Descripteurs audio:
de la simple représentation …
… aux modèles de connaissances

Geoffroy.Peeters@ircam.fr

UMR SMTS IRCAM CNRS UPMC

ircam
Centre
Pompidou

Introduction

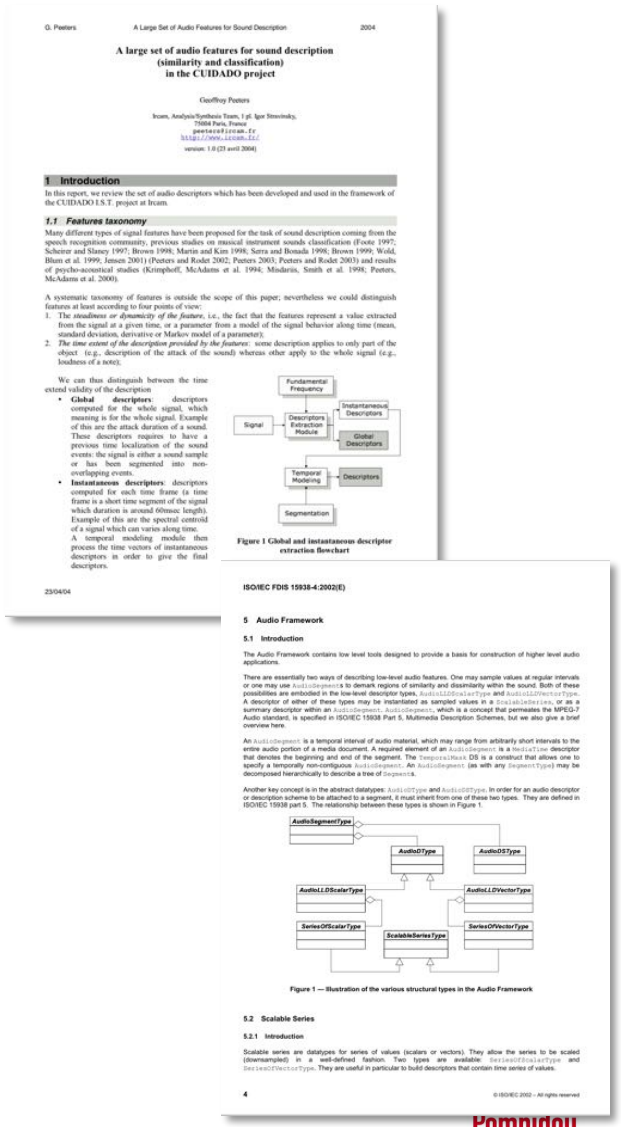
- L'analyse musicale à l'heure des outils multimédia
  - Coordination : Jean-Marc Chouvel
- Pertinence, usage et interprétation des descripteurs pour l'analyse
  - Coordination de la séance : Alain Bonardi et Pierre Couprie
  - Sous la dénomination de « descripteurs audio », un nombre considérable d'outils ont été développes ces dernières années qui proposent de mettre en évidence des aspects spécifiques du signal audio. Bien que la plupart aient été développes dans un contexte, celui de la recherche d'information musicale(M.I.R), assez éloigné des préoccupations de l'analyse, on peut toutefois interroger la pertinence de ces « descripteurs » pour la compréhension du phénomène musical. Quelles sont les catégories mises en évidence ? Quel est le potentiel de leur combinaison ? Quels aspects du message musical mettent-ils en avant ? et quels aspects leur restent inaccessibles ? Leur démultiplication est-il un gage d'adéquation ou une impuissance à délivrer une synthèse ?

. . .

- Traitement du signal
  - Extraction d'une représentation du signal permettant sa manipulation
  - Fréquence fondamentale ? Est-ce un descripteur ?

  - MPEG-7 Audio
    - Lancaster,
    - Descripteur/ Features/ Attributs Acoustique
  - Collection de descripteurs de différents domaines
    - timbreToolBox
  - Développement de descripteurs
    - ircamdescriptor
  - Abstraction progressive
    - Ircam{descriptor,beat,chord,keymode,structure}

  - Question fréquente: « quel descripteur utiliser pour faire ceci? »
  - Descripteurs
    - génériques (applicable à tout type de signaux)
    - descripteurs spécifiques (suppose un son isolé, modèle harmonique, musique composée temporellement)



G. Peeters A Large Set of Audio Features for Sound Description 2004

A large set of audio features for sound description (similarity and classification) in the CUIDADO project

Geoffroy Peeters

Ircam, Analysis/Synthesis Team, 1 pl. Igor Stravinsky, 75004 Paris, France
peeters@ircam.fr
http://www.ircam.fr/
version. 1.0 (23 avril 2004)

1 Introduction

In this report, we review the set of audio descriptors which has been developed and used in the framework of the CUIDADO I.S.T. project at Ircam.

1.1 Features taxonomy

Many different types of signal features have been proposed for the task of sound description coming from the speech recognition community, previous studies on musical instrument sounds classification (Foote 1997; Scheirer and Slaney 1997; Brown 1998; Martin and Kim 1998; Serra and Bonada 1998; Brown 1999; Wold, Blum et al. 1999; Jensen 2001) (Peeters and Rodet 2002; Peeters 2003; Peeters and Rodet 2003) and results of psycho-acoustical studies (Krimphoff, McAdams et al. 1994; Misdariis, Smith et al. 1998; Peeters, McAdams et al. 2000).

A systematic taxonomy of features is outside the scope of this paper; nevertheless we could distinguish features at least according to four points of view:

1. The *steadiness or dynamicity of the feature*, i.e., the fact that the features represent a value extracted from the signal at a given time, or a parameter from a model of the signal behavior along time (mean, standard deviation, derivative or Markov model of a parameter);
2. *The time extent of the description provided by the features*: some description applies to only part of the object (e.g., description of the attack of the sound) whereas other apply to the whole signal (e.g., loudness of a note);

We can thus distinguish between the time extend validity of the description

- **Global descriptors**: descriptors computed for the whole signal, which meaning is for the whole signal. Example of this are the attack duration of a sound. These descriptors requires to have a previous time localization of the sound events: the signal is either a sound sample or has been segmented into non-overlapping events.
- **Instantaneous descriptors**: descriptors computed for each time frame (a time frame is a short time segment of the signal which duration is around 60msec length). Example of this are the spectral centroid of a signal which can varies along time. A temporal modeling module then process the time vectors of instantaneous descriptors in order to give the final descriptors.

Figure 1 Global and instantaneous descriptor extraction flowchart

23/04/04

ISO/IEC FDIS 15938-4:2002(E)

5 Audio Framework

5.1 Introduction

The Audio Framework contains low level tools designed to provide a basis for construction of higher level audio applications.

There are essentially two ways of describing low-level audio features. One may sample values at regular intervals or one may use `AudioSegment`s to demark regions of similarity and dissimilarity within the sound. Both of these possibilities are embodied in the low-level descriptor types, `AudioLLDScalarType` and `AudioLLDVectorType`. A descriptor of either of these types may be instantiated as sampled values in a `ScalableSeries`, or as a summary descriptor within an `AudioSegment`. `AudioSegment`, which is a concept that permeates the MPEG-7 Audio standard, is specified in ISO/IEC 15938 Part 5, Multimedia Description Schemes, but we also give a brief overview here.

An `AudioSegment` is a temporal interval of audio material, which may range from arbitrarily short intervals to the entire audio portion of a media document. A required element of an `AudioSegment` is a `MediaTime` descriptor that denotes the beginning and end of the segment. The `TemporalMask` DS is a construct that allows one to specify a temporally non-contiguous `AudioSegment`. An `AudioSegment` (as with any `SegmentType`) may be decomposed hierarchically to describe a tree of `Segments`.

Another key concept is in the abstract datatypes: `AudioDType` and `AudioDSType`. In order for an audio descriptor or description scheme to be attached to a segment, it must inherit from one of these two types. They are defined in ISO/IEC 15938 part 5. The relationship between these types is shown in Figure 1.

Figure 1 — Illustration of the various structural types in the Audio Framework

5.2 Scalable Series

5.2.1 Introduction

Scalable series are datatypes for series of values (scalars or vectors). They allow the series to be scaled (downsampled) in a well-defined fashion. Two types are available: `SeriesOfScalarType` and `SeriesOfVectorType`. They are useful in particular to build descriptors that contain time series of values.

4 © ISO/IEC 2002 – All rights reserved

Représentation visuelle/ numérique

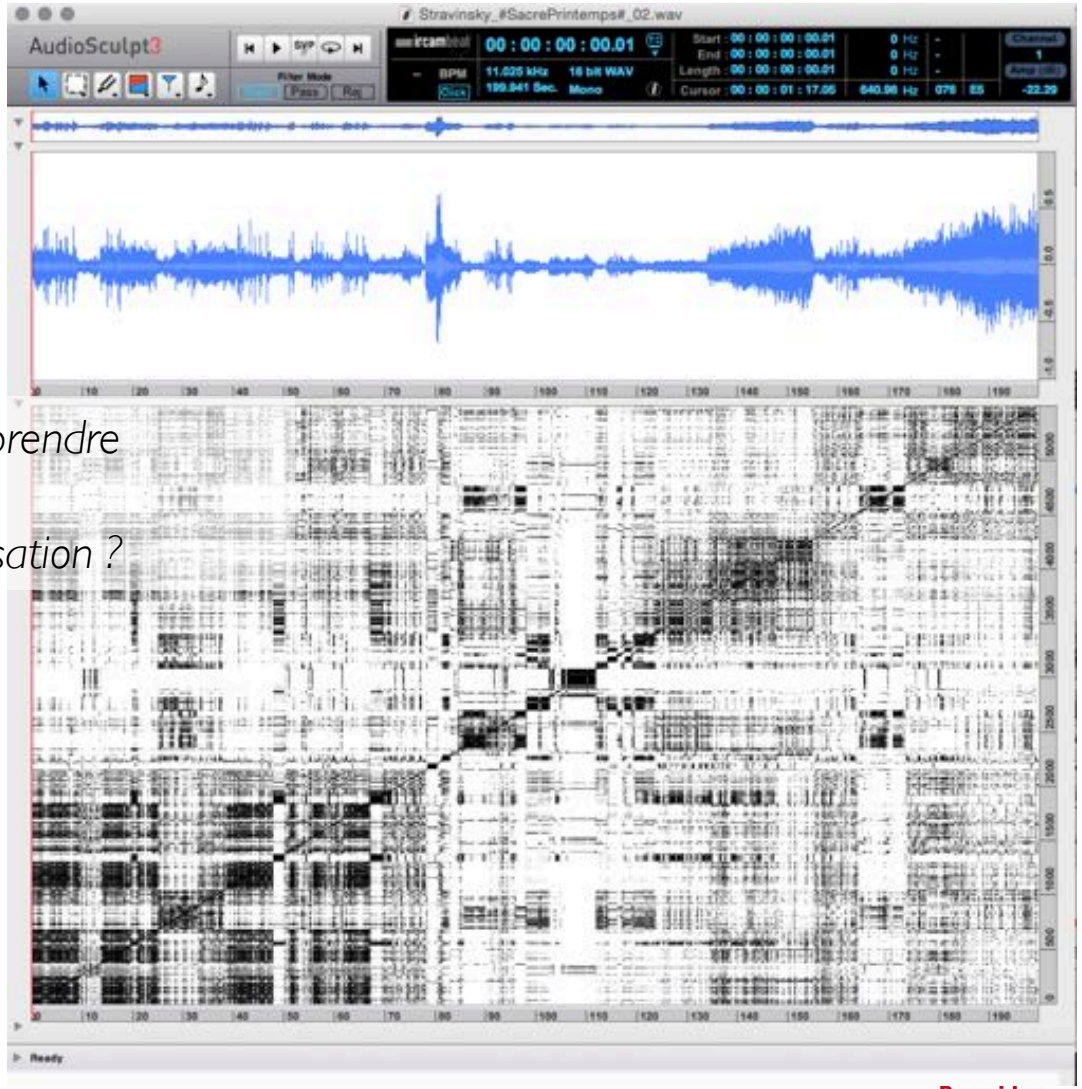
- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

Le spectre est un descripteur (multidimensionnel)

Pompidou

Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl



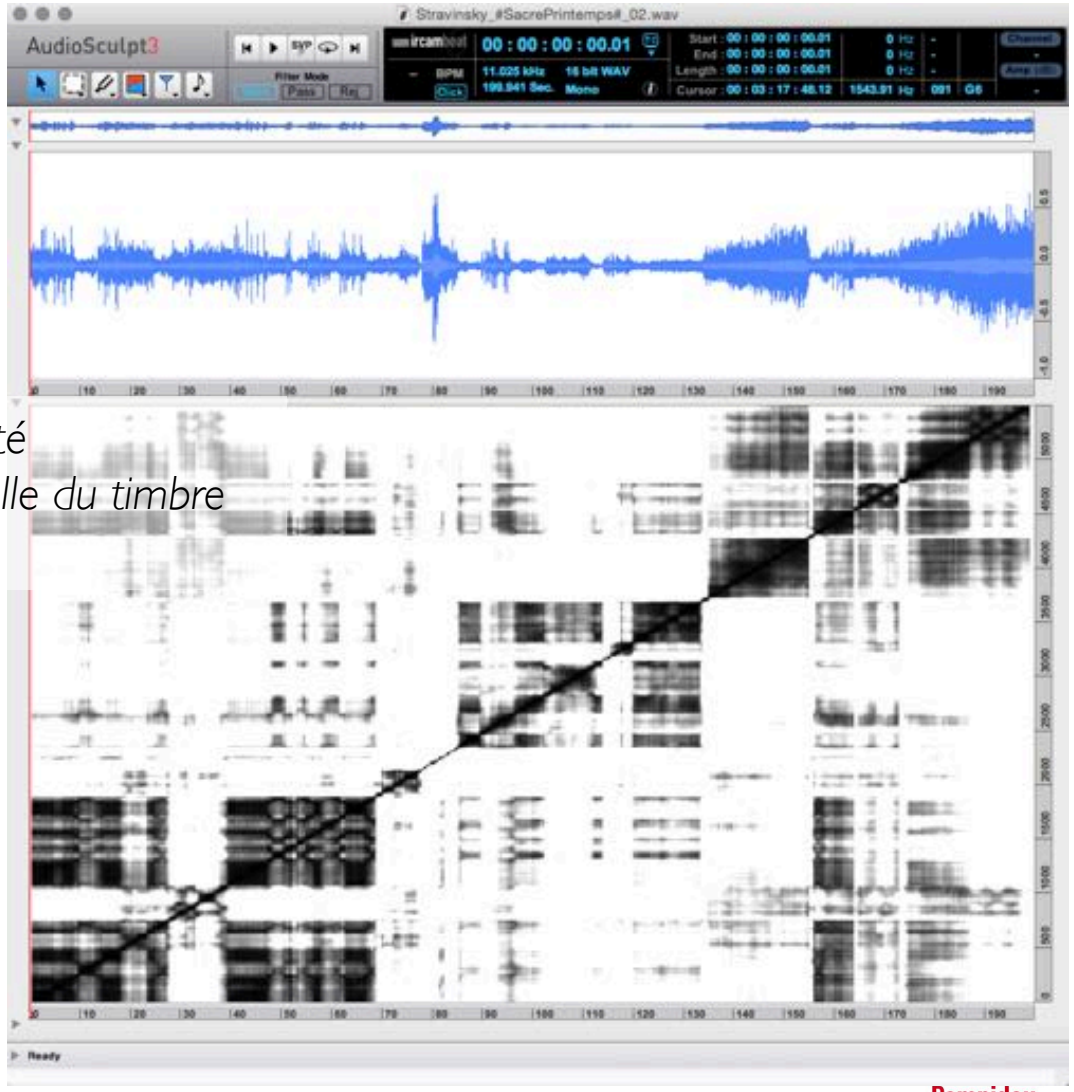
Matrice d'auto-similarité (aide à comprendre l'organisation temporelle),
Est-ce un descripteur ? Ou une visualisation ?

Pompidou

Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

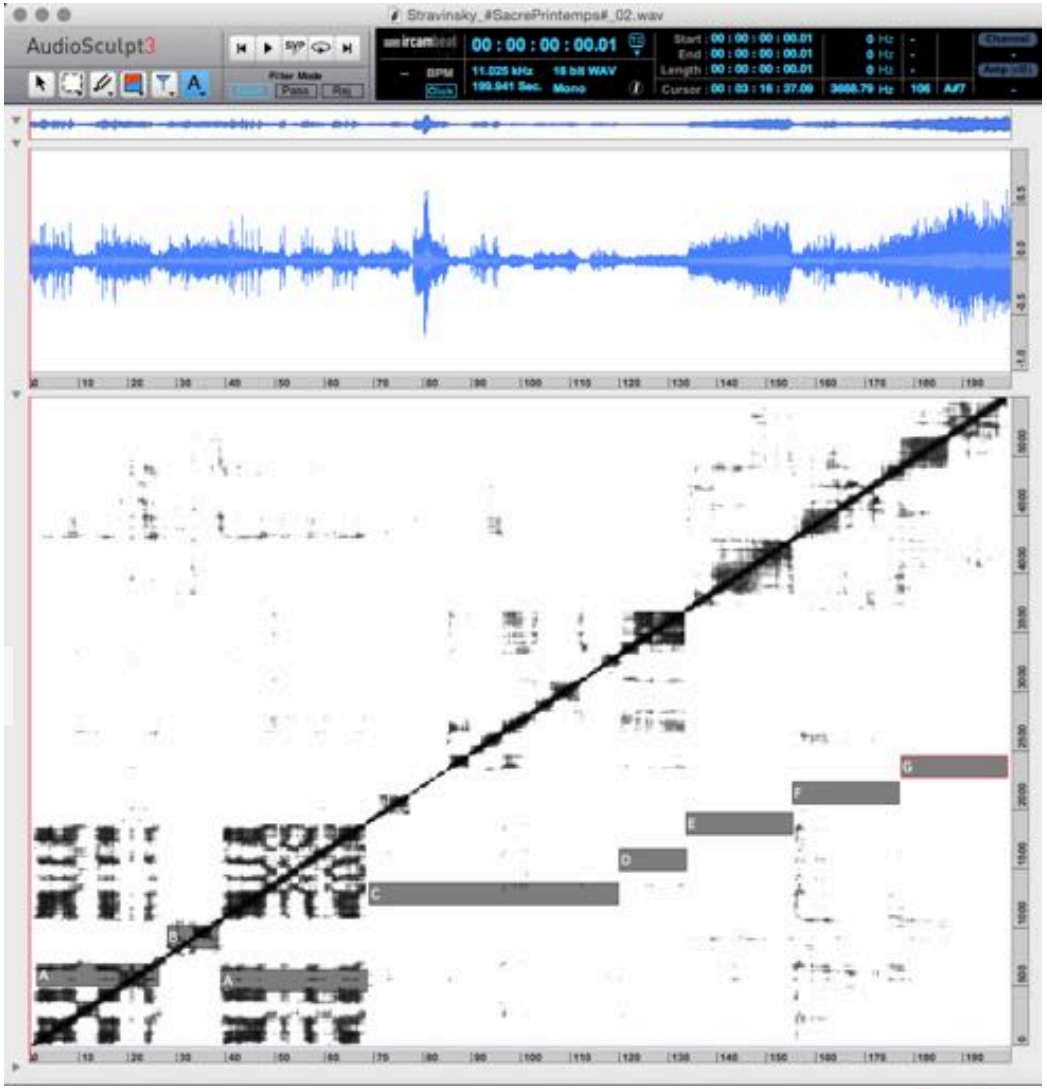
Egalement une matrice d'auto-similarité mais cette fois de la variation temporelle du timbre (paramétrisation des descripteurs)



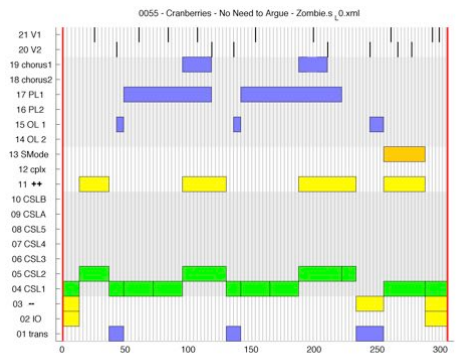
Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

D'une visualisation de l'organisation temporelle à son interprétation automatique en en terme de structure, A,B,C,D,E,…



Exemple annotations en XML

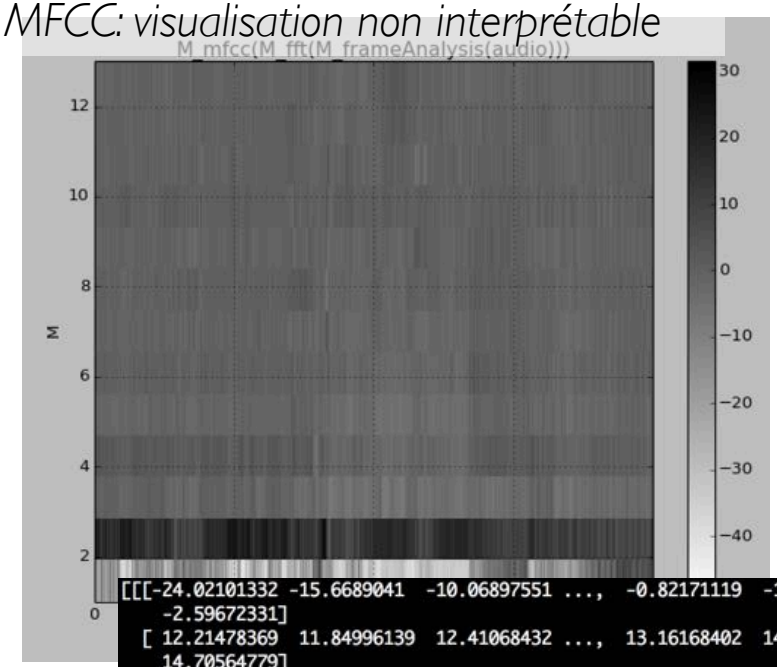


Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

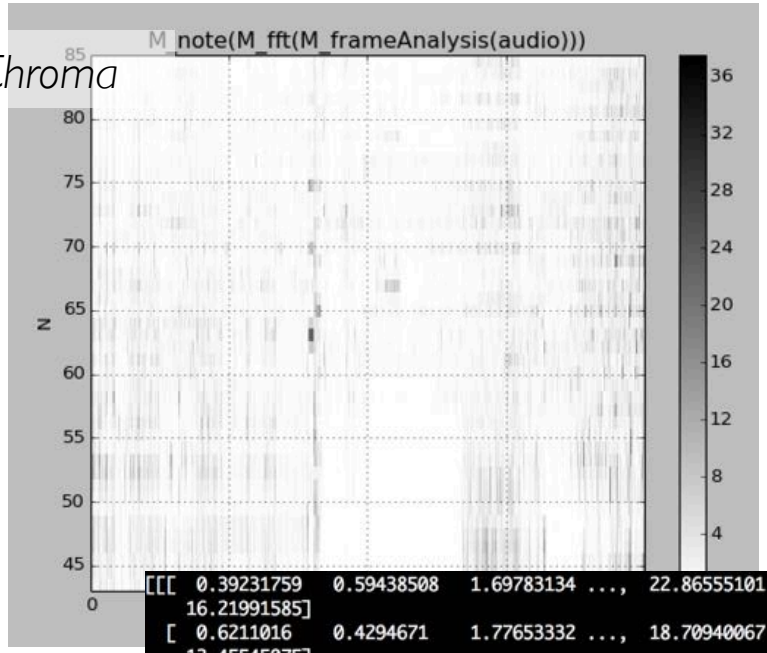
Tous les descripteurs ne peuvent pas être visualisés

MFCC: visualisation non interprétable




```
[[[-24.02101332 -15.6689041  -10.06897551 ...,  -0.82171119  -1.10674157
   -2.59672331]
  [ 12.21478369  11.84996139  12.41068432 ...,  13.16168402  14.40642901
   14.70564779]
  [ -1.77866866  -0.22262451   1.22473912 ...,  -1.43189666  -1.00244673
   -0.99117163]
  ...,
  [  0.10291074   0.48611309   0.30621258 ...,   0.45295809   0.04691121
   -0.10189695]
  [  0.02943441  -0.10793704   0.21461394 ...,   0.14901765   0.17947462
    0.54331357]
  [ -0.74765662  -1.09759493  -1.45451008 ...,   1.30288005   0.69034685
   -0.06535942]]]
```

Chroma




```
[[[  0.39231759   0.59438508   1.69783134 ...,  22.86555101  22.43387537
    16.21991585]
  [  0.6211016    0.4294671    1.77653332 ...,  18.70940067  17.4759172
    13.45545075]
  [  0.6211016    0.4294671    1.77653332 ...,  18.70940067  17.4759172
    13.45545075]
  ...,
  [  0.11508825   0.26115764   0.44664189 ...,   1.31711814   1.44949199
     1.18626625]
  [  0.13690403   0.26011007   0.35950545 ...,   3.14657322   3.19713983
     2.72775906]
  [  0.46877125   0.69442424   0.72139293 ...,   1.24502597   1.18285085
     0.96912804]]]
```

Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

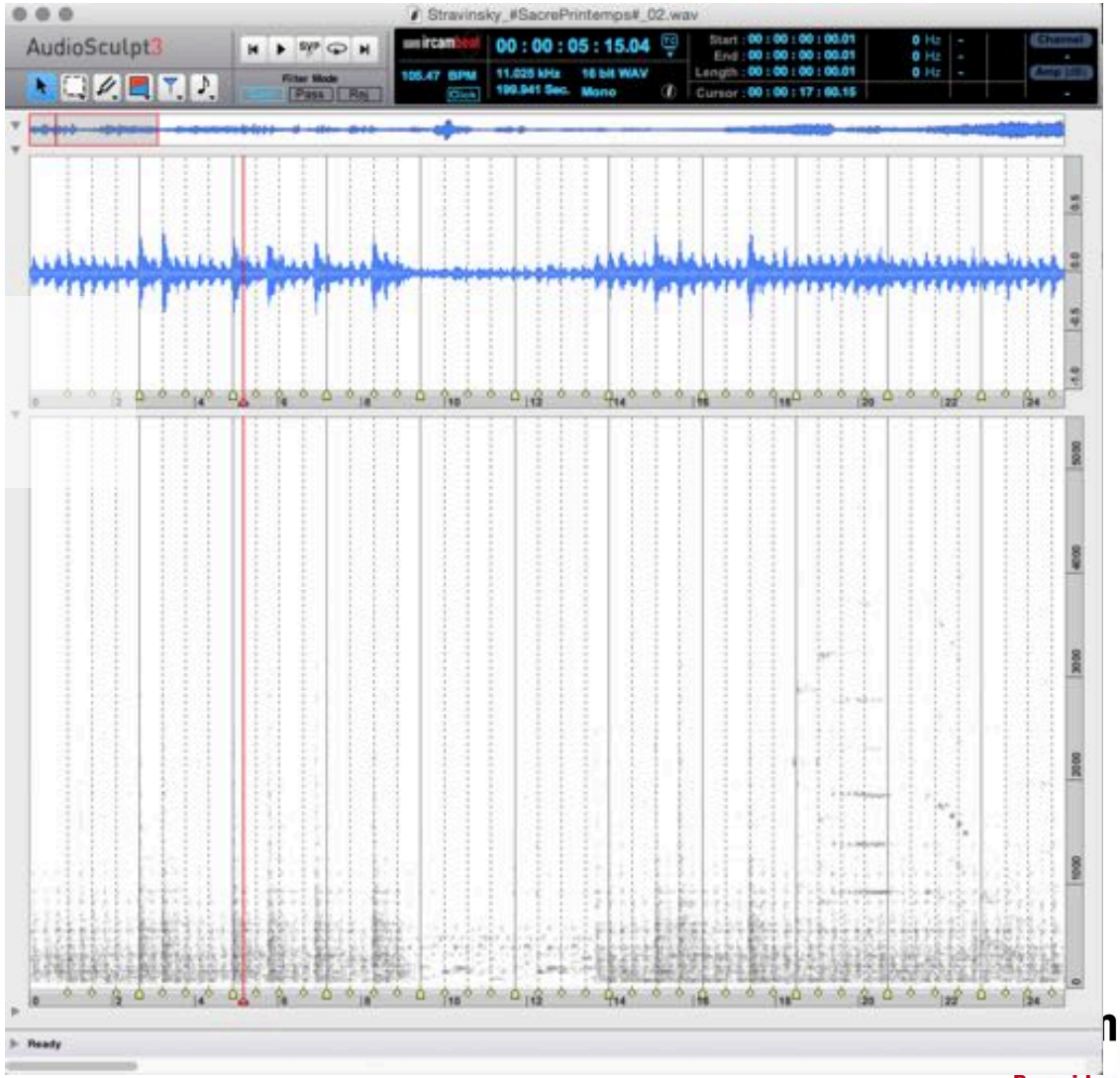
Tous les descripteurs ne peuvent pas être visualisés. Pour Audiosculpt -> choix d'un sous-ensemble d'ircamdescriptor

```
;; TextureWindowsFrames *        HopSize
; and the hopsize is
; TextureWindowsHopFrames * HopSize;
TextureWindowsFrames = -1
TextureWindowsHopFrames = -1

;~~~~~~~~~~~~~~~~~~descriptors~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
; temporal
SignalZeroCrossingRate = ShortTime MeanAndDeviation Delta DeltaDelta Median
TotalEnergy = ShortTime MeanAndDeviation Delta DeltaDelta Median

; spectral
AutoCorrelation = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralCentroid = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralSpread = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralSkewness = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralKurtosis = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralSlope = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralDecrease = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralRolloff = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralVariation = ShortTime MeanAndDeviation Delta DeltaDelta Median

; perceptual
Loudness = ShortTime MeanAndDeviation Delta DeltaDelta Median
RelativeSpecificLoudness = ShortTime MeanAndDeviation Delta DeltaDelta Median
Spread = ShortTime MeanAndDeviation Delta DeltaDelta Median
Sharpness = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralCentroid = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralSpread = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralSkewness = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralKurtosis = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralSlope = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralDecrease = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralRolloff = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralVariation = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualSpectralDeviation = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualOddToEvenRatio = ShortTime MeanAndDeviation Delta DeltaDelta Median
PerceptualTristimulus = ShortTime MeanAndDeviation Delta DeltaDelta Median

MFCC = ShortTime MeanAndDeviation Delta DeltaDelta Median

SpectralFlatness = ShortTime MeanAndDeviation Delta DeltaDelta Median
SpectralCrest = ShortTime MeanAndDeviation Delta DeltaDelta Median

; harmonic
FundamentalFrequency = ShortTime MeanAndDeviation Delta DeltaDelta Median
NoiseEnergy = ShortTime MeanAndDeviation Delta DeltaDelta Median
Noisiness = ShortTime MeanAndDeviation Delta DeltaDelta Median
Inharmonicity = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicEnergy = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralCentroid = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralSpread = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralSkewness = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralKurtosis = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralSlope = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralDecrease = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralRolloff = ShortTime MeanAndDeviation Delta DeltaDelta Median
HarmonicSpectralVariation = ShortTime MeanAndDeviation Delta DeltaDelta Median
```

Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

Descripteur « sémantique » <u>local</u>
Les battements/ premiers temps
modèle de connaissance

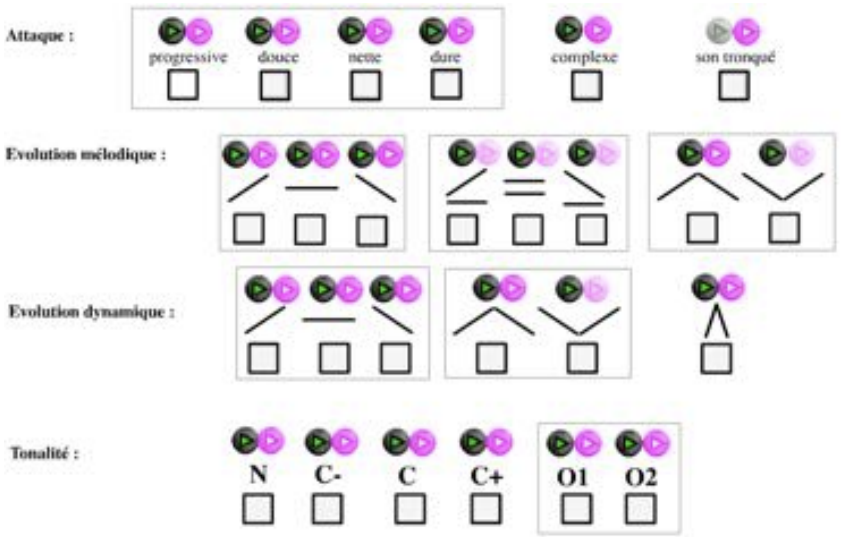


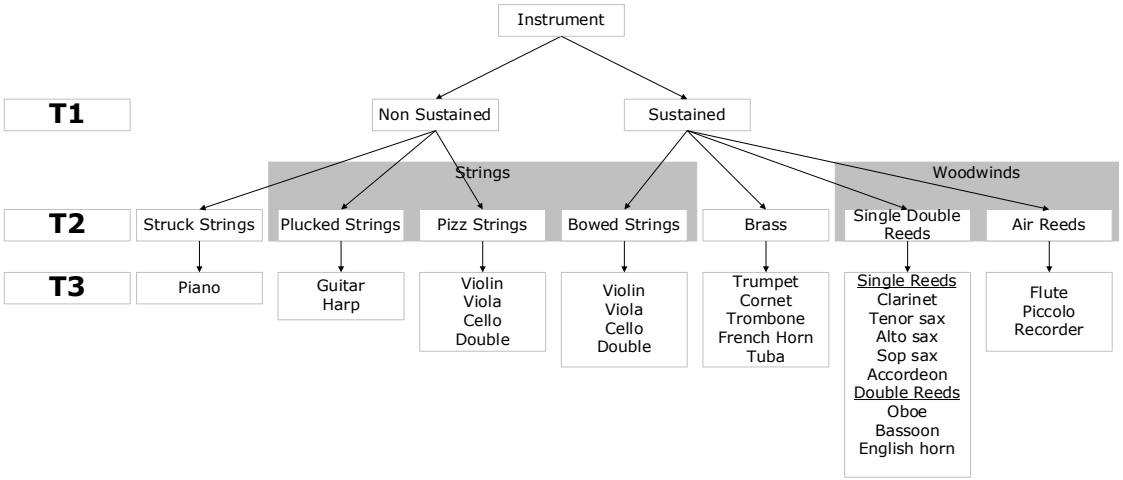
Pompidou

Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

Descripteur « sémantique » global
Profils temporelles de son
Classe d'instrument





ircam Centre Pompidou

Représentation visuelle/ numérique

- Stravinsky
  - The Rite of Spring
    - Dances of the Young Girl

```
ircamchord -i /Users/peeters/_sound/_collection/local_beat_perso/_FULL_style_classical/Stravinsky_#SacrePrintemps#_02.wav -o ./test.xml
```

Descripteurs « sémantiques » <u>locaux et globaux</u>, modèle de connaissance et apprentissage machine
descripteurs locaux (musicologie)
descripteurs globaux (indexation base de données)

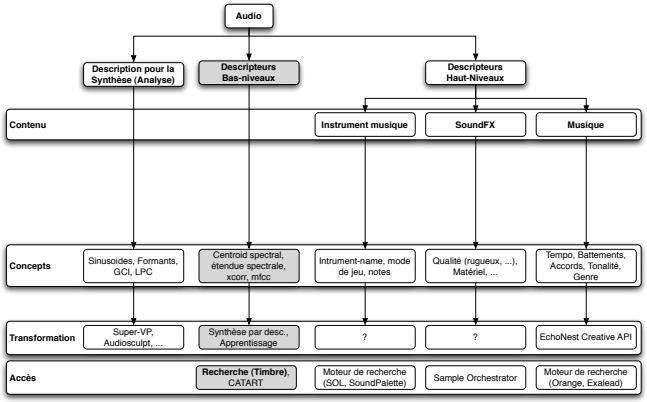
```
</descriptiondefinition>
<global sourcetrack="0">
    <tagtype id="2" value="Brass" confidence="0.7070876838" />
</global>
```

```
<segment time="0.0000000000" length="6.9123455770" sourcetrack="0">
    <chordtype id="1" value="Emaj" />
</segment>
<segment time="6.9123455770" length="2.2523356009" sourcetrack="0">
    <chordtype id="1" value="Amaj" />
</segment>
<segment time="9.1646811779" length="1.1435827664" sourcetrack="0">
    <chordtype id="1" value="Emaj" />
</segment>
<segment time="10.3082639443" length="3.3785034014" sourcetrack="0">
    <chordtype id="1" value="Emin" />
</segment>
<segment time="13.6867673457" length="6.8092517007" sourcetrack="0">
    <chordtype id="1" value="Emaj" />
</segment>
<segment time="20.4960190463" length="2.2581405896" sourcetrack="0">
    <chordtype id="1" value="Dbmin" />
</segment>
<segment time="22.7541596359" length="2.2523356009" sourcetrack="0">
    <chordtype id="1" value="Emaj" />
</segment>
<segment time="25.0064952368" length="0.2757369615" sourcetrack="0">
    <chordtype id="1" value="Abmin" />
</segment>
<segment time="25.2822321983" length="0.2757369615" sourcetrack="0">
    <chordtype id="1" value="Ebmaj" />
</segment>
<segment time="25.5579691597" length="1.6776417234" sourcetrack="0">
    <chordtype id="1" value="Bbmaj" />
</segment>
<segment time="27.2356108831" length="2.2465306122" sourcetrack="0">
    <chordtype id="1" value="Fmaj" />
</segment>
<segment time="29.4821414953" length="2.2581405896" sourcetrack="0">
    <chordtype id="1" value="Cmaj" />
</segment>
<segment time="31.7402820849" length="2.2871655329" sourcetrack="0">
    <chordtype id="1" value="Cmin" />
</segment>
```

cam
e
mpidou

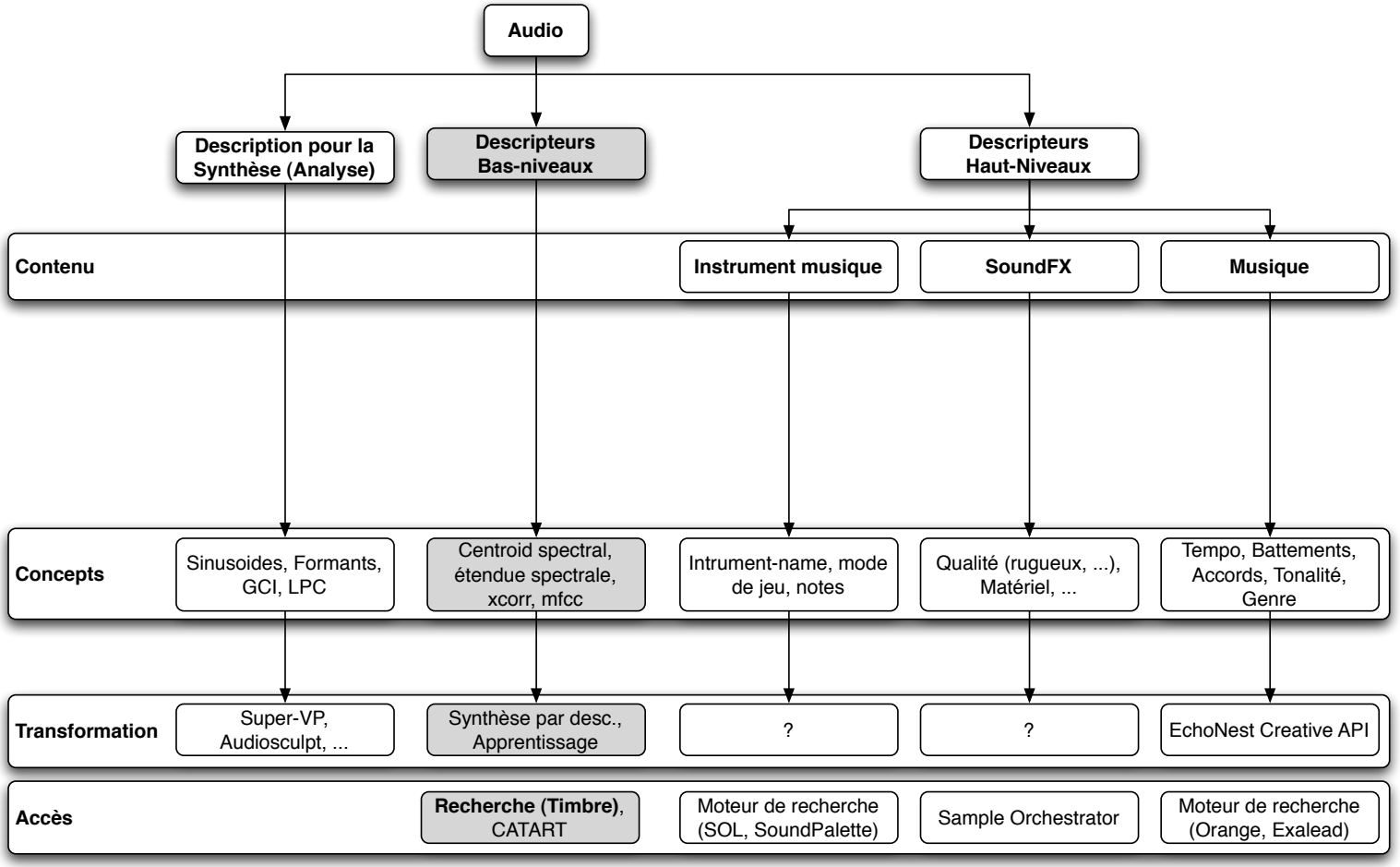
Différentes catégorisations possibles des descripteurs

- Descripteurs de bas/ de haut-niveau (cnfr MPEG-7 Audio: D: Descriptor, DS: Descriptor Scheme)
  - Niveau de complexité ? Niveau d'abstraction ?
  - Centroid Spectral -> Fréquence Fondamentale/ Pitch ? -> Accord ? -> Genre/ Humeur ?
- Descripteurs pour qui ?
  - Musicologue
  - Chercheurs
  - Amateurs de musique
  - Industriels
- Descripteurs pour quoi ?
  - Décrire un morceau de musique
    - Musicologie (descripteurs locaux en temps, compréhensibles)
  - Décrire une collection de morceau de musique
    - Gestion de base de données, moteurs de recherche (apprentissage machine)
  - Décrire des flux de signaux audio inconnus
  - Navigation visuelle dans une base de données
  - Synthèse/ manipulation
- Descripteurs sur quels types de sons ?
  - Descripteurs génériques
  - Descripteurs spécifiques (hypothèse d'harmonicité, d'unicité)



ircam Centre Pompidou

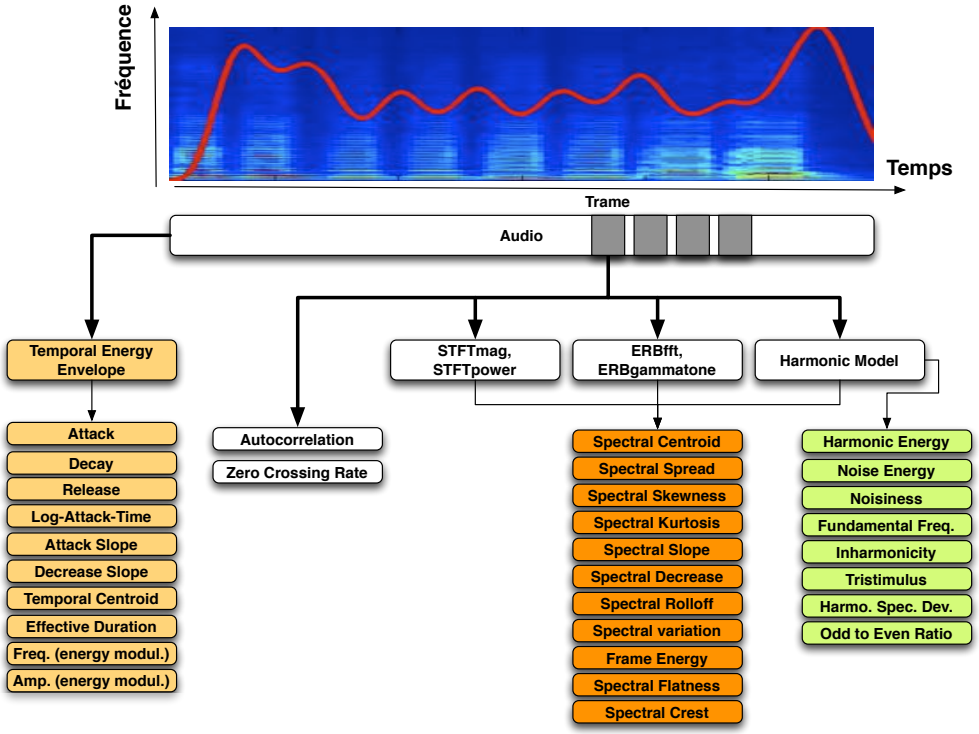
Différentes catégorisations possibles des descripteurs



ircam Centre Pompidou

Différentes catégorisations possibles des descripteurs

Beaucoup de variations sur un même concept …

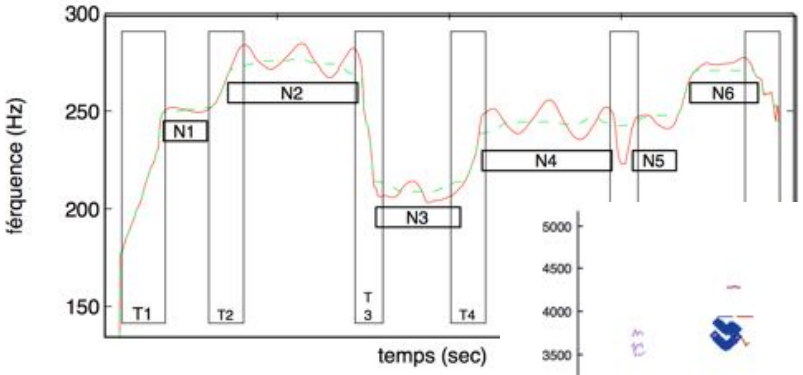


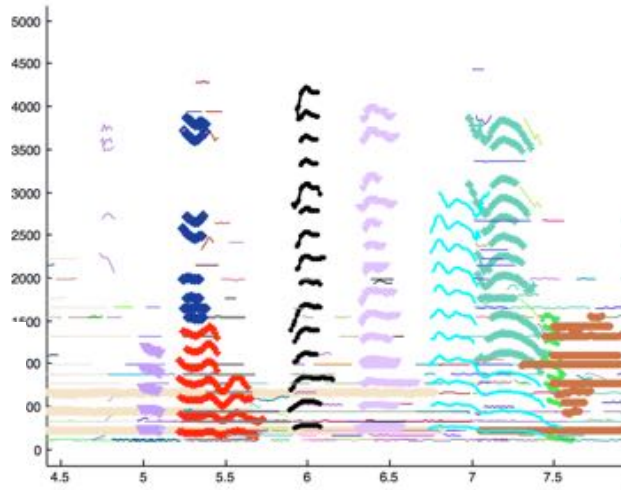
… mais également des dénomination communes (flux spectral, variation harmonique) pour des mesures très différentes

ircam Centre Pompidou

Différentes catégorisations possibles des descripteurs

Exemple de descripteurs spécifiques
Descripteurs intonatifs [Régnier]





Descripteurs stéréo [Tardieu]

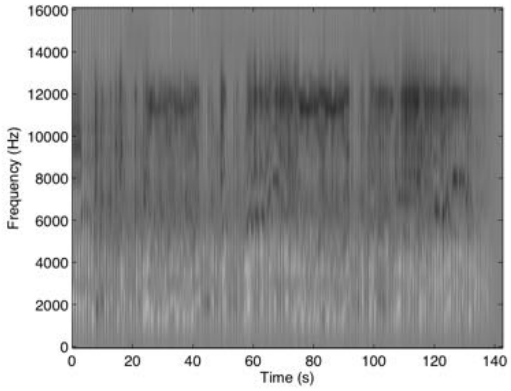


Figure 5: Cochleagram difference for the *While My Guitar Gently Weeps* from *The Beatles*. Color ranges from -.3 (white) representing right channel to .3 (black) representing left channel.

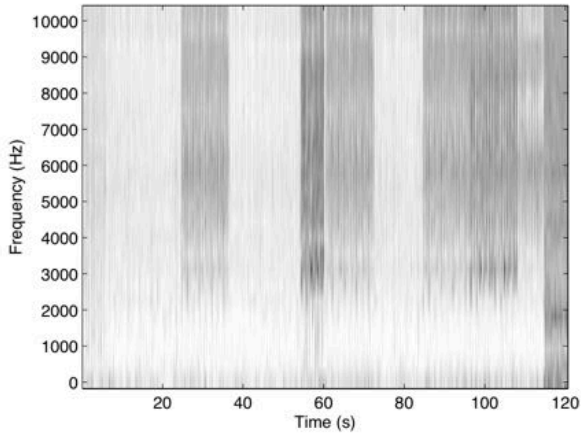


Figure 9: Cochleagram Difference (Top) and Spectral Stereo Phase Spread (Bottom) for the song *Gangsta's Paradise* by *Coolio*. In the Cochleagram Difference, Color ranges from -.3 (white) representing right channel to .3 (black) representing left channel. In the Spectral Stereo Phase Spread lighter colors represent higher phase

ircam
Centre Pompidou

Descripteurs et modèle de connaissance

- Différents types de descripteurs

  - 1) Série d'opérateurs mathématiques
    - Centroid spectral: Signal -> spectre -> barycentre
    - Chroma/ Pitch-Class-Profile: Signal -> spectre -> banc de filtres de chroma
    - Mel Frequency Cepstral Coefficient: Signal -> spectre -> bancs de filtres Mel -> DCT

  - 2) Estimateur (prise de décision, erreurs possibles)
    - Fréquence fondamentale (introduction de connaissance, série harmonique des pics du spectre)
    - Détection d'onset (maxima local, seuil)

  - 3) Modèle de connaissance
    - Connaissance acquise par un humain
    - Connaissance acquise par une machine

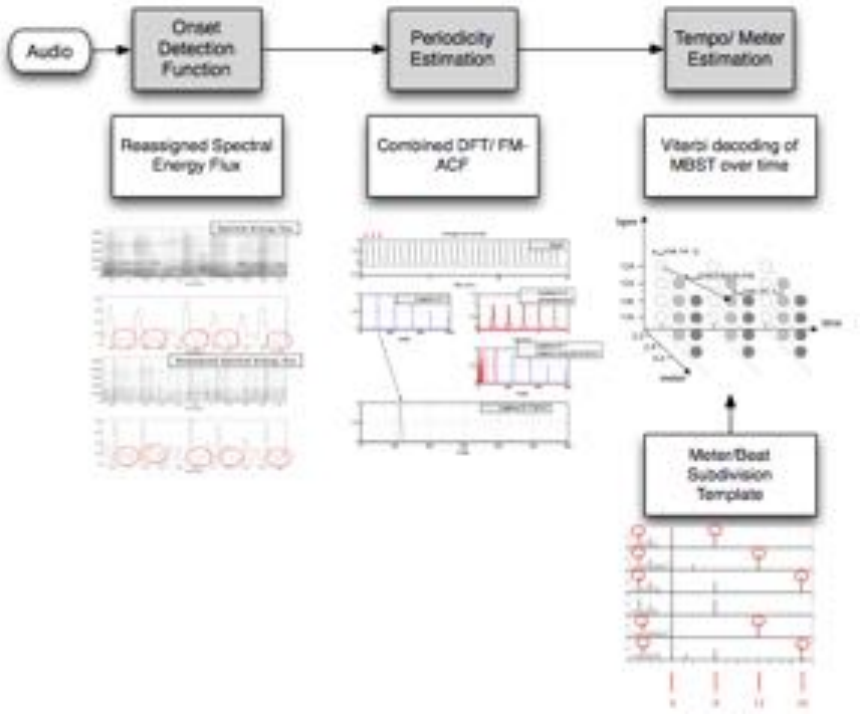
    - Accord
    - Tempo/Métrique/Premier-temps
    - Profils temporels, classe de rythme, genre, style, humeur

    - ATTENTION: les 2) et 3) peuvent se tromper ! Il est important de bien comprendre les descripteurs !

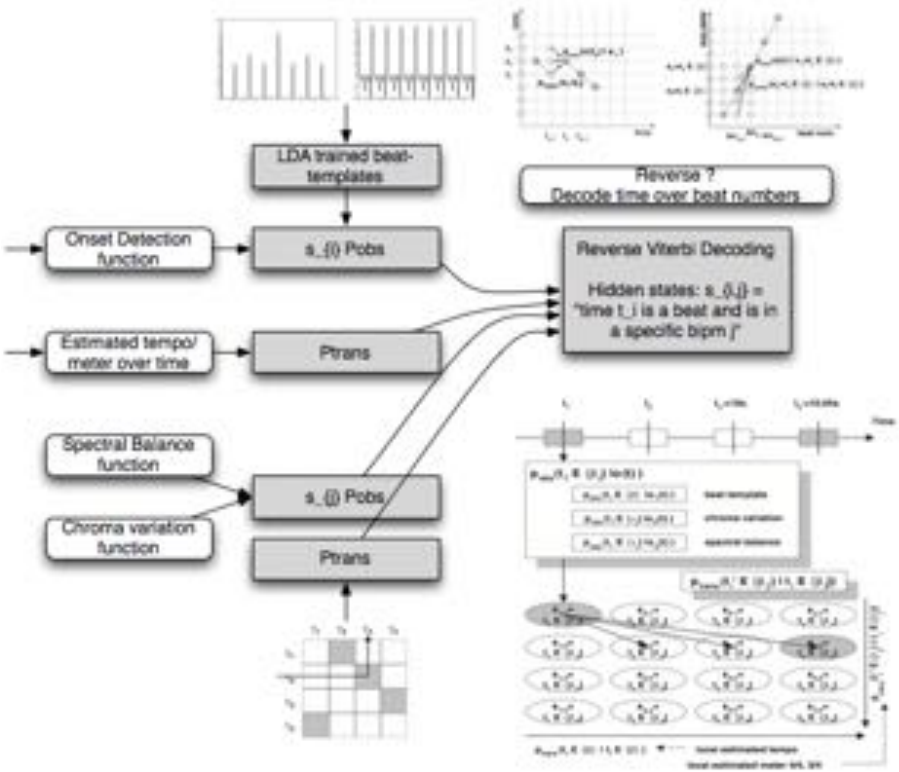
Descripteurs et modèle de connaissance

ircambeat

Détection de tempo/ métrique
-> Patterns de métrique



Détection de tempo/ métrique
-> Patterns d'énergie
-> variation harmonique, balance spectrale

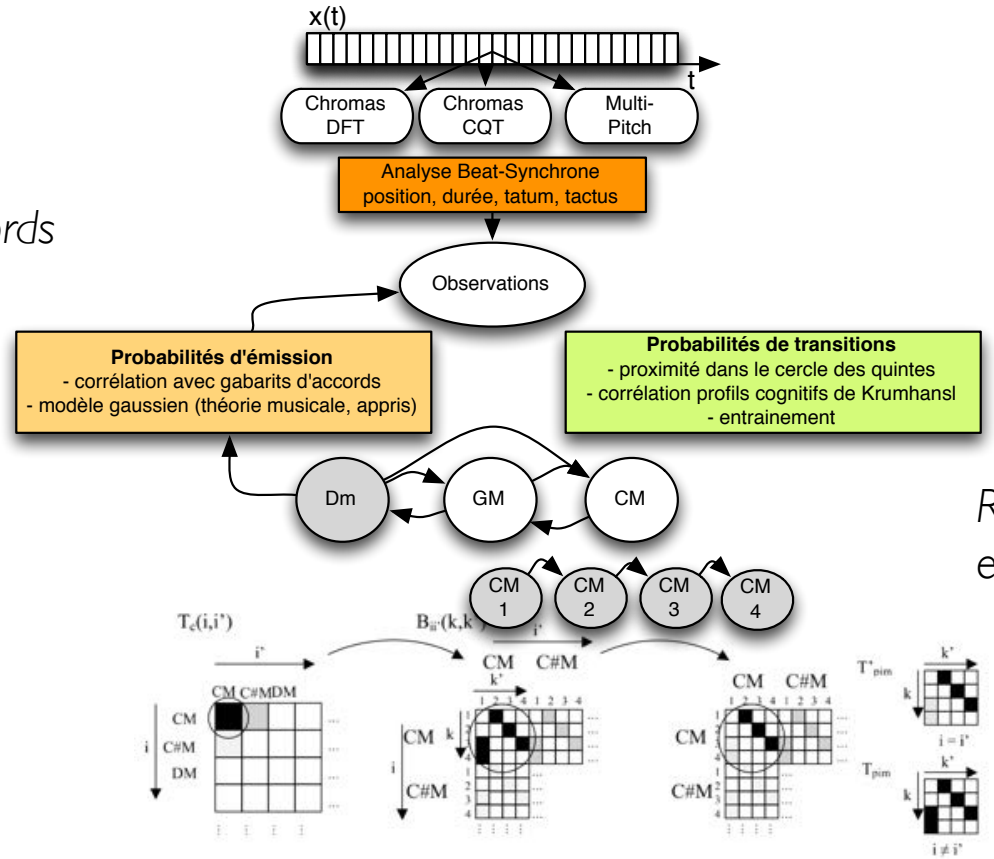


Centre Pompidou

Descripteurs et modèle de connaissance

ircamchord

Gabarits d'accords



Règles de transition entre accords

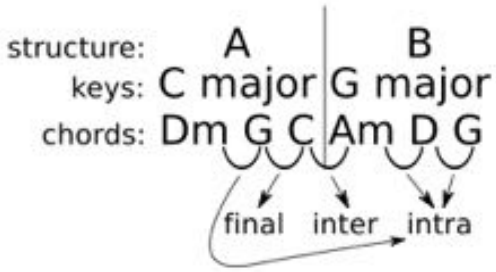
Position de changement harmoniques

ircam Centre Pompidou

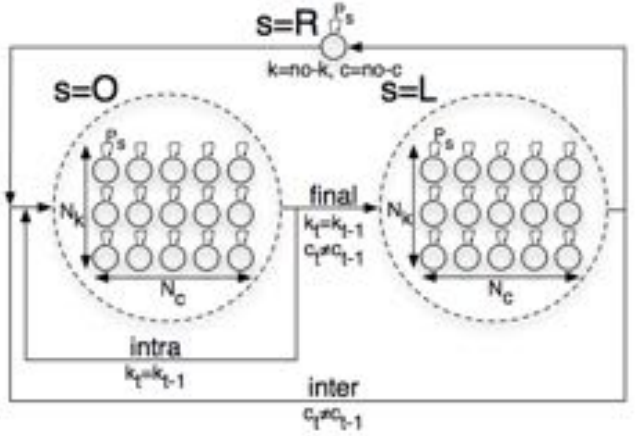
Descripteurs et modèle de connaissance

ircamstructure

Complexité (perplexity) harmoniques décroit en fin de section / structure



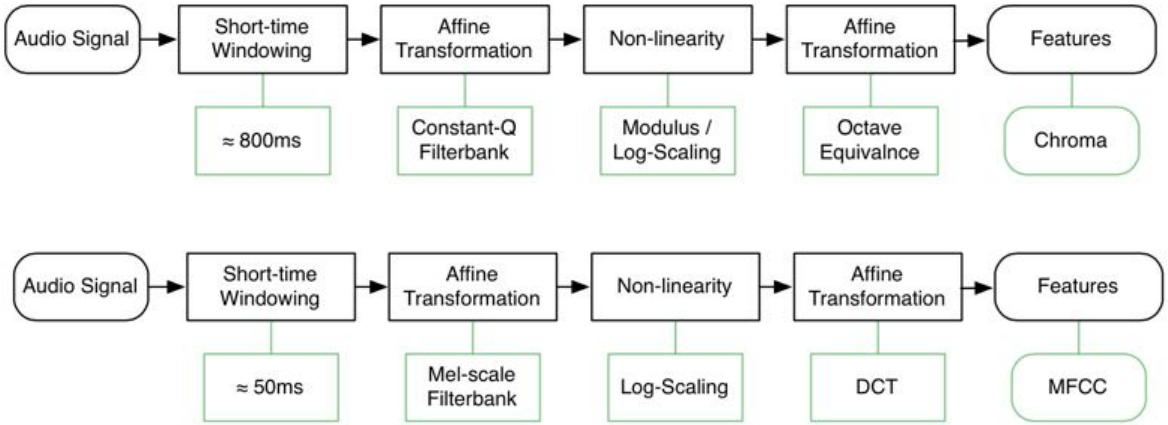
	Isophonics		Quaero	
	major	minor	major	minor
intra	6.17	6.25	4.29	4.98
inter	3.91	4.52	2.99	2.37
final	2.91	3.51	2.36	3.18



rcam
Centre Pompidou

Comment créer des descripteurs ?

- 1) A la main
  - En perception, en parole, en image, … pour mettre en évidence des propriétés spécifiques

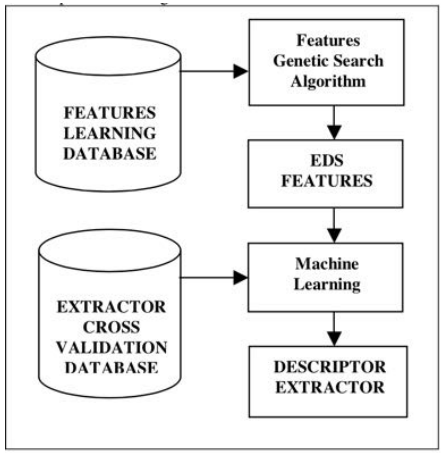


- Développer un descripteur à la main: recherche de généralisation d'un phénomène spectro-temporel observé localement à un corpus plus large
- Modélisation de l'évolution temporelle (moyenne/ vairnace à court-terme, modèle d'évolaution)
- Pour un esemble de descripteurs, lequel utiliser ?
  - Sélection manuelle ou sélection automatique de descripteurs

ircam Centre Pompidou

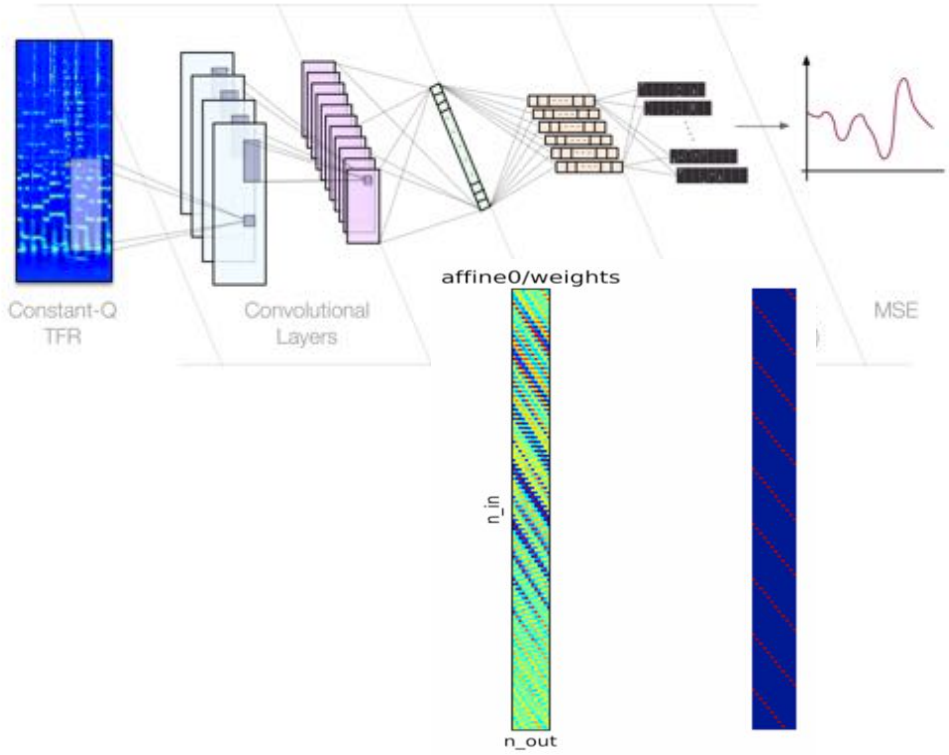
Comment créer des descripteurs ?

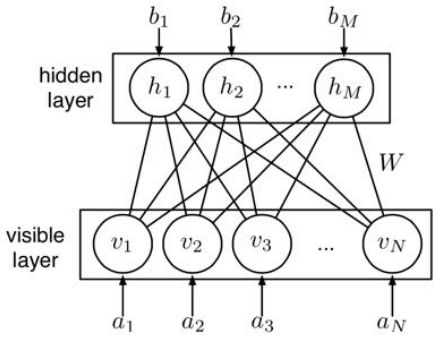
- 2) Automatiquement de manière supervisée
  - Extractor Discovery System (EDS) [Zils, Pachet]
  - Genetic Programming
  - *MEAN (MAX (FTT (SPLIT (HPFILER (Signal, 1000Hz), 10ms)))*

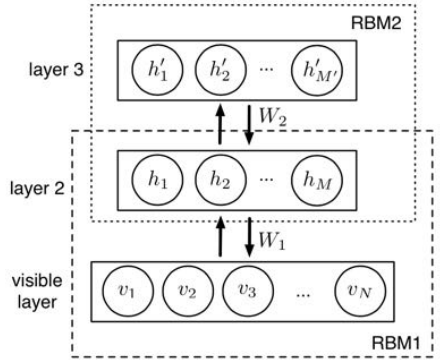


Comment créer des descripteurs ?

- <u>3) Automatiquement de manière non supervisée</u>
  - Deep Believe Network [Humphrey, Battenberg, …]
  - Layering restricted Boltzmann machines



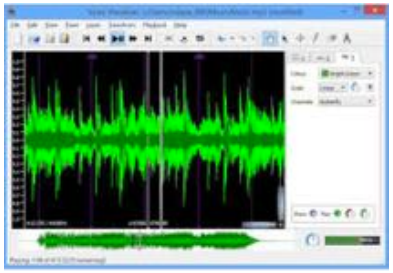




ircam Centre Pompidou

Logiciels d'extraction de descripteurs

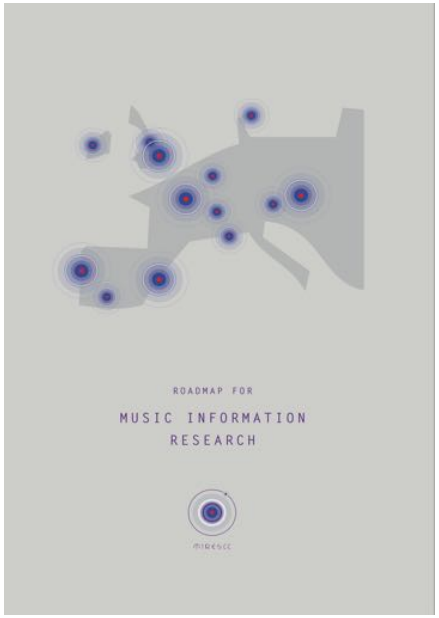
- Il en existe beaucoup
  - Marsyas
  - Yaafe
  - MIR Toolbox
  - Zsa.Descriptors
  - Vamp plug-in
  - TimbreToolBox
  - Ircamdescriptor, ircambeat, ircamkeymode, ircamchord, …
  - …
- Optimisation de l'extraction >< Flexibilité ircamdescriptor)
- Démultiplication des descripteurs ?
  - Beaucoup de variations autour d'un même thème
  - Beaucoup d'implémentations différentes (C++, Matlab, Python, JavaScript)
  - des besoins différents
- Faut-il normaliser les descripteurs ?
  - Cnfr MPEG-7 Audio
  - Standard SDIF [Burred]

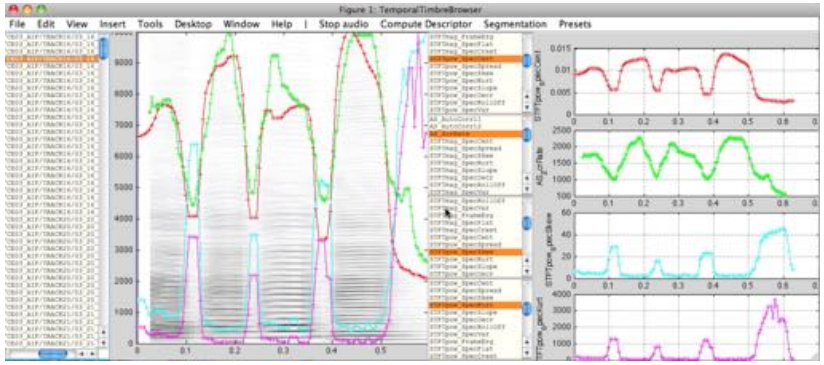


ircam Centre Pompidou

– Roadmap for Music Information Research

– Merci
– Questions ?





ircam
Centre
Pompidou