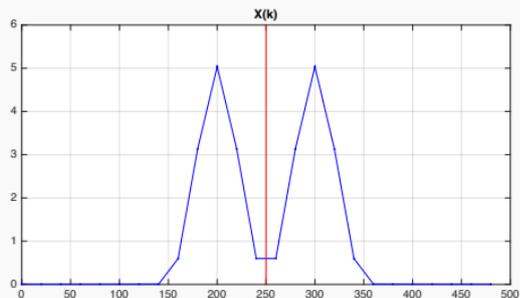
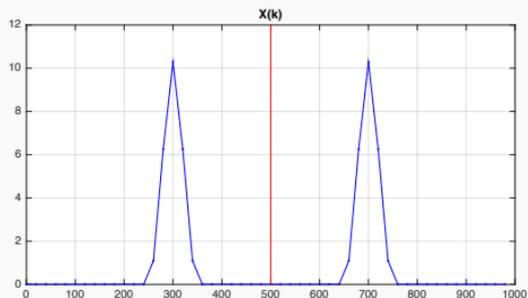
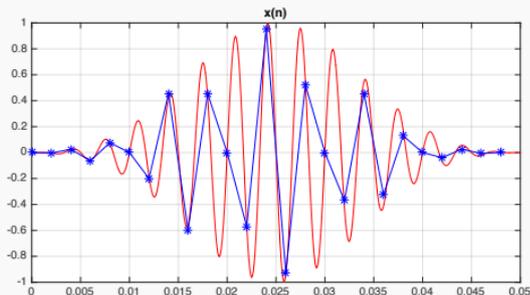
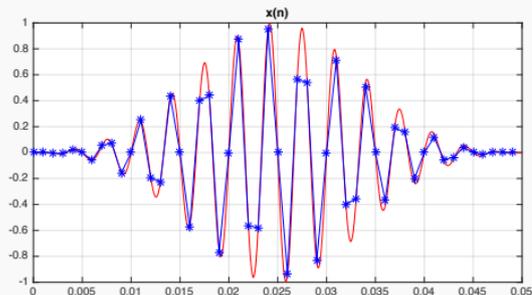
- Variables :
 - n le numéro d'**échantillon**
 - k les **fréquences discrètes**
- Fréquence d'échantillonnage (sampling rate) sr
 - sr définit à quelle fréquence le signal temporel va être échantillonné
 - Exemple :
 - Compact Disc $sr = 44100$ Hz
 - La distance temporelle entre deux échantillons (le pas d'échantillonnage) est de $\Delta t = \frac{1}{44100} = 0.000023$ s.
- sr doit être $>$ à deux fois la f_{\max} présente dans le signal
 - Sinon : repliement spectral
 - exemple : captation d'une roue d'une voiture accélérant dans les films
 - **Fréquence de Nyquist** : $f_{Nyquist} = \frac{sr}{2} > f_{\max}$

3- Théorie : Traitement du signal

3.2- Transformée de Fourier (temps et fréquences discrets)

$$f_{\max} = 300, sr = 1000$$

$$f_{\max} = 300, sr = 500$$



3- Théorie : Traitement du signal

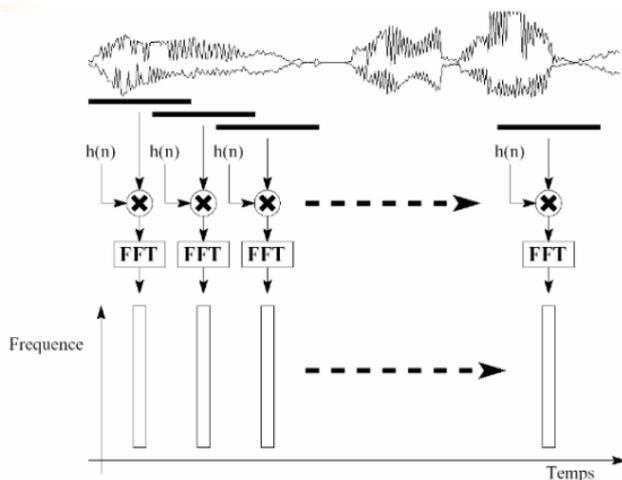
3.3- Transformée de Fourier (à Court Terme) : TFCT

$$X(k, n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \quad \forall k \in [0, N]$$

- Application de la TFD à une portion du signal centrée autour de l'échantillon n

Pourquoi la TFCT ?

- Signal audio = non-stationnaire
 - ses propriétés varient au cours du temps
- **Stationnaires "localement"** (en temps)
 - sur une durée de $\pm 40\text{ms}$
- TFCT = suite d'analyses de Fourier sur des durées de $\pm 40\text{ms}$
 - = analyse à Court Terme ("trames/frames" en vidéo)



source : Jean Laroche

3- Théorie : Traitement du signal

3.3- Transformée de Fourier (à Court Terme) : TFCT

$$X(k, n) = \sum_{m=0}^{N-1} x(m)w(n - m)e^{-j2\pi\frac{k}{N}m} \quad \forall k \in [0, N]$$

Fenêtre de pondération $w(t)$

- $x(t) \cdot w(t) \Leftrightarrow X(f) \circledast W(f)$
 - $w(t)$ est appelé "**fenêtre de pondération**"
 - $w(t)$ différents **types** de fenêtre
 - $w(t)$ définie sur un horizon fini (**longueur temporelle**) $[0, L]$.
 - Choix du type et de la longueur détermine les caractéristiques spectrales
 - Largeur de bande fréquentielle (à $-6dB_{20}$) : $Bw = \frac{Cw}{L}$
 - Hauteur des lobes secondaires

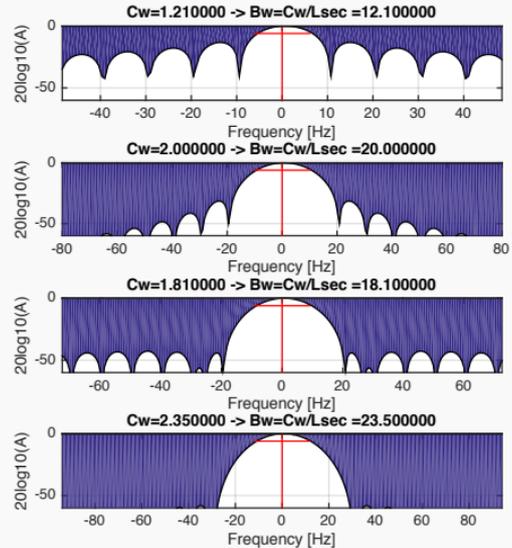
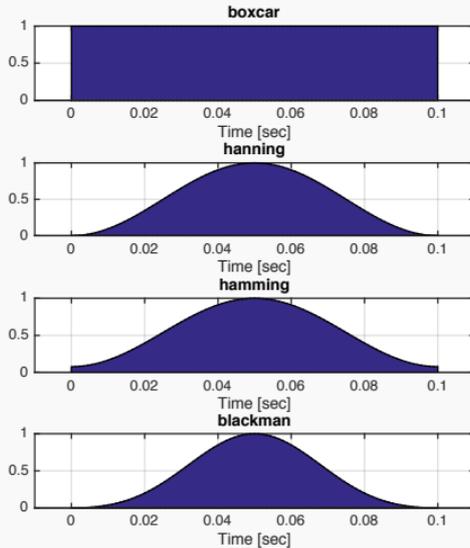
Choix du **type** de la fonction :

- rectangulaire
 - $w(n) = 1$
 - $Bw = 1.21$
- hanning
 - $w(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1}))$
 - $Bw = 2$
- hamming
 - $w(n) = 0.54 - 0.46 \cos(\frac{2\pi n}{N-1})$
 - $Bw = 1.81$
- blackman
 - $w(n) = a_0 - a_1 \cos(\frac{2\pi n}{N-1}) + a_2 \cos(\frac{4\pi n}{N-1})$
 - $Bw = 2.35$

3- Théorie : Traitement du signal

3.3- Transformée de Fourier (à Court Terme) : TFCT

Influence du **type** de la fonction



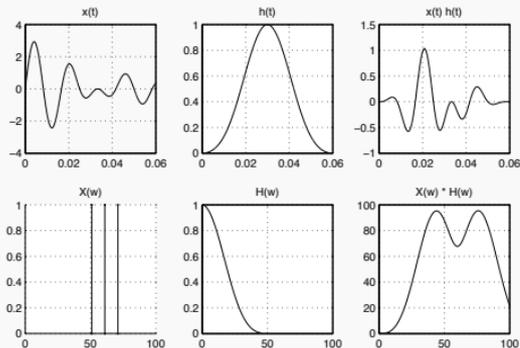
Choix de la **longueur temporelle** L :

- Au plus la fenêtre est courte,
 - au plus on observe précisément les temps.
- Au plus la fenêtre est longue,
 - au plus on observe précisément les fréquences.

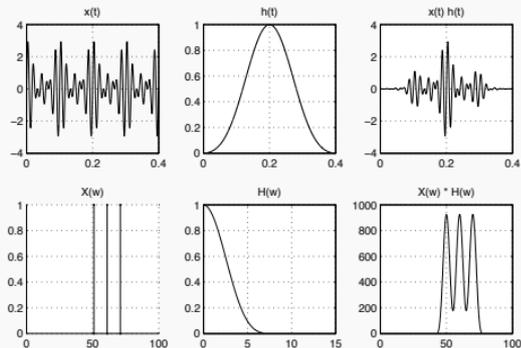
3- Théorie : Traitement du signal

3.3- Transformée de Fourier (à Court Terme) : TFCT

Influence de la **longueur temporelle** L
($L = 0.06\text{s.}$)



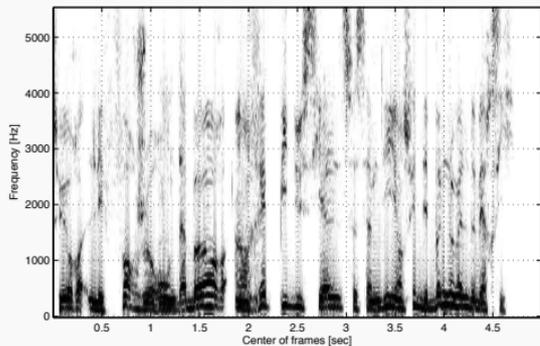
Influence de la **longueur temporelle** L
($L = 0.4\text{s.}$)



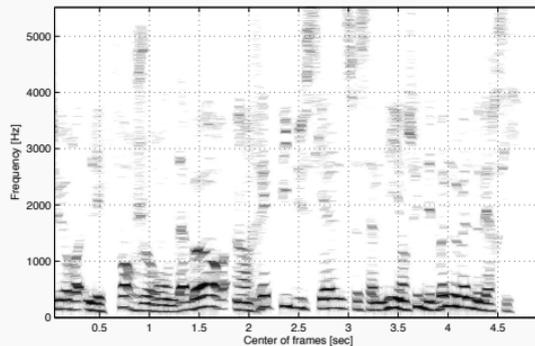
3- Théorie : Traitement du signal

3.3- Transformée de Fourier (à Court Terme) : TFCT

Influence de la **longueur temporelle** L
($L = 0.01s.$)



Influence de la **longueur temporelle** L
($L = 0.1s.$)

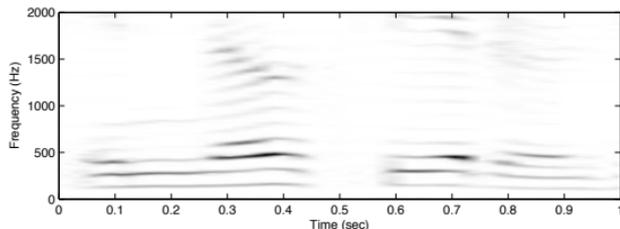
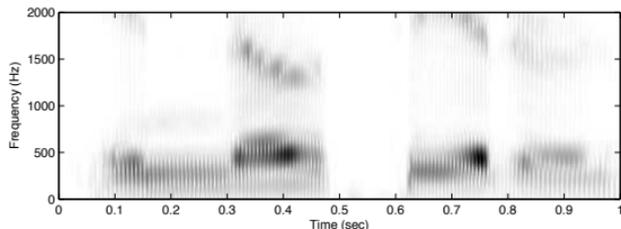
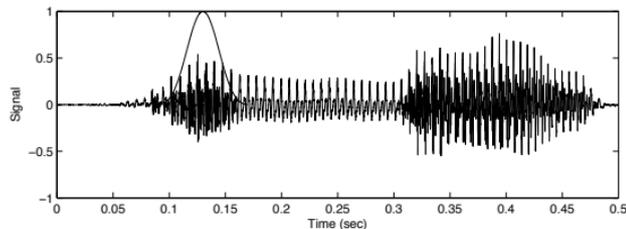
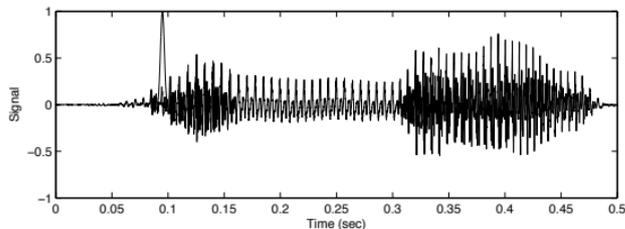


3- Théorie : Traitement du signal

3.3- Transformée de Fourier (à Court Terme) : TFCT

Paradoxe temps/ fréquence

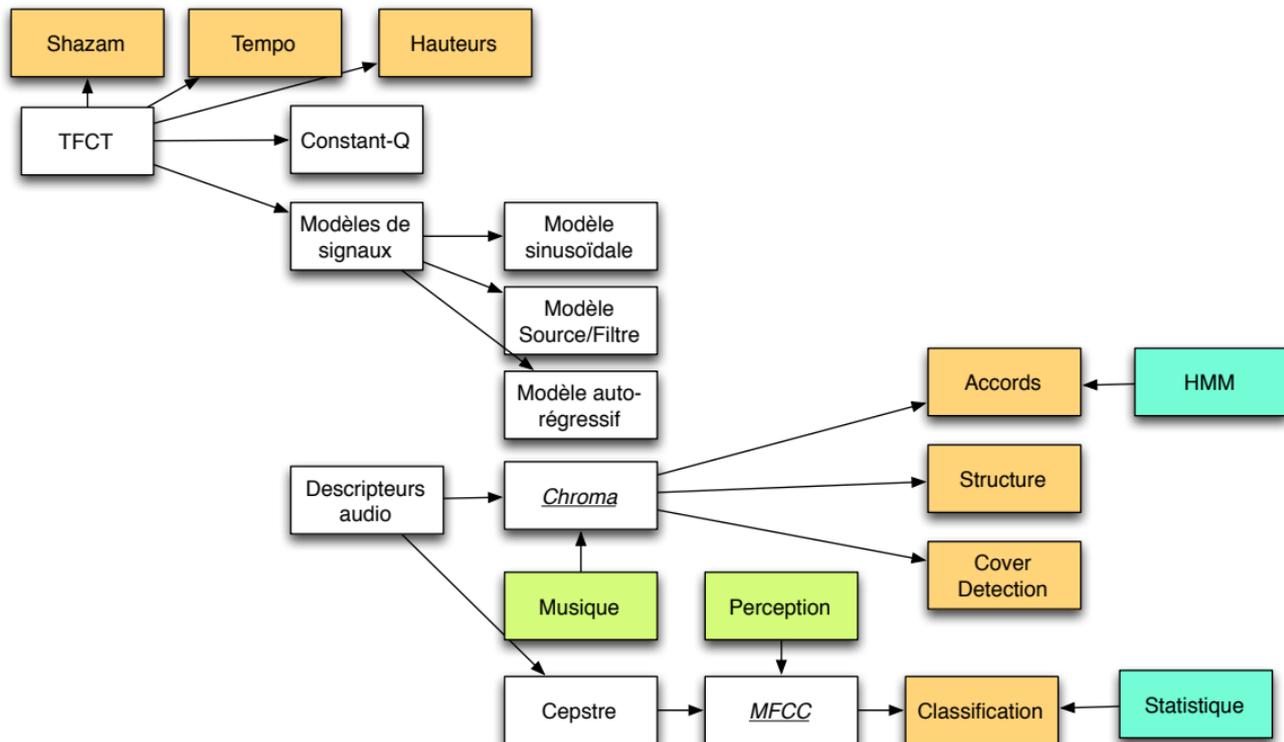
- Pas possible d'avoir simultanément une bonne localisation en temps et en fréquence !



- Comme résoudre ce problème ?
 - Utiliser d'autres transformées que celle de Fourier

3- Théorie : Traitement du signal

3.3- Transformée de Fourier (à Court Terme) : TFCT



Identification audio

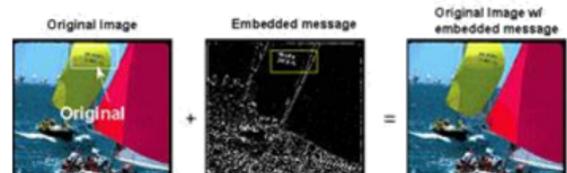
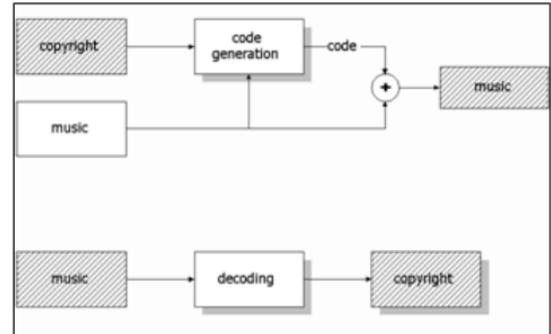
- Objectif :
 - Reconnaître un morceau diffusées sur radio, télé, Internet, bar, discothèque, ...
 - Identifier l'enregistrement (ISRC), pas l'oeuvre (ISWC)

3- Théorie : Traitement du signal

3.4- Identification audio

Méthode du Watermarking

- Codage :
 - introduction d'un code identifiant robuste mais inaudible dans le signal sonore
- Décodage :
 - pour un nouveau signal : extraction du code (si il est présent) et recherche de ce code dans une base de données

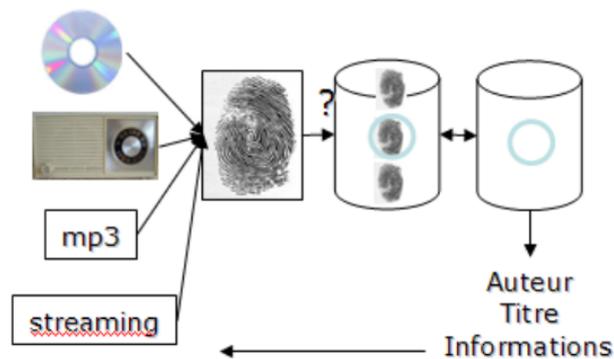


3- Théorie : Traitement du signal

3.4- 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



3- Théorie : Traitement du signal

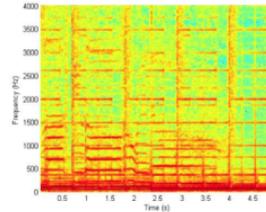
3.4- Identification audio

Algorithme de Fingerprint de Shazam

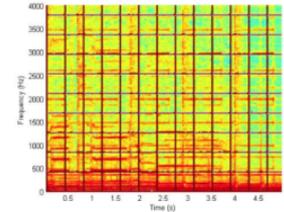
[A. L.-C. Wang. *An industrial strength audio search algorithm*. In *Proc. of ISMIR*, 2003.]

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

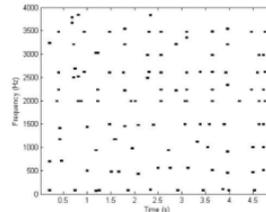
- Extraction de points saillants dans le plan temps/fréquence
 - Calcul du spectrogramme
 - fenêtre de Hamming, $L=64$ ms, $S=32$ ms
 - Dans chaque pavé du spectrogramme ($\Delta t=0.4$ s, Δf) :
 - détection du maximum \rightarrow valeur = 1
 - = "constellation points"



(a)



(b)



(c)

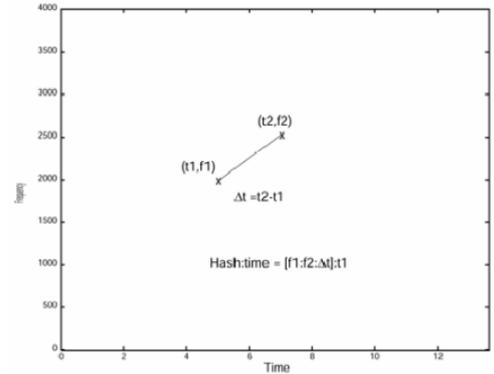
source : Sebastien Fenet

3- Théorie : Traitement du signal

3.4- Identification audio

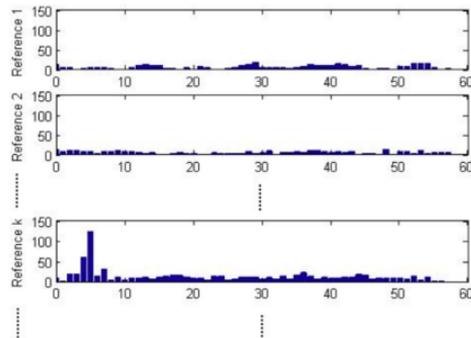
Partie stockage

- 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
 - $f_2 - f_1 < \Delta f_{\max} = 350Hz$
 - $t_2 - t_1 < \Delta T_{\max} = 3s$
- Stockage des triplets
 - $[f_1, f_2, t_2 - t_1]$ stocké sur 32 bits



Partie matching

- 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 $t_{k,r_0} = t_{k,u} + d$
 - 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
- 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 : Sébastien Fenet

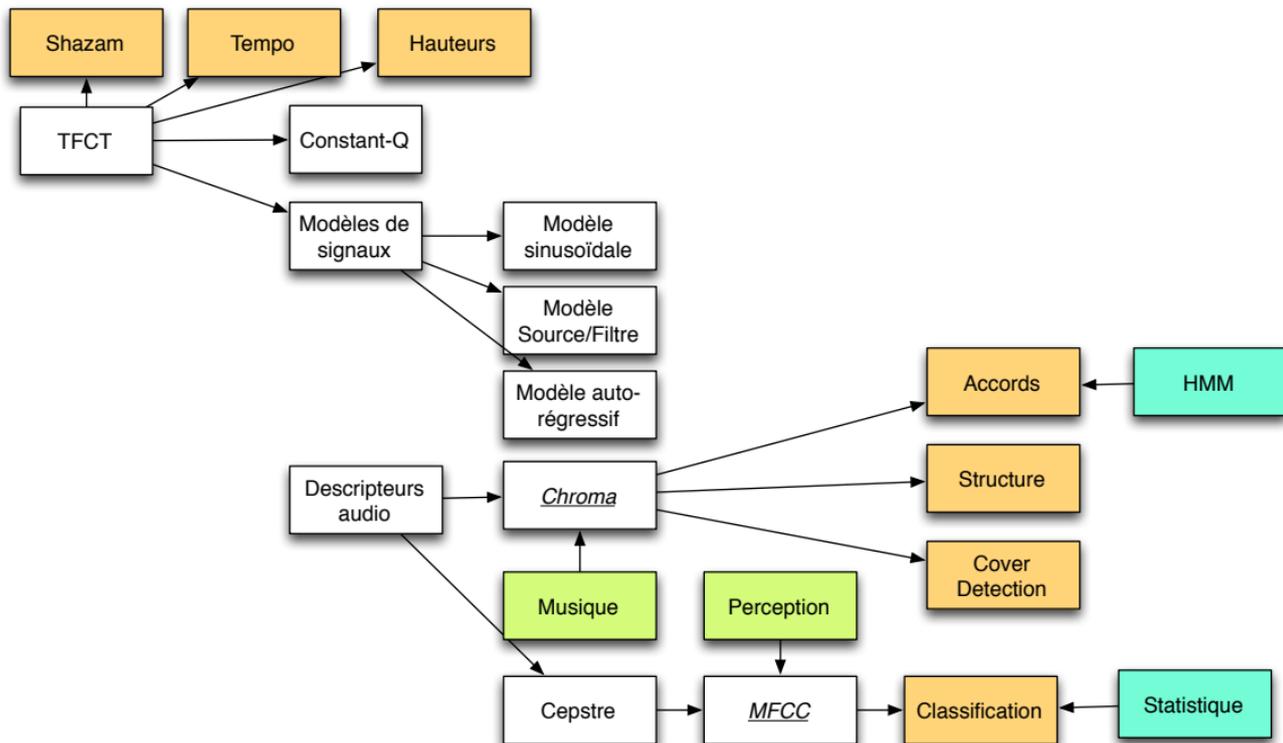
3- Théorie : Traitement du signal

3.4- Identification audio



3- Théorie : Traitement du signal

3.4- Identification audio



3- Théorie : Traitement du signal

3.5- Estimation du tempo

Rythme ?

- Tempo (beat)
 - indiquer sur une partition
 - "vitesse moyenne à laquelle les gens tapent du pied en écoutant la musique"
- Subdivision du rythme
 - mesure
 - entre deux barres, le groupement des noires
 - tactus
 - généralement la noire → le tempo
 - tatum
 - la vitesse la plus rapide
 - la subdivision de la noire en croches, triple-croches, double-croches

Andante grazioso (♩ = 120)

p

Beats Beats divided Beat subdivided

1 2 3 1 2 3 1 2 3

Beats Beats divided Beat subdivided

1 2 3 1 2 3 1 2 3

Beats Beats divided Beats subdivided

1 2 3 4 1 2 3 4 1 2 3 4

Beats Beats divided Beats subdivided

1 2 3 4 1 2 3 4 1 2 3 4

3- Théorie : Traitement du signal

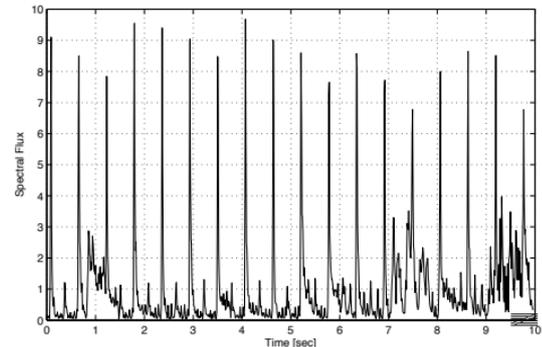
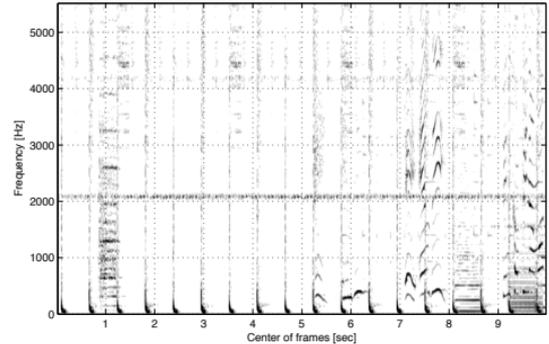
3.5- Estimation du tempo

Estimation du tempo ?

- Détecter la périodicité des évènements dans le signal audio

Détection des évènements ?

- Début des évènements = onsets
- Méthode 1
 - détecter les maxima locaux de la fonction d'énergie du signal
 - $ener(m) = \sum_k X(k, m)^2$
- Méthode 2
 - détecter les maxima locaux du flux spectral :
 - $flux(m) = \sum_k \text{HWR}[X(k, m) - X(k, m - 1)]$
 - $\text{HWR}(x) = x$ si $x > 1$
 - $\text{HWR}(x) = 0$ sinon

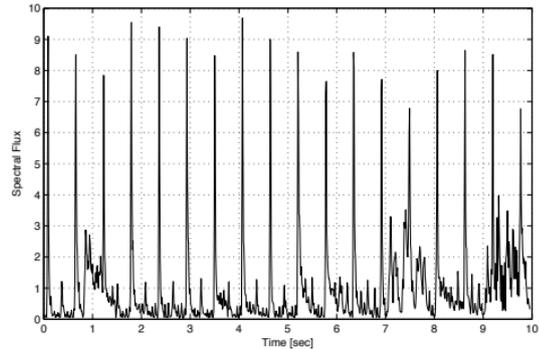


3- Théorie : Traitement du signal

3.5- Estimation du tempo

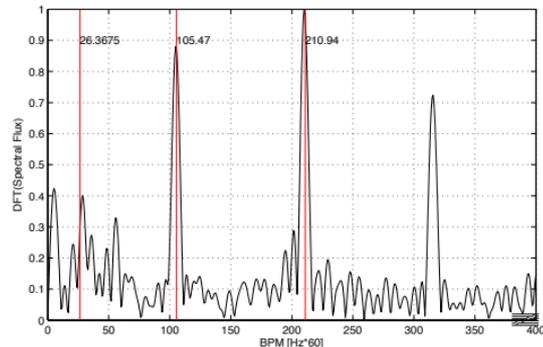
Périodicité des évènements ?

- Calcul de la transformée de Fourier (DFT) du flux spectral :
 - $FLUX(k) = \sum_{n=0}^{N-1} flux(n) \cdot \exp(j2\pi \frac{k}{N}n), \forall k$
- Calcul de la fonction d'auto-corrélation du flux spectral



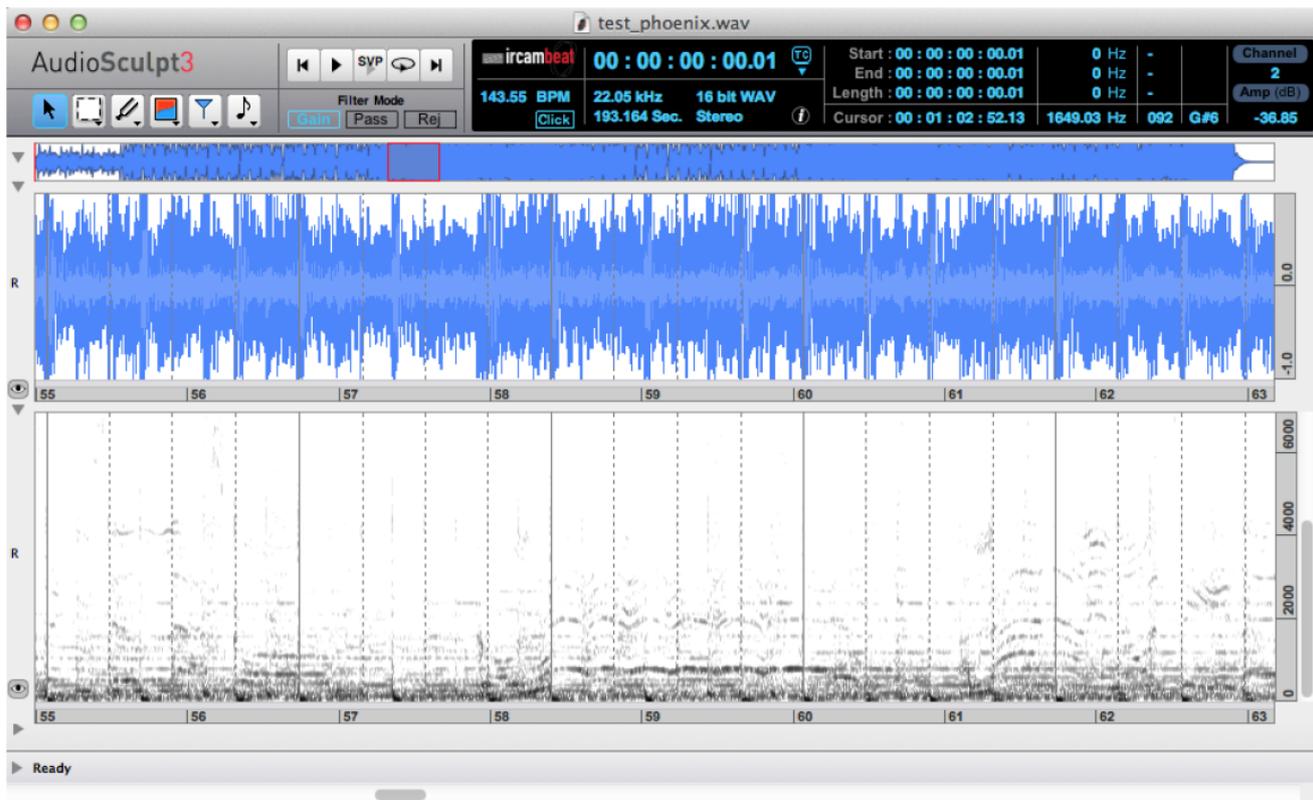
Estimation du tempo ?

- Détecter le peak (fréquence f_k de la DFT) correspondant au tempo
 - Tempo = $60f_k$ (BPM : Battement par Minute)
 - Peaks correspondant à la mesure, au tactus, au tatum



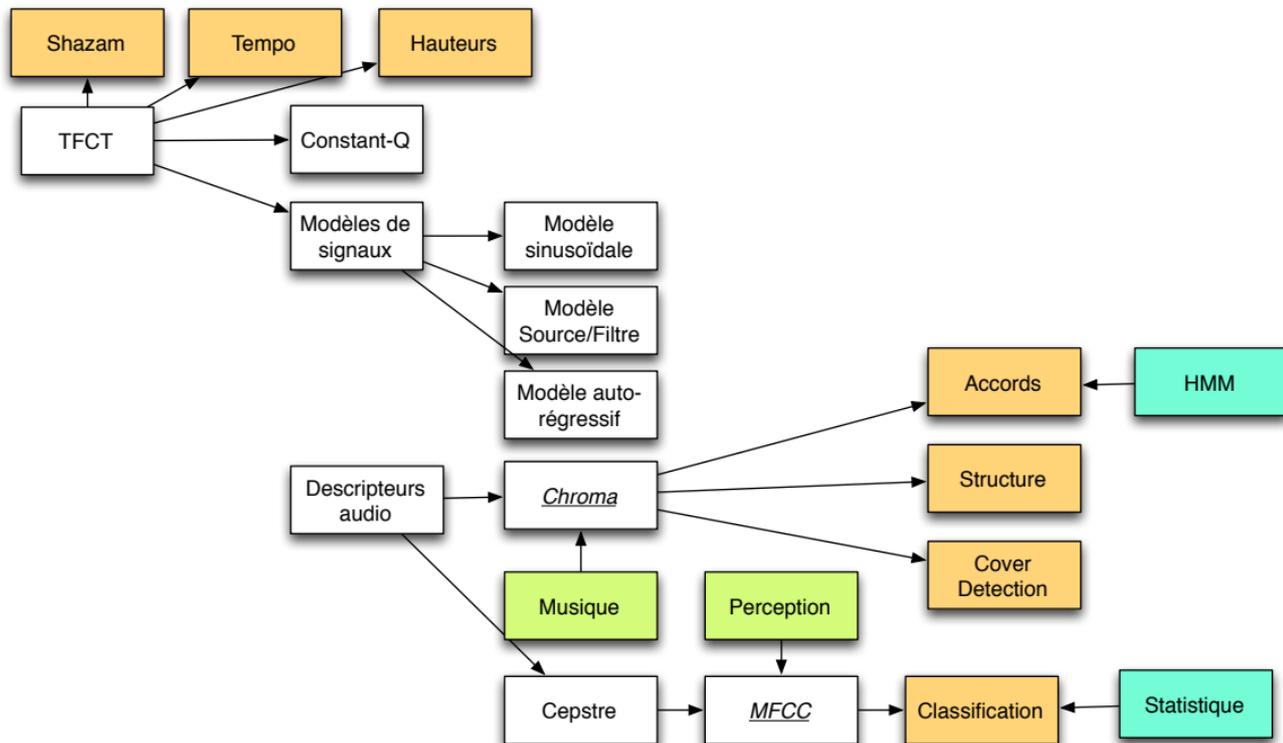
3- Théorie : Traitement du signal

3.5- Estimation du tempo



3- Théorie : Traitement du signal

3.5- Estimation du tempo

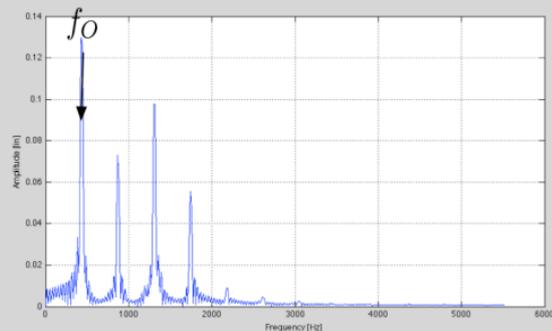
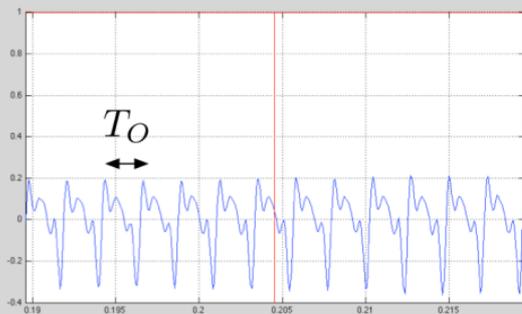


3- Théorie : Traitement du signal

3.6- Estimation de la hauteur d'une note

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.

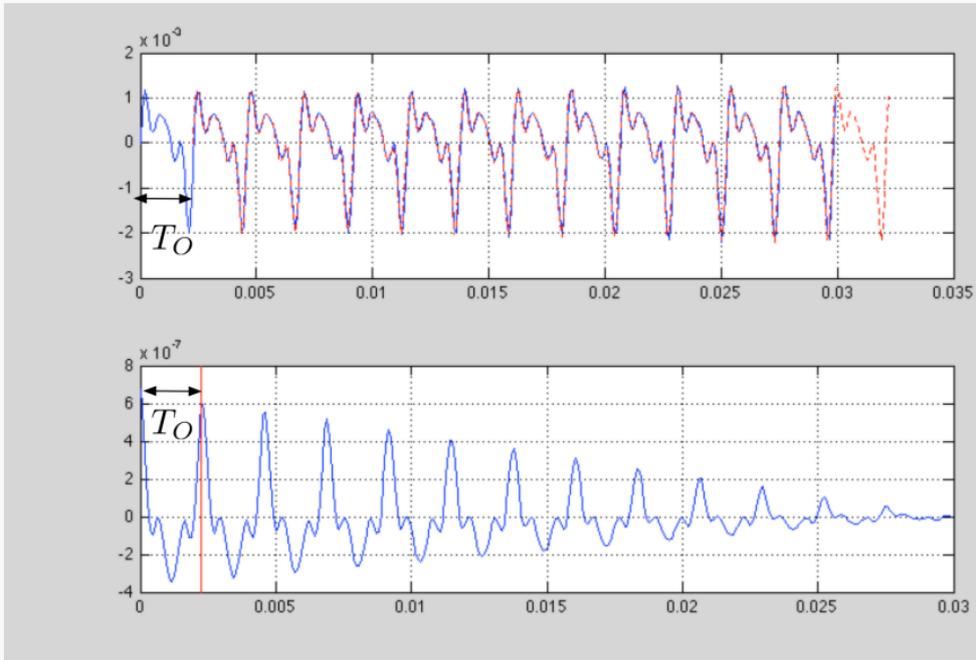


3- Théorie : Traitement du signal

3.6- Estimation de la hauteur d'une note

Méthode de l'autocorrélation

- $r(m) = \frac{1}{N} \sum_{n=0}^{N-1-m} x(n)x(n+m)$ si $m \geq 0$
- $r(m)$ est maximum si $m = T_0$

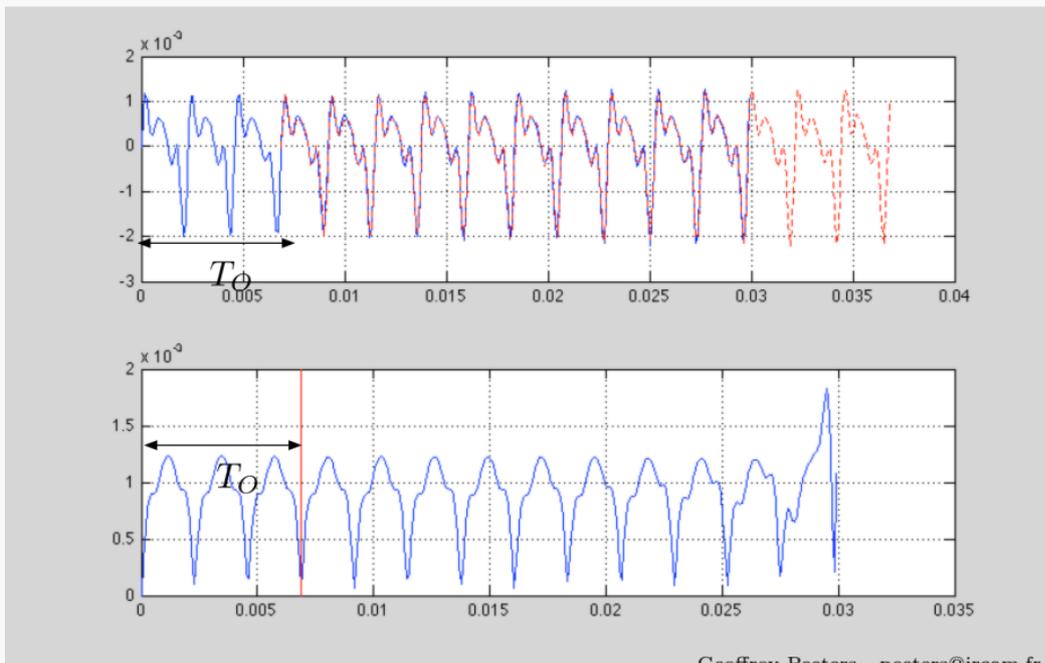


3- Théorie : Traitement du signal

3.6- Estimation de la hauteur d'une note

Méthode de l'Average Mean Difference Function (AMDF)

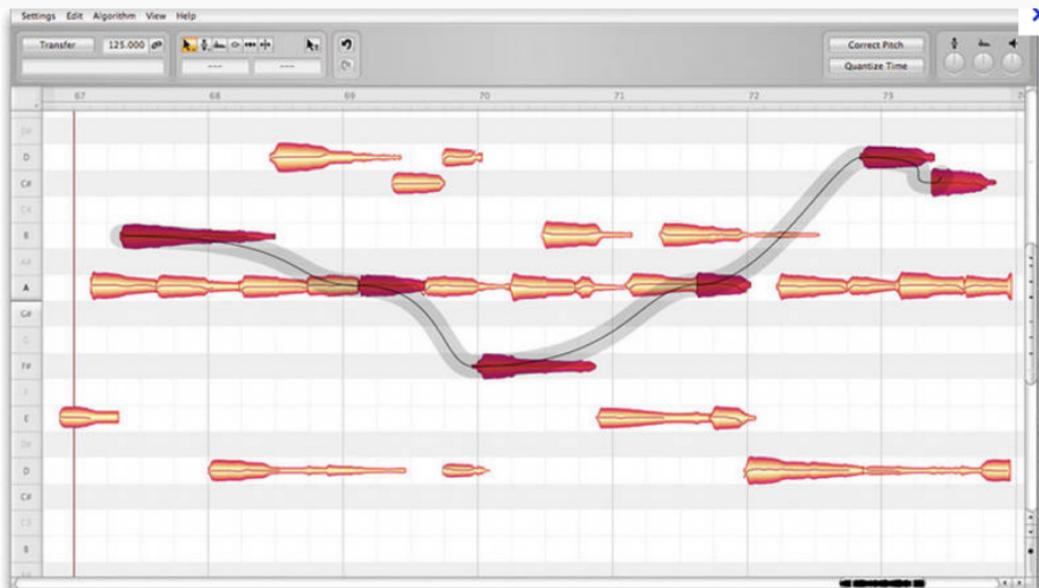
- $AMDF(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x(n) - x(n+m)|$
- $AMFT(m) = 0$ si $m = T_0$



3- Théorie : Traitement du signal

3.6- Estimation de la hauteur d'une note

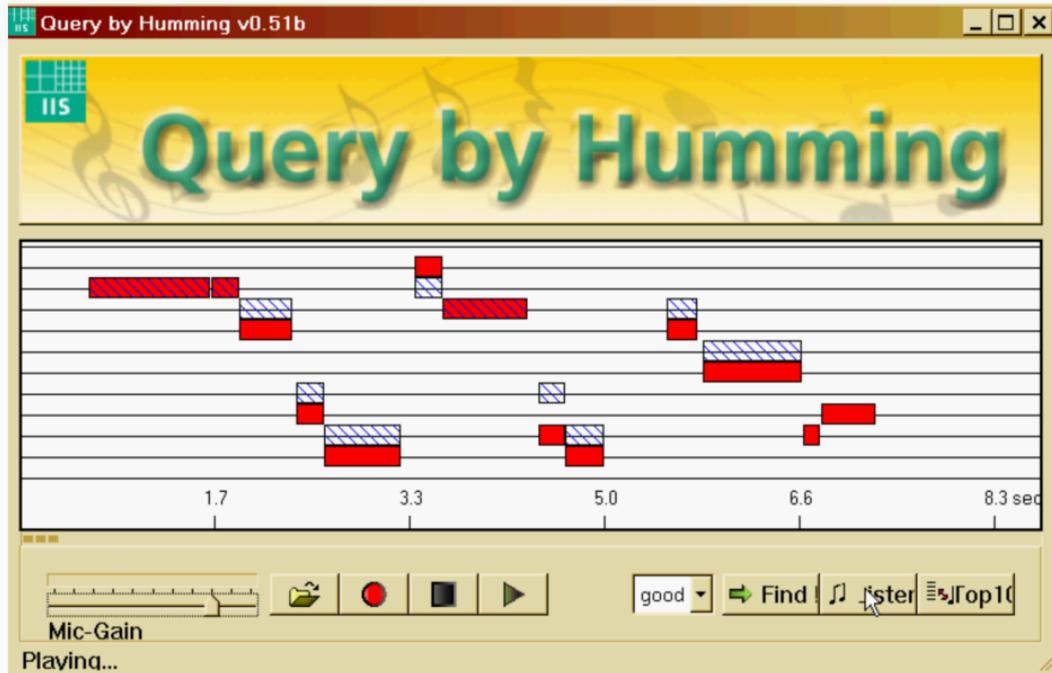
Application : Edition Audio



3- Théorie : Traitement du signal

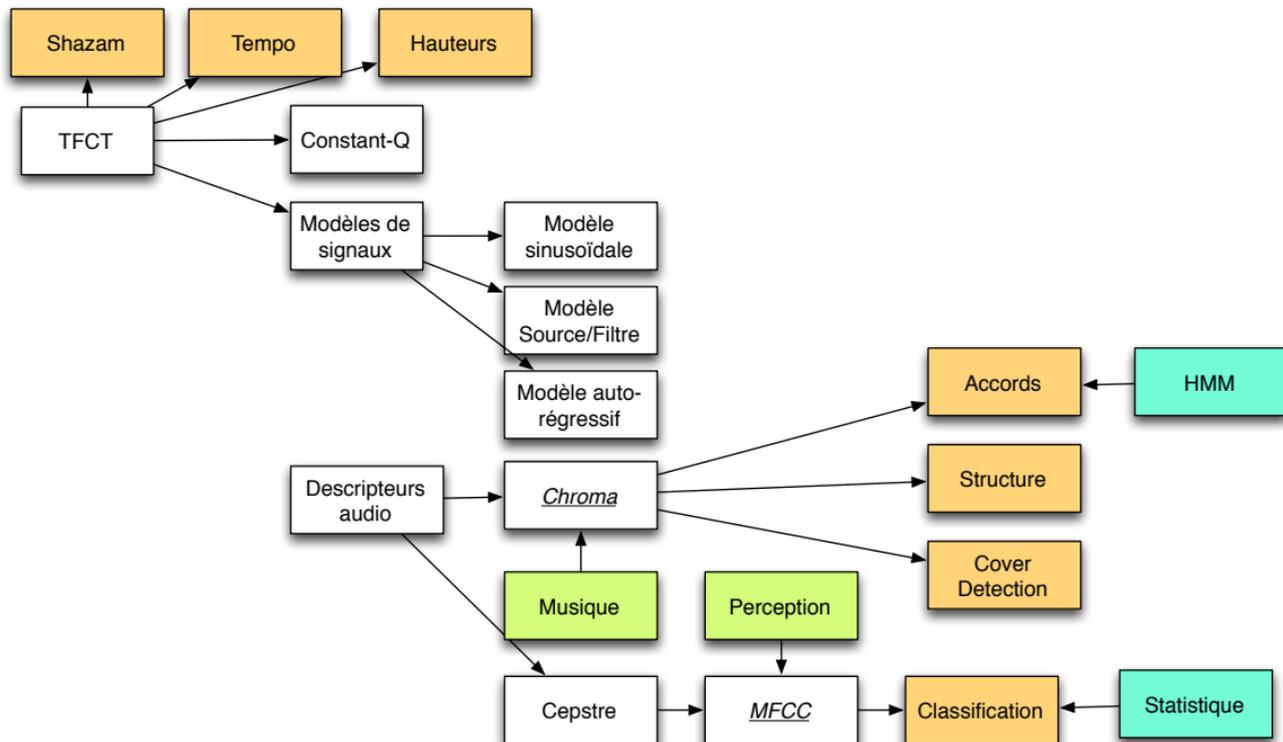
3.6- Estimation de la hauteur d'une note

Application : Query by Humming



3- Théorie : Traitement du signal

3.6- Estimation de la hauteur d'une note



4- Modèles de signaux

4.1- Modèle source/ filtre

Modèle source/ filtre

- Hypothèse :

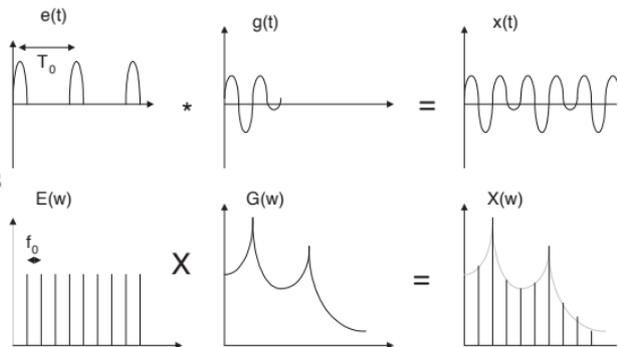
- le signal $x(t)$ est le résultat du passage d'une excitation (un pulse, une série de pulse) dans un filtre (résonnant)
 - Exemples : le signal de parole, certains instruments de musique (trompette)

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

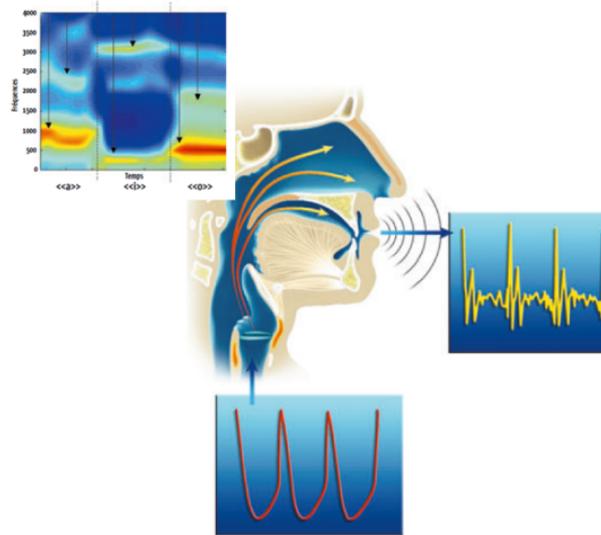


4- Modèles de signaux

4.1- Modèle source/ filtre

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/ nez)
 - filtrage résonant/ anti-résonant

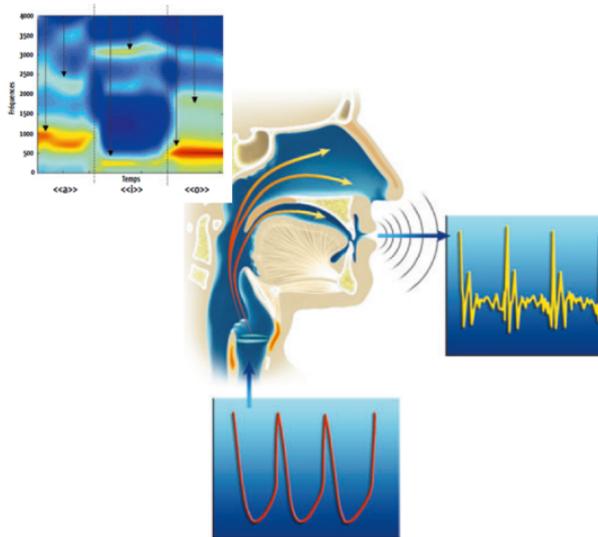


4- Modèles de signaux

4.1- Modèle source/ filtre

Production du signal vocal

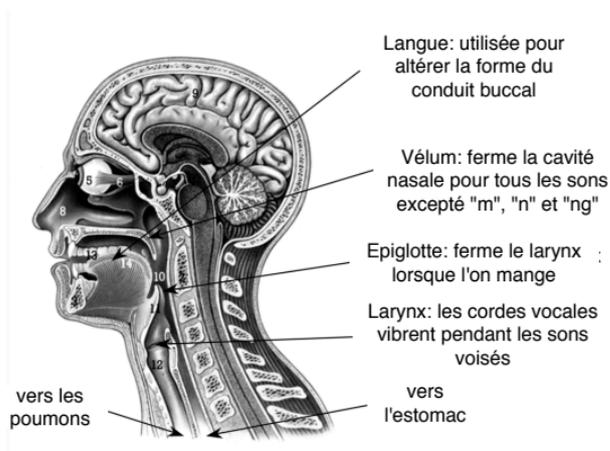
- Ouverture/fermeture périodique des cordes vocales
 - détermine la hauteur
 - Hauteur de 100Hz ? pulses d'air sont espacés de $T_0 = 1/f_0 = 1/100 = 10ms$.
 - Appelé signal d'excitation (ou signal source), $e(t)$.
- Conduit bucco-nasal
 - créer les différentes voyelles pour une hauteur donnée en renforçant (résonance) et retirant (anti-résonances) certains fréquences.
 - Filtre résonant (AR : Auto-Regressif) et anti-résonant (MA : Moving Average) : un filtre dit "ARMA".



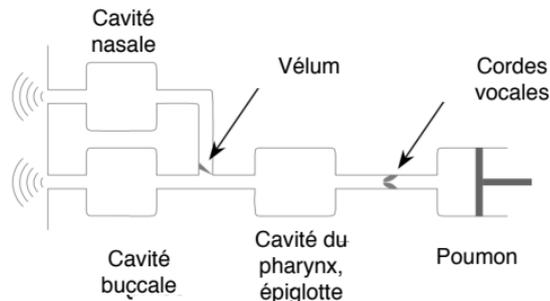
4- Modèles de signaux

4.1- Modèle source/ filtre

Production du signal vocal



source : Mike Brookes

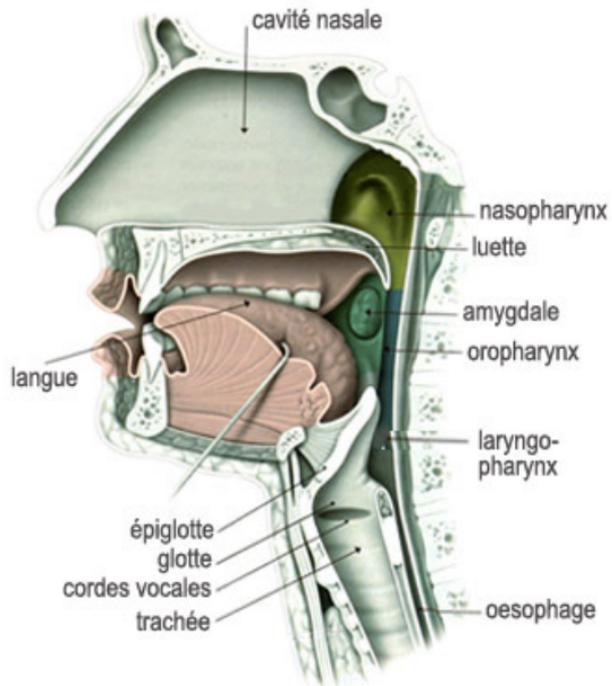


source : Mike Brookes

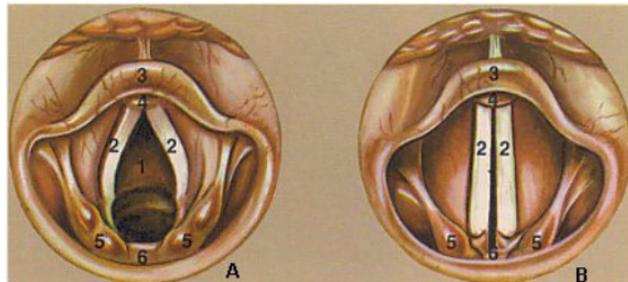
4- Modèles de signaux

4.1- Modèle source/ filtre

Production du signal vocal



source : outilsrecherche.over-blog.com

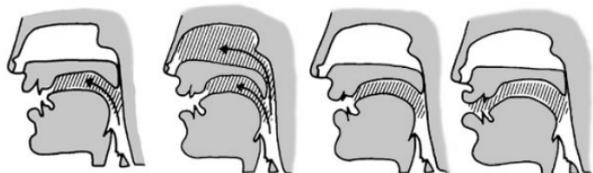


source : outilsrecherche.over-blog.com

4- Modèles de signaux

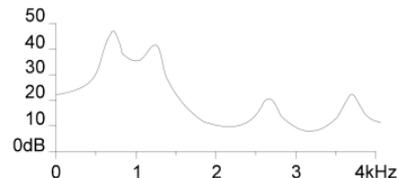
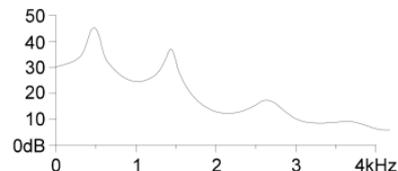
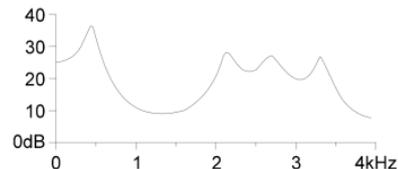
4.1- Modèle source/ filtre

Production du signal vocal



Voyelles orales Voyelles nasales Voy. non-arrondies Voy. arrondies

source : outilsrecherche.over-blog.com

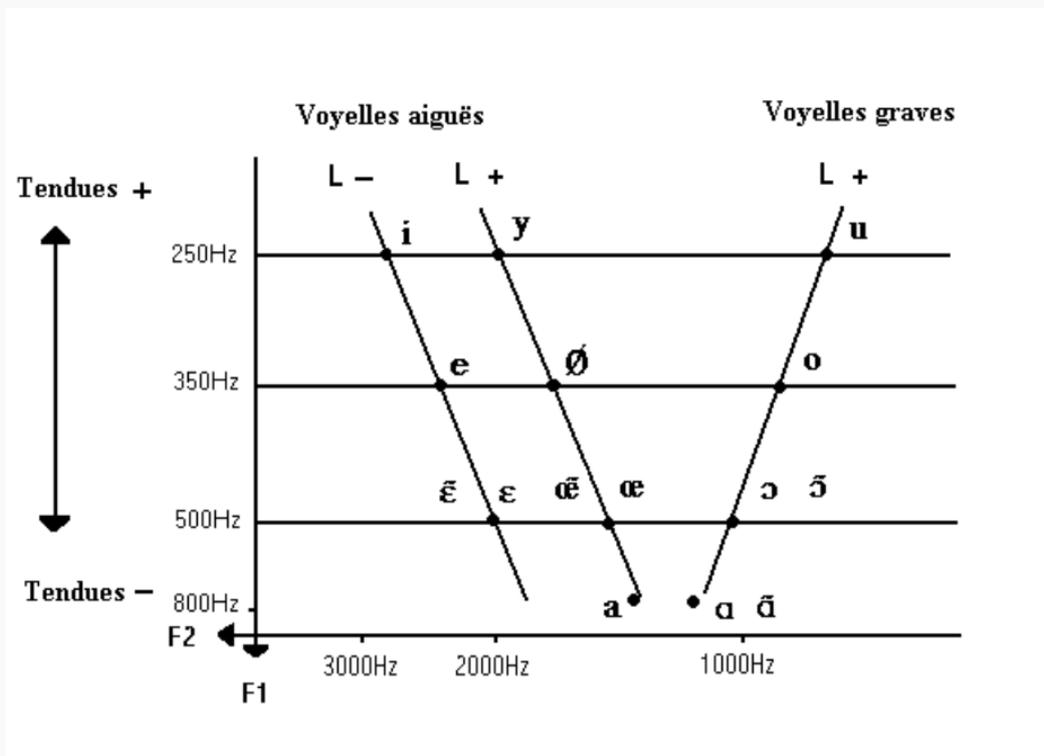


source : Mike Brookes

4- Modèles de signaux

4.1- Modèle source/ filtre

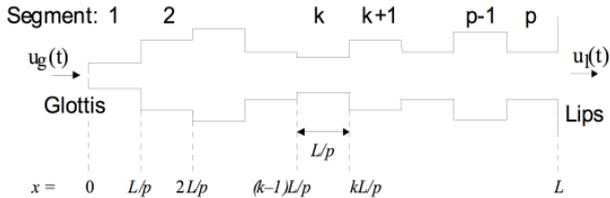
Fréquences des formants pour les différentes voyelles



4- Modèles de signaux

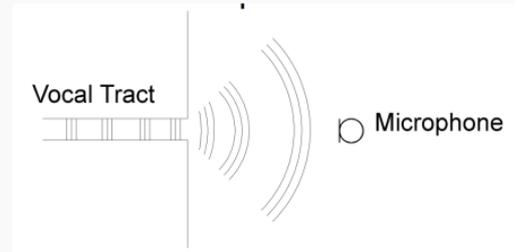
4.1- Modèle source/ filtre

Représentation sous-forme de tube



source : Mike Brookes

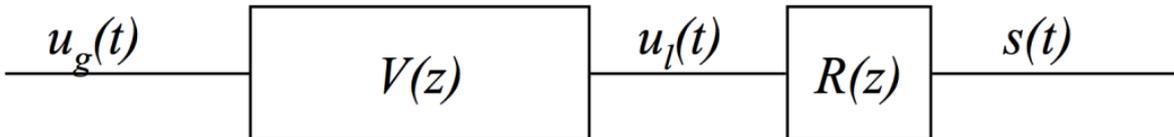
Radiation des lèvres



source : Mike Brookes

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

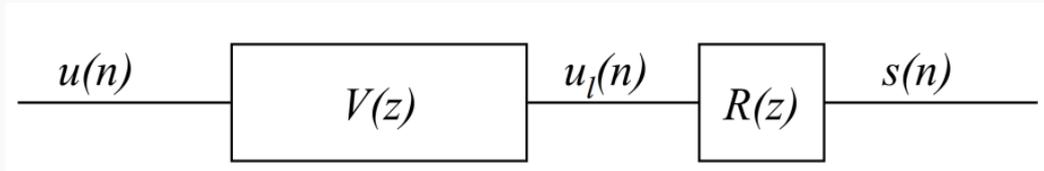
Système équivalent



4- Modèles de signaux

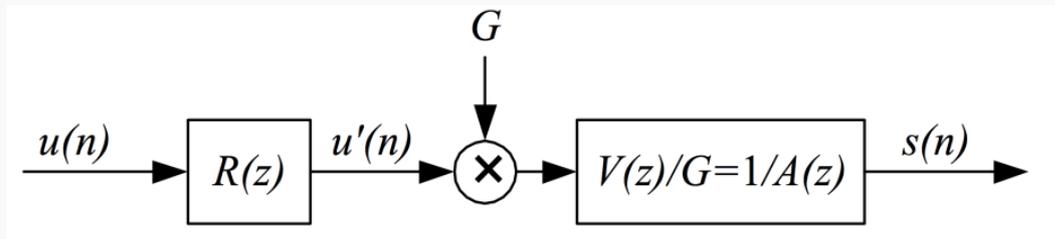
4.1- Modèle source/ filtre

La prédiction linéaire, modèle auto-régressif



source : Mike Brookes

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



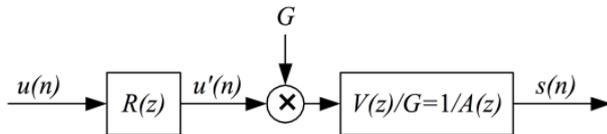
source : Mike Brookes

4- Modèles de signaux

4.1- Modèle source/ filtre

La prédiction linéaire, modèle auto-régressif

- $x(n) = Gu'(n) + \sum_{j=1}^p a_j x(n-j)$
 - 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 $s(n)$ comme combinaison linéaire des échantillons passés de la voix



source : Mike Brookes

4- Modèles de signaux

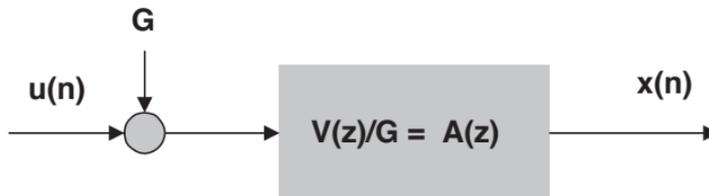
4.1- Modèle source/ filtre

La prédiction linéaire, modèle auto-régressif

- Le signal à l'instant n peut être prédit à partir des instants précédents

$$\begin{aligned}x(n) &= a_1x(n-1) + a_2x(n-2) + a_3x(n-3)\dots + a_Px(n-P) \\ &= \sum_{p=1}^P a_px(n-p) + G \cdot u(n)\end{aligned}\tag{2}$$

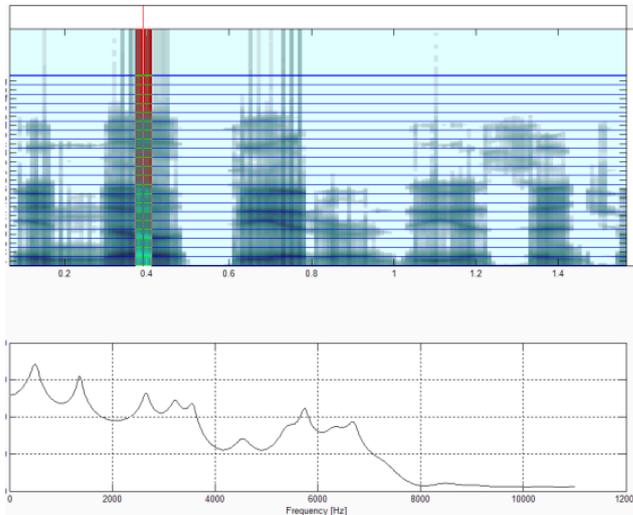
- Equivalent à passer le signal dans un filtre FIR tout-pôle : $V(z) = \frac{G}{1 + \sum_{p=1}^P a_p z^{-p}}$
- Objectif de la prédiction linéaire ?
 - déterminer le filtre $V(z)$ (donc les résonances ou les formants dans le cas de la voix) à partir du signal $x(n)$ (signal de pression micro)



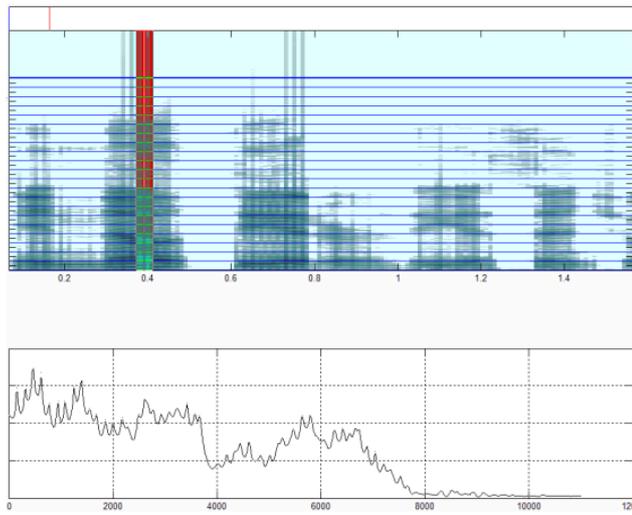
4- Modèles de signaux

4.1- Modèle source/ filtre

Choix du nombre de pôle P



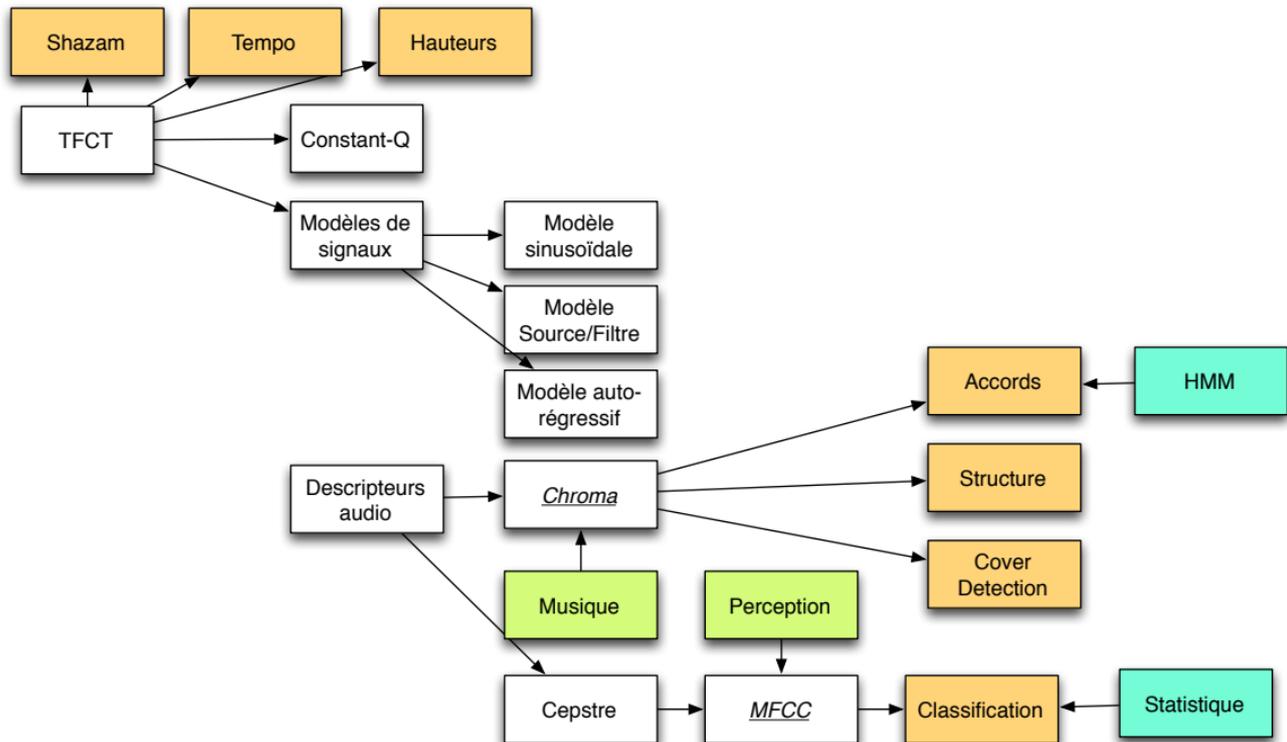
$P = 40$



$P = 200$

4- Modèles de signaux

4.1- Modèle source/ filtre



5- Descripteurs audio

5- Descripteurs audio

5.1- Introduction

Les descripteurs audio

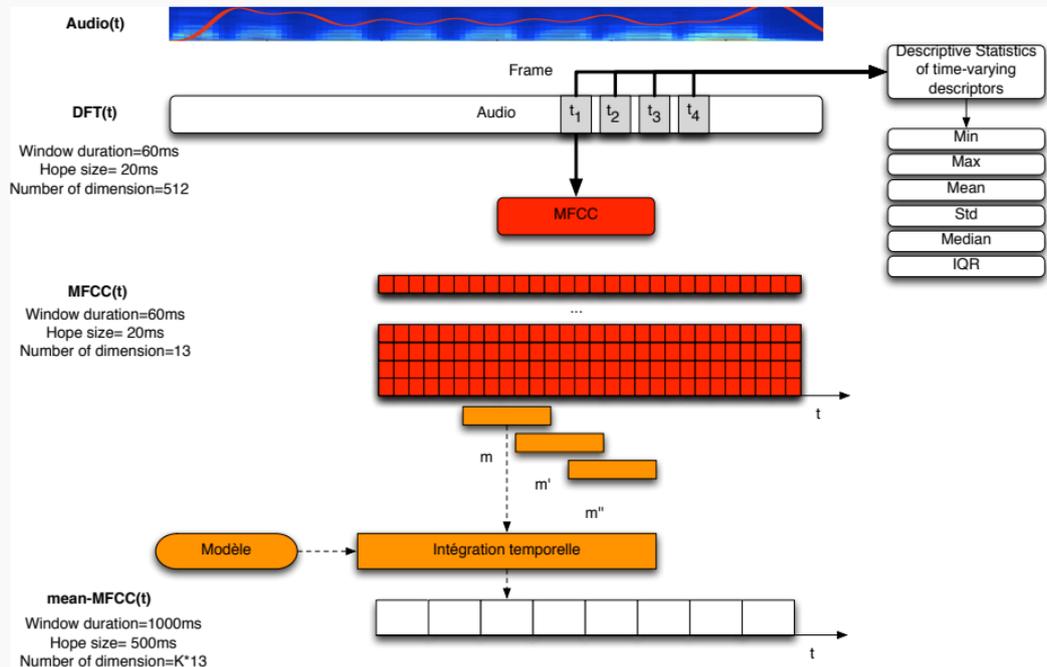
[G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado project report, Ircam, 2004.]

- 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

5.1- Introduction

Les descripteurs audio



5- Descripteurs audio

5.5- Mel Frequency Cepstral Coefficients (MFCCs)

Objectif

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

Cepstre complexe

- **Cepstre complexe** :

$$\text{cepstre} = c(\tau) = TF^{-1}(\log(X(\omega))) \frac{1}{2\pi} \int_{\omega} \log(X(\omega)) e^{j\omega\tau} d\omega \quad (3)$$

- τ est appelé "céfrence"
- $x(t) \rightarrow X(\omega) \rightarrow \log(X(\omega)) \rightarrow TF^{-1} \rightarrow \text{cepstre } c(l)$
- Modèle source/ filtre :

$$\begin{aligned} x(t) &= e(t) \circledast g(t) \xrightarrow{TF} X(\omega) = E(\omega) \cdot G(\omega) \\ X(\omega) &\xrightarrow{\log} \log(X(\omega)) = \log(E(\omega)) + \log(G(\omega)) \end{aligned} \quad (4)$$

- Le cepstre fournit une information relative à la variation du spectre à travers les fréquences

5- Descripteurs audio

5.5- Mel Frequency Cepstral Coefficients (MFCCs)

Cepstre réel

- **Cepstre réel :**

- Cepstre calculé sur la partie réelle du log-spectrum \rightarrow la log-amplitude
- si on note $X(\omega) = A(\omega) \cdot e^{j\phi(\omega)}$

$$\begin{aligned}\text{cepstre réel} &= TF^{-1}(\Re(\log(X(\omega)))) \\ &= TF^{-1}(\Re(\log(A(\omega)) + j\phi(\omega))) \\ &= TF^{-1}(\log(A(\omega)))\end{aligned}\tag{5}$$

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

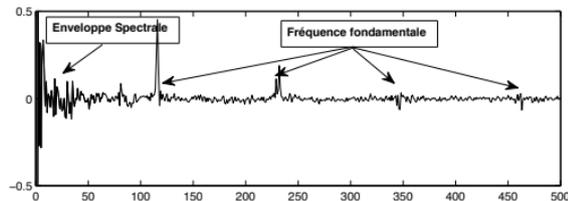
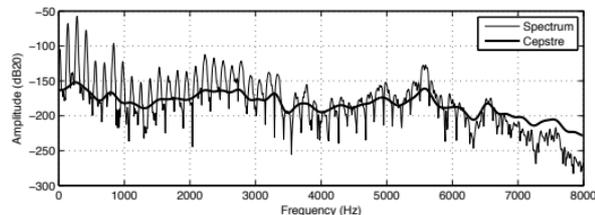
- Le spectre d'amplitude étant réel et symétrique
 - sa TF se réduit à sa partie réelle
 - donc à la projection de $\log(A(\omega))$ sur un ensemble de cosinus \rightarrow DCT

5- Descripteurs audio

5.5- Mel Frequency Cepstral Coefficients (MFCCs)

Cepstre complexe

- Les **coefficients** (cepstraux)
- d'**ordre faibles** (proches de zéro)
 - = décomposition du spectre sur des cosinus à variation lente
 - représente le spectre du filtre $G(\omega)$ (résonances/ anti-résonances),
 - la $TF^{-1}(\log(G(\omega)))$ varie lentement à travers ω
- d'**ordre élevés**
 - = décomposition du spectre sur des cosinus à variation rapide
 - représente le spectre de la source $E(\omega)$ (fréquence fondamentale),
 - la $TF^{-1}(\log(E(\omega)))$ varie rapidement à travers ω



5- Descripteurs audio

5.5- 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 perceptible (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 ?

5- Descripteurs audio

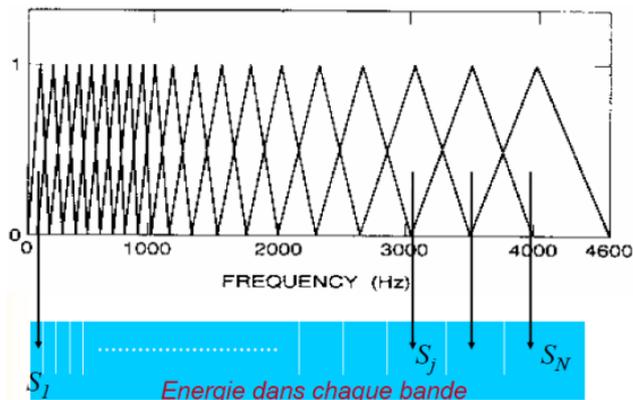
5.5- Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs)

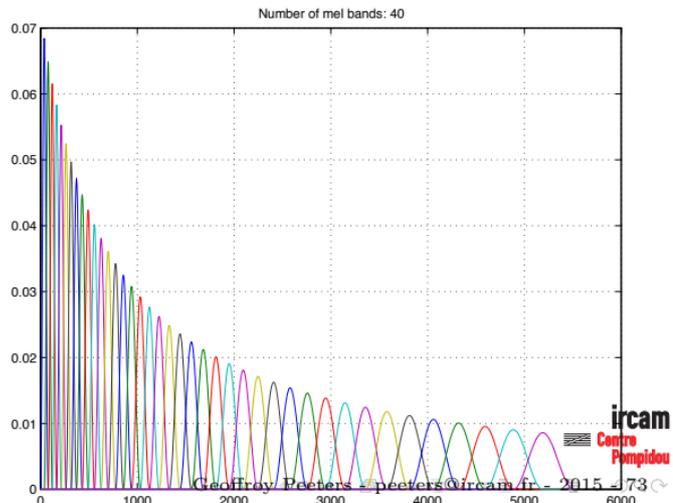
- Echelle de Mel :

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

$$M = f_c \left(1 + \log_{10} \left(\frac{f}{f_c} \right) \right) \text{ pour } f \geq 1000\text{Hz} \quad (6)$$



source : Gaël Richard

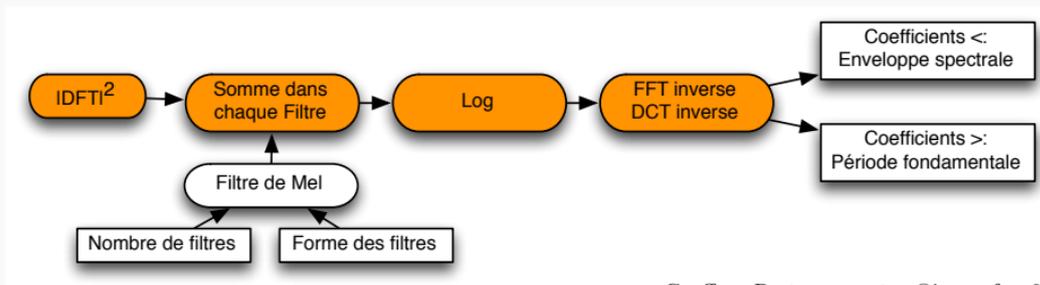


5- Descripteurs audio

5.5- Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs)

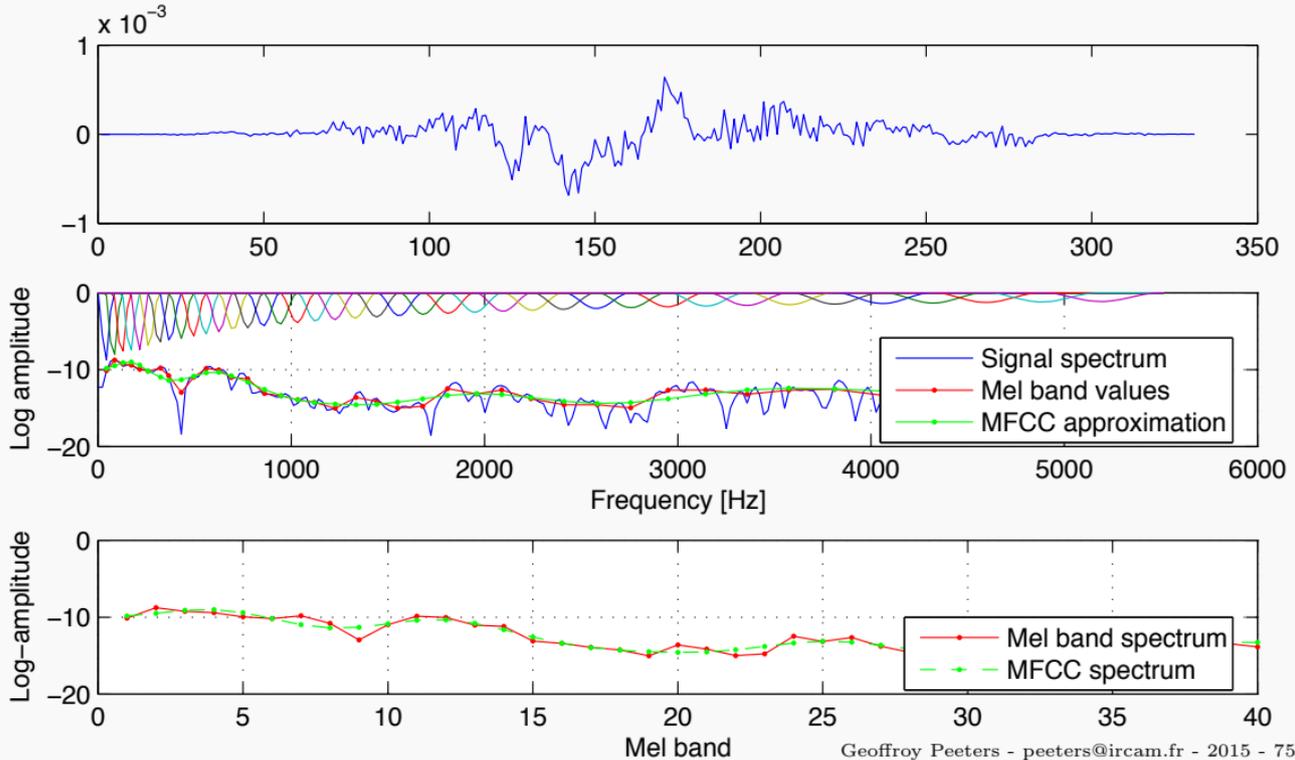
- Calcul du spectre de puissance : $|X(\omega)|^2$
- Calcul des filtres de Mel : $H_b(\omega)$ avec $b \in [1, B]$
 - choix du nombre de filtres B : 40
 - choix de la forme des filtres : triangulaire, hanning, tanh, ...
- Conversion du spectre de puissance en bandes de Mel : $S(b) = \sum_{\omega} |X(\omega)|^2 \cdot H_b(\omega)$
- 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

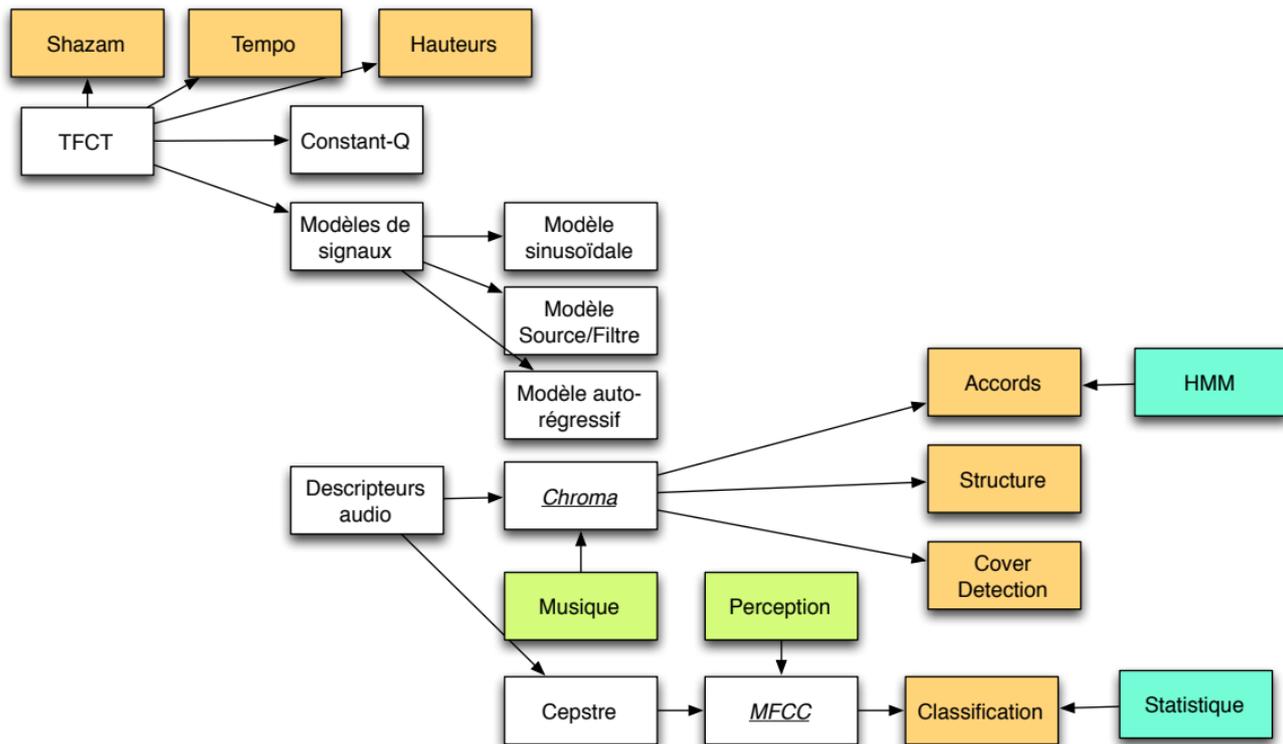
5.5- Mel Frequency Cepstral Coefficients (MFCCs)

Exemple de calcul de MFCCs



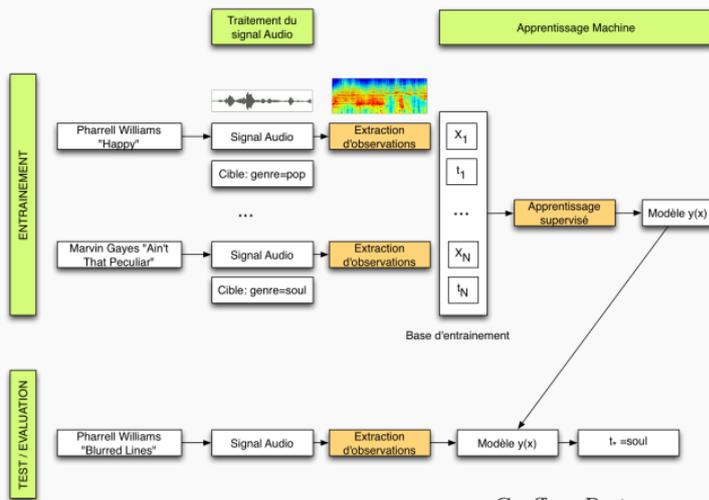
6- Applications

6.0-



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 temps en catégories paroles/musiques, musique instrumentale/chantée

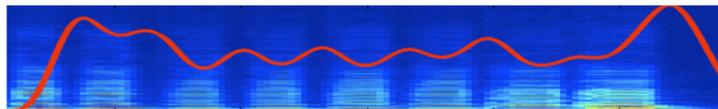


7- Classification Audio

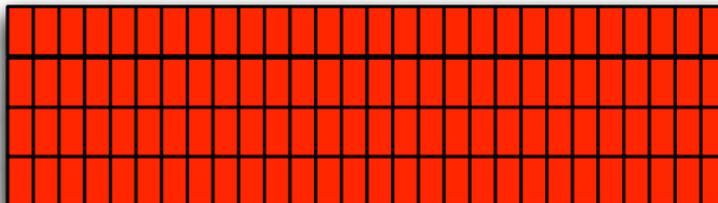
7.1- Extraction des descripteurs

Extraction des descripteurs sur un fichier audio

Morceau 1



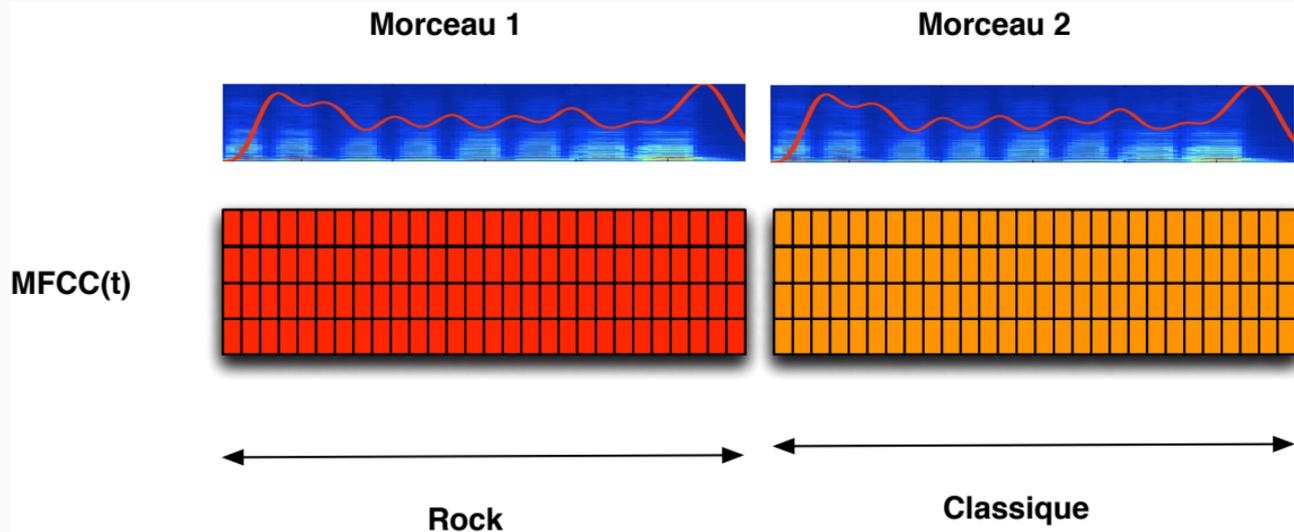
MFCC(t)



7- Classification Audio

7.1- Extraction des descripteurs

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

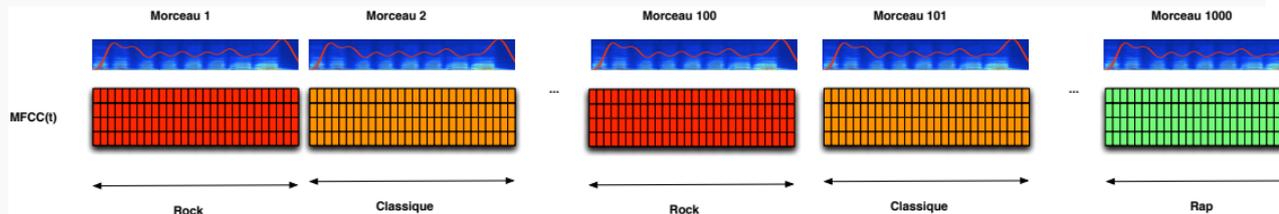


7- Classification Audio

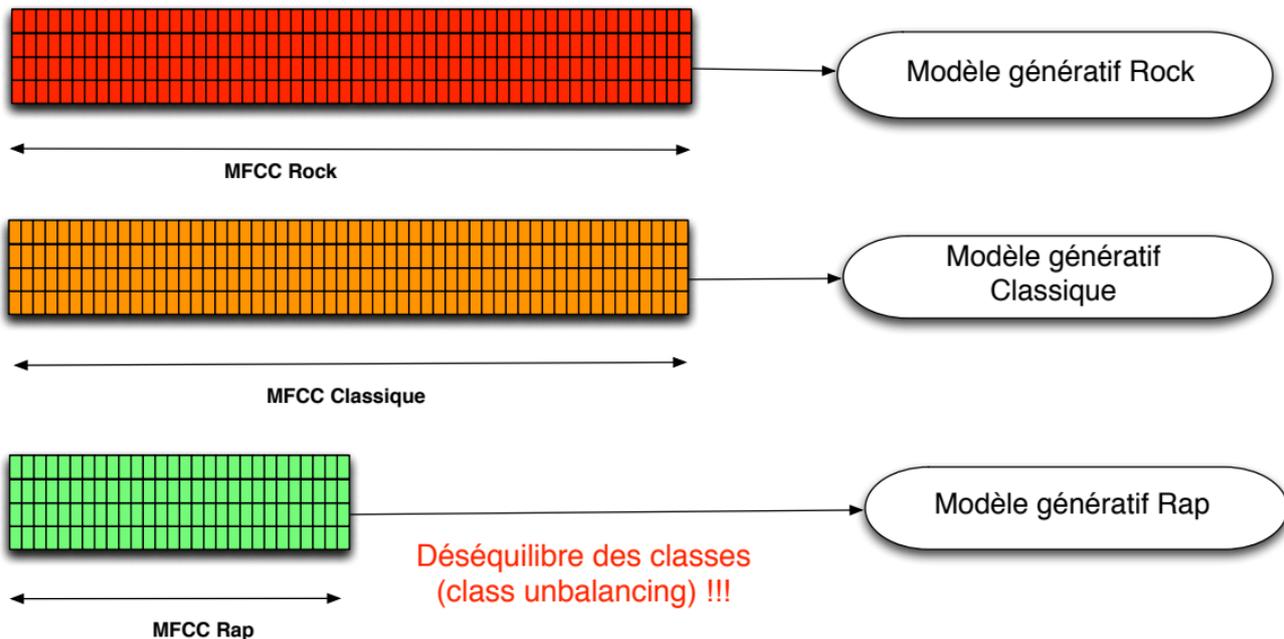
7.1- Extraction des descripteurs

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)



Apprentissage



7- Classification Audio

7.2- Apprentissage

Algorithmes d'apprentissage supervisé

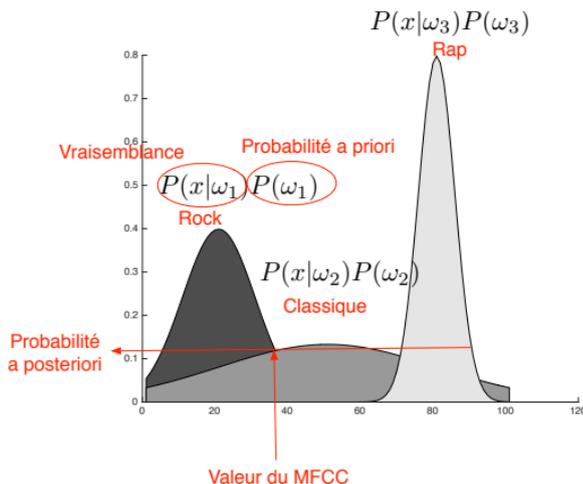
- Modèles génératifs : Gaussien, GMM
- Modèles discriminants : LDA, SVM
- Approche par exemplification : KNN

Modèle génératif gaussien

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

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

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



5- Descripteurs audio

5.6- Chroma - Pitch Class Profile (PCP)

Définition des Chroma - Pitch Class Profile (PCP)

- **Objectif :**

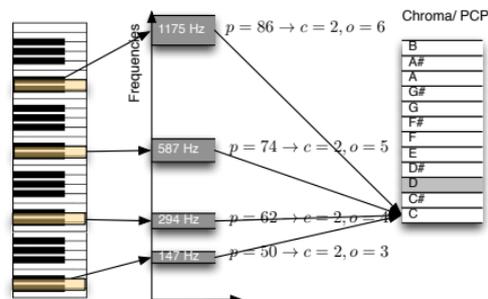
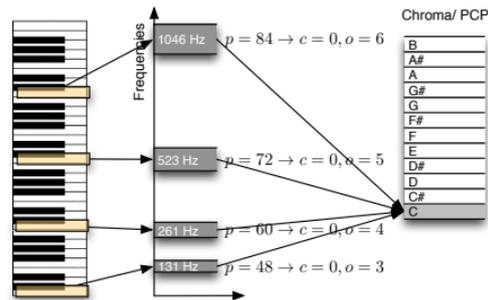
- le spectre à l'instant n : $X(k, n)$
- représenter son contenu harmonique sous forme d'un vecteur : $C(c, n)$ $c \in [0, 12[$

- 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),

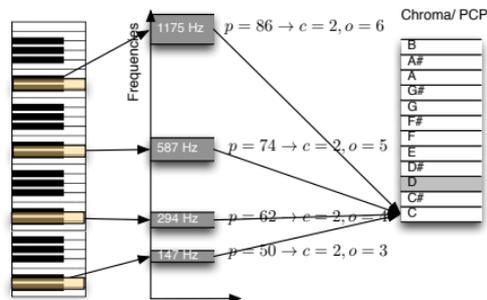
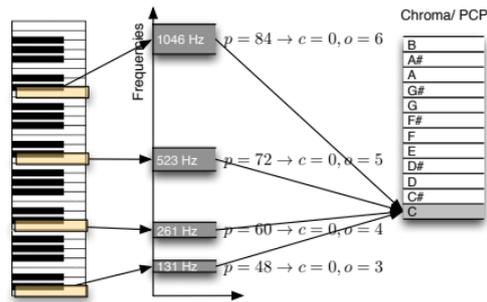


5- Descripteurs audio

5.6- Chroma - Pitch Class Profile (PCP)

Calcul des Chromas - Pitch Class Profile (PCP)

- Relation entre les fréquences f_k de la DFT et les hauteurs de note p (hauteurs de demi-tons en échelle de notes MIDI)
 - $p(f_k) = 12 \log_2 \left(\frac{f_k}{440} \right) + 69, p \in \mathbb{R}^+$
 - $f(p) = 440 \cdot 2^{\frac{p-69}{12}}$
- Calcul des chromas $C(c, n)$
 - On additionne toutes les valeurs du spectre $X(k, n)$ tel que f_k correspondent à un c donné
 - Hard-mapping
 - Soft-mapping



5- Descripteurs audio

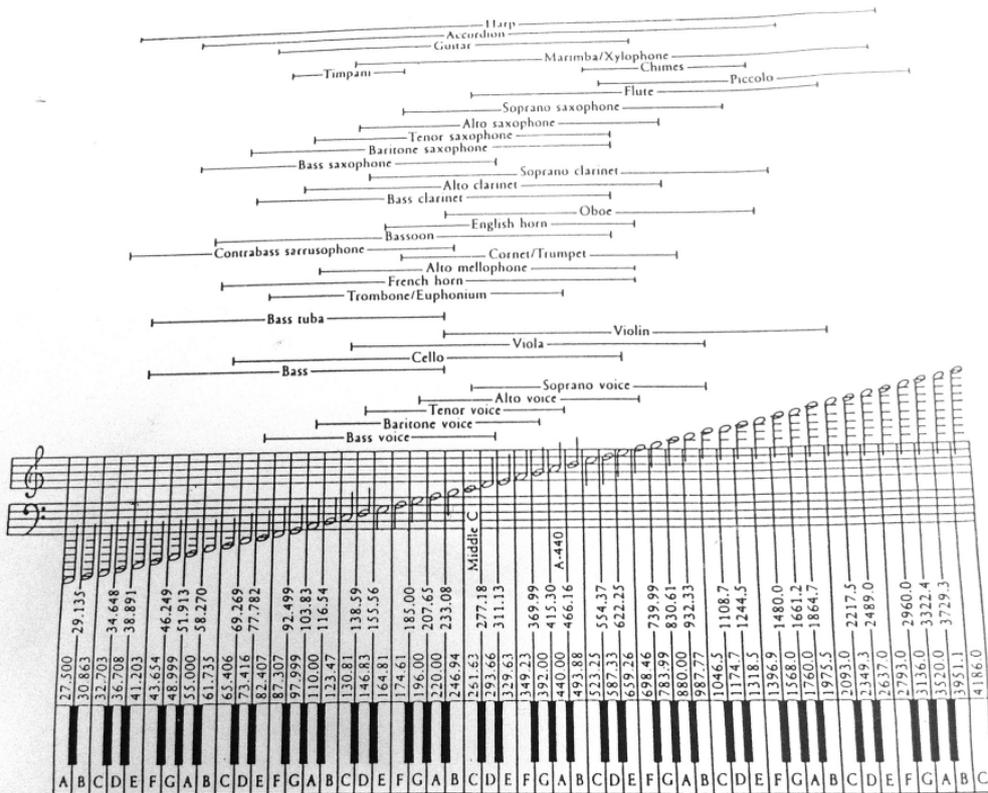
5.6- Chroma - Pitch Class Profile (PCP)

Calcul des Chromas - Pitch Class Profile (PCP)

- Résolution fréquentielle ?
 - Elle doit permettre la séparation des notes voisines
 - On définit la largeur (à -6 dB) : $Bw = \frac{Cw}{L_{sec}}$
 - Si f_{\min} (la fréquence la plus basse considérée dans le spectre) est 50 Hz
 - on veut séparer G#1 (51.91Hz) et A1 (55Hz) →
$$L_{sec} = \frac{Cw}{Bw} = \frac{2.35}{3.0869Hz} = 0.7613s$$
 - Si f_{\min} est 100 Hz
 - on veut séparer G#2 (103.82Hz) de A2 (110Hz) →
$$L_{sec} = \frac{Cw}{Bw} = \frac{2.35}{6.1738Hz} = 0.3806s$$
- Deux possibilités :
 - Choisir L_{sec} en fonction f_{\min}
 - Choisir f_{\min} en fonction de L_{sec}

5- Descripteurs audio

5.6- Chroma - Pitch Class Profile (PCP)



5- Descripteurs audio

5.6- Chroma - Pitch Class Profile (PCP)

Calcul des Chromas - Pitch Class Profile (PCP)

- Calcul des chromas $C(c, n)$
 - On additionne toutes les valeurs du spectre $X(k, n)$ tel que f_k correspondent à un c donné

5- Descripteurs audio

5.6- Chroma - Pitch Class Profile (PCP)

Limitations des Chromas - Pitch Class Profile (PCP)

- Présence des harmoniques supérieures de chaque note
 - En pratique pour une note C on a pas $[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$
 - mais plutôt $[a_1 + a_2 + a_4, 0, 0, 0, a_5, 0, 0, a_4, 0, 0, 0, 0]$
- Influence de l'enveloppe spectrale

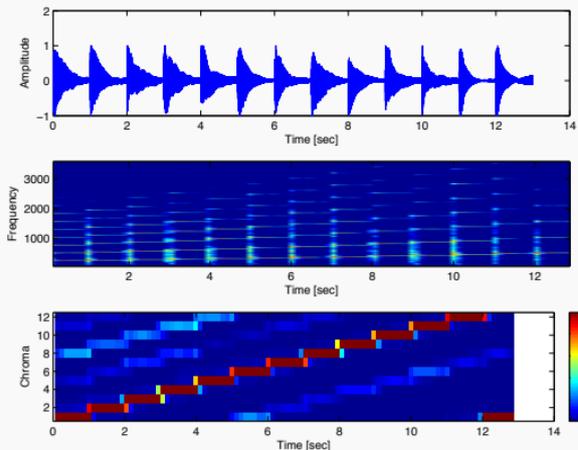
Pitch	Harmonic	Frequency f_μ	MIDI-scale m_μ	Chroma/PCP p
c3	f_0	130.81	48	1 (=c)
	$2f_0$	261.62	60	1 (=c)
	$3f_0$	392.43	67.01	8.01 (\simeq g)
	$4f_0$	523.25	72	1 (=c)
	$5f_0$	654.06	75.86	4.86 (\simeq e)

5- Descripteurs audio

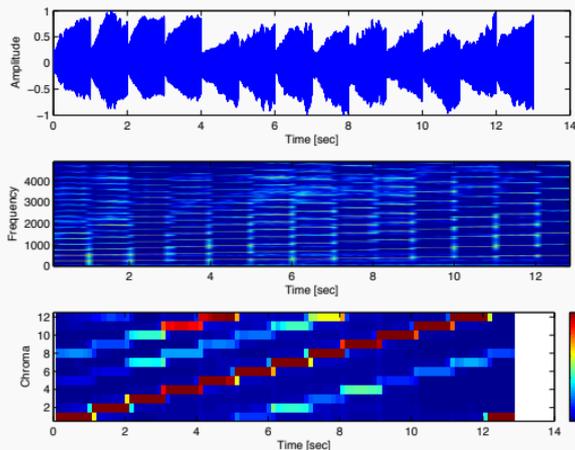
5.6- Chroma - Pitch Class Profile (PCP)

Limitations des Chromas - Pitch Class Profile (PCP)

Exemple piano

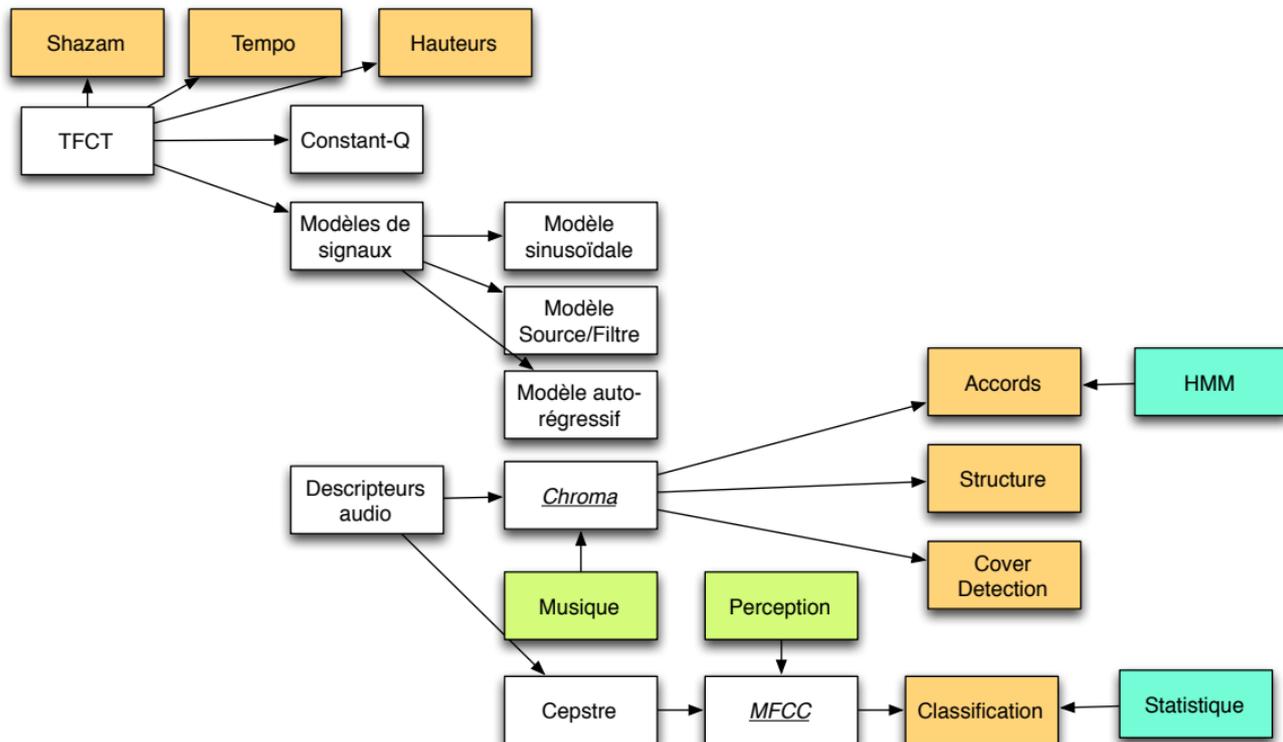


Exemple violon



7- Classification Audio

7.2- Apprentissage

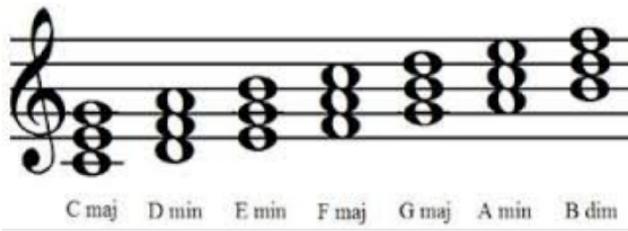


7- Classification Audio

7.3- Estimation d'accords

Les accords ?

- réduction du contenu harmonique à un instant donné d'un morceau
 - ensemble de notes jouées (quasi) simultanément
- différents types d'accords



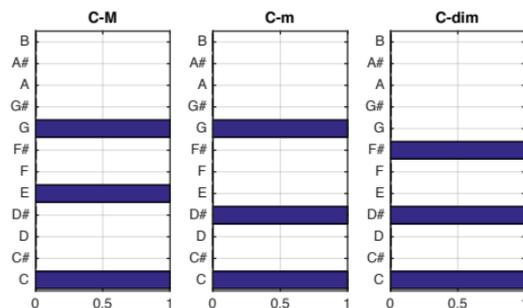
Root-note	Type of the chord
c, c#, d, d# ... b	Triads: major (C-M: c,e,g) , minor (C-m: c, e♭, g), suspended (C-sus2: c, d, g / C-sus4: c, f, g), augmented (C-aug: c, e, g#), diminished (C-dim: c, e♭, g♭)
	Tetrads: major 7 (C-M7: c, e, g, b), minor 7 (C-m7: c, e♭, g, b♭), dominant 7 (C-7: c, e, g, b♭), major 6 (C-M6: c, e, g, a), minor 6 (C-m6: c, e♭, g, a) ...
	Pentads: major 9 (C-M9: c, e g, b, d), dominant 9 (C-9: c, e, g, b, d) ...

7- Classification Audio

7.3- Estimation d'accords

Estimation des accords ?

- Représentation des accords dans un ordinateur
 - Création d'un vecteur à 12 dimensions (les 12 chromas)
 - valeur 1 si le chroma est présent dans l'accord
 - valeur 0 si le chroma est absent
 - Gabarits $G_a(c)$
 - a le nom de l'accord
 - $a \in [C-M, C-m, C\#-M, C\#-m, \dots]$
 - c le chroma
 - $c \in [0, 12[$
 - Vecteur de Chroma $C(c, n)$ à l'instant n

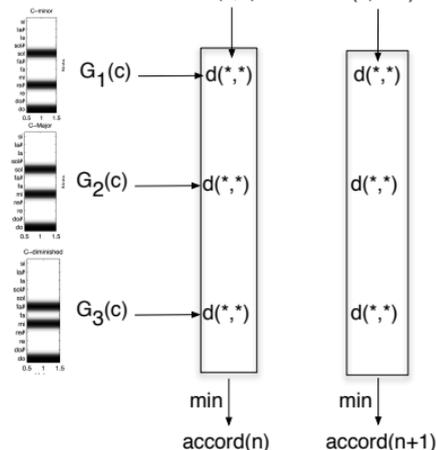
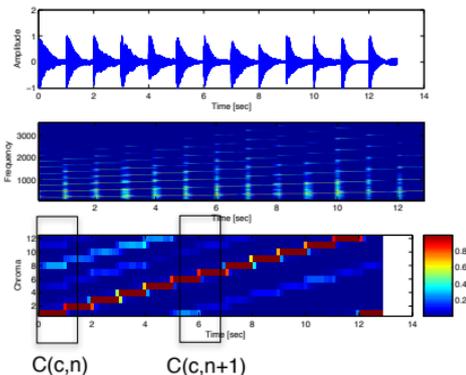


7- Classification Audio

7.3- Estimation d'accords

Estimation des accords par trame

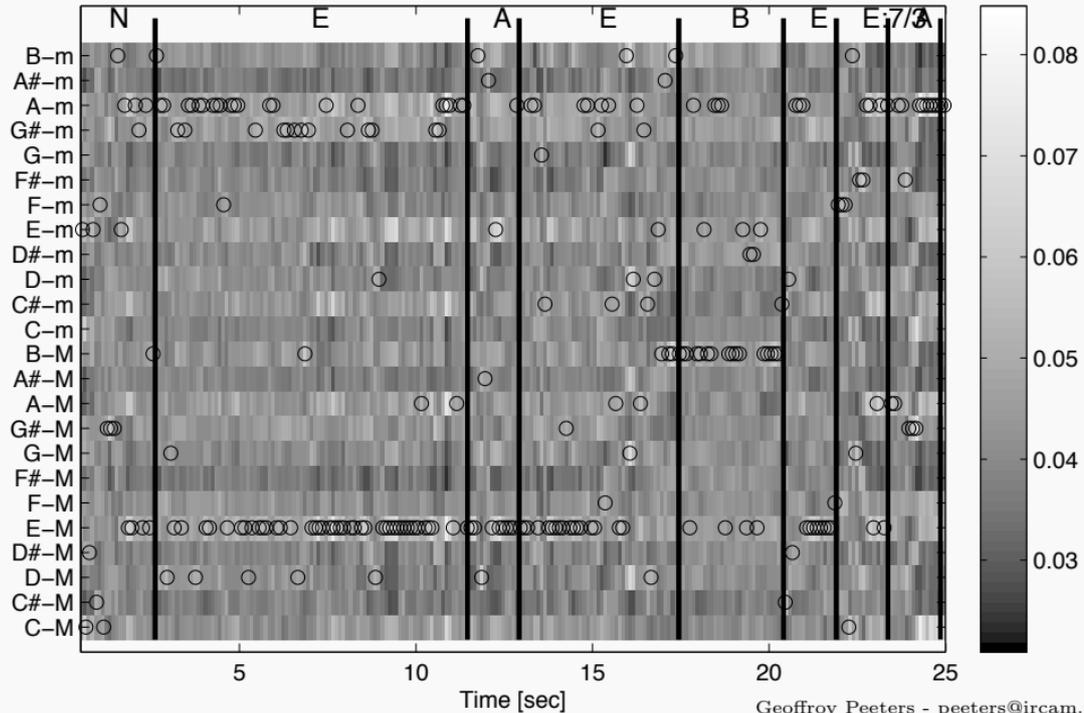
- A chaque instant n calcul de la distance entre gabarit d'accord $G_a(c)$ et vecteur de chroma $C(c, n)$
- Choix de l'accord donnant la distance la plus petite
- Choix de la distance $d(C(c, n), G_a(c))$
 - distance Euclidéenne
 - distance cosinusoidale



7- Classification Audio

7.3- Estimation d'accords

Estimation des accords par trame



7- Classification Audio

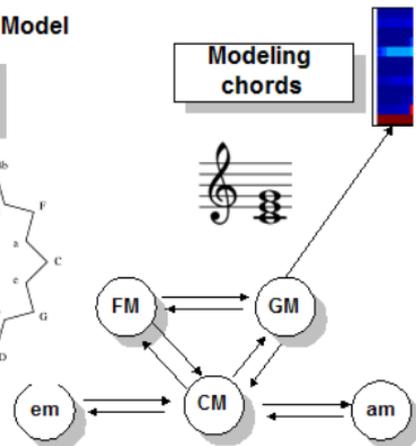
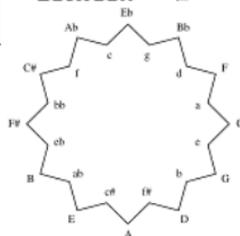
7.3- Estimation d'accords

Estimation des accords par trame par modèle de Markov caché (HMM)

- **Observations** :
 - on extrait la séquence de descripteurs audio chroma/PCP
- **Etats cachés** :
 - les 24 accords
- **Probabilités initiales** :
- **Probabilité d'émission** des accords :
 - $p(S = C-M|\text{chroma})$,
 - $p(S = C\#-M|\text{chroma})$, ...

Hidden Markov Model

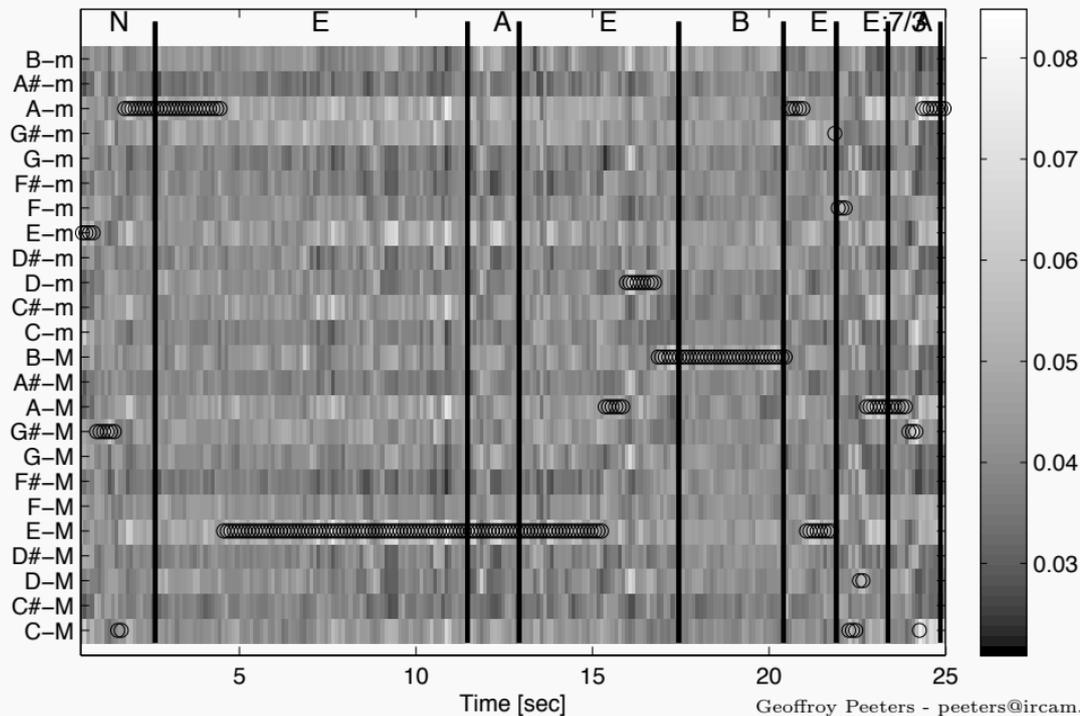
Modeling transition between



7- Classification Audio

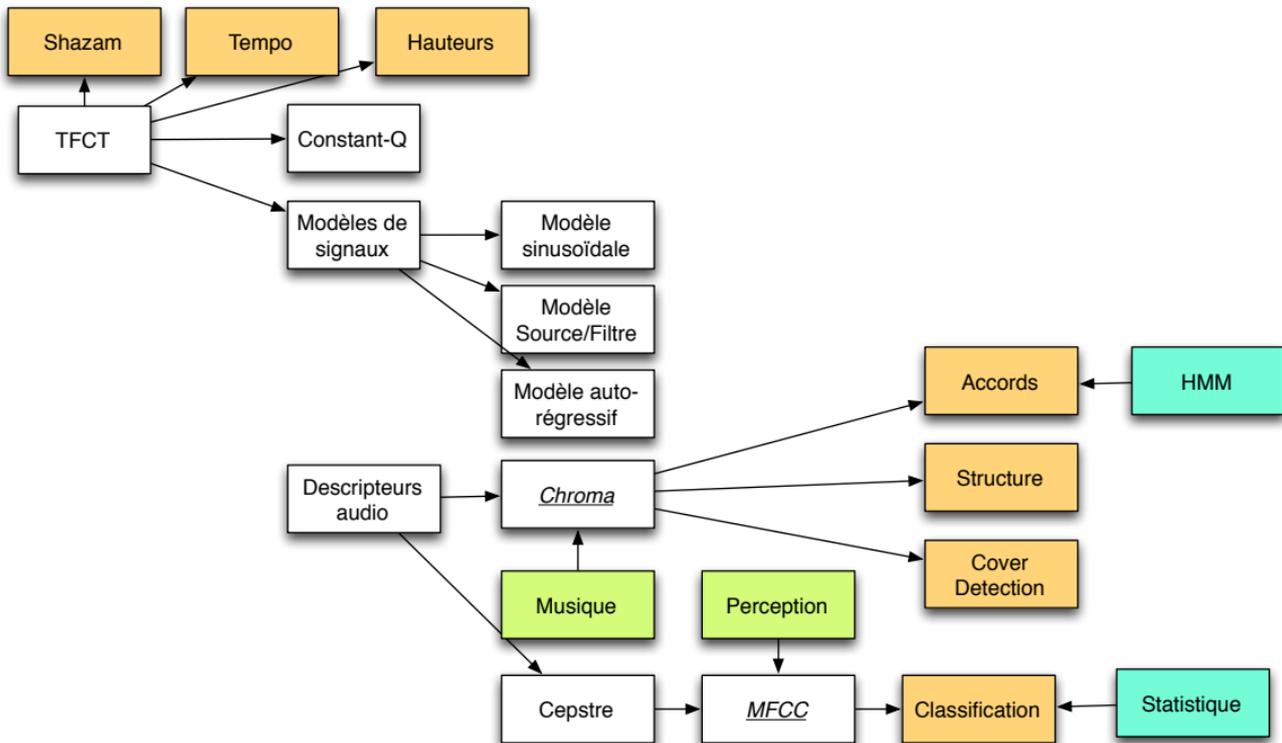
7.3- Estimation d'accords

Estimation des accords par trame par modèle de Markov caché (HMM)



7- Classification Audio

7.3- Estimation d'accords



7- Classification Audio

7.4- Détection de cover-version

Cover-version ?

- Une reprise
- "Let it be" par The Beatles, Aretha Franklin Joan Baez, ...

Caractéristiques d'une cover-version :

- généralement la même suite harmonique
 - même suite d'accords,
 - même mélodie
- éventuellement transposée

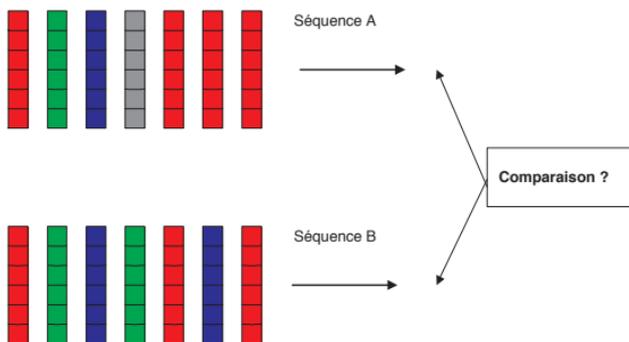
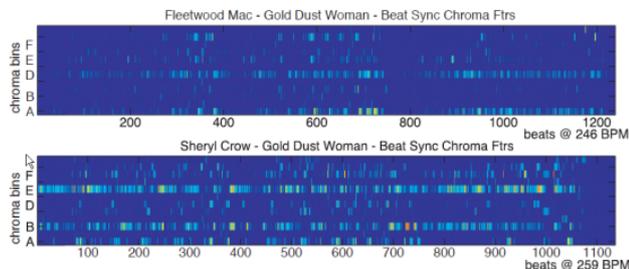
Titre	Artiste	Album	D.	Pop.	
Let It Be	▼ The Beatles Recovered Band	30 Beatles Top Hits	03:50		□
Let It Be	▼ The Hit Co., The Tribute Co.	A Tribute to the Beatles: The Lat...	03:42		□
Let It Be	▼ Labrinth	Let It Be	03:05		□
Let It Be Me	▼ Ray LaMontagne	Gossip In The Grain	04:41		□
Let It Be - The Beatles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

7- Classification Audio

7.4- Détection de cover-version

Méthode

- Chaque morceau est représenté par la séquence temporelle de ses chromas/PCP : $C(c, n)$
 - c est le chroma (pitch-class) et
 - n est le temps
- Pour une collection de morceaux, on compare les morceaux deux à deux
- Comparaison de deux morceaux A et B
 - Calcul du coût pour aligner la séquence de chroma $C_A(c, n)$ et $C_B(c, n)$
 - Si le coût d'alignement est faible, il s'agit vraisemblablement d'une cover ou ... d'un plagia
 - Technique utilisée :
 - Alignement Dynamique du Temps

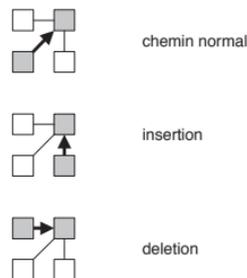
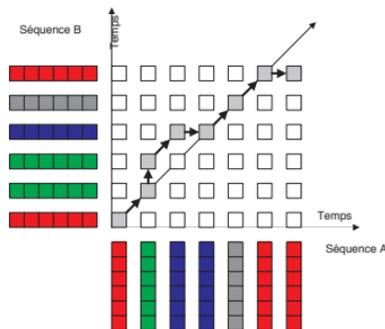


7- Classification Audio

7.4- Détection de cover-version

Alignement de deux séquences A et B

- Hypothèse :
 - temps de début et de fin en correspondances
- Méthode
 - Parcours progressif de tous les points (a, b) de la matrice d'alignement
 - En un point (a, b) , recherche du meilleur point précédent (de coût local minimal) parmi :
 - $(a - 1, b - 1) \rightarrow (a, b)$: chemin normal
 - $(a, b - 1) \rightarrow (a, b)$: insertion d'un élément de B
 - $(a - 1, b) \rightarrow (a, b)$: deletion d'un élément de B
 - Coût **local** cumulé associé à (a, b) =
 - coût (cumulé) du point précédent
 - + coût du chemin (défavorable si ins. ou del.)
 - + coût intrinsèque : $d(C_A(c, a), C_B(c, b))$
 - Coût **global** de l'alignement de A et B
 - = coût cumulé en $(a=\text{end}, b=\text{end})$

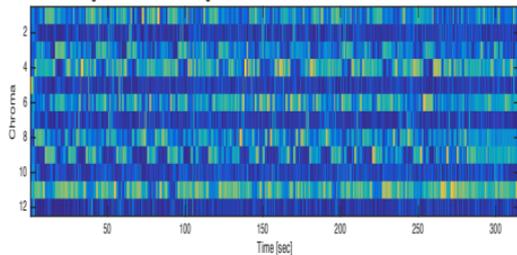


7- Classification Audio

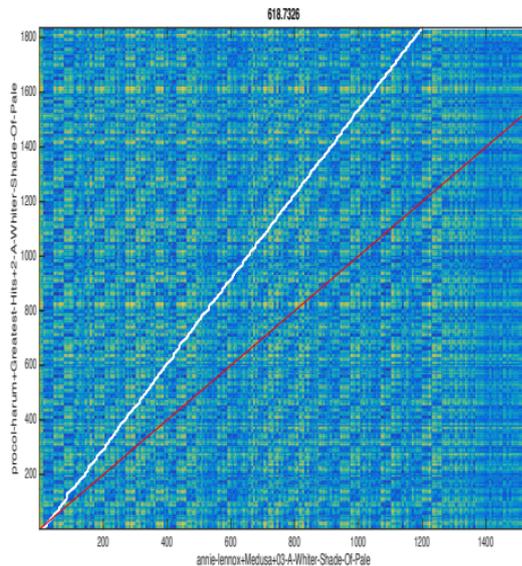
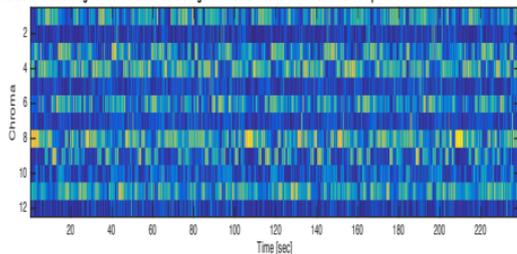
7.4- Détection de cover-version

Exemple d'alignement de coût faible

/Users/peeters/sound-collection/global-cover/201302-Cover/Songs-DanEllis-covers32k/A-Whiter-Shade-Of-Pale/annie-lennox+Medusa+03-A-Whiter-Shade-Of-Pale.mpt



/Users/peeters/sound-collection/global-cover/201302-Cover/Songs-DanEllis-covers32k/A-Whiter-Shade-Of-Pale/procol-harum+Greatest-Hits+2-A-Whiter-Shade-Of-Pale.m

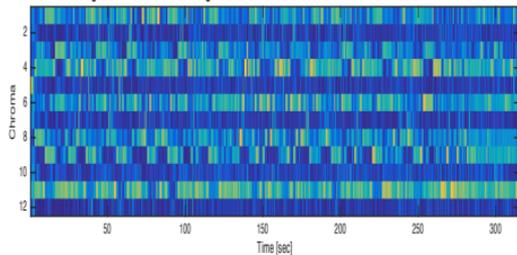


7- Classification Audio

7.4- Détection de cover-version

Exemple d'alignement de coût élevé

/Users/peeters-sound-collection/global-cover-201302-CoverSongs-DanEllis-covers32k/A-Whiter-Shade-Of-Pale/annie-lennox+Medusa+03-A-Whiter-Shade-Of-Pale.mp3



/Users/peeters-sound-collection/global-cover-201302-CoverSongs-DanEllis-covers32k/Abracadabra/steve-miller-band+Steve-Miller-Band-Live+09-Abracadabra.mp3

