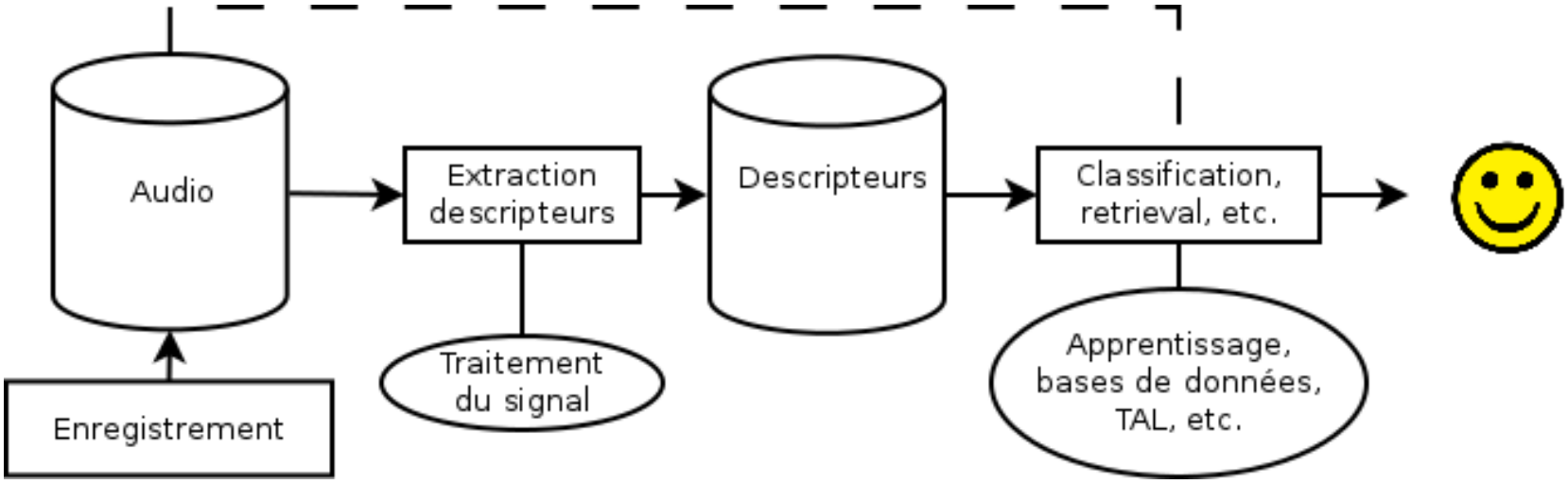
Indexation audio

Recherche d'information et indexation multimedia

M2 RIM, Aix-Marseille Université
Valentin Emiya (prenom.nom@lif.univ-mrs.fr)

27 novembre 2012

Indexation audio : de quoi parle-t-on ?



- Objectif du cours : sensibilisation à l'indexation audio.
- Les données : audio.
- Motivation : les applications.

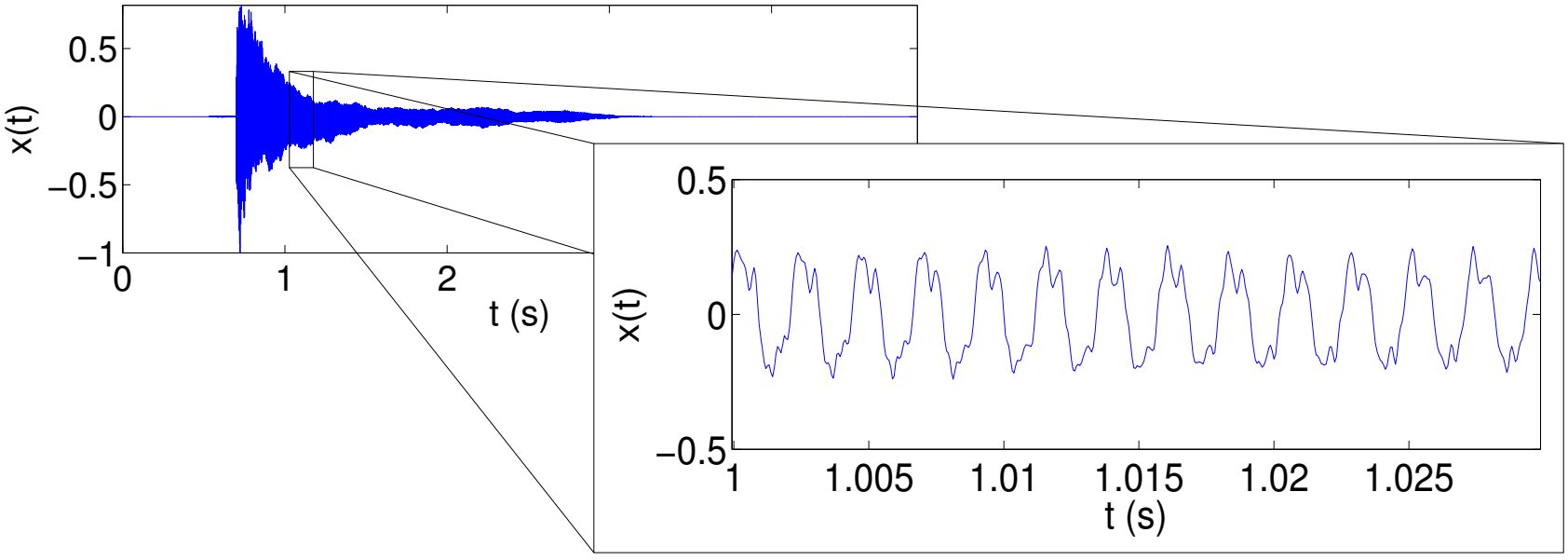
$\Rightarrow$ Comment manipuler l'information dans les données audio ?

NB : cette séance est orientée « musique » mais les principes se généralisent aux autres types de sons.

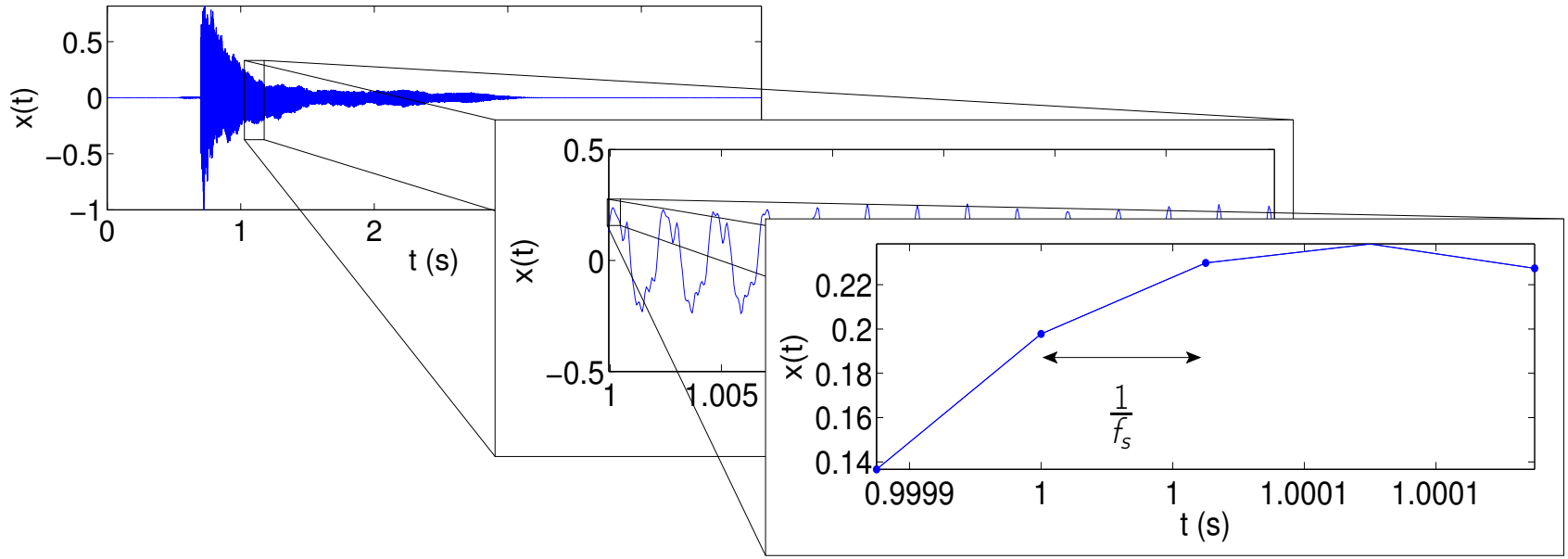
Les données audio

- Nature : tout enregistrement sonore (numérique).
- Provenance :
  - production professionnelle (studio, concert, audiovisuel),
  - particuliers,
  - capteurs divers : téléphonie, surveillance, etc.
- Contenu :
  - voix parlée,
  - musique,
  - sons de l'environnement,
  - bruits,
  - etc.
- Autres media parfois associés : métadonnées, multimédia.

Le signal audio



Le signal audio



Signal audio lu à partir d'un fichier .wav :

- vecteur $\mathbf{x} = (\mathbf{x}(0), \ldots, \mathbf{x}(N_\mathbf{x} - 1))^T \in \mathbb{R}^{N_\mathbf{x}}$ de longueur $N_\mathbf{x}$.
- f_s : fréquence d'échantillonnage (nombre de coefficients par seconde)

Pourquoi indexer ? Les applications.

- Détection d'événements
- Recommandation, génération de playlist
- Recherche par le contenu
- Recherche par similarité
- Identification par fredonnement
- Accompagnement automatique, interaction musicien/machine, synchronisation
- Extraction de partition, tablature, accords,

La recherche : exemples de tâches applicatives

Le challenge Music Information Retrieval Evaluation eXchange

Classification

- Audio Artist Identification
- Audio Genre Classification
- Audio Music Mood Classification
- Audio Classical Composer Identification
- Symbolic Genre Classification
- Audio Cover Song Identification
- Query by Singing/Humming
- Query by Tapping
- Audio Music Similarity and Retrieval

Alignement

- Real-time Audio to Score Alignment
- Symbolic Melodic Similarity

Extraction de descripteurs

- Multiple Fundamental Frequency Estimation and Tracking
- Audio Chord Estimation
- Audio Melody Extraction
- Audio Beat Tracking
- Audio Tempo Extraction
- Audio Tag Classification

Segmentation

- Audio Onset Detection
- Audio Key Detection
- Symbolic Key Detection
- Structural Segmentation
- Audio Drum Detection

Autres champs utilisant de l'indexation audio

- Traitement de la parole, des langues
- Signal : séparation de sources, restauration
- Multimedia : fusion multimodale
- Audition : perception, cognition
- Robotique
- Sécurité/surveillance
- Cris des animaux
- etc.

Notions d'acoustique et d'audition : pourquoi ?

Les données audio sont enregistrées à mi-chemin entre :

- en amont, la génération du son (acoustique)
- en aval, la perception du son (audition)

Buts de l'analyse computationnelle : perception par l'ordinateur, analyse de la production sonore.

Notions d'acoustique

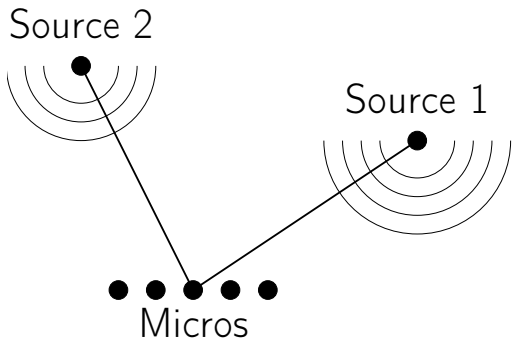
Production

sources harmonique
sources de bruit
source-filtre : la voix, les instruments

Propagation

Son direct : délai, atténuation
Reflexions, réverbération : convolution avec réponse impulsionnelle

Notions d'acoustique



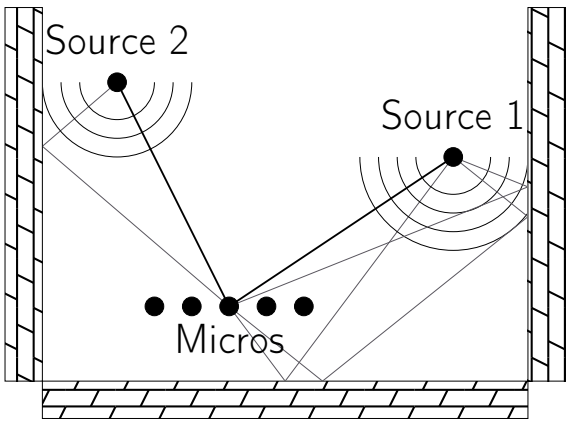
Production

sources harmonique
sources de bruit
source-filtre : la voix, les instruments

Propagation

Son direct : délai, atténuation
Reflexions, réverbération : convolution avec réponse impulsionnelle

Notions d'acoustique



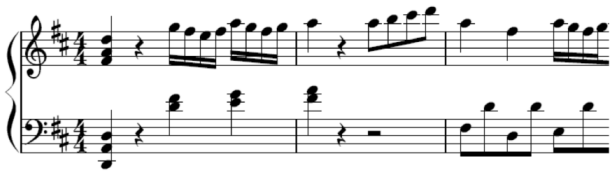
Production

sources harmonique
sources de bruit
source-filtre : la voix, les instruments

Propagation

Son direct : délai, atténuation
Reflexions, réverbération : convolution avec réponse impulsionnelle

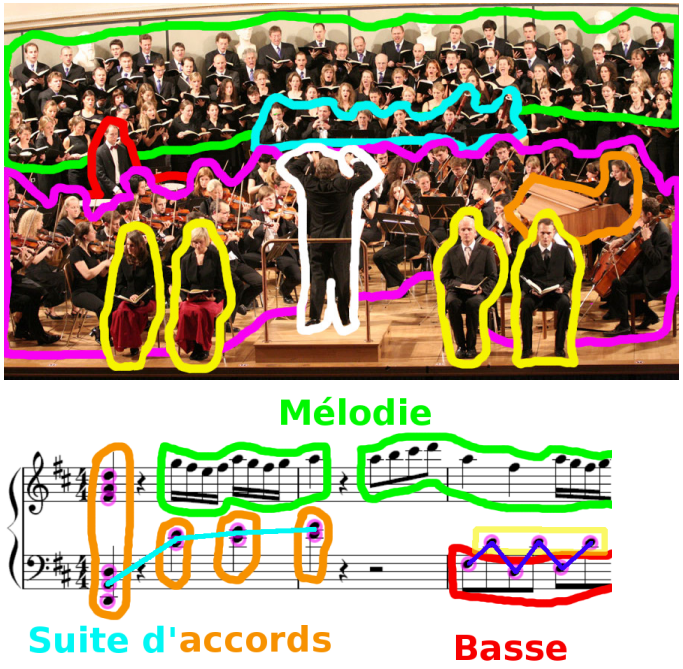
Notions d'audition



Comment décrire un son perçu ?

- intensité : faible/fort (échelle non-linéaire)
- localisation : distance, angle, étendue
- hauteur : grave/aigu
- temporalité : durée, rythme, tempo
- identification : timbre, locuteur, etc.
- notion d'objet, de forme (fusion, perception active)
- rond, rugueux, clair, etc.

Notions d'audition



Comment décrire un son perçu ?

- intensité : faible/fort (échelle non-linéaire)
- localisation : distance, angle, étendue
- hauteur : grave/aigu
- temporalité : durée, rythme, tempo
- identification : timbre, locuteur, etc.
- notion d'objet, de forme (fusion, perception active)
- rond, rugueux, clair, etc.

Notions d'audition

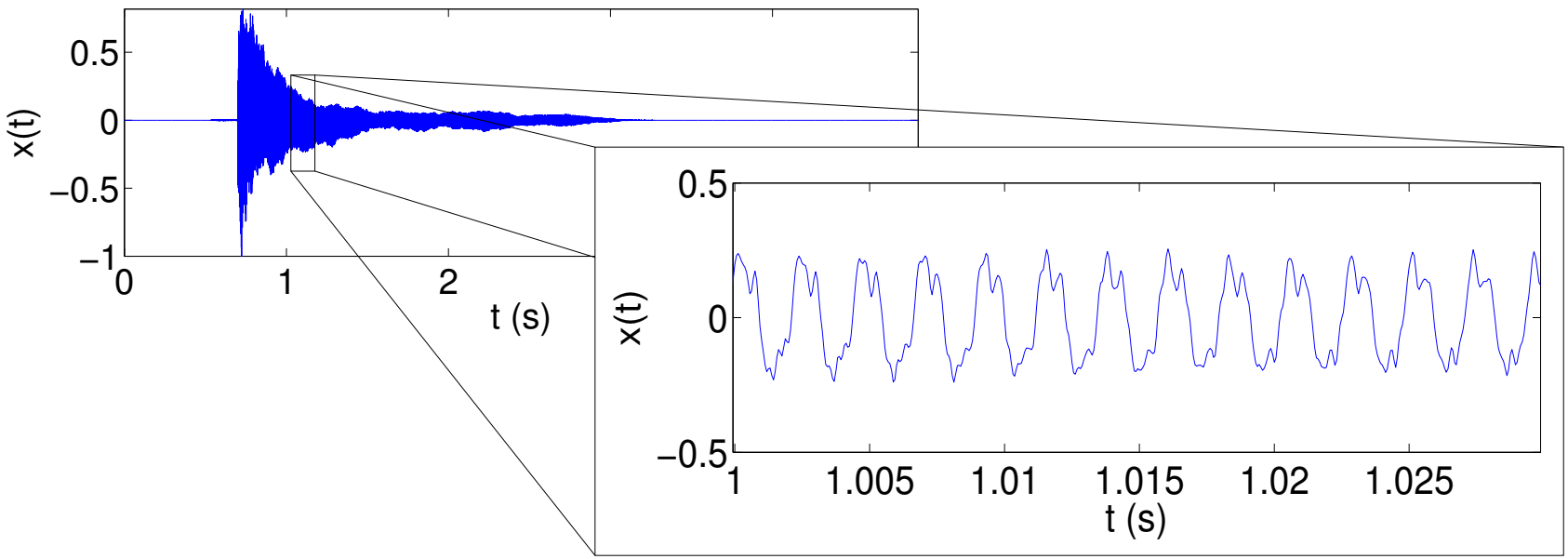
⇒ La description des sons est complexe

- fortement multidimensionnelle,
- subjective,
- souvent difficile à verbaliser.

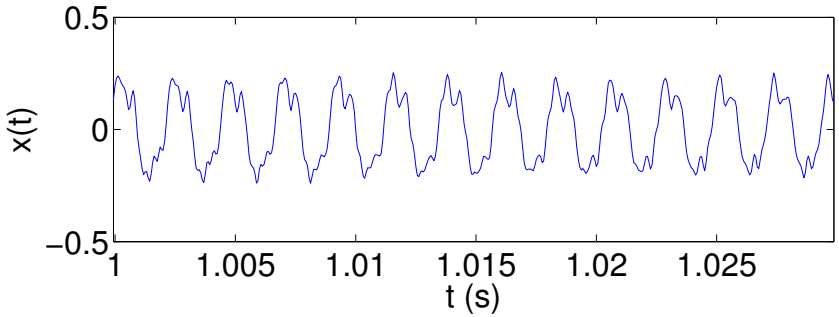
Comment décrire un son perçu ?

- intensité : faible/fort (échelle non-linéaire)
- localisation : distance, angle, étendue
- hauteur : grave/aigu
- temporalité : durée, rythme, tempo
- identification : timbre, locuteur, etc.
- notion d'objet, de forme (fusion, perception active)
- rond, rugueux, clair, etc.

Complément sur la hauteur d'un son périodique



Complément sur la hauteur d'un son périodique

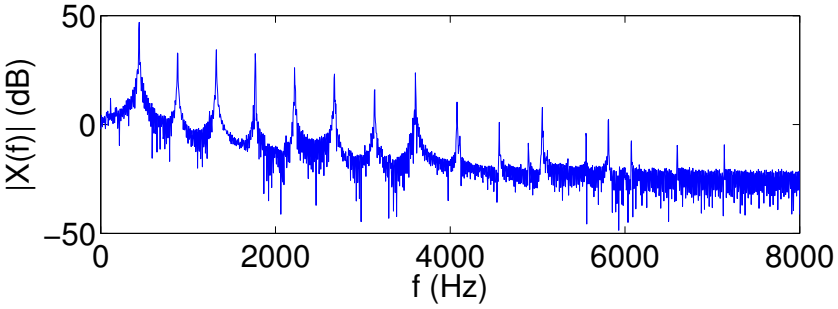


$\Downarrow$

Tranformée de Fourier discrète (TFD)

$$X(f) = \sum_n \mathbf{x}(n) e^{-2i\pi f \frac{n}{f_s}}$$
$$= \langle \mathbf{x}, \text{sinus}_{\mathbb{C}}(f) \rangle$$

$\Downarrow$



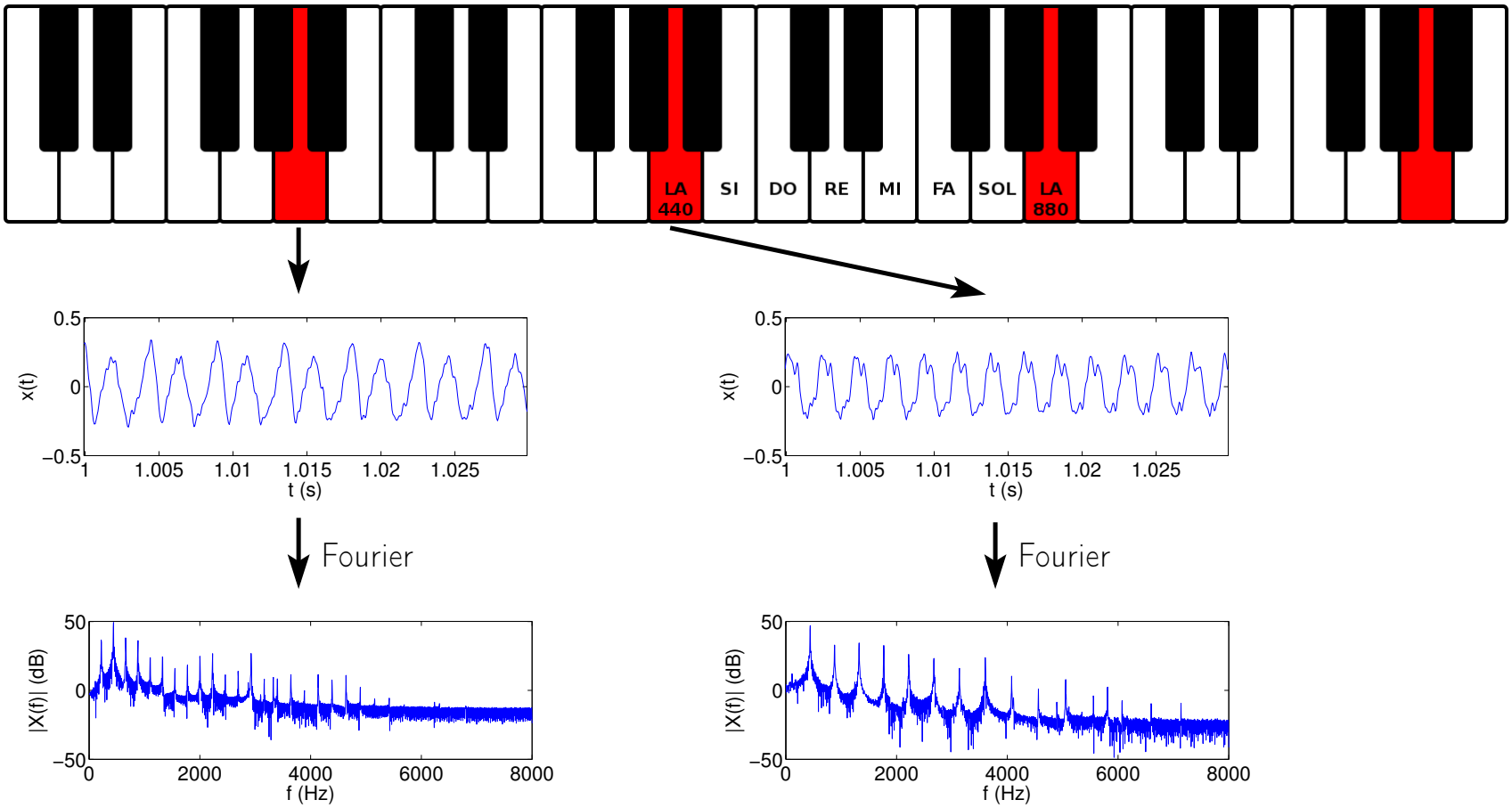
Hauteur caractérisée par la fréquence fondamentale f_0, inverse de la période.

Dans le domaine fréquentiel

- TFD : énergie en fonction de la fréquence.
- *Peigne harmonique :* énergie aux fréquences
$$f_h = hf_0$$

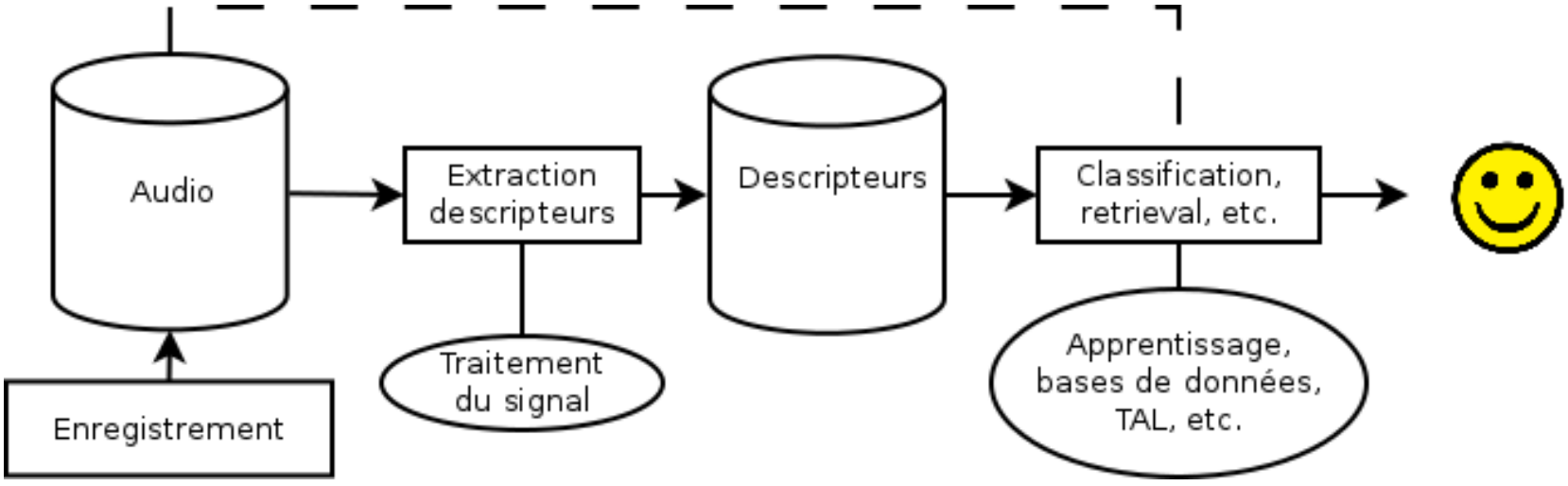
.

Hauteur d'un son périodique : exemples



[jouer sons pianoteq]

La foule des descripteurs



Caractère multidimensionnel de la description des sons

+

Nombreuses applications

=

Grand nombre de descripteurs

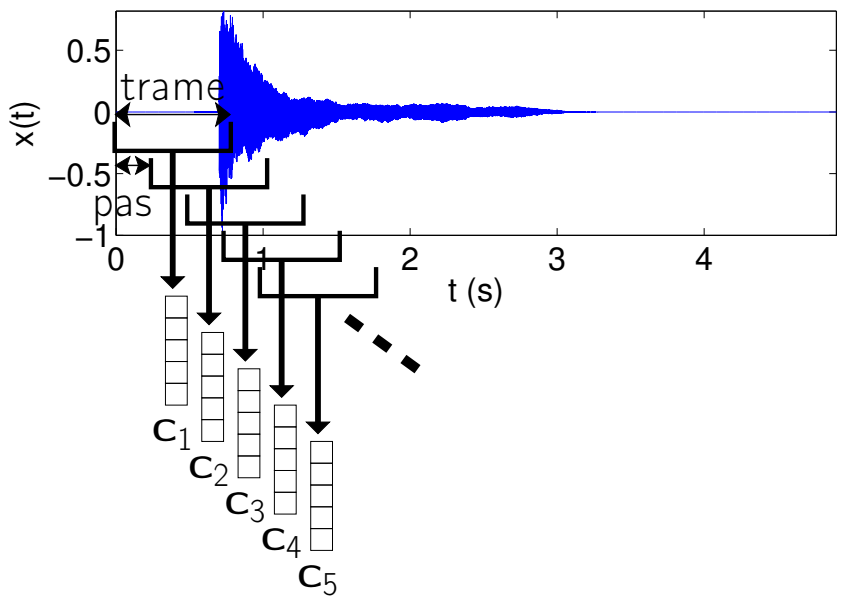
Il existe des toolboxes d'extraction de descripteurs audio, des méthodes de sélection de descripteurs, etc.
Comment s'y retrouver ?

Catégories de descripteurs

(sources : Peeters et al., 2003 ; Mathieu et al., 2010)

Selon l'objet décrit :

- descripteur global : décrit un son entier.
- descripteur instantané/de trame : décrit une trame de son ; une suite temporelle de descripteurs est associée au son.
- descripteurs de haut niveau/bas niveau



Catégories de descripteurs

(sources : Peeters et al., 2003 ; Mathieu et al., 2010)

Selon l'objet décrit :

- descripteur global : décrit un son entier.
- descripteur instantané/de trame : décrit une trame de son ; une suite temporelle de descripteurs est associée au son.
- descripteurs de haut niveau/bas niveau

Selon la nature de l'information :

- temporelle,
- spectrale,
- énergétique,
- perceptive.

Descripteurs instantanés

Descripteurs temporels

Signal Auto-correlation function
Zero-crossing rate

Descripteurs énergétiques

Total energy
Total harmonic energy
Total noise energy

Descripteurs spectraux

Spectral centroid
Spectral spread
Spectral skewness
Spectral kurtosis
Spectral slope
Spectral decrease
Spectral rolloff
Spectral variation
MFCC
Delta MFCC
Delta Delta MFCC

Descripteurs harmoniques

Fundamental frequency
Noisiness
Inharmonicity
Harmonic Spectral Deviation
Odd to Even Harmonic Ratio
Harmonic Tristimulus
HarmonicSpectral centroid
HarmonicSpectral spread
HarmonicSpectral skewness
HarmonicSpectral kurtosis
HarmonicSpectral slope
HarmonicSpectral decrease
HarmonicSpectral rolloff
HarmonicSpectral variation

chroma
multif0

Descripteurs instantanés

Descripteur perceptifs

Loudness
RelaitveSpecific Loudness
Sharpness
Spread
Perceptual Spectral centroid
Perceptual Spectral spread
Perceptual Spectral skewness
Perceptual Spectral kurtosis
Perceptual Spectral Slope
Perceptual Spectral Decrease
Perceptual Spectral Rolloff
Perceptual Spectral Variation
Odd to Even Band Ratio
Band Spectral Deviation
Band Tristimulus

Descripteurs globaux

Descripteurs temporels

Log Attack Time
Temporal increase
Temporal Decrease
Temporal Centroid
Effective Duration

Descripteurs énergétiques

Total energy Modulation (frequency, amplitude)

Descripteurs harmoniques

Fundamental fr. M. (frequency, amplitude)

Descripteurs subjectifs

score subjectif
j'aime/j'aime pas
« Cool », « 80s »,

Tags

Genre
Sources : nature, identité
Condition enregistrement (intérieur, extérieur, mix)
Date, lieu enregistrement
etc.

Descripteurs globaux

Descripteurs temporels

Log Attack Time
Temporal increase
Temporal Decrease
Temporal Centroid
Effective Duration

Descripteurs énergétiques

Total energy Modulation (frequency, amplitude)

Descripteurs harmoniques

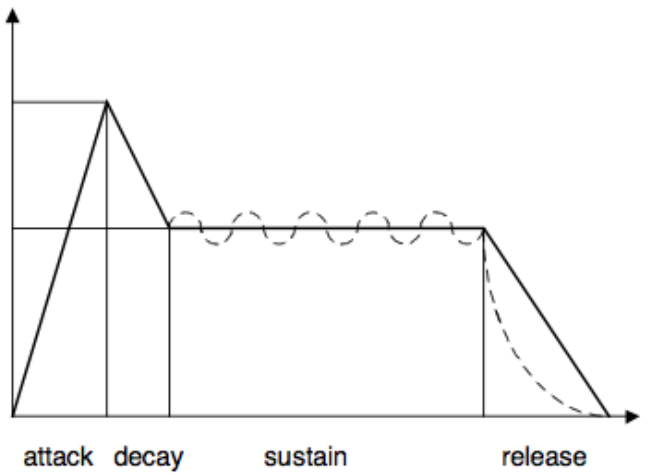
Fundamental fr. M. (frequency, amplitude)

Descripteurs subjectifs

score subjectif
j'aime/j'aime pas
« Cool », « 80s »,

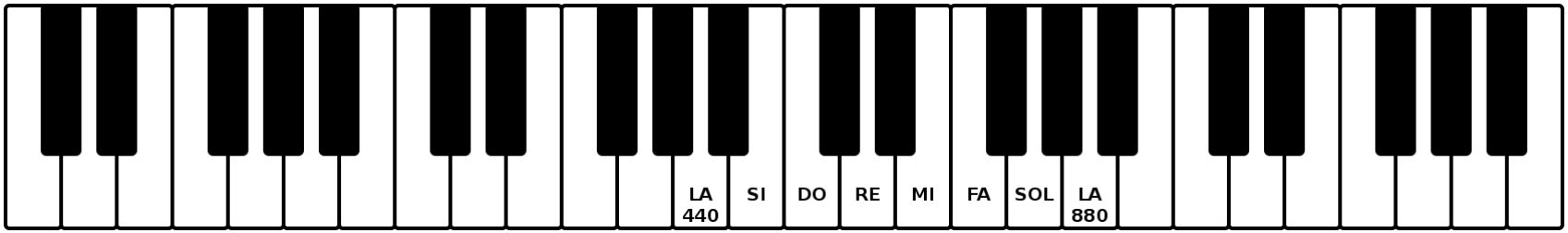
Tags

Genre
Sources : nature, identité
Condition enregistrement (intérieur, extérieur, mix)
Date, lieu enregistrement
etc.



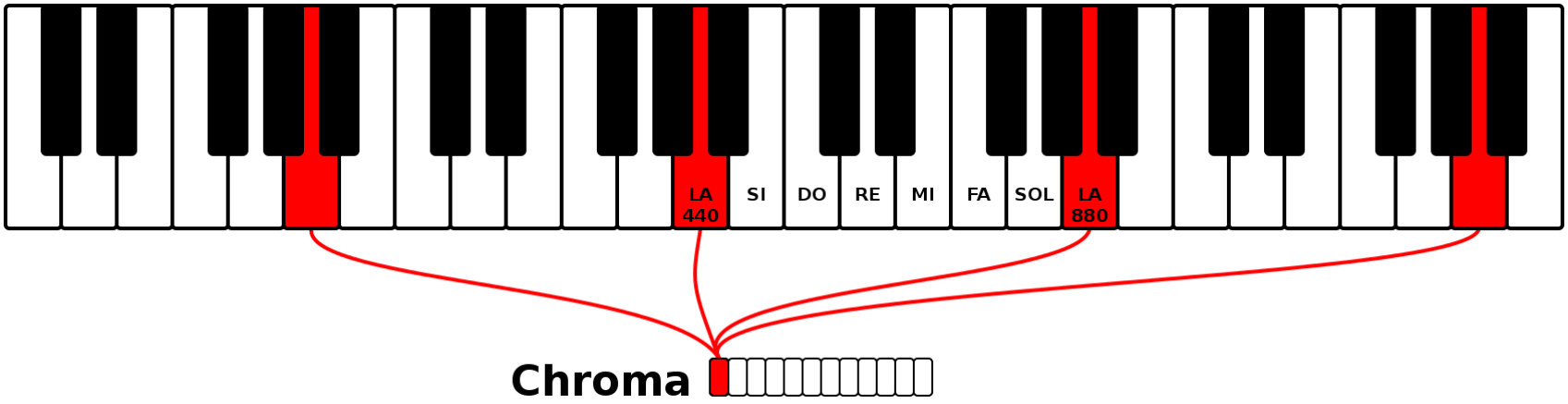
(source : Peeters et al., 2003)

Descripteurs : vecteur de chroma



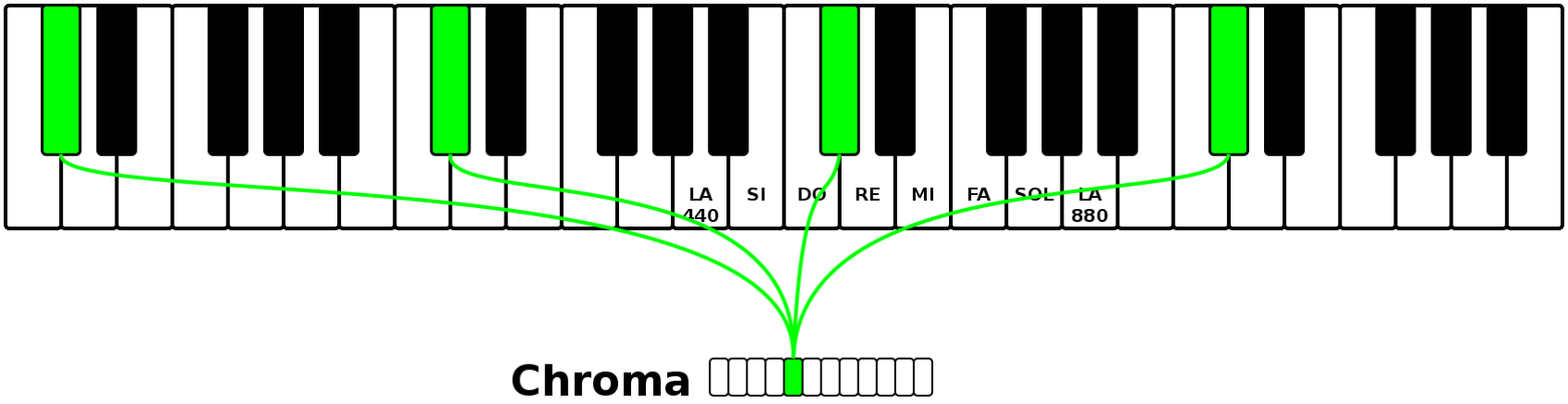
Principe : avoir une description synthétique des notes jouées.

Descripteurs : vecteur de chroma



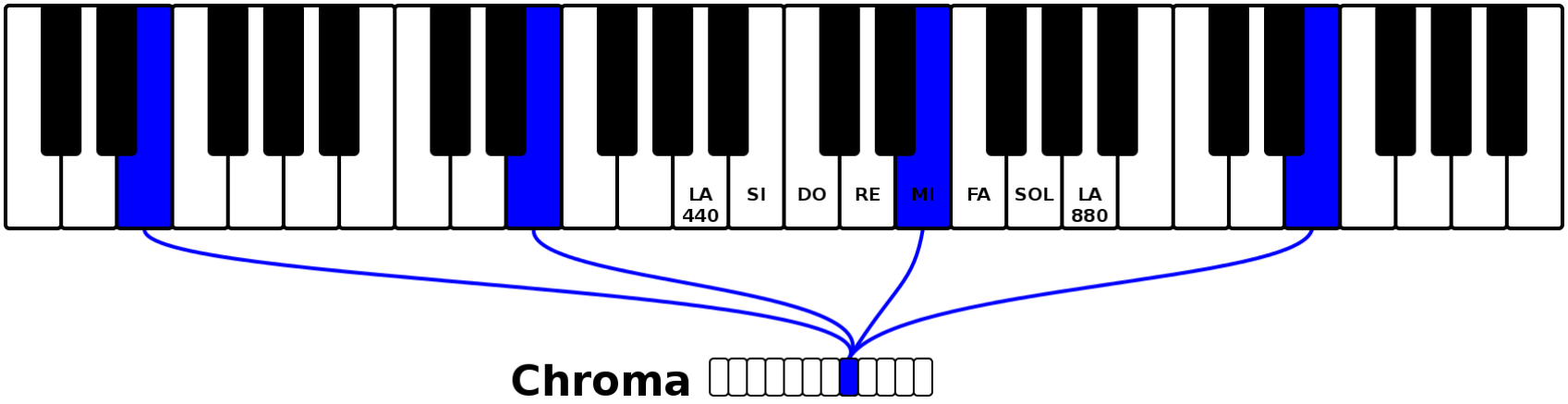
Principe : avoir une description synthétique des notes jouées.
Idée : invariance d'octave.

Descripteurs : vecteur de chroma



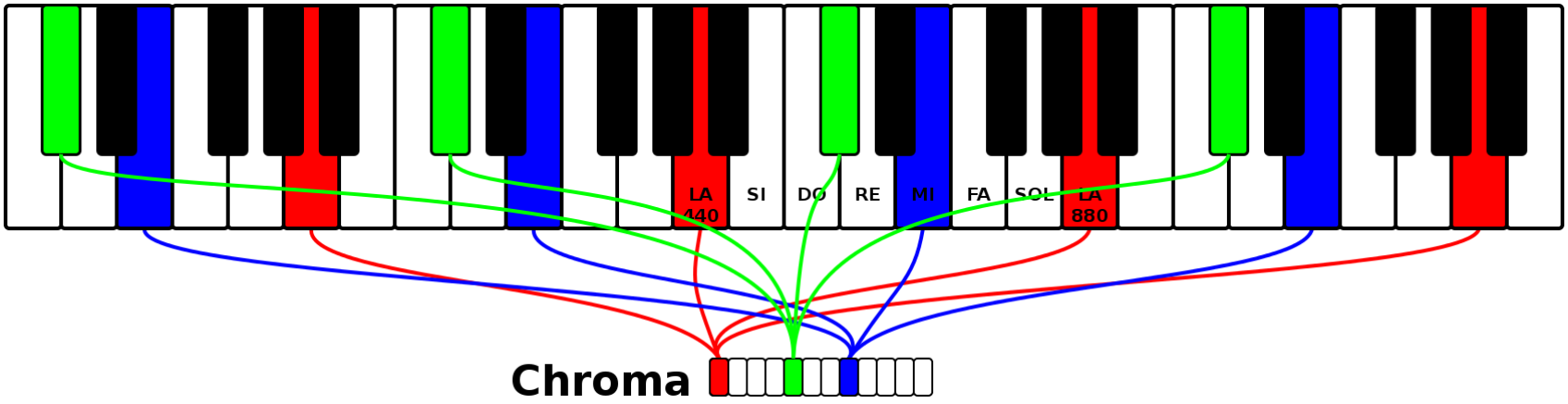
Principe : avoir une description synthétique des notes jouées.
Idée : invariance d'octave.

Descripteurs : vecteur de chroma



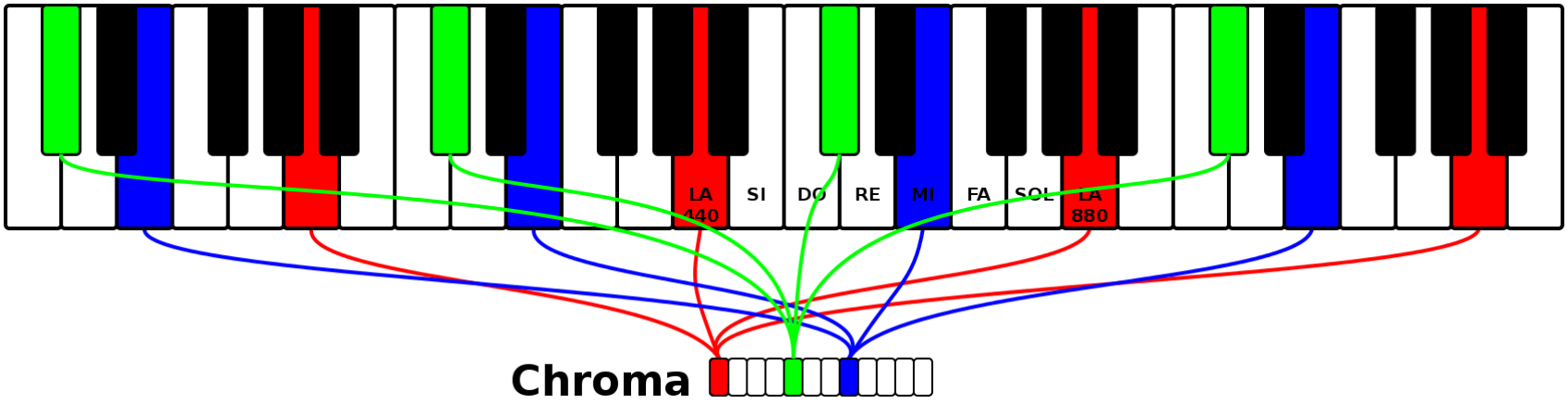
Principe : avoir une description synthétique des notes jouées.
Idée : invariance d'octave.

Descripteurs : vecteur de chroma



Principe : avoir une description synthétique des notes jouées.
Idée : invariance d'octave.
Résultat : description approximative du contenu harmonique.

Descripteurs : vecteur de chroma



Principe : avoir une description synthétique des notes jouées.
Idée : invariance d'octave.
Résultat : description approximative du contenu harmonique.
Pour une trame m, le coefficient k est

$$\mathbf{c}_m(k) \triangleq \sum_{p=p_{\mathbf{min}}}^{p_{\mathbf{max}}} \left| \mathbf{X}_m \left(\left[f_0^{(k,p)} \frac{N_{\text{fft}}}{f_s} \right] \right) \right|$$

avec

p octave, $N_{\mathbf{c}}$ taille de $\mathbf{c}_m$

$$f_0^{(k,p)} \triangleq 440 \times 2^{\frac{k}{N_{\mathbf{c}}}} \times 2^p$$

Références sur les descripteurs

B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software, proceedings of ISMIR, 2010.

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

T. Bertin-Mahieux, D. P.W. Ellis, B. Whitman, P. Lamere, The Million Song Dataset, proceedings ISMIR, 2011.

Objectif : estimation de la structure musicale d'un morceau

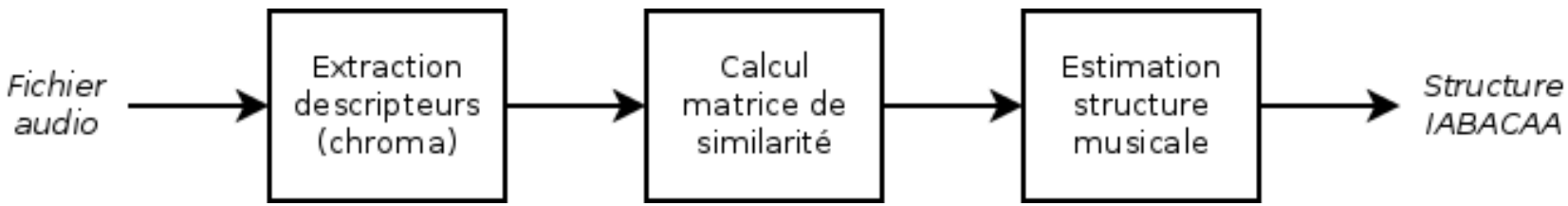
Éléments de la structure d'un morceau

- Nature : Intro, refrain, couplets, break, etc.
- Propriété 1 : répétition de chaque élément
- Propriété 2 : homogénéité de chaque élément

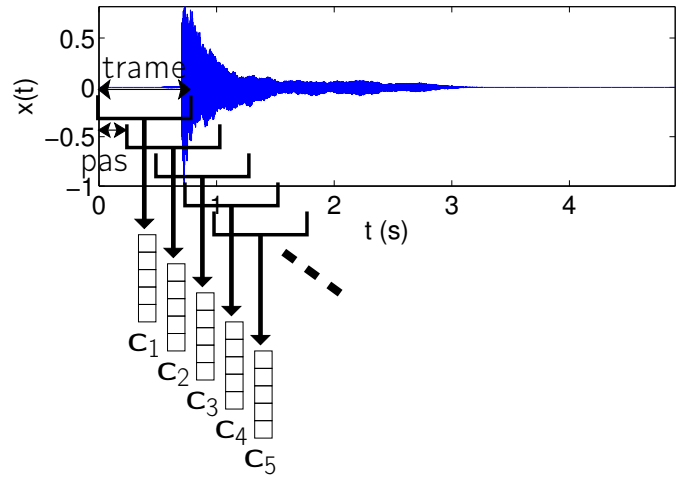
Objectif

Segmenter le morceau selon sa structure d'auto-similarité.

Principales étapes du système



Lecture du son, calcul des descripteurs



Son $\mathbf{x} = [\mathbf{x}(0), \ldots, \mathbf{x}(N_\mathbf{x} - 1)]^T \in \mathbb{R}^{N_\mathbf{x}}$, de fréquence d'échantillonnage f_s.
Trames de longueur N, indexée par $m \in [\![0, M-1]\!]$:

$$\mathbf{x}_m \triangleq [\mathbf{x}(n_m), \ldots, \mathbf{x}(n_m + N - 1)]$$

avec indice de début de trame $n_m \triangleq mh$ et pas h (*hop size*).

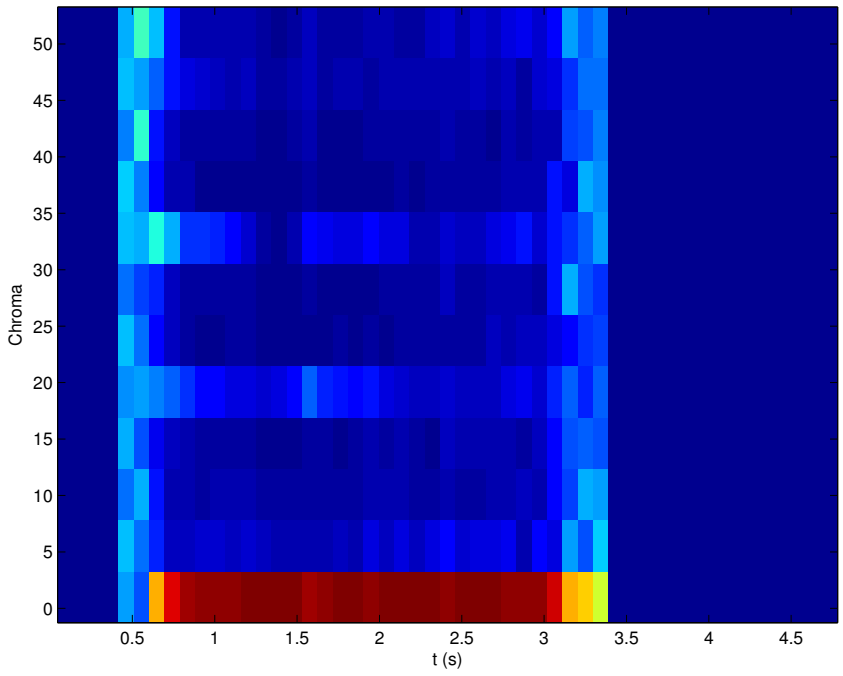
Instant milieu de trame $t_m \triangleq \frac{n_m + \frac{N}{2}}{f_s}$.

Calcul du descripteur $\mathbf{c}_m$ pour chaque trame $\mathbf{x}_m$.

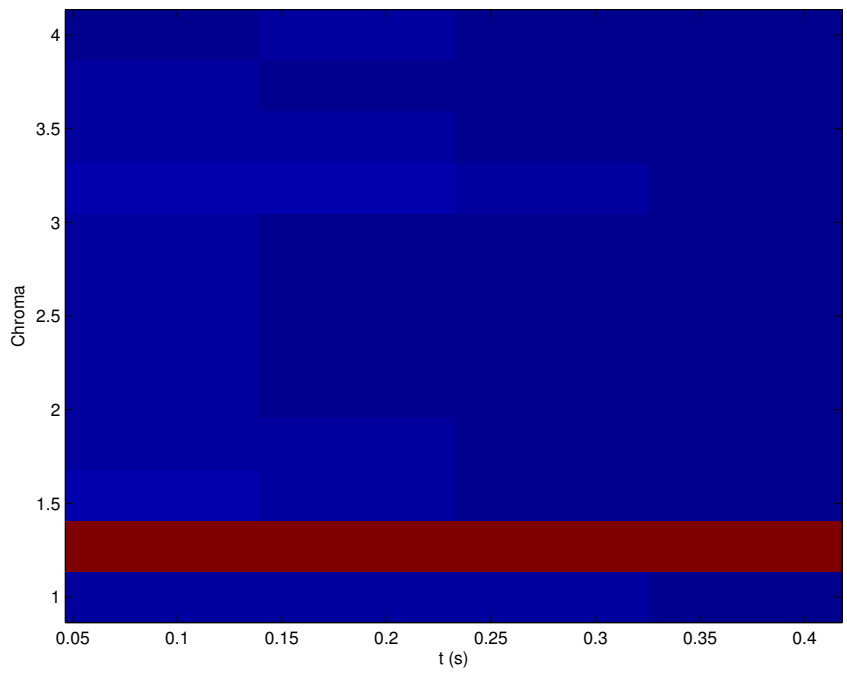
Test de l'extraction des chromas

Affichage du résultat pour deux sons simples (notes isolées).

Exemple avec un La



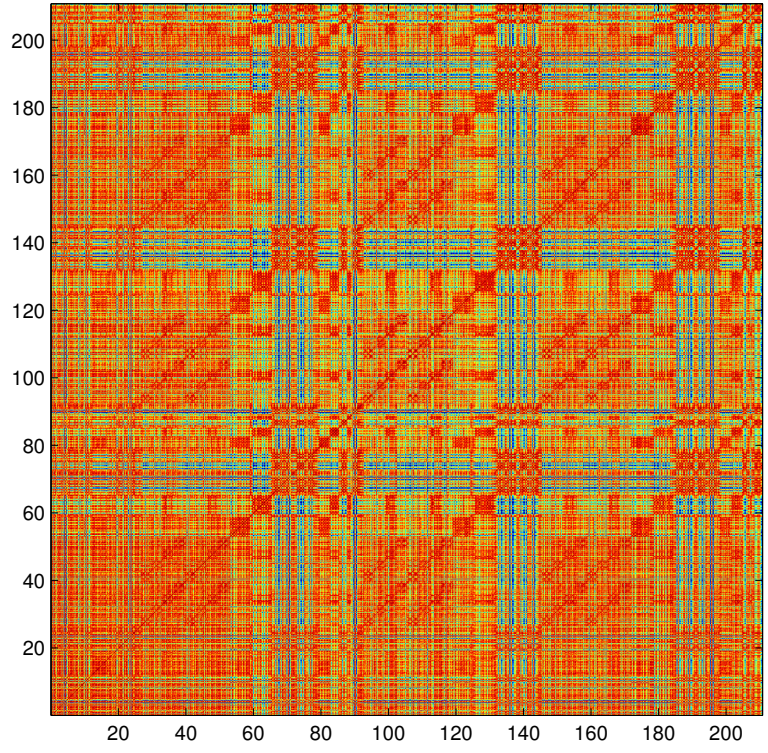
Exemple avec un La#



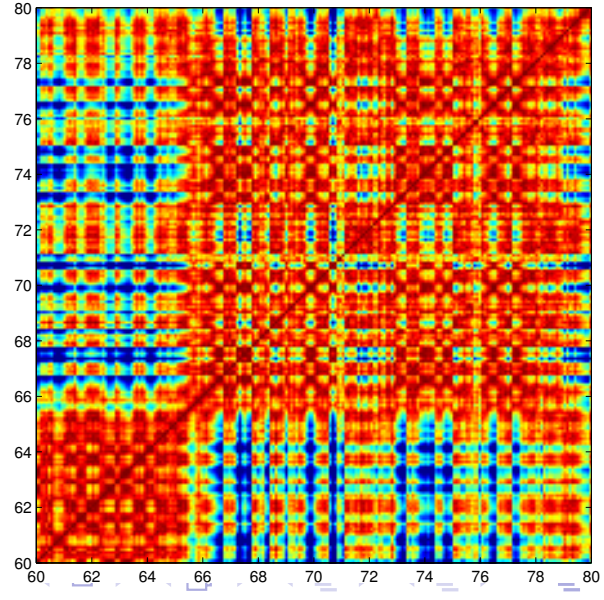
Matrice d'autosimilarité : définition

Matrice de similarité $\mathbf{\Gamma} \in \mathbb{R}^{M \times M}$:

$$\mathbf{\Gamma}\left(m_1, m_2\right) \triangleq \frac{\left\langle \mathbf{c}_{m_1}, \mathbf{c}_{m_2} \right\rangle}{\left\| \mathbf{c}_{m_1} \right\|_2 \left\| \mathbf{c}_{m_2} \right\|_2}$$



Zoom :

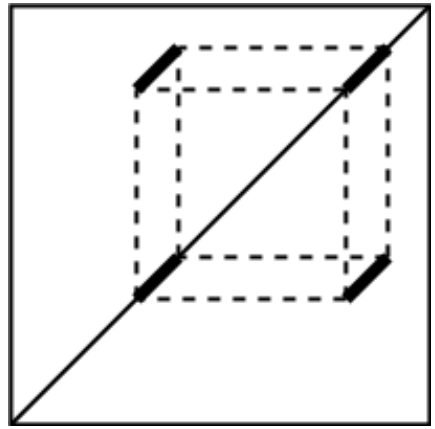


Matrice d'autosimilarité : utilité

Une séquence répétée

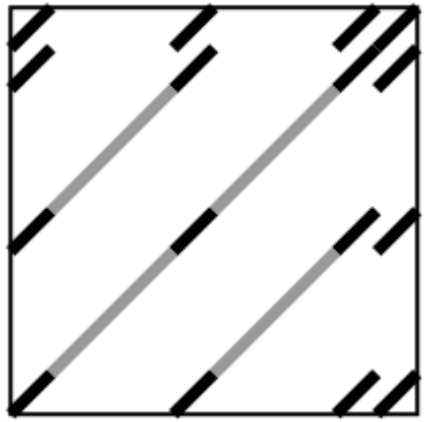
=

Un segment de diagonale



Répétition d'une séquence

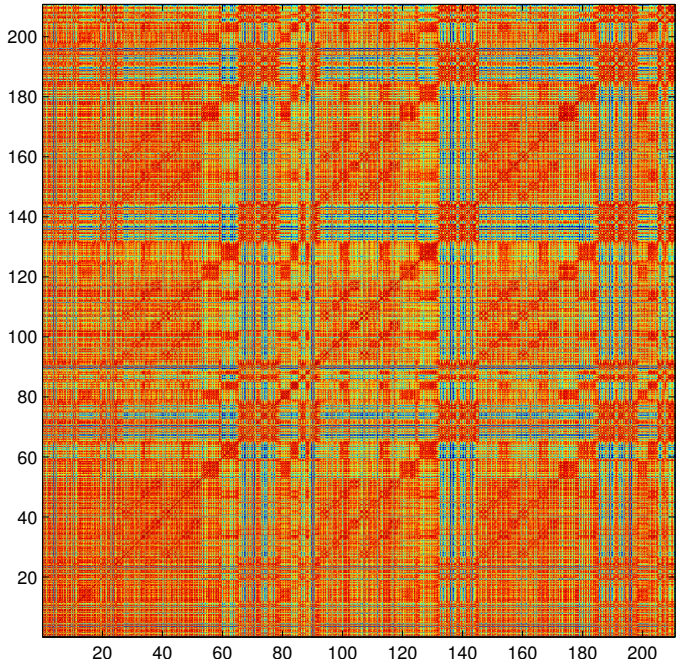
$$\mathbf{\Gamma}\left(m_1, m_2\right) \triangleq \frac{\left\langle \mathbf{c}_{m_1}, \mathbf{c}_{m_2} \right\rangle}{\left\| \mathbf{c}_{m_1} \right\|_2 \left\| \mathbf{c}_{m_2} \right\|_2}$$



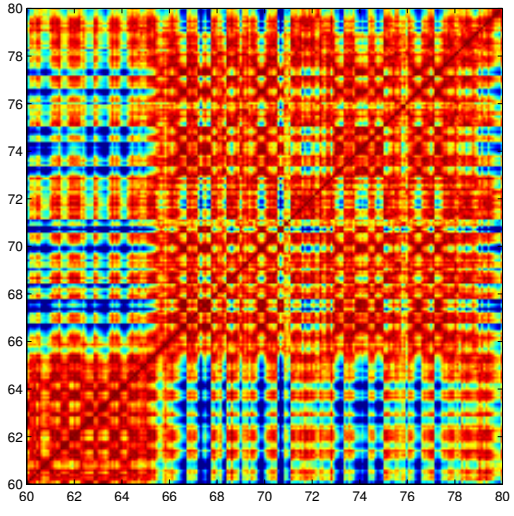
ABABAA

Réhaussement des motifs diagonaux

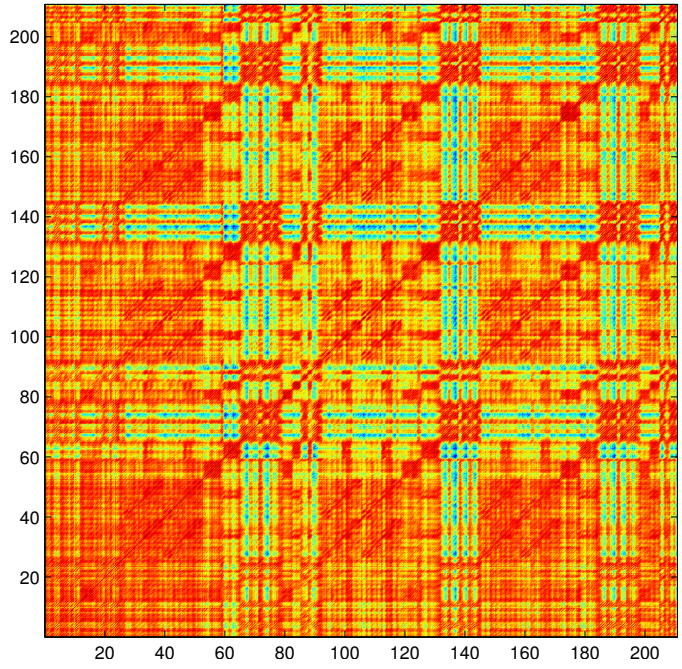
Avant filtrage



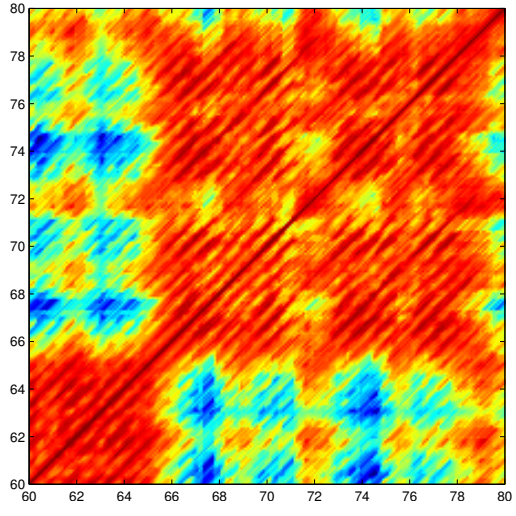
Zoom :



Après filtrage



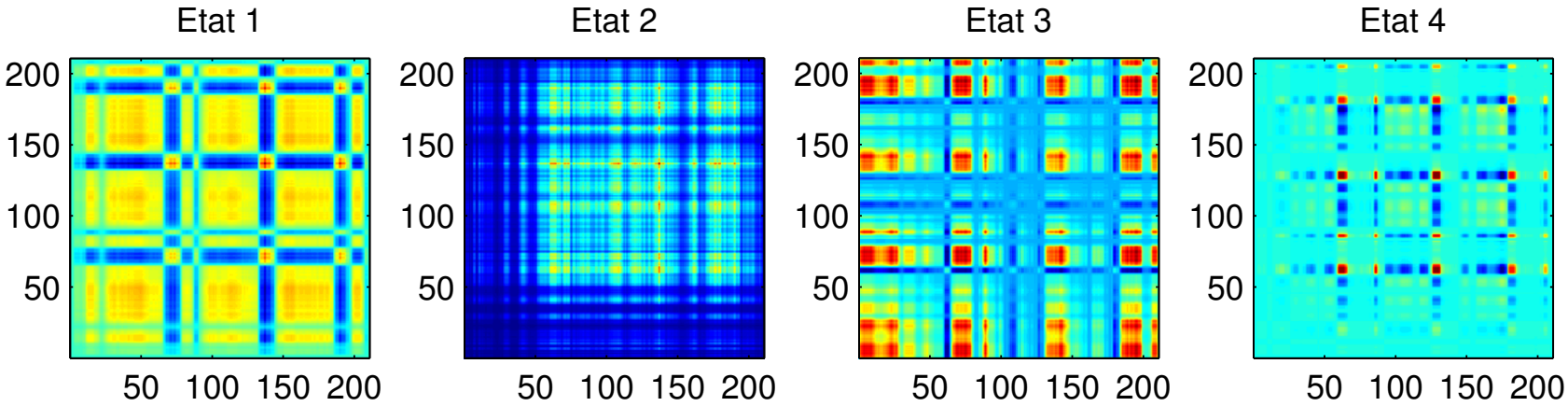
Zoom :



Estimation des segments

Principe : approximation de la matrice d'autosimilarité comme une somme de matrices de rang 1.

Méthode : décomposition en valeurs singulières.



Questions ?