Similarity 4 Audio

Mathieu Lagrange 

ircam
Centre
Pompidou

June 4, 2013

Outline

Motivation

Let humans access audio data in a way that makes sense for them

Outline

Motivation

Let humans access audio data in a way that makes sense for them

Means

explore different means of representing sound to quantify the notion of resemblance between sounds as experienced by humans

- in musical corpora
- for environmental sounds

Outline

Motivation

Let humans access audio data in a way that makes sense for them

Means

explore different means of representing sound to quantify the notion of resemblance between sounds as experienced by humans

- in musical corpora
- for environmental sounds

Challenges

- semantic representations
- human perception processes
- mathematical representation
- computational tractability

Outline

Motivation

Let humans access audio data in a way that makes sense for them

Means

explore different means of representing sound to quantify the notion of resemblance between sounds as experienced by humans

- in musical corpora
- for environmental sounds

Challenges

- semantic representations
- human perception processes
- mathematical representation
- computational tractability

Music Information Retrieval (MIR)

As in every multimedia retrieval task, the main issue is to bridge the semantic gap.
Depending on the data at hand, the difficulty of the task ranges from impossible to hardly doable

1. raw data (signal)

Music Information Retrieval (MIR)

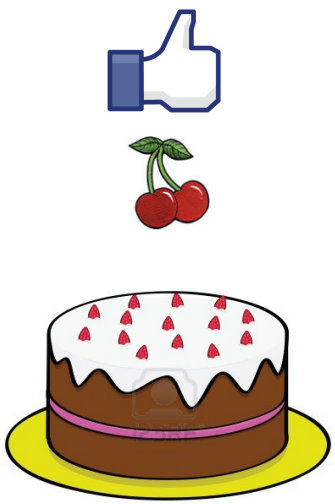
As in every multimedia retrieval task, the main issue is to bridge the semantic gap.
Depending on the data at hand, the difficulty of the task ranges from impossible to hardly doable

1. raw data (signal)
2. meta data (tags: genre)

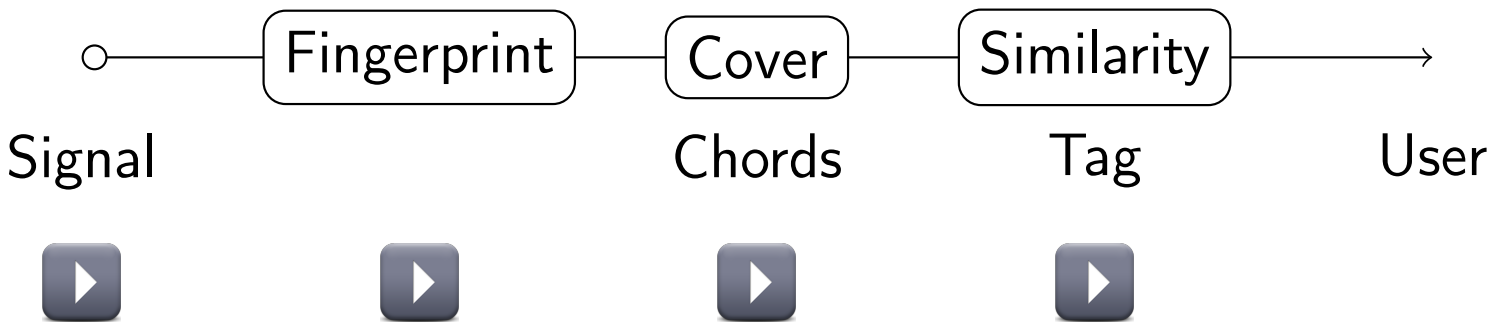
Music Information Retrieval (MIR)

As in every multimedia retrieval task, the main issue is to bridge the semantic gap.
Depending on the data at hand, the difficulty of the task ranges from impossible to hardly doable

1. raw data (signal)
2. meta data (tags: genre)
3. user ratings (likes)



Content-based Similarity in Music



Fingerprinting: the quest of the cherry

How ?

- for each item of the database, compute several fingerprints
- for a query, do the same
- match the fingerprints.

Fingerprinting: the quest of the cherry

How ?

- for each item of the database, compute several fingerprints
- for a query, do the same
- match the fingerprints.

The design of a good fingerprint is the key:

- noisy channel paradigm
- express the tolerable distortions induced by the channel to the signal
- define a compact representation [Ramona'11] that
  - is robust to those degradations,
  - preserves a good precision.

Pitfall: 

Fingerprinting: the quest of the cherry

How ?

- for each item of the database, compute several fingerprints
- for a query, do the same
- match the fingerprints.

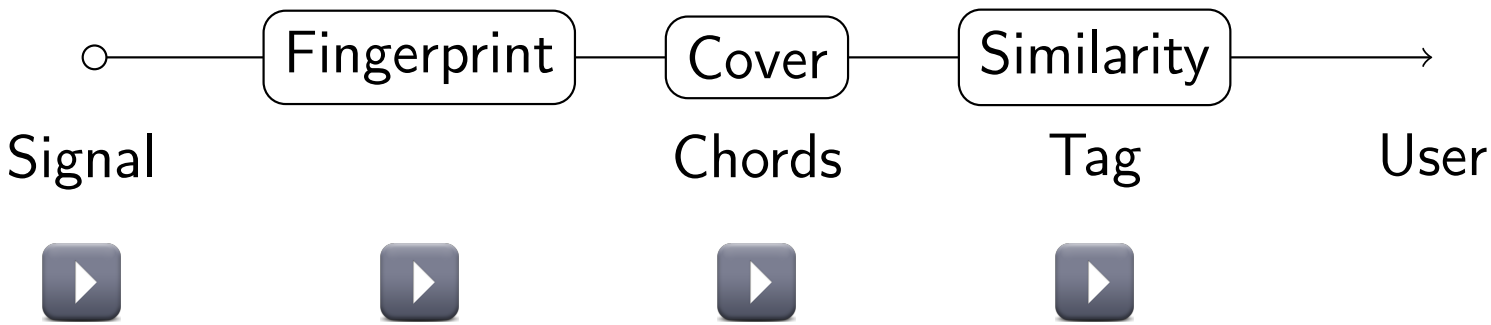
The design of a good fingerprint is the key:

- noisy channel paradigm
- express the tolerable distortions induced by the channel to the signal
- define a compact representation [Ramona'11] that
  - is robust to those degradations,
  - preserves a good precision.

Pitfall: 

- The database may not be big enough ☹

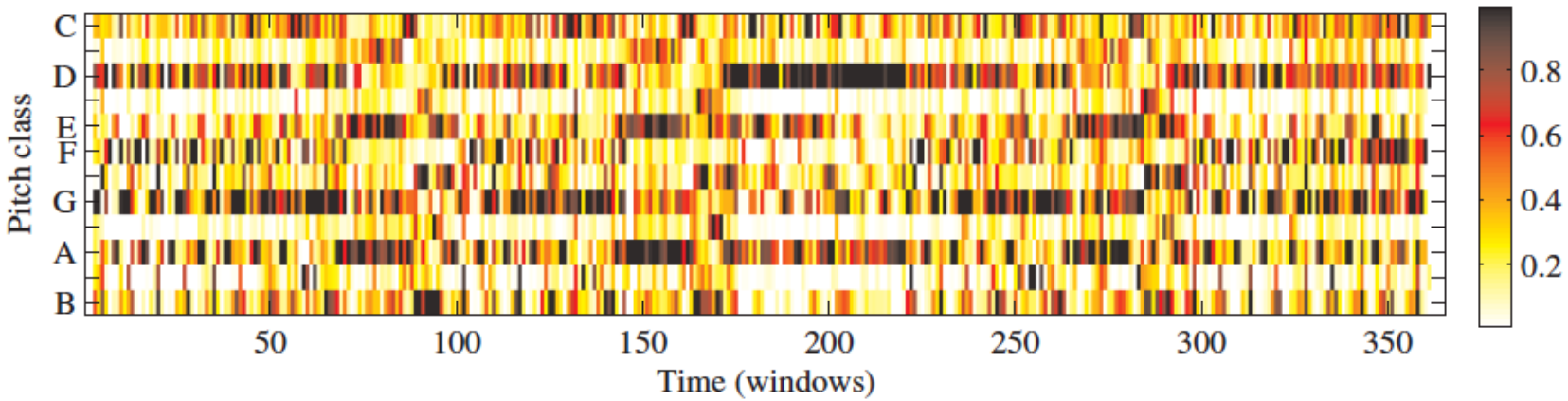
Content-based Similarity in Music



Cover detection

Principle:

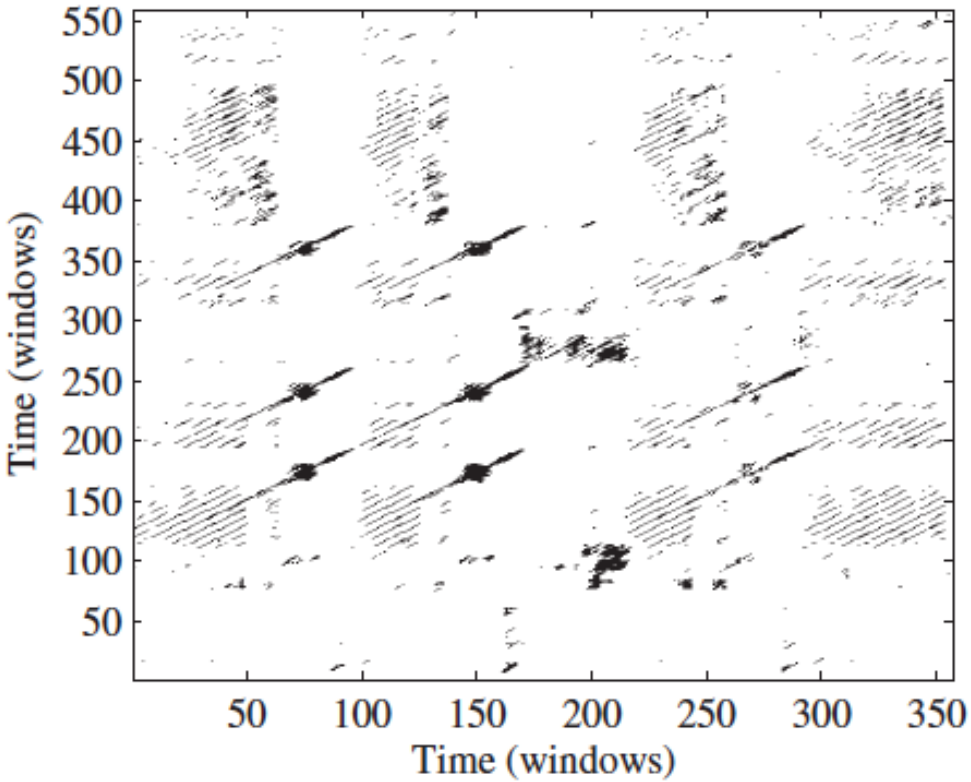
- compute chromagrams (octave-folded spectrograms)



Cover detection

Principle:

- compute chromagrams (octave-folded spectrograms)
- align sub-sequences using Dynamic Time Warping (DTW) techniques $O(n^2)$



Cover detection

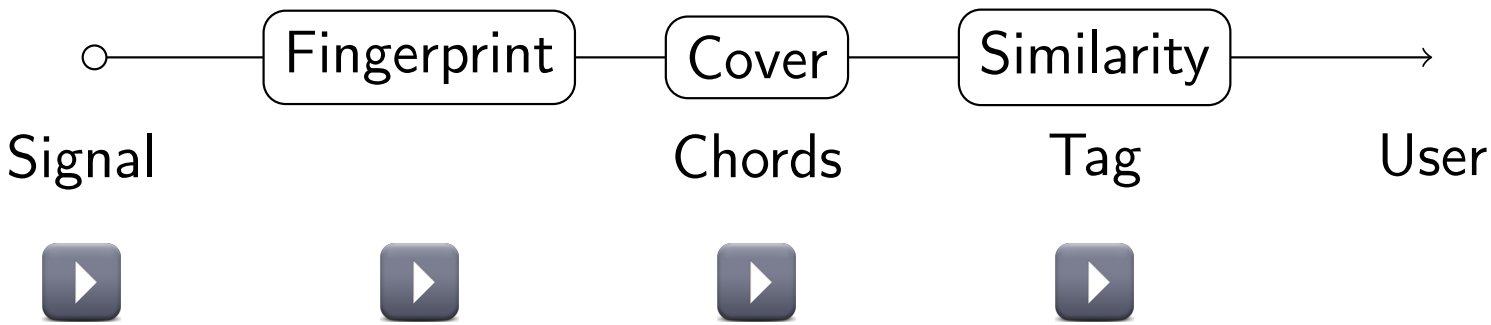
Principle:

- compute chromagrams (octave-folded spectrograms)
- align sub-sequences using Dynamic Time Warping (DTW) techniques $O(n^2)$

Challenge: Large scale

- chromas are not selective enough by themselves
- need a way to encode temporality
- hash-based system report an average rank of 308 369 on the Million Song Dataset ! [Bertin-Maheux'11]
- lost battle ?

Content-based Similarity in Music



Content-based Music Similarity

Measure: Artist-filtered Genre
How:

- compute in an unsupervised way an abstract representation: Bag of Frames (BOF)
- add supervision:
  - inclusion of auto-taggers output
  - learn the metric based on known tags

Content-based Music Similarity

Measure: Artist-filtered Genre
How:

- compute in an unsupervised way an abstract representation: Bag of Frames (BOF)
- add supervision:
  - inclusion of auto-taggers output
  - learn the metric based on known tags

Yet, it is far from reaching the use of user ratings [Slaney]. This scheme is only useful to tackle the cold start problem, *i.e.* when you do not have user ratings.

Content-based Music Similarity

Measure: Artist-filtered Genre
How:

- compute in an unsupervised way an abstract representation: Bag of Frames (BOF)
- add supervision:
  - inclusion of auto-taggers output
  - learn the metric based on known tags

Yet, it is far from reaching the use of user ratings [Slaney]. This scheme is only useful to tackle the cold start problem, *i.e.* when you do not have user ratings. Challenge: find an elegant way to fuse informations about the piece of music from very disparate channels.

Why am I uneasy with MIR ?

Music is fascinating and playing with this kind of data is kind of cool.

Why am I uneasy with MIR ?

Music is fascinating and playing with this kind of data is kind of cool. But music is

- like speech, it is special, *i.e.* engineered
- very diverse
- very complex, both intrinsically and in terms of human usage
  - a strong identity vector
  - every functional areas of the brain are involved (oro-linguistic, logico-mathematic, kinestetic, ...)

Why am I uneasy with MIR ?

Music is fascinating and playing with this kind of data is kind of cool. But music is

- like speech, it is special, *i.e.* engineered
- very diverse
- very complex, both intrinsically and in terms of human usage
  - a strong identity vector
  - every functional areas of the brain are involved (oro-linguistic, logico-mathematic, kinestetic, ...)

That it is hard to formulate any scientific problem without narrowing drastically the scope of experimentation.
On the practical side, in order to tackle reasonable scale problems, we have to resort to numerous kind of numerical approximations that limits the significance of the studies.

Environmental sounds

To better understand better how human listen to his environment (which include music), I chose to

Environmental sounds

To better understand better how human listen to his environment (which include music), I chose to

- abandon music, sigh ☹

Environmental sounds

To better understand better how human listen to his environment (which include music), I chose to

- abandon music, sigh ☹
- focus on environmental sounds, yipee ☺

The process of hearing: making sense of the input

According to Gaver [Gaver], there are 2 modes of listening

- everyday listening
- musical listening

The process of hearing: making sense of the input

According to Gaver [Gaver], there are 2 modes of listening

- everyday listening
- musical listening

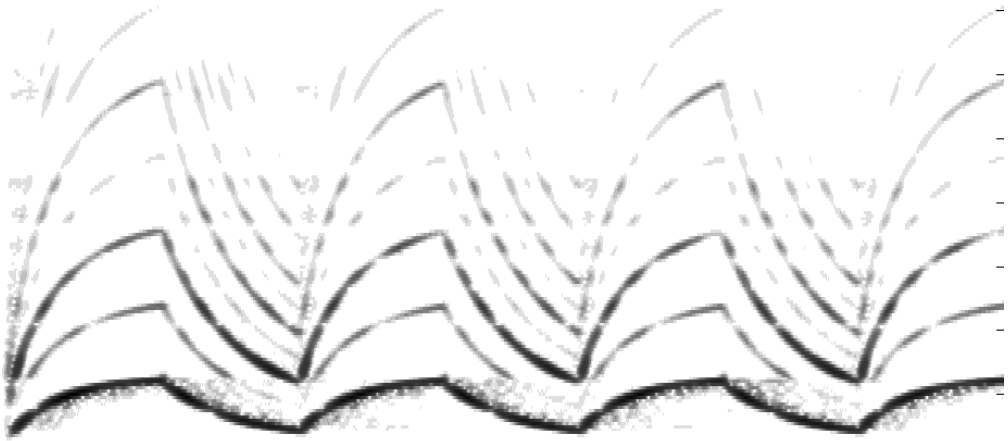
They can be reformulated as

- holistic listening: fast screening based on pattern matching (low power processes)
- analytical listening: intensive search of correlation between various cues (high power processes)

What is a good representation of sounds ?

Seek

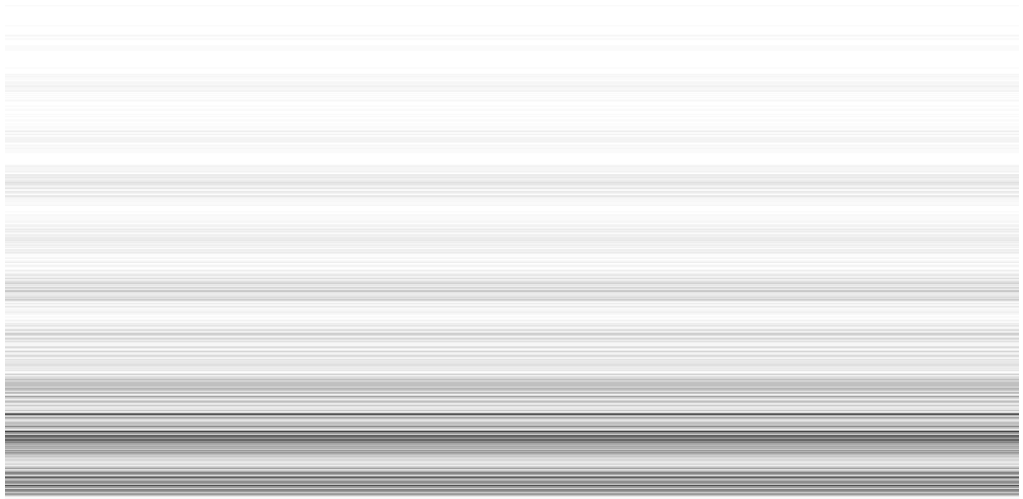
- invariance
  - in time



What is a good representation of sounds ?

Seek

- invariance
  - in time



What is a good representation of sounds ?

Seek

- invariance
  - in time
  - in frequency

What is a good representation of sounds ?

Seek

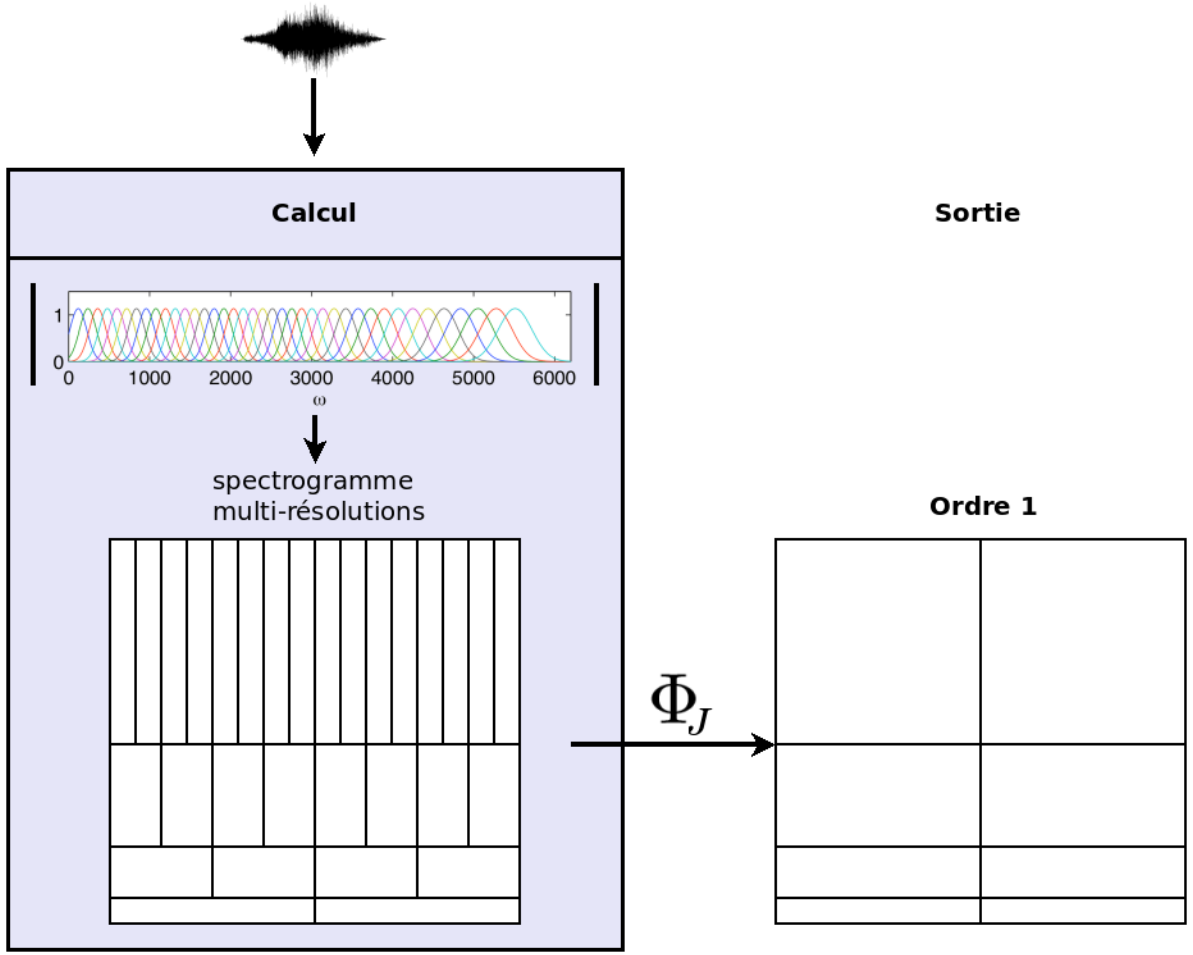
- invariance
  - in time
  - in frequency

What is a good representation of sounds ?

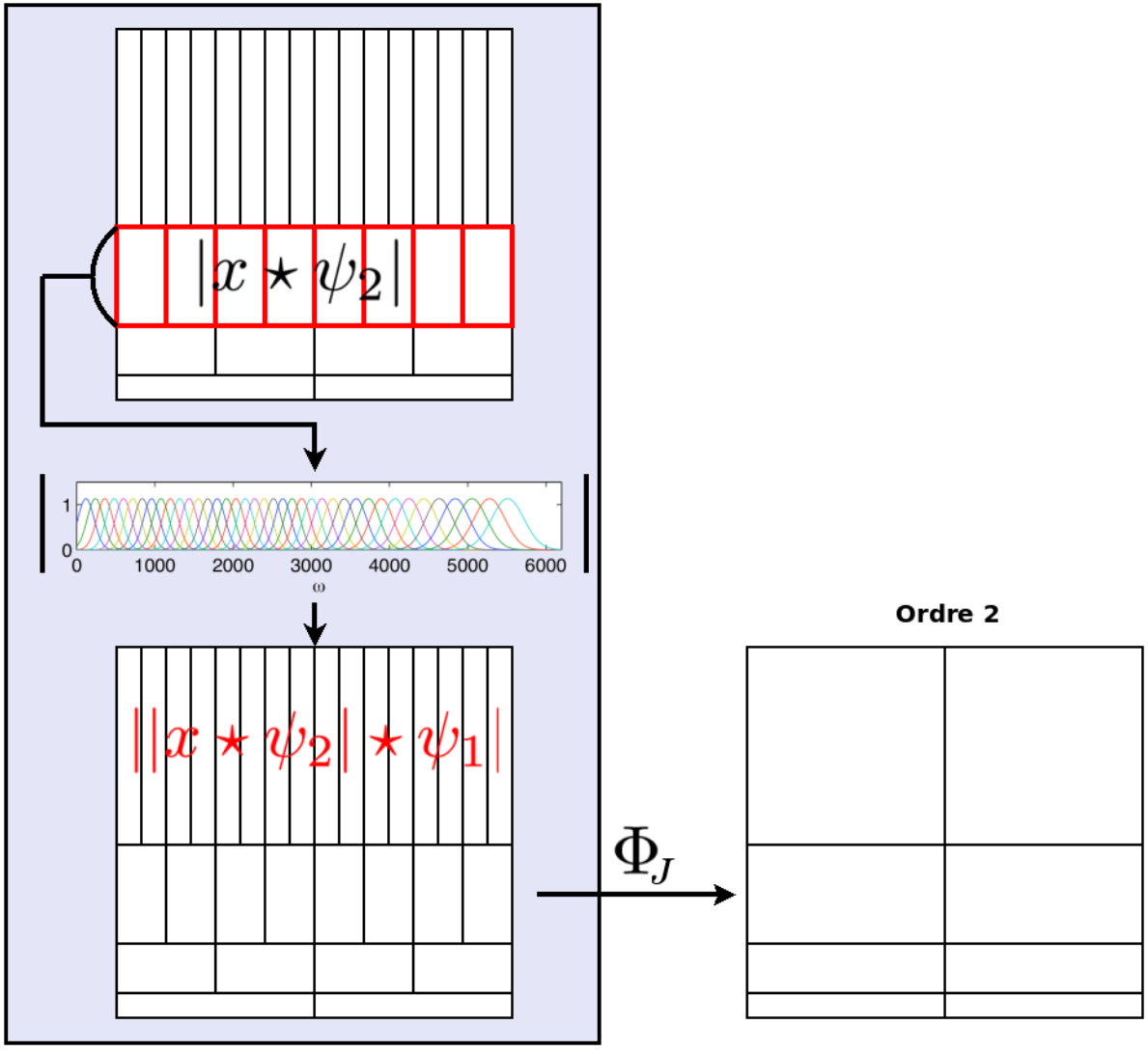
Seek

- invariance
  - in time
  - in frequency
- compacity

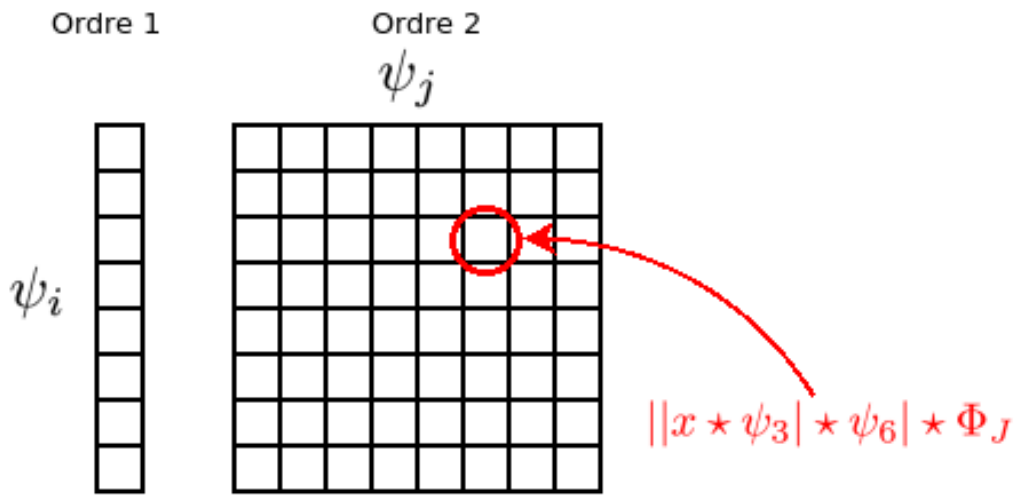
The scattering in a nutshell [Anden11]



The scattering in a nutshell [Anden11]

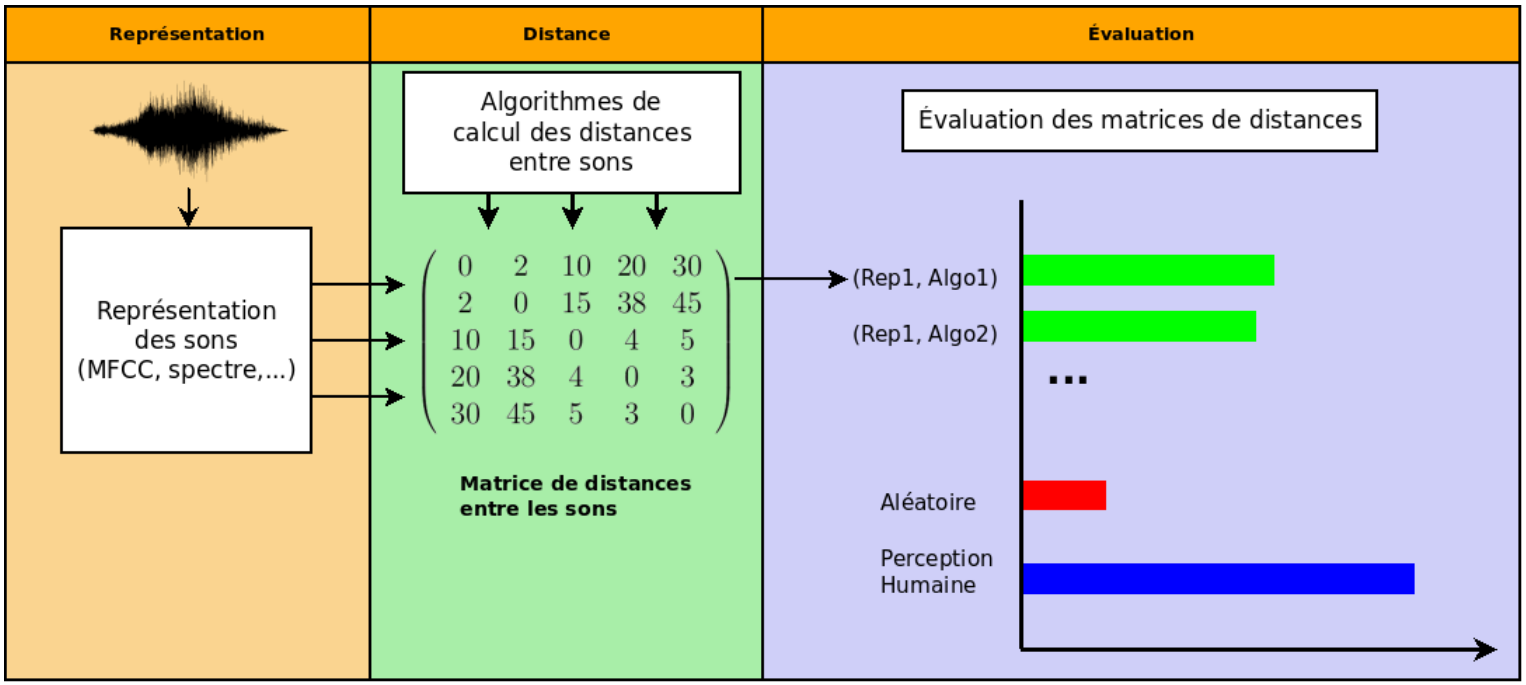


The scattering in a nutshell [Anden11]



The Cosine Log Scattering roughly consist in a DCT step over the log scattering coefficients.
Seek cheap decorrelation to achieve a good compacity (as with the MFCCs).

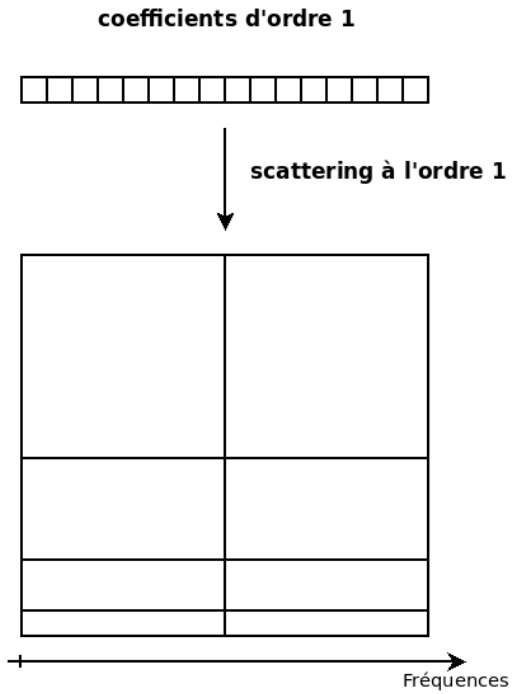
Experimental protocol



Some results

	ALEA	*BOF*	*DTW*	*CLSo1*	*CLSo2*
gygi	5.1	31.8	25.8	23.9	39.3
gygiExt	3.6	20.9	19.3	19.4	28.4
houix1	43.6	54.6	55.5	54.8	53.4
iowa	8.4	29.8	32.0	47.0	50.4
rwc	8.9	30.0	30.2	38.6	44.8

Do it again : the scattering combined

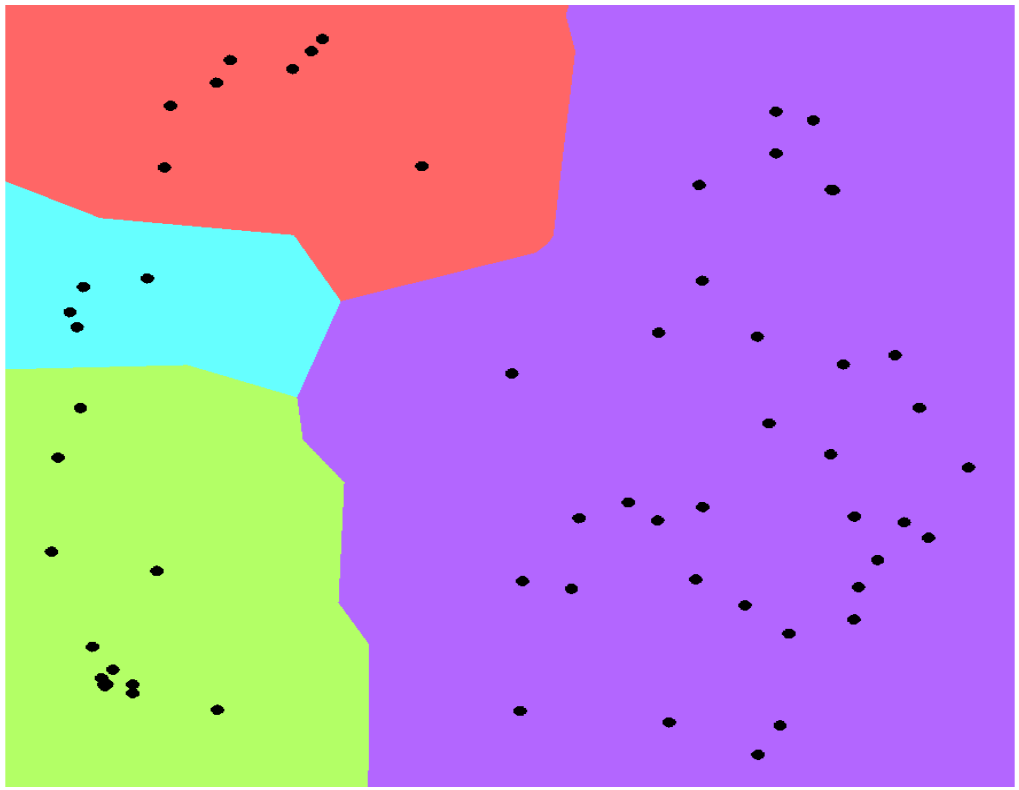


Replace the linearly spaced bins of the DCT by some logarithmic ones to achieve frequency axis invariance at the higher order scattering levels.

More results

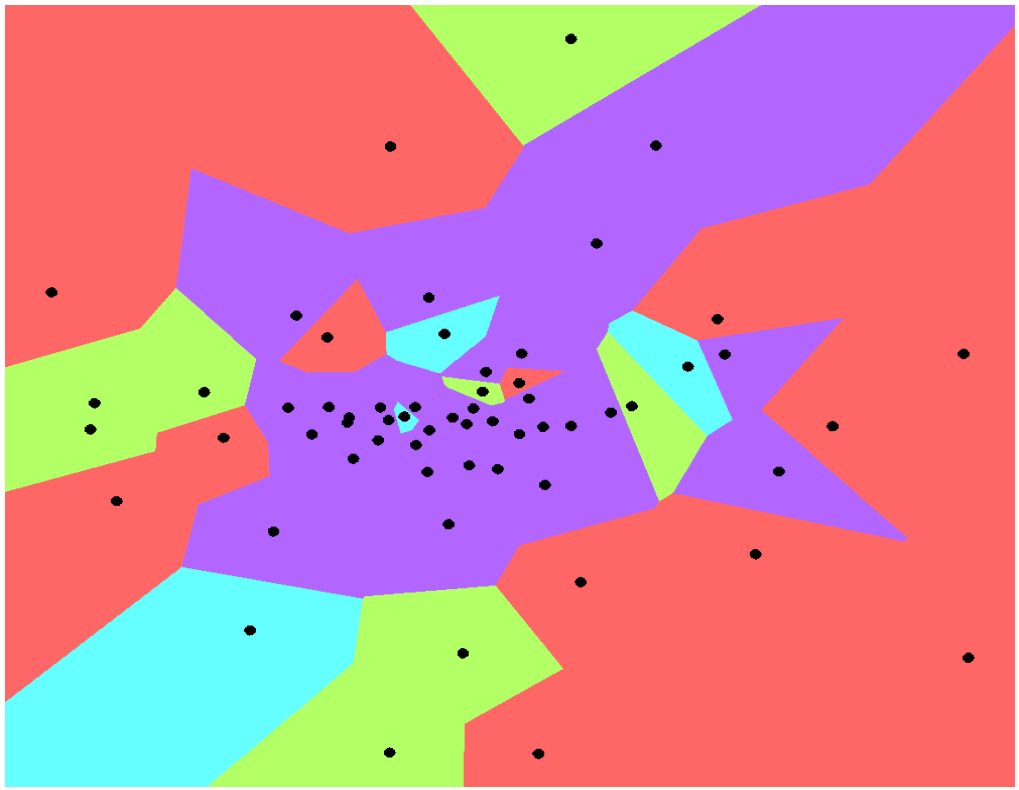
	ALEA	*BOF*	*DTW*	*CLSo*1	*CLSo*2	*COo*1	*COo*2
gygi	5.1	31.8	25.8	23.9	39.3	30.0	44.4
gygiExt	3.6	20.9	19.3	19.4	28.4	20.9	38.9
*houix*1	43.6	54.6	55.5	54.8	53.4	52.0	59.0
iowa	8.4	29.8	32.0	47.0	50.4	35.7	39.9
rwc	8.9	30.0	30.2	38.6	44.8	40.5	39.5

MDS visualization on the Houix1 Database



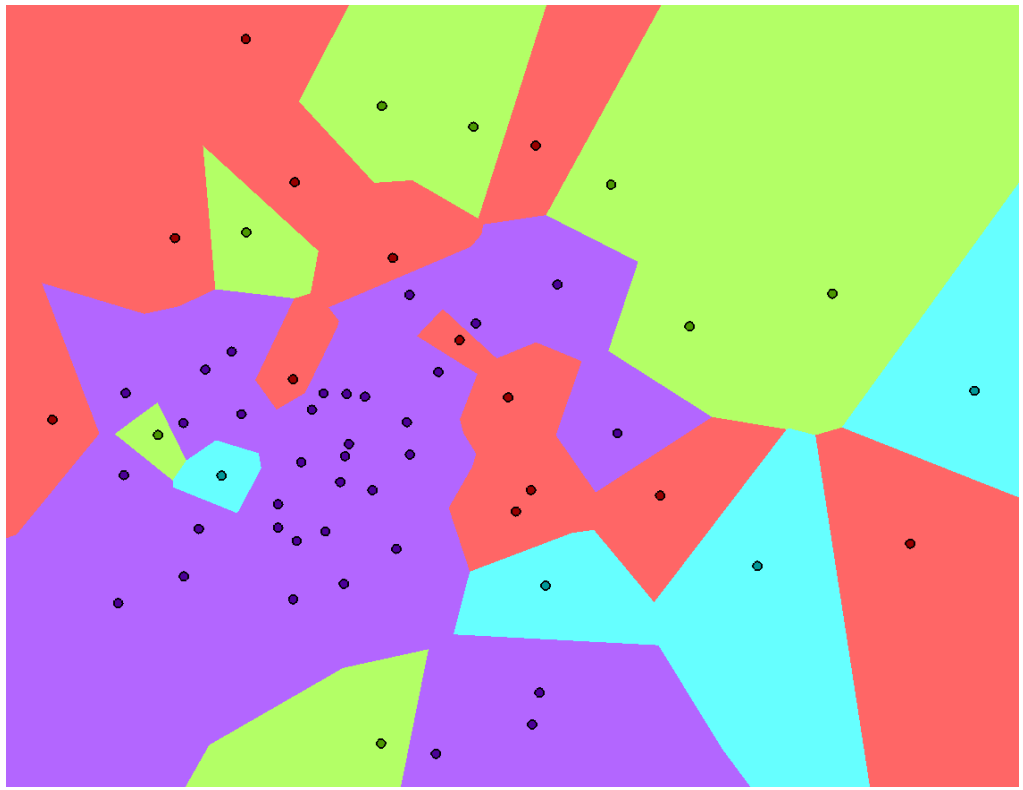
human (MAP=94%)

MDS visualization on the Houix1 Database



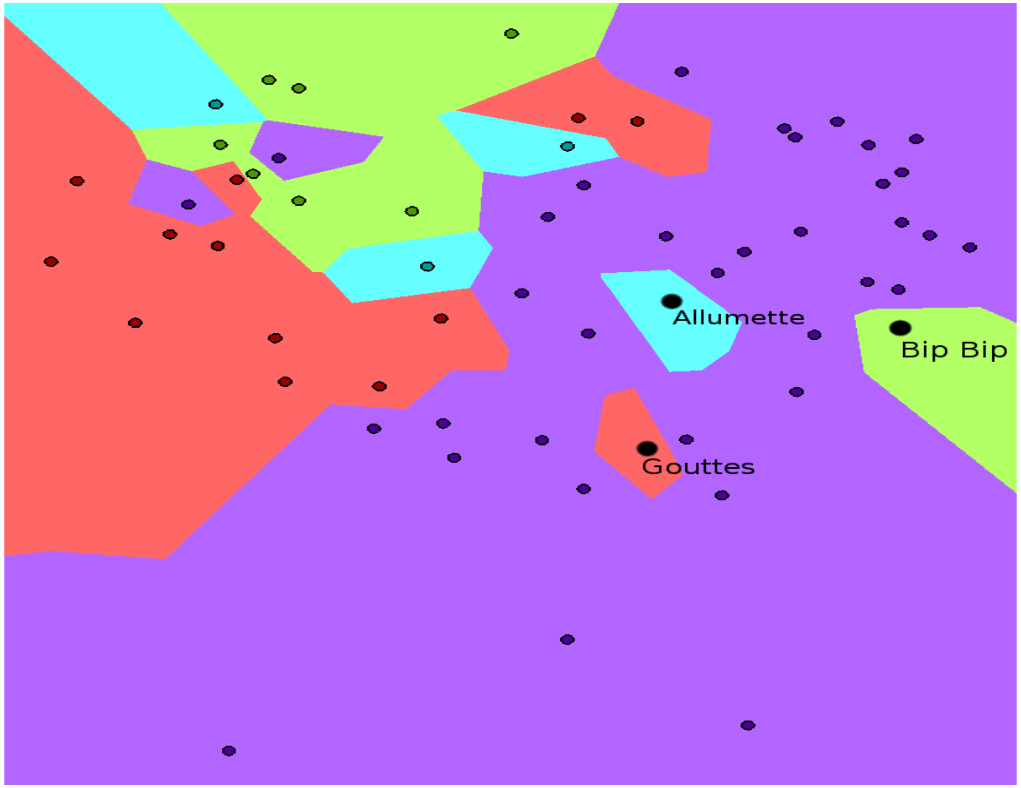
DTW (MAP=55.4%)

MDS visualization on the Houix1 Database



CLS order 2 (MAP=56.7%)

MDS visualization on the Houix1 Database



combined order 2 (MAP=59%)

MDS visualization on the Houix1 Database

Matlab Gui code available at: `http://recherche.ircam.fr/equipes/analyse-synthese/lagrange/research/vizuMds`.

Conclusion I: Music vs. Environmental Sounds

Studying music has some flaws:

- hard to get open data
- the scale is huge
- mainly engineering solutions

Studying environmental sounds has a lot of advantages

- easier to collect and share open data
- likely to involve simpler cortical processes
- great application potential

Conclusion II: Similarity vs. Classification

Classification: task oriented

- compact representation of data (tags)
- ease of manipulation
- loose a lot of informations
- specify only one level of granularity

Conclusion II: Similarity vs. Classification

Classification: task oriented

- compact representation of data (tags)
- ease of manipulation
- loose a lot of informations
- specify only one level of granularity

Similarity: view oriented

- rich representation
- less direct usage and interpretation (work to be done ?)

Conclusion II: Similarity vs. Classification

Classification: task oriented

- compact representation of data (tags)
- ease of manipulation
- loose a lot of informations
- specify only one level of granularity

Similarity: view oriented

- rich representation
- less direct usage and interpretation (work to be done ?)

Practical benefits of using similarity based tasks:

- lighter experiments
- remove any dependencies to the nature of the chosen classifier.

People

Carlo Baugé

People

Carlo Baugé

Mathias Rossignol

People

Carlo Baugé

Joakim Anden

Mathias Rossignol

People

Carlo Baugé

Joakim Anden

Mathias Rossignol

Stéphane Mallat

Food 4 thoughts

Kamini

Food 4 thoughts

Kamini

Child crowd in Africa

Thank you !!

Announcements

2nd CASA workshop took place last Saturday in Porto (vids soon available) !

- Cognition and Neurosciences (Emmanuel Bigand, Shihab Shamma)
- Mathematics (Joakim Andén)
- Computer Science (Tom Walters, George Tzanetakis)

IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events

- web based environmental scene synthesizer based on Freesound
- Audio input needed for backgrounds and events, please contribute !
  `http://soundthings.org/challenge-contrib`