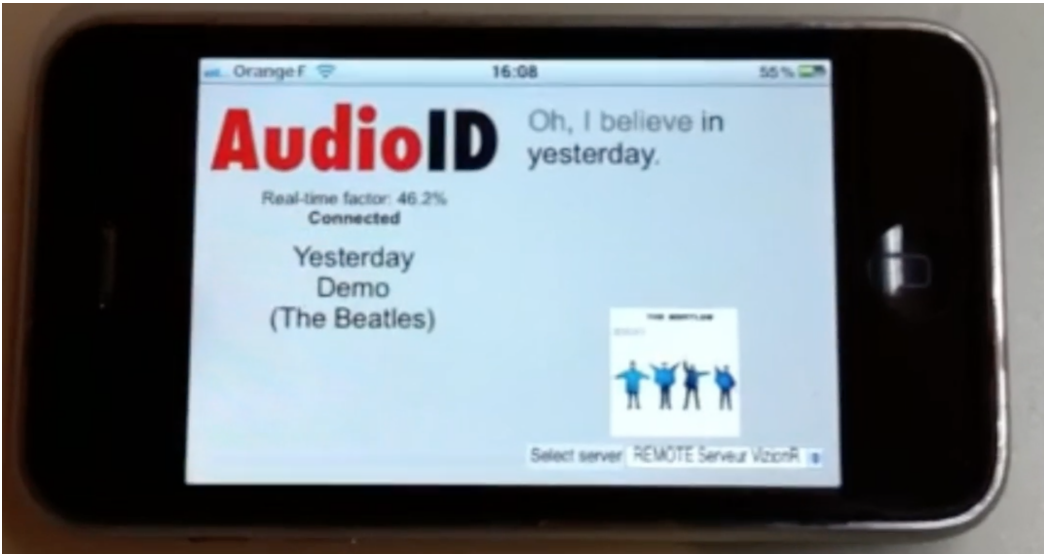
ENSEA 2ème AuPAM (2016-2017)
Traitement du signal audio musical, descripteurs et estimation

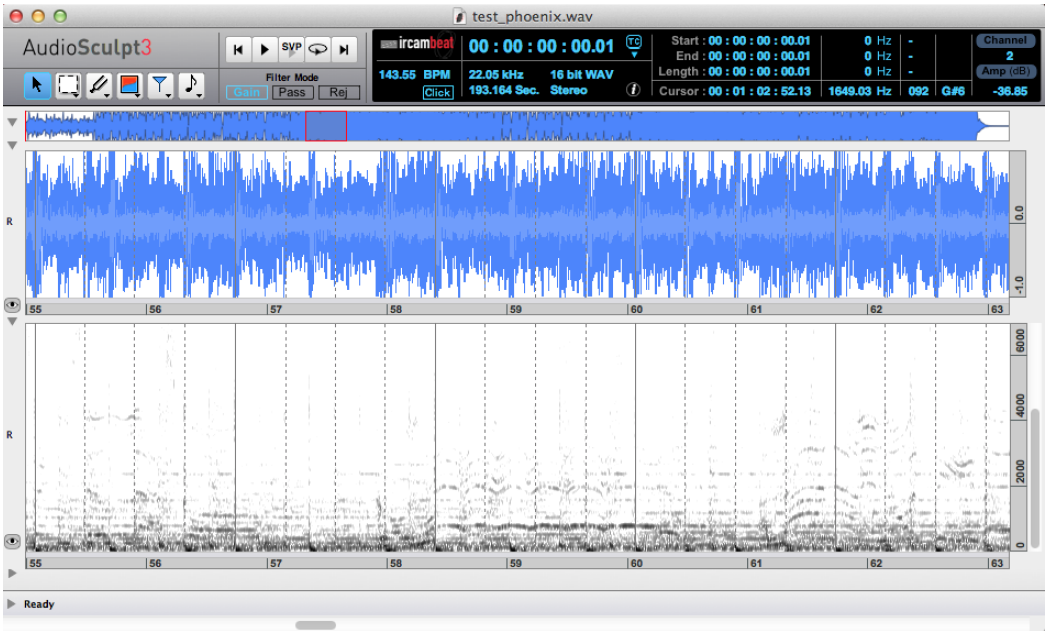
Geoffroy.Peeters@ircam.fr
UMR SMTS 9912 (IRCAM CNRS UPMC)

source : Gaël Richard

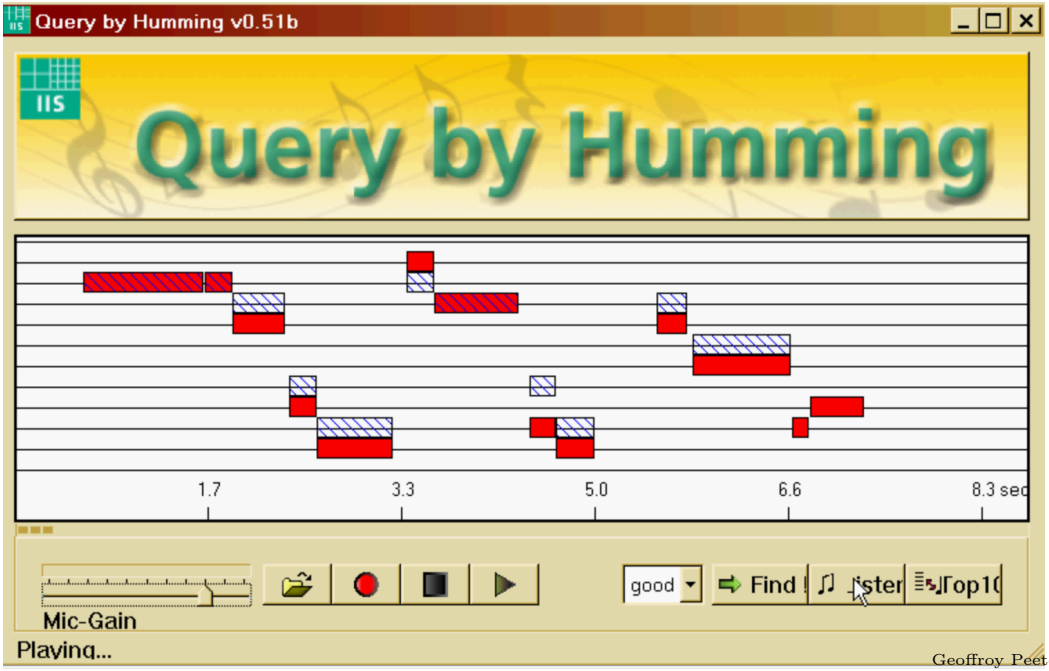
- Identification audio
  - recherche de doublons, gestion de copyright, attacher des méta données à une instance d'un morceau



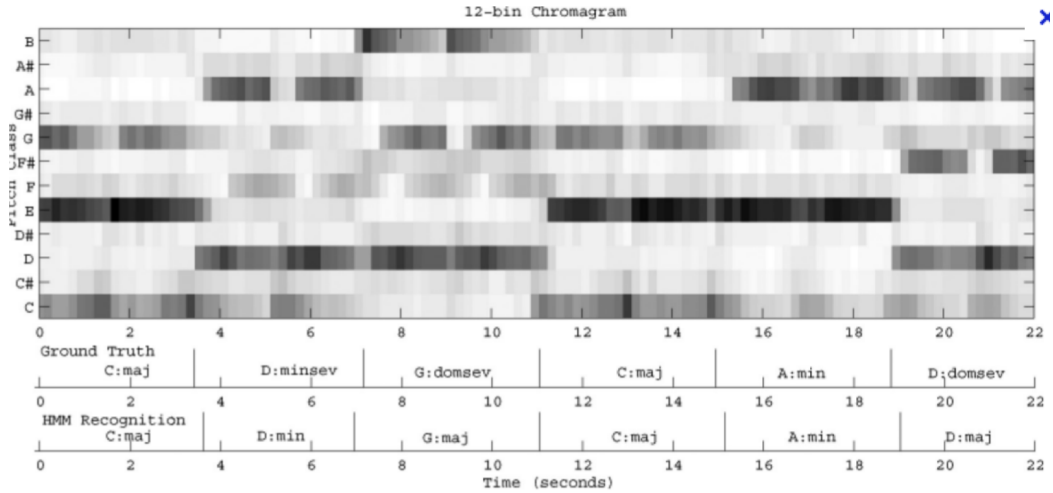
- Estimation du tempo, de la position des temps/ premier-temps
  - DJing, mainpulation du contenu (add swing ...)



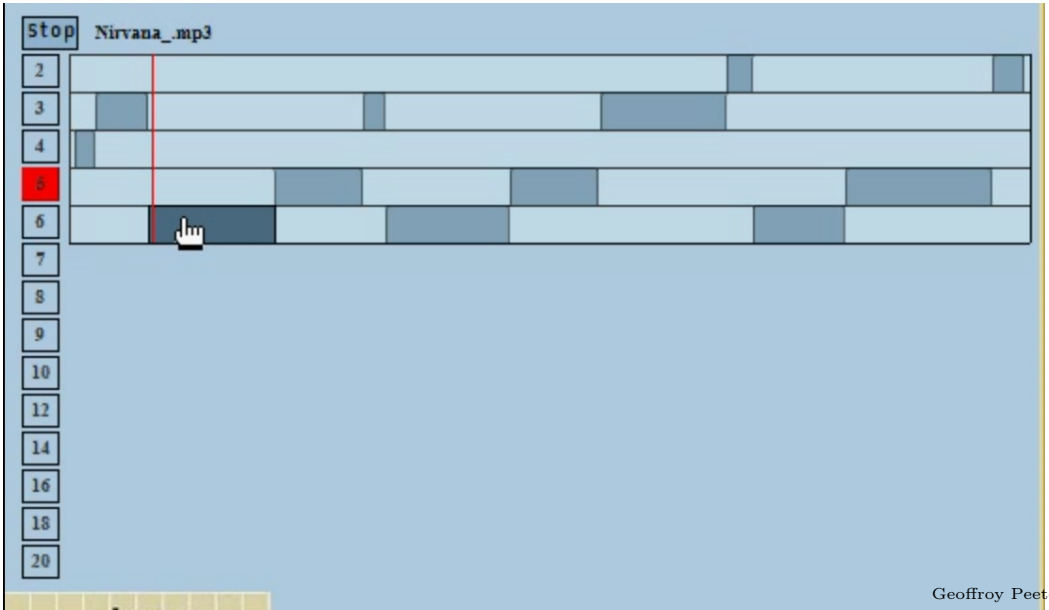
- Nouveaux modes de recherche :
  - par chantonnement/ sifflement



- Estimation des accords
  - obtenir des guitar-tab automatiquement



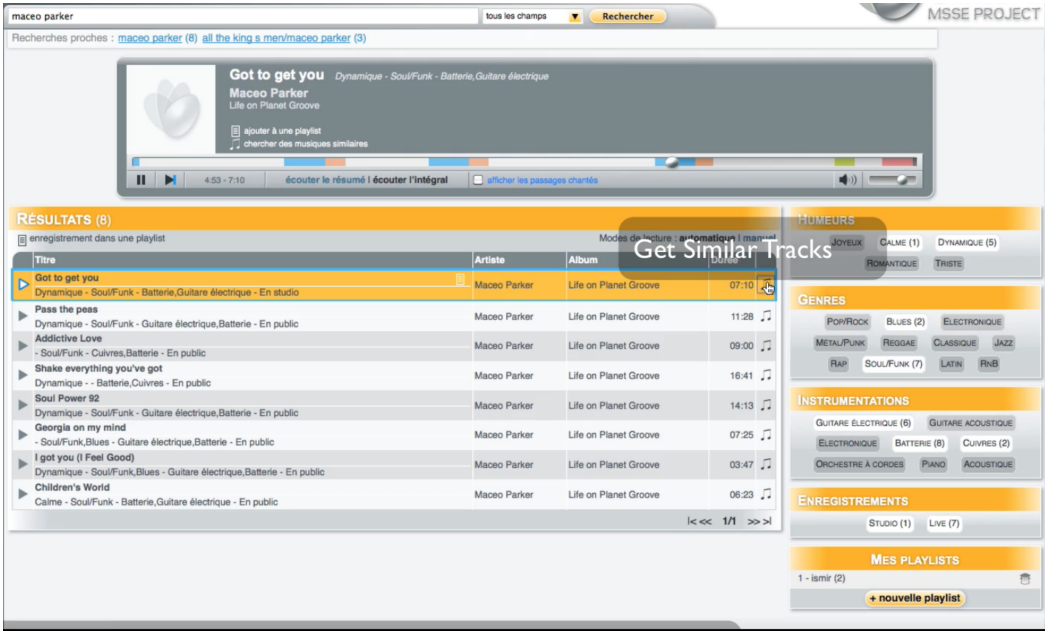
- Navigation à l'intérieur d'un morceau de musique par couplet/refrain
  - Génération automatique de résumé audio
- Dé-linéarisation d'un flux audio :
  - segmentation de flux radio, télé et étiquetage des parties

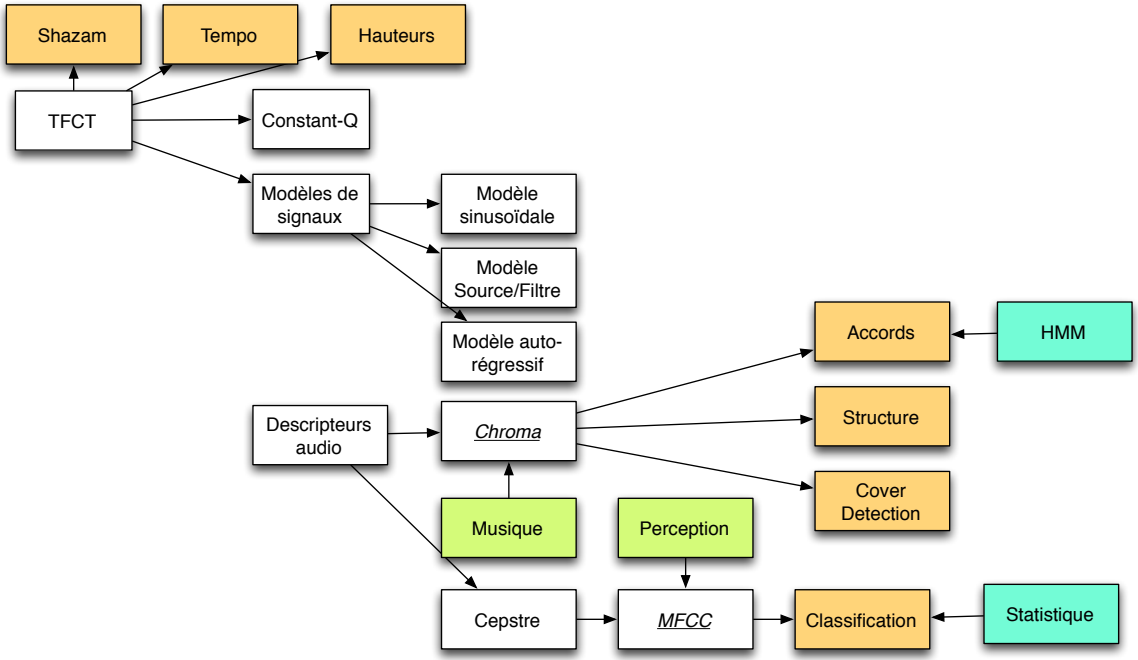


- Détection des cover, reprises ou ... des plagias

Titre	Artiste	Album	D.	Pop.
Let It Be	The Beatles Recovered Band	30 Beatles Top Hits	03:50	
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	

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

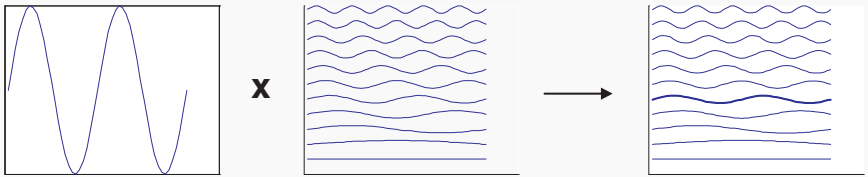



Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

Transformée de Fourier (temps et fréquences continus)

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

- Variables :
  - t est le **temps**
  - $\omega = 2\pi f$ les **fréquences continues** exprimées en radian,
  - $\exp(j2\pi f t) = \cos(2\pi f t) + j \cdot \sin(2\pi f t)$.
- 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



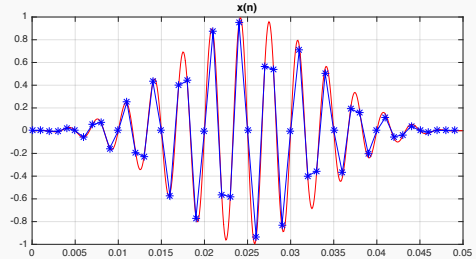
Propriété de la Transformée de Fourier (temps et fréquences continus)

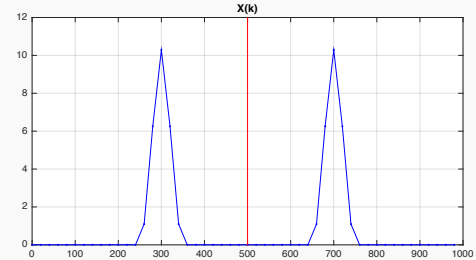
Propriétés	$x(t)$	$X(f)$
Similitude	$x(at)$	$\frac{1}{\|a\|}X(\frac{f}{\|a\|})$
Linéarité	$ax(t) + by(t)$	$aX(f) + bY(f)$
Translation	$x(t - t_0)$	$X(f)\exp(-j2\pi f t_0)$
Modulation	$x(t)\exp(j2\pi f_0 t)$	$X(f - f0)$
Convolution	$x(t) \circledast y(t)$	$X(f)Y(f)$
Produit	$x(t)y(t)$	$X(f) \circledast Y(f)$
Parité	réelle paire	réelle paire
	réelle impaire	imaginaire paire
	imaginaire paire	imaginaire paire
	imaginaire impaire	réelle impaire
	complexe paire	complexe paire
	complexe impaire	complexe impaire
	réelle	$X(f) = X^\star(-f)$
		$\Re(X(f))$ est paire
		$\Im(X(f))$ est impaire
	$x^\star(t)$	$X^\star(f)$

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

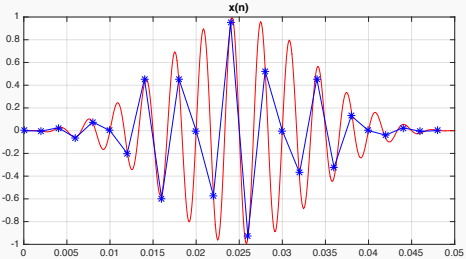
- Variables :
  - m le numéro d'**échantillon**
  - k les **fréquences discrètes**
- Fréquence d'échantillonnage (sampling rate) sr
  - sr définit à quelle fréquence le signal temporel va être échantillonné
  - Exemple :
    - Compact Disc $sr = 44100$ Hz
    - La distance temporelle entre deux échantillons (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** : $f_{Nyquist} = \frac{sr}{2} > f_{\max}$

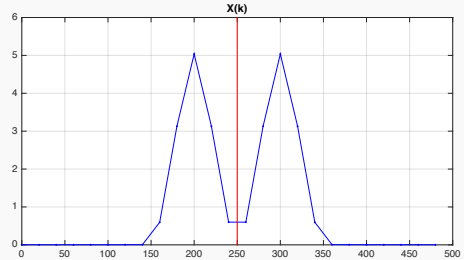
$f_{\max} = 300,\ sr = 1000$





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



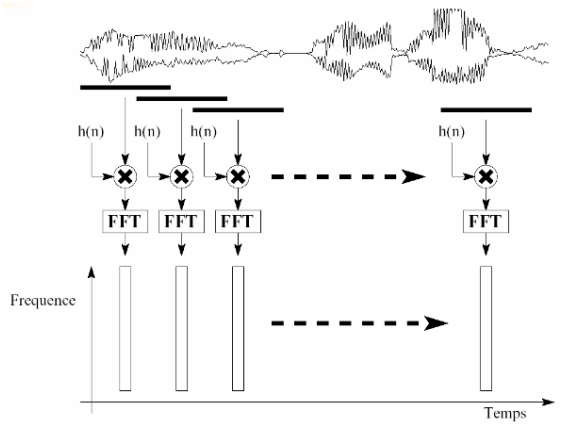


$$X(k,n) = \sum_{m=0}^{N-1} x(m)w(n-m)e^{-j2\pi\frac{k}{N}m} \qquad \forall k \in [0,N]$$

- 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 $\pm$ 40ms
- TFCT = suite d'analyses de Fourier sur des durées de $\pm$ 40ms
  - = analyse à Court Terme ("trames/frames" en vidéo)



source : Jean Laroche

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

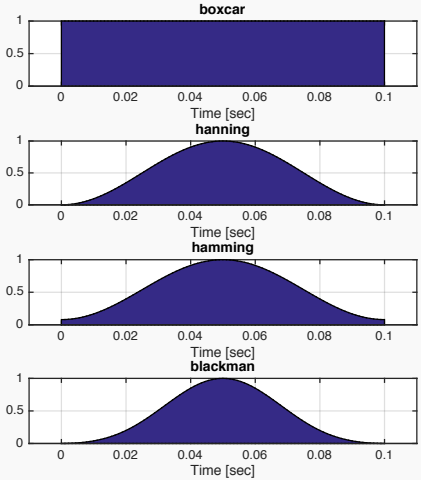
Fenêtre de pondération $w(t)$

- $x(t) \cdot w(t) \rightleftharpoons X(f) \circledast W(f)$
  - $w(t)$ 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
  - $w(n) = a_0 - a_1\cos(\frac{2\pi n}{N-1}) + a_2\cos(\frac{2\pi n}{N-1})$
  - $Bw = 2.35$

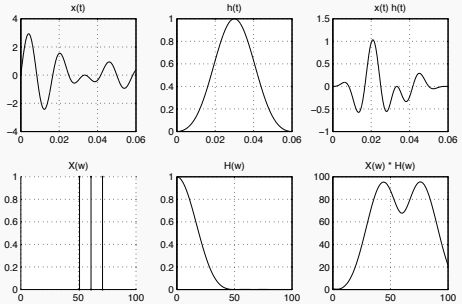
Influence du **type** de la fonction



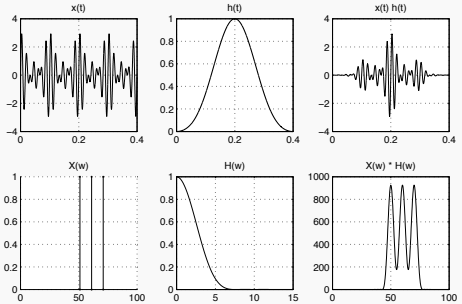
Choix de la **longueur temporelle** L :

- 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.

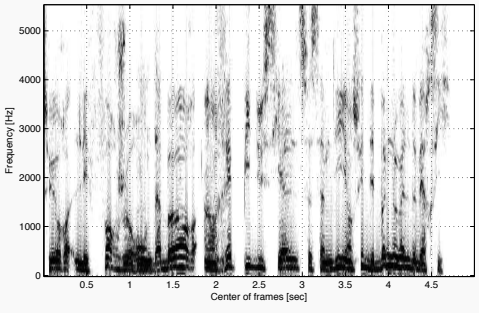
Influence de la **longueur temporelle** L ($L = 0.06$s.)



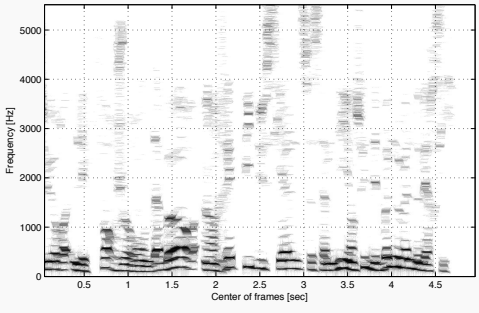
Influence de la **longueur temporelle** L ($L = 0.4$s.)



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

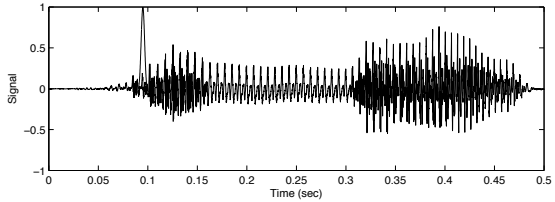


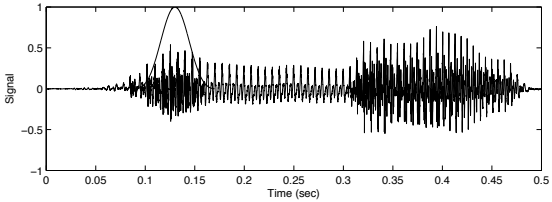
Influence de la **longueur temporelle** L ($L = 0.1$s.)

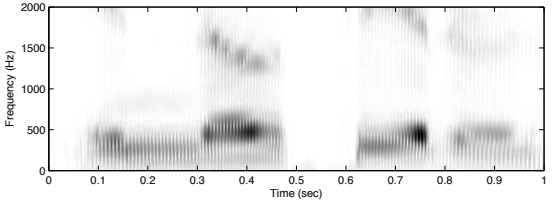


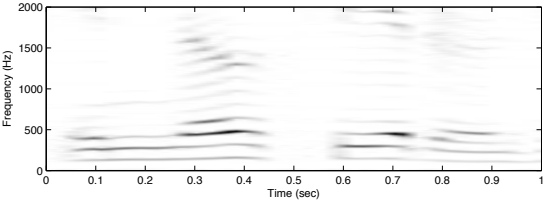
Paradoxe temps/ fréquence

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

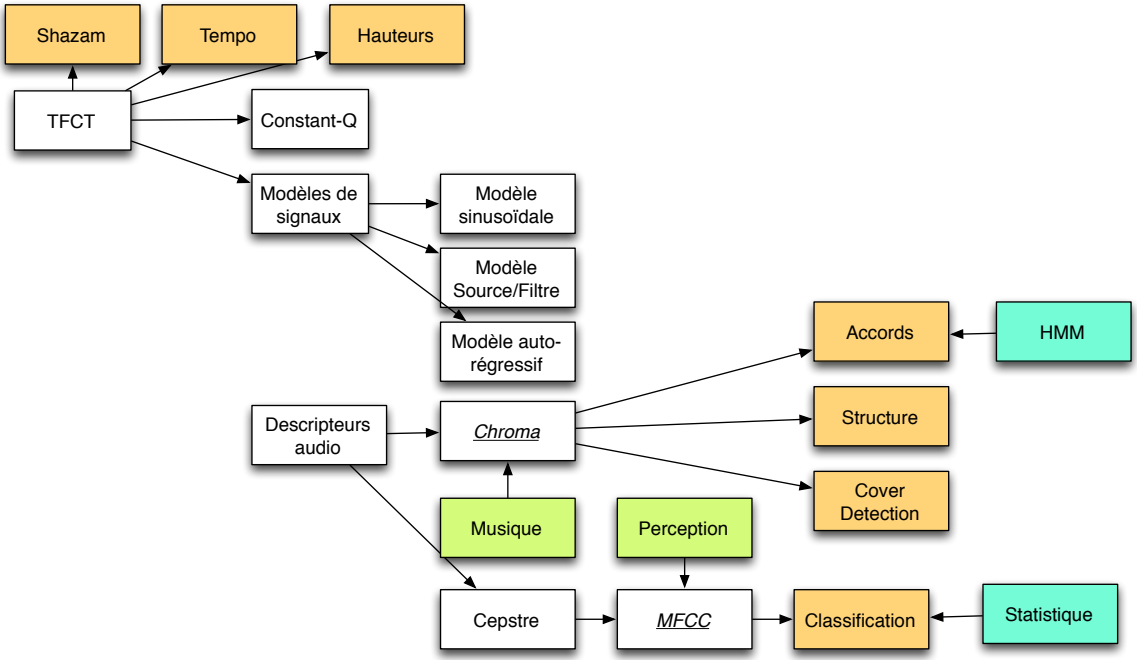








- Comme résoudre ce problème ?
  - Utiliser d'autres tranformées que celle de Fourier

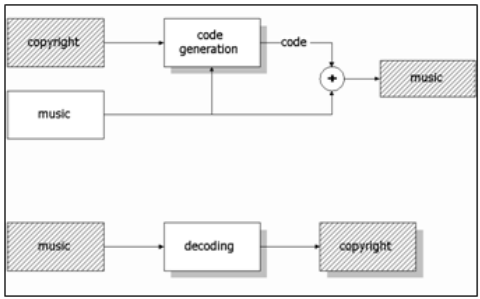

Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

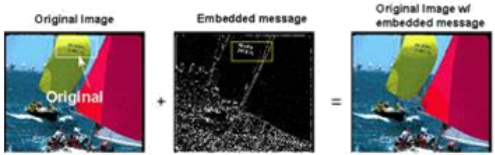
Identification audio

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

Méthode du Watermarking

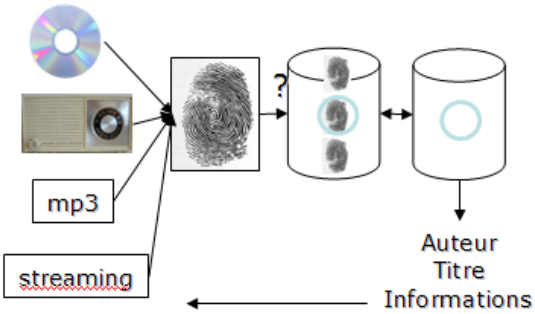
- Codage :
  - introduction d'un code identifiant robuste mais inaudible dans le signal sonore
- Décodage :
  - pour un nouveau signal : extraction du code (si il est présent) et recherche de ce code dans une base de données





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

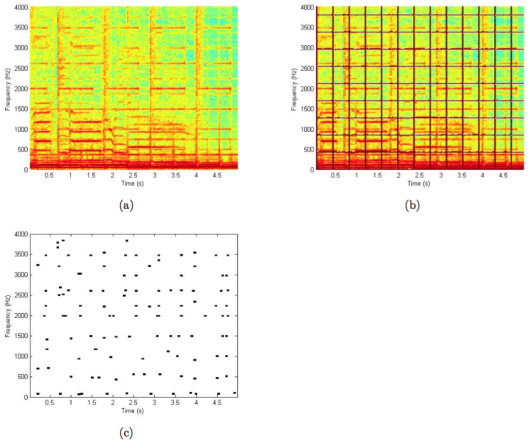


Algorithme de Fingerprint de Shazam

[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.]

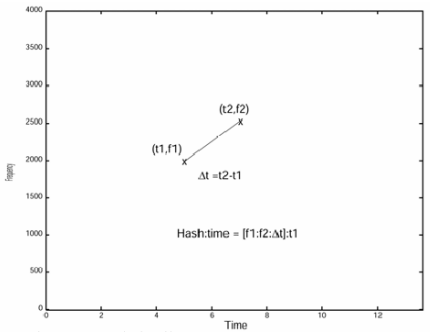
- Extraction de points saillants dans le plan temps/fréquence
  - Calcul du spectrogramme
    - fenêtre de Hamming, L=64 ms, S=32 ms
  - Dans chaque pavé du spectrogramme (Δt=0.4 s, Δf) :
    - détection du maximum → valeur = 1
  - = "constellation points"



source : Sebastien Fenet

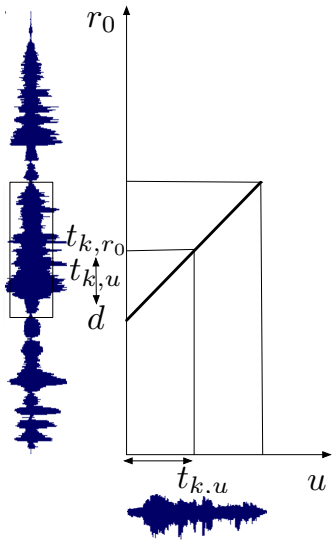
A) Partie stockage de signature

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



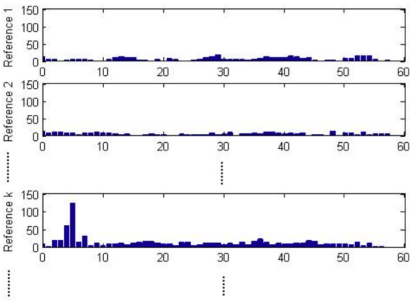
B) Partie matching de signature

- 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

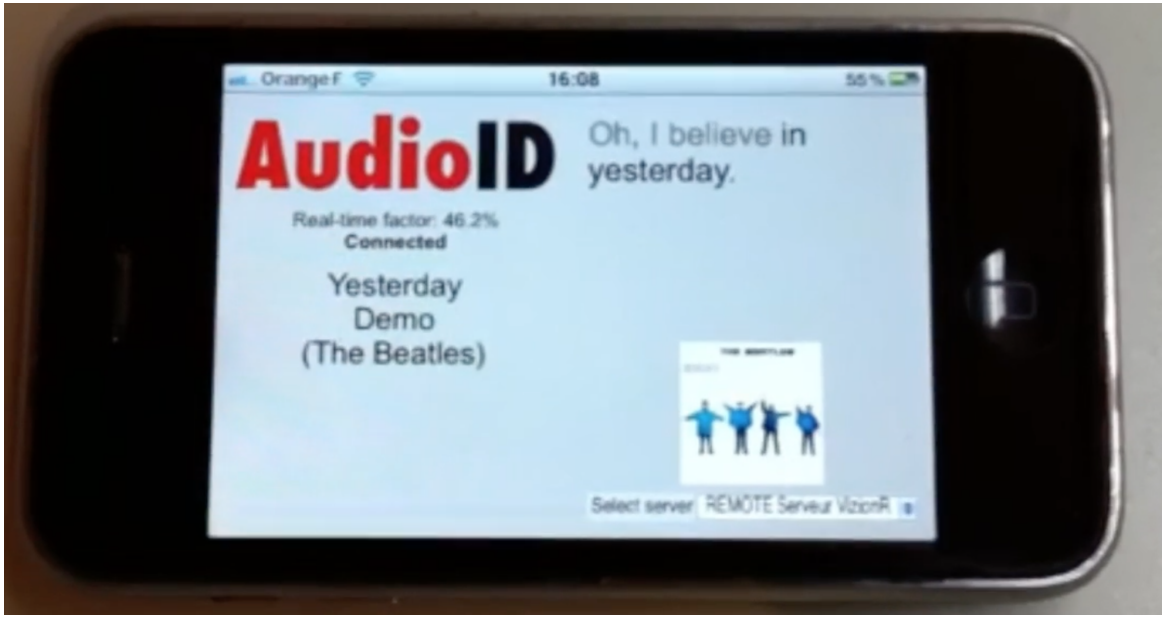


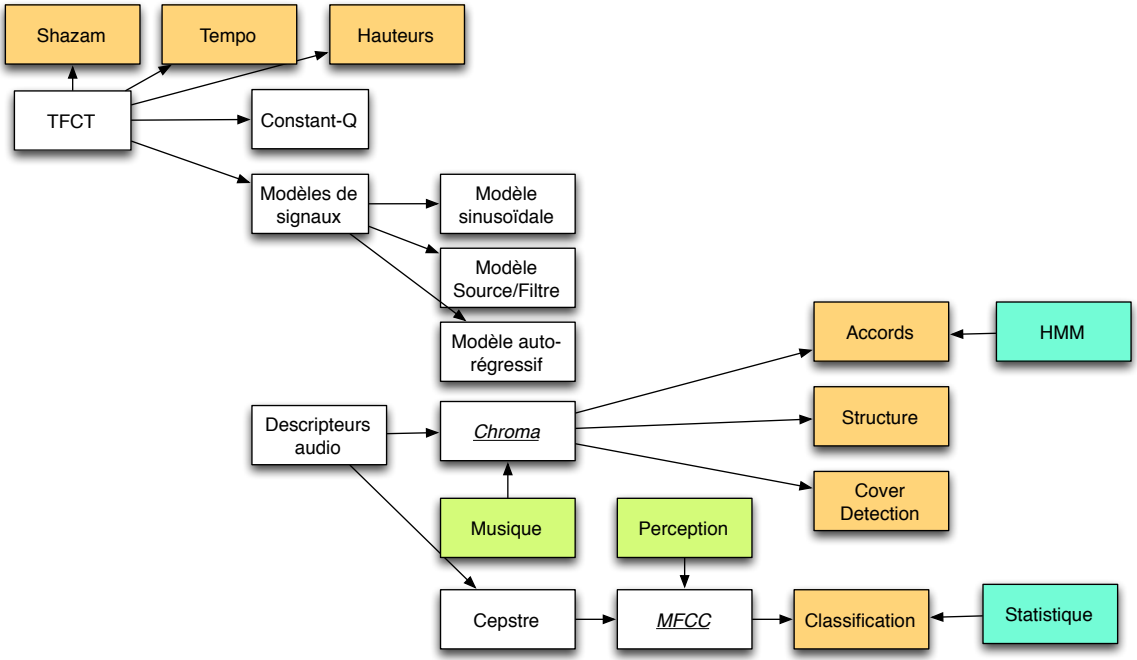
B) Partie matching de signature

- 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 histogram donne le point de démarrage d dans le signal de référence



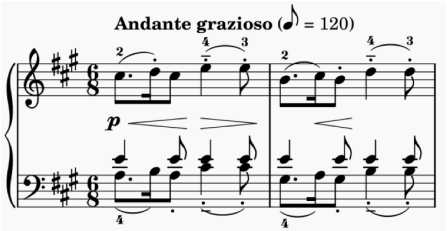
source : Sebastien Fenet

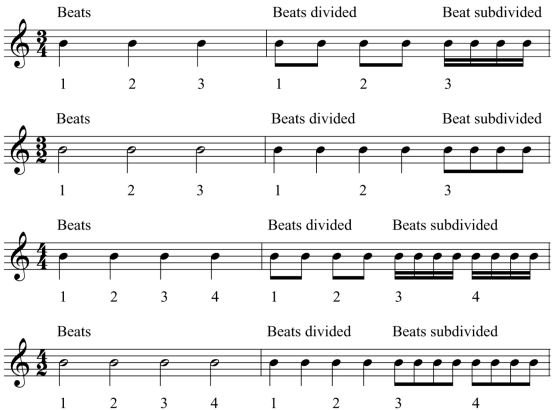

Orange F
16:08
55 %
AudioID
Oh, I believe in yesterday.
Real-time factor: 46.2%
Connected
Yesterday
Demo
(The Beatles)
Select server
REMOTE Serveur VizcomR


Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

Rhythme ?

- 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



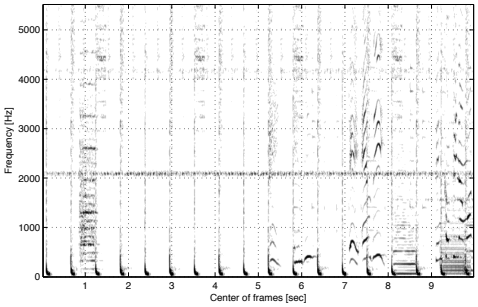


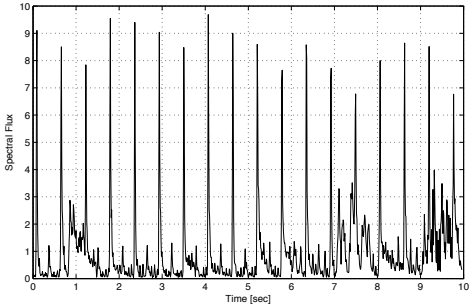
Estimation du tempo ?

- Détecter la périodicité des évènements dans le signal audio

Détection des évènements ?

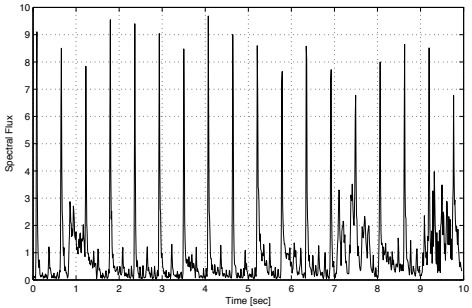
- Début des évènements = onsets
- Méthode 1
  - détecter les maxima locaux de la fonction d'énergie du signal
    - $ener(m) = \sum_k X(k,m)^2$
- Méthode 2
  - détecter les maxima locaux du flux spectral :
    - $flux(m) = \sum_k \mathrm{HWR}[X(k,m) - X(k,m-1)]$
    - $\mathrm{HWR}(x) = x \quad si \quad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





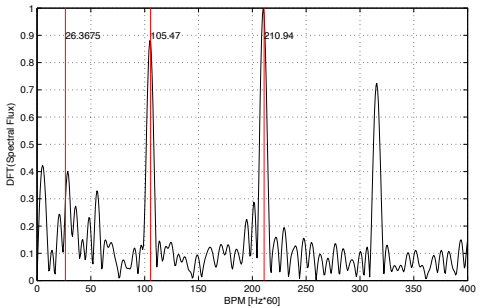
Périodicité des évènements ?

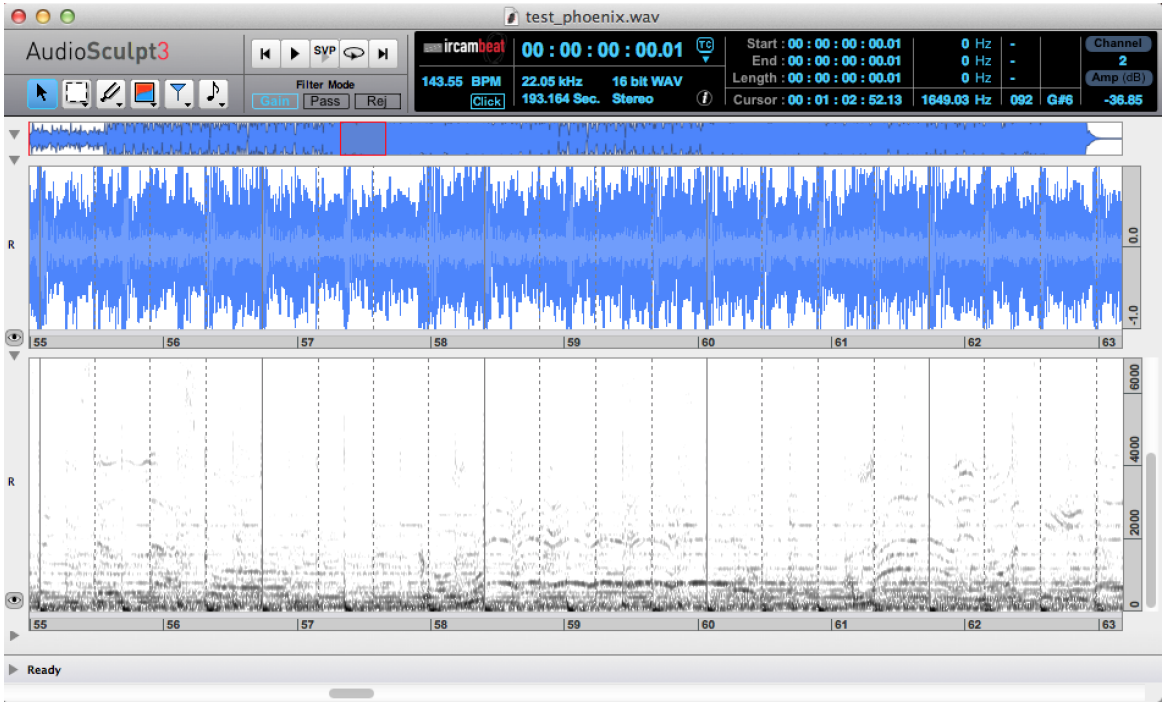
- Calcul de la transformée de Fourier (DFT) du flux spectral :
  - $FLUX(k) = \sum_{n=0}^{N-1} flux(n) \cdot \exp(j2\pi \frac{k}{N} n), \forall k$
- Calcul de la fonction d'auto-correlélation du flux spectral

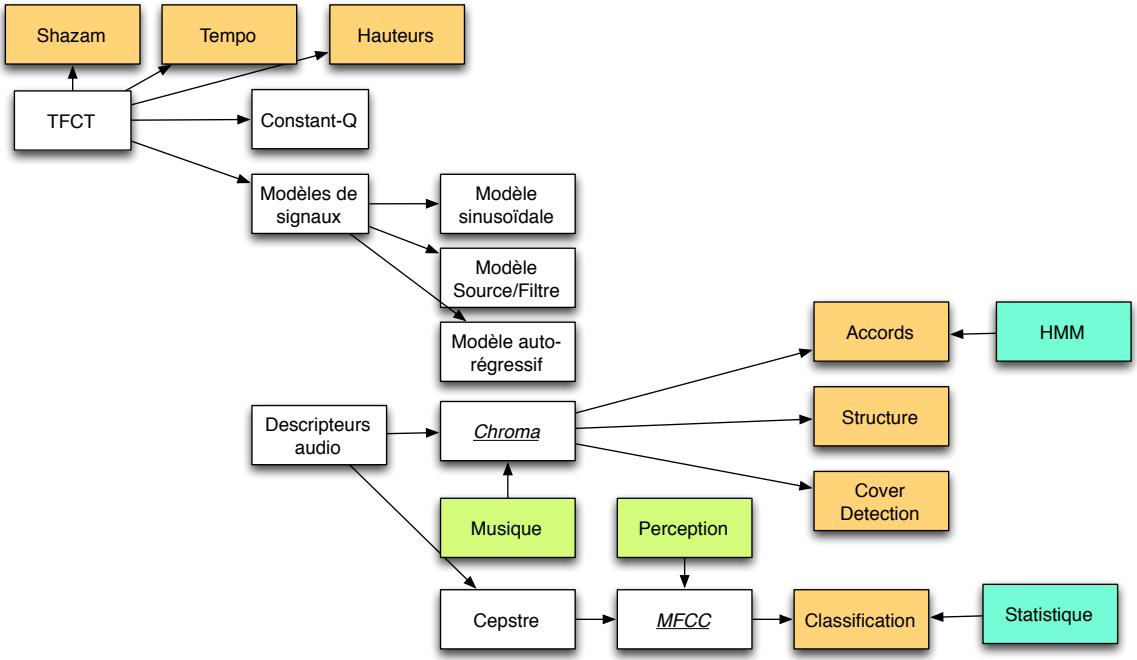


Estimation du tempo ?

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

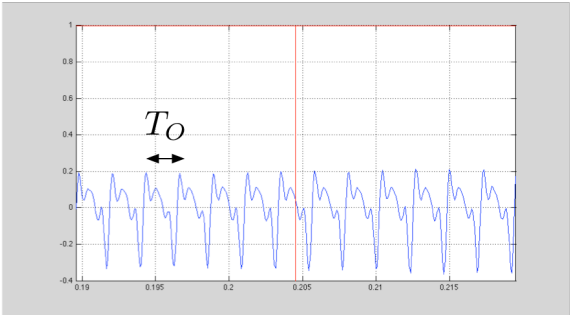


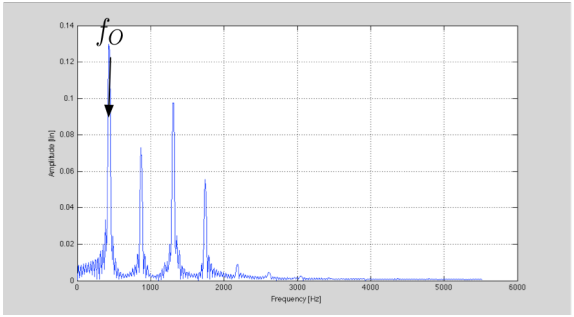

test_phoenix.wav
AudioSculpt3
ircambeat
00 : 00 : 00 : 00.01
143.55 BPM
22.05 kHz
16 bit WAV
193.164 Sec.
Stereo
Click
Filter Mode
Gain
Pass
Rej
Start : 00 : 00 : 00 : 00.01
End : 00 : 00 : 00 : 00.01
Length : 00 : 00 : 00 : 00.01
Cursor : 00 : 01 : 02 : 52.13
Channel
2
Amp (dB)
-36.85
1649.03 Hz
092
G#6
55
56
57
58
59
60
61
62
63
6000
4000
2000
0
Ready


Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

Période fondamentale T_0 ou fréquence fondamentale f_0

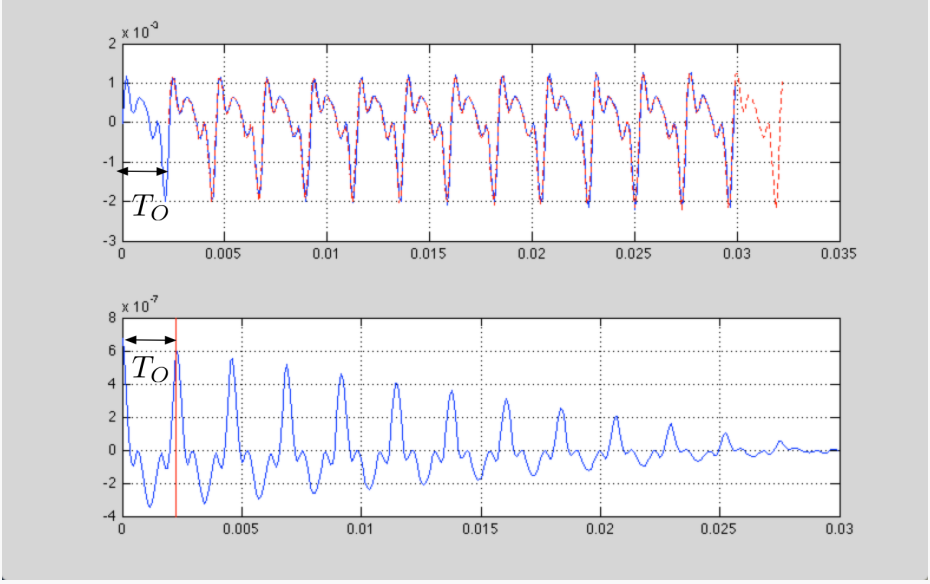
- f_0 : fréquence fondamentale en Hz
  - exemple La3/A4= 440Hz
- $T_0 = \frac{1}{f0}$: période fondamentale en secondes
  - exemple La3/A4 = 0.0023s.





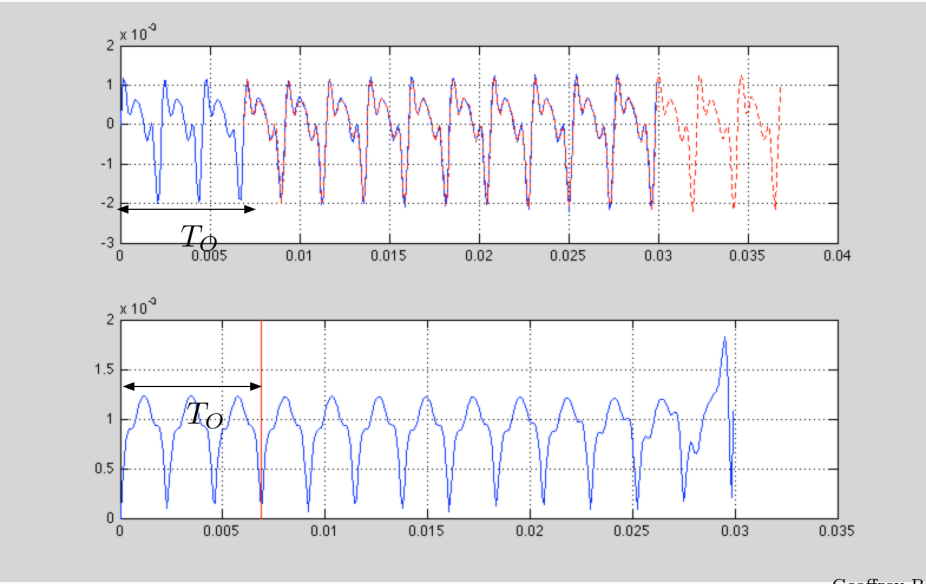
Méthode de l'autocorrélation

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

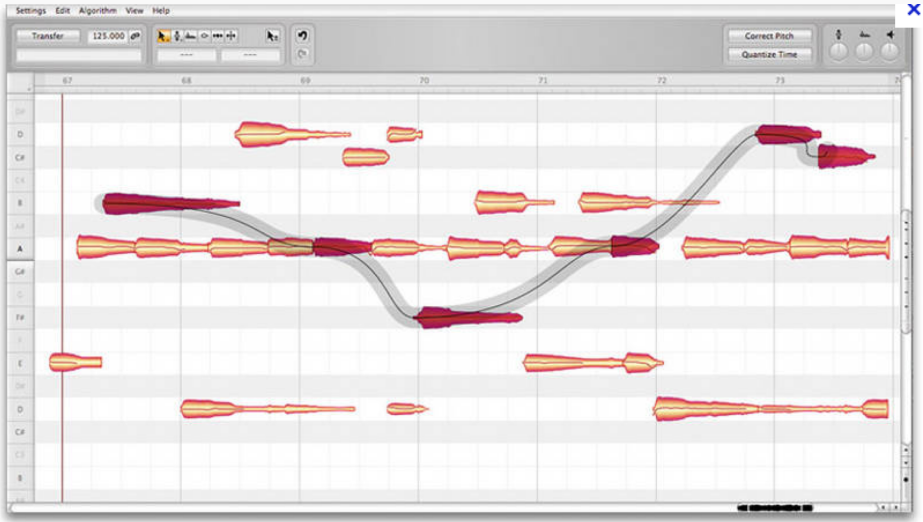


Méthode de l'Average Mean Difference Function (AMDF)

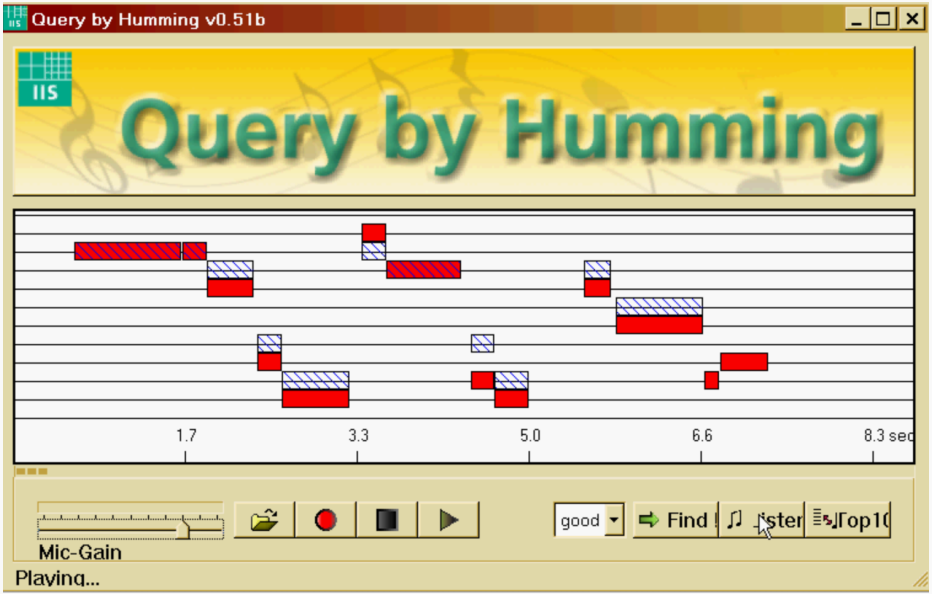
- $AMDF(m) = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x(n) - x(n+m)|$
- $AMFT(m) = 0$ si $m = T_0$

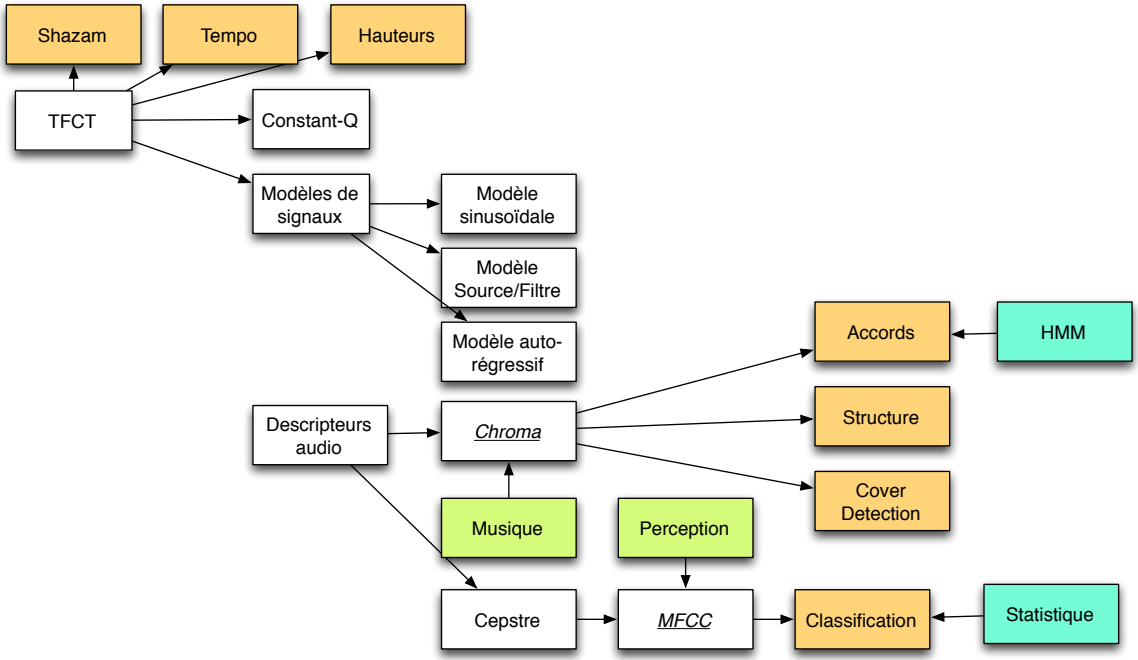


Application : Edition Audio



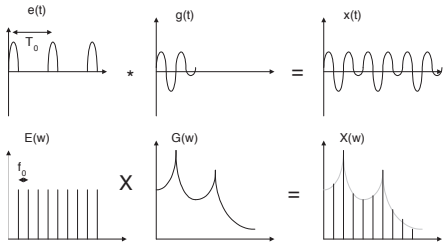
Application : Query by Humming




Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

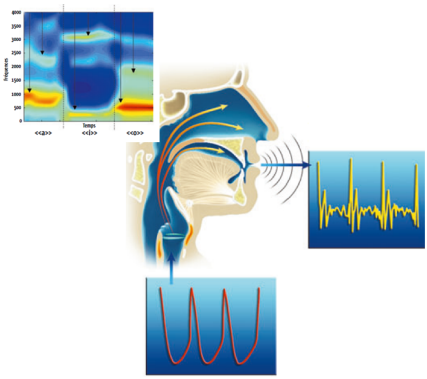
Modèle source/ filtre

- 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 $g(t)$:
  - $x(t) = e(t) \circledast g(t)$
- Modélisation fréquentielle
  - la multiplication de la TF du signal d'excitation (source) par la TF du filtre.
  - $X(\omega) = E(\omega) \cdot G(\omega)$



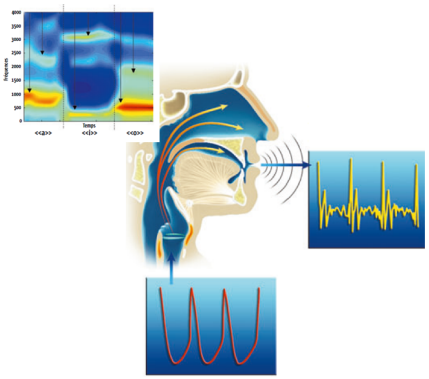
Production du signal vocal

- 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

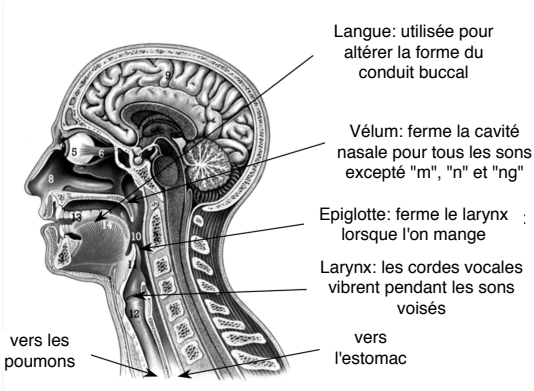


Production du signal vocal

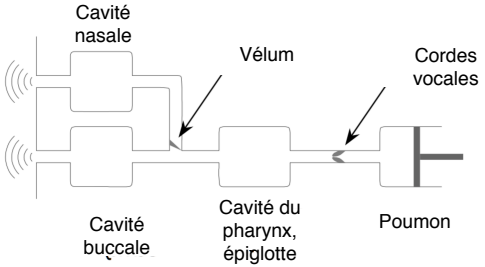
- Ouverture/fermeture périodique des cordes vocales
  - détermine la hauteur
    - Hauteur de 100Hz ? pulses d'air sont espacés de $T_0 = 1/f_0 = 1/100 = 10ms$.
  - Appelé signal d'excitation (ou signal source), $e(t)$.
- Conduit bucco-nasal
  - créer les différentes voyelles pour une hauteur donnée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



Production du signal vocal

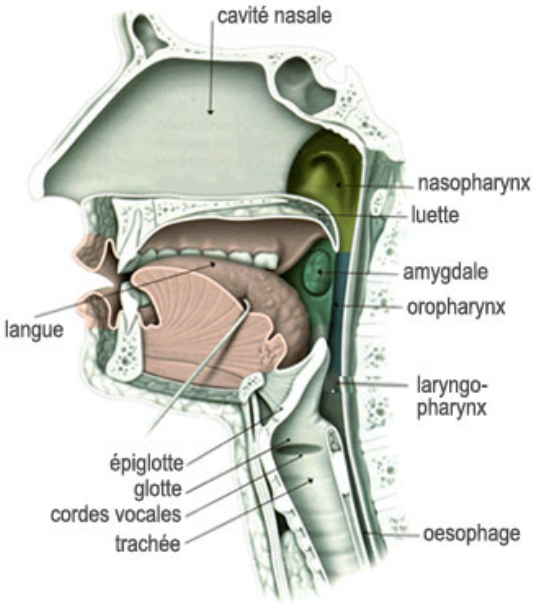


source : Mike Brookes

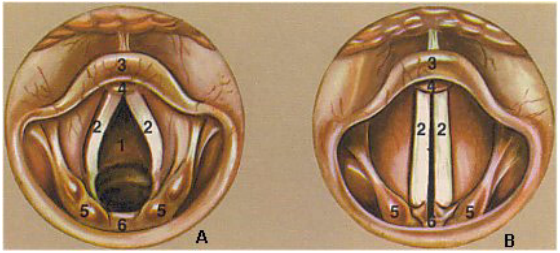


source : Mike Brookes

Production du signal vocal

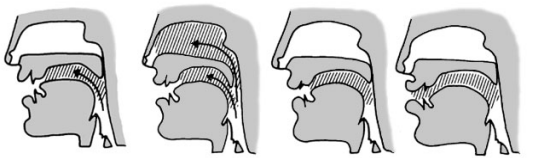


source : outilsrecherche.over-blog.com

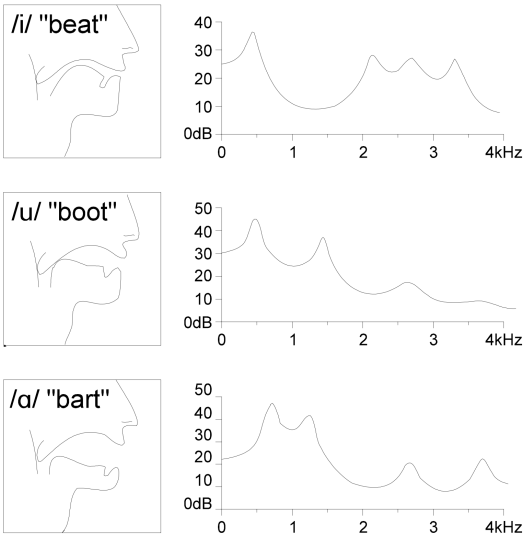


source : outilsrecherche.over-blog.com

Production du signal vocal

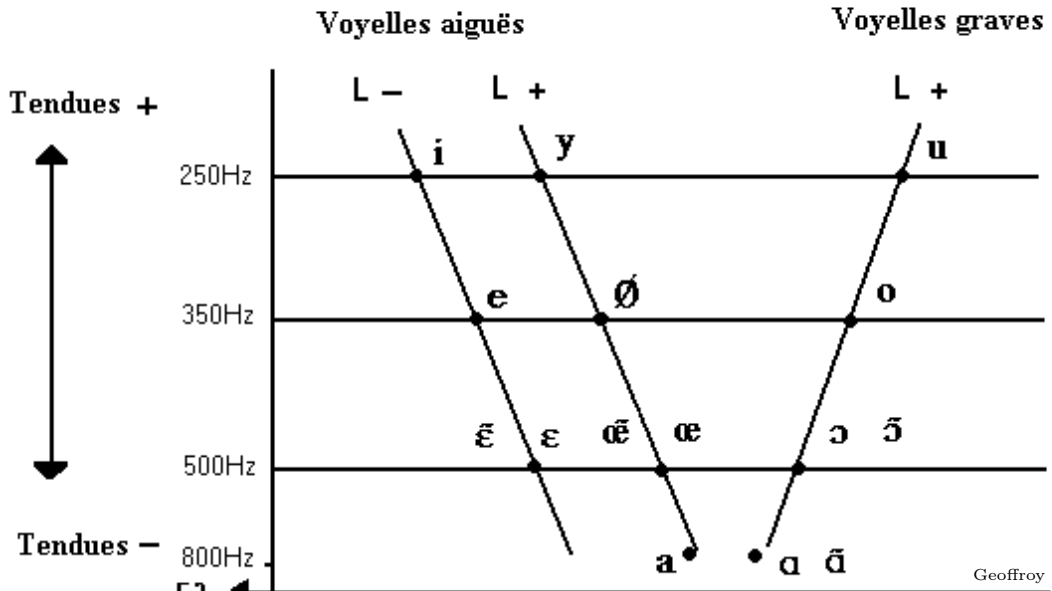


source : outilsrecherche.over-blog.com

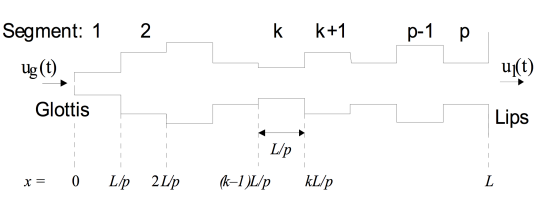


source : Mike Brookes

Fréquences des formants pour les différentes voyalles

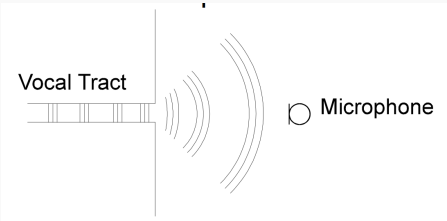


Représentation sous-forme de tube



source : Mike Brookes

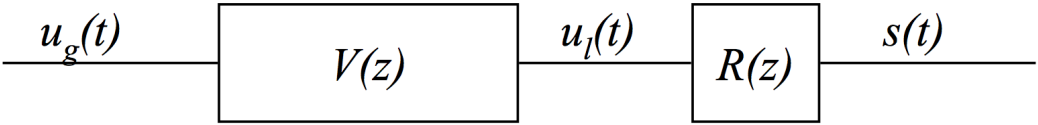
Radiation des lèvres



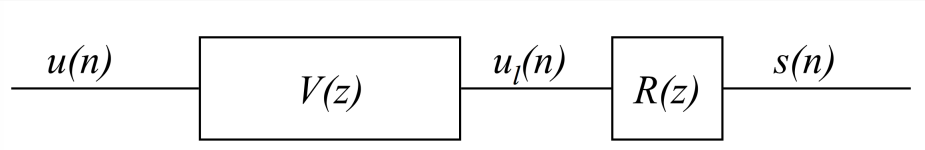
source : Mike Brookes

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

Système équivalent

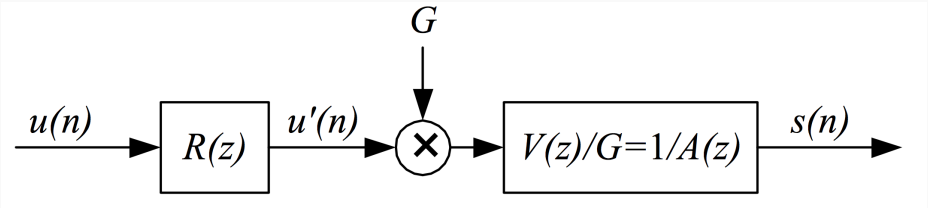


La prédiction linéaire, modèle auto-régressif



source : Mike Brookes

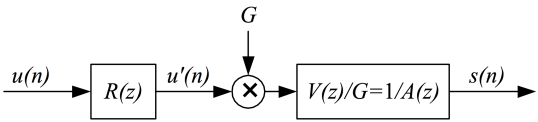
- Inversion de l'ordre de $V(z)$ et $R(z)$
  - puisque linéaire et $V(z)$ ne change pas significativement durant la réponse impulsionnelle de $R(z)$ et inversement



source : Mike Brookes

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



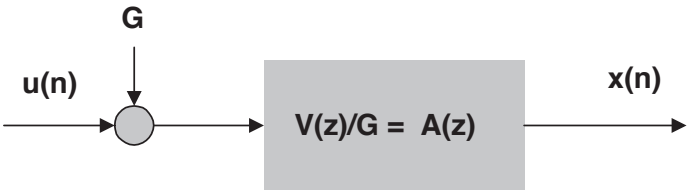
source : Mike Brookes

La prédiction linéaire, modèle auto-régressif

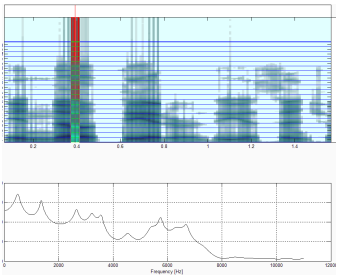
- Le signal à l'instant n peut être prédit à partir desinstants précédents

$$\begin{aligned} x(n) &= a_1 x(n-1) + a_2 x(n-2) + a_3 x(n-3) ... + a_P x(n-P) \\ &= \sum_{p=1}^{P} a_p x(n-p) + G \cdot u(n) \end{aligned} \tag{1}$$

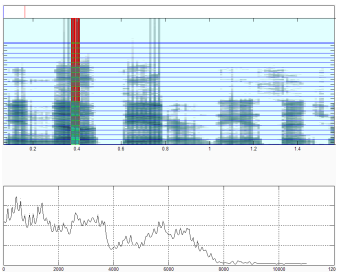
- Equivalent à passer le signal dans un filtre FIR tout-pôle :$V(z) = \frac{G}{1+\sum_{p=1}^{P} a_p z^{-p}}$
- Objectif de la prédiction linéaire ?
  - déterminer le filtre $V(z)$ (donc les résonances ou les formants dans le cas de la voix) à partir du signal $x(n)$ (signal de pression micro)



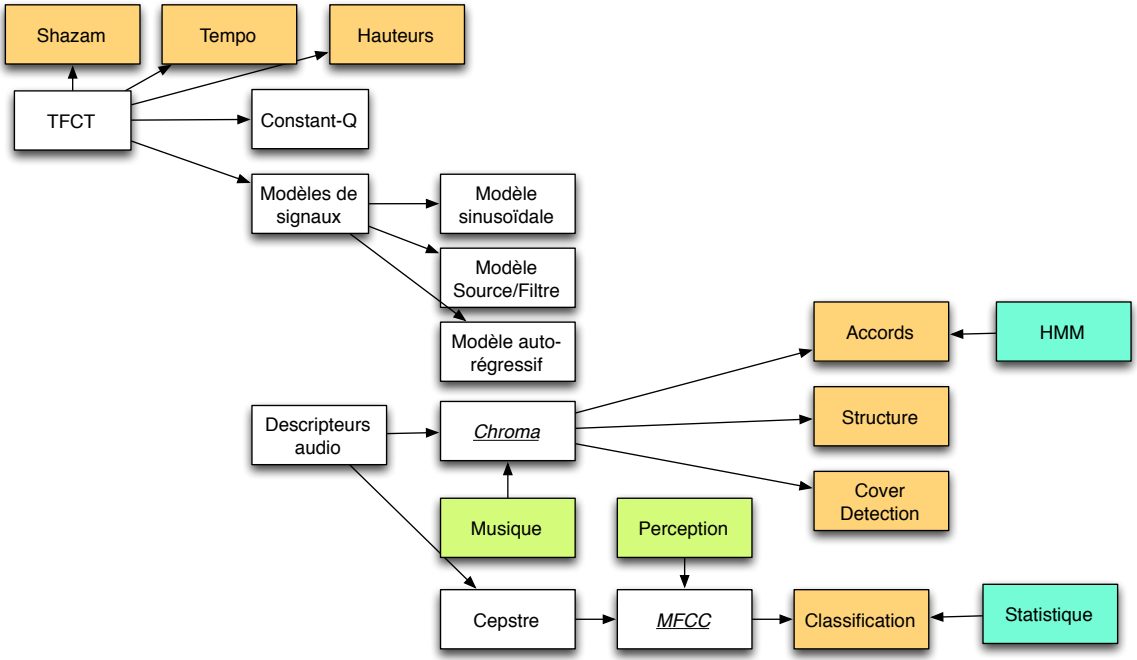
Choix du nombre de pôle P



$P = 40$



$P = 200$

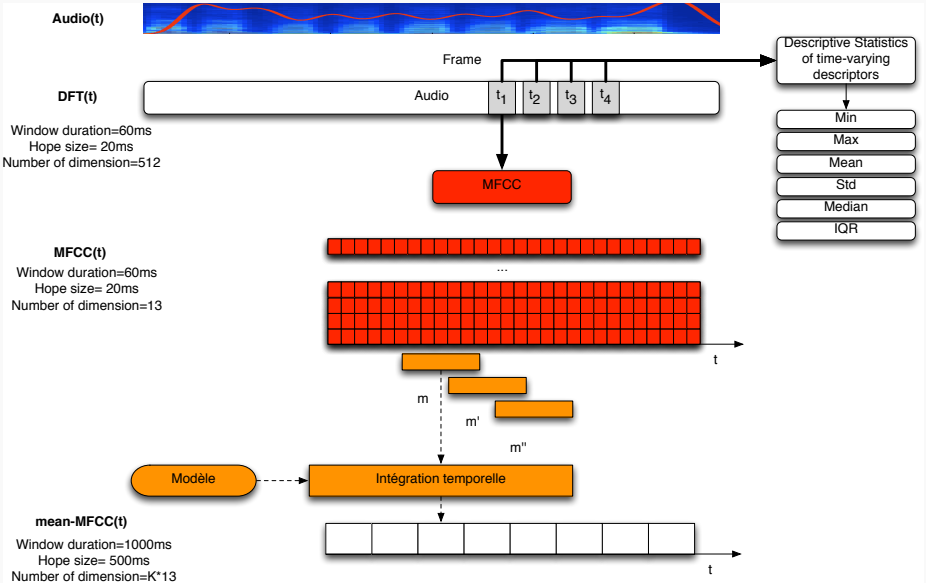

Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

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, diffiçile à lire, trop grande dimension
- Contrainte :
  - on veut le même nombre de dimensions pour toutes les données
- Extraction ?
  - Algorithme d'estimation
  - Opérateurs mathématique

Les descripteurs audio



Objectif

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

Cepstre complexe

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

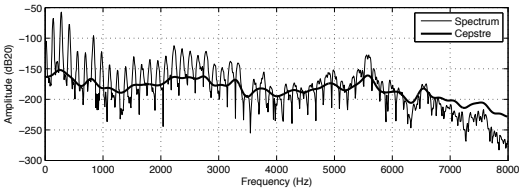
$$\begin{aligned} c(\tau) &= TF^{-1}\left[\log(X(\omega))\right] \\ &= \frac{1}{2\pi}\int_{\omega}\log(X(\omega))e^{j\omega\tau}d\omega \end{aligned} \tag{2}$$

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

Cepstre complexe

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

$$\begin{aligned} x(t) &= e(t) \circledast g(t) \\ \xrightarrow{TF} X(\omega) &= E(\omega) \cdot G(\omega) \end{aligned} \tag{3}$$



$$\begin{aligned} \xrightarrow{\log} \log(X(\omega)) = & \underbrace{\log(E(\omega))}_{\text{variation rapide à travers } \omega} + \underbrace{\log(G(\omega))}_{\text{variation lente à travers } \omega} \\ \xrightarrow{TF^{-1}} TF^{-1}\left[\log(X(\omega))\right] = & \underbrace{TF^{-1}\left[\log(E(\omega))\right]}_{\text{énergie aux céfrences } \tau >>} + \underbrace{TF^{-1}\left[\log(G(\omega))\right]}_{\text{énergie aux céfrences } \tau <<} \end{aligned} \tag{4}$$

Cepstre réel

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

$$\begin{aligned} X(\omega) &= A(\omega) \cdot e^{j\phi(\omega)} \\ \log(X(\omega)) &= \log(A(\omega)) + j\phi(\omega) \\ \Re(\log(X(\omega))) &= \log(A(\omega)) \end{aligned} \tag{5}$$

$$\begin{aligned} \text{cepstre réel} &= TF^{-1}\left[\Re(\log(X(\omega)))\right] \\ &= TF^{-1}\left[\log(A(\omega))\right] \\ c(\tau) &= \frac{1}{2\pi}\int_{\omega} \log(A(\omega)) e^{j\omega\tau} d\omega \end{aligned} \tag{6}$$

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

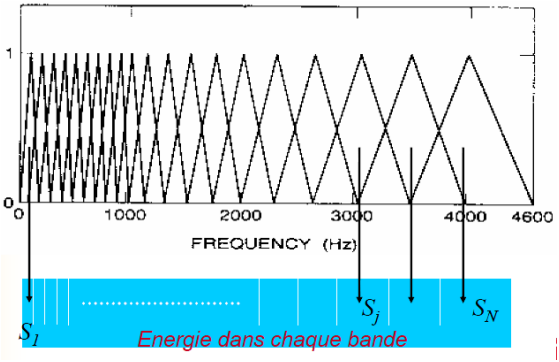
Mel Frequency Cepstral Coefficients (MFCCs)

- **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)
- Pouquoi ?
  - 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,

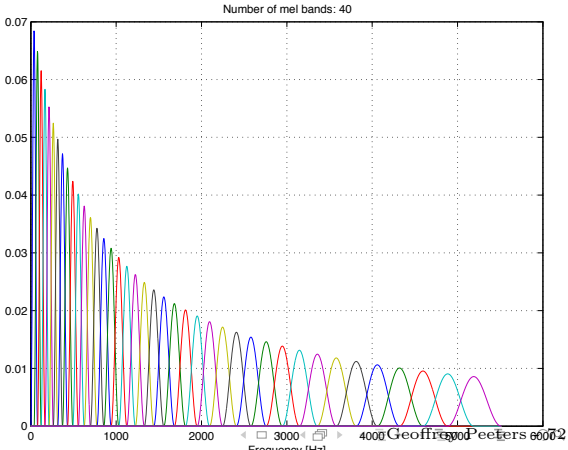
Mel Frequency Cepstral Coefficients (MFCCs)

- Echelle de Mel :

$$M = f \quad \text{pour} \quad f < 1000Hz$$
$$M = f_c \left(1 + \log_{10}\left(\frac{f}{f_c}\right)\right) \quad \text{pour} \quad f \geqslant 1000Hz \tag{7}$$

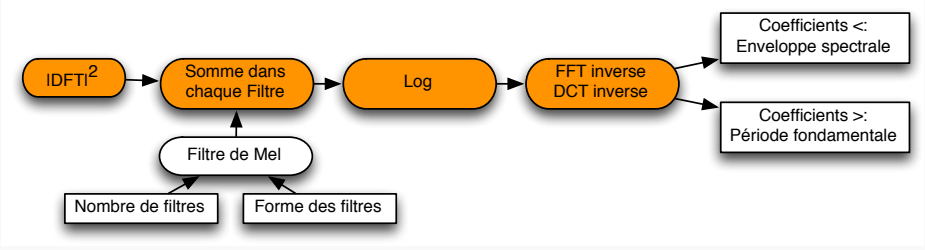


source : Gaël Richard

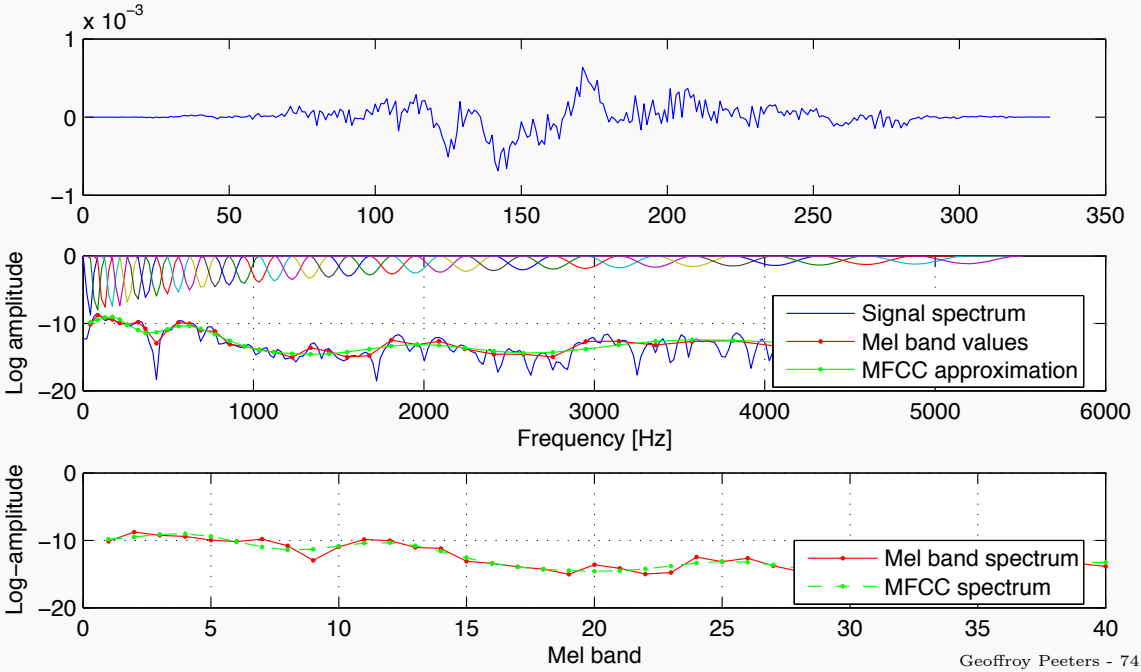


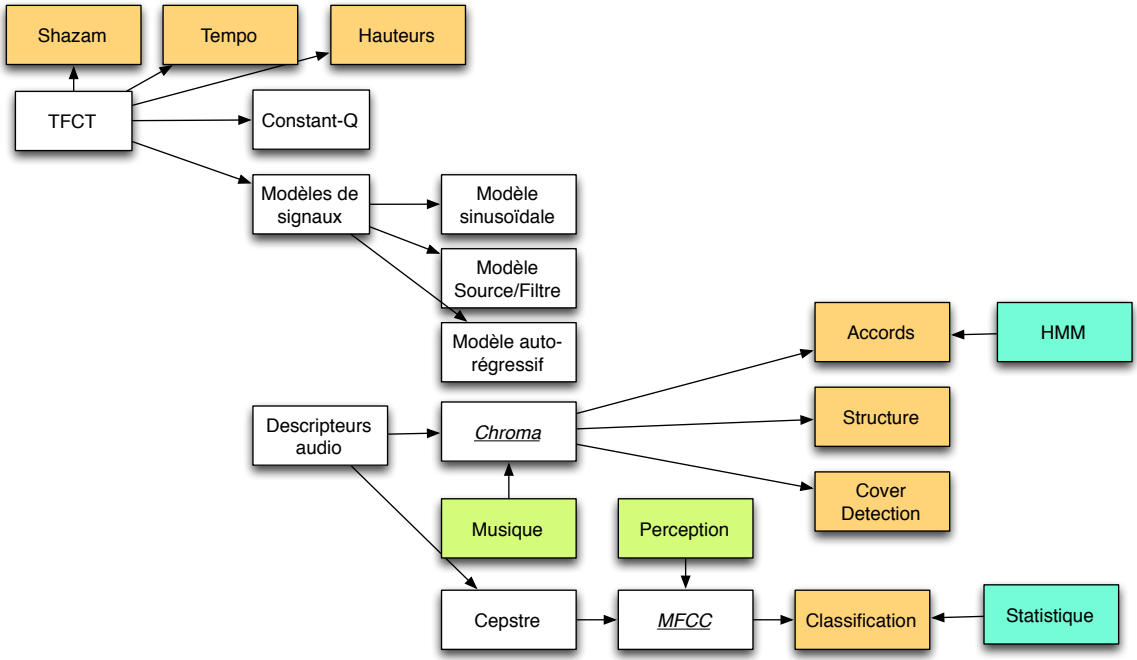
Mel Frequency Cepstral Coefficients (MFCCs)

- 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



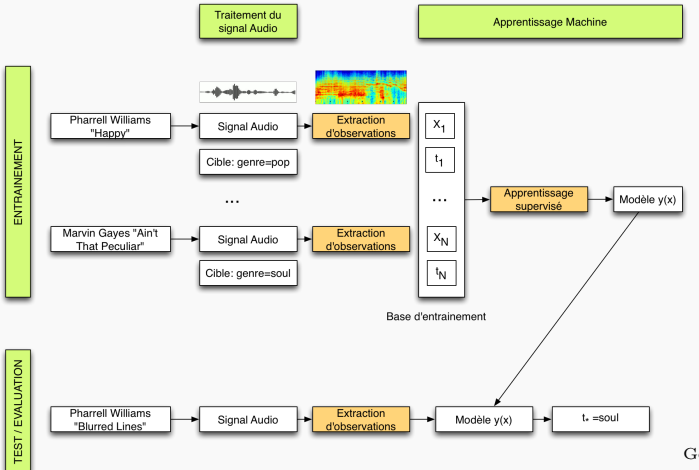
Exemple de calcul de MFCCs



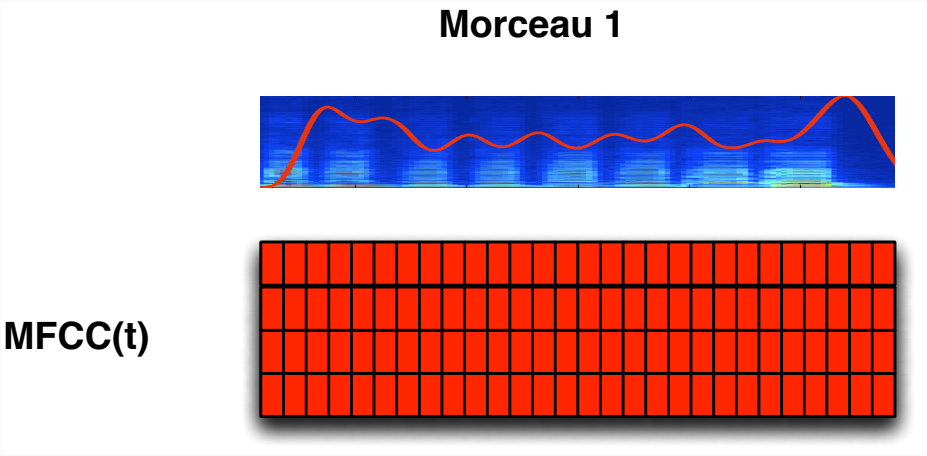

Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

Classification Audio

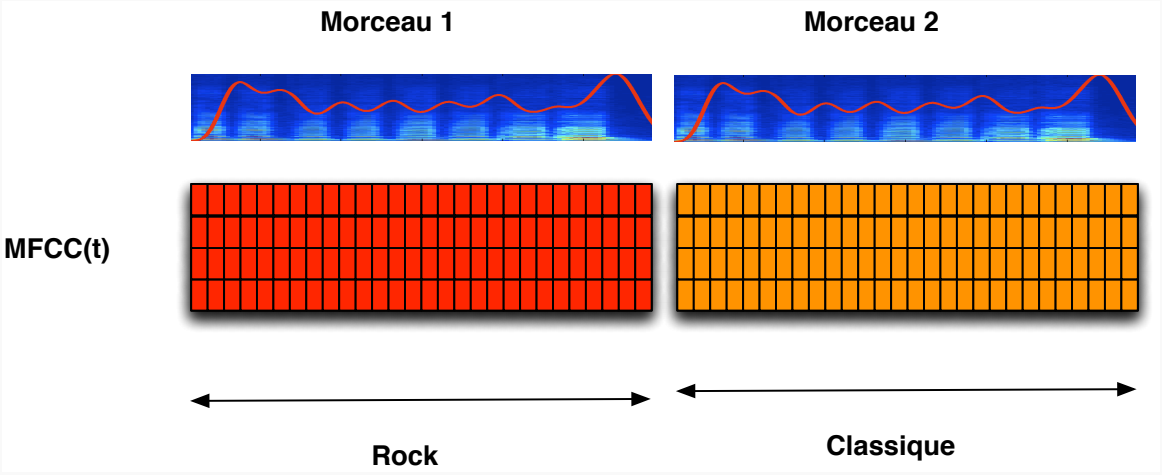
- Utilisation des descripteurs audio en entrée d'un algorithme d'**apprentissage supervisé**
- Exemples d'utilisation :
  - auto-tagging en genre, en mood (humeur), en instrumentation
  - segmentation d'un flux tempos en catégories paroles/musiques, musique instrumentale/chantée



Extraction des descripteurs sur un fichier audio

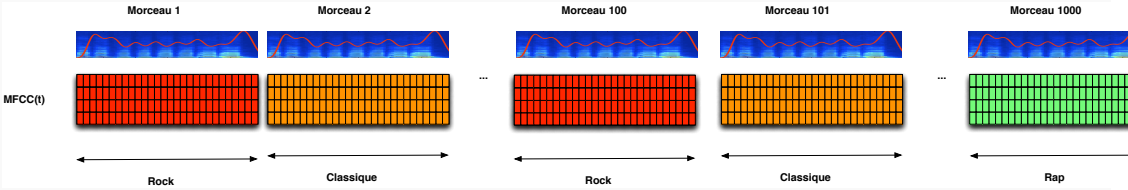


Extraction des descripteurs sur plusieurs fichier audio + assignation des labels de classes aux données

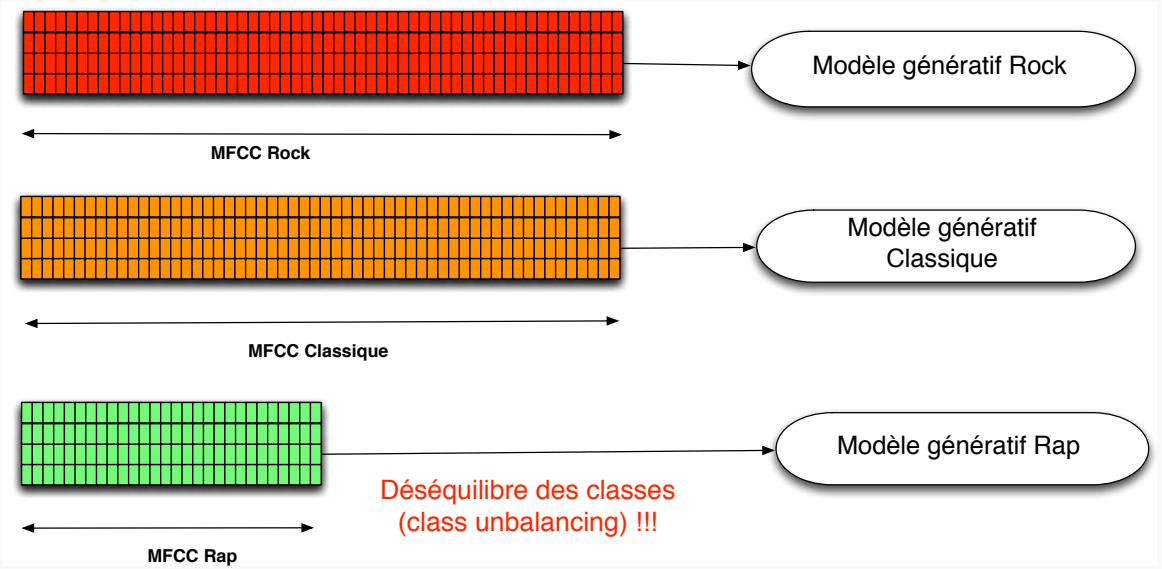


Extraction des descripteurs sur une collection de fichiers + assignation des labels de classes aux données

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



Apprentissage



Algorithmes d'apprentissage supervisé

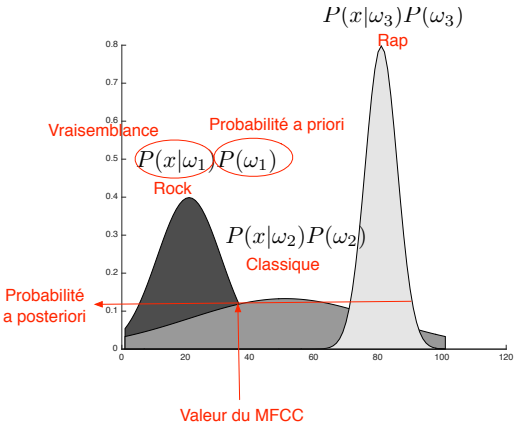
- Modèles génératifs : Gaussien, GMM
- Modèles discriminants : LDA, SVM
- Approche par exemplification : KNN

Modèle génératif gaussien

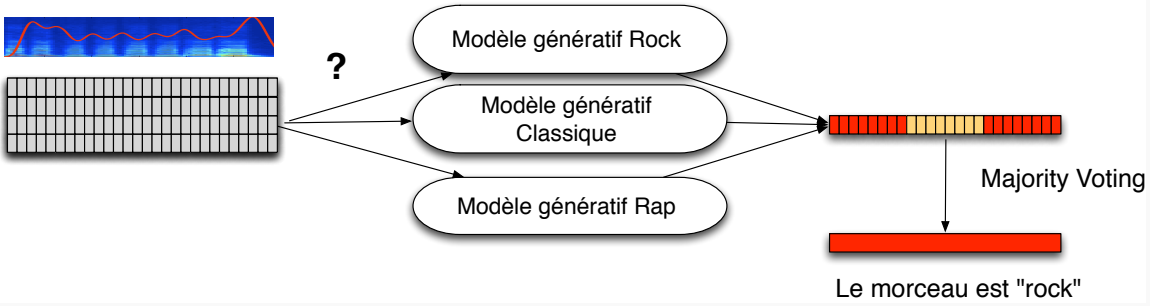
- On modélise chaque classe ω_i par une densité de probabilité gaussienne
  - $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\Sigma^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}$
- On applique la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

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

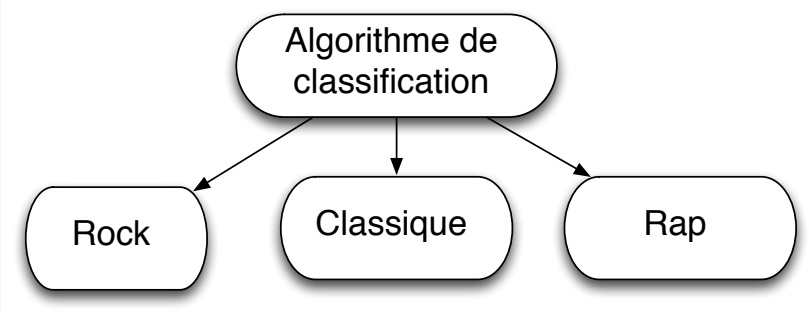
$$\text{posterior} = \text{prior} \cdot \frac{\text{vraisemblance}}{\text{evidence}}$$



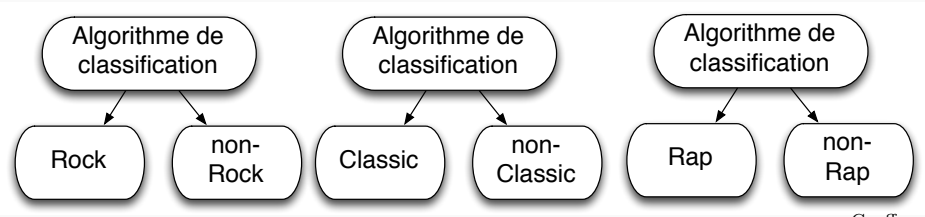
Estimation du label de classe d'un fichier audio inconnu



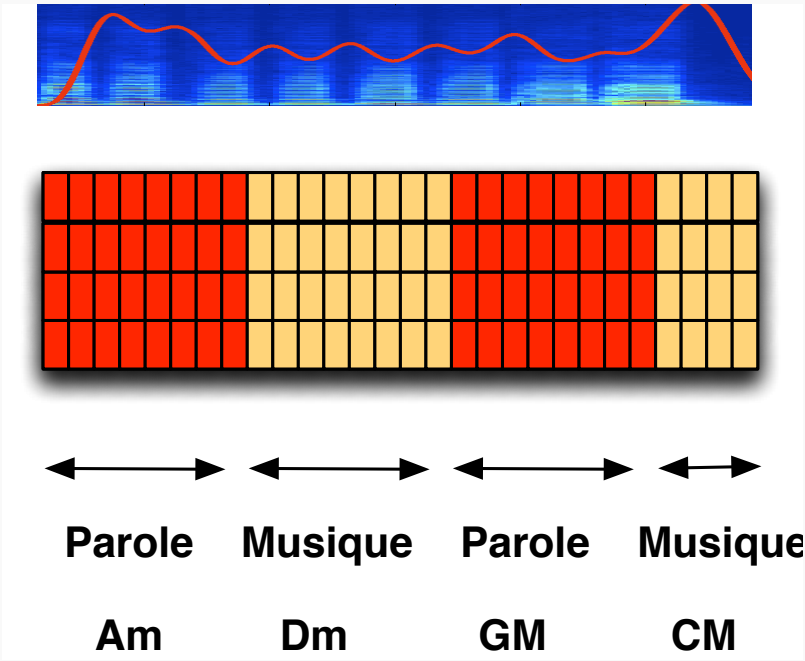
Classificateur multi-classes



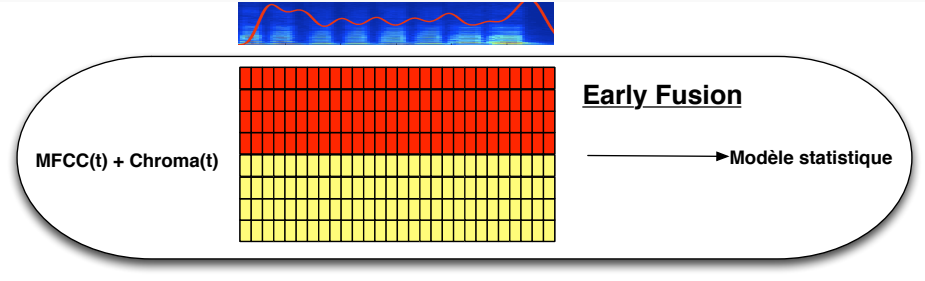
Classificateur binaire (One versus All)



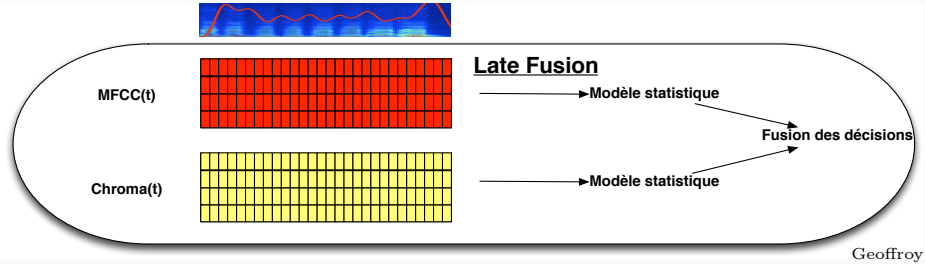
Segmentation = classification local en temps

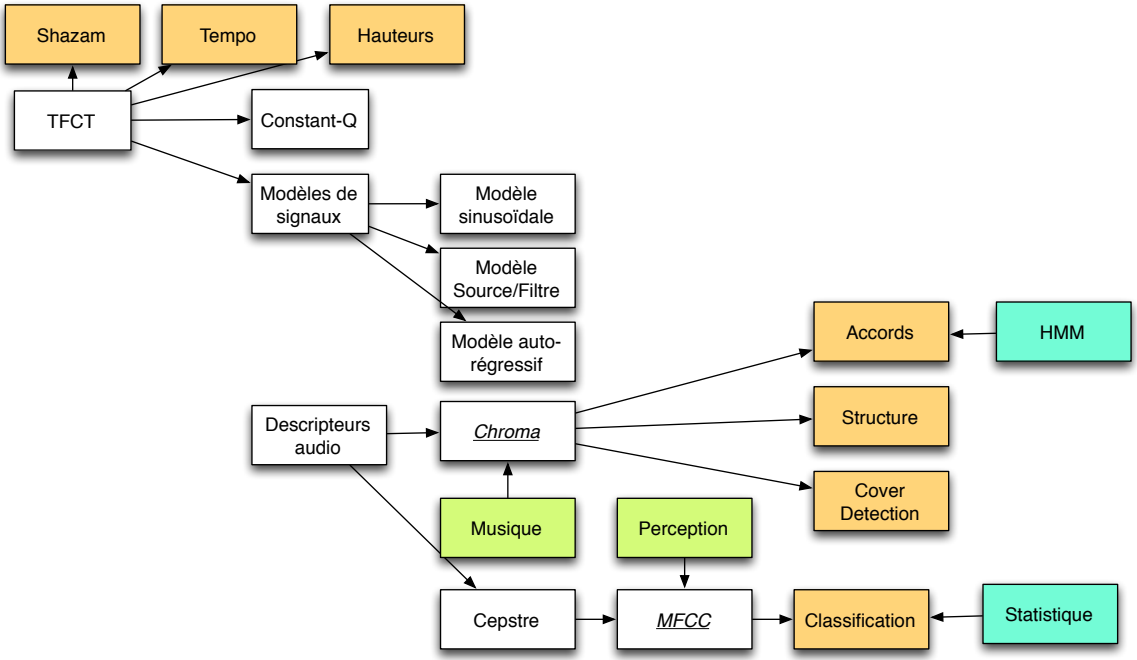


Early Fusion



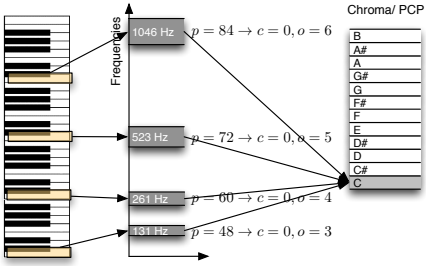
Late Fusion

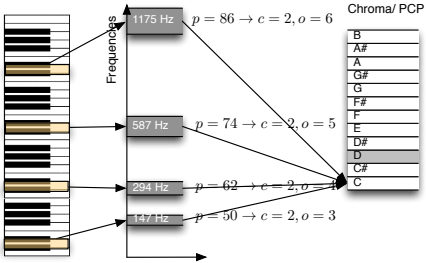



Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

Définition des Chroma - Pitch Class Profile (PCP)

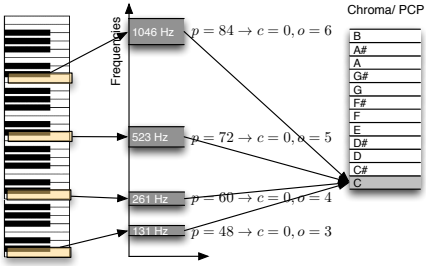
- **Objectif :**
  - le spectre à l'instant n : $X(k,n)$
  - représenter son contenu harmonique sous forme d'un vecteur : $C(c,n)$ $c \in [0,12[$
- Utilisations :
  - reconnaissance de tonalité,
  - reconnaissance de suite d'accords,
  - détection de "cover versions"
- Shepard-1964 :
  - représenter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

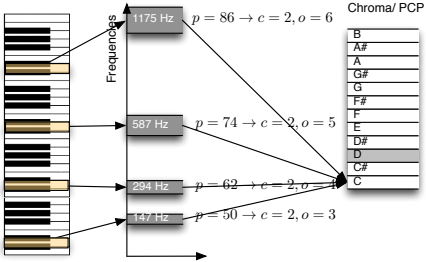




Calcul des Chromas - Pitch Class Profile (PCP)

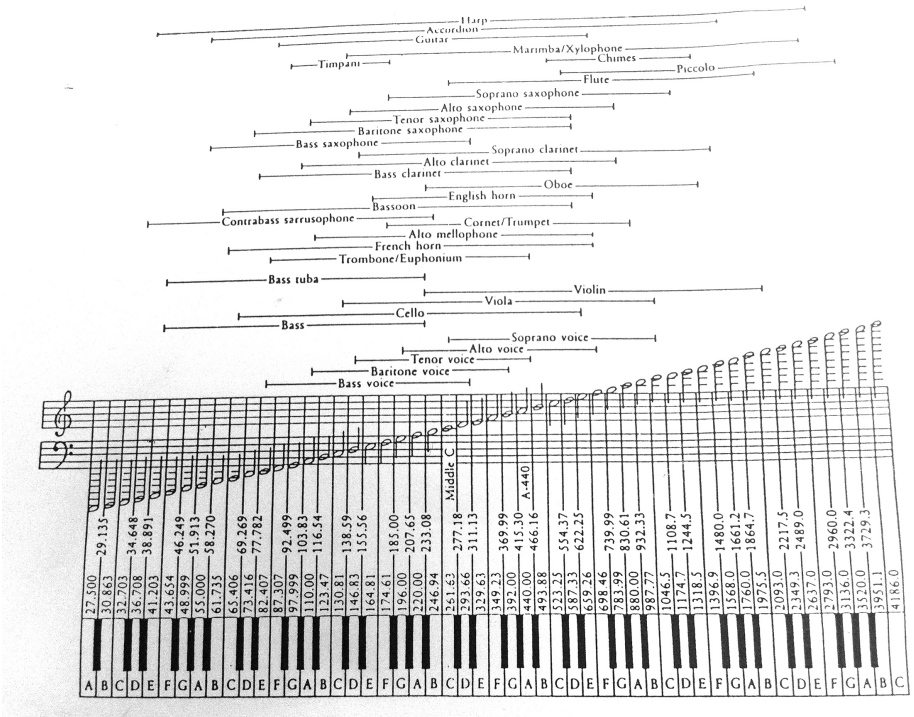
- Relation entre les fréquences f_k de la DFT et les hauteurs de note p (hauteurs de demi-tons en échelle de notes MIDI)
  - $p(f_k) = 12 \log_2 \left(\frac{f_k}{440} \right) + 69, \quad p \in \mathbb{R}^+$
  - $f(p) = 440 \cdot 2^{\frac{p-69}{12}}$
- Calcul des chromas $C(c, n)$
  - On additionne toutes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





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 sectre) 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}

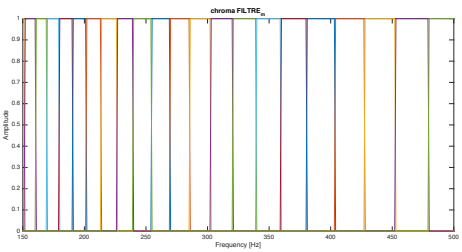

Harp
Accordion
Guitar
Marimba/Xylophone
Timpani
Chimes
Piccolo
Flute
Soprano saxophone
Alto saxophone
Tenor saxophone
Baritone saxophone
Bass saxophone
Soprano clarinet
Alto clarinet
Bass clarinet
Oboe
English horn
Bassoon
Contrabass sarrusophone
Cornet/Trumpet
Alto mellophone
French horn
Trombone/Euphonium
Bass tuba
Violin
Viola
Cello
Bass
Soprano voice
Alto voice
Tenor voice
Baritone voice
Bass voice
Middle C
A-440
A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C D E F G A B C
27.500 29.135 30.863 32.703 34.648 36.708 38.891 41.203 43.654 46.249 48.999 51.913 55.000 58.270 61.735 65.406 69.269 73.416 77.782 82.407 87.307 92.499 97.999 103.83 110.00 116.54 123.47 130.81 138.59 146.83 155.56 164.81 174.61 185.00 196.00 207.65 220.00 233.08 246.94 261.63 277.18 293.66 311.13 329.63 349.23 369.99 392.00 415.30 440.00 466.16 493.88 523.25 554.37 587.33 622.25 659.26 698.46 739.99 783.99 830.61 880.00 932.33 987.77 1046.5 1108.7 1174.7 1244.5 1318.5 1396.9 1480.0 1568.0 1661.2 1760.0 1864.7 1975.5 2093.0 2217.5 2349.3 2489.0 2637.0 2793.0 2960.0 3136.0 3322.4 3520.0 3729.3 3951.1 4186.0

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

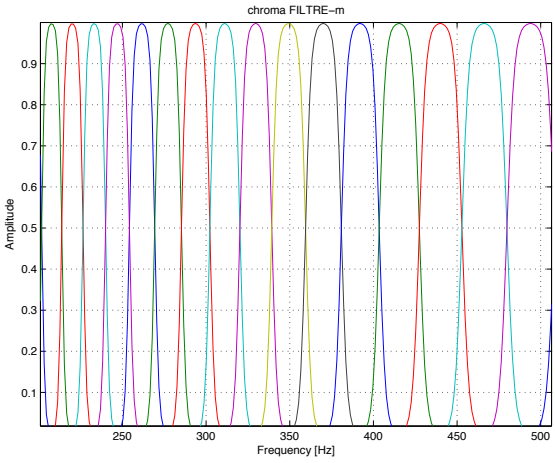
Hard-mapping

- Hard-mapping ?
  - Une fréquence f_k de la DFT contribue uniquement à la note la plus proche
  - Par exemple,
    - l'énergie à f_k=452 Hz ($p(f_k)$=69.4658) contribue entièrement à la note p=69 (c=10)
    - alors que f_k=453 Hz ($p(f_k)$=69.5041) à p=70 (c=11).
- Création d'un banc de filtres $H_{p'}$ centrés sur les hauteurs de demi-tons $p' \in [43, 44, \ldots, 95]$:



Soft-mapping

- Soft-mapping ?
  - Une fréquence f_k de la DFT contribue à différents chroma avec un poids inversement proportionnel à la distance entre $p(f_k)$ et les p les plus proches
  - Par exemple,
    - l'énergie à f_k=452 Hz ($p(f_k)$=69.4658) contribuera de manière presque égale à p=69 (c=10) et p=70 (c=11).
- Création d'un banc de filtres $H_{p'}$ centrés sur les hauteurs de demi-tons $p' \in [43, 44, \ldots, 95]$:

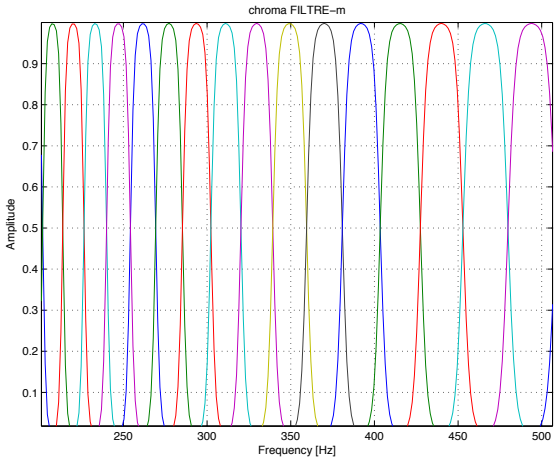


Soft-mapping

- Création d'un banc de filtres $H_{p'}$ centrés sur les hauteurs de demi-tons $p' \in [43, 44, \ldots, 95]$:
  - Chaque filtre est défini par la fonction

$$H_{p'}(f_k) = \frac{1}{2}\tanh\left(\pi(1-2x)\right) + \frac{1}{2}$$

    dans lequel x= distance relative entre centre du filtre et fréquences de la TF $x = R\,|p' - p(f_k)|$.
  - Les filtres sont équi-répartis et symétriques sur l'échelle logarithmique des hauteurs de demi-tons, non-nulles entre $p'-1$ et $p'+1$ et à valeur maximale en p'.



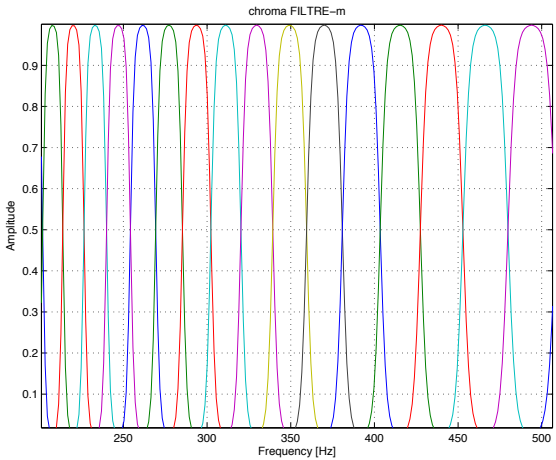
Calcul des Chromas - Pitch Class Profile (PCP)

- La valeur du spectre de hauteur de demi-ton $N(n')$ est obtenue en multipliant les valeurs de la transformée de Fourier $A(f_k)$ par l'ensemble des filtres $H_{n'}$:

$$P(p') = \sum_{f_k} H_{p'}(f_k) A(f_k)$$

- Le mapping entre les hauteurs de demi-tons n et les classes de hauteurs de demi-ton (chroma) c est défini par $c(p) = mod\,(p, 12)$.
- La valeur du vecteur de chroma est obtenue en additionnant les valeurs de classes de hauteur équivalentes

$$C(c) = \sum_{p' \text{ tel que } c(p')=l} P(n') \quad c \in [0, 12[$$



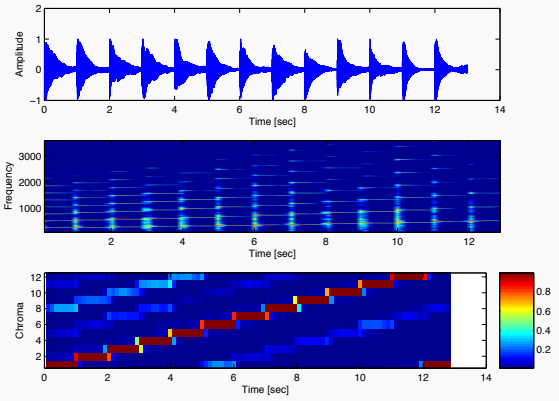
Limitations des Chromas - Pitch Class Profile (PCP)

- 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

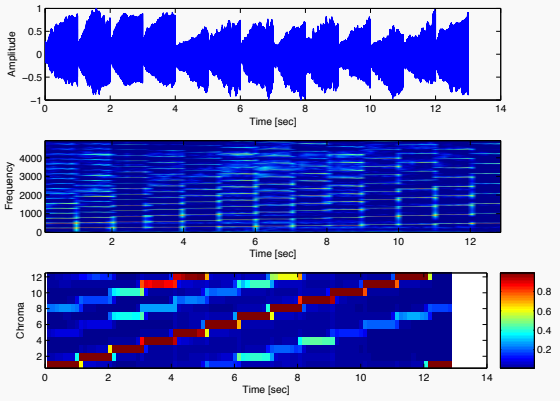
Pitch	Harmonic	Frequency f_μ	MIDI-scale m_μ	Chroma/PCP p
c3	f_0	130.81	48	1 (=c)
	$2f_0$	261.62	60	1 (=c)
	$3f_0$	392.43	67.01	8.01 ($\simeq$ g)
	$4f_0$	523.25	72	1 (=c)
	$5f_0$	654.06	75.86	4.86 ($\simeq$ e)
	...	...	...	...

Limitations des Chromas - Pitch Class Profile (PCP)

Exemple piano

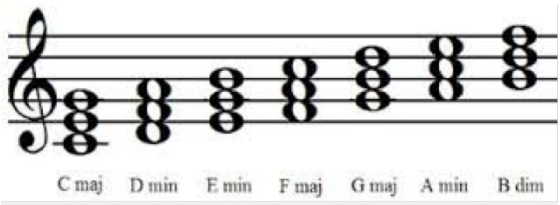


Exemple violon

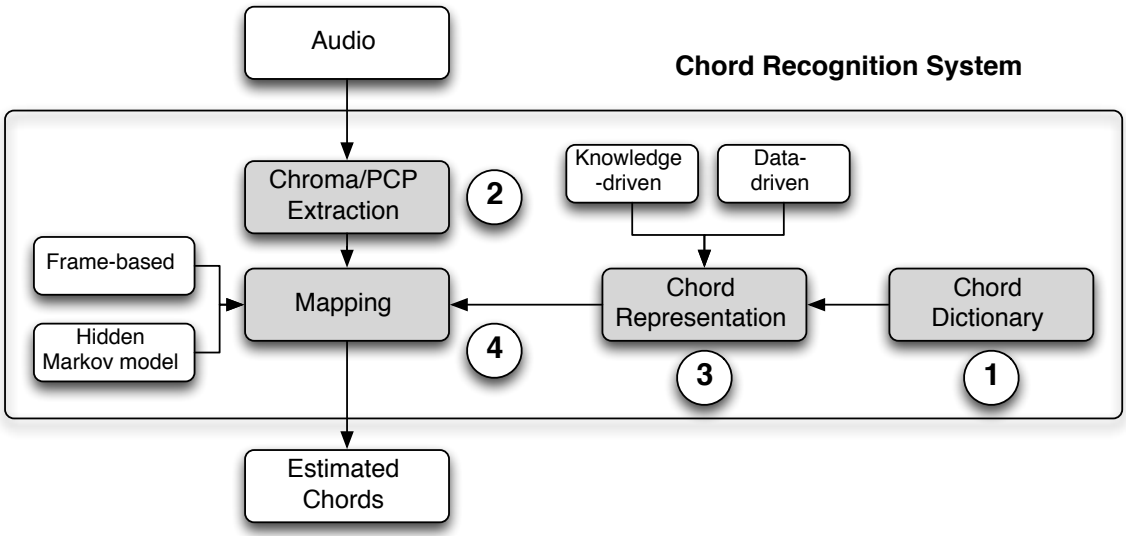


Les accords ?

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

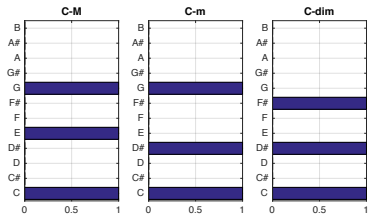


Root-note	Type of the chord
c, c#, d, d# . . . b	**Triads**: major (C-M: c,e,g) , minor (C-m: c, e♭ , g), suspended (C-sus2: c, d, g / C-sus4: c, f, g), augmented (C-aug: c, e, g#), diminished (C-dim: c, e♭, g♭)
	Tetrads: major 7 (C-M7: c, e, g, b), minor 7 (C-m7: c, e♭, g, b♭), dominant 7 (C-7: c, e,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


Audio
Chord Recognition System
Chroma/PCP Extraction
2
Knowledge -driven
Data-driven
Frame-based
Hidden Markov model
Mapping
4
Chord Representation
3
Chord Dictionary
1
Estimated Chords

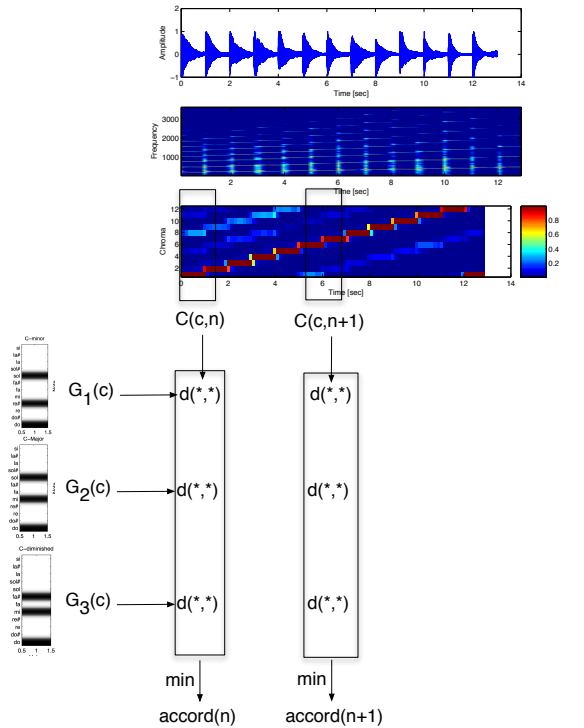
Estimation des accords ?

- 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, ...]
    - c le chroma
      - $c \in [0, 12[$
  - Vecteur de Chroma $C(c, n)$ à l'instant n

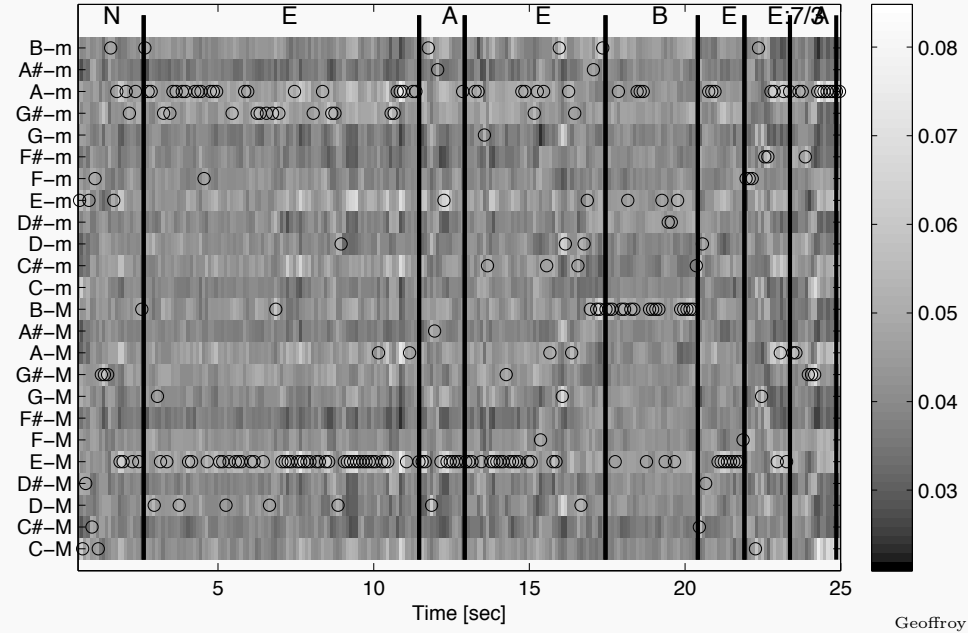


Estimation des accords par trame

- 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'acoord donnant la distance la plus petite
- Choix de la distance $d(C(c,n), G_a(c))$
  - distance Euclidéenne
  - distance cosinusoidale

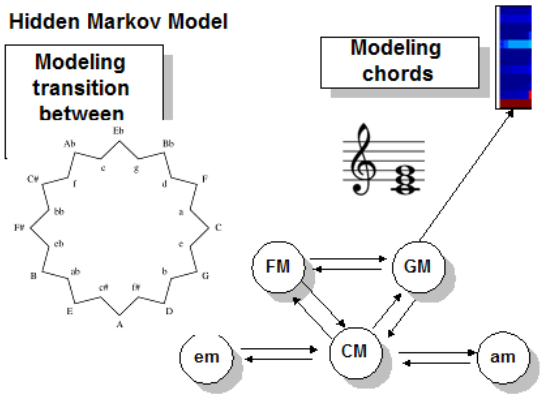


Estimation des accords par trame



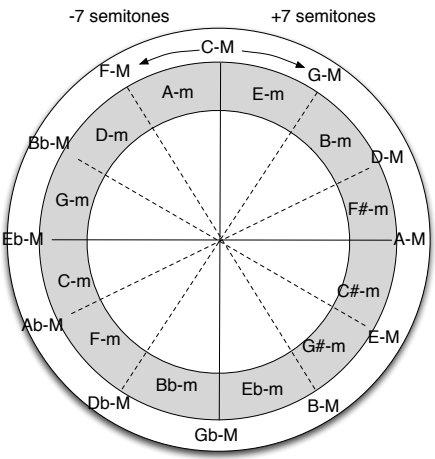
Estimation des accords par trame par modèle de Markov caché (HMM)

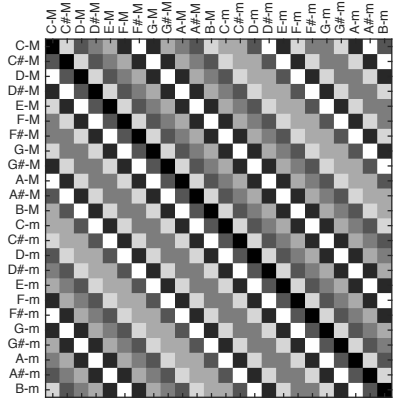
- **Observations** :
  - on extrait la séquence de descripteurs audio chroma/PCP
- **Etats cachés** :
  - les 24 accords
- **Probabilités initiales :**
- **Probabilité d'émission** des accords :
  - $p(S = \text{C-M}|\text{chroma})$, $p(S = \text{C\#-M}|\text{chroma})$, ...



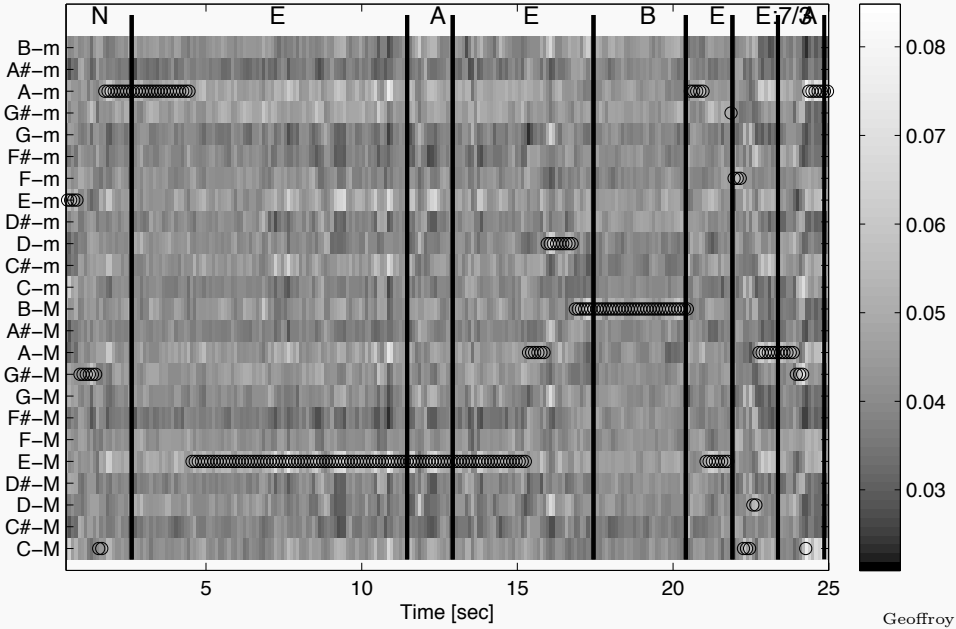
Estimation des accords par trame par modèle de Markov caché (HMM)

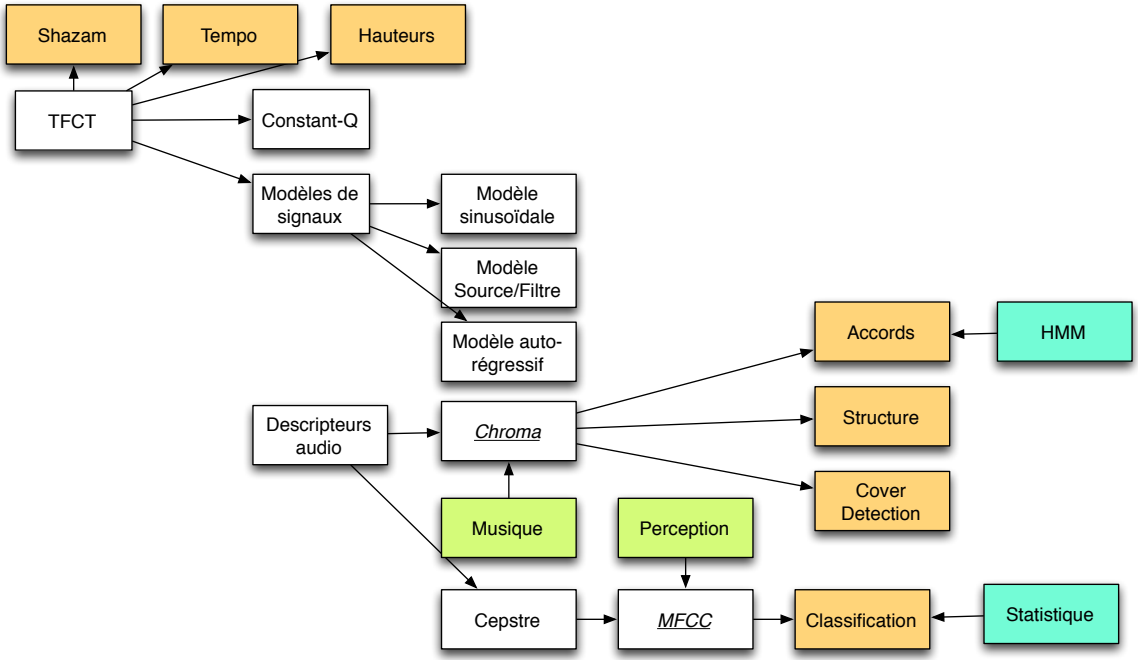
- **Probabilitiés de transition** entre accords :
  - suivent la théorie musicale (cercle des quintes, relatifs majeur-mineur) :
    - G-M vers C-M (consonance),
    - G-M vers C#-M (dissonance)





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



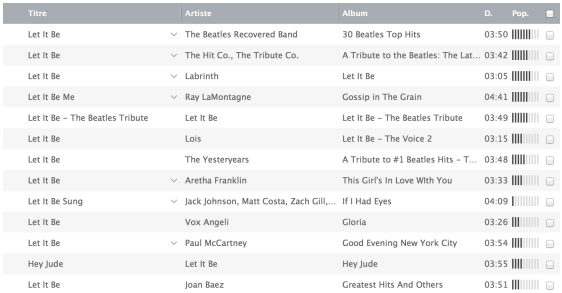

Shazam
Tempo
Hauteurs
TFCT
Constant-Q
Modèles de signaux
Modèle sinusoïdale
Modèle Source/Filtre
Modèle auto-régressif
Accords
HMM
Descripteurs audio
Chroma
Structure
Cover Detection
Musique
Perception
Cepstre
MFCC
Classification
Statistique

Cover-version ?

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

Caractéristiques d'une cover-version :

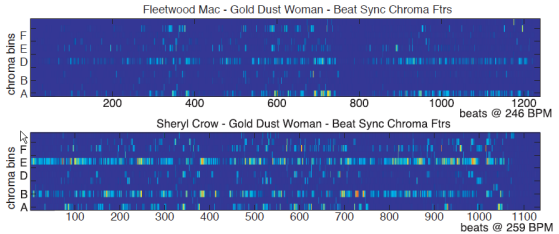
- généralement la même suite harmonique
  - même suite d'accords,
  - même mélodie
- éventuellement transposée

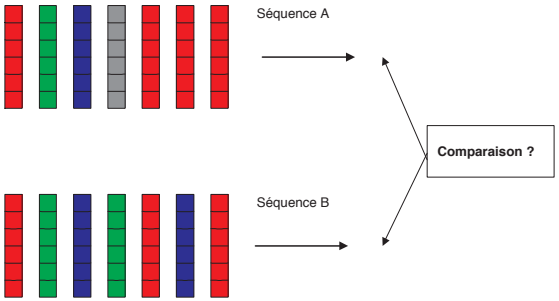


Titre	Artiste	Album	D.	Pop.
Let It Be	The Beatles Recovered Band	30 Beatles Top Hits	03:50	
Let It Be	The Hit Co., The Tribute Co.	A Tribute to the Beatles: The Lat...	03:42	
Let It Be	Labrinth	Let It Be	03:05	
Let It Be Me	Ray LaMontagne	Gossip in The Grain	04:41	
Let It Be – The Beatles Tribute	Let 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	

Méthode

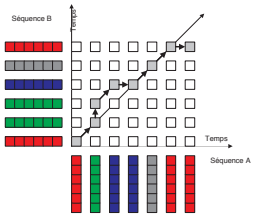
- Chaque morceau est représenté par la séquence temporelle de ses chromas/PCP : $C(c, n)$
  - c est le chroma (pitch-class) et
  - n est le temps
- Pour une collection de morceaux, on compare les morceaux deux à deux
- Comparaison de deux morceaux A et B
  - Calcul du coût pour aligner la séquence de 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





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=end, b=end)

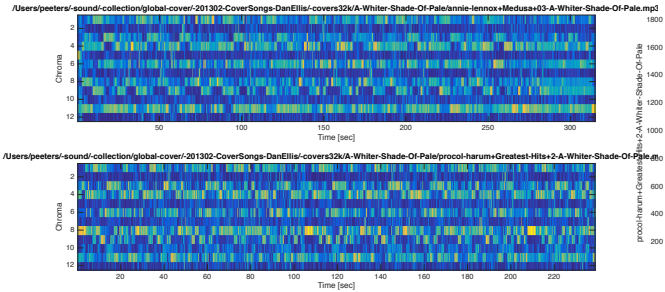


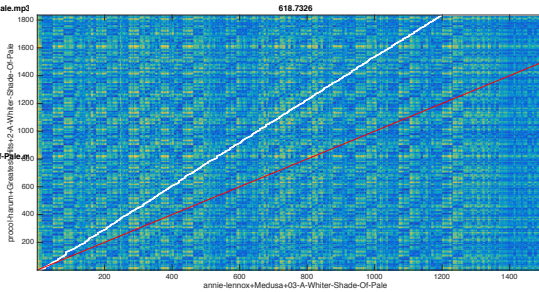

chemin normal


insertion


deletion

Exemple d'alignement de coût faible





Exemple d'alignement de coût élevé