SI350
Indexation Audio

Gaël RICHARD

Télécom ParisTech

Département Traitement des signaux et des images

Juin 2012

Contenu

- **Introduction**
- **Outils pour l'indexation audio**
  - Exemple d'architecture d'un système d'indexation
  - Paramétrisation
  - Classification
- **Quelques exemples d'application de l'indexation audio**
  - Identification/classification des instruments de musique
  - Extraction du rythme
  - Identification audio
  - Extraction de fréquences fondamentales multiples
  - Applications aux signaux percussifs (batterie)
- **Conclusion**

Merci à Olivier Gillet pour certains transparents

Indexation audio : intérêts

- **Nouveaux challenges pour la société de l'information:**
  - Volume considérable de données numériques multimedia disponibles
  - L'accroissement rapide et continu de ces données numériques (qu'elles se trouvent sur le réseau Internet ou dans des bases personnelles)
  - Généralisation de leur utilisation pour de nombreuses applications

  Diminution de « l'accessibilité » des données
  Un fort besoin pour de nouvelles méthodes efficaces d'indexation, de classification et d'accès par le contenu.

- **L'indexation automatique vise ainsi à extraire du flux numérique multimedia des descripteurs de haut niveau permettant de réaliser par la suite une classification ou un accès à l'information par son contenu.**

Recherche par le contenu …..

Google™ Audio BETA

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

Search Audio

Audio Preferences
Audio Help

Steve Jobs interview
7 min 14 sec
Speech

Metric - Raw Sugar
3 min 47 sec
Music - Indie Pop

Grenade explosion
23 sec
Sound effect

similarly random recordings »

Google Labs - Discuss - Terms of use - About Google Audio - Submit your recording

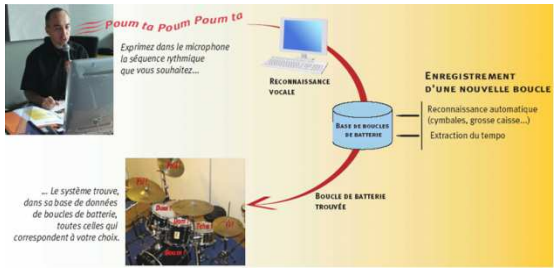
TELECOM ParisTech

Pourquoi analyser le signal musical ?

- **Rechercher par le contenu**
  - À partir d'un morceau …
  - À partir d'un chantonnement…
  - De nouveaux morceaux à partir de ce que j'aime….
  - Une nouvelle version d'un air connu ..
  - Une vidéo qui « va bien » avec l'audio
  - …

- **Nouvelles applications**
  - Playlist « sémantiques » (jouer des morceaux de plus en plus rapide…)
  - Karaoké « intelligent » (l'accompagnement suit le chanteur…)
  - Prédire le potentiel succès d'un titre
  - Aide au mixage, Djing,
  - Ecoute active,..

Recherche à la voix



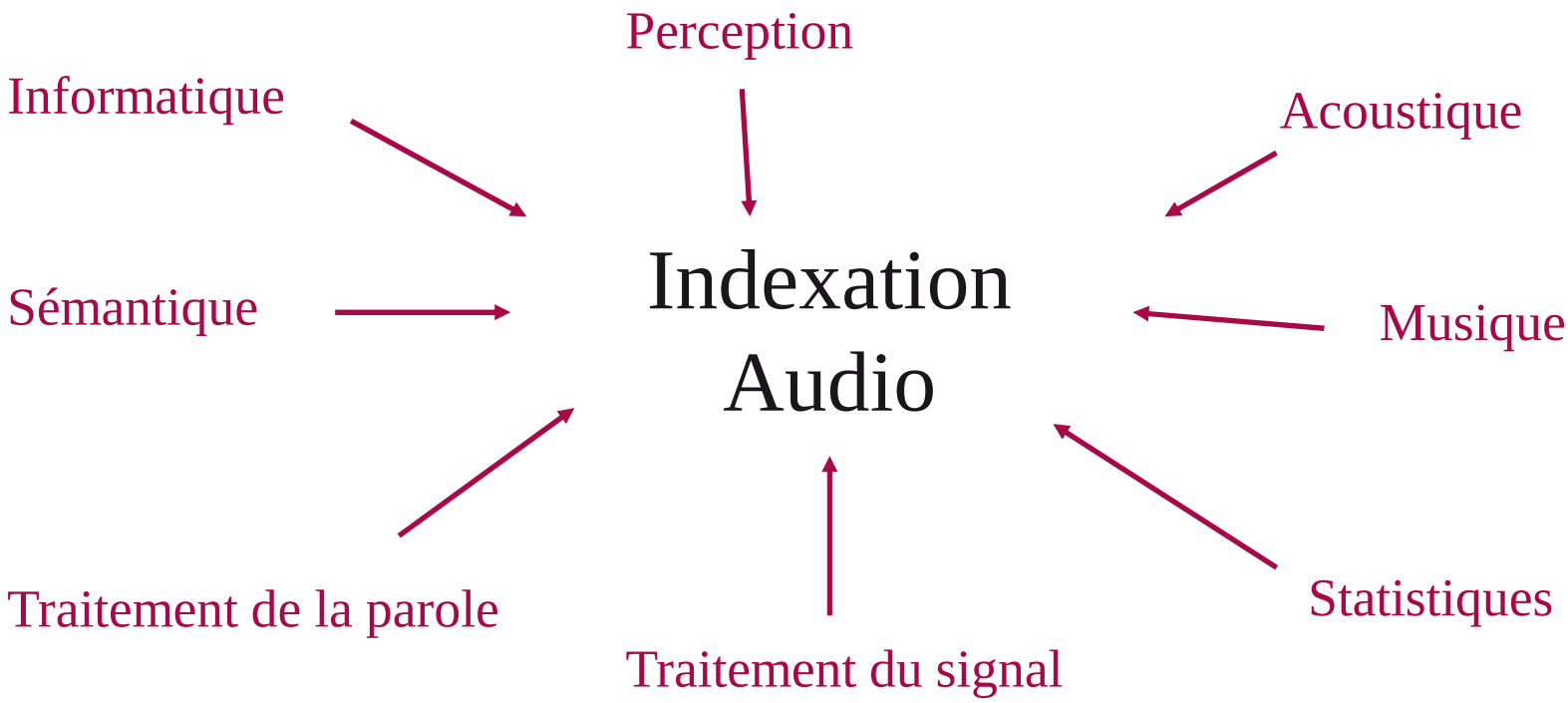
Jogging musical

Modifications synchrones

Playlist, « espace musical »

L'indexation audio…

- **… un domaine multi-disciplinaire.**

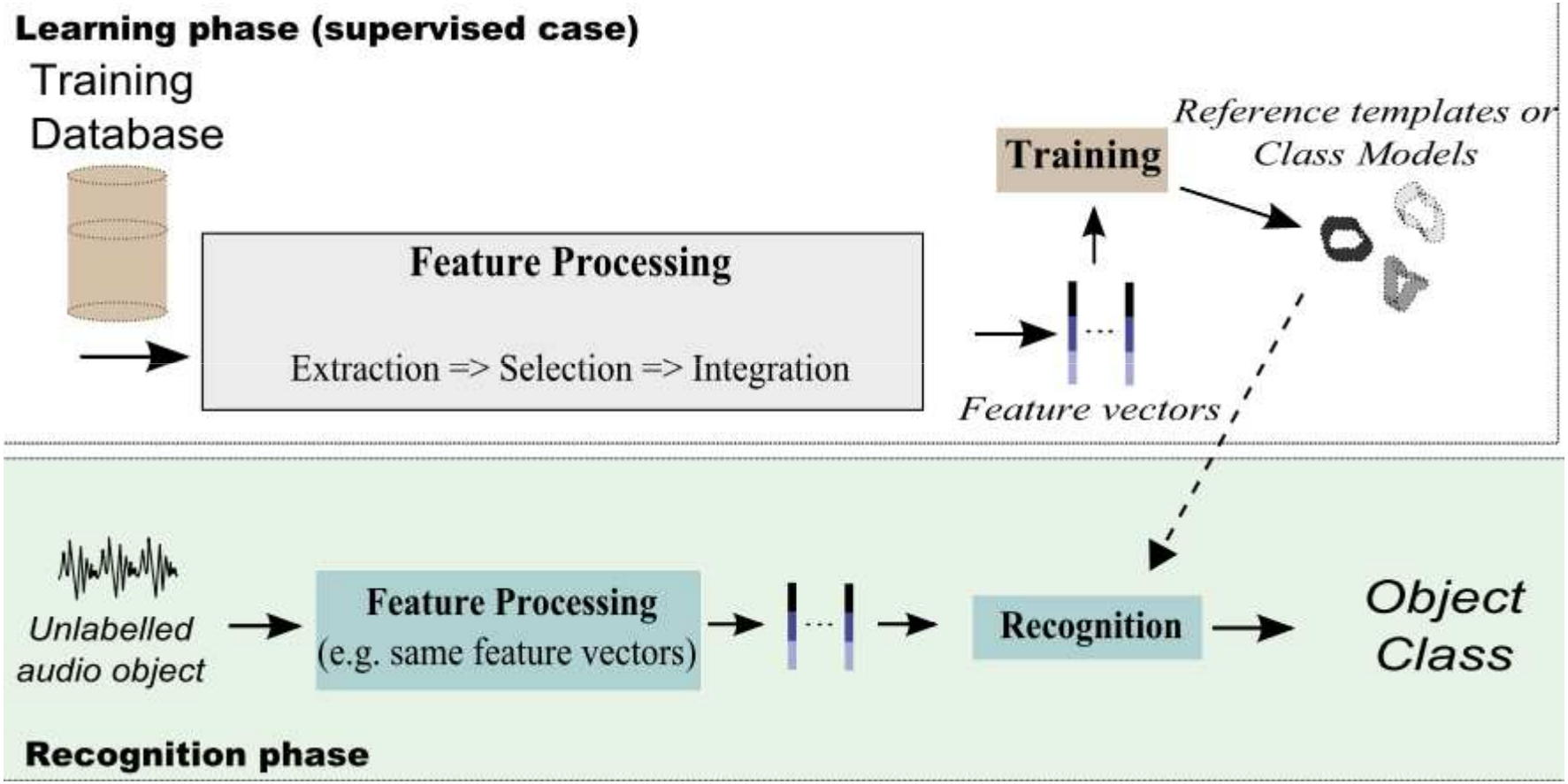


Systèmes de classification

- **Plusieurs problèmes, une même approche**

  - Reconnaissance automatique du genre musical.
  - Reconnaissance des instruments utilisés dans un morceau.
  - Classification d'échantillons sonores.
  - Etiquetage d'une bande son (dialogues, scènes d'action, musique, effets spéciaux).
  - Organisation d'une collection musicale selon les habitudes d'écoute.
  - Détection de "hits" potentiels.

Exemple d'architecture pour un système de classification



From G. Richard, S. Sundaram, S. Narayanan, "Perceptually-motivated audio indexing and classification", submitted to Proc. of the IEEE.

Architecture (2)

- **Prétraitements pour...**
  - Réduire la quantité de données à traiter.
  - Découper le signal en segments uniformes.

- **Paramétrisation**
  - Résumer les propriétés perceptuelles du signal en un vecteur de paramètres réels.
- **Classification**
  - **Supervisée :** Attribuer à chaque signal une classe selon une taxonomie définie à l'avance.
  - **Non-supervisée** (clustering) : Identifier des groupes disjoints de signaux qui se ressemblent.

Quelques dimensions du signal musical…

Hauteurs, Harmonie,..

Tempo, rythme,…

Timbre, instruments,…

Polyphonie, mélodie, ….

Qu‘est-ce que le “Timbre“

- *Une définition possible:* « l’attribut de la sensation auditive qui permet de différencier 2 sons de même hauteur et de même intensité »

- Lié à l’identification de sources sonores

- Examples de sons avec la même hauteur et le même niveau sonore (en terme d‘énergie globale) mais avec des timbres différents:




- Quelques thèses récentes sur la reconnaissance des instruments de musique: [Essid06], [Kitahara-07], [Eronen-09]

Le Timbre „polyphonique“

- Fait référence au timbre global (the „global sound“) d‘une pièce de musique [Alluri-10]
- Principalement porté par l‘instrumentation
- Example
  - “Bohemian rhapsody” par Queen
  - “Bohemian rhapsody” par le London Symphony Orchestra

- Retrouver automatiquement le timbre est une tâche proche de celle pour retrouver le genre musical [Scaringella-06] ou les tags musicaux [Scaringella-06] and music tagging [Turnbull-08]

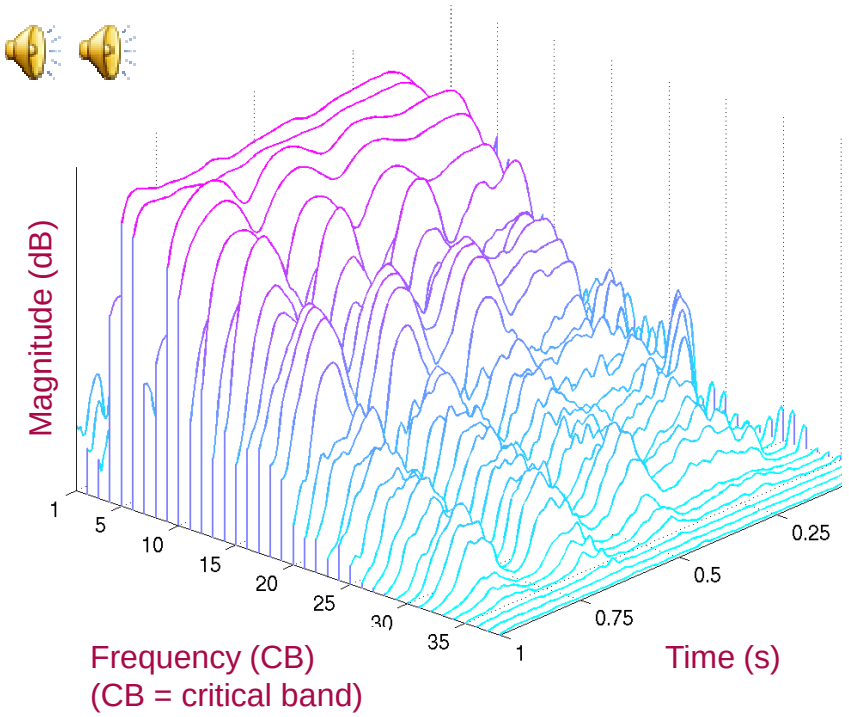
Différentes facettes du timbre

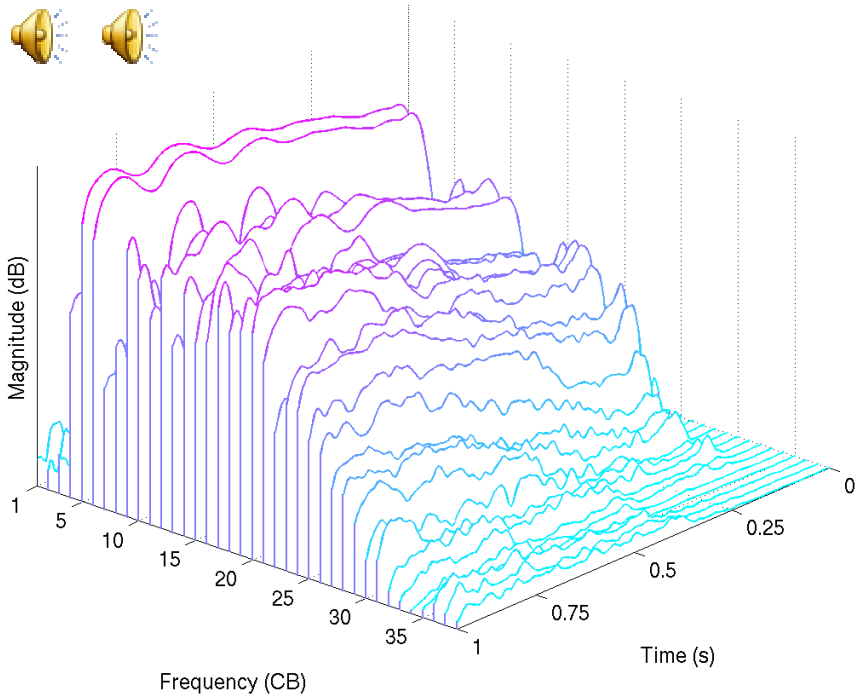
- Le timbre est un concpet multidimensionel
- De nombreux paramètres (spectraux et temporels) sont nécessaires pour le décrire
- Schouten [1968] avait listé 5 paramètres majeurs:

1. Position sur une échelle tonal vs bruité
2. Enveloppe spectrale
3. Enveloppe temporelle
4. Changement dynamique de l'enveloppe spectrale et du pitch
5. Différence entre l'attaque et la partie tenue

Variations temporelles de l‘enveloppe

- Illustration du timbre (représenté ici comme le niveau d’énergie dans les bandes critiques en fonction du temps)
- Flute (Gauche) et violon (droite)

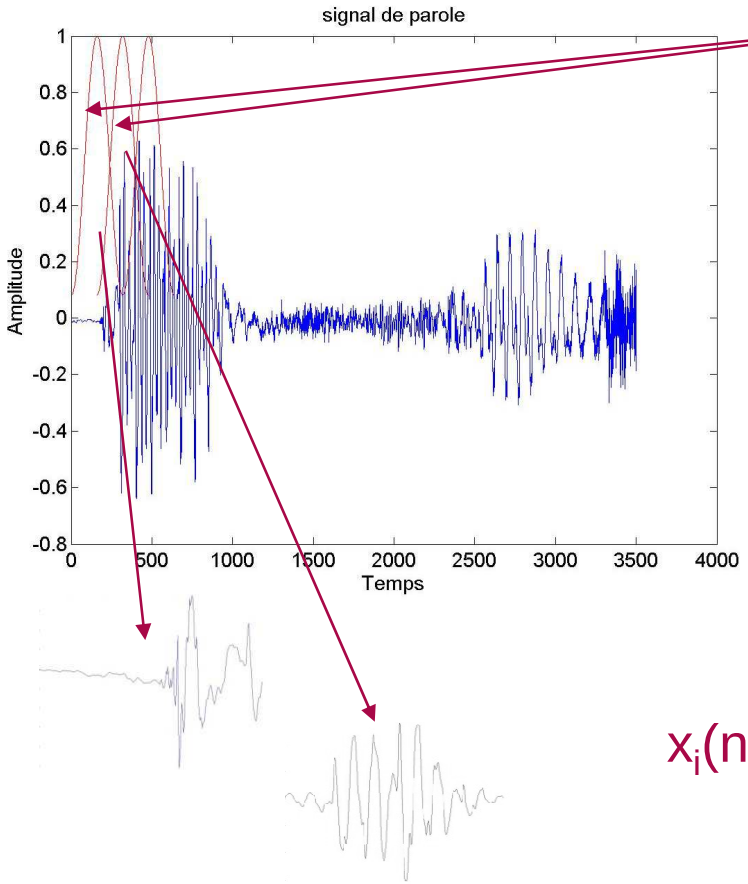


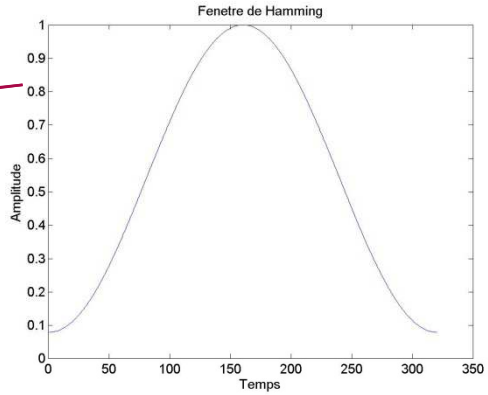


Paramètres acoustiques….

Prétraitements

- **Découper le signal en segments uniformes (« fenêtrage »)**





$x_i(n) = x(n+n_i)$. $h(n)$ pour $n=1:N$

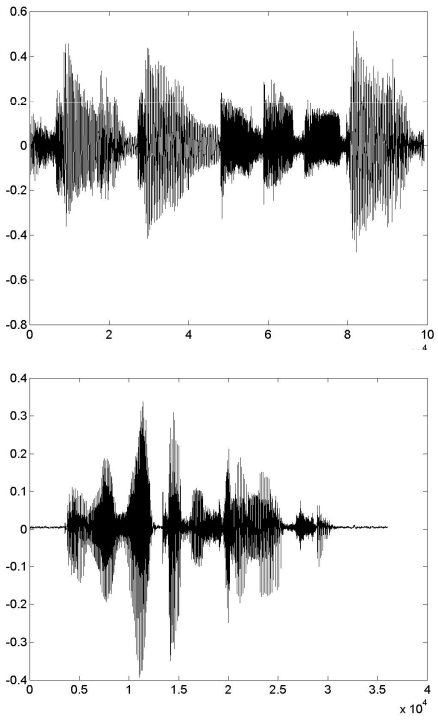
Paramétrisation

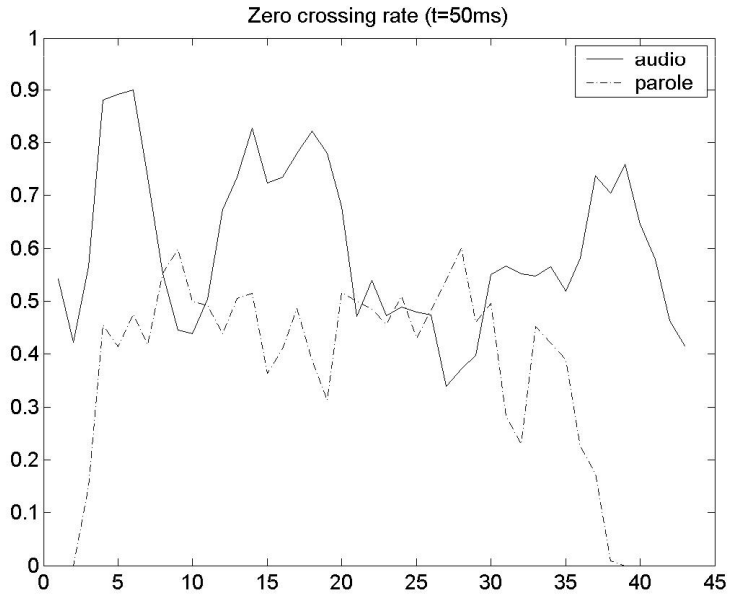
- **Paramètres temporels**
  - Taux de passage par zéro

$$Zcr = 0.5 * \sum_{n=1}^{N} \left| sign(x[n]) - sign(x[n-1]) \right|$$

Audio

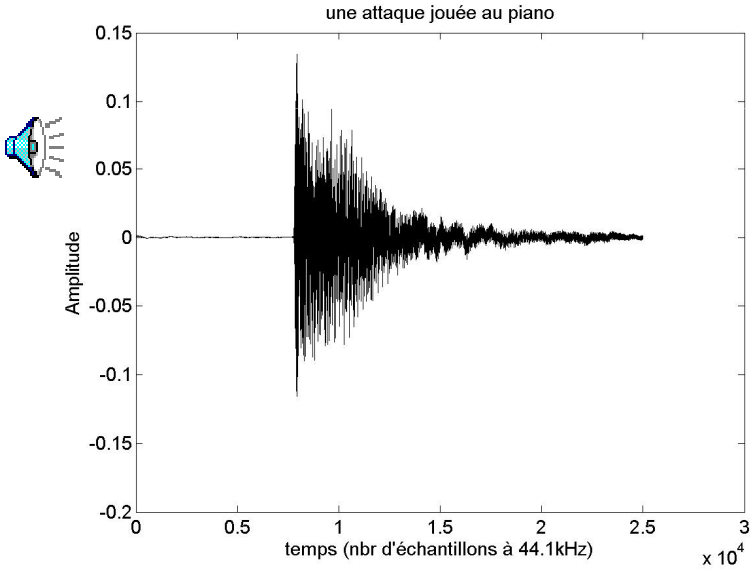
Parole

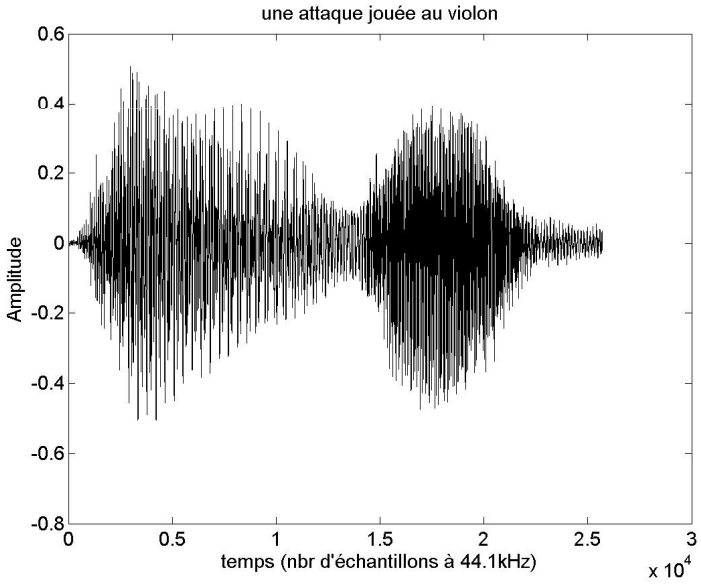




Paramétrisation: Paramètres temporels

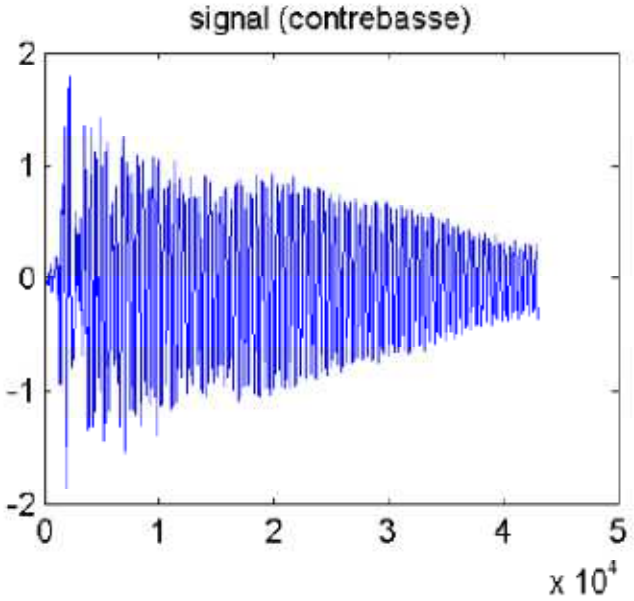
- **Evolution temporelle:**
  - Enveloppe de l'attaque est caractéristique du type de son (corde frottée, frappée, grattée…
    - Paramètres possibles: durée de l'attaque (ou impulsivité de l'attaque)

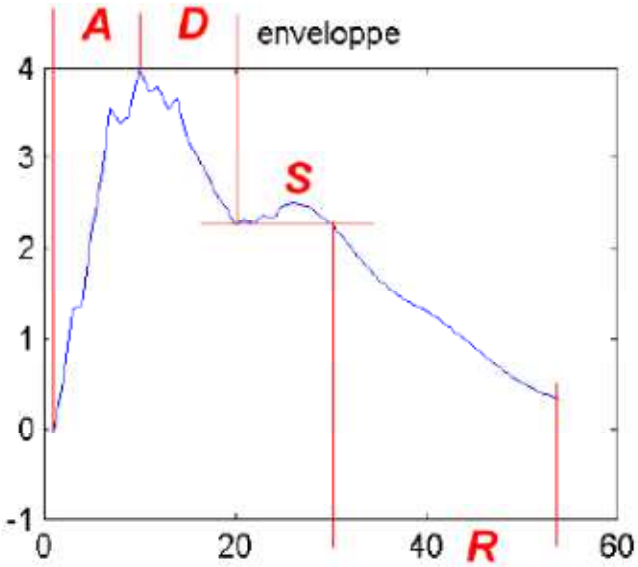




Paramétrisation: Paramètres temporels

- **Enveloppe: possibilité d'utiliser un modèle**
  - Modèle ADSR
  - Enveloppe peut être obtenue par filtrage passe-bas de l'énergie



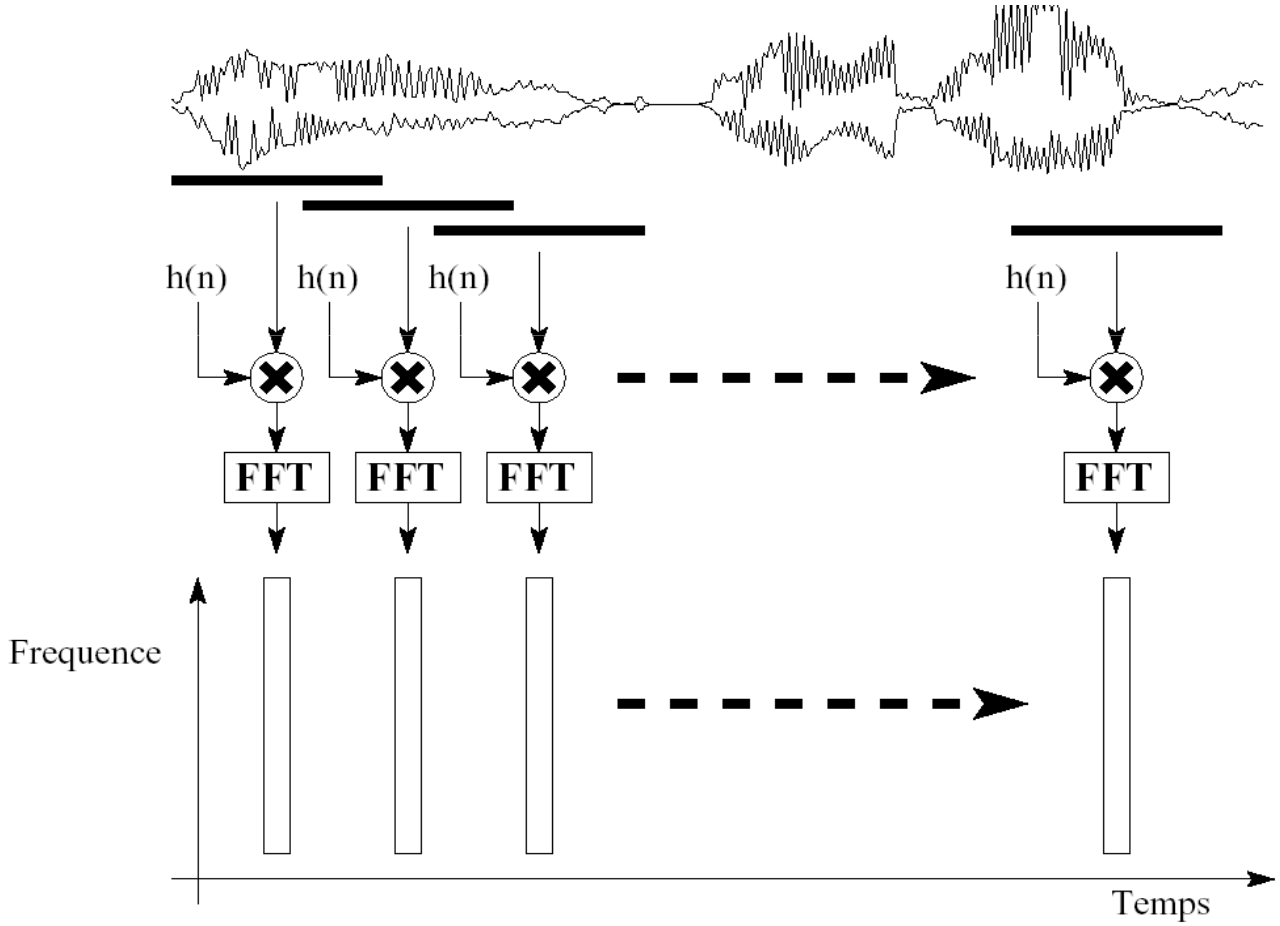


Paramétrisation: Paramètres temporels

- **Autres paramètres temporels utilisés:**
  - Modulation d'amplitude (4 Hz, ou 10-40 Hz)
  - Facteur crête
  - Impulsivité du signal (moment d'ordre 4)
  - Période fondamentale (ou inversement fréquence fondamentale)
  - …..

Paramétrisation: paramètres spectraux

- **Paramétrisation spectrale: analyse d'un signal audio (d'après Laroche)**



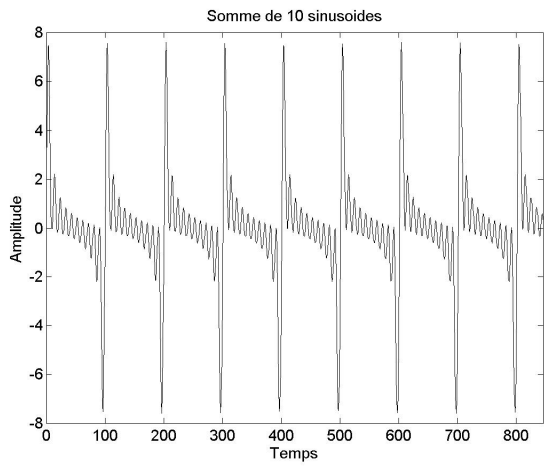
Paramètres spectraux

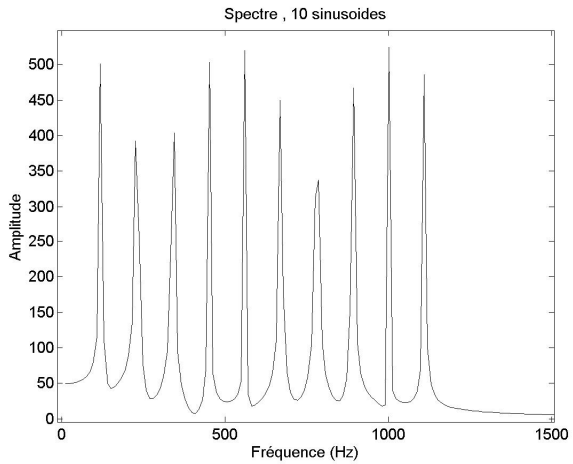
- **Représentation temps-fréquence : Transformée de Fourier**

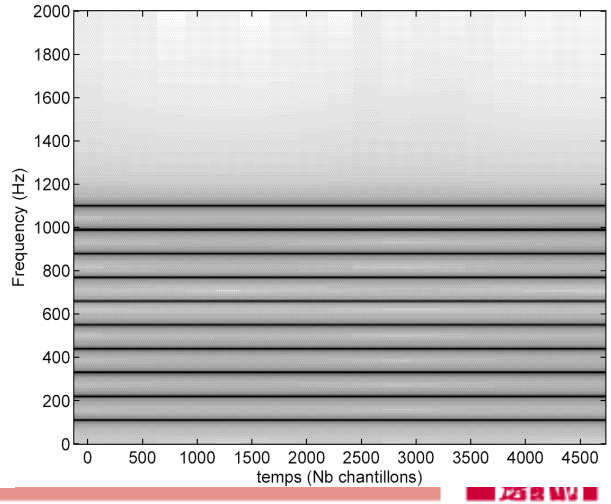
$$X_k = \sum_{n=0}^{N-1} x_n e^{-2j\pi nk/N}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{2j\pi nk/N}$$

x_n $|X_k|$ *Spectrogramme*





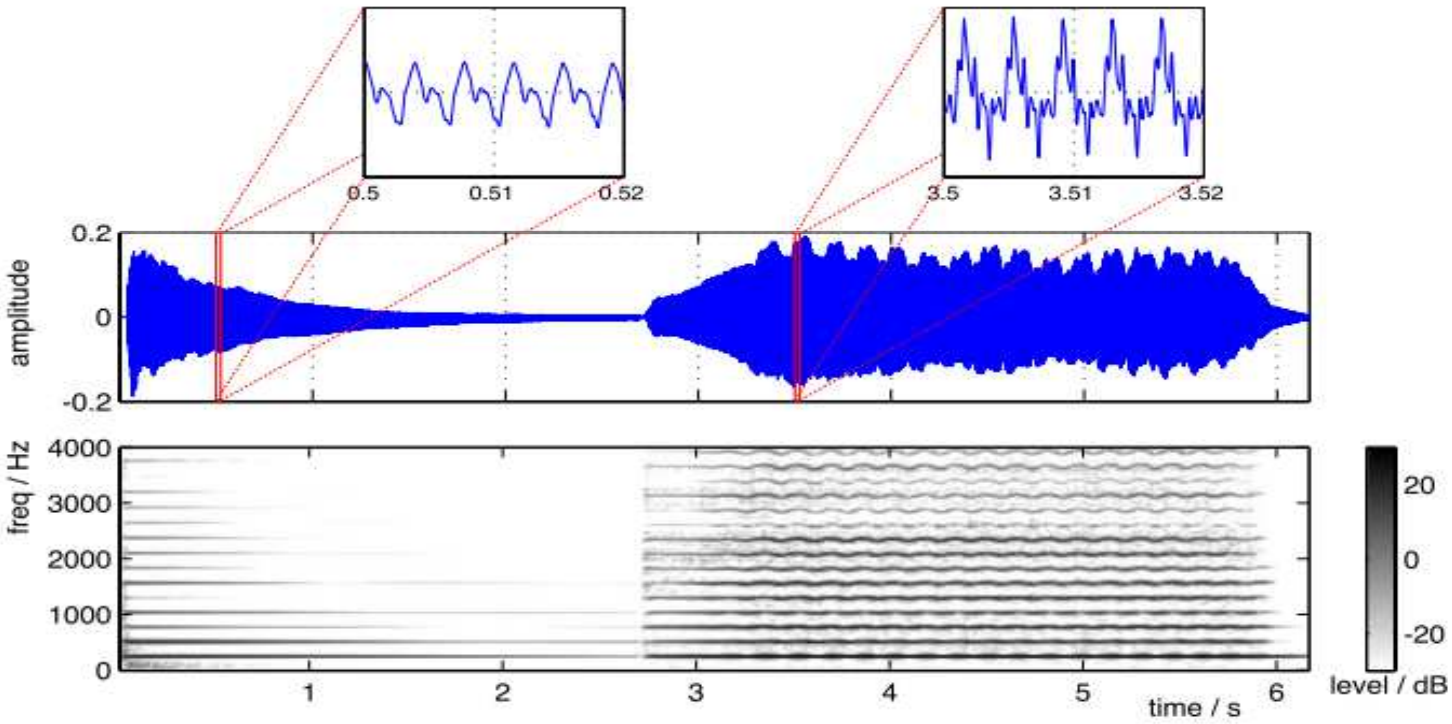


Représentations du signal audio

- **Exemple sur un signal audio: note Do (262 Hz) jouée par un piano et un violon.**

Signal temporel

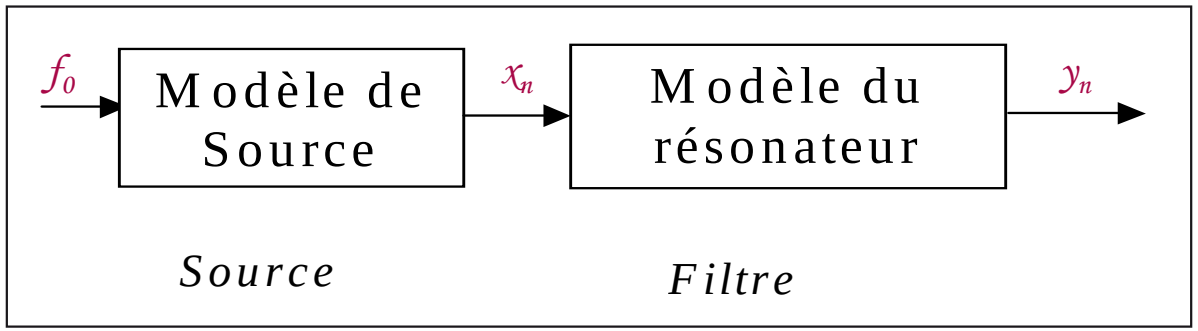
Spectrogramme

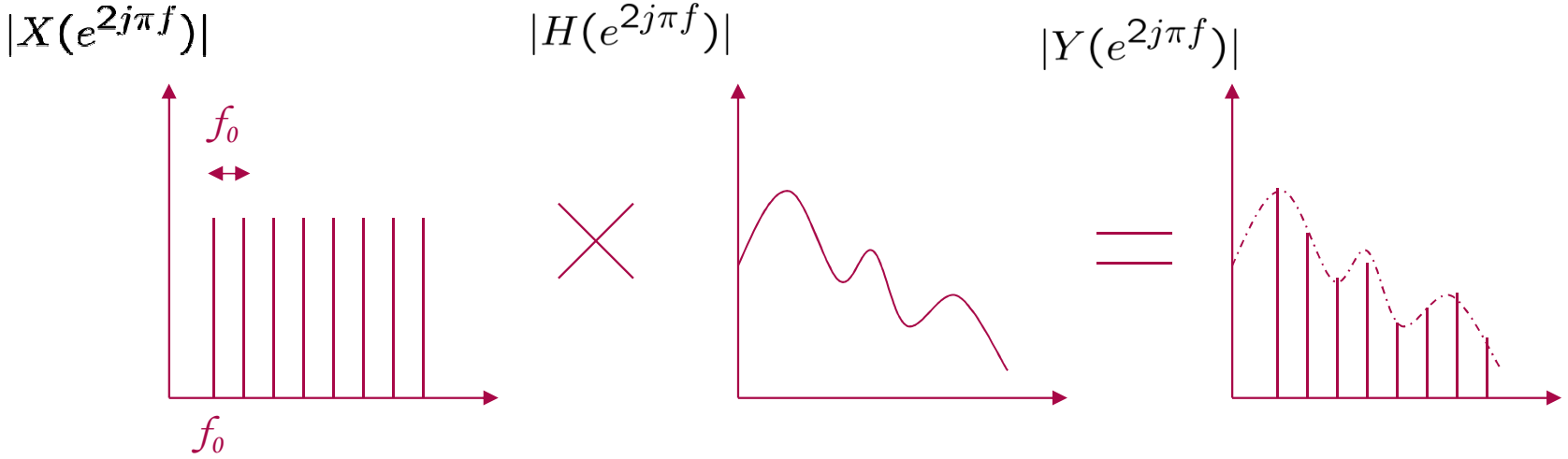


D'après M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Modèle source-filtre

- enveloppe spectrale, source





Paramétrisation spectrale

- **Le Centre de Gravité Spectral (CGS)**

$$CGS = \frac{\sum_{k=1}^{N} k.|X_k|}{\sum_{k=1}^{N} |X_k|}$$

  - CGS élevé: son brillant
  - CGS faible: son chaud, rond

- **Le flux spectral (« variation temporelle du contenu spectral »)**

$$Flux = \sum_{k=1}^{N} (|X_k(t)| - |X_k(t-1)|)^2$$

Paramétrisation (suite)

- **Rayon de Giration Spectral**

$$RGS = \sqrt{\frac{\sum_{k=1}^{N}(k - CGS)^2.|X_k|}{\sum_{k=1}^{N}|X_k|}}$$

  - RGS faible, le timbre est « compact »

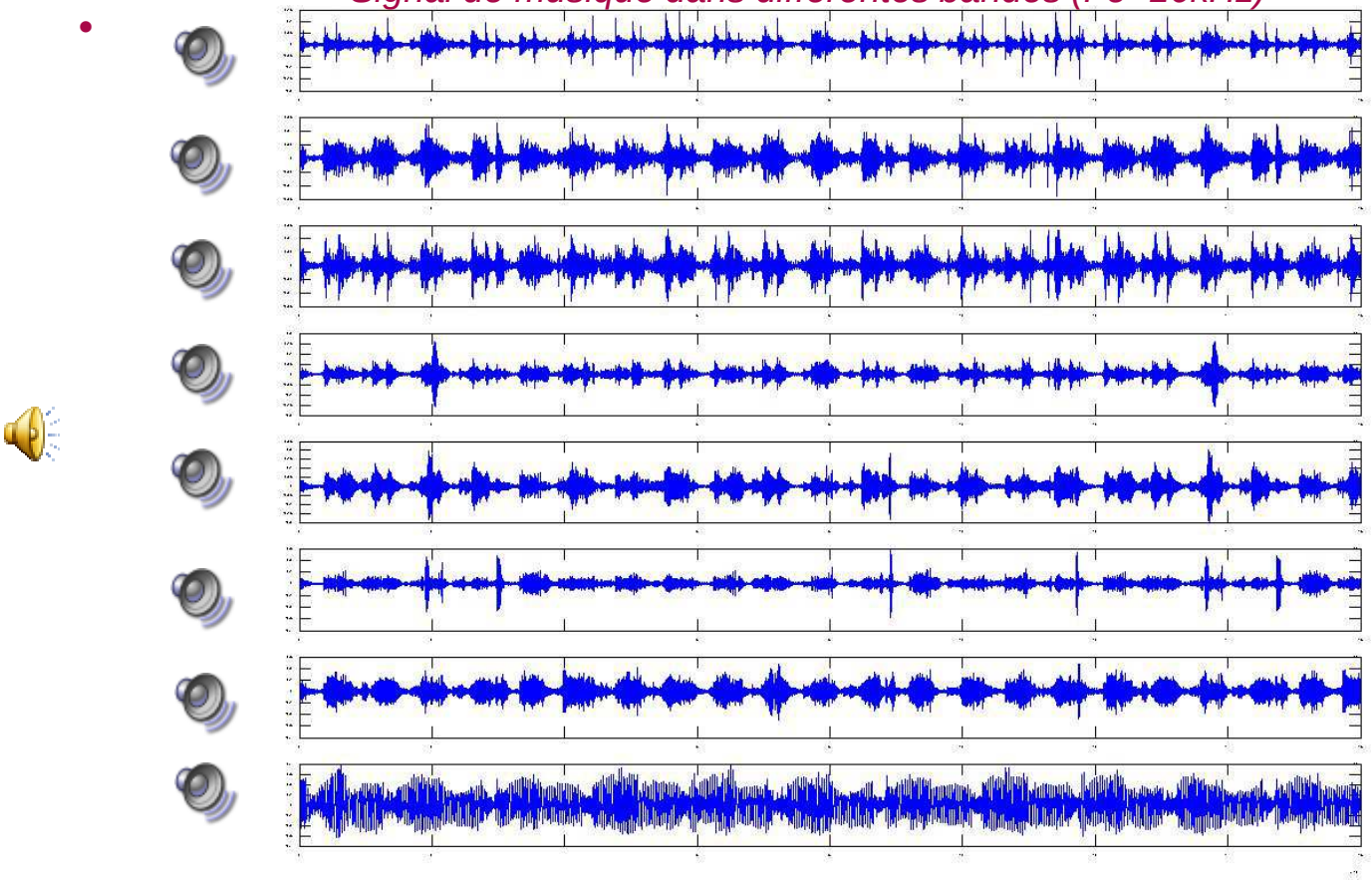
- **« Coupure spectrale » (*Spectral Roll off*) définit la fréquence R_t au dessous de laquelle 85% de la distribution spectrale est concentrée:**

$$\sum_{k=1}^{R_t}|X_k| = 0.85 \times \sum_{k=1}^{N}|X_k|$$

Utiliser un banc de filtres: intérêt

- **Décomposer le signal en bandes de fréquences…**

Signal de musique dans différentes bandes (Fe=16kHz)



Paramétrisation

- **Intérêts d'une analyse par un banc de filtres**

  - Permet de séparer les informations localisées en fréquence
  - Permet une réduction de complexité (sous-échantillonnage dans chaque bande)
  - Cas particulier: FFT

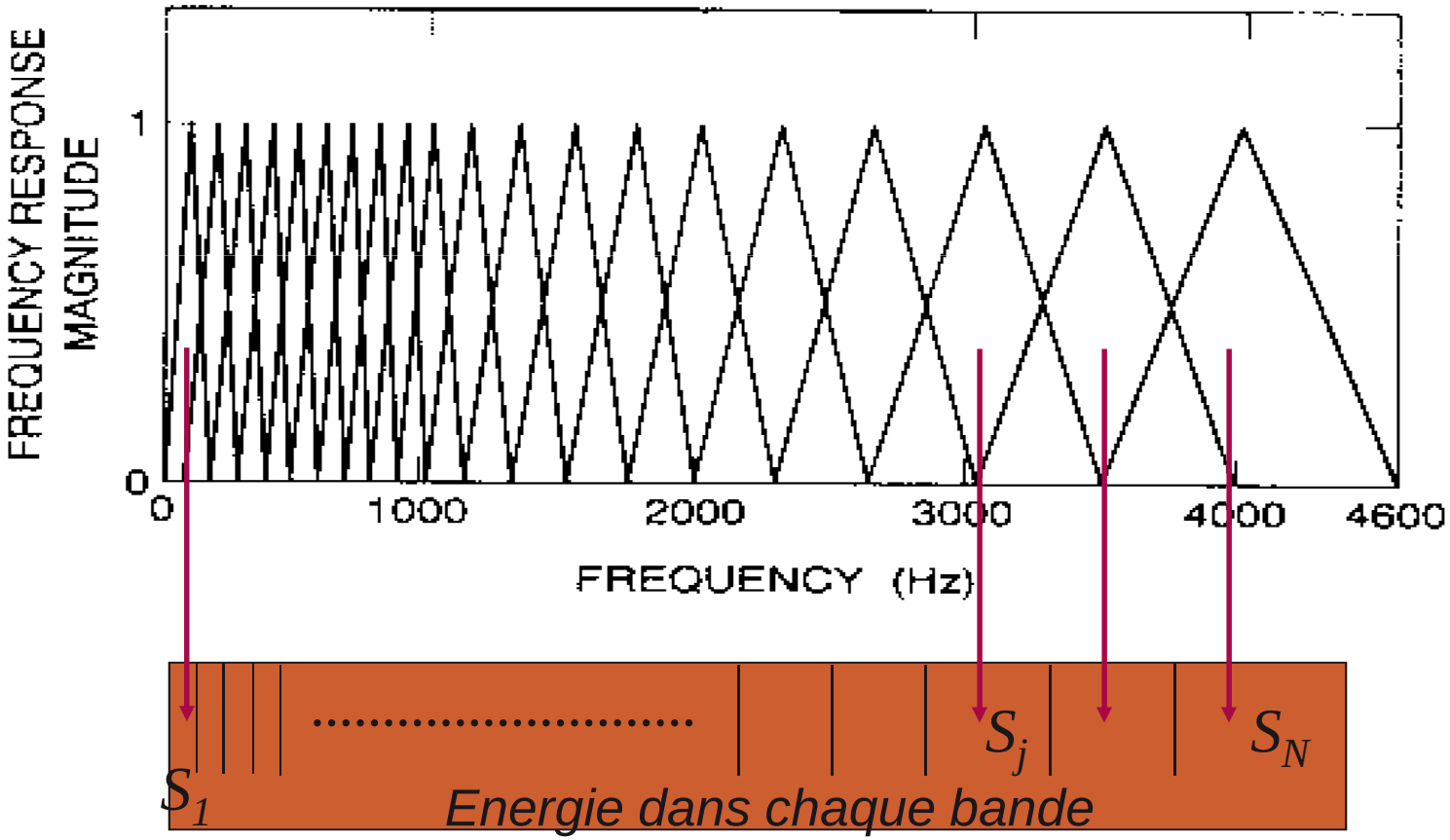
  - Possibilité d'utiliser des échelles de fréquences « perceptives »

    - Echelle Mel: Correspond à une approximation de la sensation psychologique de hauteur d'un son (Tonie)

$$mel(f) = 1000 \log_2(1 + \frac{f}{1000})$$

Filtre en échelle Mel

- **Filtrage Mel (d'après Rabiner93)**



Echelle Mel

- **Correspond à une approximation de la sensation psychologique de hauteur d'un son (Tonie)**

- **Existence de formules analytiques:**

$$mel(f) = 1000 \log_2(1 + \frac{f}{1000})$$

- **Exemples:**
  - Gamme mel

Gamme Hertz

Représentation cepstrale

- **Intérêt**
  - Modèle source filtre de la parole/des signaux musicaux

$$s(t) = g(t) * h(t)$$

✓ Modèle source filtre dans le domaine spectral

$$S(\omega) = G(\omega)H(\omega)$$

✓ Cepstre (réél): somme de 2 termes

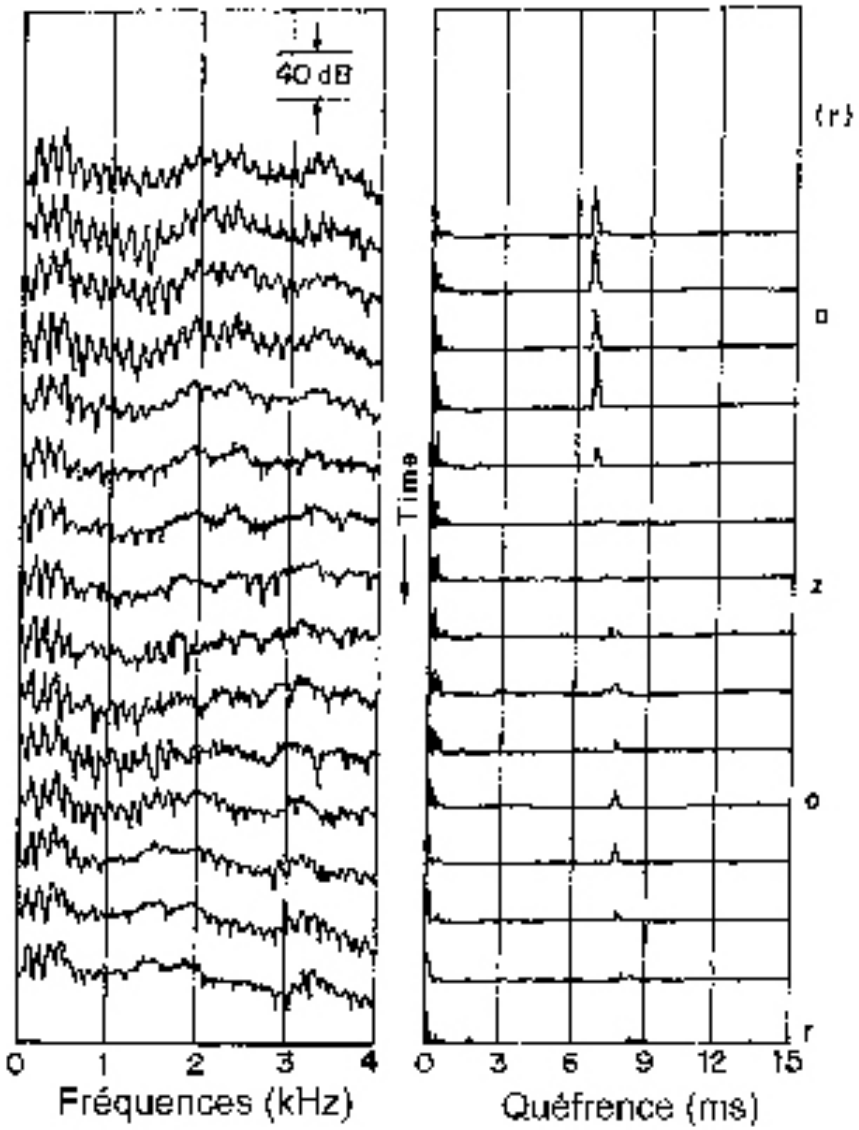
$$c(\tau) = FFT^{-1} \log |S(\omega)| = FFT^{-1} \log |G(\omega)| + FFT^{-1} \log |H(\omega)|$$

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{2j(\pi)kn/N}$$

Représentation cepstrale(d'après Furui2001)

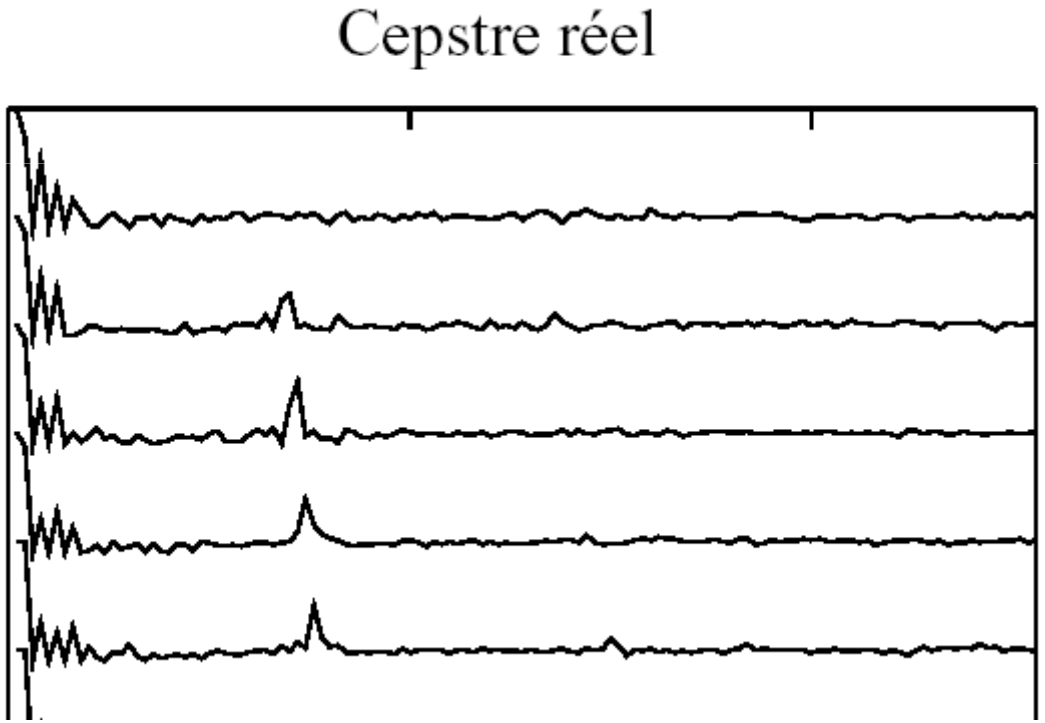
- **Exemples:**
  - de Spectres à court terme (gau
  - et de cepstre c(τ) (droite)

- **τ est homogène à un temps**



Représentation cepstrale

- **Séparation de la contribution du filtre (conduit vocal ou instrument) et de la source par liftrage**



Représentation cepstrale

- **Contribution de la source**

$$p_n = \sum_{i=-\infty}^{\infty} \alpha_i \delta(n - iT) \quad \longrightarrow \quad \hat{p}_n = \sum_{i=-\infty}^{\infty} \beta_i \delta(n - iT)$$

- **Contribution du conduit vocal/instrument**

(hypothèse: filtre causal, stable, minimum de phase)

$$F(z) = K \frac{\prod_{j=1}^{M}(1 - a_j z^{-1})}{\prod_{j=1}^{N}(1 - b_j z^{-1})} \qquad \begin{aligned} |a_j| &< 1 \\ |b_j| &< 1 \end{aligned}$$

Représentation cepstrale

- **Contribution du conduit vocal/instrument**

- **Développem**

$$\log(F(z)) = \sum_{n=0}^{\infty} c_n z^{-n}$$

$$\log(1-a) = -\sum_{n=1}^{\infty} a^n/n \quad pour \quad |a| < 1$$

$$\hat{c}_n = \begin{cases} \log(K) & n = 0 \\ -\sum_{j=1}^{M} \frac{a_j^n}{n} + \sum_{j=1}^{N} \frac{b_j^n}{n} & n > 0 \end{cases} \qquad |z| > |a_j|, |b_j|$$

Représentation cepstrale

- **Exemples de liftres (d'après Calliope89)**

(1) filtre rectangulaire

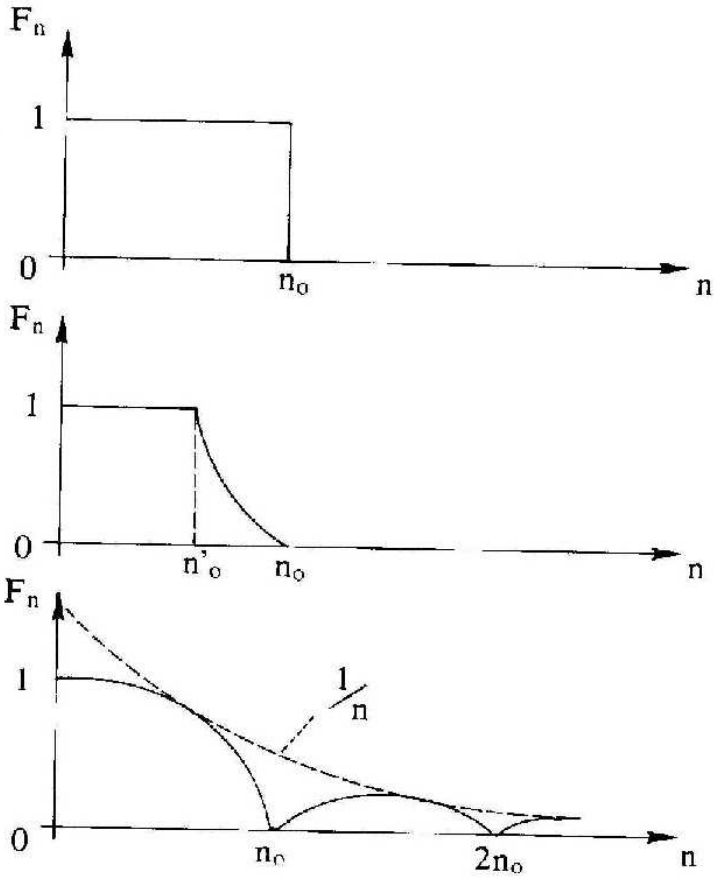
$$\begin{cases} F_n = 1 & \text{si } n < n_0 \\ F_n = 0 & \text{si } n \geqslant n_0 \end{cases}$$

ou (2) filtre adouci

$$\begin{cases} F_n = 1 & \text{si } n < n'_0 < n_0 \\ F_n = 1 - e^{-\alpha(n-n'_0)} & \text{si } n \geqslant n'_0 \end{cases}$$

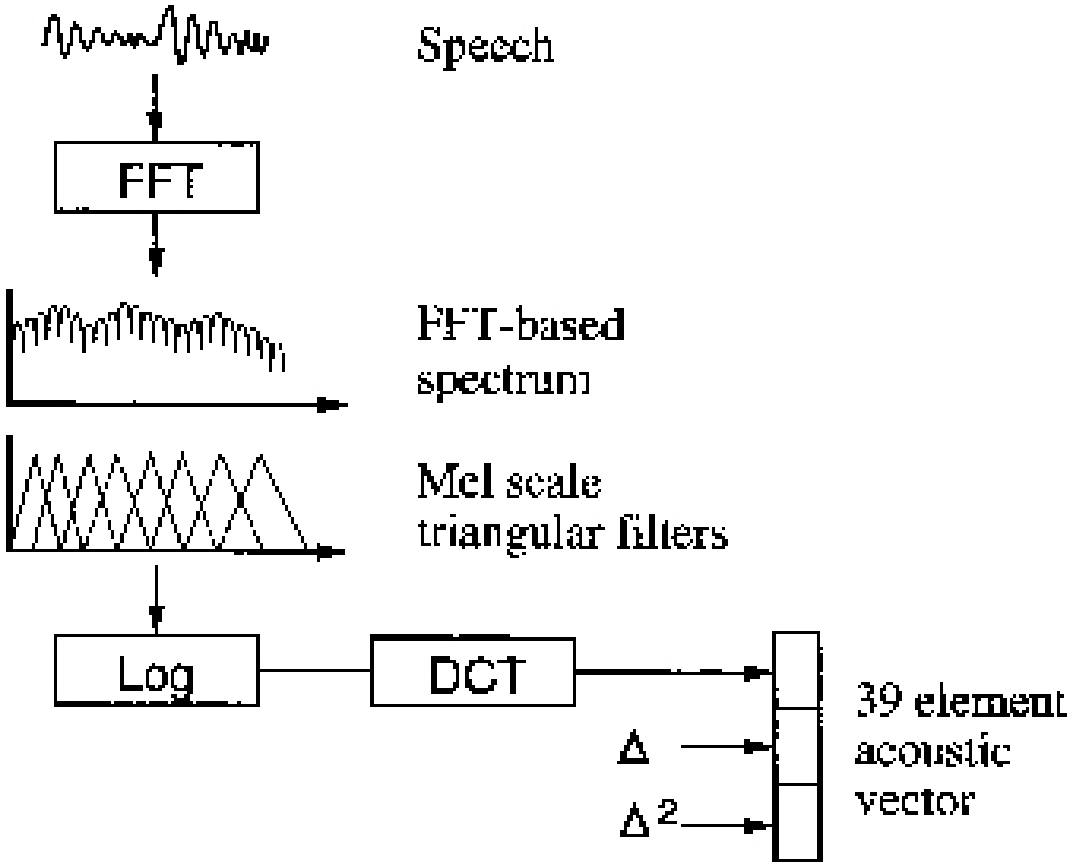
ou (3) filtre de Combs

$$F_n = \hat{C}_n - C_{n-n_0}$$



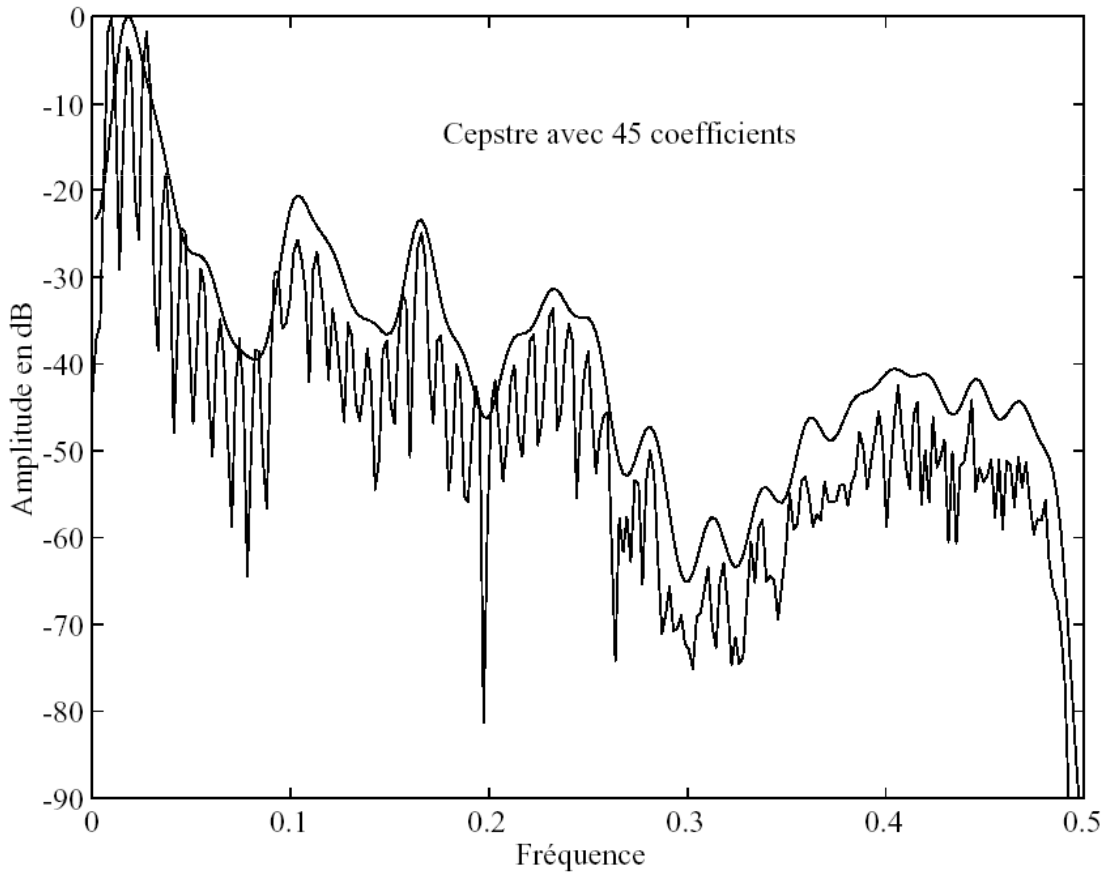
Paramétrisation (*issue de la reconnaissance vocale*)

- **Paramétrisation MFCC (*Mel-Frequency Cepstral Coefficients*)**



Lissage cepstral

- **Estimation de l'enveloppe par le cepstre:**
  - Calcul du cepstre réel $C_{n,}$ puis lifrage basses quéfrences
  - Reconstruction de l'enveloppe spectrale d'amplitude $E = FFT(C_n)$



Modélisation « Sinusoïdes + bruit »

- **Basé sur le modèle *Exponentially Damped Sinusoidal (EDS)***
  avec

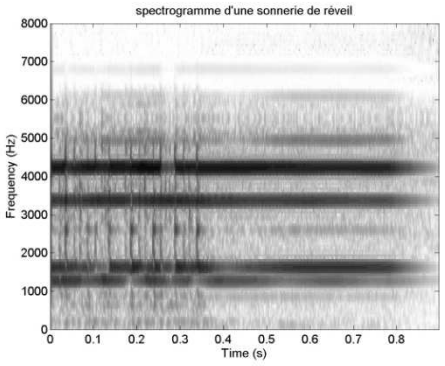
$$s(t) = \sum_{i=1}^{N} e^{-\alpha_i t} e^{j\omega_i t + \phi_i}$$

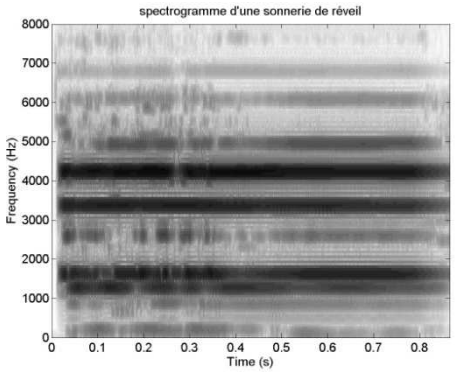
$$x(t) = s(t) + w(t)$$

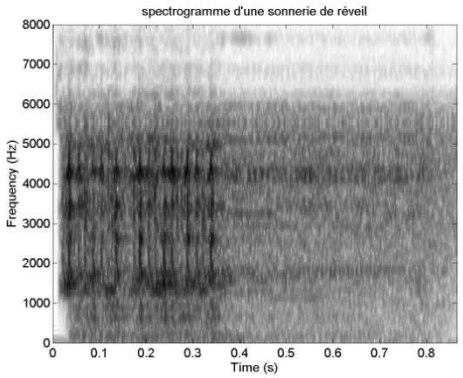
- *Original = Somme de sinusoïdes + Bruit*

(.wav)

(.wav)

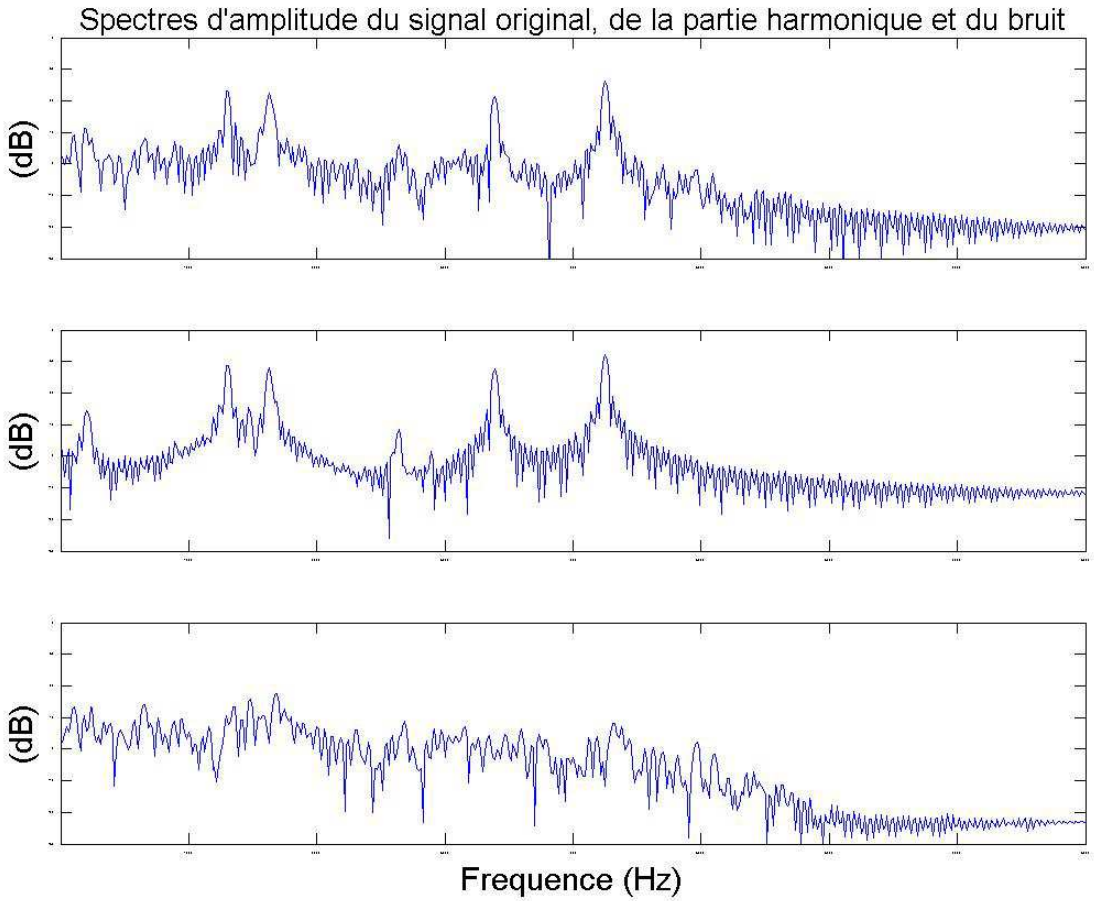






Paramétrisation (suite…)

- **Séparation harmoniques / bruit**



Séparation harmoniques / bruit

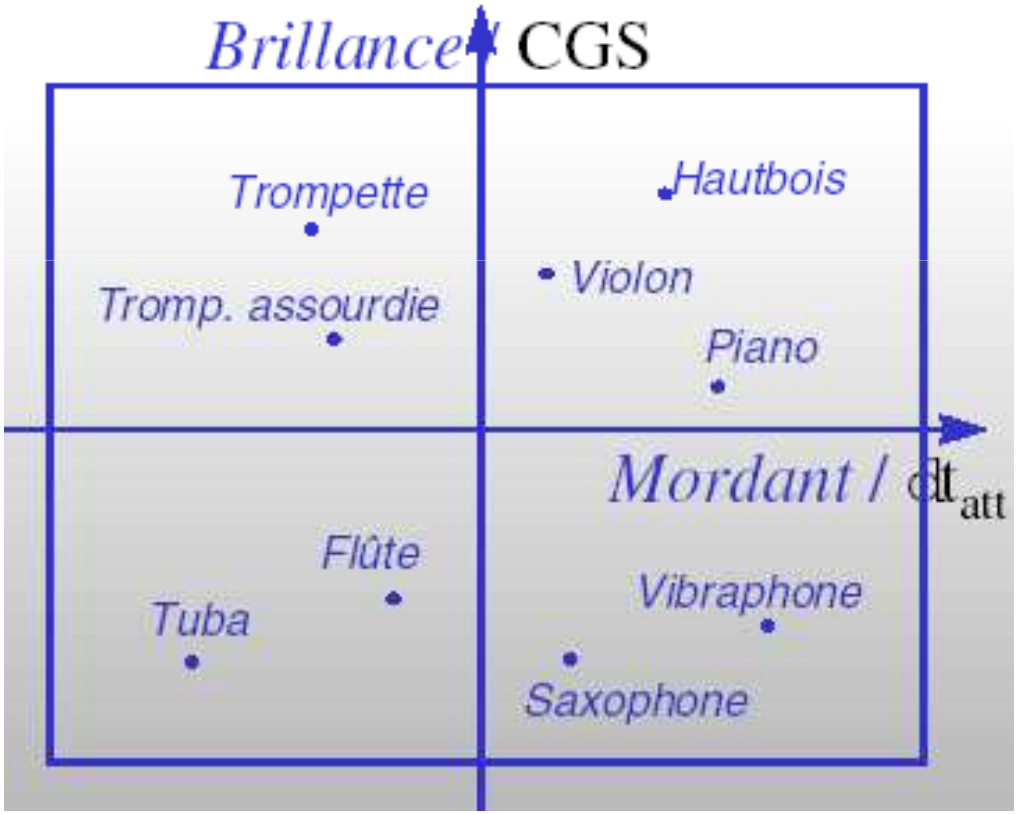
- **Approche à partir de la STFT**
  - Découpage en fenêtres
  - Calcul de la STFT
  - Localisation et Détection des maxima (« harmoniques »);
  - Synthèse des parties harmoniques et bruit par *Overlap and Add*

  - *Précautions:*

    - *Zero padding pour limiter l'effet de la convolution circulaire*

Le timbre des instruments de musique

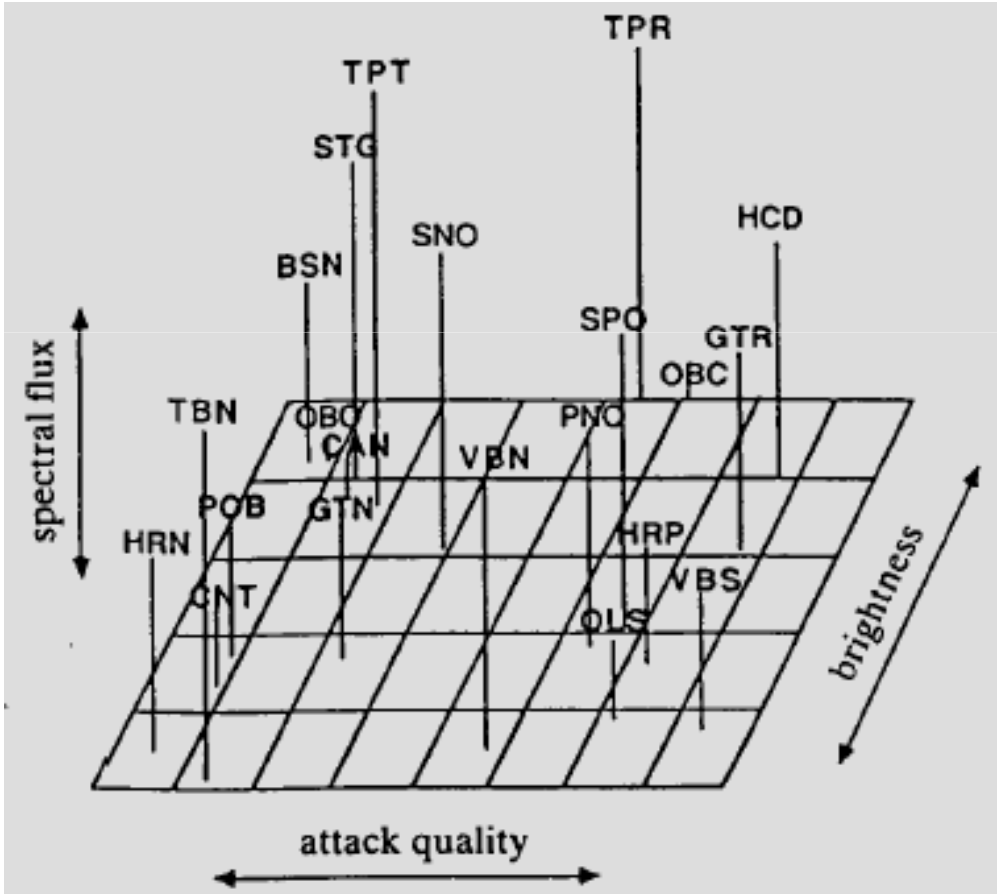
- « Espace » de timbre



Le timbre des instruments de musique

- Espace 3D (Krumhansl, 1989, McAdams, 1992)

- BSN = basson
- CAN = cor anglais
- CNT = clarinet
- GTR = guitar
- HRN = cor
- HRP = harpe
- TPT = trompette
- PNO = piano
- VBS = vibraphone



Autres paramètres utilisés en indexation audio

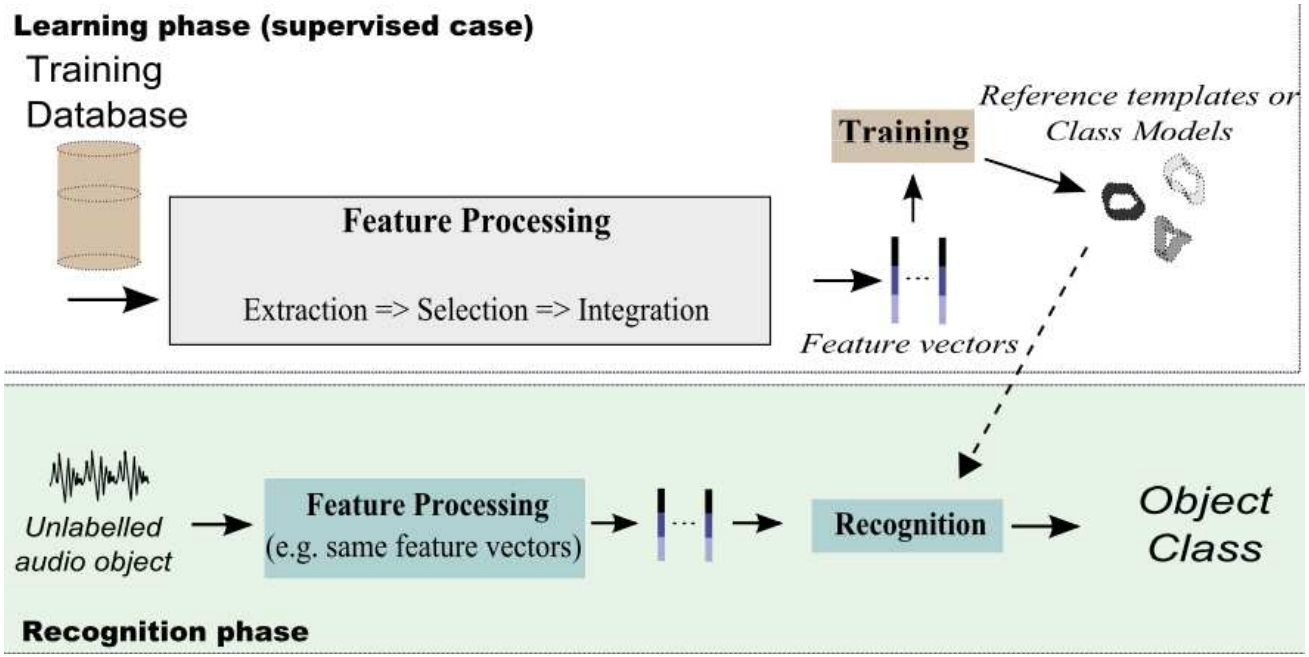
- **Warped Linear prediction Cepstral coefficients**
- **« Asynchronie » fréquentielle des attaques**
- **Coefficients d'ondelettes**
- **Séparation harmonique-bruit**
- **Entropie,**
- **Variation de l'entropie,**
- **….**
  - Pas de réel consensus sur la bonne paramétrisation

↳Une voie de recherche: utiliser un nombre important de caractéristiques (features) et utiliser des algorithmes de sélection de « features »

Eléments de classification

Exemple de la reconnaissance automatique des instruments de musique

- ***Objectif de la classification:***
  - Permettre de retrouver la classe (i.e l'instrument) à partir des paramètres extraits du signal



Classification

- **Quels sont les paramètres appropriés pour une tâche donnée?**
  - Pas de consensus
- **Une approche possible:**
  - Utiliser un nombre important de caractéristiques
  - Utiliser des techniques d'analyse de caractéristiques ou/et de sélection de caractéristiques pour réduire la dimension des vecteurs de caractéristiques.

Analyse des caractéristiques

- ***Principal Component Analysis* (PCA)**
  - But: « débruiter les données » et « réduire l'espace des paramètres »

- Etape 1: Décomposition en valeurs singulières (SVD)

$$\mathbf{R_x} = \mathbf{U}\mathbf{D}\mathbf{V}^t$$

$\mathbf{R_x}$ Matrice de covariance

$\mathbf{D}$ Matrice des valeurs singulières

- Etape 2: Les modèles sont entraînés sur les données « transformées »:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

$$\mathbf{W} = \mathbf{V}^t$$
$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$$

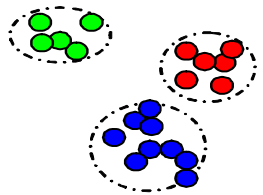
Sélection de caractéristiques

- **De nombreux algorithmes existent**
- **Un algorithme simple mais efficace basé sur le discriminant de Fisher**

Le principe est intuitif: « Sélectionner les caractéristiques une par une qui permettent une bonne séparation entre les classes en conservant une dispersion intra-classe minimale».

- Ainsi, la caractéristique sélectionnée correspond au plus fort ratio:

$$r_i = \frac{B_i}{R_i} = \frac{\sum_{k=1}^{K} \frac{N_k}{N} \|\mathbf{m}_{i,k} - \mathbf{m}_i\|}{\sum_{k=1}^{K} \left(\frac{1}{N_k} \sum_{n_k=1}^{N_k} \|\mathbf{x}_{i,n_k} - \mathbf{m}_{i,k}\| \right)}$$



Où Bi est l'inertie entre les classes et Ri le rayon moyen de la dispersion de chaque classe

Principe des systèmes de classification

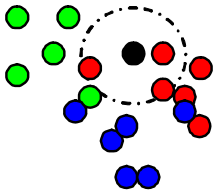
- **Définition du problème**
  - Objets décrits par un vecteur de paramètres *x* de R^n et par une classe *y* dans [1,C].
  - Ensemble d'apprentissage (exemples), $A = (x_i, y_i)_{i \in [1,N]}$

- **Apprentissage**
  - Associe à un ensemble d'apprentissage une fonction de décision $f : R^n \rightarrow [1, C] :$
    - **Couvrant les exemples, ie** $f(x_i) = y_i$
    - **Généralisant,** $f(x) = f(x_i)$ **si *x* est associé au même phénomène que xi**.
  - La fonction de décision *f* va permettre de classer de nouveaux objets qui ne figurent pas parmi les exemples.

Algorithme des k-plus proches voisins

- **Choix d'une distance dans l'espace de description**
  - Ex: une distance euclidienne

$$D(x_i, x_j) = \sqrt{\sum_{k=1}^{N} (x_i^k - x_j^k)^2}$$



- **Choix du nombre de voisins *K* à considérer**

- **Algorithme: Soit x_k le vecteur à classer, *X* l'ensemble des vecteurs de la base d'apprentissage:**
  - Chercher $p_1 .. p_k$ les ***K*** plus proches voisins de x_k
  - La classe reconnue est donnée par:

$$\underset{c \in C}{\operatorname{argmax}} \sum_{k=1}^{K} \delta(c, Y(x_k))$$

Les k-plus proches voisins

- **Exemples de Performance (Eronen2001) :**

  - Base de données: 1500 exemples sonores monophoniques
  - Identification de la famille d ’instruments: 94 %
  - Identification de l ’instrument 80 %

- **Avantages**
  - Simplicité de mise en œuvre

- **Inconvénients**
  - Pas de généralisation (seulement basée sur une information locale)
  - Très sensibles aux éléments extrêmes
  - Nécessitent d’avoir tous les vecteurs d ’entraînement en mémoire et donc complexes en temps calcul
  - Pas de mesure de confiance: impossible de savoir si le classifieur est sûr de lui ou non

Classifieurs: Décision bayésienne

- **Idée générale**
  - On aimerait pouvoir calculer *P(y = c|x),* c'est à dire la probabilité que l'objet à classifier appartienne à une classe donnée *c*, connaissant son vecteur de paramètres x.
  - Il serait alors possible d'effectuer une décision avec la règle suivante : on associe la classe la plus probable, conditionnellement aux paramètres observés :

$$y = argmax_{c \in [1;C]} P(y = c|x)$$

  - *Question: Comment calculer P(y = c|x) ?*

Classifieurs: décision bayésiennes

- **Règle de Bayes**

$$P(y=c|x) = \frac{P(y=c)p(x|y=c)}{p(x)}$$

- **Simplifications**
  - Comme on veut maximiser *P(y=c|x) (e.g. maximum a posteriori),* on peut ignorer *p(x).*
  - *Si on suppose les classes équiprobables [P(y=c) = constante*], on se ramène au maximum de vraisemblance.

- *et l*e problème se résume donc à calculer *p(x|y=c)*, c'est à dire à estimer **la densité de probabilité** du vecteur de paramètres pour chacune des classes

Classifieurs: approche paramétrique

- **Approche paramétrique:**
  - On suppose que *p(x|y = c)* suit une loi connue, dont on va déterminer les paramètres.

- **Loi Gaussienne**

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} |\mathbf{\Sigma}_i|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)\right]$$

Approche par Mélanges de Gaussiennes

(Voir Cours O. Cappé http://tsi.enst.fr/~ocappe/em_tap.pdf)

Modèle de mélange

$$f(\mathbf{x}) = \sum_{i=1}^{M} \pi_i f_i(\mathbf{x})$$

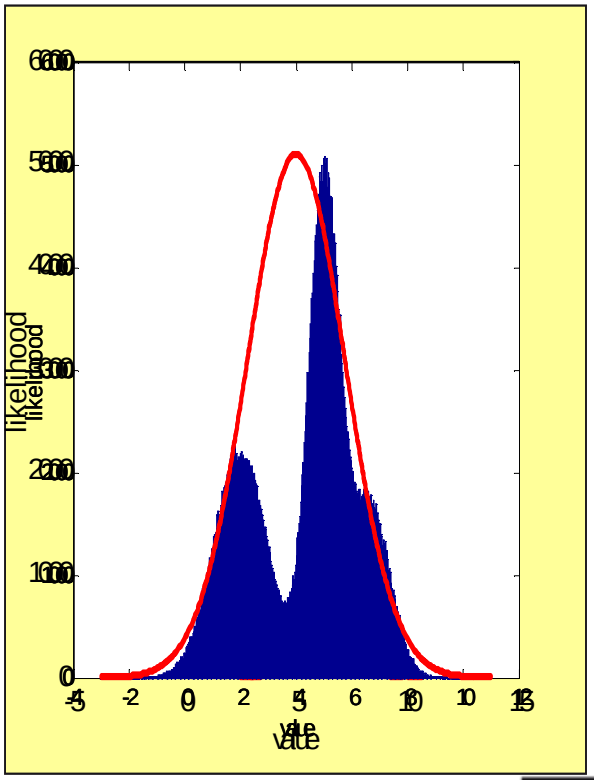
$f_i(\mathbf{x})$ est une densité de probabilité

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} |\mathbf{\Sigma}_i|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

π_i sont des scalaires positifs

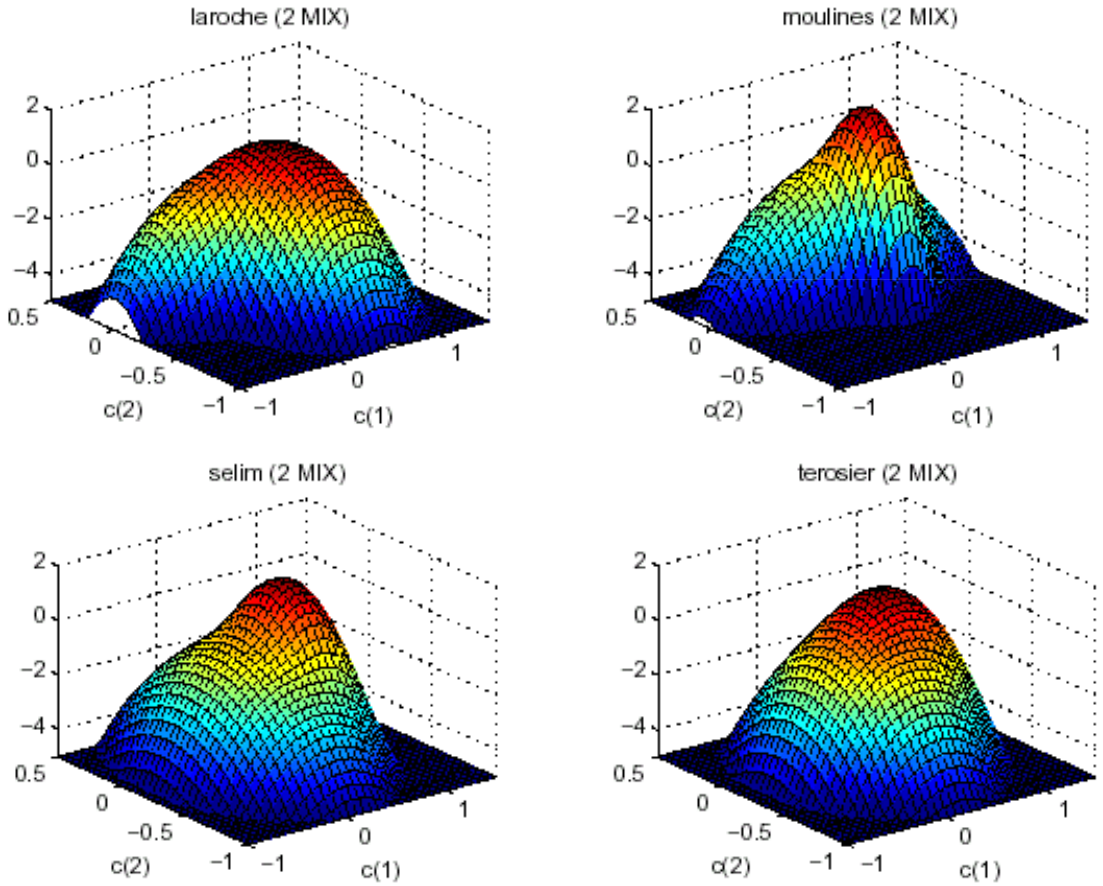
$$\sum_{i=1}^{M} \pi_i = 1$$

- Exemple à 2 dimensions avec 1, 2 puis 3 gaussiennes



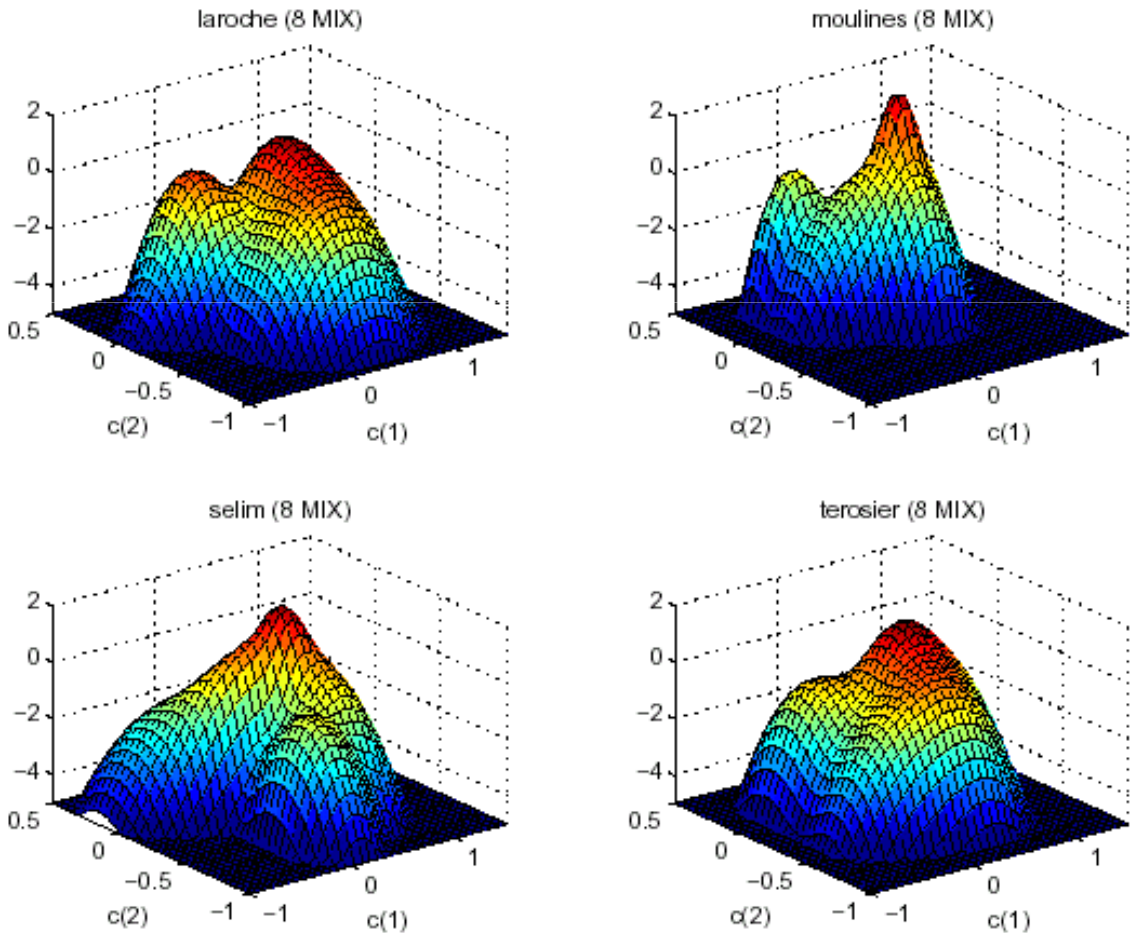
Approche par Mélanges de Gaussiennes (GMM)

- **Exemple de modèles à 2 composantes**



Approche par Mélanges de Gaussiennes (GMM)

- **Exemple de modèles à 2 composantes**



Identification/classification des instruments de musique

- **Classification bayesienne: mélange de gaussiennes (GMM)**
- **Paramétrisation: coefficients cepstraux (MFCC) ou obtenue par sélection de caractéristiques**
- **Modélisation des classes d'instruments:**
  - Chaque classe est représentée par un certain nombre de clusters (obtenus par l'algorithme K-means)
  - Chaque classe est représentée par une somme de gaussiennes
  - On peut alors attacher un instrument par classe

Reconnaissance des instruments de musique

- **Améliorations possibles et voies de recherche**
  - Utilisation de classifieurs à vaste marge (SVM)
  - Utilisation de sélections statistiques des paramètres
  - Modèles paramétriques avancés (décomposition parcimonieuses)

- **Vers la reconnaissance des instruments pour des ensembles instrumentaux**

Quelques dimensions du signal musical…

Hauteurs, Harmonie,..

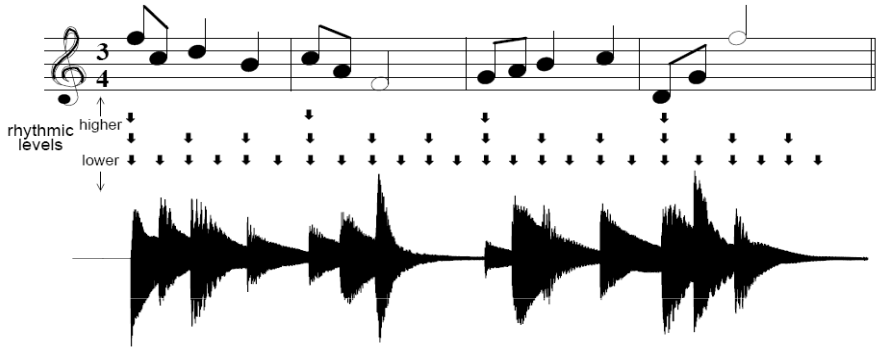
Tempo, beat, rythme,…

Timbre, instruments,…

Polyphonie, mélodie, ….

Extraction du rythme ou du Tempo

- **Le rythme: concept musical intuitivement simple à comprendre mais difficile à définir !!**

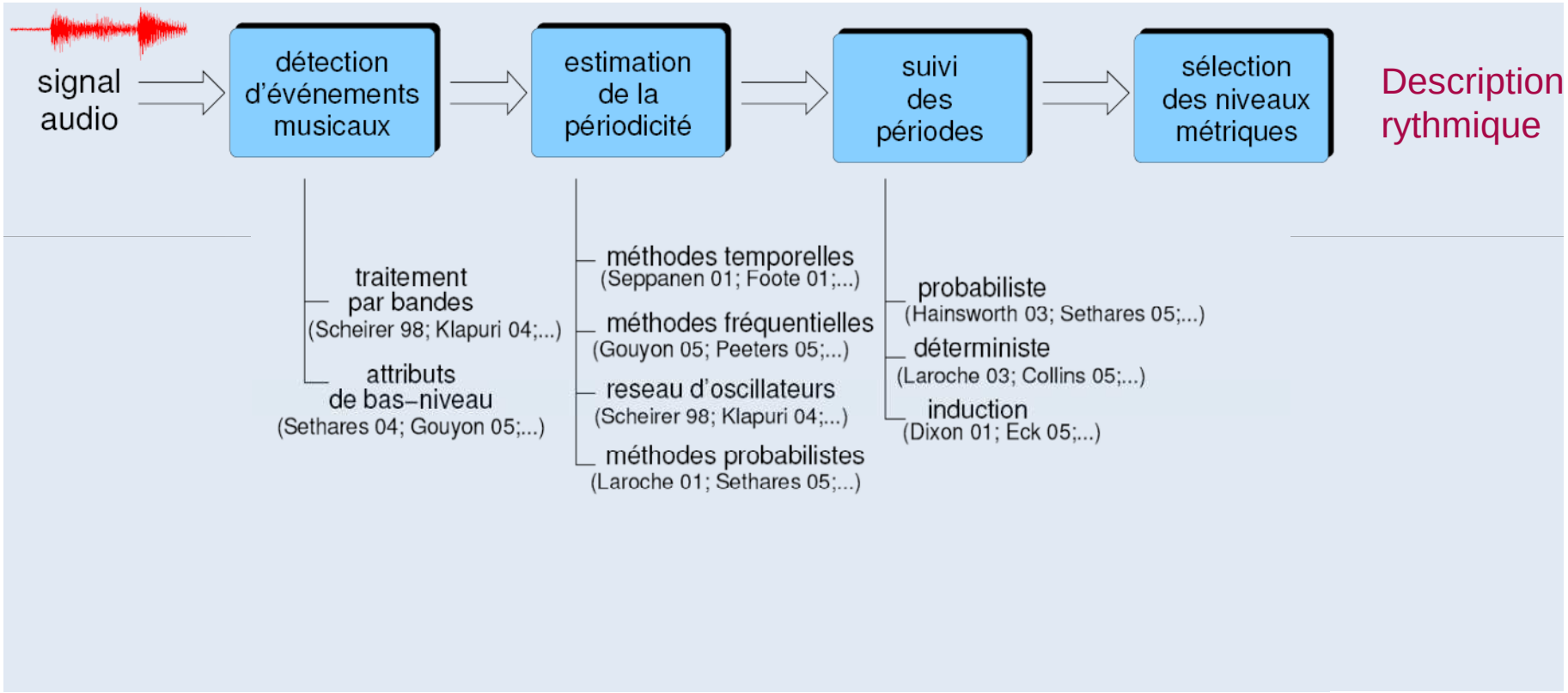


- **Handel (1989): *« The experience of rhythm involves movement regularity, grouping and yet accentuation and differentiation »***

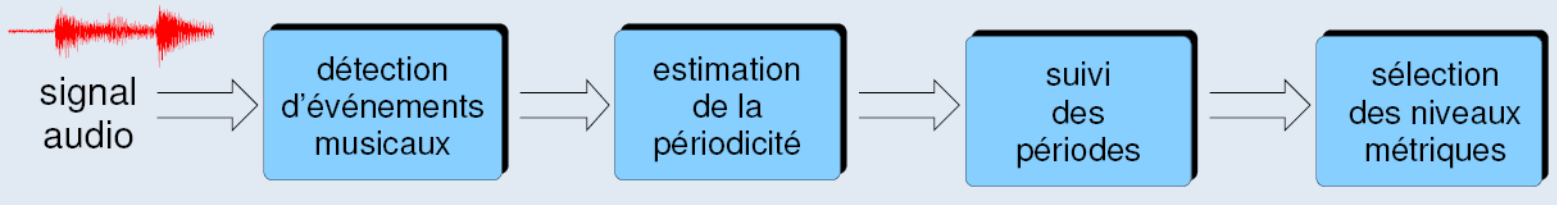
- **le rythme d'un signal écouté n'a pas nécessairement une interprétation unique !!**
- **On définit fréquemment la pulsation (beat en anglais)**

Extraction du rythme ou du Tempo

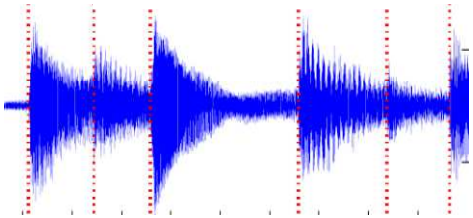
- **Principe Général**



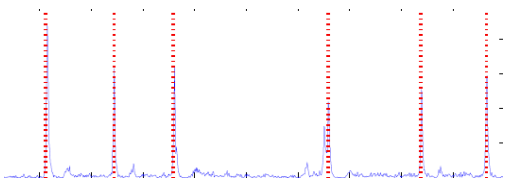
Extraction du rythme ou du Tempo



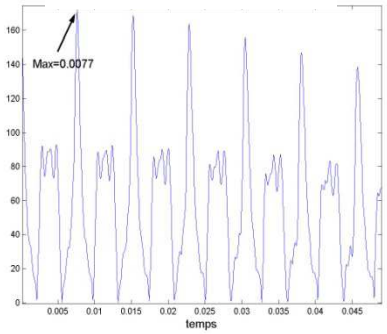
Signal + Onsets



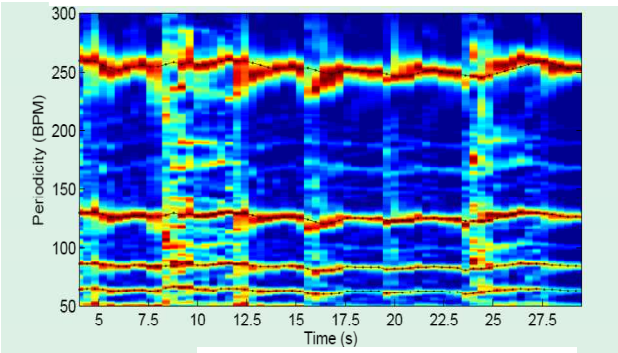
« Fonction de détection »



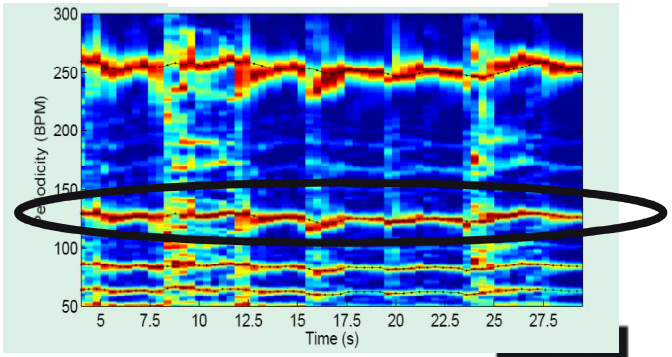
Autocorrélation



Suivi du tempo (« tempogramme»)



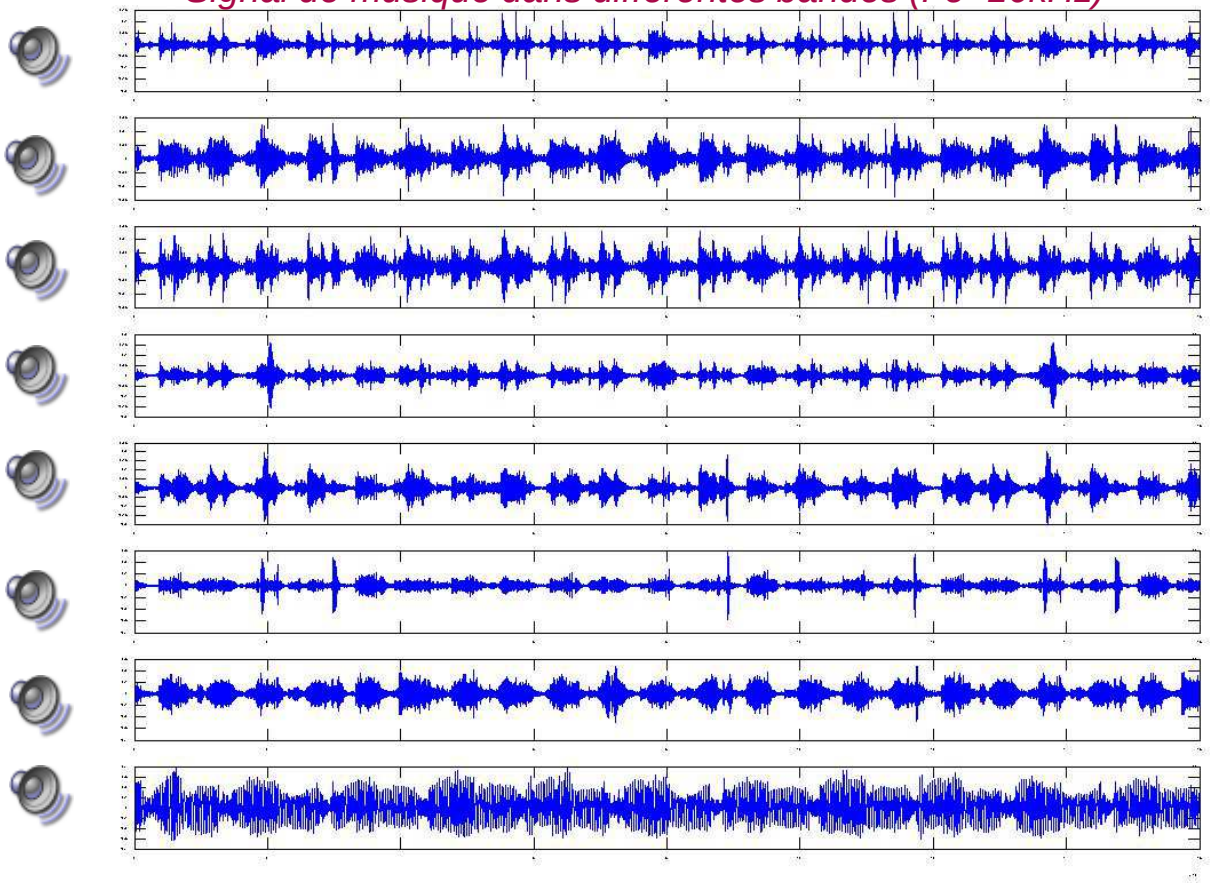
Tempo « à la noire »



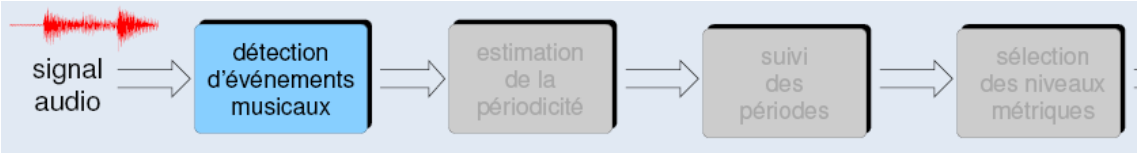
Découvrir l'information rythmique

- **Décomposer le signal en bandes de fréquences…**

Signal de musique dans différentes bandes (Fe=16kHz)

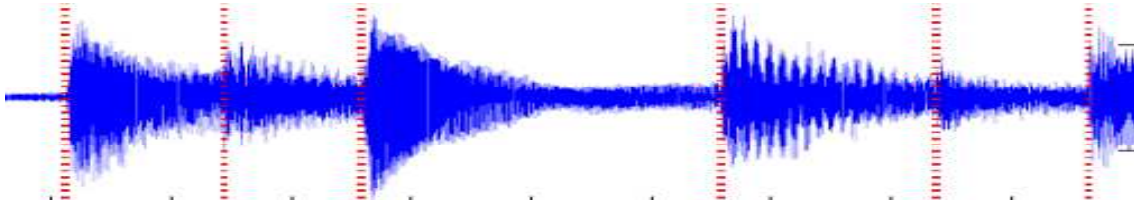


Extraction du rythme ou du Tempo

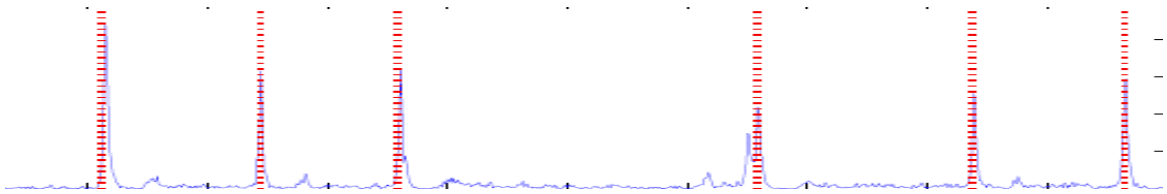


- **Principe de la détection d'événements musicaux**

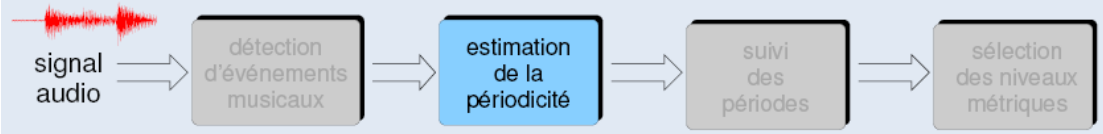
  - A partir du signal audio (dans une bande fréquentielle….)



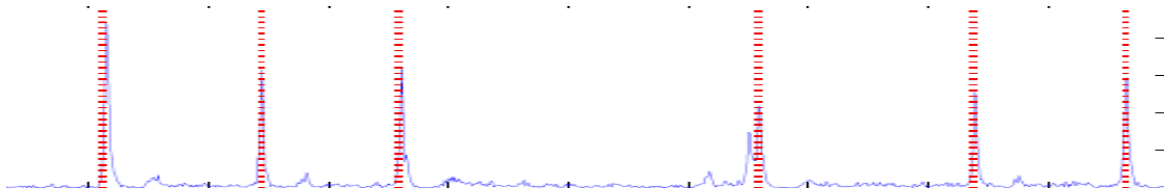
  - *…obtenir une fonction de détection*



Extraction du rythme ou du Tempo



- **Principe de l'estimation de périodicité**
  - A partir d'une *fonction de détection*



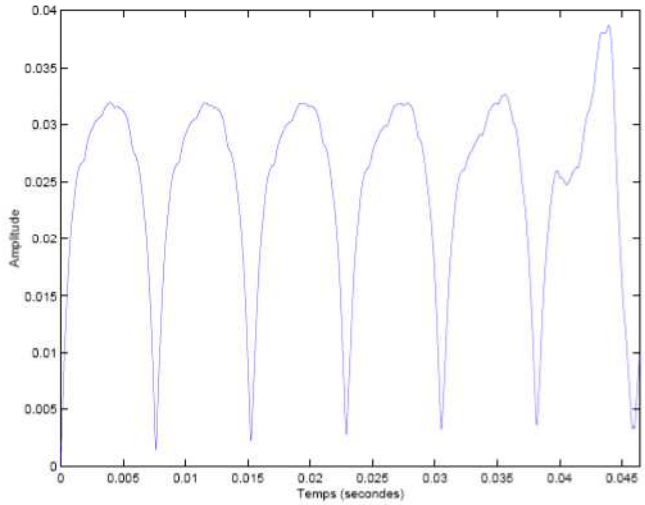
  - *… obtenir le tempo (ou « beats »)*
    - *Par estimation de la périodicité de cette fonction*

Estimation d'une périodicité

- **Une méthode temporelle**
  - Average magnitude difference function (AMDF)

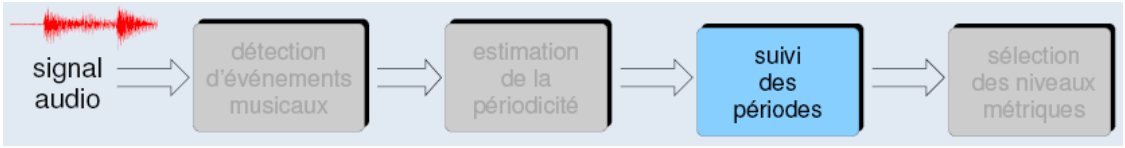
$$\mathrm{AMDF}[m] = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x[n] - x[n+m]|$$

$$\mathrm{AMDF}[m] = 0 \text{ ssi } x \text{ est de période } T_0 = m$$



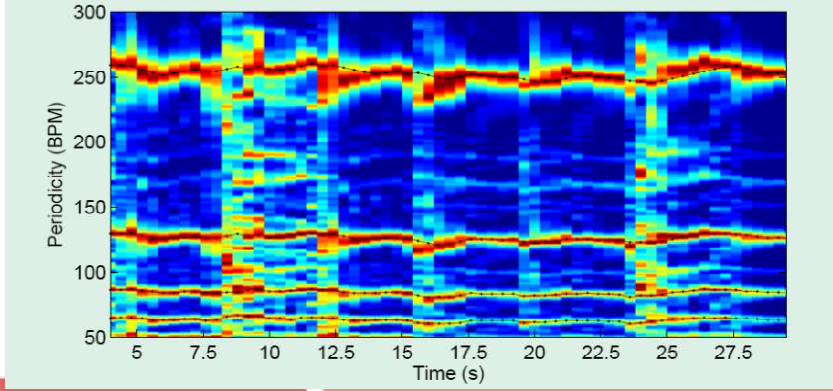
  - *De nombreuses méthodes spectrales ou cepstrales existent aussi*

Extraction du rythme ou du Tempo



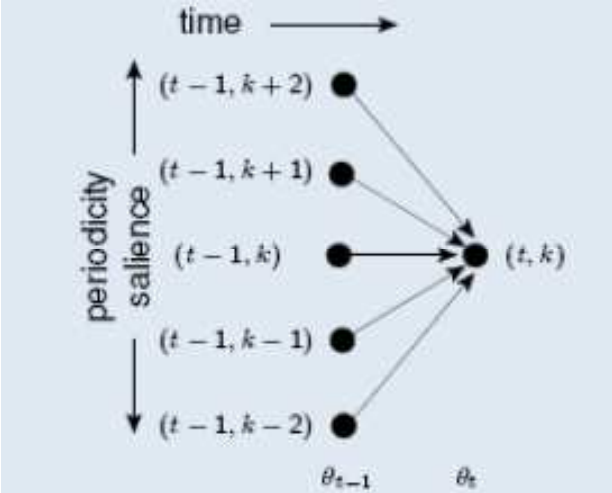
Principe du suivi de périodicités

- A partir d'une analyse des périodicités au cours du morceau
- *… obtenir la variation temporelle du tempo (ou « beats »)*
  - *Par exemple par programmation dynamique*

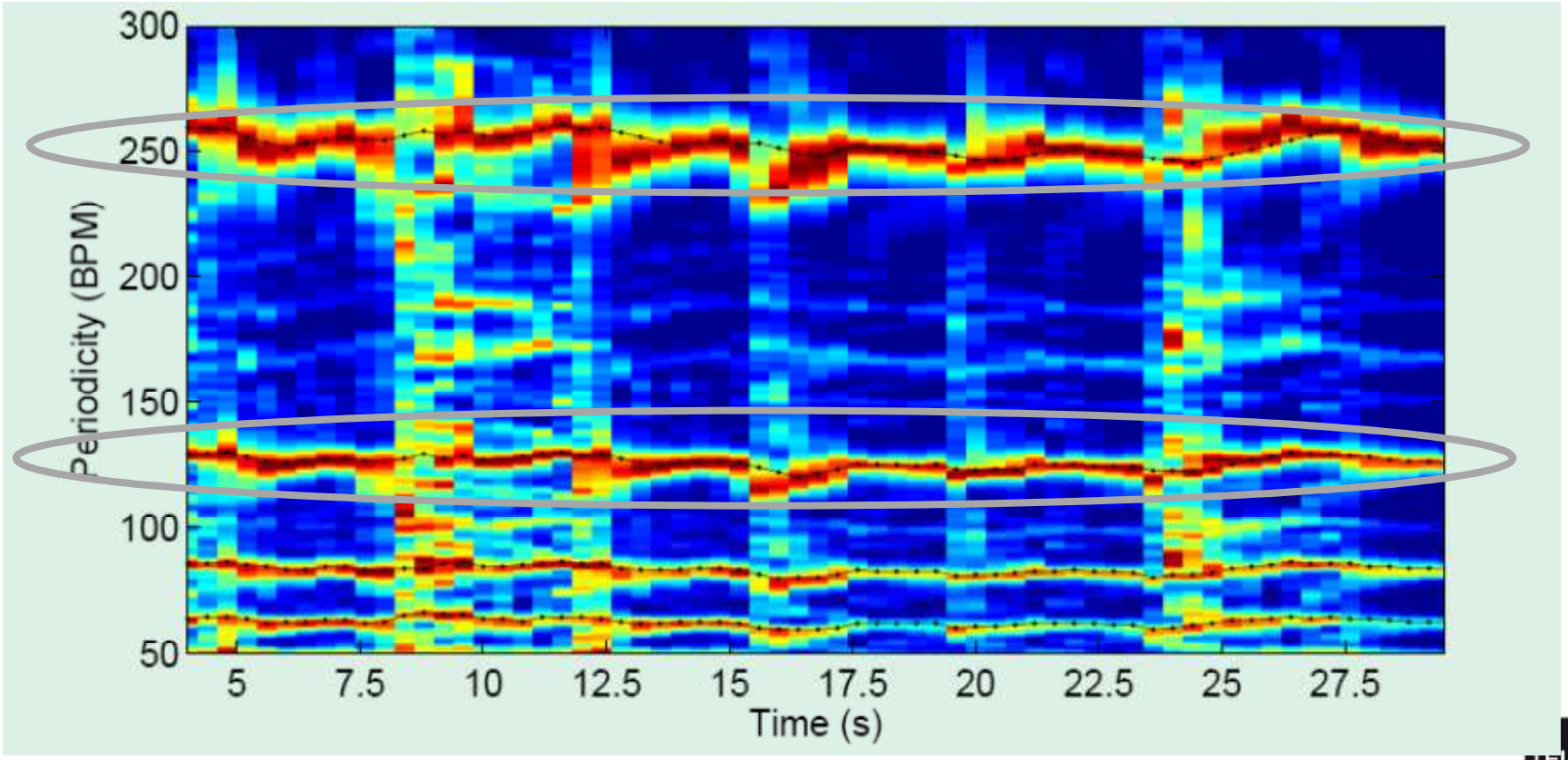


Autres améliorations

- **Suivi dynamique du tempo**
  - Utilisation de la programmation dynamique
  - Adaptation pour le suivi de la structure rythmique
  - Utilisation de contraintes de variation

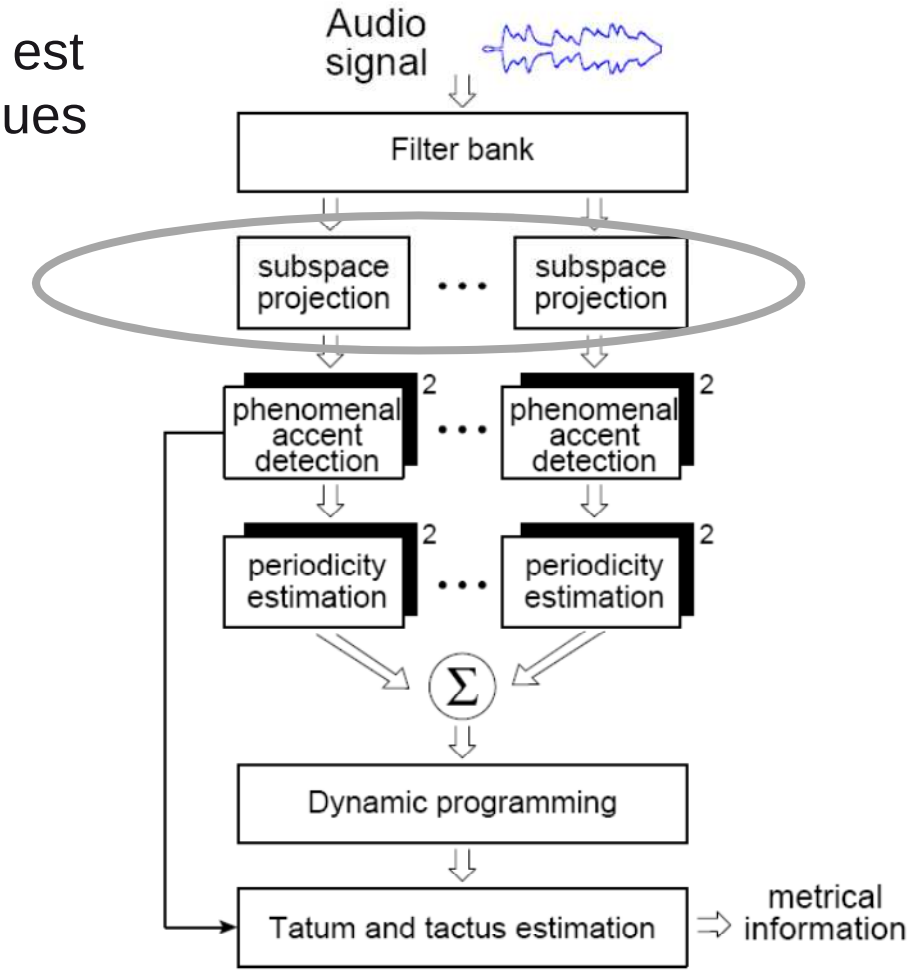


Suivi dynamique du tempo: exemple



Extraction robuste du tempo (From Alonso et al.)

- **Améliorations possibles :**
  - En exploitant le fait que le rythme est principalement porté par les attaques
  - En utilisant une décomposition harmoniques / bruit

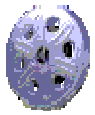


Exemples de résultats

- Evaluation internationale réalisée à MIREX'06 (http://www.music-ir.org/mirex/)

MIREX 2006 Audio Tempo Extraction Summary Results

Contestant	At least 1 tempo correct	Both tempi correct
klapuri	94.29%	61.43%
davies	92.86%	45.71%
alonso 2	89.29%	43.57%

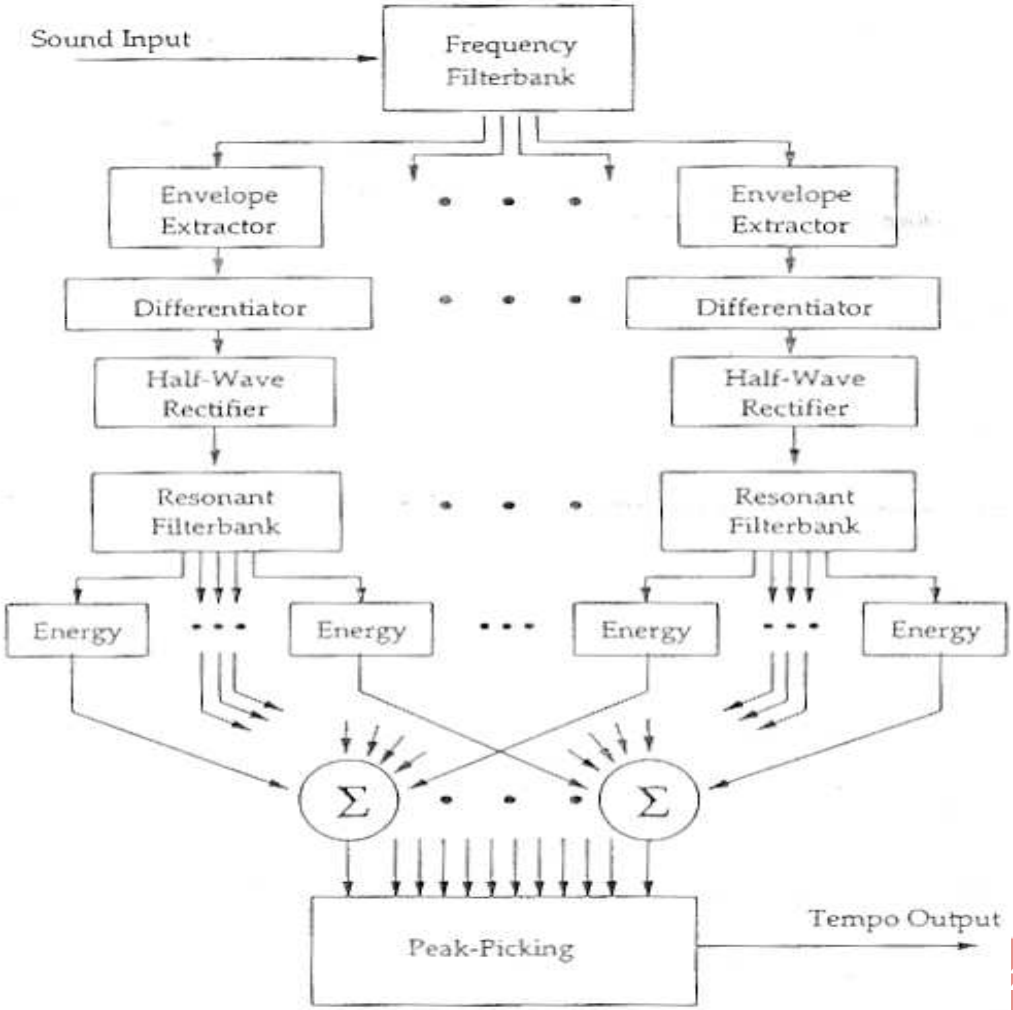


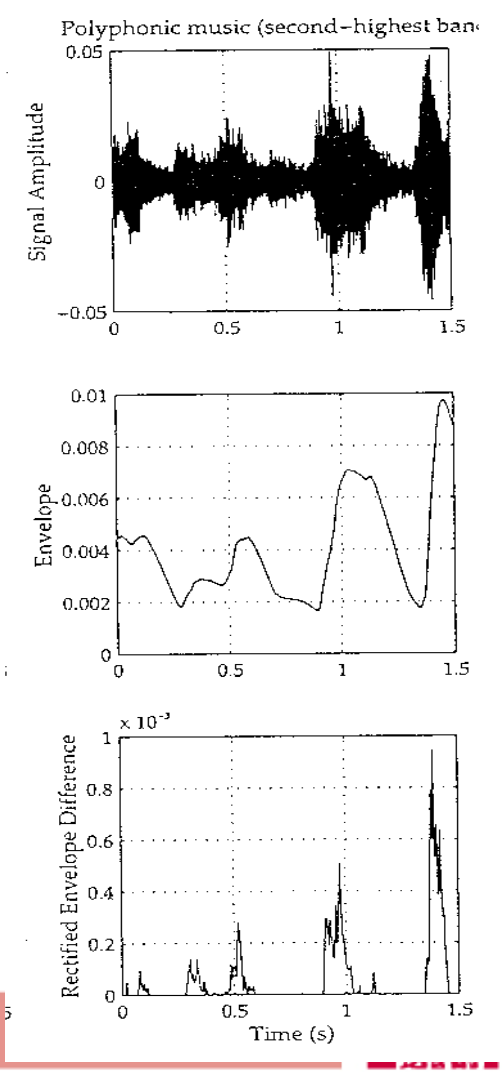
video_tracking.mp4

Demo

Extraction du rythme: Exemple de travaux précurseurs

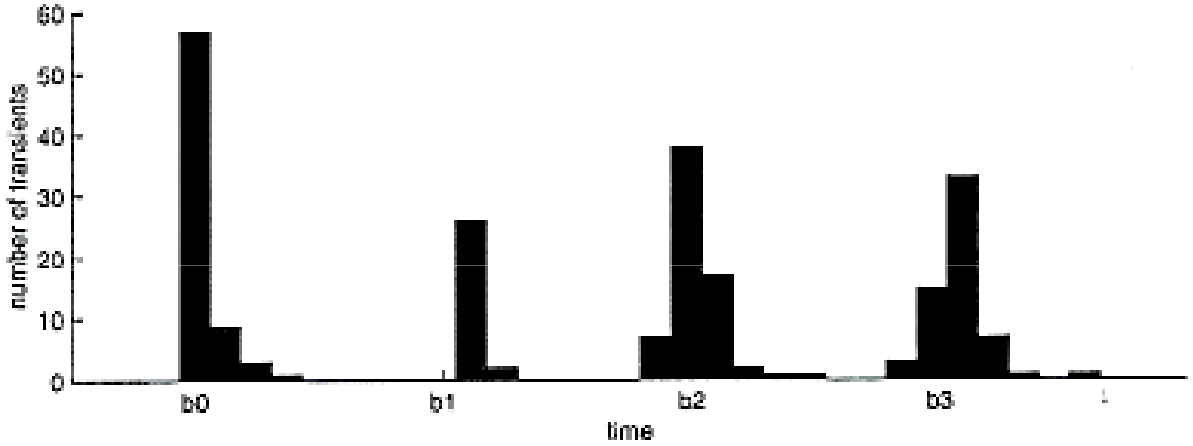
- **Approche par banc de filtres (Scheirer, 1998)**





Extraction du rythme

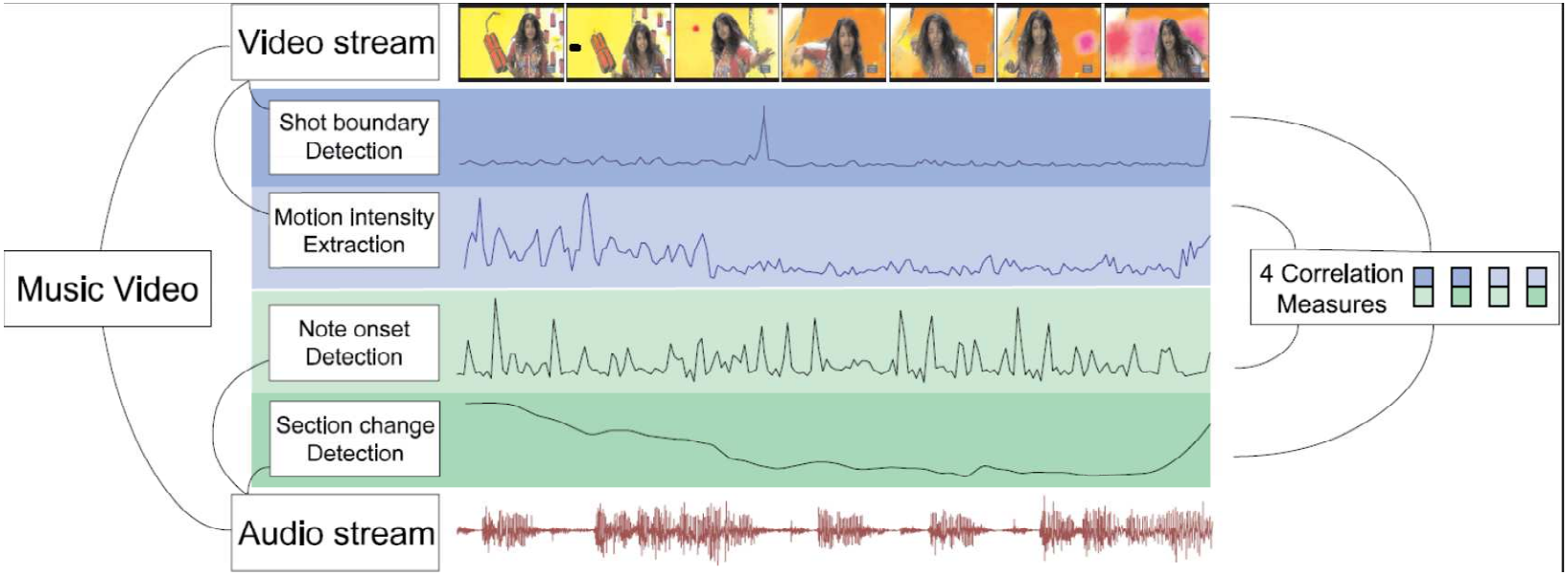
- **Le rythme: un indice intéressant pour la classification des styles musicaux**



Histogramme de la position des attaques sur un signal de musique techno (Laroche2001)

Extraction du rythme: un indice intéressant

- Recherche de musique par similarité (ou classification par genre) genre classification)
- Recherche de contenus audio adaptés à la vidéo



Demo

- Reference

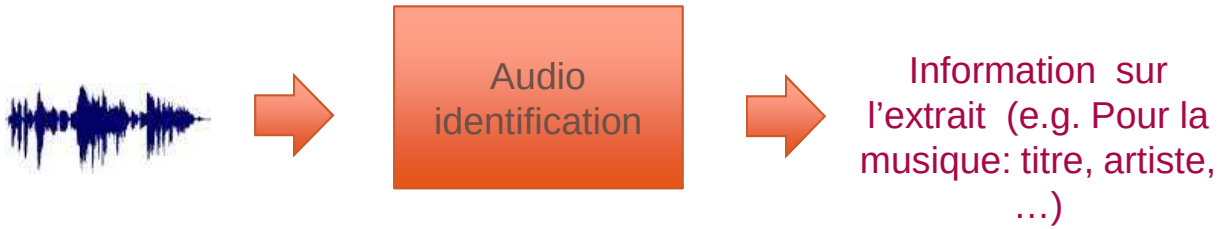
O. Gillet and G. Richard, « On the Correlation of Automatic Audio and Visual Segmentations of Music Videos » I EEE Trans. on CSVT, 2007

Autres Applications
Identification Audio

(Merci à S. Fenêt pour les transparents)

Audio Identification ou AudioID

- **Audio ID = retrouver des métadonnées haut niveau à partir d'un son/morceau**



- **Challenges:**
  - Efficacité en conditions adverses (distorsion, bruits,..)
  - Passage à l'échelle (bases > 100.000 titres)
  - Rapidité / Temps réel

- **Example de produit: Shazam**

Audio fingerprinting

- **Audio Fingerprinting: une approche pour l'Audio-ID**
- **Le principe :**
  - Pour chaque référence, une empreinte audio unique.
  - Identification d'un son: calculer son empreinte et comparaison avec une base d'empreintes de références.

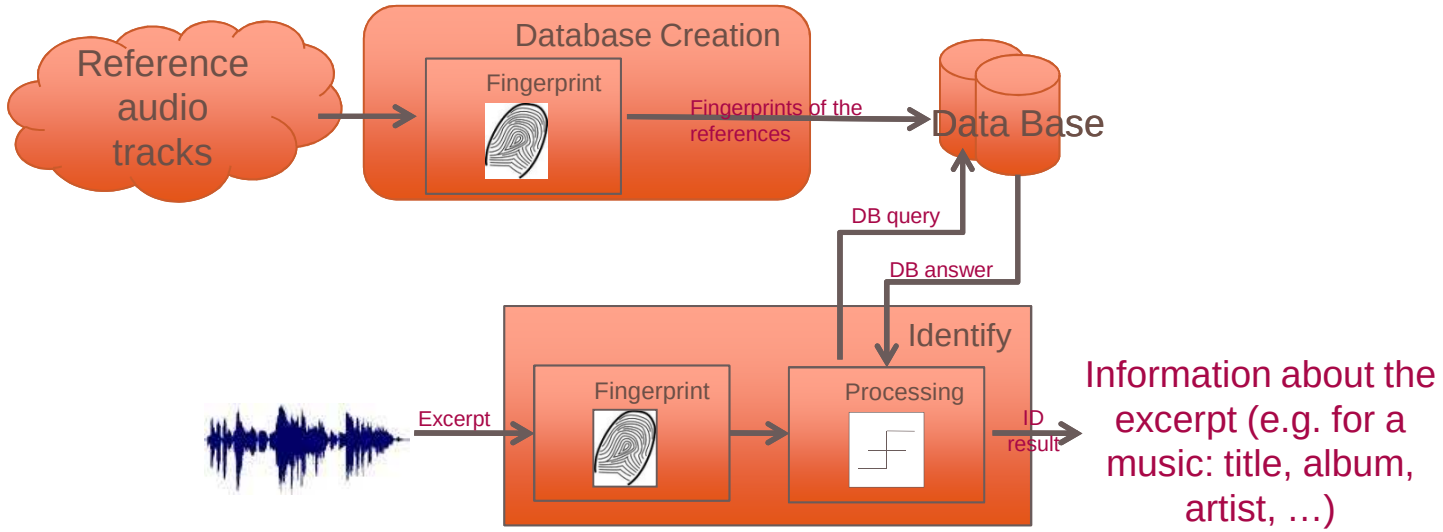
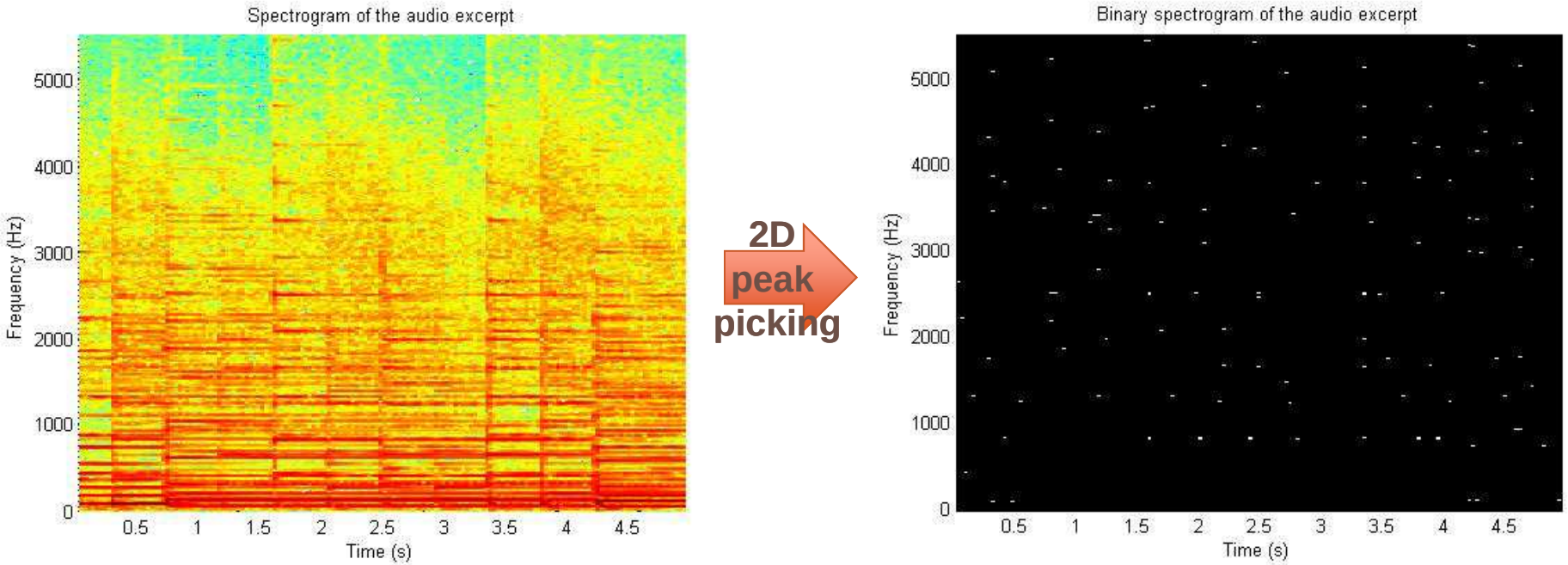


Schéma d'après Sébastien Fenêt

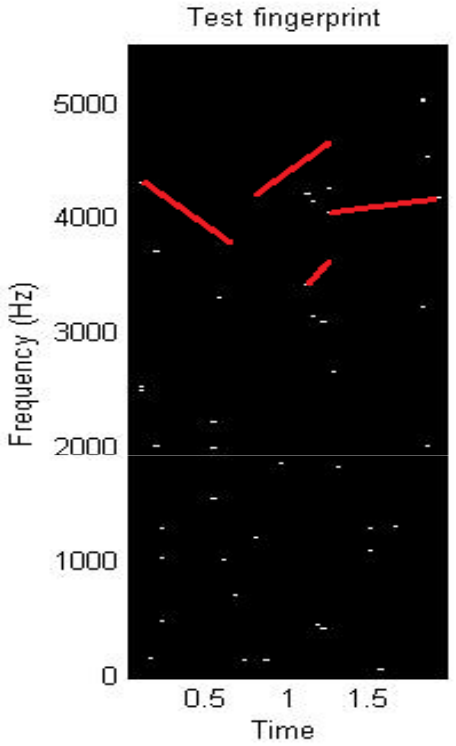
Modèle de signal utilisé

- ‘Binarisation’ du spectrogramme (2D-peak-picking):



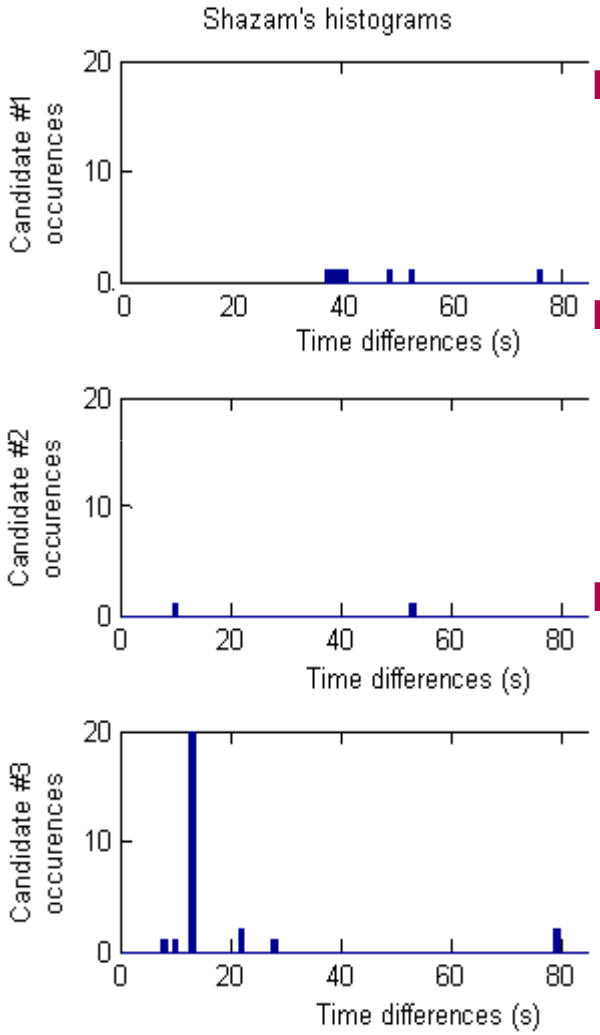
Stratégie de recherche efficace

- **Extrait inconnu à identifier dans une base de + de 100.000 titres**



- **Stratégies possibles**
  - Comparaison directe avec chaque référence de la base (avec tous les décalages temporels possibles)
  - Utiliser la localisation des points blancs comme index
  - Utiliser les paires de points comme index

Trouver la meilleure référence



- **Pour chaque paire, une requête à la base: “quelle référence possède cette paire, et à quel instant cette paire apparaît”**
- **Si la paire apparaît à T1 dans l’extrait inconnu et à T2 dans la référence, on définit le décalage temporel :**

$$\Delta T(pair)=T2-T1$$

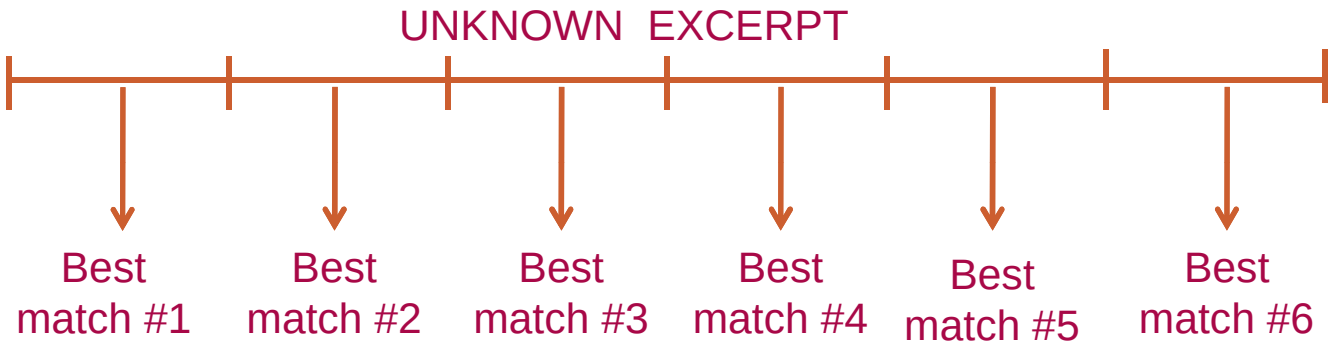
- **Algorithme pour trouver la meilleure référence:**

```
For each pair:
      Get the references having the pair;
      For each reference found:
            Store the time-shift;

Look for the reference with the most frequent time-
   shift;
```

Rejet d'un extrait hors-base: Fusion de décisions locales

- L'extrait inconnu est divisé en sous-segments
- Pour chaque segment, l'algorithme retourne un meilleur candidat



- Si une référence apparaît de manière prépondérante (ou un nombre de fois supérieure à un seuil), l'extrait est identifié
- Sinon, la requête est jugée hors-base
- Taux de bonne détection proche de 90% (pour base de 7500 references) ; 97% avec contrôle du pitch-shifting

Quelques dimensions du signal musical…

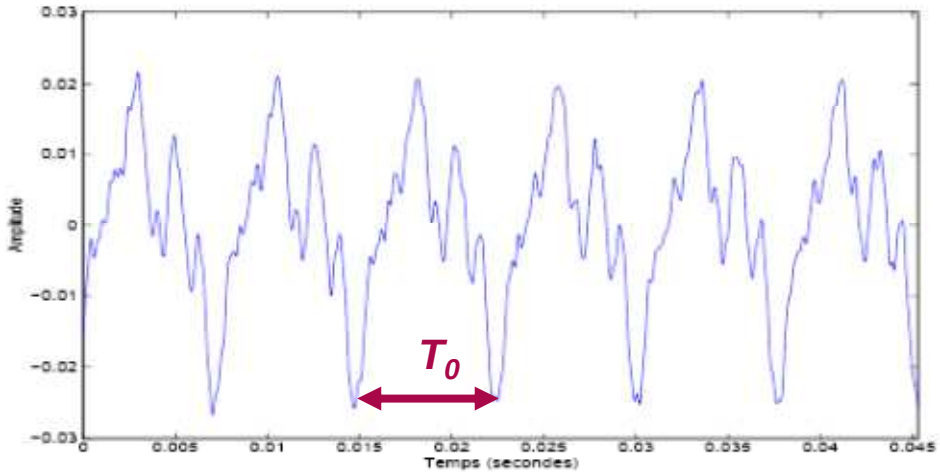
Hauteurs, Harmonie,..

Tempo, rythme,…

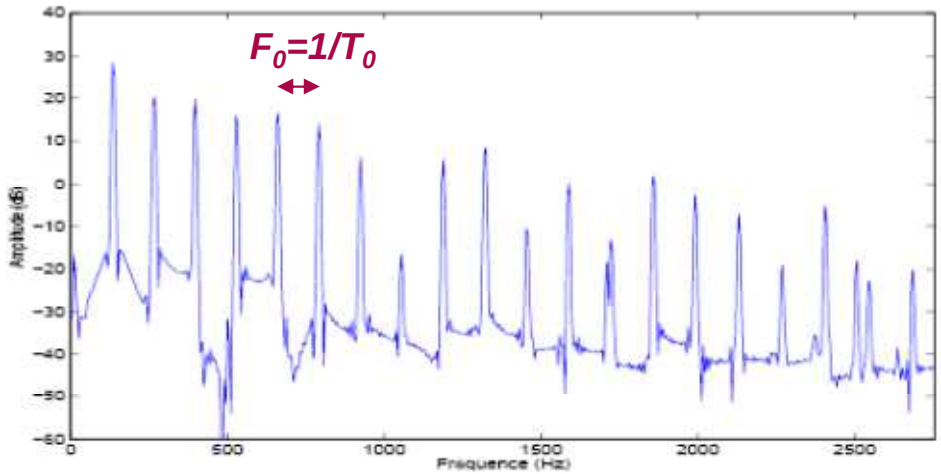
Timbre, instruments,…

Polyphonie, mélodie, ….

Un son quasi-périodique



Son de piano (C3)

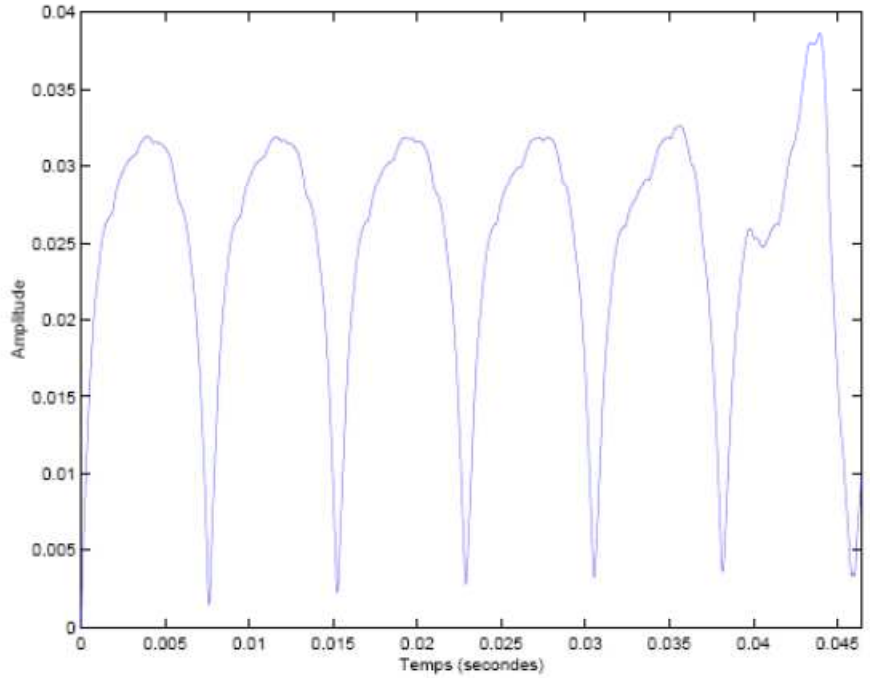


Spectre du son de piano

Average magnitude difference function (AMDF)

$$\mathrm{AMDF}[m] = \frac{1}{N-m} \sum_{n=0}^{N-1-m} |x[n] - x[n+m]|$$

$$\mathrm{AMDF}[m] = 0 \text{ ssi } x \text{ est de période } T_0 = m$$



Détection de fréquences fondamentales multiples

- **Objectif: extraire l'ensemble des notes d'un enregistrement polyphonique**
- **Problème important lorsque les notes sont en rapport harmonique (ce qui est souvent le cas en musique…!!)**
- **Nécessité de traiter le caractère non parfaitement harmonique des notes jouées par un instrument.**

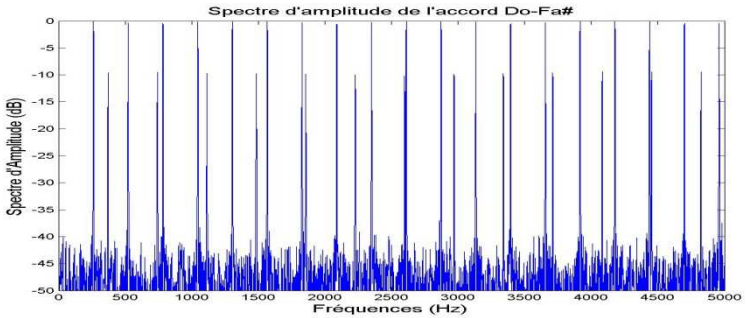
Une possibilité: Reconnaître itérativement chaque note …

- **Un exemple avec la transcription polyphonique…**

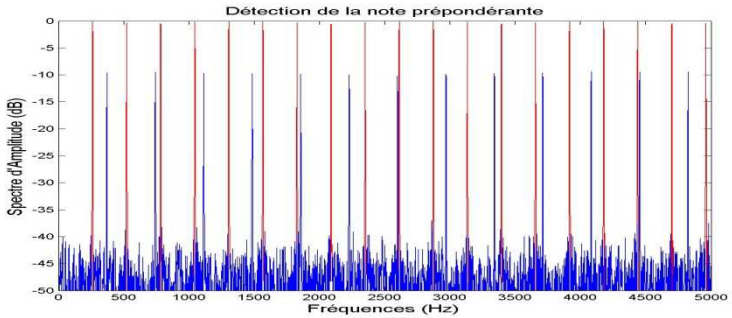
  - Reconnaître la note la plus forte …
  - La soustraire de l'accord
  - Reconnaître la note suivante
  - La soustraire de l'accord
  - Etc… tant qu'il y a des notes dans l'accord

transcription polyphonique…

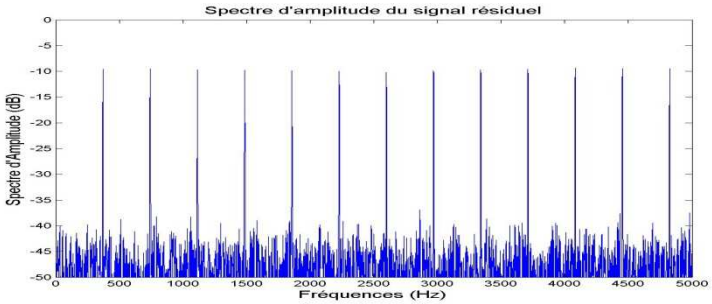
Accord de deux notes de synthèse Do – Fa#



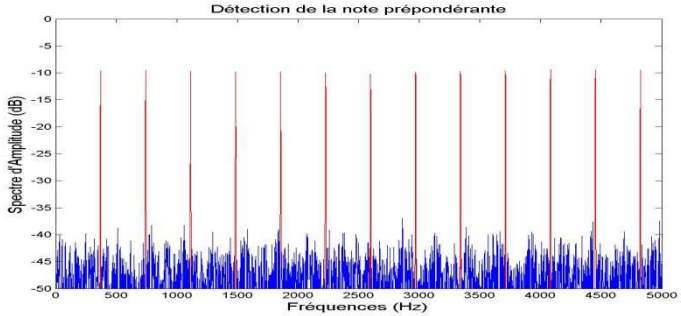
Détecter la note prépondérante (en rouge)



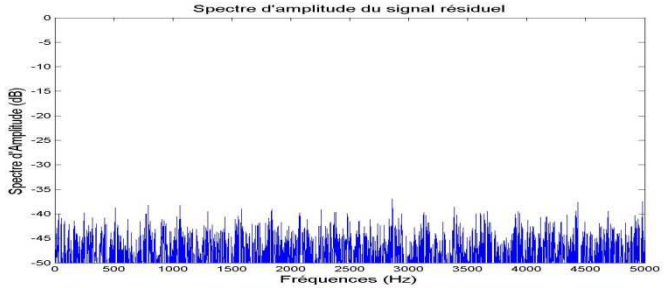
Soustraire la note détectée



Détection la note suivante….

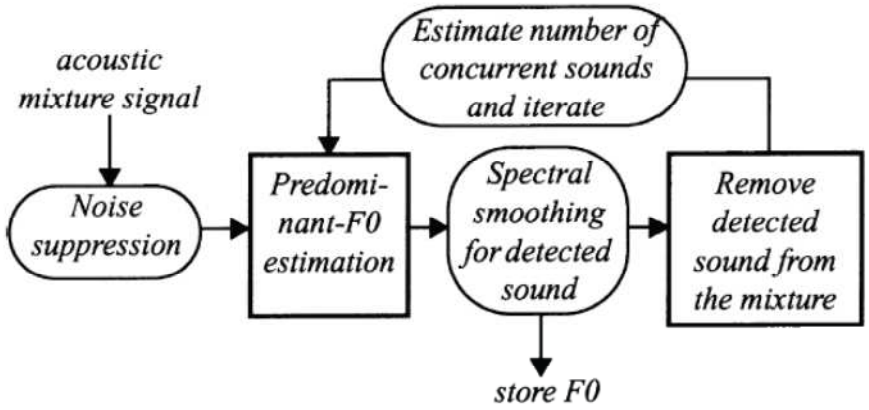


Il n'y a plus de plus de notes…l'accord do Fa# a été reconnu



Détection de fréquences fondamentales multiples

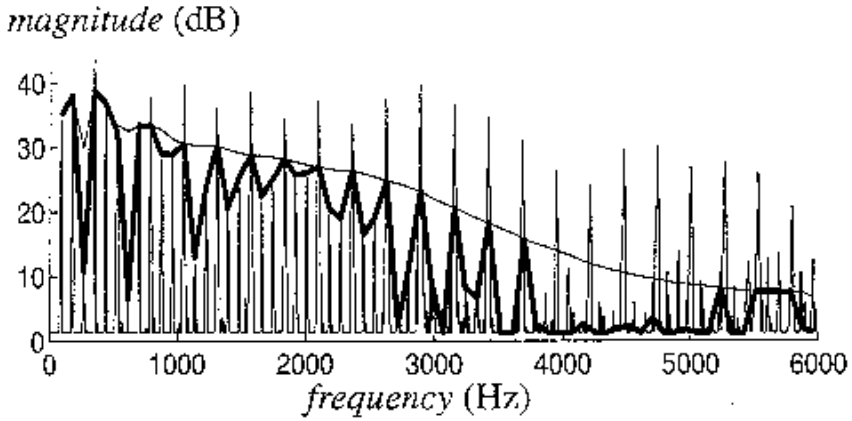
- **Approche par soustraction itérative (Klapuri)**



Principe de lissage spectral

$a_h = min(a_h, m_h)$

où m_h est la moyenne sur une fenêtre d 'un octave autour du partiel



Exemple d’application aux sons percussifs (batterie)

TELECOM PARIS
école nationale supérieure des télécommunications

Comment retrouver des musiques par leur rythme ?

Poum ta Poum Poum ta

Exprimez dans le microphone la séquence rythmique que vous souhaitez...

RECONNAISSANCE VOCALE

ENREGISTREMENT D'UNE NOUVELLE BOUCLE

Reconnaissance automatique (cymbales, grosse caisse...)

Extraction du tempo

BOUCLE DE BATTERIE TROUVÉE

... Le système trouve, dans sa base de données de boucles de batterie, toutes celles qui correspondent à votre choix.

CONTACT : *Olivier Gillet, Gaël Richard*

POUR EN SAVOIR PLUS : *www.telecom-paris.fr*

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Transcription de boucles de batterie
(d'après Gillet et al.)

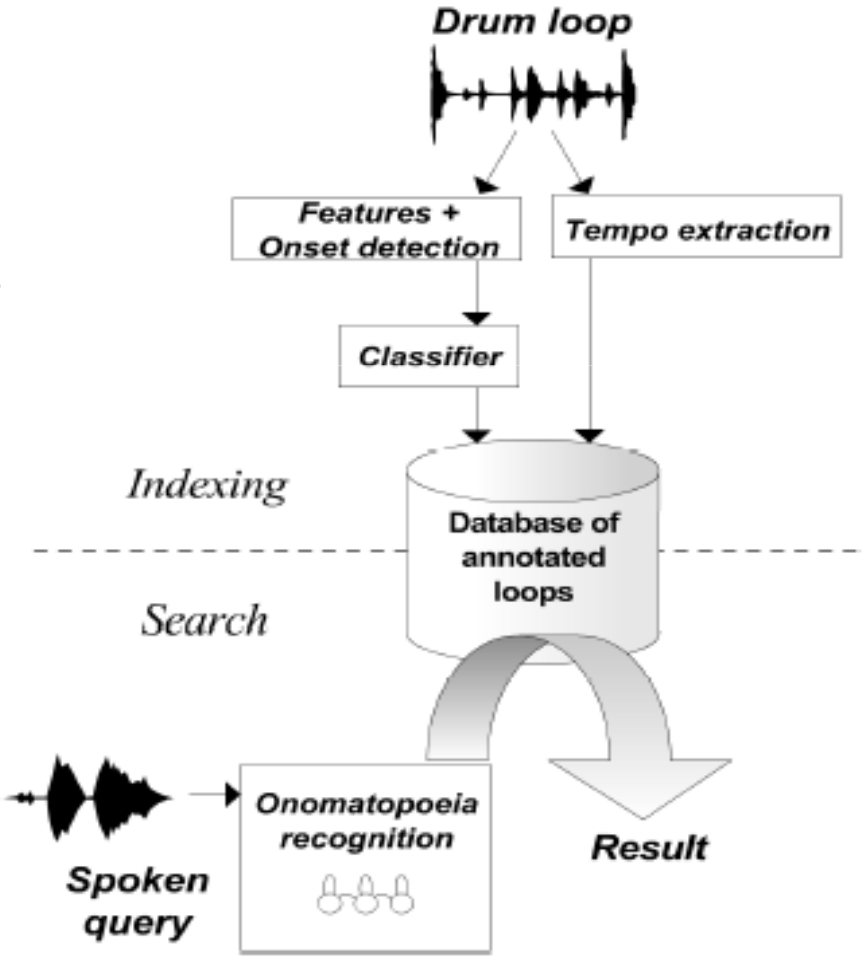
- Un système de transcription de boucles de batterie

Light_1.mp3

- Un moteur simple de reconnaissance vocale

Heavy_1.mp3

- Un moteur de recherche par similarité sur les transcriptions



Quelques éléments techniques

- **Paramétrisation:**
  - 23 paramètres (MFCC, moments spectraux,..)
- **Classification :**
  - Séparateurs à vaste marge
- **Utilisation d'un modèle de séquences**

Problème de l ’évaluation en général

- **Domaine ne possédant pas encore la maturité de la reconnaissance vocale**
- **Pas de bases de données et de protocoles communs pour l’évaluation des techniques mais si un effort important est réalisé avec MIREX**
- **Nécessité d’une annotation préalable du signal pour vérifier le taux de reconnaissance/identification des algorithmes:**
  - Utilisation de standards de description (MIDI, MPEG4-SA, MPEG7 etc…)

Conclusion

- **L ’indexation audio apparaît comme un domaine de plus en plus important:**
  - Nombreuses applications
  - poussée par l’Internet et la croissance des données audio sur la toile

- **Le « rêve » d ’extraire la partition de musique (incluant les paroles) directement du signal audio reste encore un rêve…..**

Quelques References

- **Estimation du rythme/Tempo**
  - M. Alonso, G. Richard, B. David, "*Accurate tempo estimation based on harmonic+noise decomposition", EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 82795, 14 pages, 2007.
  - Scheirer E., 1998, "*Tempo and Beat Analysis of Acoustic Musical Signals*", Journal of the Acoustical Society of America (1998), Vol. 103, No. 1, pp. 588-601. 50
  - Laroche, 2001] J. Laroche. Estimating Tempo, Swing, and Beat Locations in Audio Recordings. Dans Proc. of WASPAA'01, New York, NY, USA, octobre 2001
- **Statistiques, apprentissage**
  - R. Duda, P. Hart and D. Stork, *Pattern Classification*, Wiley-Interscience, 2001
  - B. Schölkopf and A. Smola, Learning with kernels. The MIT Press, Cambridge, MA, 2002
  - L. Rabiner, Fundamentals of Speech Processing, Prentice Hall Signal Processing Series, 1993
  - T. Hastie and R. Tibshirani Classification by pairwise coupling, in Advances in Neural Information Processing Systems, vol 10, The MIT Press, 1998.
- **Reconnaissance des instruments de musique**
  - S. Essid, G. Richard, B. David. *Instrument recognition in polyphonic music based on automatic taxonomies*. IEEE Trans. on Audio, Speech, and Language Proc. 14 (2006), no. 1
  - Eronen, « comparison of features for musical instrument recognition », Proc of IEEE-WASPAA'2001.
  - S. Essid, G. Richard, B. David. *Musical Instrument recognition by pairwise classification strategies*. IEEE Trans. on Audio, Speech and Language Proc. 14 (2006), no. 4
  - O. Gillet et G. Richard , « *Extraction and Remixing of Drum tracks from polyphonic music signals », IEEE-WASPAA'05,* New Paltz, NY, 2005
  - O. Gillet et G. Richard , « *Drum loops retrieval from spoken queries* », Journal of Intelligent Information Systems, 24:2/3, pp 159-177, Springer Science, 2005
  - O. Gillet, G. Richard. *Transcription and separation of drum signals from polyphonic music*. accepted in IEEE Trans. on Audio, Speech and Language Proc. (2008)
- **Indexation audio, paramètres**
  - A. Klapuri A. M. Davy, Methods for Music Transcription M. Springer New York 2006
  - G. Peeters, "Automatic classification of large musical instrument databases usign hierarchical classifiers with inertia ratio maximization, in 115th AES convention, New York, USA, Oct. 2003.
  - G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM (2004)

Quelques liens et papiers

- **Estimation du rythme/Tempo**
  - M. Alonso, G. Richard, B. David, "*Accurate tempo estimation based on harmonic+noise decomposition", EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 82795, 14 pages, 2007.
  - Scheirer E. "*Tempo and Beat Analysis of Acoustic Musical Signals*", Journal of the Acoustical Society of America (1998), Vol. 103, No. 1, pp. 588-601. 50
  - Laroche, 2001] J. Laroche. Estimating Tempo, Swing, and Beat Locations in Audio Recordings. Dans Proc. of WASPAA'01, New York, NY, USA, oct. 2001
- **Statistiques, apprentissage**
  - R. Duda, P. Hart and D. Stork, *Pattern Classification*, Wiley-Interscience, 2001
  - B. Schölkopf and A. Smola, Learning with kernels. The MIT Press, Cambridge, MA, 2002
  - L. Rabiner, Fundamentals of Speech Processing, Prentice Hall Signal Processing Series, 1993
  - T. Hastie and R. Tibshirani Classification by pairwise coupling, in Advances in Neural Information Processing Systems, vol 10, The MIT Press, 1998.
- **Reconnaissance des instruments de musique**
  - S. Essid, G. Richard, B. David. *Instrument recognition in polyphonic music based on automatic taxonomies*. IEEE Trans. on Audio, Speech, and Language Proc. 14 (2006), no. 1
  - Eronen, « comparison of features for musical instrument recognition », Proc of IEEE-WASPAA'2001.
  - S. Essid, G. Richard, B. David. *Musical Instrument recognition by pairwise classification strategies*. IEEE Trans. on Audio, Speech and Language Proc. 14 (2006), no. 4
  - O. Gillet et G. Richard , « *Extraction and Remixing of Drum tracks from polyphonic music signals », IEEE-WASPAA'05,* New Paltz, NY, 2005
  - O. Gillet et G. Richard , « *Drum loops retrieval from spoken queries* », Journal of Intelligent Information Systems, 24:2/3, pp 159-177, Springer Science, 2005
  - O. Gillet, G. Richard. *Transcription and separation of drum signals from polyphonic music*. accepted in IEEE Trans. on Audio, Speech and Lang. Proc. (2008)
- **Indexation audio, paramètres**
  - A. Klapuri A. M. Davy, Methods for Music Transcription M. Springer New York 2006
  - G. Peeters, "Automatic classification of large musical instrument databases usign hierarchical classifiers with inertia ratio maximization, in 115th AES convention, New York, USA, Oct. 2003.
  - G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM (2004)
- **Synthèse sonore**
  - CHOWNING (John M.), « *The synthesis of complex audio spectra by means of frequency modulations* », Journal of the Audio Engineering Society (J.A.E.S.), vol. 21, n° 7, septembre 1973 .
  - Bensoam: http://www.ircam.fr/equipes/instruments/bensoam/PageHtml/illustration/
  - http://membres.lycos.fr/hhh/SYNTHES/Divers/Histoire.htm, http://www.ircam.fr

- http://www.enst.fr/~grichard/ , http://www.enst.fr/~rbadeau/ , http://www.enst.fr/~gillet/, http://www.enst.fr/~bedavid/