Master M2 - DataScience

Audio and music information retrieval

Lecture on

Machine Listening, Music recognition, Decomposition models

Gaël RICHARD

Télécom Paris

February 2021

Content

- **Introduction**
  - What is Machine listening / audio recognition ?
  - Some applications

- **Machine listening: DCASE**

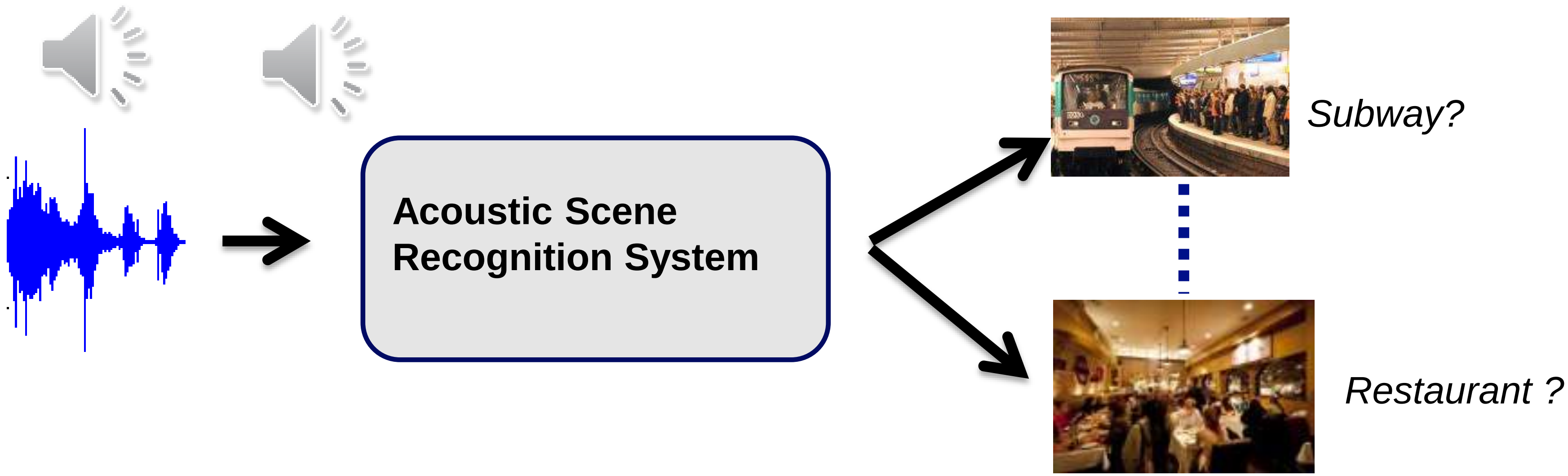
- **Signal decomposition models**
  - Sinusoidal models
  - Decomposition models (matching pursuit, NMF)
  - Exploitation of such models in scene analysis

- **Audiofingerprint or Music recognition**

Acoustic scene and sound event recognition

Acoustic scene recognition:

- « associating a semantic label to an audio stream that identifies the environment in which it has been produced »

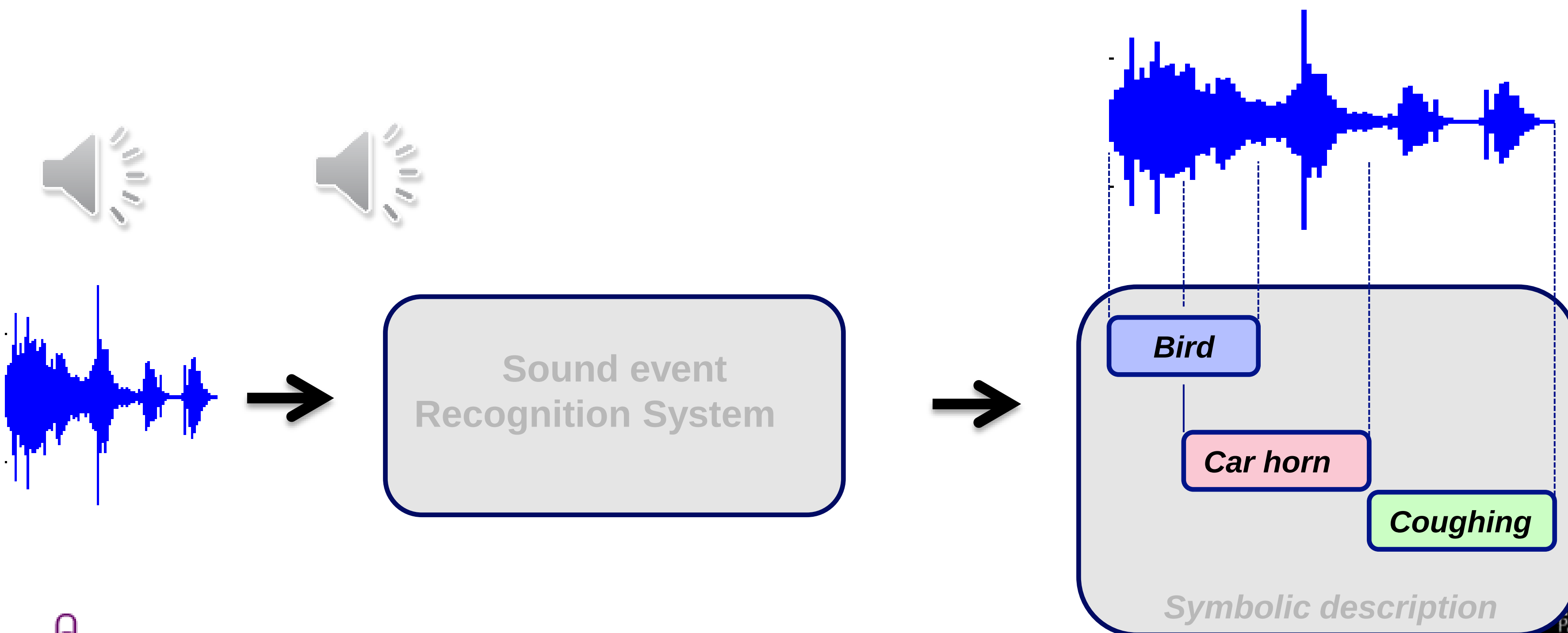


- Related to CASA (*Computational* Auditory Scene Recognition) and SoundScape cognition (*psychoacoustics*)

D. Barchiesi, D. Giannoulis, D. Stowell and M. Plumbley, « Acoustic Scene Classification », IEEE Signal Processing Magazine [16], May 2015

Acoustic scene and sound event recognition

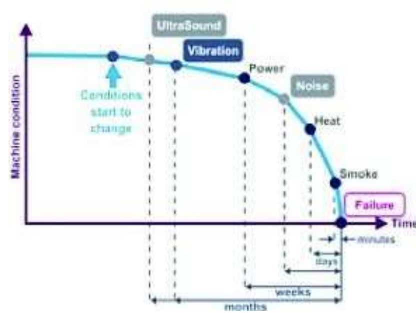
- **Sound event recognition**
  - “aims at transcribing an audio signal into a symbolic description of the corresponding sound events present in an auditory scene”.

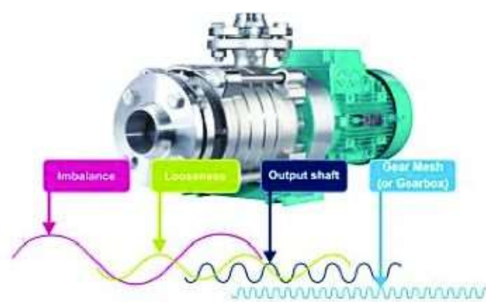


Applications of scene and events recognition

- **Smart hearing aids (Context recognition for adaptive hearing-aids, Robot audition,..)**
- **Security**
- **indexing,**
- **sound retrieval,**
- **predictive maintenance,**
- **bioacoustics,**
- **environment robust speech recognition,**
- **ederly assistance, smart homes**
- **…..**

The Rowe Wildlife Acoustic lab





From ST Microelectronics

Classification systems

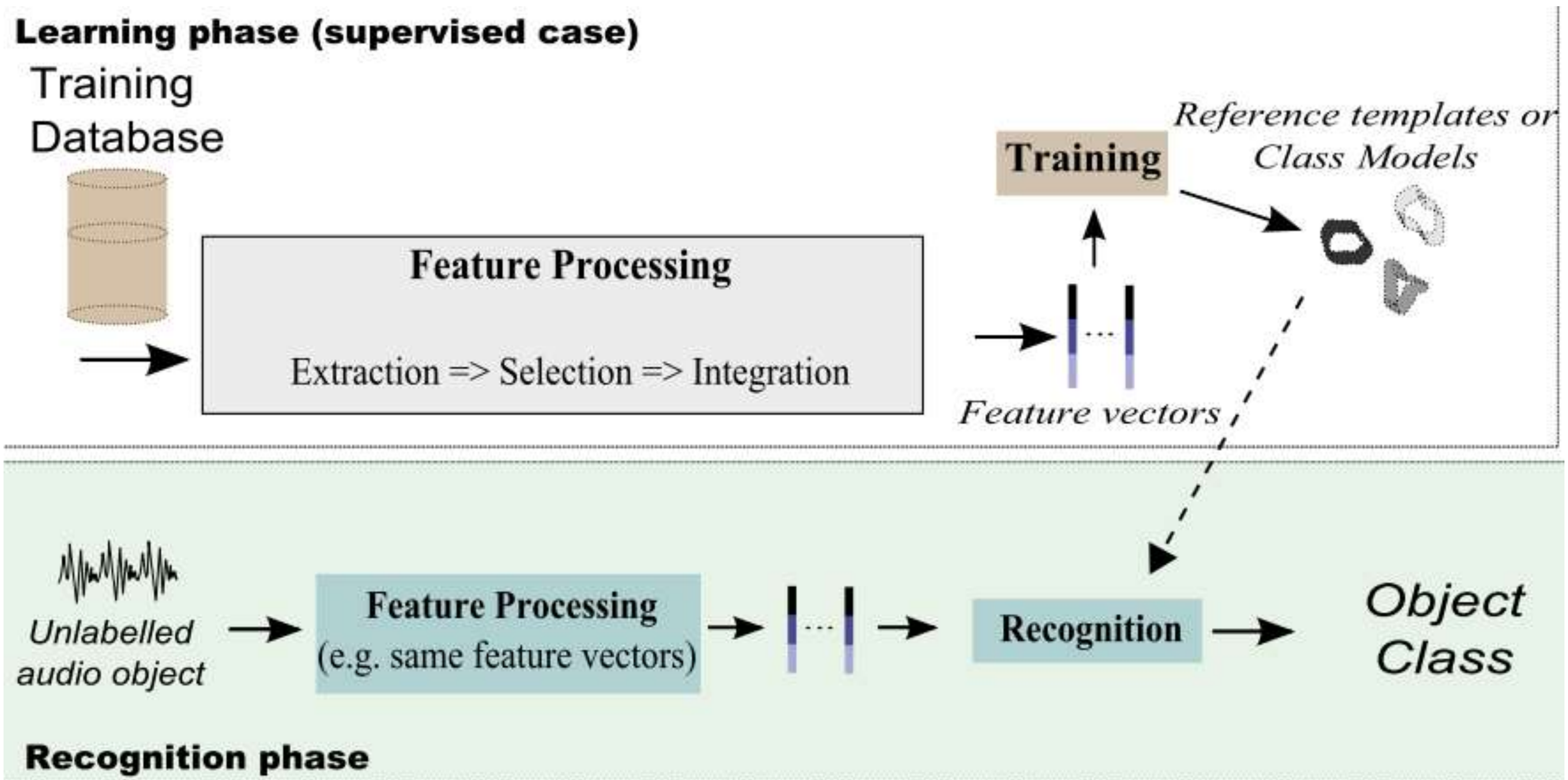
- **Several problems, a similar approach**

  - Speaker identification/recognition
  - Automatic musical genre recognition
  - Automatic music instruments recognition.
  - Acoustic scene recognition
  - Sound samples classification.
  - Sound track labeling (speech, music, special effects etc…).
  - Automatically generated Play list
  - Hit predictor...

Some challenges in Audio listening

- **Huge databases of recordings and sounds**
- **But …. few recordings are precisely annotated**
  - Ex. *label is « bird song » while the bird song last 2s in a 1 mn recording*
- ***The individual sources composing the scene are rarely available.***
  - *Complexifies the learning paradigm*
- ***In Predictive maintenance, the abnormal event is very rare (sometimes never observed)***
  - *Importance of the few-shot learning paradigms, weakly supervised schemes.*

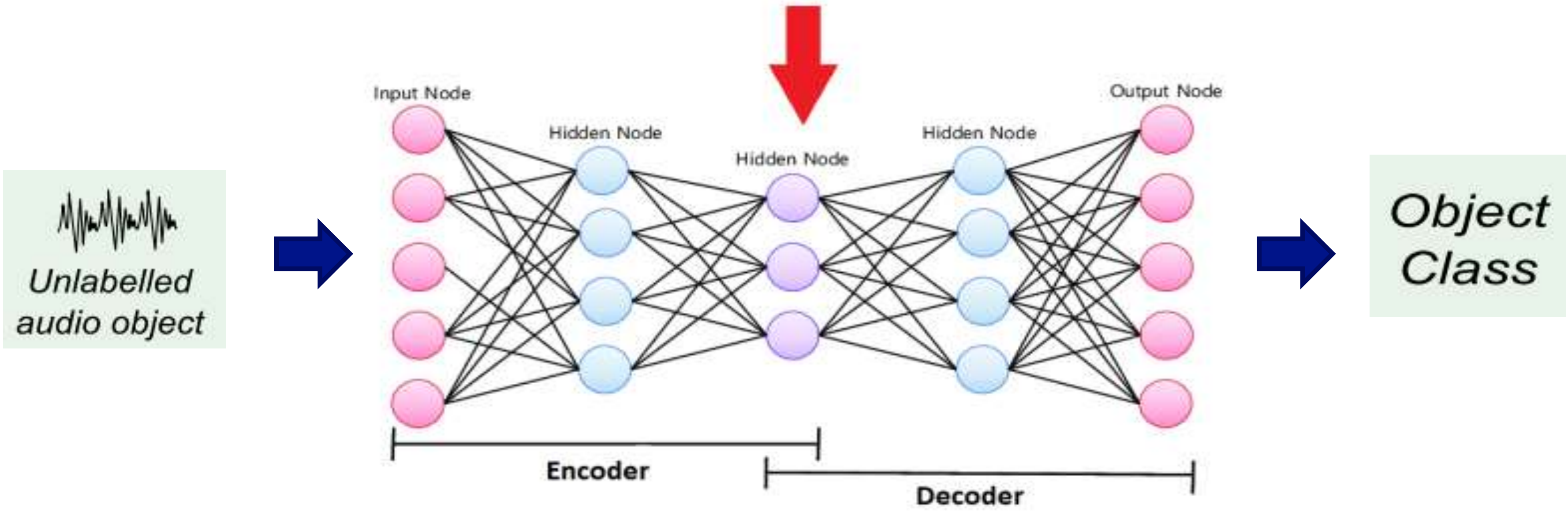
Traditional Classification system



From G. Richard, S. Sundaram, S. Narayanan, "Perceptually-motivated audio indexing and classification", Proc. of the IEEE, 2013

Current trends in audio classification

- **Deep learning now widely adopted**
  - For example under the form of encoder/decoder for representation learning

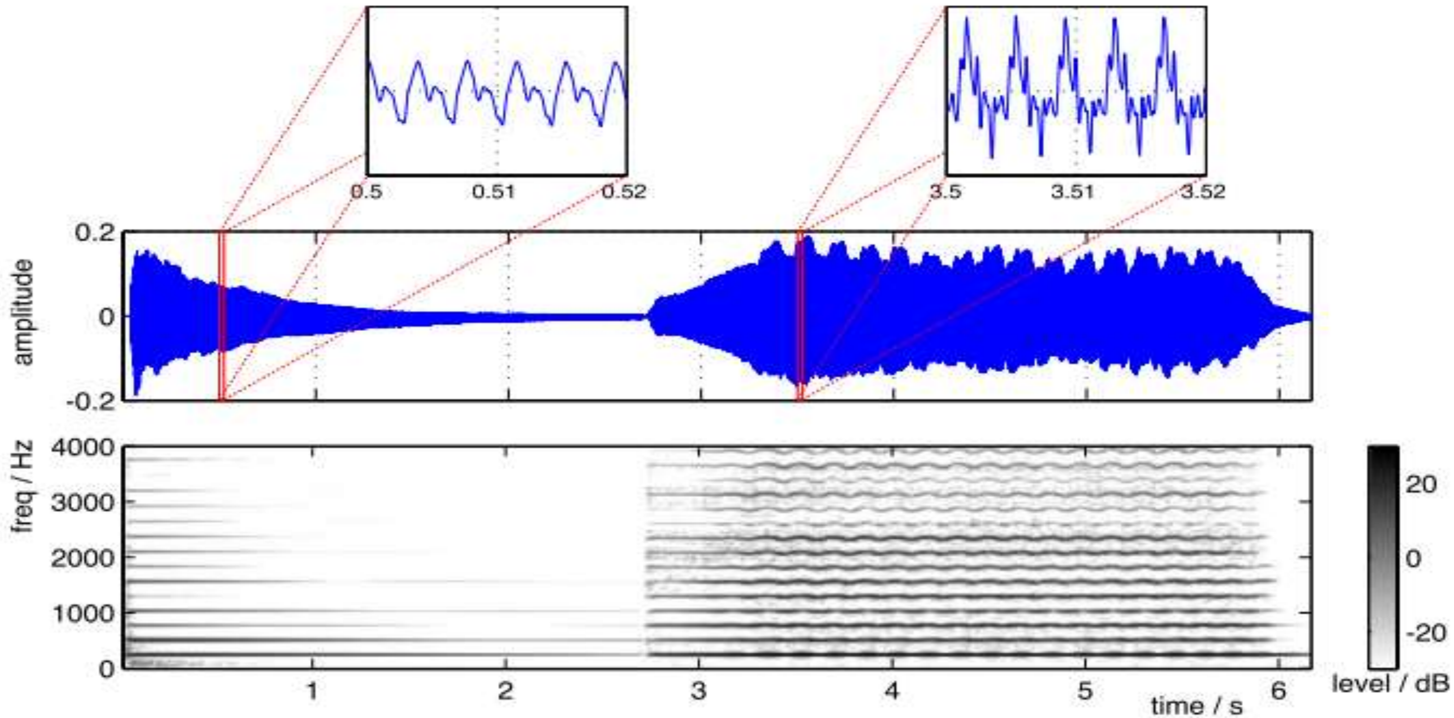


Audio signal representations

- **Example on a music signal: note C (262 Hz) produced by a piano and a violin.**

Temporal Signal

Spectrogram

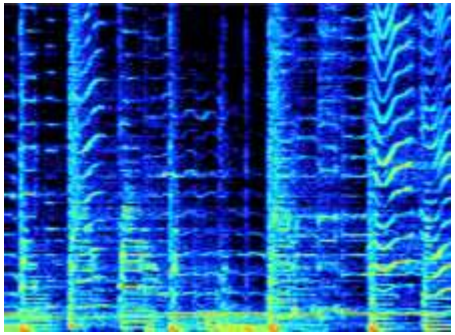


From M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Deep learning for audio

Differences between an image and audio representation

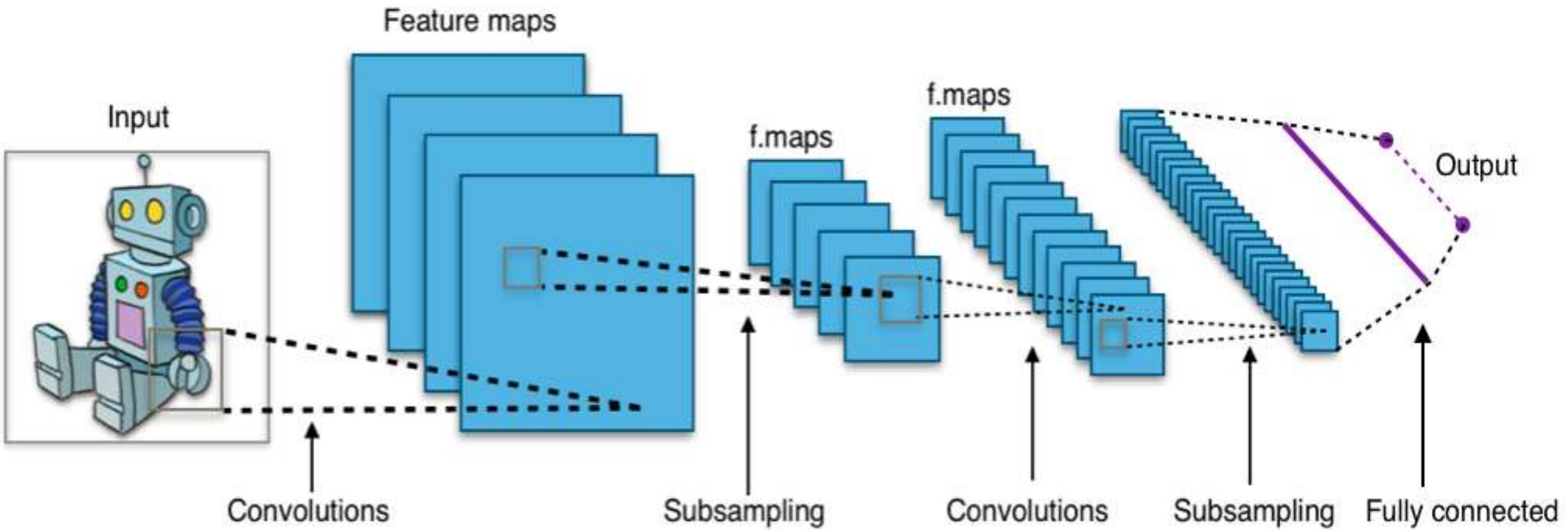
- x and y axes: **same concept** (spatial position).
- Image elements (cat's ear) : **same meaning** independently of their positions over x and y.
- **Neighbouring pixels** : often correlated, often belong to the same object
- **CNN are appropriate :**
  - Hidden neurons locally connected to the input image,
  - Shared parameters between various hidden neurons of a same feature map
  - Max pooling allows spatial invariance



- x and y axes: **different concepts** (time and frequency).
- Spectrogram elements (e.g. a time-frequency area representing a sound source): **same meaning** independently in time **but not over frequency**.
- No invariance over y (even with log-frequency representations): neighboring pixels of a spectrogram are not necessarily correlated since an harmonic sound can be distributed overt he whole frequency in a sparse way
- **CNN not as appropriate than it is for natural images**

G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021 (to appear)

A typical CNN



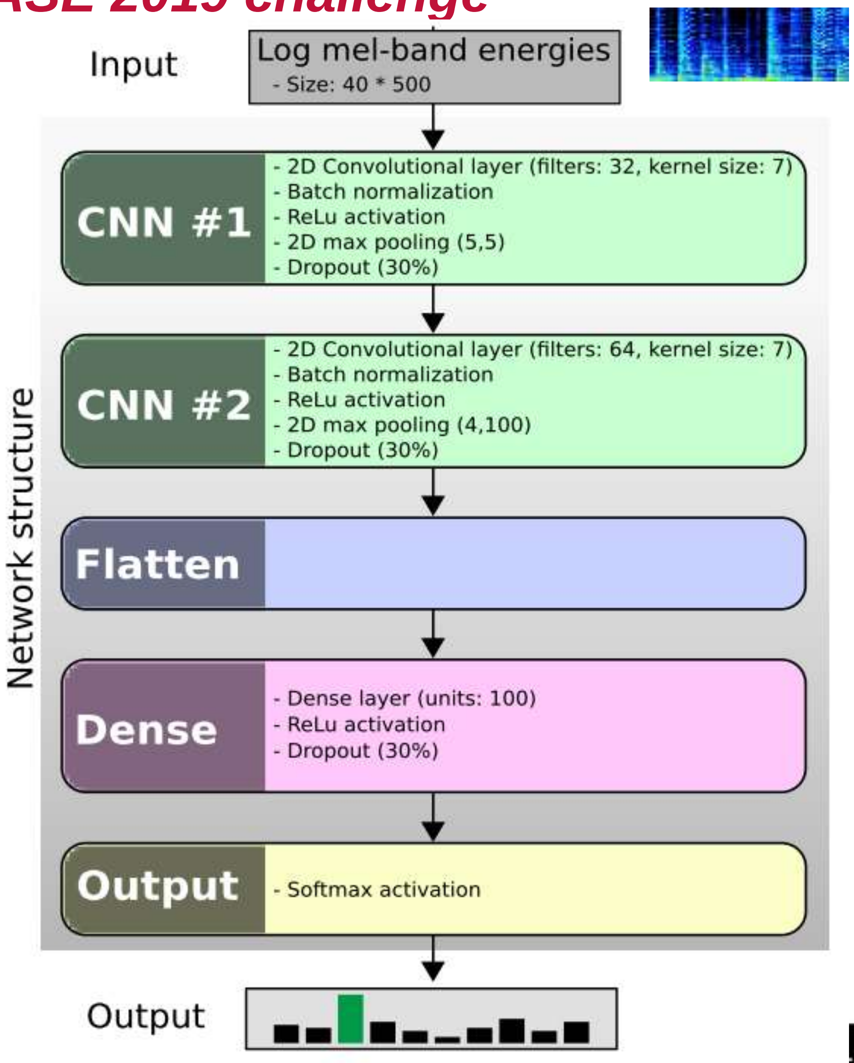
From https://en.wikipedia.org/wiki/Convolutional_neural_network

Acoustic scene recognition:
an example from the DCASE 2019 challenge

- **Baseline model**

  - **Input:** 10s audio file
  - **Analysis frame:** 40ms, 50% overlap
  - log mel-band energies extracted in 40 bands

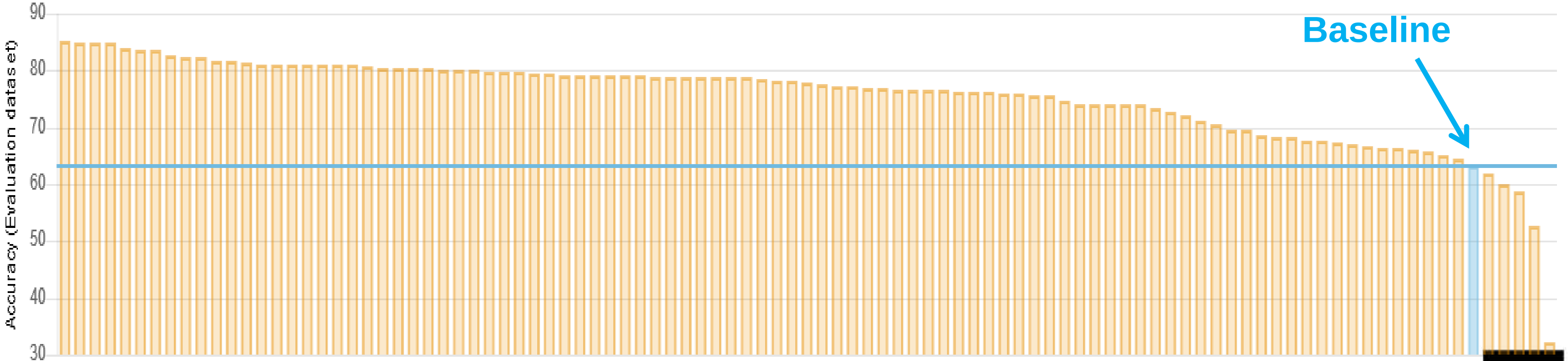
➡ input size: 40x500



A. Mesaros, T. Heittola, and T. Virtanen. *A multi-device dataset for urban acoustic scene classification.* In Proc. of DCASE 2018.

Acoustic scene recognition:
an example from the DCASE 2019 challenge

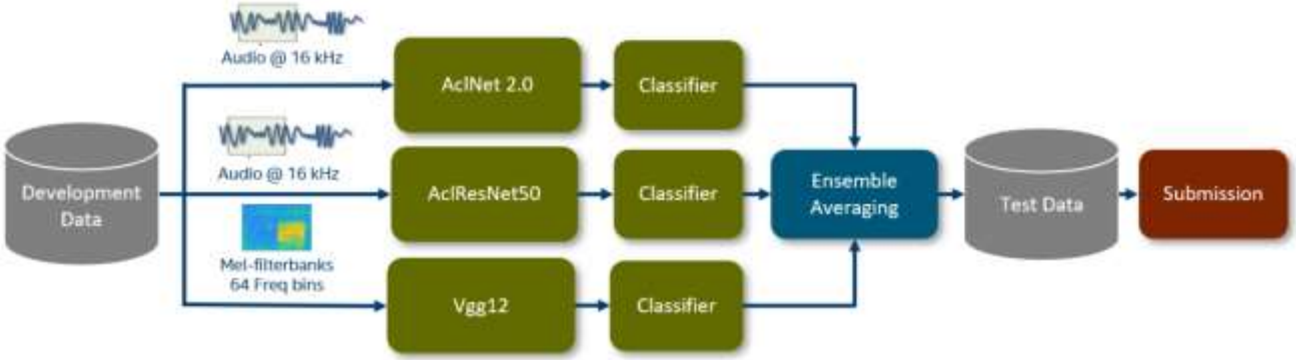
- **10 classes** (Airport, Indoor shopping mall, Metro station, Pedestrian street, Public square, Street with medium level of traffic, Travelling by a tram, Travelling by a bus, Travelling by an underground metro, Urban park)
- **12 cities** (10 only kept for training)
- **Training set: 40h of recordings**
- **Test set: 20h, from 12 cities** (2 not encountered in training)
- **The same recording device** for training and test sets (task 1A)



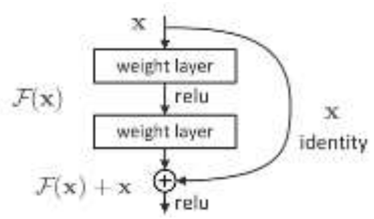
http://dcase.community/challenge2019/task-acoustic-scene-classification#results

Acoustic scene recognition:
How to improve ?

- **Some trends and tricks**
  - Use ensemble techniques



  - Use Data augmentation (*mix up, random cropping, channel confusion, Spectrum augmentation, spectrum correction, reverberation, pitch shift, speed change, random noise, mix audios, ...*)
  - Use large networks (> 17 layers), Resnets
  - Use signal or audio models (NMF, ..)



Hu Hu & al. Device-Robust Acoustic Scene Classification Based on Two-Stage Categorization and Data Augmentation,. in DCASE 2020 Acoustic Scene Classification Challenge
Lopez & al. “Ensemble of Convolutional Neural Networks”, in DCASE 2020 Acoustic Scene Classification Challenge

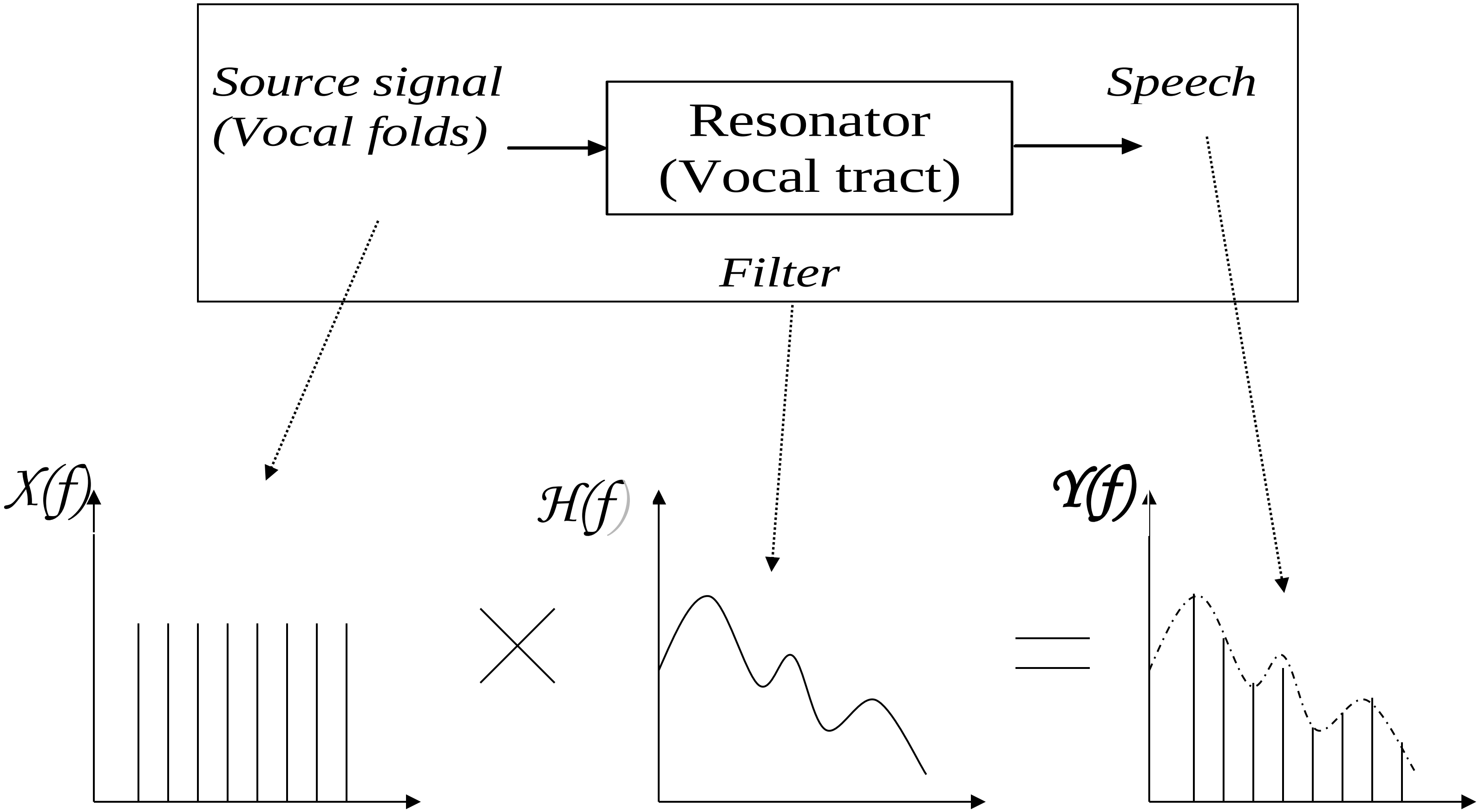
Acoustic scene recognition:
Why using signal or perceptual models

- **Using perceptual models**
  - Example: Mel specrogram, MFCC, CQT,..
  - The classifier does not learn what is not audible

- **Using signal models**
  - Example: Harmonic + noise, Source filter, NMF, …
  - *e.g The classifier does not learn what is not typical of an audio signal*

- **With such models**
  - The training may be simpler (faster convergence)
  - The need for data may be far less (frugality in data)
  - The need for complex architecture may be lower (frugality in computing power)

A widely used model: the source filter model



Signal models

- **Sinusoidal models**
- **Harmonic + noise models**
- **Other « decomposition » models**
  - Sparse representations
  - Non-negative matrix factorization

Sinusoidal models

- **Generic sinusoidal model**

$$x(n) = \sum_{i=1}^{I} A_i.sin(2\pi\nu_i n + \phi_i), \quad \nu_i \in [0, 1[$$

- **Harmonic + noise model**

$$x(n) = \sum_{i=1}^{I} A_i.sin(2\pi k_i \nu_0 n + \phi_i), \quad k_i\nu_0 \in [0, 1[$$

- **Model with modulated sinusoids and modulated noise**

$$x(n) = \sum_{i=1}^{I} A_i(n).sin(2\pi\nu_i n + \phi_i) + m(n).b(n)$$

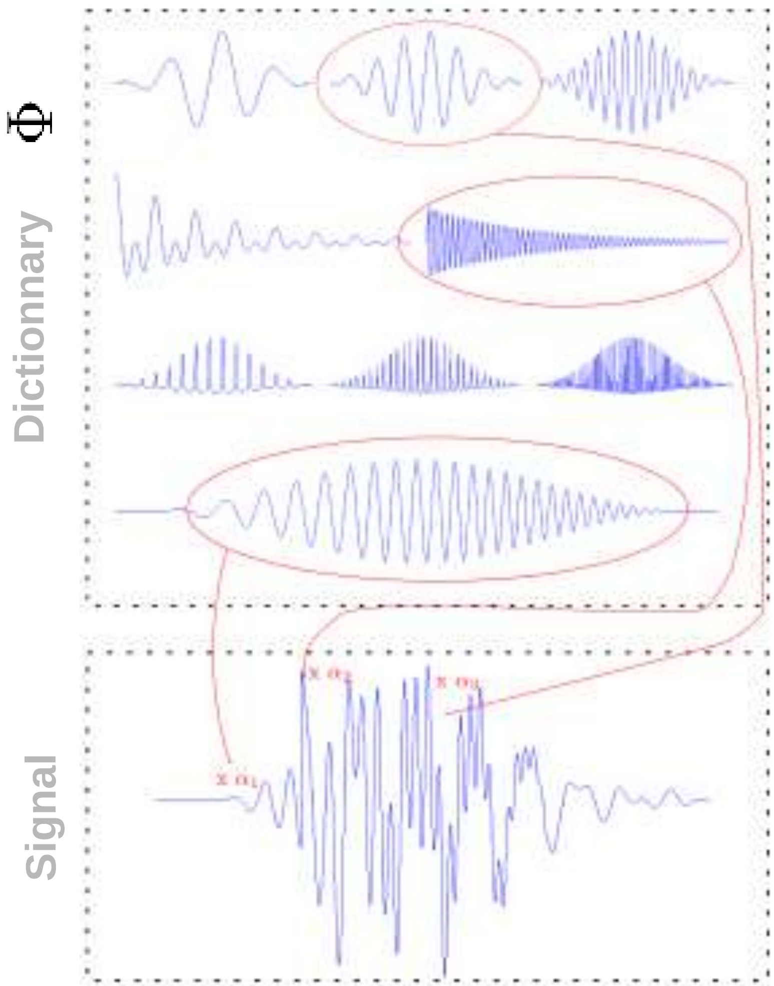
Sparse representation

- **Audio signal :**
  - Is a vector of high dimension: $x \in \mathbb{R}^N$

- **Definition:**
  - We have a set of atoms : $\{\phi_i\} \in \mathbb{R}^N$
    - Atoms can be time-frequency atoms, wavelets, modulated sinusoids …
  - And a dictionary of atoms: $\Phi = \{\phi_i\}_{i \in [0..M-1]}$
  - The sparse representation is expressed as a linear combination of only few atoms

$$x = \sum_{k=1}^{K} \alpha_k \phi_k$$

Sparse representation of an audio signal



- **Standard formulation**
- Let $x \in \mathbb{R}^N$, find the sparsest linear expression f on the dictionary $\Phi = \{\phi_i\}_{i \in [0..M-1]}$

Or

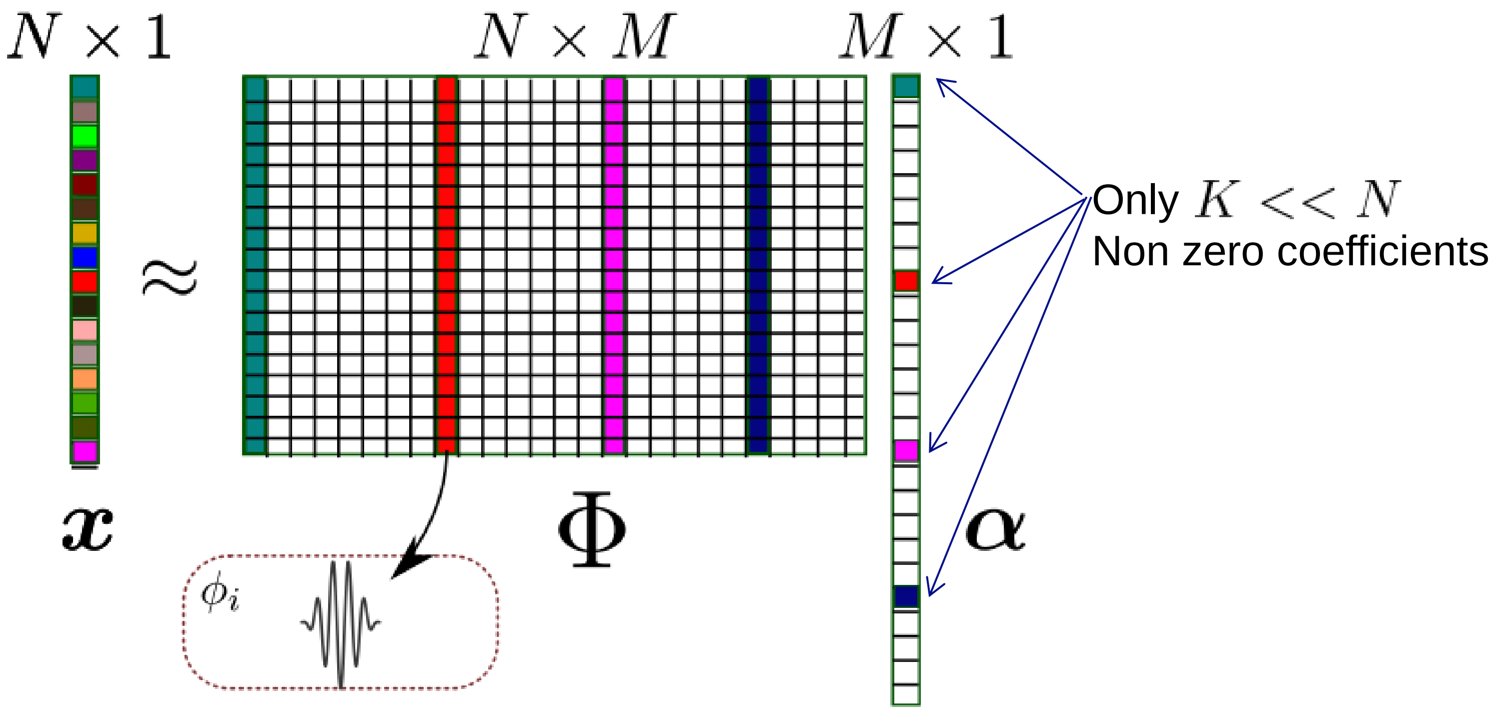
$$\min \|\alpha\|_0 \text{ s.t. } x = \Phi\alpha$$

Or alternatively

$$\min K \text{ s.t. } x = \sum_{k=1}^{K} \alpha_k \phi_k$$

Sparse representation of an audio signal

- Parsimony



Complexity of sparse approximation

- **Brute force approach: an exhaustive search amongst all potential combinations**

$$\min_x ||x - \Phi\alpha||_2 \quad \text{s.t.} \quad \text{support}(\alpha) = I$$

- **It can be shown that the l_0 minimisation problem (v. Davies et al, Natarajan) is NP-hard**

- **An alternative approach**
  - Greedy approaches

« Matching Pursuit »: a greedy approach

- **The atomic decomposition is obtained by « matching pursuit »**
  - The most correlated atom with the signal is first extracted and subtracted from the original signal
  - The process is iterated until a predefined number of atoms have beend subtracted (*or until a predefined Signal to noise ratio is reached*)

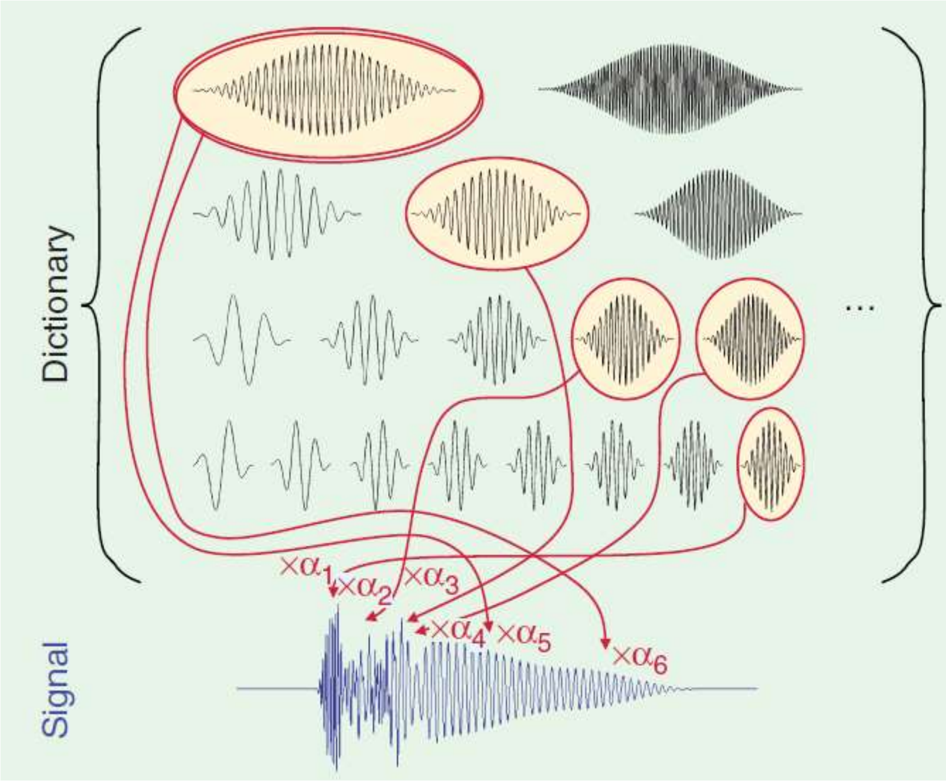
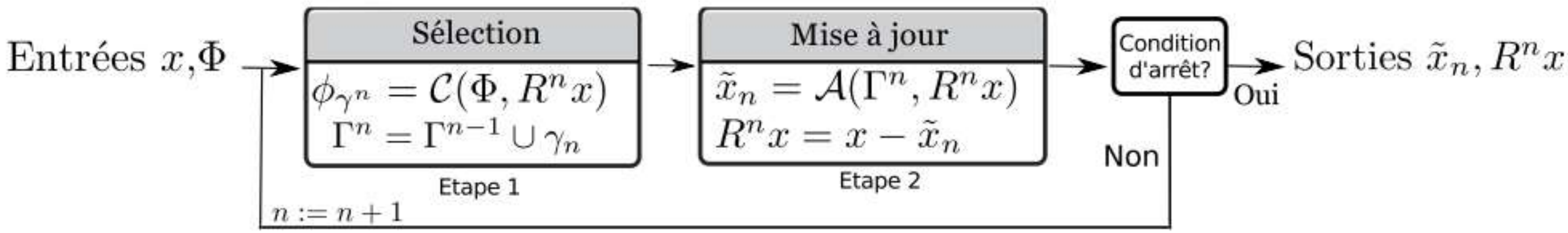


Figure from L. Daudet: *Audio Sparse Decompositions in Parallel,* IEEE Signal Processing Magazine, 2010

Standard Matching pursuit



- **Selection : the most correlated atom with the residual**

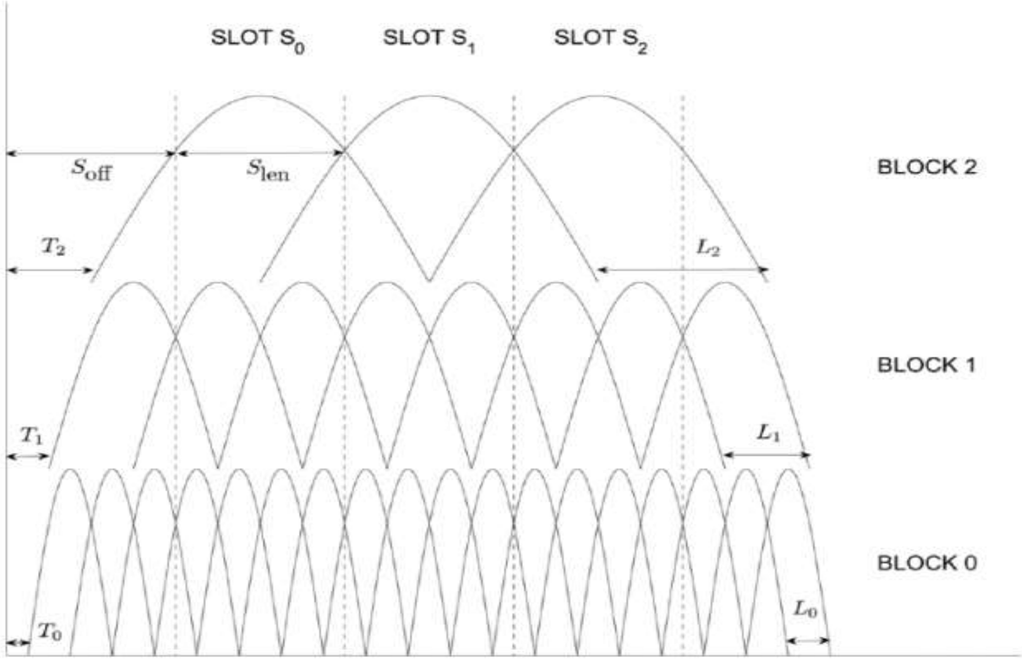
$$\phi_{\gamma^n} = \arg\max_{\phi_i \in \Phi} |\langle R^n x, \phi_i \rangle|$$

- **Update : subtraction**

$$R^{n+1}x = R^n x - \langle R^n x, \phi_{\gamma^n} \rangle \phi_{\gamma^n}$$

Union of MDCT bases

- **Possibility to build redundant dictionnaries : Union of MDCT MDCT (Modified Discrete Cosine Transform)** **(from E. Ravelli & al. 2008)**



Several variants exist

- Orthogonal matching pursuits (OMP)
- Cyclic Matching Pursuit (CMP)
- Weak Matching Pursuit
- Stagewise Greedy algorithms
- Stochastic Matching Pursuit
- Random Matching Pursuit

- ……

Use in music transcription

- **Idea: use a dictionary of “informed” atoms**

- **Music instrument recognition**
  - Build a dictionary with “characteristic” atoms of given instruments
  - For example, a set of atoms for each pitch and each instrument (obtained for example by VQ)

- **Multipitch extraction**
  - Build a dictionary with “characteristic” atoms of given pitches (note height)

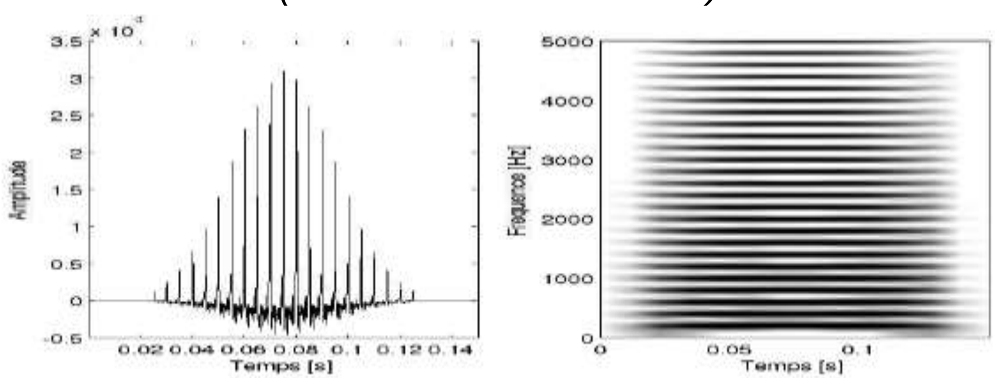
Use in music transcription

- **Harmonic atoms**

$$h_{s,u,f_0,c_0,A,\Phi}(t) = \sum_{m=1}^{M} a_m \, e^{j\phi_m} g_{s,u,m\times f_0,m\times c_0}(t)$$

- a_m (resp ϕ_m) amplitudes (resp. phases) des partiels
- s paramètre d'échelle
- u localisation temporelle
- f_0 (resp c_0) fundamental frequency and chirp rate

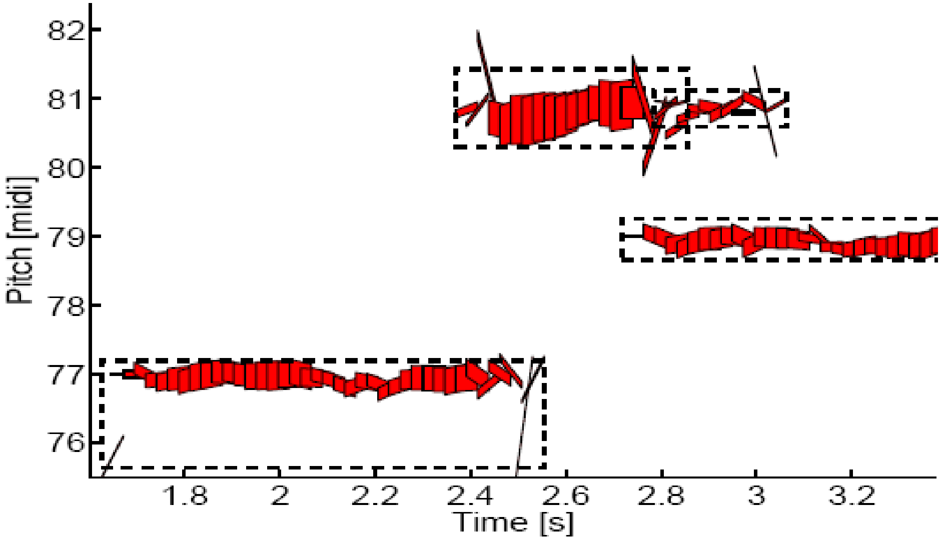
(from P. Leveau & al.2008)

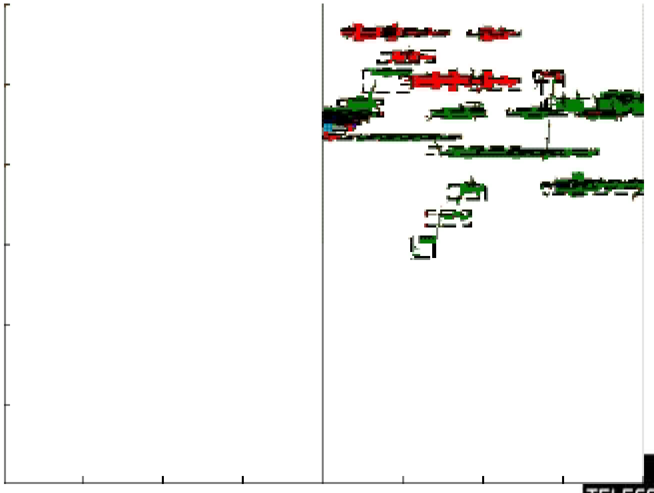


Use in music transcription

- **For example in music instrument recognition**
  - With atoms indexed by pitch/instrument
  - Possibility to build “molecules” (succession of “similar atoms)

Demo from P. Leveau





Non-negative Matrix Factorization (NMF)

- Use of non-supervised decomposition methods (for example Non-Negative Factorization methods or NMF)

- **Principle of NMF :**

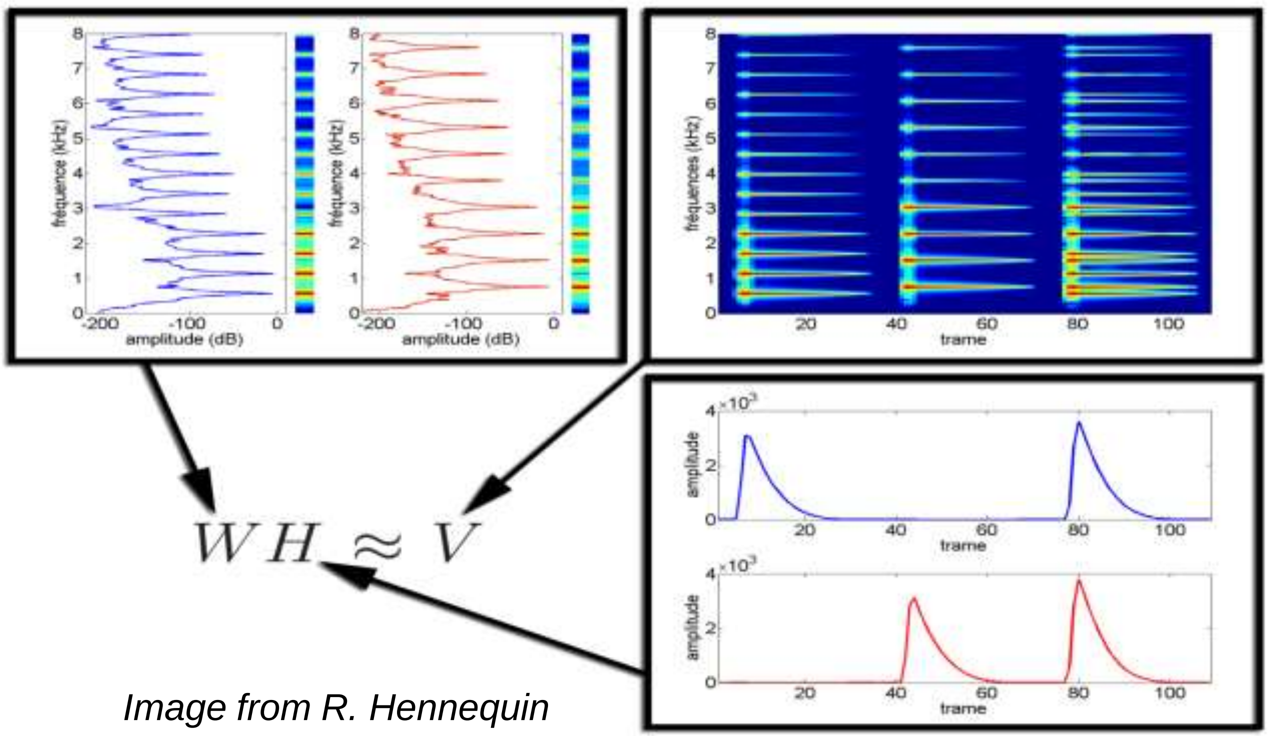
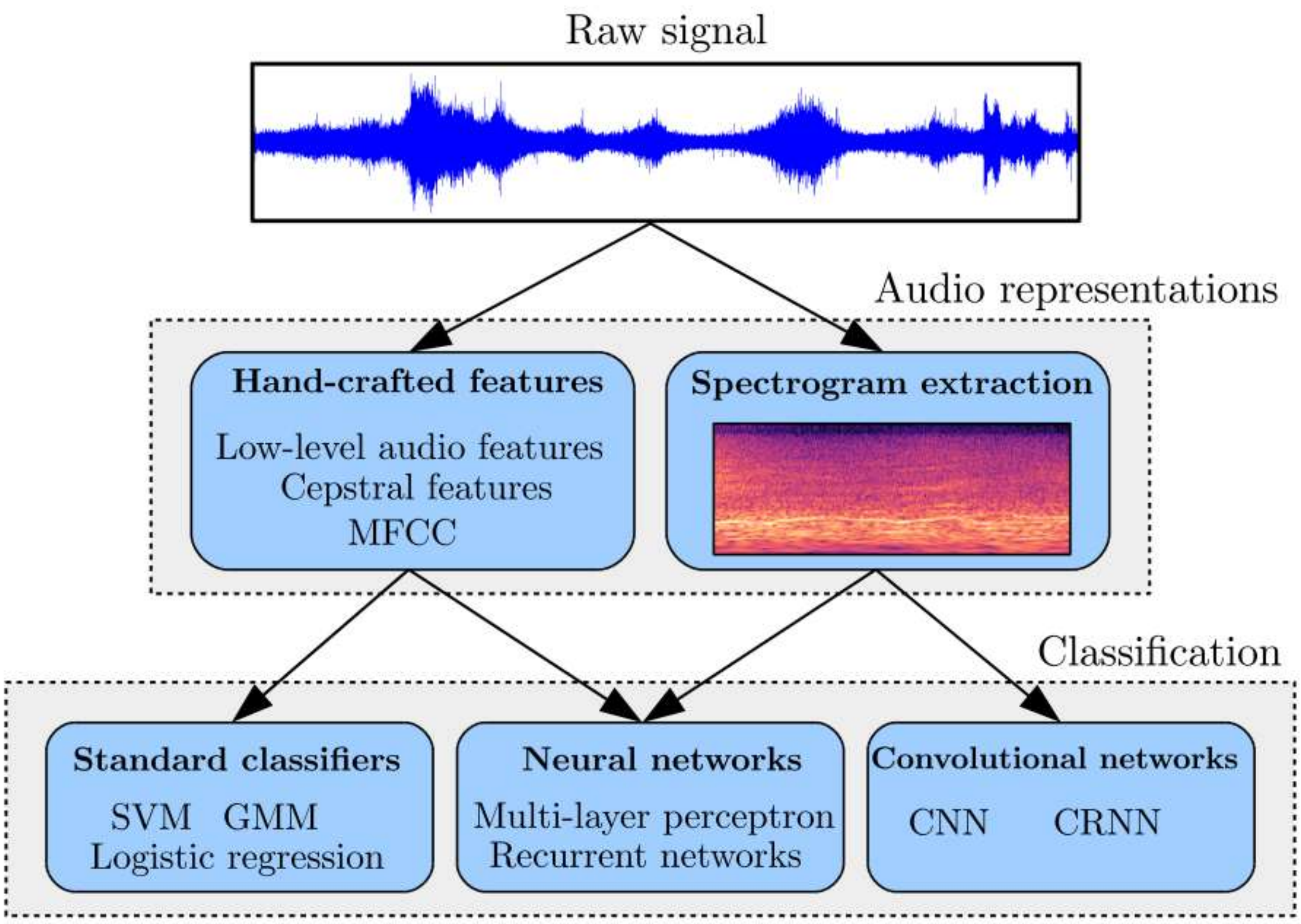
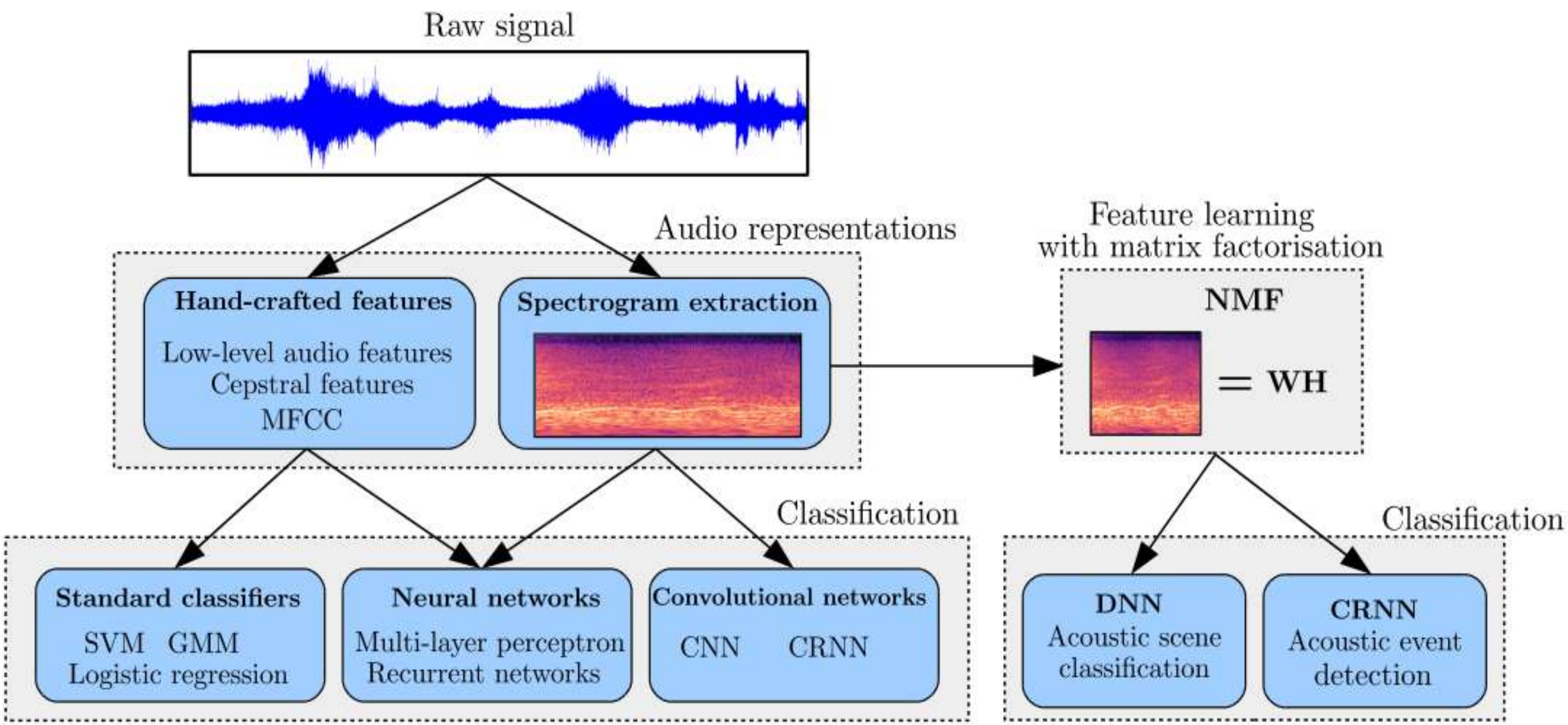


Image from R. Hennequin

Recent approaches for Audio scene and event recognition



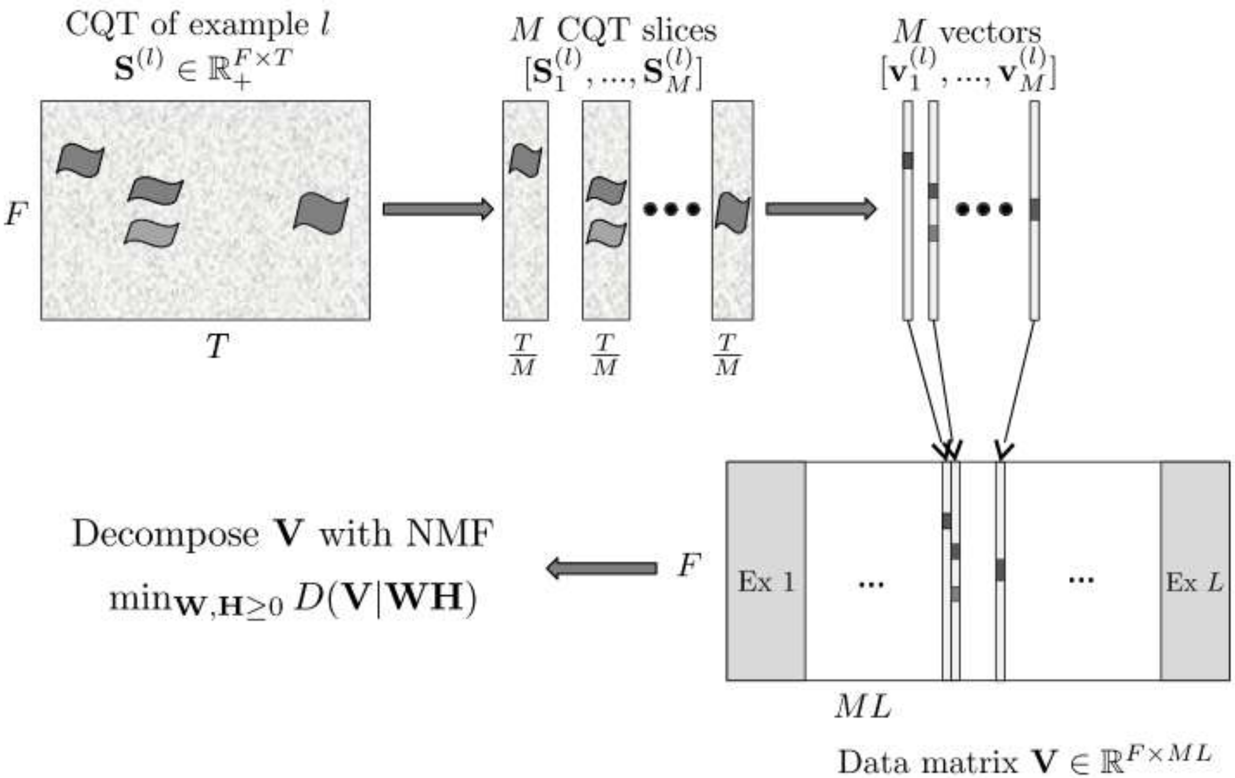
A recent framework for Audio scene and event recognition (Bisot & al. 2017)



V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),
V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tokyo,

Example for scene classification

From time-frequency representations to dictionary learning

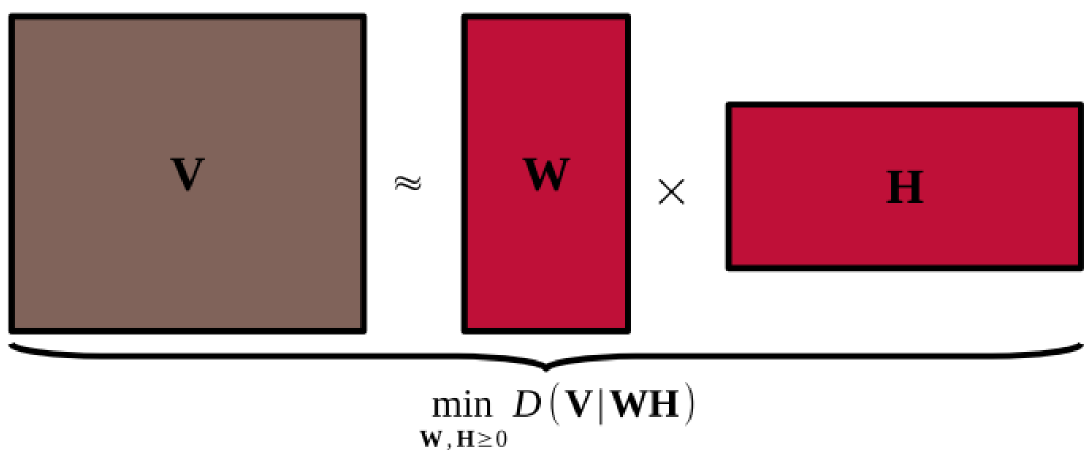


Unsupervised NMF for acoustic scene recognition

Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})$ with $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$

Dictionary learning with NMF

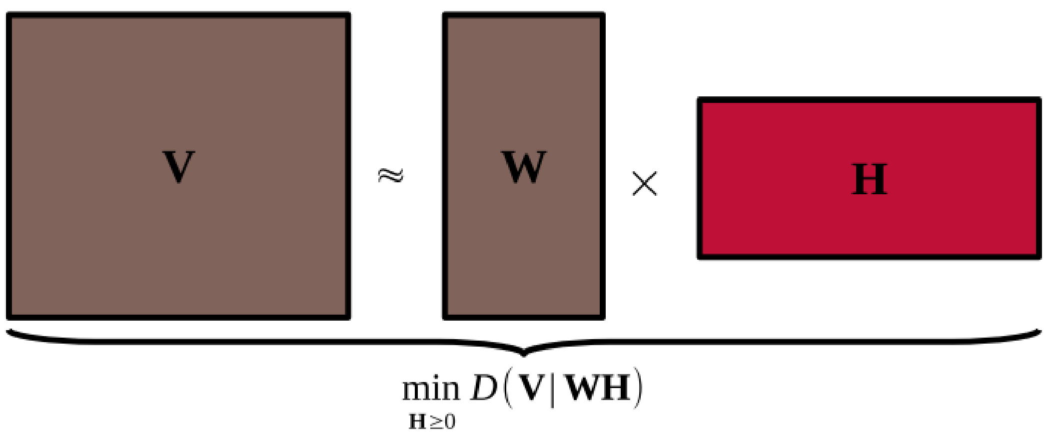


Unsupervised NMF for acoustic scene recognition

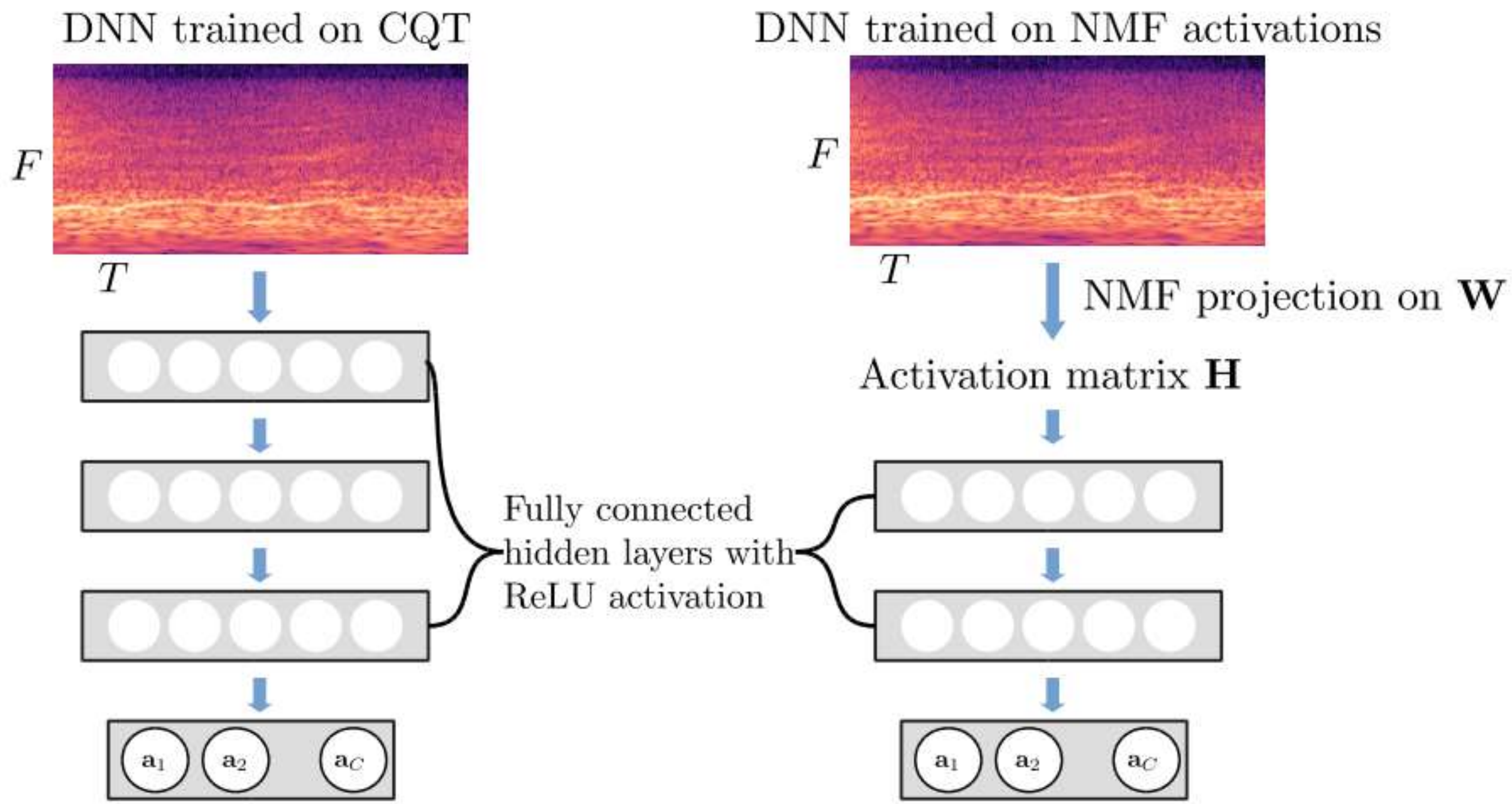
Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH})$ with $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$

Feature extraction $\longrightarrow$ project on learned dictionary

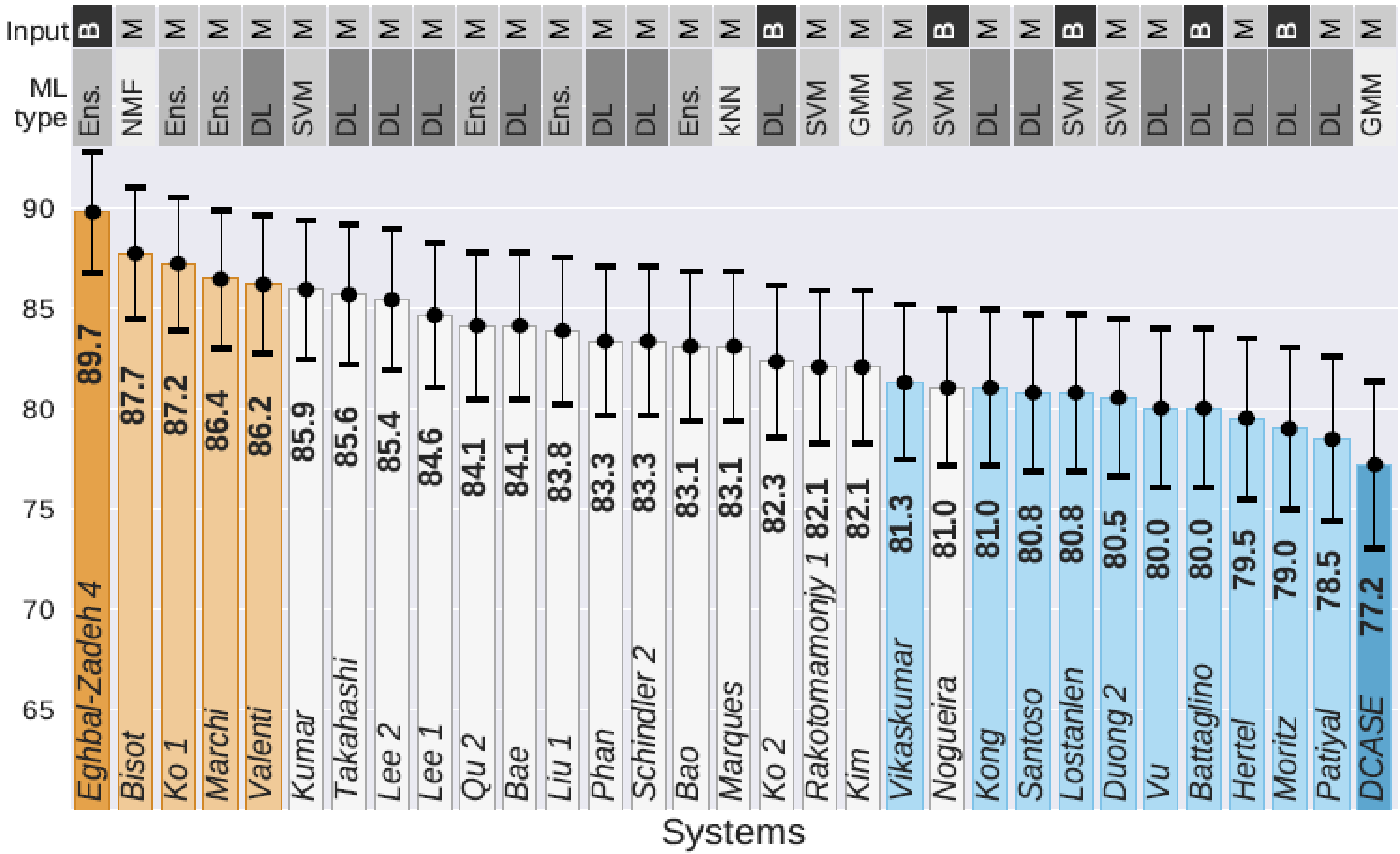


Example with DNN: acoustic scene recognition



V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),
V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tokyo,

Typical performances of Acoustic scene recognition (challenge DCASE 2016)



- *A Mesaros & al. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 challenge IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2), 379-393*

Audiofingerprint
(Reconnaissance musicale)

Audio Identification ou AudioID

- **Audio ID = find high-level metadata from a music recording**

- **Challenges:**
  - Efficiency in adverse conditions (distorsion, noises,..)
  - Scale to “Big data” (bases > millions of titles)
  - Rapidity / Real time

- **Product example : Shazam**

Audio fingerprinting

- **Audio Fingerprinting: One possible approach**
- **Principle :**
  - For each reference, a unique “fingerprint” is computed
  - Music recordings recognition: compute its “fingerprint” and comparison with a database of reference fingerprints .

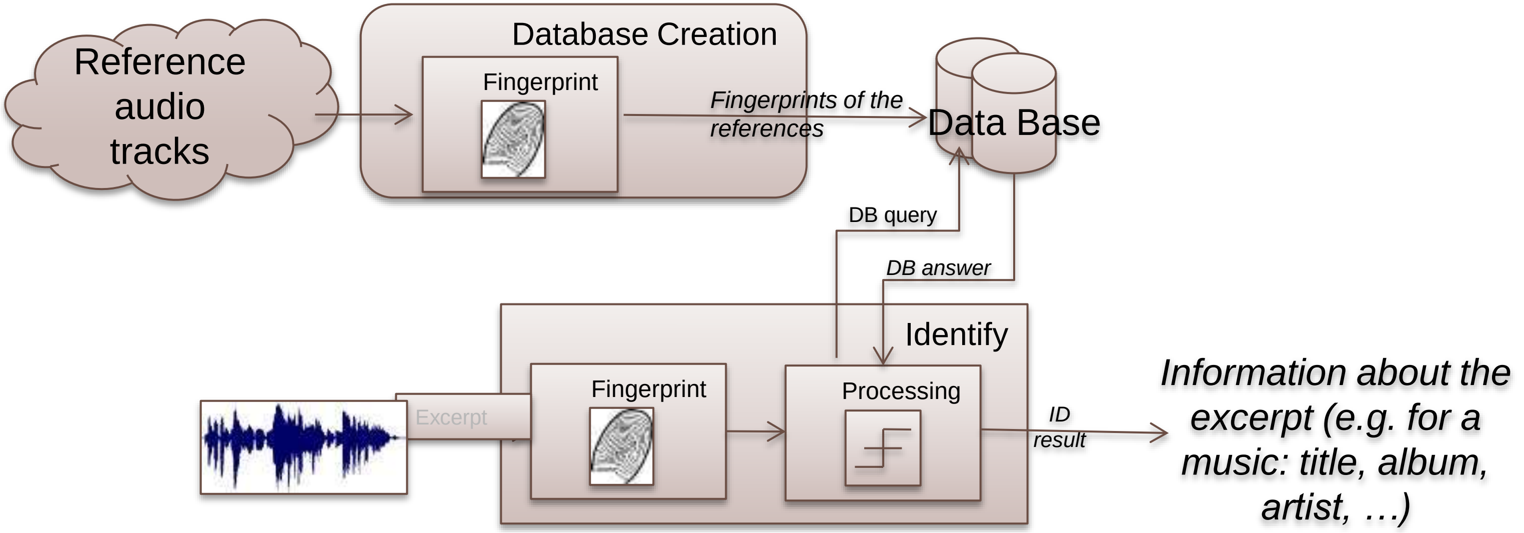
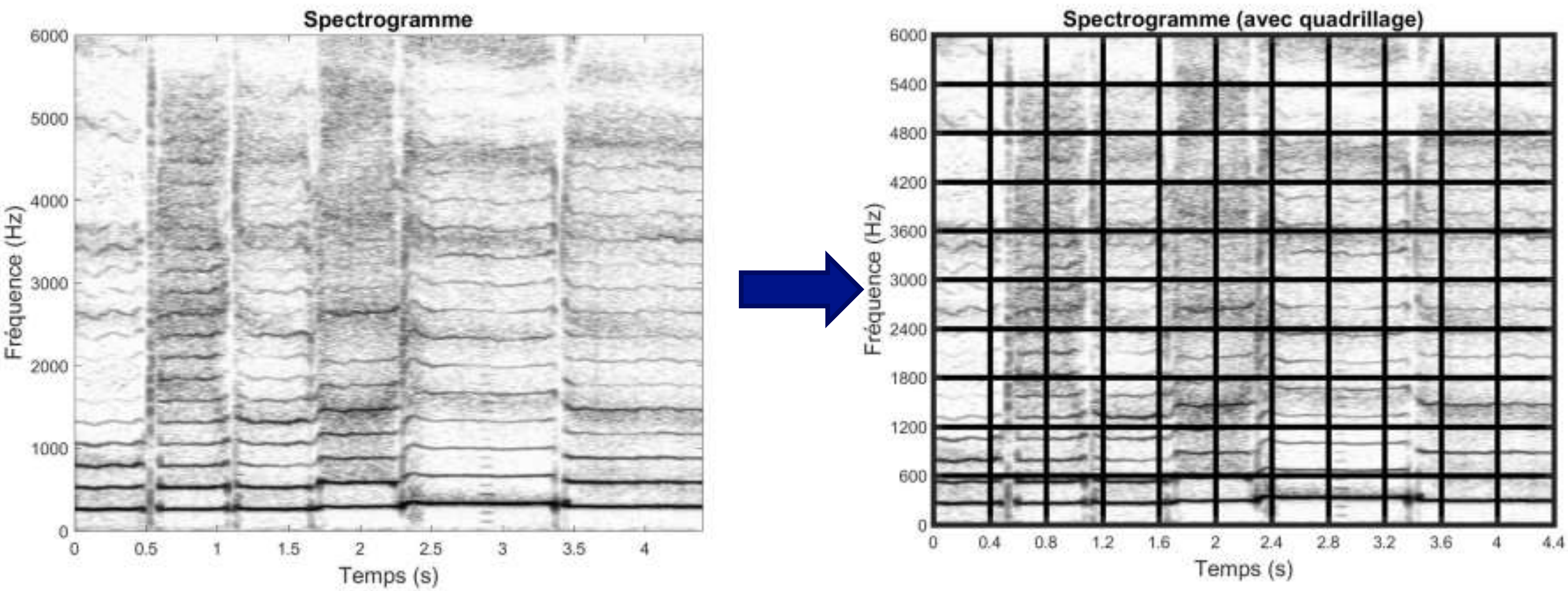


Figure from Sébastien Fenêt

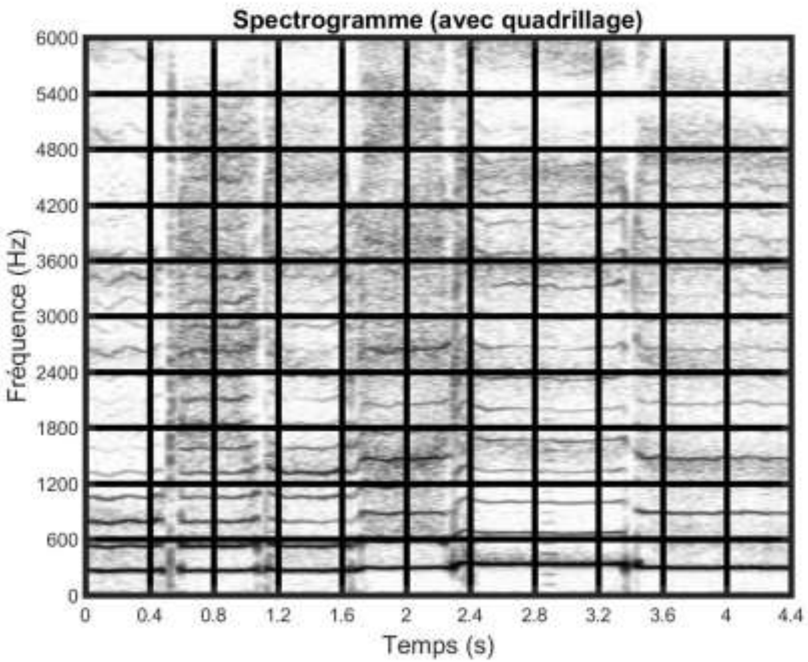
Signal model : from spectrogram to “schematic binary spectrogram”

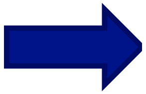
- **1st step: split the spectrogram in time-frequency zones**

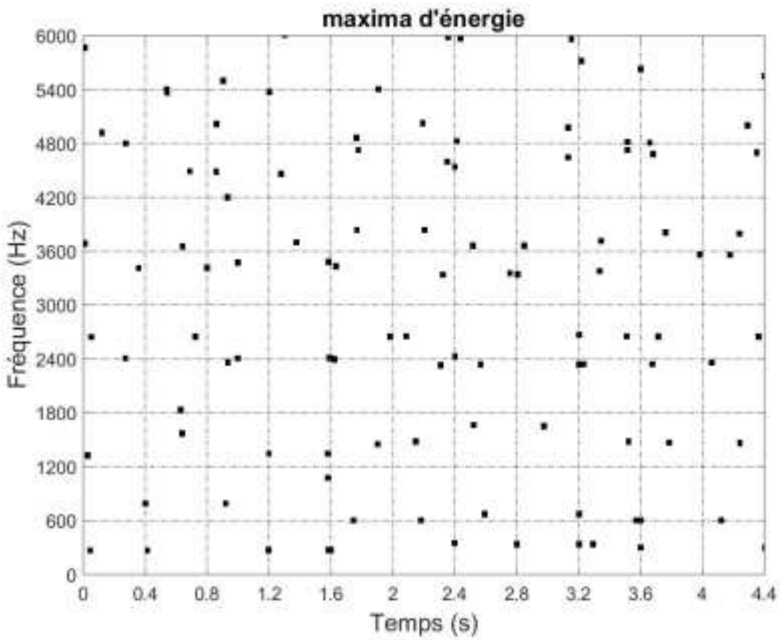


Signal model : from spectrogram to "schematic binary spectrogram"

- **2nd step: peak one maximum per zone**







Efficient research strategy

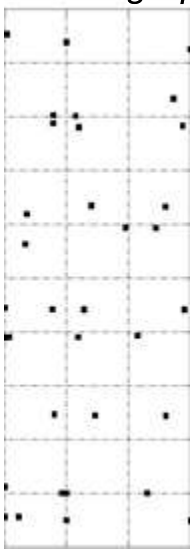
- **Towards identifying an Unknown recording using a large database of known references**

Potential strategies

- Direct comparison with each reference of the database (with all possible time-shifts)
- Use “black dots” as index (see figure)

- Alternative: ?

Test fingerprint



Efficient research strategy

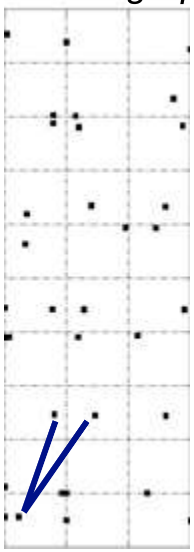
- **Towards idetifying an Unknown recording using a large database of known references**

Potential strategies

- Direct comparison with each reference of the database (with all possible time-shifts)
- Use “white dots” as index (see figure)

- Alternative: Use pairs of “white dots”

Test fingerprint



Find the best reference

- **To be efficient: necessity to rely on an « index »**
- **For each pair, a query is made in the database for obtaining all references who have this pair, and at what time it appears**
- **If the pair appears at T1 in the unknown recording and at T2 in the reference, we have a time shift of:**
  - ΔT(pair)=T2-T1

- **In summary, the algorithm is :**

```
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;
```

Find the best reference

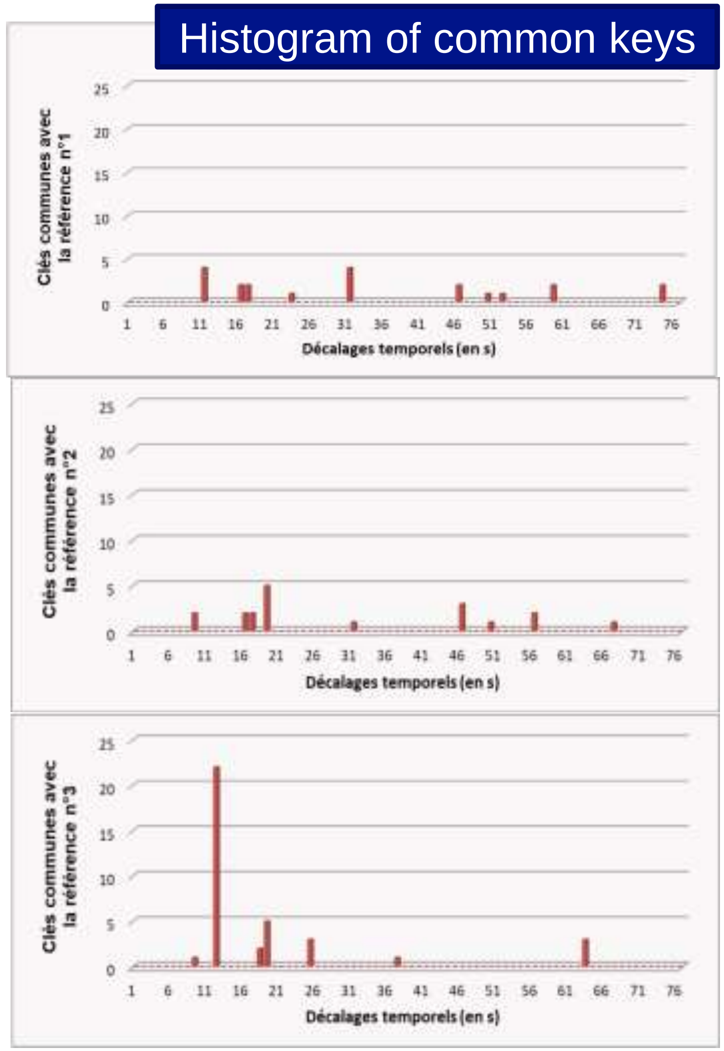
- **The three main steps for the recognition:**

  1. **Extraction of pair maxima (with their position in time) from the unknown recording**. Each pair is a « key » and is encoded as a vector [f_1, f_2,t_2 −t_1] where ($f_1 t_1$) (resp. (f_2,t_2) is the time-spectral position of the first (resp. second) maximum

  2. **Search in the database for all candidate references** (e.g. those who have common pairs with the unknown recording). For each key, the time shift $\Delta t = t_1 - t_{ref}$ where t_1 and t_{ref} are respectively the time instant of the first maximum of the key in the unknown and in the reference recording.

  3. **Recognition:** The reference which has the most keys in common at a constant Δt is the recognized recording

Find the best reference :Illustration of the histogram of Δt with 3 references



Reference 1

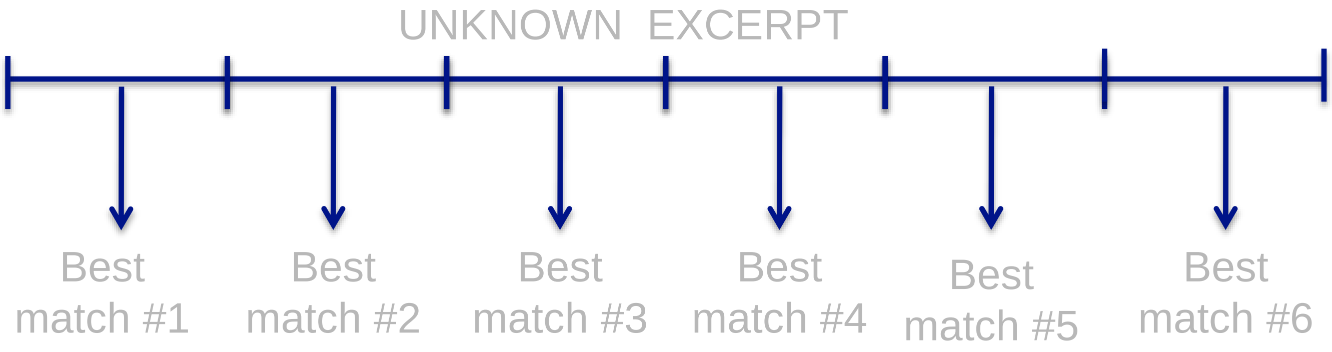
Reference 2

Reference 3

Recognized recording

Detection of an "out-of-base" recording : local decision fusion

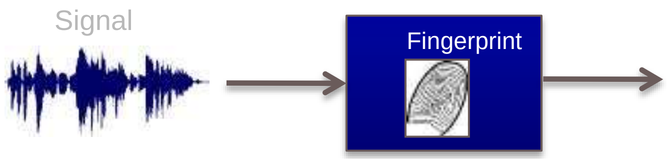
- **The unknown recording is divised in sub-segments**
- **For each sub-segment, the algorithm gives back a best candidate**



- **If a reference appears predominantly (or more than a predefined number of time), it is a valid recording to be recognized**
- **Otherwise, the query is rejected**
- **High rate can be achieved (over 90%)**

An alternative with different time-frequency representations: use of Matching pursuit

- **Most systems relay on “fingerprints” computation**



$$\begin{matrix}(key_1, t_1)\\ (key_2, t_2)\\ \dots \\ (key_n, t_n)\end{matrix}$$

Signal fingerprints

- **Possibility: use MP with time-frequency coverage constraints to obtain fingerprints.**

$$\mathcal{C}_{\mathcal{M}}(R^n x, \Phi) = \arg\max_{\phi_i \in \Phi} \left(|\langle R^n x, \phi_i \rangle| \mathcal{M}(\phi_i | \Gamma^n) \right)$$

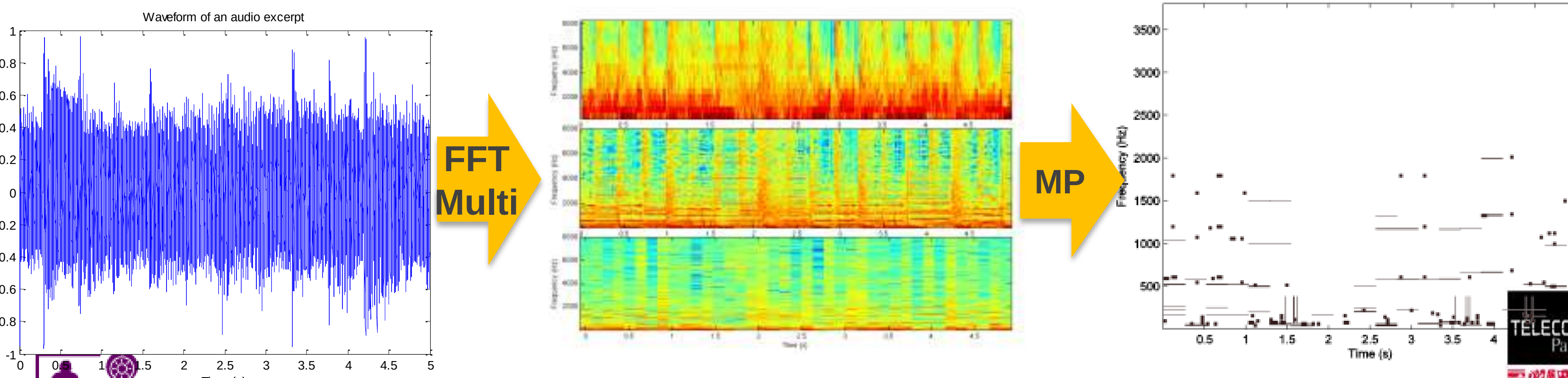
$$\mathcal{M}(\phi_i|\Gamma^n) = 1 - \max_{\gamma \in \Gamma^n} |\langle \phi_i, \phi_\gamma \rangle|$$

Audio fingerprints obtained by MP

- **use MP with time-frequency coverage constraints to obtain fingerprints.**
  - One key = one atom (scale and frequency)

$$\mathcal{C}_{\mathcal{M}}(R^n x, \Phi) = \arg\max_{\phi_i \in \Phi} \left(|\langle R^n x, \phi_i \rangle| \mathcal{M}(\phi_i | \Gamma^n) \right)$$

$$\mathcal{M}(\phi_i|\Gamma^n) = 1 - \max_{\gamma \in \Gamma^n} |\langle \phi_i, \phi_\gamma \rangle|$$



Performance examples (Evaluation – recurrent events detection) - Quaero 2012

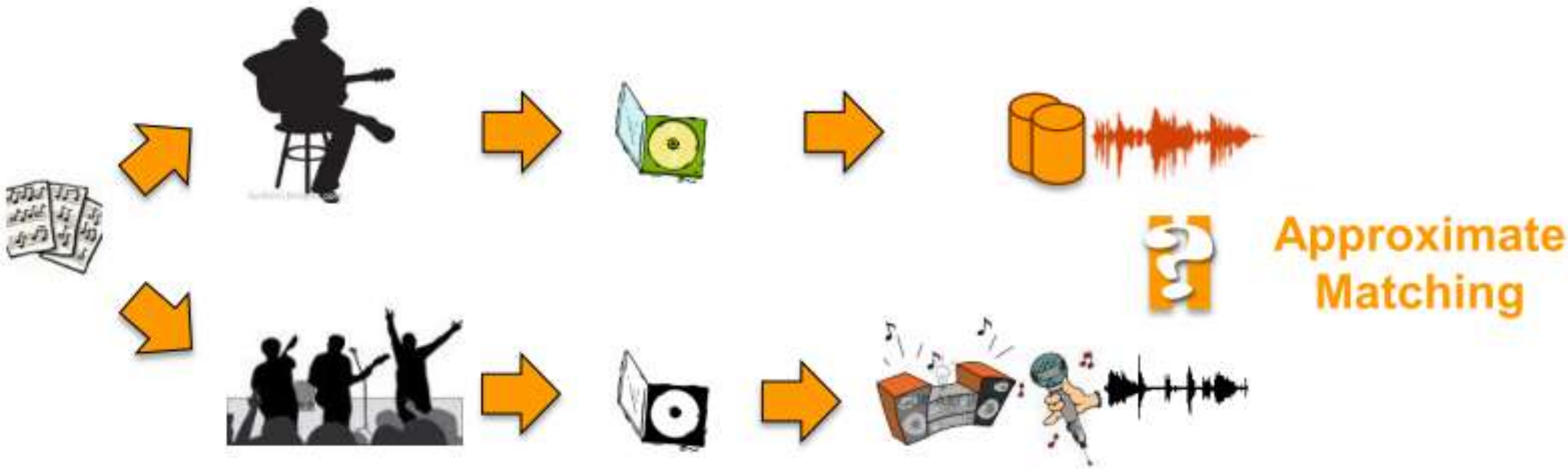
- **2 real world corpora:**
  - 3 days of the same radio (72 h)

Algorithm	Télécom - CQT	Télécom - MP
Recall	1.00	0.95
Precision	0.99	0.99

  - The same day for 3 different radios (72 h)

Algorithm	Télécom - CQT	Télécom - MP
Recall	0.97	0.78
Precision	0.99	1.00

Extension : « Approximate » Real-time Audio identification
(Fenet & al.)



- **Audio recordings recognition**
  - Identical
  - Approximate (live vs studio)

  - For music recommendation, second screen applications, …

G. Richard & al. “De Fourier à reconnaissance musicale”, Revue Interstices, Fev. 2019, online at: https://interstices.info/de-fourier-a-la-reconnaissance-musicale/ (in French)
S. Fenet & al. An Extended Audio Fingerprint Method with Capabilities for Similar Music Detection. ISMIR 2013

A few additional references…

- **Audio classification / Music signal processing**
  - M. Mueller, D. Ellis, A. Klapuri, G. Richard, Signal Processing for Music Analysis", IEEE Journal on Selected Topics in Signal Processing, October 2011.
  - G. Richard, S. Sundaram, S. Narayanan "An overview on Perceptually Motivated Audio Indexing and Classification", Proceedings of the IEEE, 2013.
  - M. Mueller, Fundamentals of Music Processing, “Audio, Analysis, Algorithms, Applications, Springer, 2015
  - A. Klapuri A. M. Davy, Methods for Music Transcription M. Springer New York 2006
  - G. Peeters, “Automatic classification of large musical instrument databases using 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)
  - G. Peeters, G. Richard, Deep Learning for Audio and Music, published in Multi-faceted Deep Learning: Models and Data, edited by J. Benois-Pineau, A. Zemmari, 2021, Springer

- **Signal models**
  - D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,”Nature, vol. 401, no. 6755, pp. 788–791,1999.
  - P. Leveau, E. Vincent, G. Richard, and L. Daudet, “Instrument-specific harmonic atoms for mid-level music representation,” *IEEE Trans. Audio, Speech and Language Processing, vol. 16, no. 1, pp. 116–128,* 2008.
  - S. Mallat and Z. Zhang, “Matching pursuits with timefrequency dictio-naries,”IEEE Trans. Signal Process., vol. 41, no. 12, pp. 3397–3415,Dec. 1993.
  - L. Daudet: *Audio Sparse Decompositions in Parallel,* IEEE Signal Processing Magazine, 201
  - E. Ravelli, G. Richard, L. Daudet, Union of MDCT bases for audio coding, IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, Issue 8, pp 1361-1372, Nov. 2008.
  - G. Richard, C. d’Alessandro, “Analysis/synthesis and modification of the speech aperiodic component”, Speech Communication, Vol. 19, Issue 3, September 1996, Pages 221–244

A few references…

- ***AudioFingerprint***
  - G. Richard & al. “De Fourier à reconnaissance musicale”, Revue Interstices, Fev. 2019, online at: https://interstices.info/de-fourier-a-la-reconnaissance-musicale/ (in French)
  - S. Fenet & al. An Extended Audio Fingerprint Method with Capabilities for Similar Music Detection. ISMIR 2013
  - S. Fenet, M. Moussallam, Y. Grenier, G. Richard et L. Daudet, (2012), A Framework for Fingerprint-Based Detection of Repeating Objects in Multimedia Streams, "EUSIPCO", Bucharest, Romania, pp. 1464-1468.
  - A. Wang, “An Industrial-strength Audio Search Algorithm,” in SMIR, 2003.

- ***Acoustic Scene and event recognition***
  - *V. Bisot & al., "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017),*
  - *V. Bisot & al., Leveraging deep neural networks with nonnegative representations for improved environmental sound classification IEEE International Workshop on Machine Learning for Signal Processing MLSP, Sep 2017, Tokyo,*
  - A Mesaros & al. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 challenge IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (2), 379-393
  - D. Barchiesi, D. Giannoulis, D. Stowel, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds theyproduce,”IEEE Signal Processing Magazine, vol. 32, no. 3, pp. 16–34, 2015
  - P. Lopez & al. “Ensemble of Convolutional Neural Networks”, in DCASE 2020 Acoustic Scene Classification Challenge
  - T. Virtanen, M. Plumbley, D. Ellis, Computational Analysis of Sound Scenes and Events, Springer, 2018