Similarity 4 Audio

Mathieu Lagrange

January 12, 2015

Me

CNRS researcher

- Computational Auditory Scene Analysis (CASA),
- Machine listening,
- Audio processing from signal processing theory to implementation

Some history...

2001-2004	France Telecom R&D Rennes
2004-2006	LaBRI (University Bordeaux 1)
2006-2007	University of Victoria, BC, Canada
2007-2008	McGill University, QC, Canada
2008- 2009	Telecom ParisTech
2009- 2013	IRCAM
2013- –	**ADTSI team of IRCCYN**

Rationale

"Drowning in Data yet Starving for Knowledge"

John Naisbitt (1982)

Data

Numerical data is :

- blind
- huge
- important
- needs care
- needs to be accessed
- just a material

Sound FX

- sound ideas: ¿ 450 000 files
- sounddogs: ¿ 680 000 files

Music

- Google play: database size ¿ 22 000 000
- Spotify: database size ¿ 25 000 000
  additions per day ¿ 20 000
- Deezer: database size ¿ 20 000 000
- iTunes store: database size ¿ 37 000 000
  downloads per minute: database size ¿ 15 000

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)

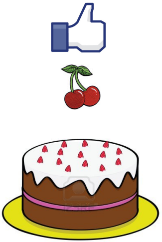
As in every multimedia retrieval task, the main issue is to bridge the semantic gap.
Depending on the data at hand, the difficulty of the task ranges from impossible to hardly doable

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

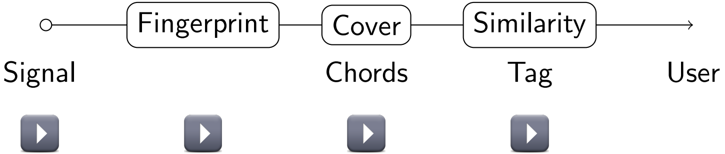
Music Information Retrieval (MIR)

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.

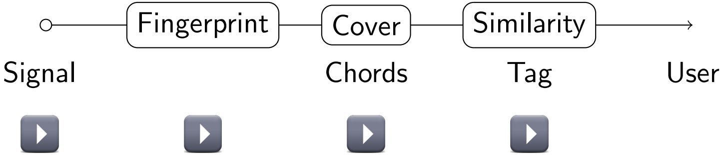
The design of a good fingerprint is the key:

- 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 ☹

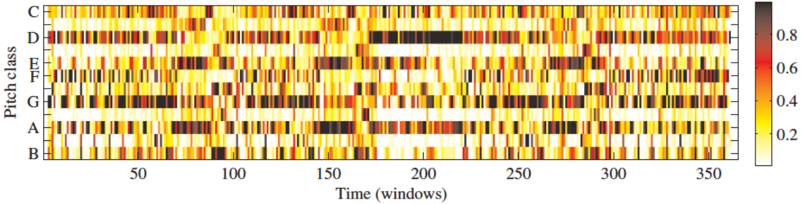
Content-based Similarity in Music



Cover detection

Principle:

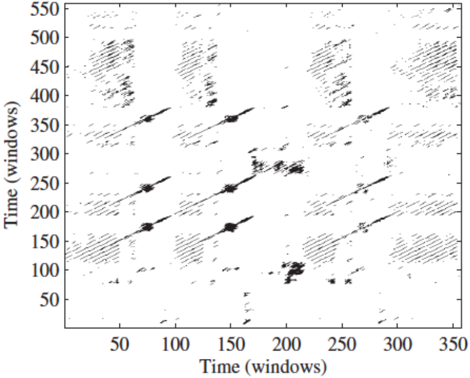
- compute chromagrams (octave-folded spectrograms)



Cover detection

Principle:

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



Cover detection

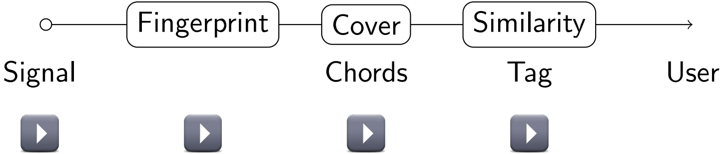
Principle:

- 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.

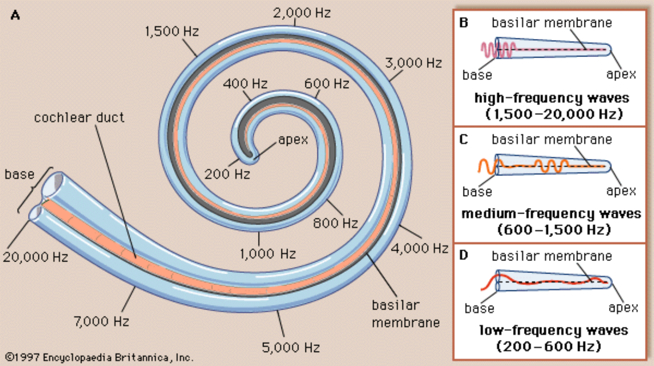
Content-based Music Similarity

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.

The process of hearing: making sense of the input

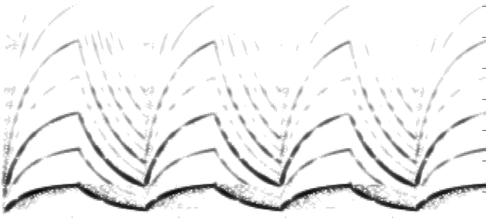


The process of hearing: making sense of the input

What is a good representation of sounds ?

Seek

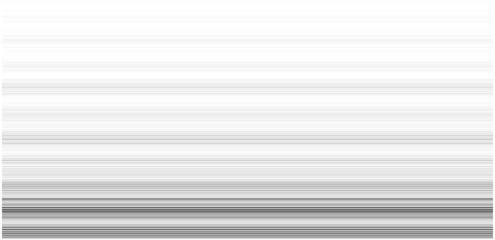
- invariance
  - in time



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

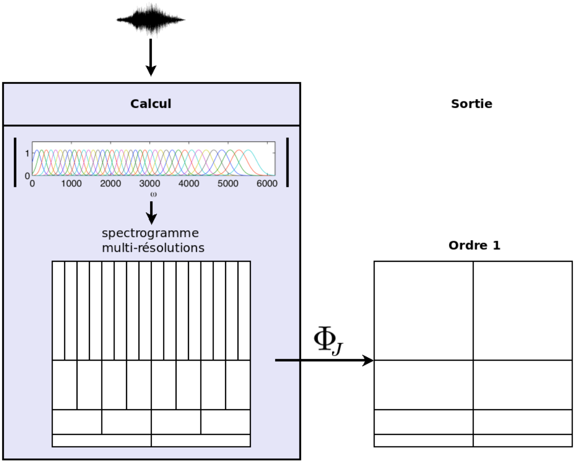
- invariance
  - in time
  - in frequency

What is a good representation of sounds ?

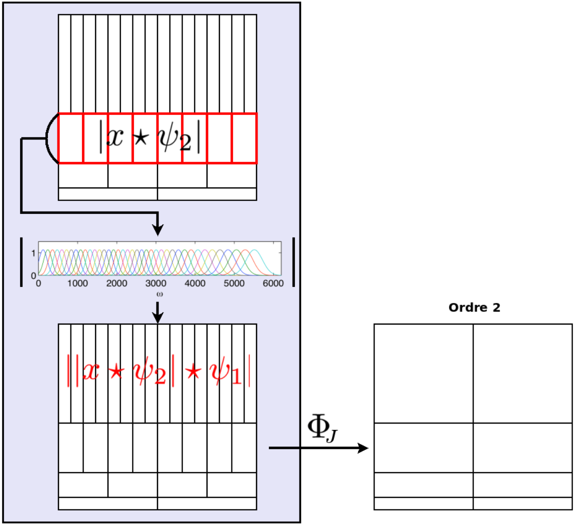
Seek

- invariance
  - in time
  - in frequency
- compacity

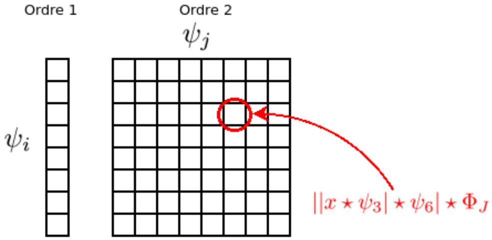
The scattering in a nutshell [Anden11]



The scattering in a nutshell [Anden11]

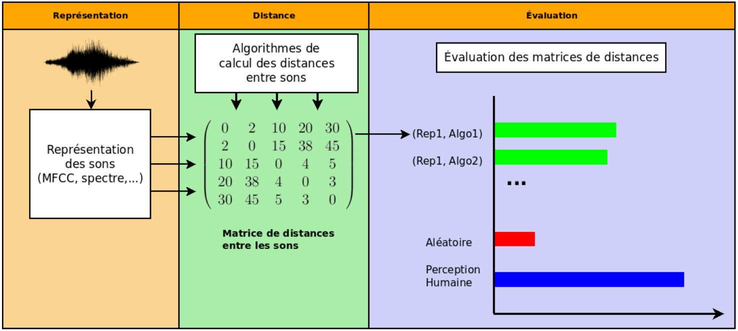


The scattering in a nutshell [Anden11]



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

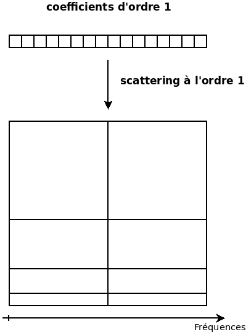
Experimental protocol



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

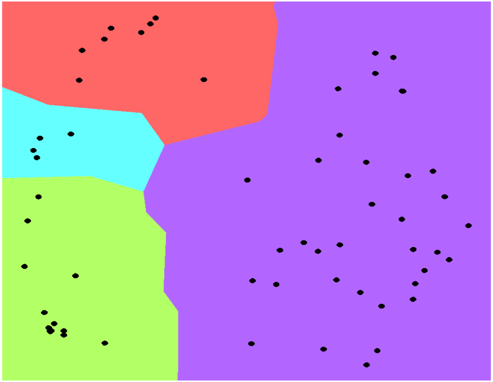


Replace the linearly spaced bins of the DCT by some logarithmic ones to achieve frequency axis invariance at the higher order scattering levels.

More results

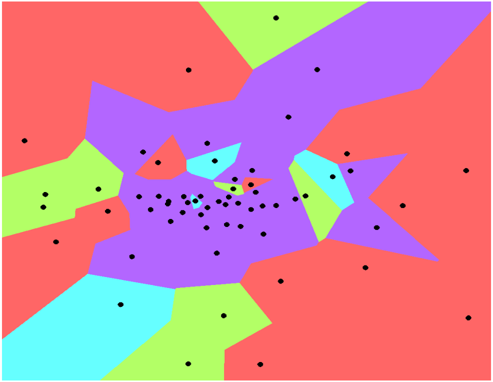
	ALEA	*BOF*	*DTW*	*CLSo1*	*CLSo2*	*COo1*	*COo2*
gygi	5.1	31.8	25.8	23.9	39.3	30.0	44.4
gygiExt	3.6	20.9	19.3	19.4	28.4	20.9	38.9
houix1	43.6	54.6	55.5	54.8	53.4	52.0	59.0
iowa	8.4	29.8	32.0	47.0	50.4	35.7	39.9
rwc	8.9	30.0	30.2	38.6	44.8	40.5	39.5

MDS visualization on the Houix1 Database



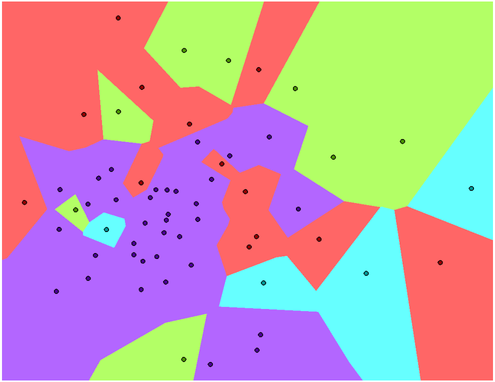
human (MAP=94%)

MDS visualization on the Houix1 Database



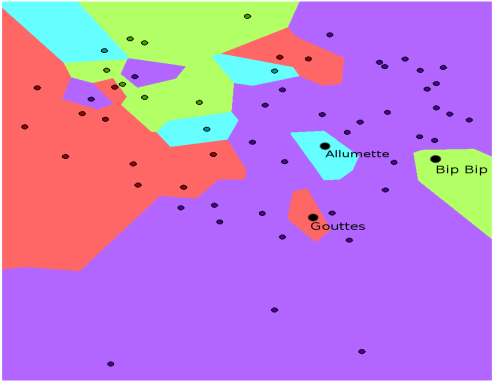
DTW (MAP=55.4%)

MDS visualization on the Houix1 Database



CLS order 2 (MAP=56.7%)

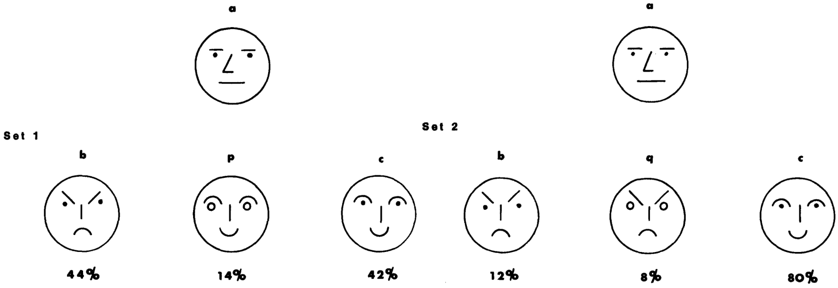
MDS visualization on the Houix1 Database



combined order 2 (MAP=59%)

Similarity: a matter of context

[Tversky 1977]



Real Sound: Context

Question the metric and dimensional assumptions that underlie the geometric representation of similarity

Real Sound: Context

Question the metric and dimensional assumptions that underlie the geometric representation of similarity

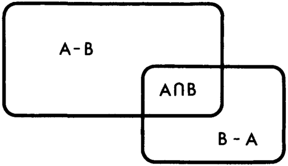
- $d(a, b) \geq d(a, a) = 0$ (identity, minimality)
- $d(a, b) = d(b, a)$ (symmetry)
- $d(a, b) + d(b, c) \geq= d(a, c)$ (triangle inequality)

Real Sound: Context

- **The set-theoretical approach to similarity: the contrast model** [Tversky 1977]

Real Sound: Context

- **The set-theoretical approach to similarity: the contrast model** [Tversky 1977]

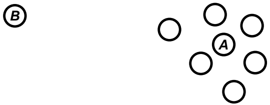


$$s(a, b) = F(A \cap B, A - B, B - A)$$

Real Sound: Context

- **The Interrelationship Between Similarity and Spatial Density** [Krumhansl 1978]

"A geometric approach may be compatible with these effects if the traditional multidimensional scaling model is augmented by the assumption that spatial density in the configuration has an effect on the similarity measure"



- $d'(a, b) = d(a, b) + \alpha\delta(a) + \beta\delta(b)$
- $\delta(a) = g[s(a, a)]$
  g is a non-negative monotonic decreasing function

Real Sound: Context

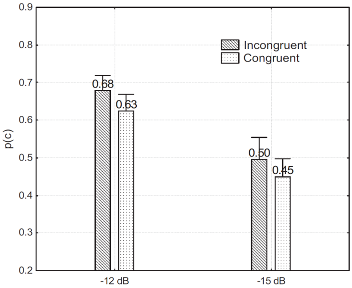
Congruency Advantage [Gygi and Shafiro 2011]

Identification: significant advantage for sounds that are contextually incongruous with the background scene (e.g., a rooster crowing in a hospital)

Real Sound: Context

Congruency Advantage [Gygi and Shafiro 2011]

Identification: significant advantage for sounds that are contextually incongruous with the background scene (e.g., a rooster crowing in a hospital)

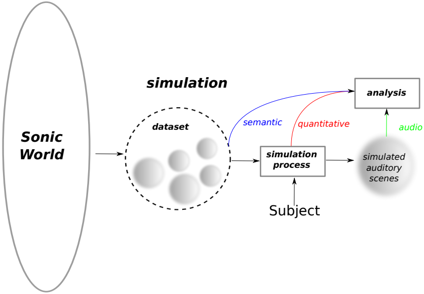


repeated 2x2 ANOVA

- So/Sc effect: $F(1, 11) = 96.04$ $p < .00001$
- Congruence: $F(1, 11) = 4.84$ $p < .05$

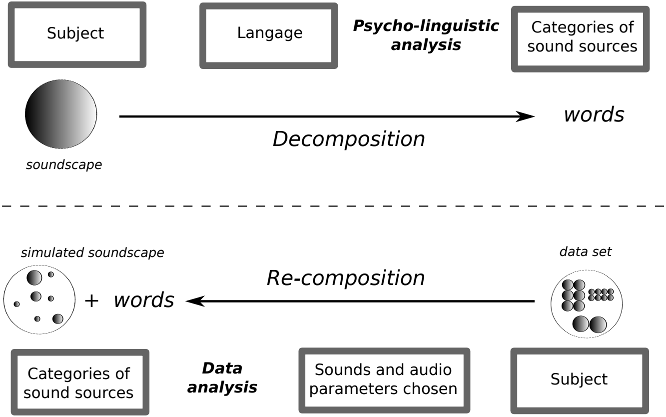
Simulation: Environmental Auditory Scenes

Does human qualitative evaluation of sounds rely on semantic attributes or quantitative properties like sound levels and sound activity ?



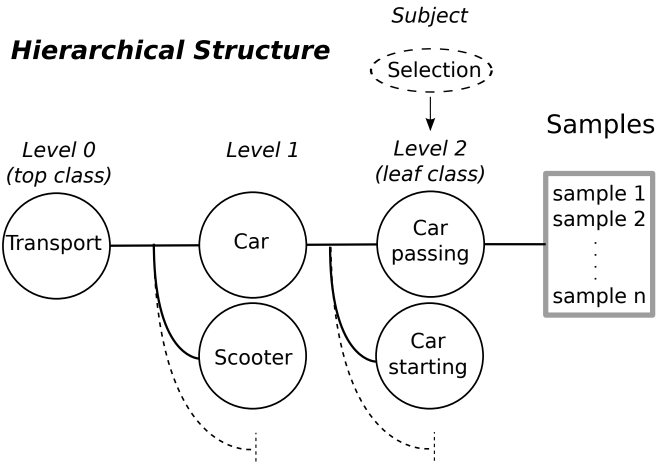
Paradigm (Cognitive Psychology)

Simulation vs. describing task



Corpus Generation

Based on sound categories

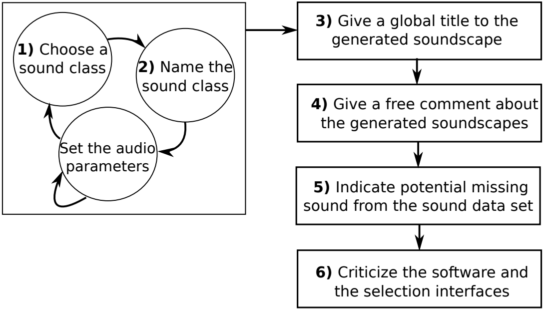


Protocol

Simulation of two urban auditory scenes: one ideal the other not ideal (40 subjects: 80 simulated scenes)

Step 1:
Iterative process for generating a soundscape

Step 2:
Verbal Description



Results: Quantitative attributes

	Event classes	Texture classes
i-scenes	-6.8 (5.4)	-2.6 (3.9)
ni-scenes	-2.4 (3.2)	-1.6 (2.6)

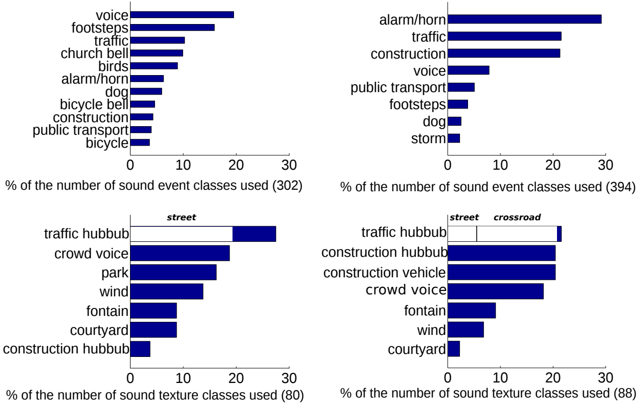
Sound levels: mean sound levels in dB averaged over the subjects ($p < 0.0001$). The deviation between the texture sound levels is not significant($p = 0.14$).

Results: Quantitative attributes

	Density of the sound events
i-scenes	53 (65)
ni-scenes	63 (64)

Density of the sound events: mean number of sound events of each scene averaged over the subjects ($p = 0.14$).

Results: Semantic attributes



left: i-scenes, *right*: ni-scenes
top: events, *bottom*: texture

Results: Semantic attributes

Does human qualitative evaluation of sounds rely on semantic attributes ?

Results: Semantic attributes

Does human qualitative evaluation of sounds rely on semantic attributes ?

- Each simulated scene is represented by a boolean vector of n dimensions $S_i = (x_1, x_2, \dots, x_n)$, $i \in [1, 80]$. Each dimension corresponds to a sound class (event and texture) of a particular semantic level

Results: Semantic attributes

Does human qualitative evaluation of sounds rely on semantic attributes ?

- Each simulated scene is represented by a boolean vector of n dimensions $S_i = (x_1, x_2, \ldots, x_n)$, $i \in [1, 80]$. Each dimension corresponds to a sound class (event and texture) of a particular semantic level
- Precision at rank 5: the average number of items of the same class among the 5 closest items to a given seed item

Results: Semantic attributes

Does human qualitative evaluation of sounds rely on semantic attributes ?

- Each simulated scene is represented by a boolean vector of n dimensions $S_i = (x_1, x_2, \ldots, x_n)$, $i \in [1, 80]$. Each dimension corresponds to a sound class (event and texture) of a particular semantic level
- Precision at rank 5: the average number of items of the same class among the 5 closest items to a given seed item
- A *Jaccard* distance is then computed between the vectors S_i

Results: Semantic attributes

Does human qualitative evaluation of sounds rely on semantic attributes ?

- Each simulated scene is represented by a boolean vector of n dimensions $S_i = (x_1, x_2, \ldots, x_n)$, $i \in [1, 80]$. Each dimension corresponds to a sound class (event and texture) of a particular semantic level
- Precision at rank 5: the average number of items of the same class among the 5 closest items to a given seed item
- A *Jaccard* distance is then computed between the vectors S_i
- mds visualization

Results: Semantic attributes

Does human qualitative evaluation of sounds rely on semantic attributes ?

Semantic level	Event and texture	Event	Texture
0	81 %	76 %	70 %
1	90 %	91 %	78 %
2	92 %	89 %	80 %
3	93 %	91 %	—

Precision at rank 5 ($P@5$) computed from the *Jaccard* distances between the scenes for different semantic levels

Sound Markers

Is there an event class which has been mostly used in one type of sound environment ?

Sound Markers

Is there an event class which has been mostly used in one type of sound environment ?

V-test at 0.001% significance level (Bonferroni Correction)

Sound Markers

Is there an event class which has been mostly used in one type of sound environment ?

V-test at 0.001% significance level (Bonferroni Correction)

Semantic level	Markers	
	i-scenes	ni-scenes
0		construction work
1	church bell	klaxon
	bicycle bell	siren
	animal	vehicle work
	footsteps	
2	church bell	klaxon
	birds	siren
	bicycle bell	vehicle work
	female laugh	
	male laugh	
3	church bell	klaxon
	birds singing	siren
	bicycle bell	vehicle work
	female laugh	
	male footsteps concrete	

Event classes are ordered using descending order of V-test values

Thank you !!

People

Carlo Baugé

Joakim Anden

Mathias Rossignol

Grégoire Lafay