Master M2 - DataScience

Audio and music information retrieval

Lecture on
Signal Models, Decomposition models, Music recognition, Scene/events recognition (DCASE)

Gaël RICHARD

Télécom Paris

March 2024

Content

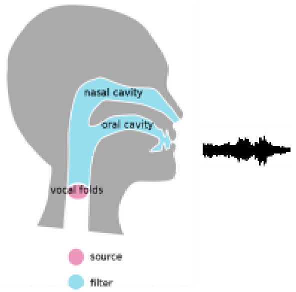
- **Introduction**
- **A Sound production model**
- **(A few) elements of sound perception**
  - Basics of perception
  - Example of perception principles in models
- **Signal decomposition models**
  - Sinusoidal models
  - Decomposition models (matching pursuit, NMF)
  - Exploitation of such models in scene analysis
- **Audiofingerprint or Music recognition**
- **Machine listening or DCASE**

Lecture 10: What you need to know

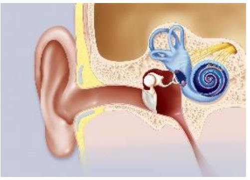
- **Models, Signal Representation**
  - What is the threshold of hearing
  - What is NMF ? How it is applied to Audio
  - What is the source-filter model of speech production ?
  - What is Matching pursuit ? How can it be applied to audio/music analysis
- **Audio Fingerprint**
  - What is an audio fingerprint ? How can it be computed ?
- **Audio events and acoustic scene recognition**
  - What is polyphonic event detection ?
  - Explain how to evaluate sound detection performances (metrics, …)

Audio Models

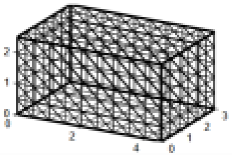
- **Audio models can represent the knowledge of**
  - How the sound is produced (sound production models)



  - How the sound is perceived (perception models)

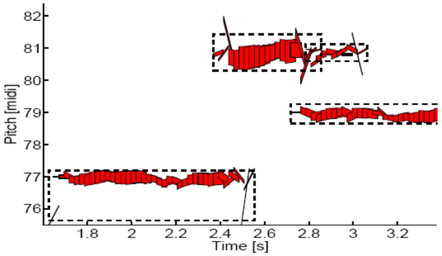


  - How the sound propagates (sound rendering or reverberation models)

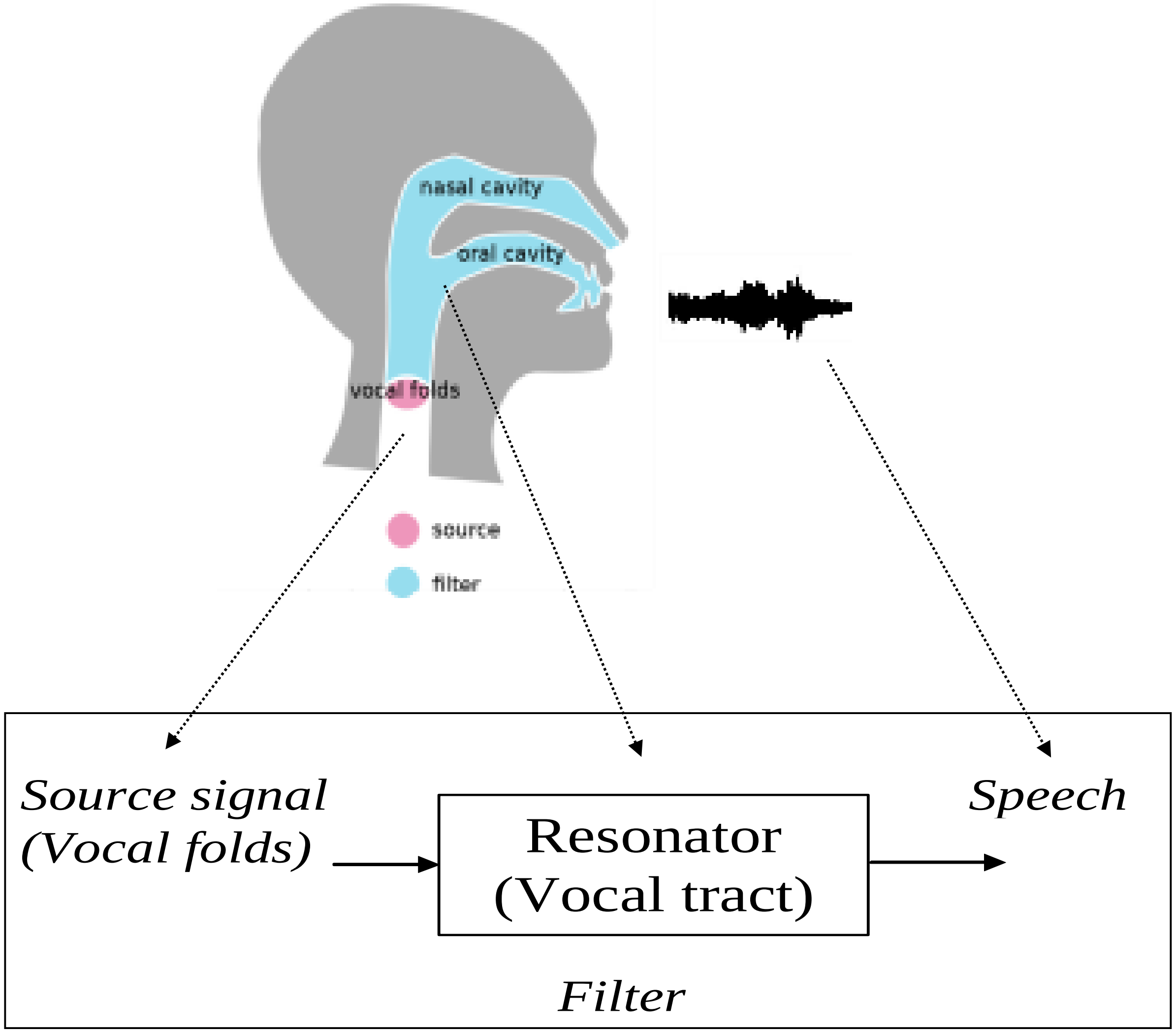


Discretized room

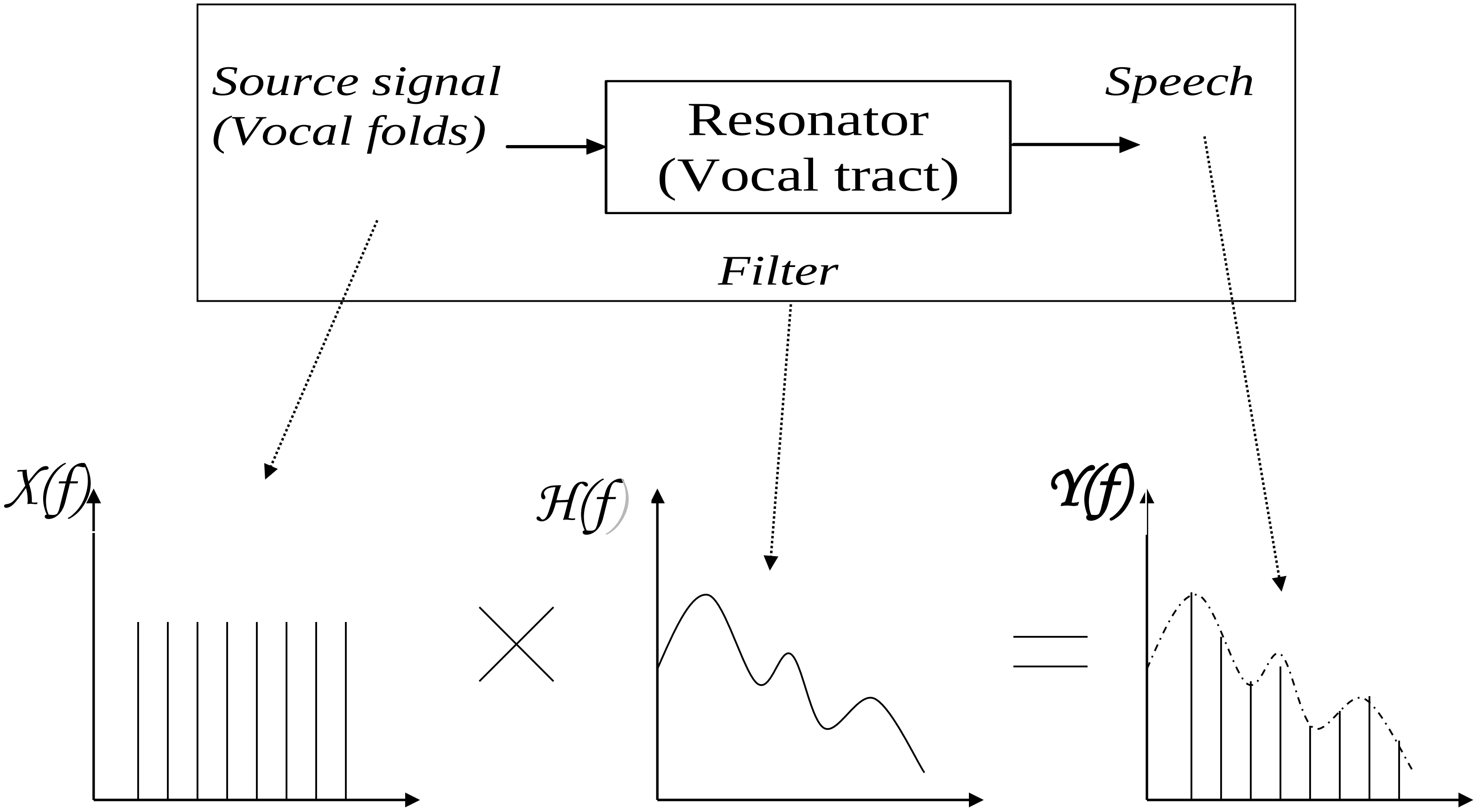
  - How the signal is structured (signal models, decomposition models)



An example of a sound production model
the (speech) source filter model

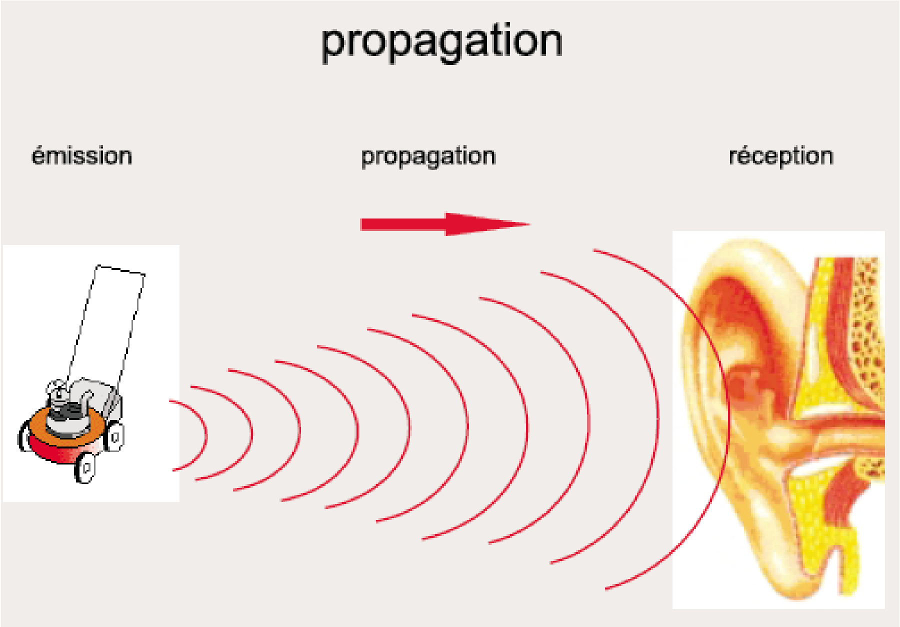


A widely used model: the source filter model



Perception and perception models

- **Sound is a wave (pressure variation)**



- **Decibel:**

$$L_{dB} = 20 \log_{10} \frac{P}{P_0}$$
$$= 10 \log_{10} \frac{I}{I_0}$$

$$I \propto P^2$$

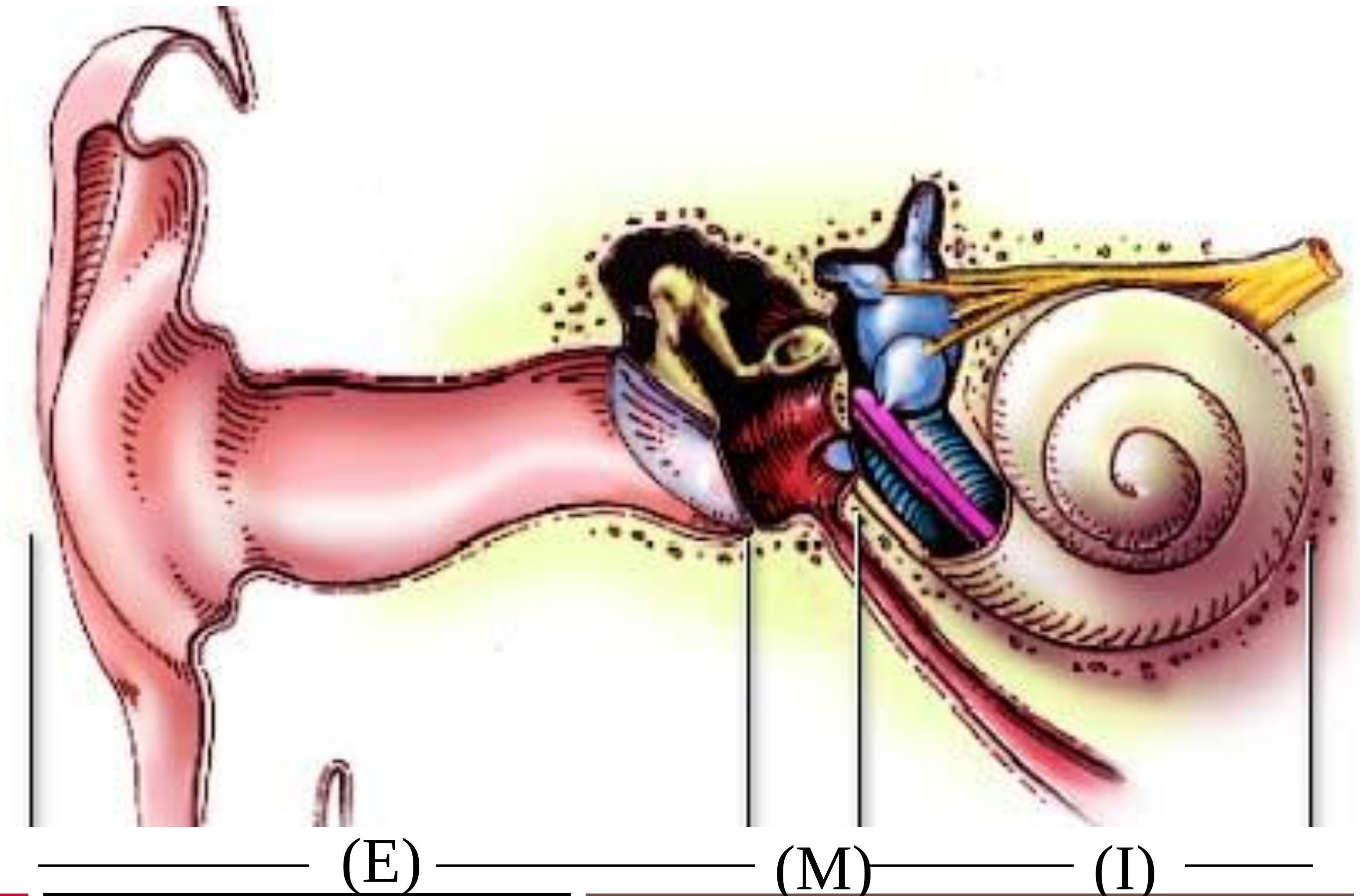
Perceptual scales

- **To each physical scale of sound, we aim to associate a subjective or perceptual scale**

Scale	Unit	Perception of	vocabulary	Physical scale	Unit
Isosonie	Phones	Intensity (same as dB @ 1 kHz)	High / low	-	dB
Sonie	Sones	Intensity/loudness		SPL (Sound pressure Level)	dB
Tonie	Tones/mels	pitch	Bass/Trebble	Frequency	Hz
	???	Timbre	« warm, brillant.. »	???	
Chronie	-	Duration	Short/long	Time	s

Audition

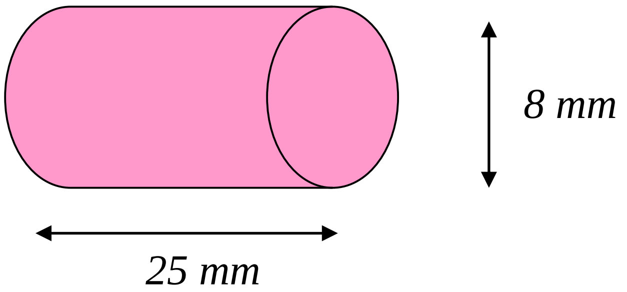
Outer ear (E), middle ear (M) and inner ear (I)



Outer ear

- **The pinna of the ear** performs the following selective filtering:
  - the direction of sound incidence
  - its frequency

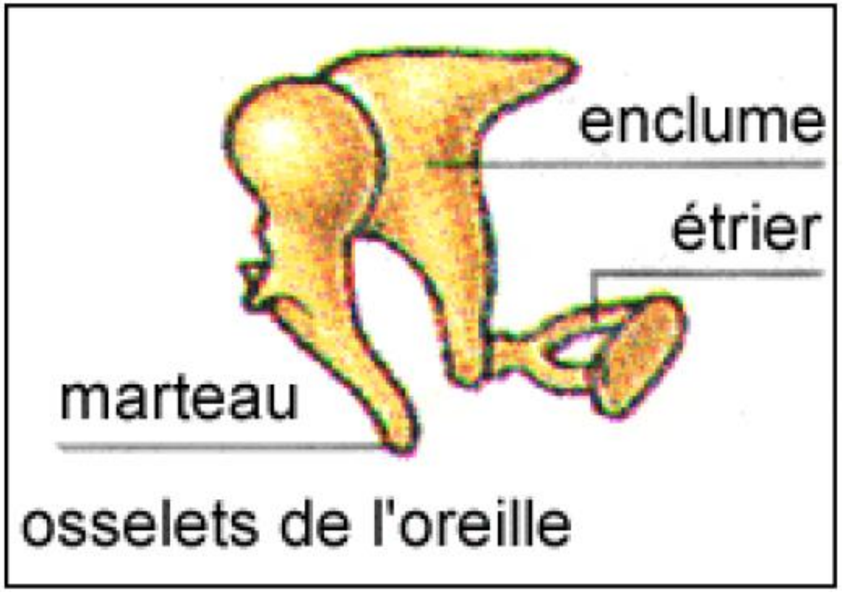
- **The External Auditory Canal** (E.A.C) = waveguide, to the eardrum



- **increased sound intensity at the eardrum**
  - of a few dB between 1.5 and 7 kHz with peaks around 5 kHz (pinna), and around 2 kHz (E.A.C)

Middle ear

- **The middle ear contains three tiny bones:**
  - Hammer (malleus) — 20g
  - Anvil (incus) (25g)
  - Stirrup (stapes) (5g)



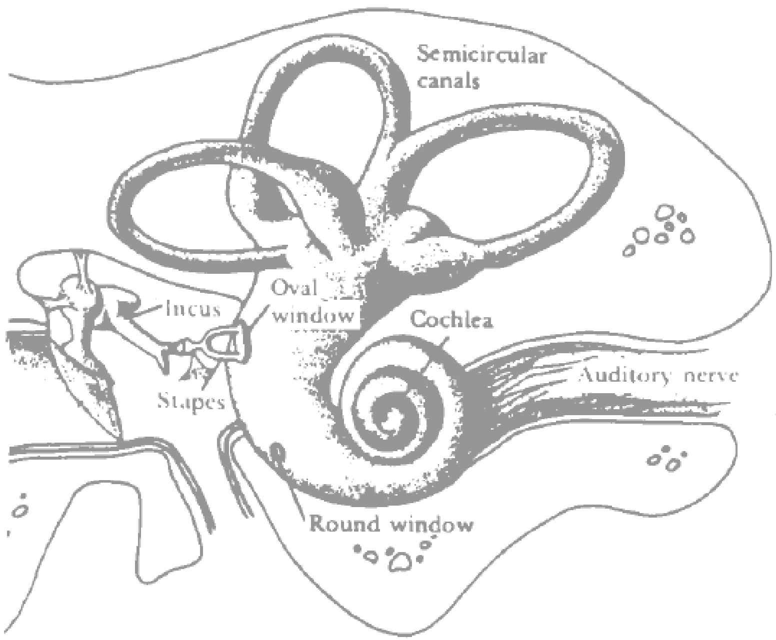
- **Hammer and Anvil attached with ligaments**

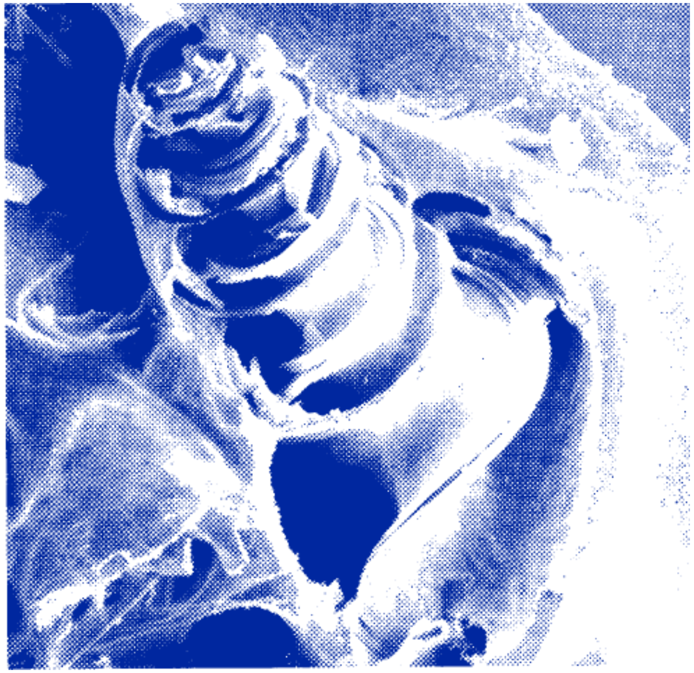
Middle ear: role

- **Amplification and impedence adaptation:**
  - Surface ratio (65 mm²) / (3 mm²) ~= 20
  - Amplification or about 20 to 30 dB between 1 and 10 kHz with a maximum at 4 kHz
    - Without this adaptaiton 99% of energy would have been reflected.
- **Protection of the inner ear:**
  - Mechanical limitation.
  - Stapedious reflex: with two muscles: one is linked to tympani and the other to the stirrus
  - Latency period: about 40ms
  - Though limited effect in amplitude (about -10 dB) and in time (muscular fatigue)

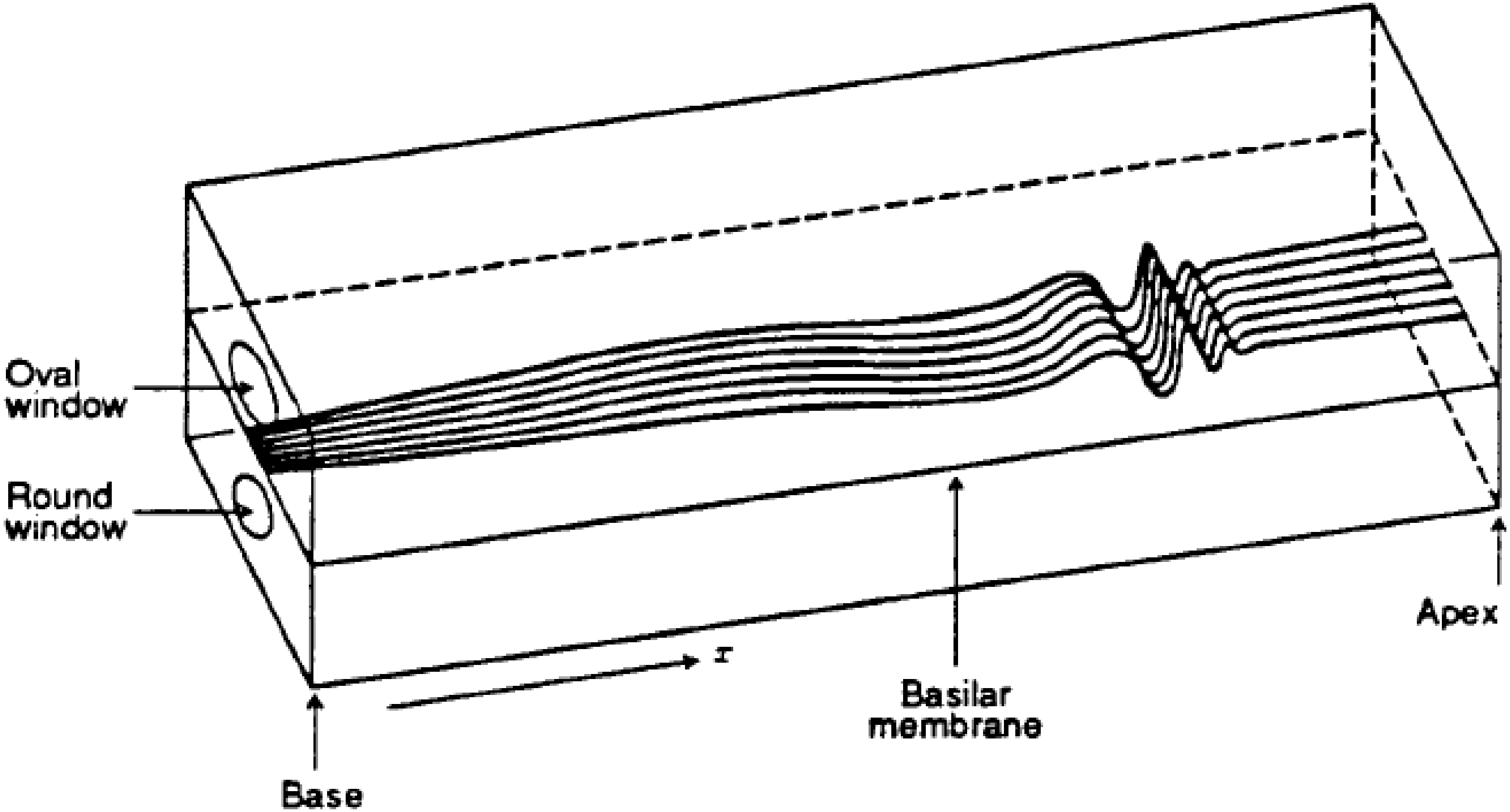
Inner ear

- **Transform mechanical energy in bio-electric energy and in nerve action potentials**

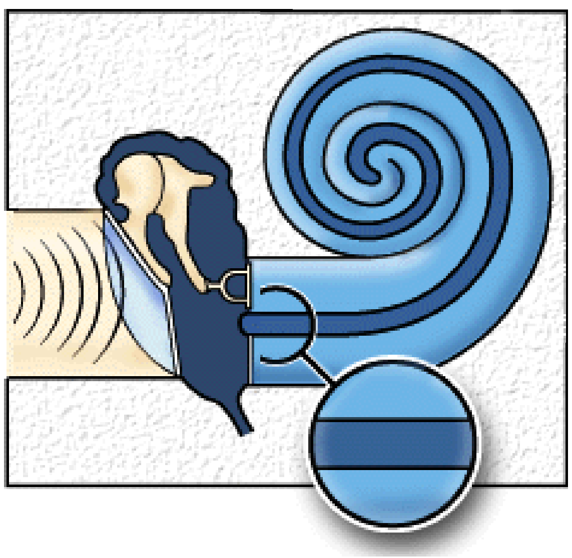




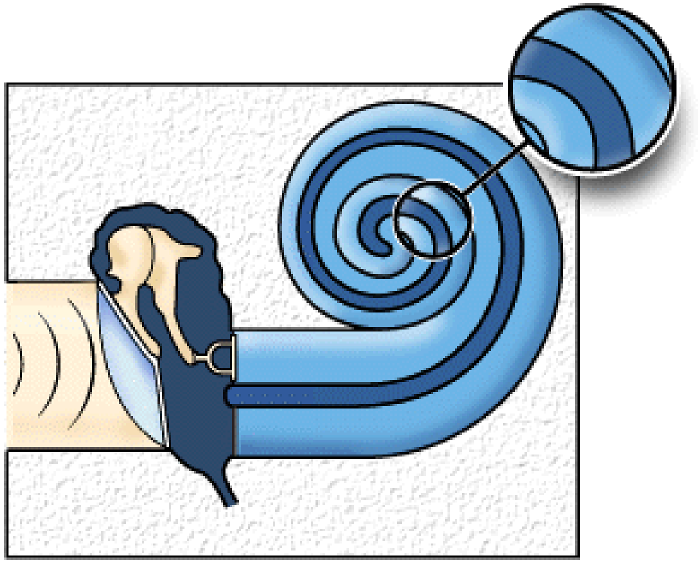
Cochlear canal



Ear: cochlear tonotopy



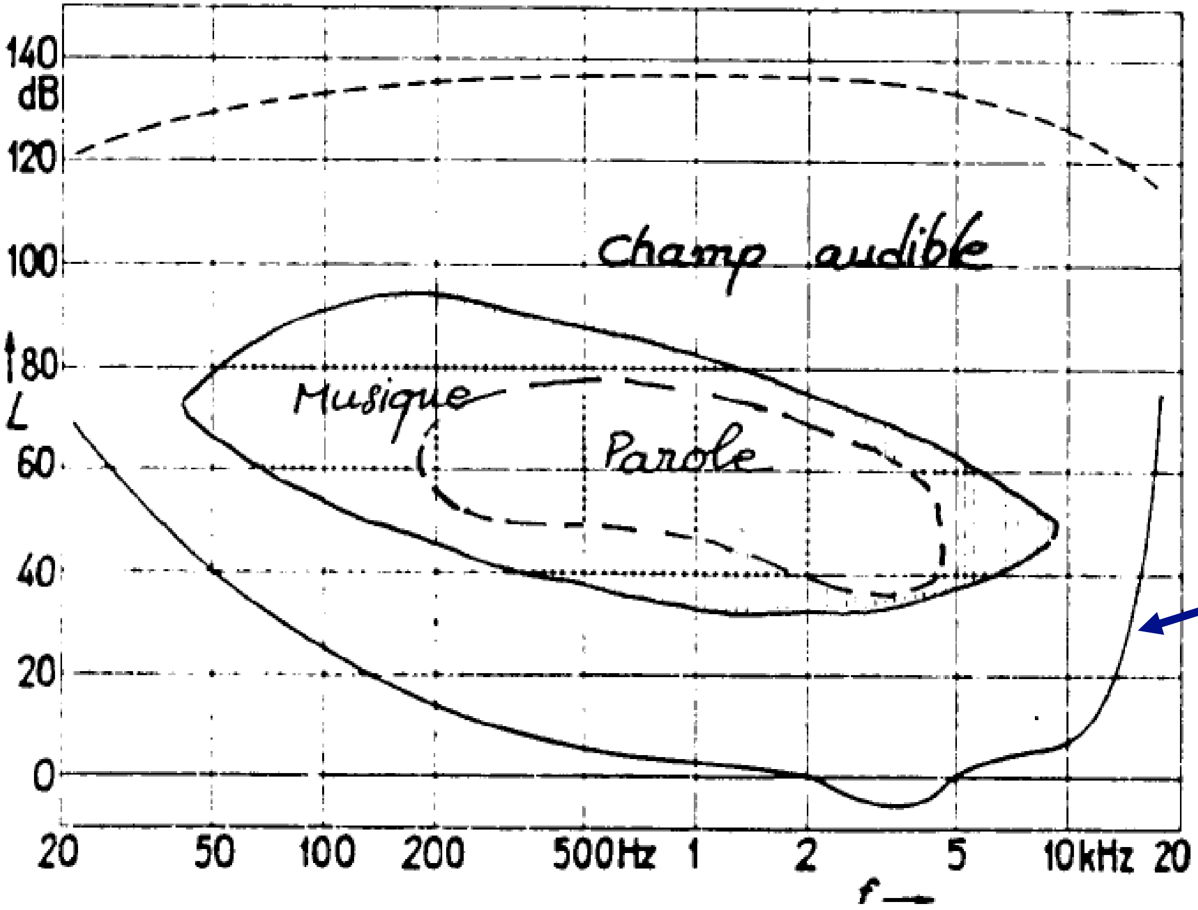
Trebble sound



Bass sound

Audition

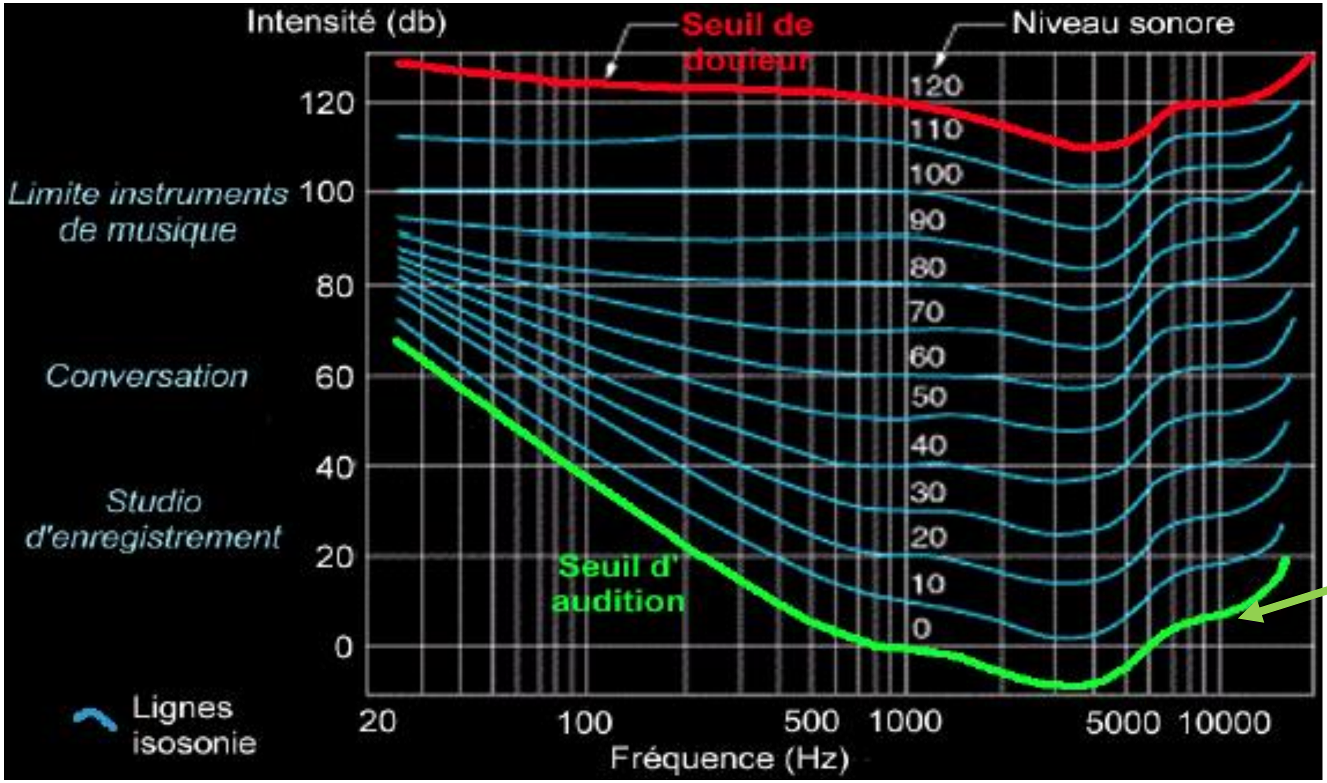
Dynamic of the ear: 120 dB!!



Hearing threshold

Isonosy : the phons

- **N phons <=> intensity of a pure sinusoid at 1 kHz of N dB.**

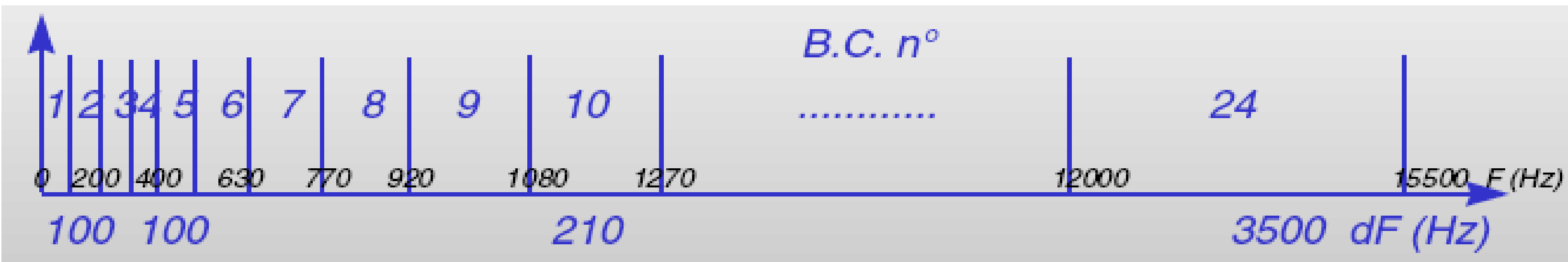


Hearing threshold

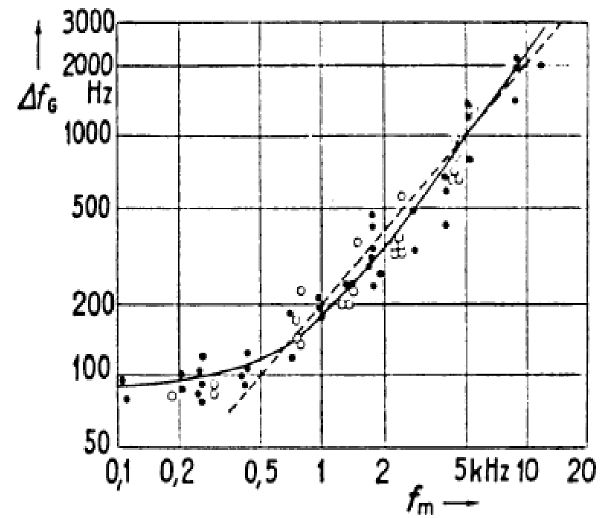
Critical bands

- ***Cochlea reacts as a filter bank***

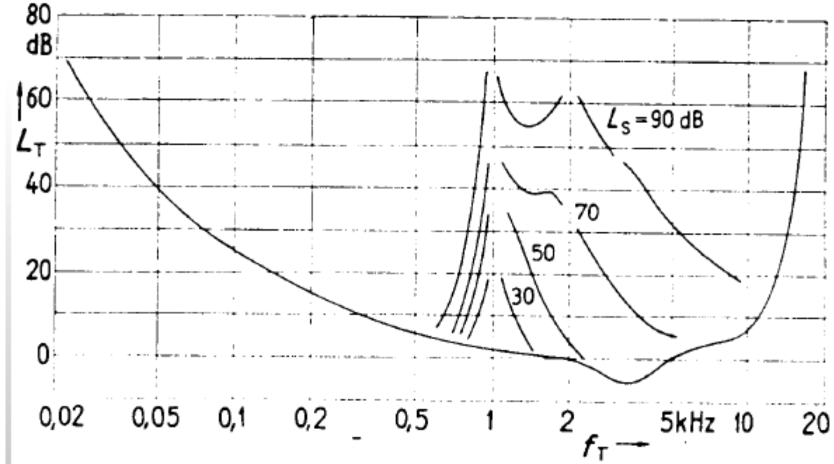
  at 1 kHz the filter has approx. 160 Hz bandwidth



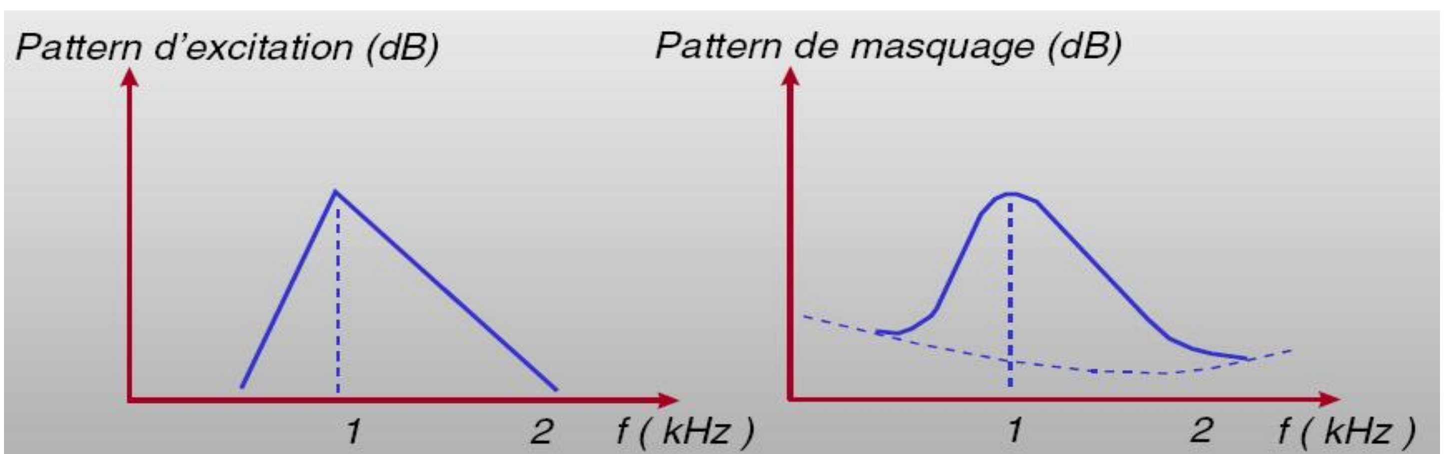
Log-variation of CB bandwidths



Masking properties of pure sinoidal sounds



Interpretation: the loudest sound *mask* the sounds below its *excitation pattern:*



An example of « perceptual » principles used in Audio and MIR

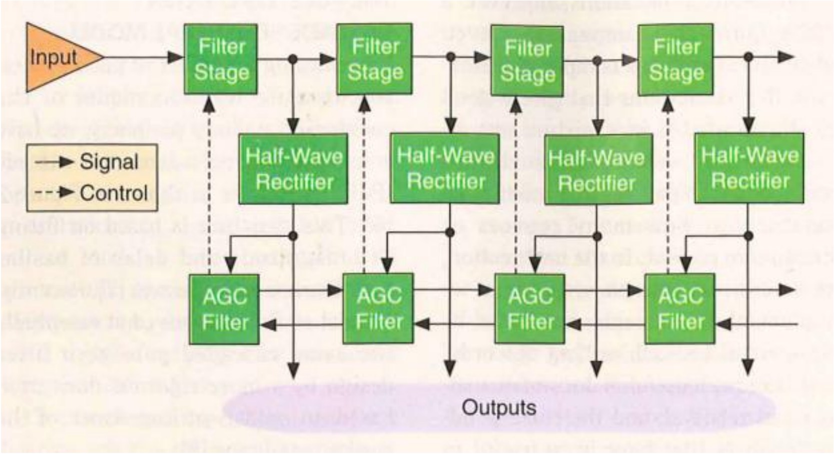
- **« Perceptual » time-frequency representations**
  - Mel-spectrograms
  - CQT (Constant Q transform)
  - Wavelets
  - Gammatone filterbanks

- **« Perceptual » features**
  - MFCC (Mel-frequency Cepstral Coefficients)

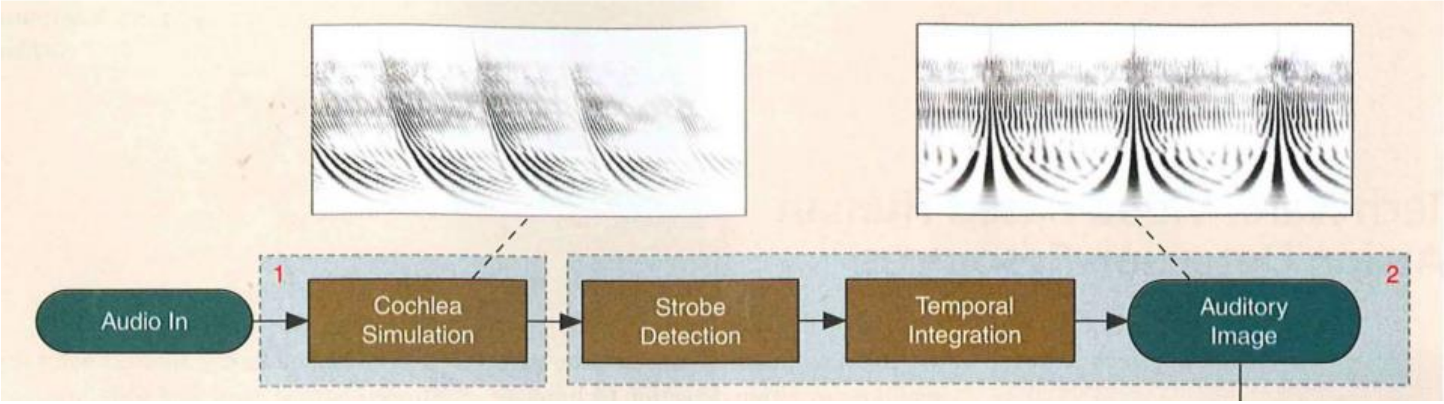
- **Psychoacoustics models**
  - In audio coding (e.g. masking patterns)

An example of a hearing model (Lyon’s)

- **The pole-zero filter cascade model of cochlea**



- **The stabilized auditory image**



R. F. Lyon, "Machine Hearing: An Emerging Field [Exploratory DSP]," in *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 131-139, Sept. 2010, doi: 10.1109/MSP.2010.937498.

Signal models

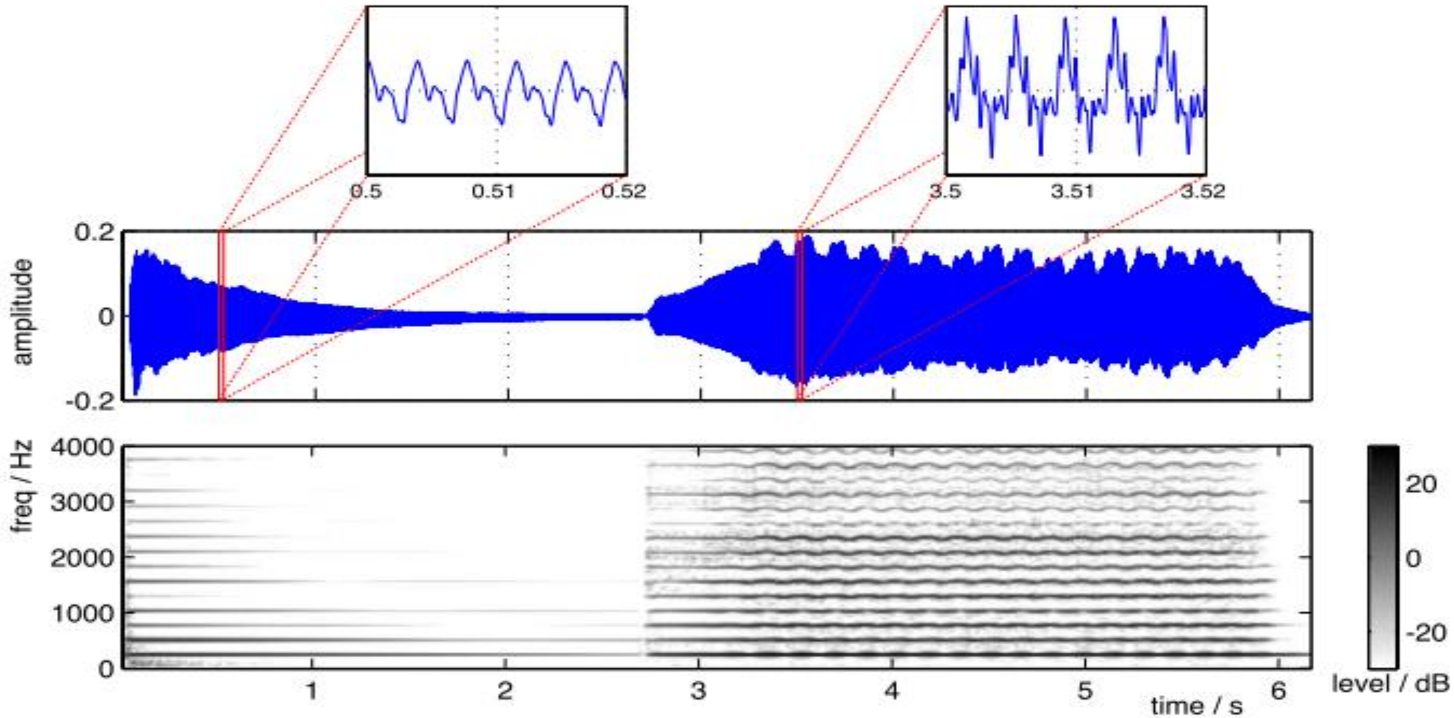
- **Sinusoidal models**
- **Harmonic + noise models**
- **Other « decomposition » models**
  - Sparse representations
  - Non-negative matrix factorization

Audio signal representations

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

Temporal Signal

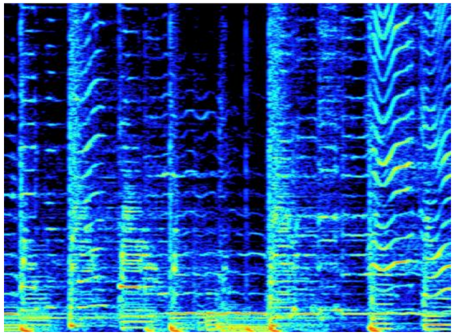
Spectrogram



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)

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**

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

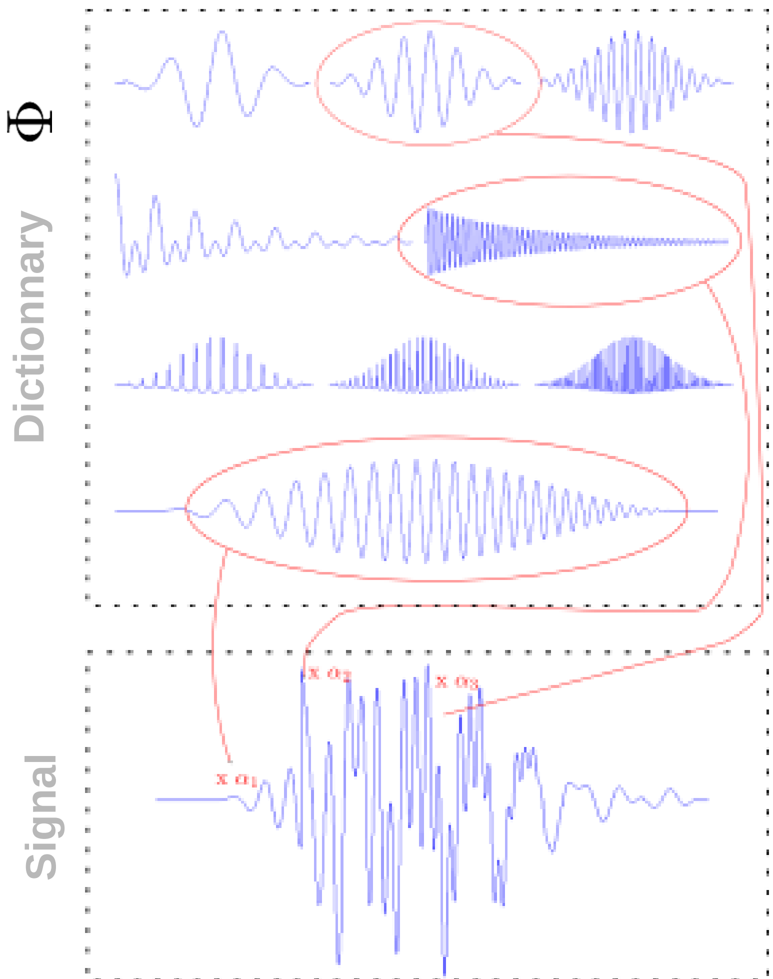
Sparse representation

- **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

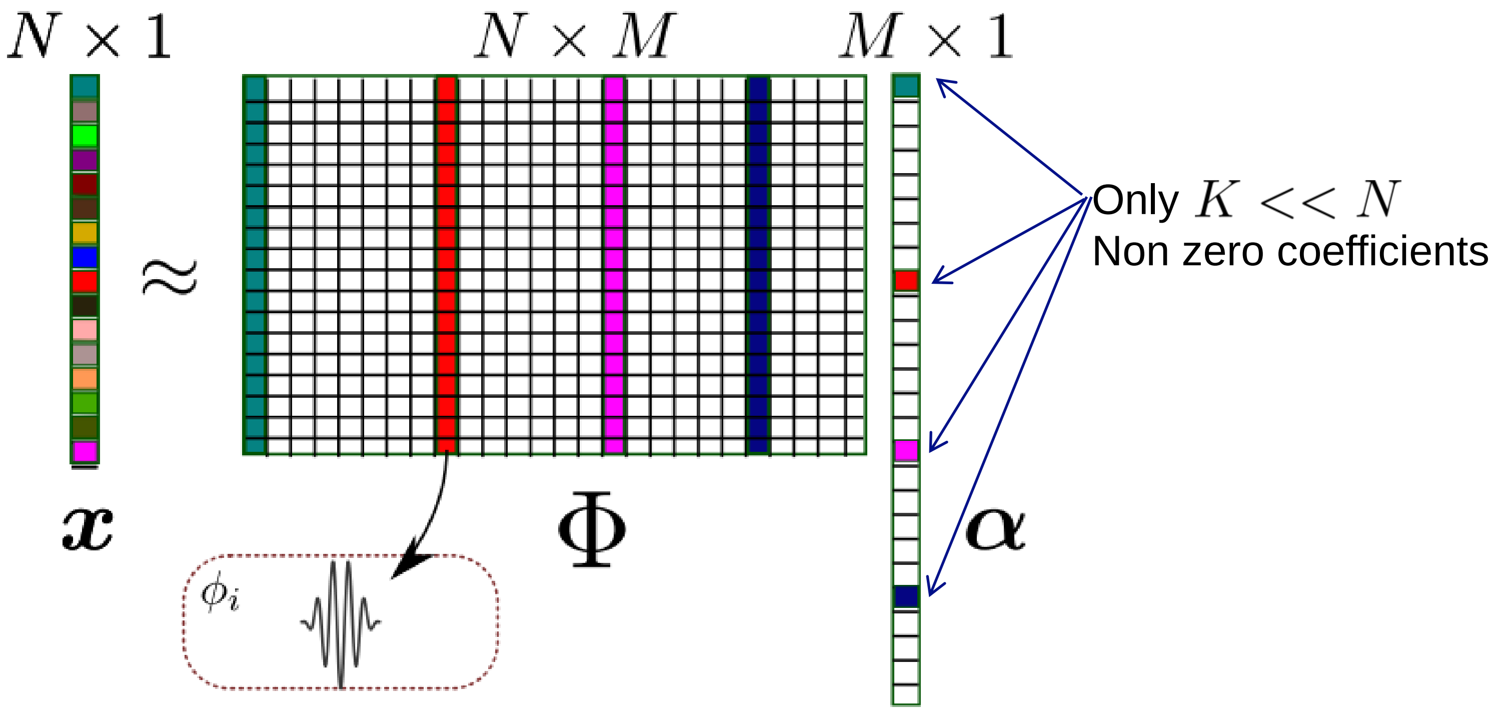
$$\min \|\alpha\|_0 \text{ s.t. } x = \Phi\alpha$$

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*)

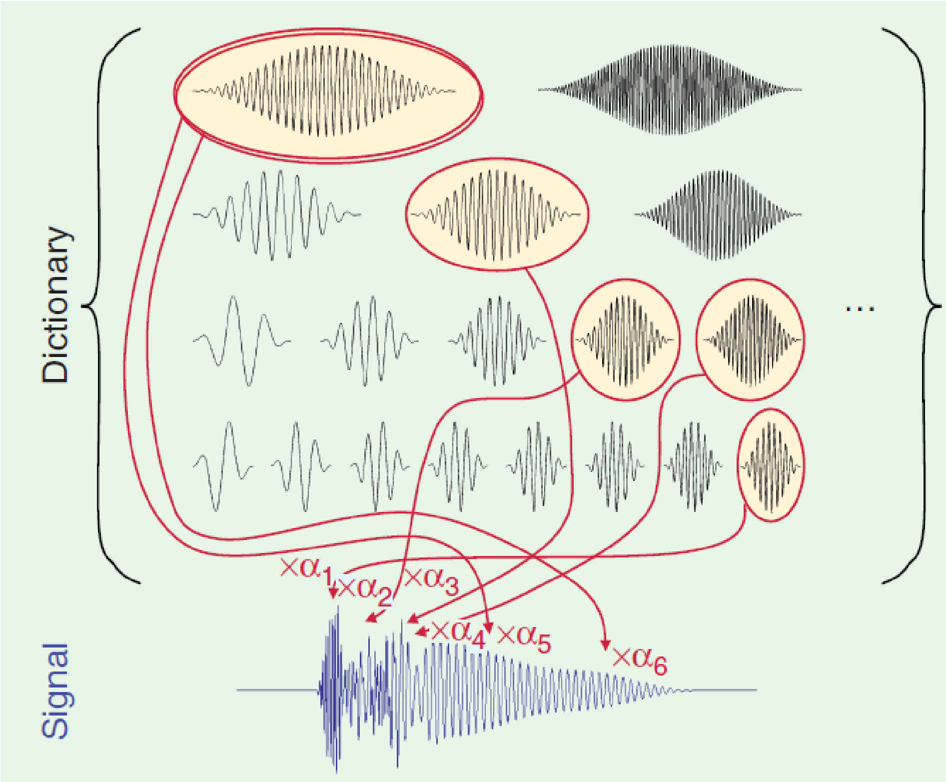
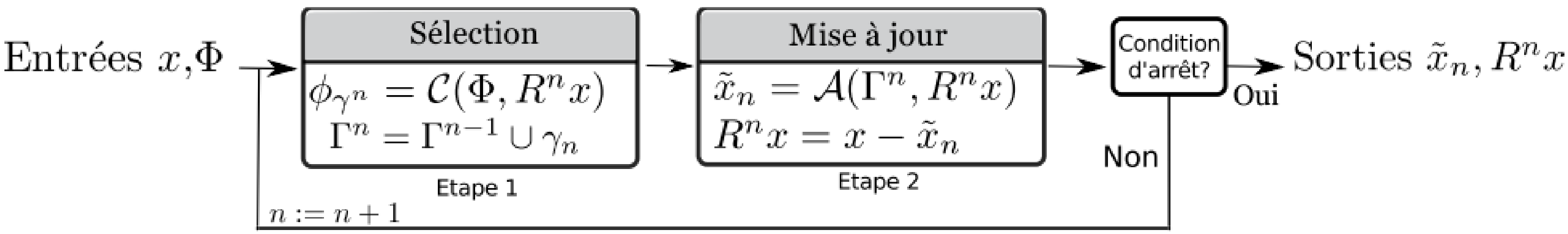


Figure from L. Daudet: *Audio Sparse Decompositions in Parallel,* IEEE Signal Processing Magazine, 2010

Standard Matching pursuit



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

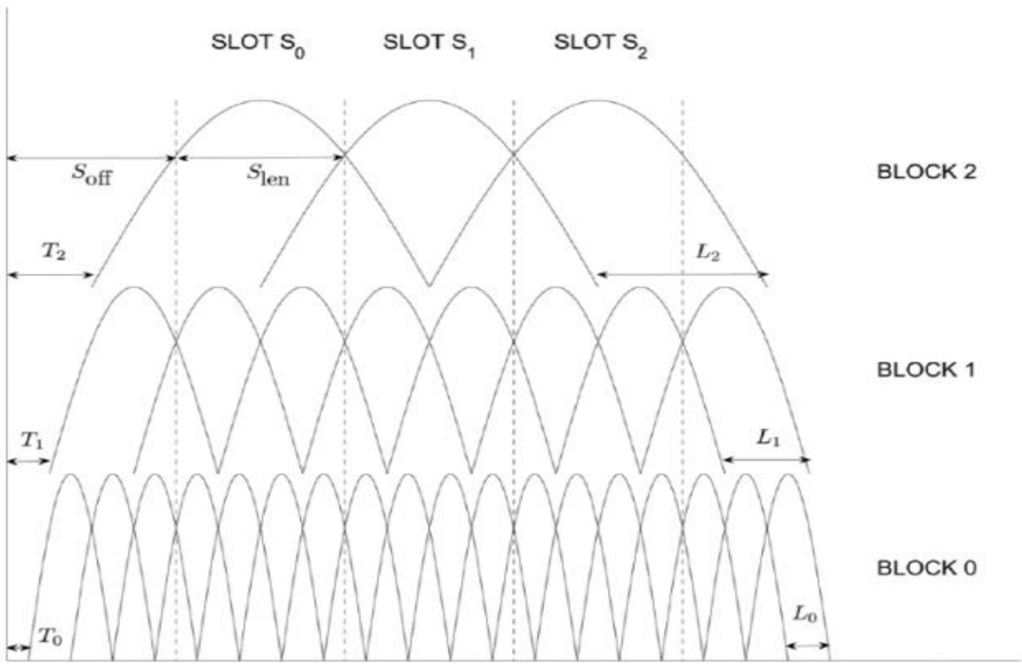
$$\phi_{\gamma^n} = \arg\max_{\phi_i \in \Phi} |\langle R^n x, \phi_i \rangle|$$

- **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 characteristics 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 characteristics atoms of given pitches (note height)

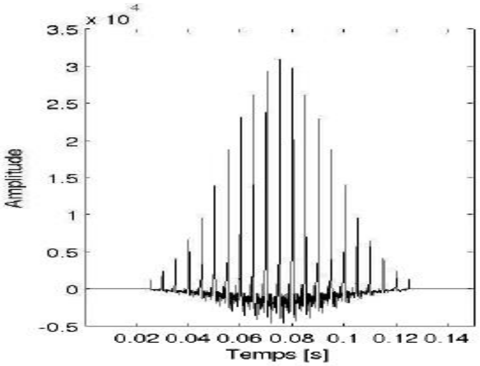
Use in music transcription

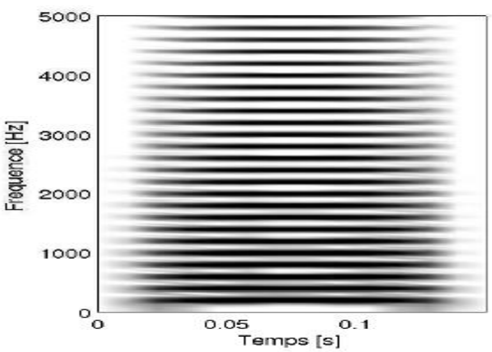
- **Harmonic atoms**

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

  - a_m (resp ϕ_m) amplitudes (resp. phases) of partials
  - s scale parameter
  - u time localisation
  - f_0 (resp c_0) fundamental frequency and chirp rate

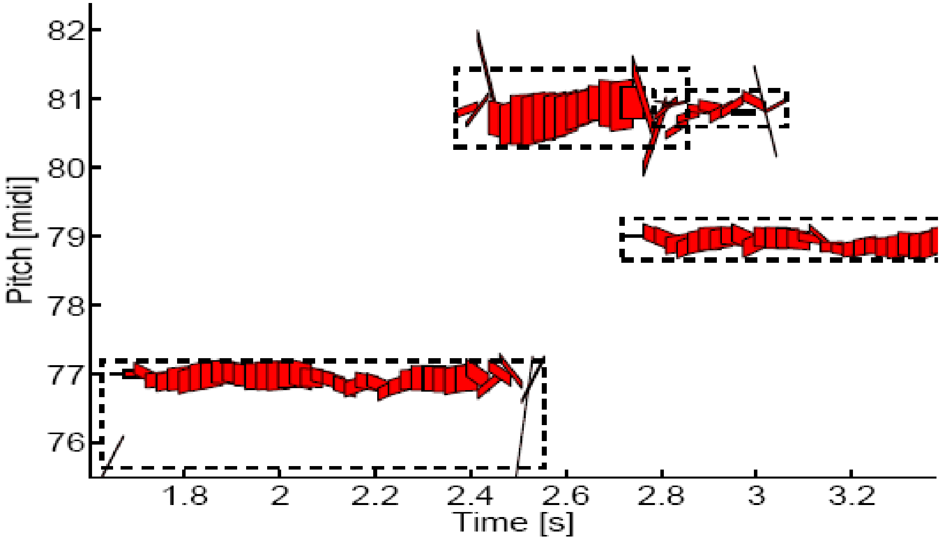
(from P. Leveau & al.2008)



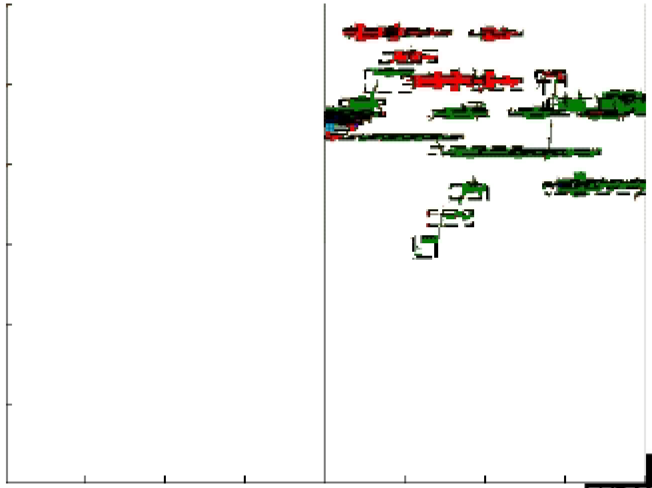


Use in music transcription

- **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 :**

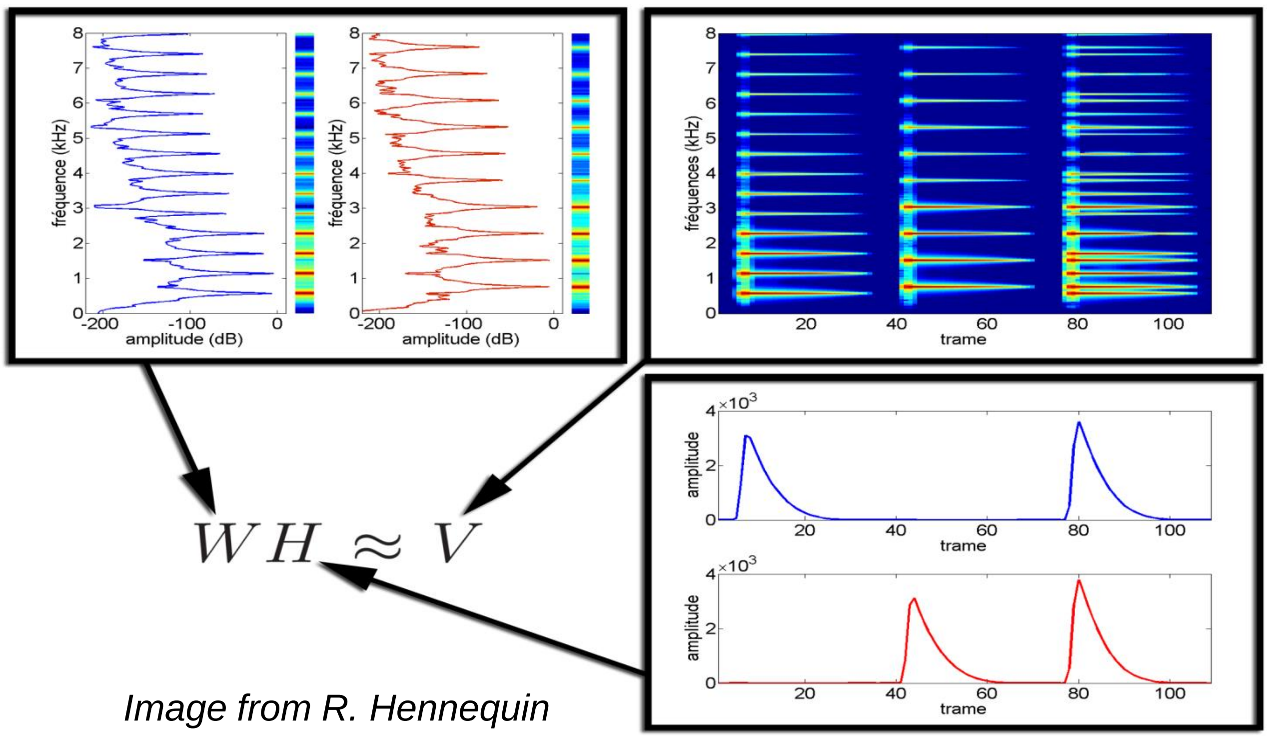


Image from R. Hennequin

Non-negative Matrix Factorization (NMF)

- **The problem**

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} = \hat{\mathbf{V}}$$

- **Solution obtained by minimizing a cost function:**

$$D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{f=1}^{F} \sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn})$$

  - Classic distances/divergences:

$$d_{EUC}(a|b) = \frac{1}{2}(a-b)^2$$

$$d_{KL}(a|b) = a \log\left(\frac{a}{b}\right) - a + b.$$

$$d_{IS}(a|b) = \frac{a}{b} - \log\left(\frac{a}{b}\right) - 1.$$

Non-negative Matrix Factorization (NMF)

- **In the most general case:**
  - The cost function is not convex in W and H
- **But is separately convex for W and H**
- **..towards altenative algorithms**
- **A possible approach (gradient descent):**
  - Compute the differential of the cost function (fixing W or H)
  - Express the gradient as the difference of two positive terms; $\nabla^{+}D$ - $\nabla^{-}D$
  - Obtention of the multiplicative update rules

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\nabla^{-}_{\mathbf{W}} D(\mathbf{V}|\mathbf{WH})}{\nabla^{+}_{\mathbf{W}} D(\mathbf{V}|\mathbf{WH})} \\ \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\nabla^{-}_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})}{\nabla^{+}_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})} \end{cases}$$

Non-negative Matrix Factorization (NMF)

- **Other optimisation approaches**
  - Alternate Least squares, projected gradient, Quasi-newton,…
- **NMF can be expressed in a probabilistic framework**
- **Numerous extension with constrained cost functions**

$$\min_{\mathbf{W},\mathbf{H}} D_r(\mathbf{V}|\mathbf{W}\mathbf{H}) + \lambda D_c(\mathbf{W},\mathbf{H})$$

  - with pitch dependant templates
  - Or enforcing sparsity of W or H
  - …

Audiofingerprint
(Music recognition)

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 .

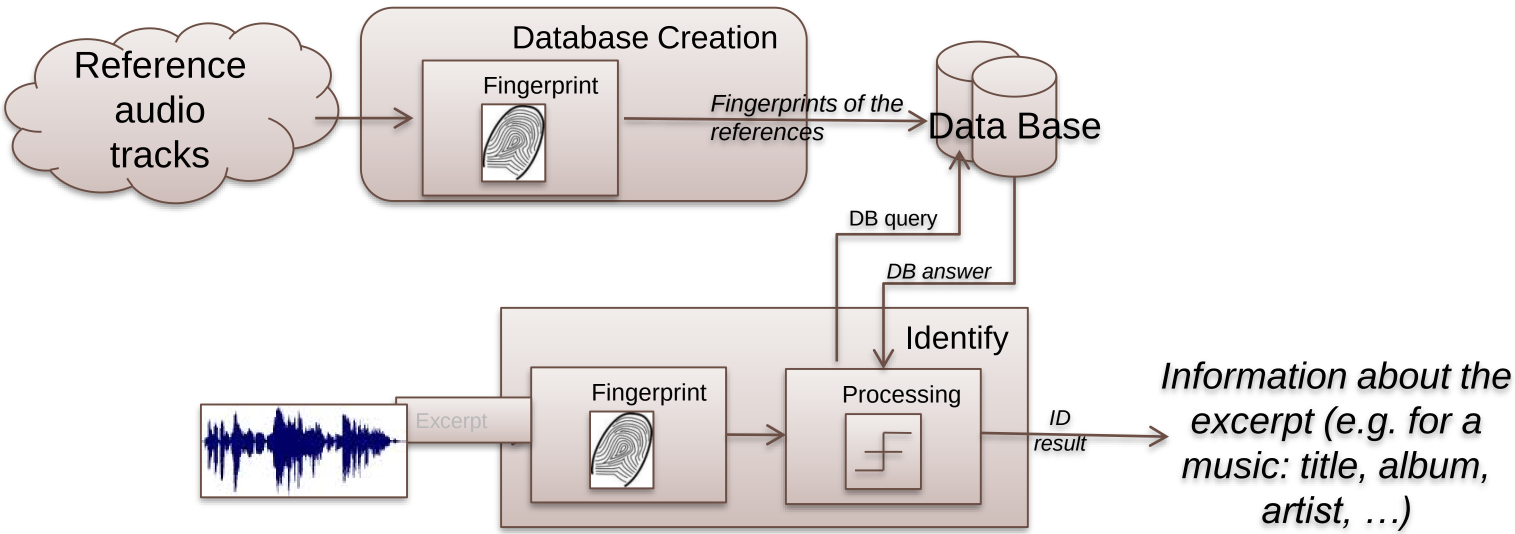
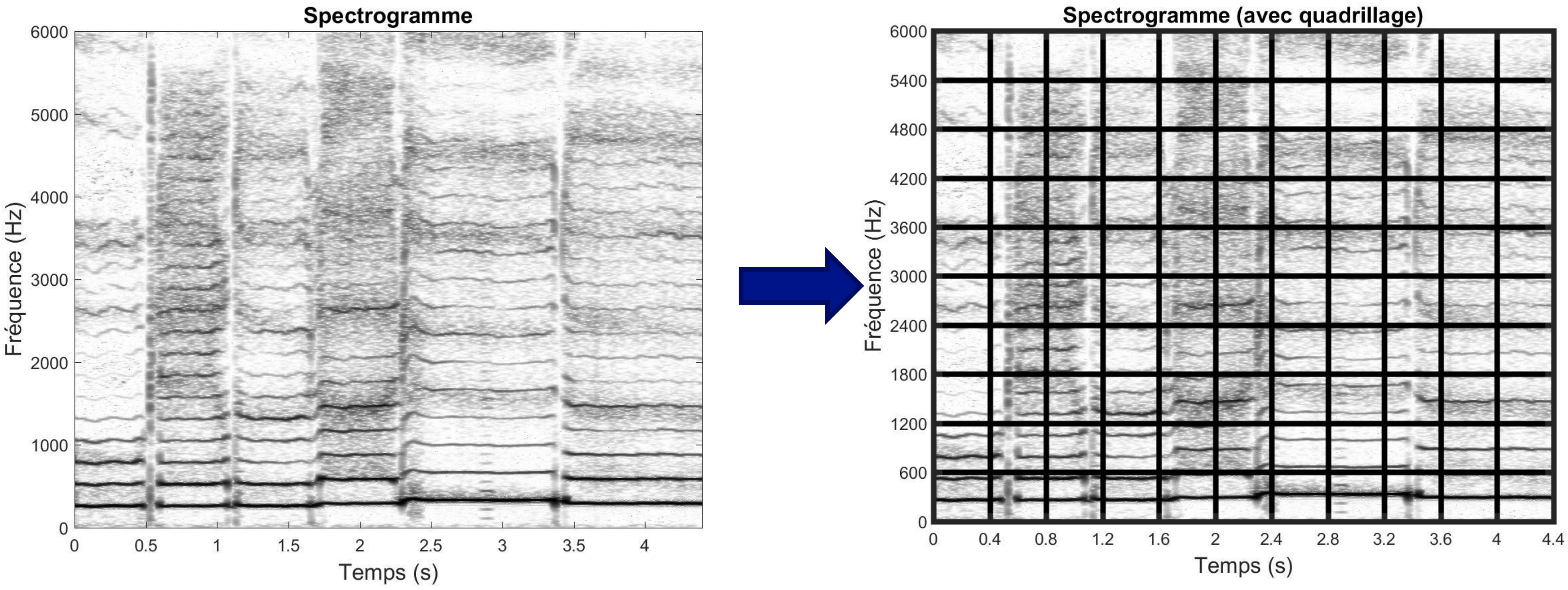


Figure from Sébastien Fenêt

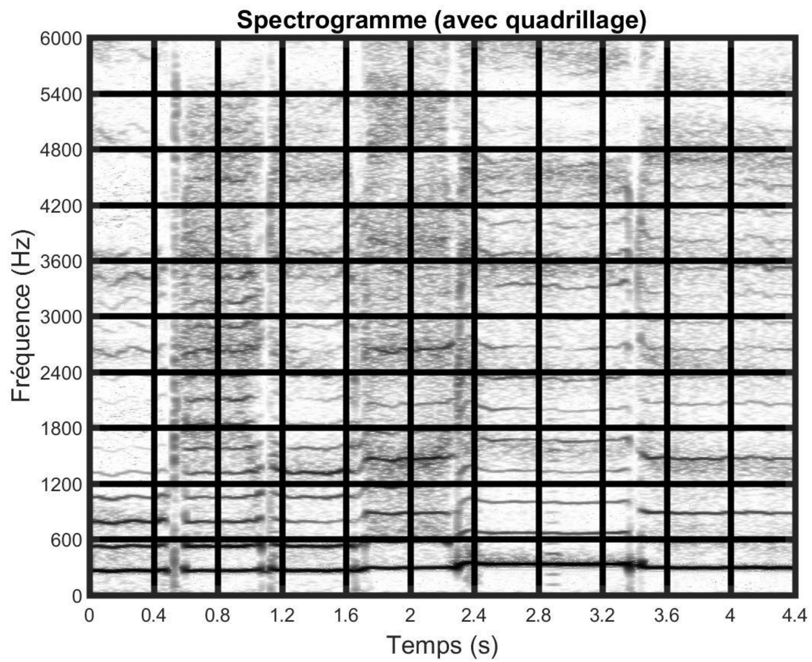
Signal model : from spectrogram to "schematic binary spectrogram"

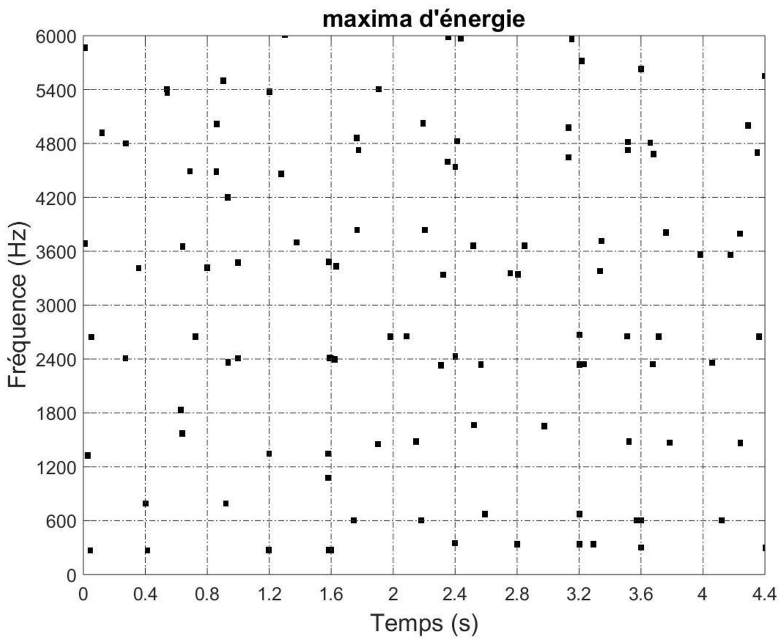
- **1st step: split the spectrogram in time-frequency zones**



Signal model : from spectrogram to “schematic binary spectrogram”

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

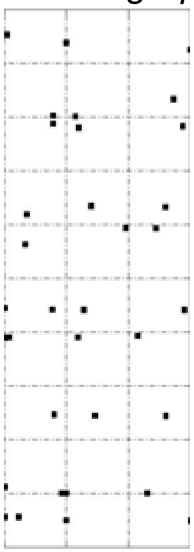




Efficient research strategy

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

Test fingerprint



Potential strategies

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

- Alternative: ?

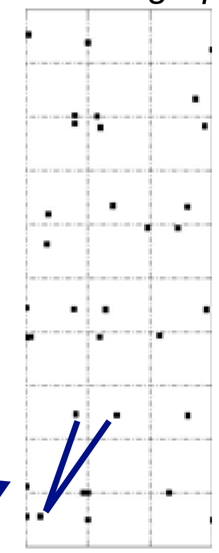
Efficient research strategy

- **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 has 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:**
  - $\Delta T(\text{pair}) = T2 - T1$

- **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.

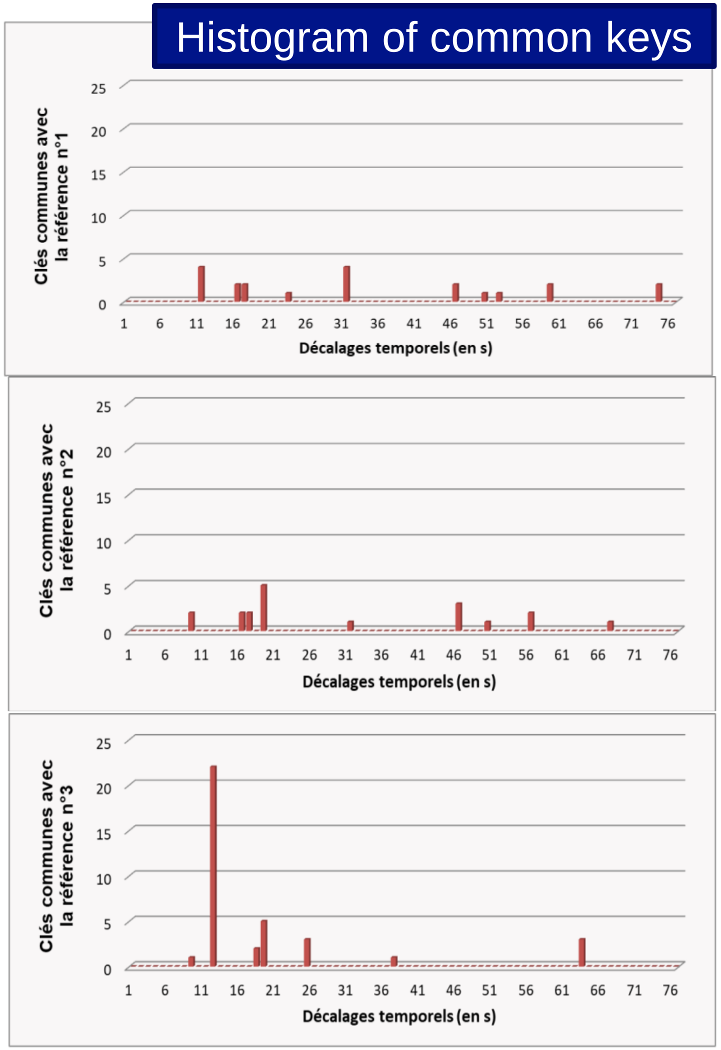
3. **Recognition:** The reference which has the most keys in common at a constant Δt is the recognized recording

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

Reference 1

Reference 2

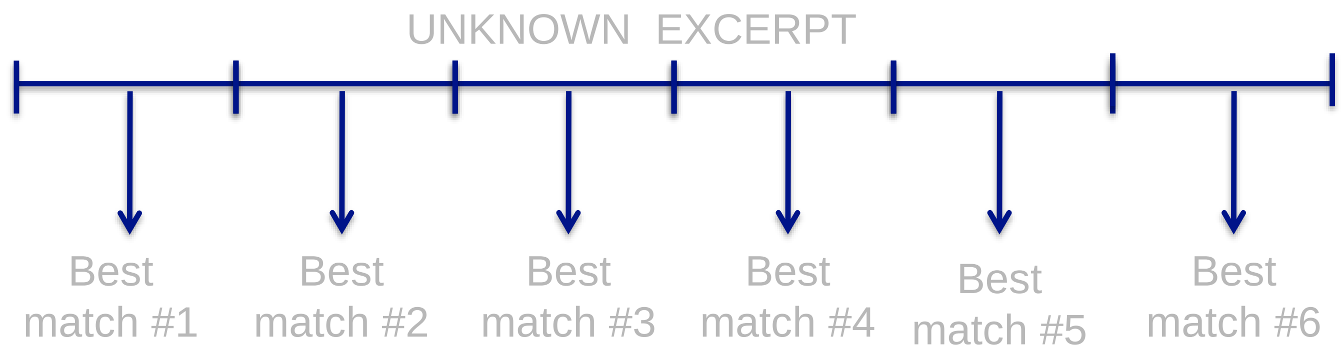
Reference 3



Recognized recording

Detection of an “out-of-base” recording : local decision fusion

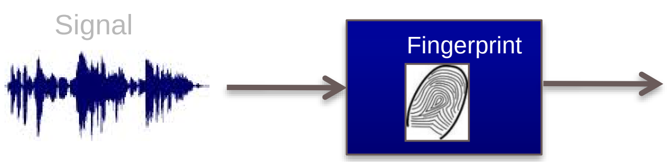
- **The unknown recording is divised in sub-segments**
- **For each sub-segment, the algorithm gives back a best candidate**



- **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 rely on "fingerprints" computation**



$$(key_1, t_1)$$
$$(key_2, t_2)$$
$$\dots$$
$$(key_n, t_n)$$

Signal fingerprints

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

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

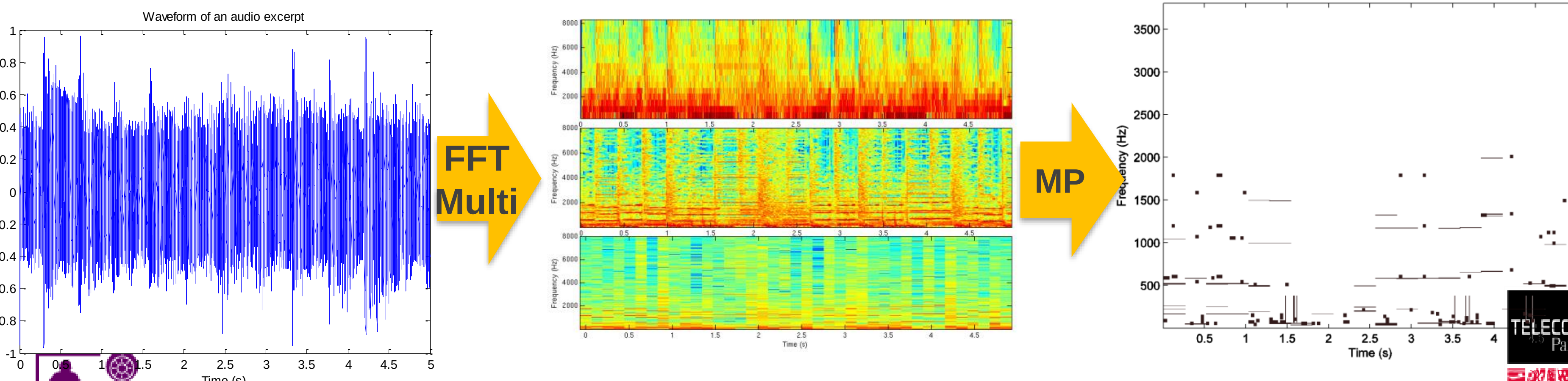
$$\mathcal{M}(\phi_i | \Gamma^n) = 1 - \max_{\gamma \in \Gamma^n} |\langle \phi_i, \phi_\gamma \rangle|$$

Audio fingerprints obtained by MP

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

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

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



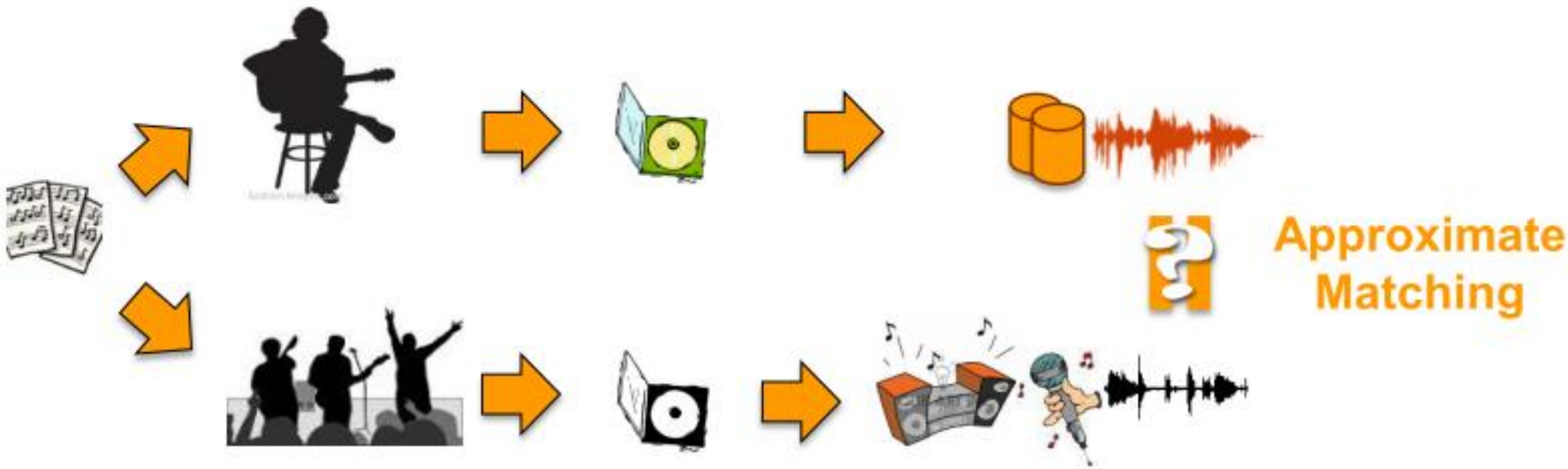
Limitations and other solutions

- **Not robust to time-scale or frequency scale transformations**
  - e.g. change of speed or transposition
  - *Solutions ?*
    - Change of the time-frequency representation (CQT, …) [1]
    - Design of a compact representation more invariant to time-frequency (*geometric hash representations of quadruples of points*) [2]
    - Exploit invariant image features (e.g. SIFT) [3]
    - Exploit evolution of energy in spectral bands [4]

- **Can only recognize the same recording**
  - *Solutions ?*
    - Approach the problem as cover song recognition
    - Approximate matching

[1] S. Fenet, G. Richard, Y. Grenier. A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting. In Proc. of ISMIR, 2011
[2] R. Sonnleitner, G. Widmer, "Robust Quad-Based Audio Fingerprinting," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 409-421, March 2016
[3] X. Zhang & al. SIFT-based local spectrogram image descriptor: a novel feature for robust music identification, "Eurasip Journal on Audio Speech and Music Processing, 2015
[4] M. Ramona and G. Peeters, "Audioprint: An efficient audio fingerprint system based on a novel cost-less synchronization scheme," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2013

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



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

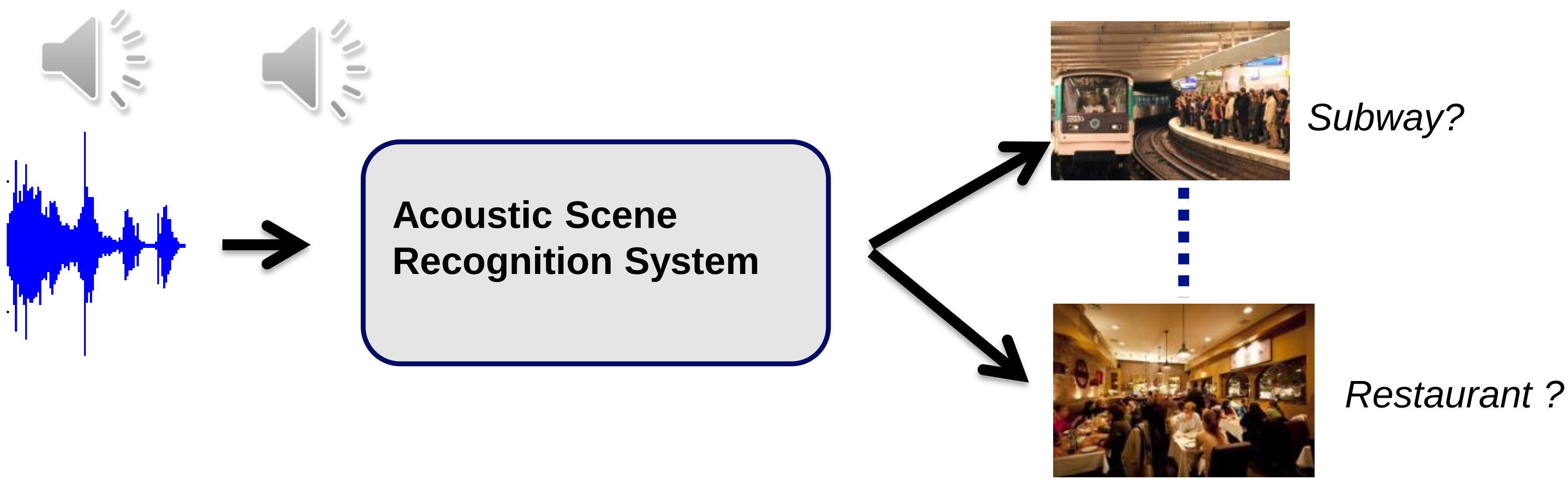
  - For music recommendation, second screen applications, …

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

Machine Listening, DCASE

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 »

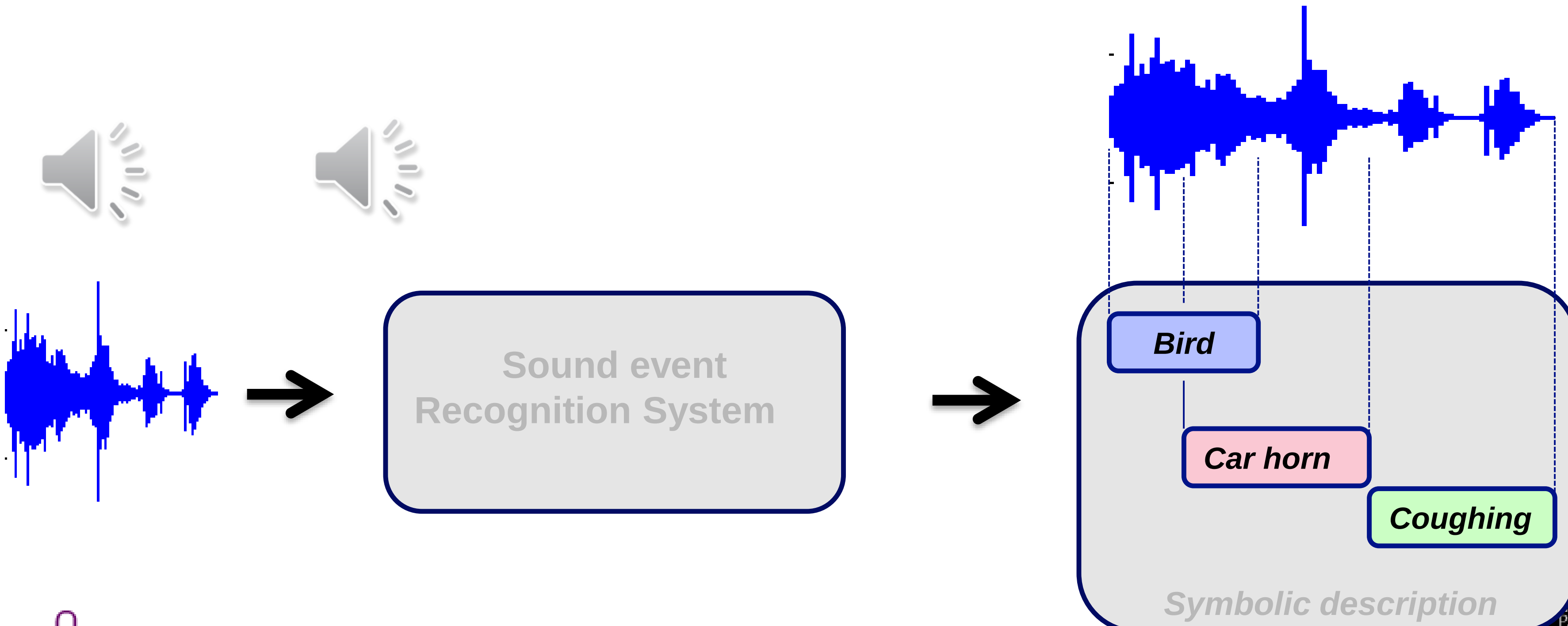


  - Related to CASA (*Computational* Auditory Scene Recognition) and SoundScape cognition (*psychoacoustics*)

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

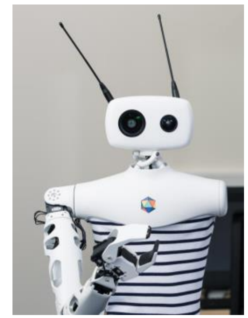
Acoustic scene and sound event recognition

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



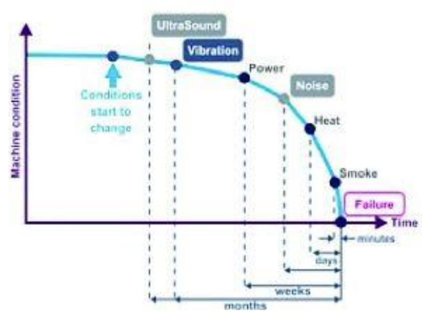
Applications of scene and events recognition

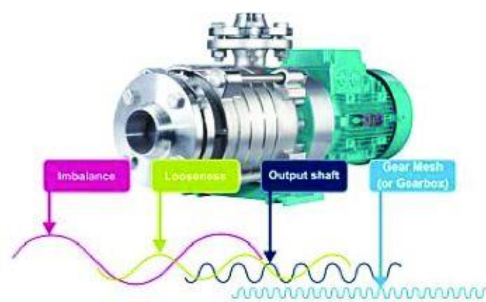
- **Smart hearing aids (Context recognition for adaptive hearing-aids, Robot audition,..)**
- **Security**
- **indexing,**
- **sound retrieval,**
- **predictive maintenance,**
- **bioacoustics,**
- **environment robust speech recognition,**
- **ederly assistance, smart homes**
- **…..**



Listening for Nature

The Rowe Wildlife Acoustic lab





From ST Microelectronics

Some challenges in Audio listening

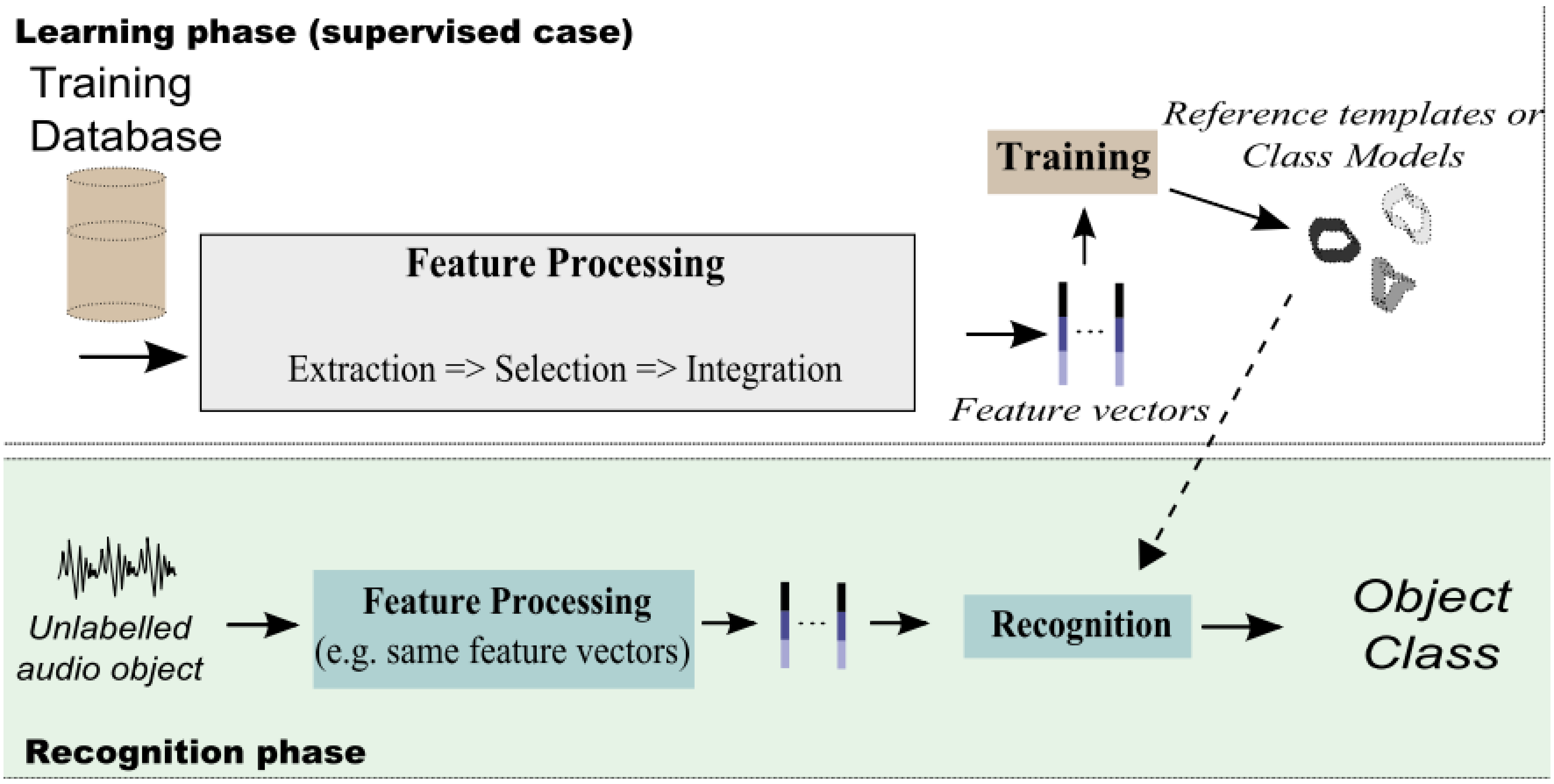
- **Huge databases of recordings and sounds**
- **But …. few recordings are precisely annotated**
  - Ex. *label is « bird song » while the bird song last 2s in a 1 mn recording*
- ***The individual sources composing the scene are rarely available.***
  - *Complexifies the learning paradigm*
- ***In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

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...

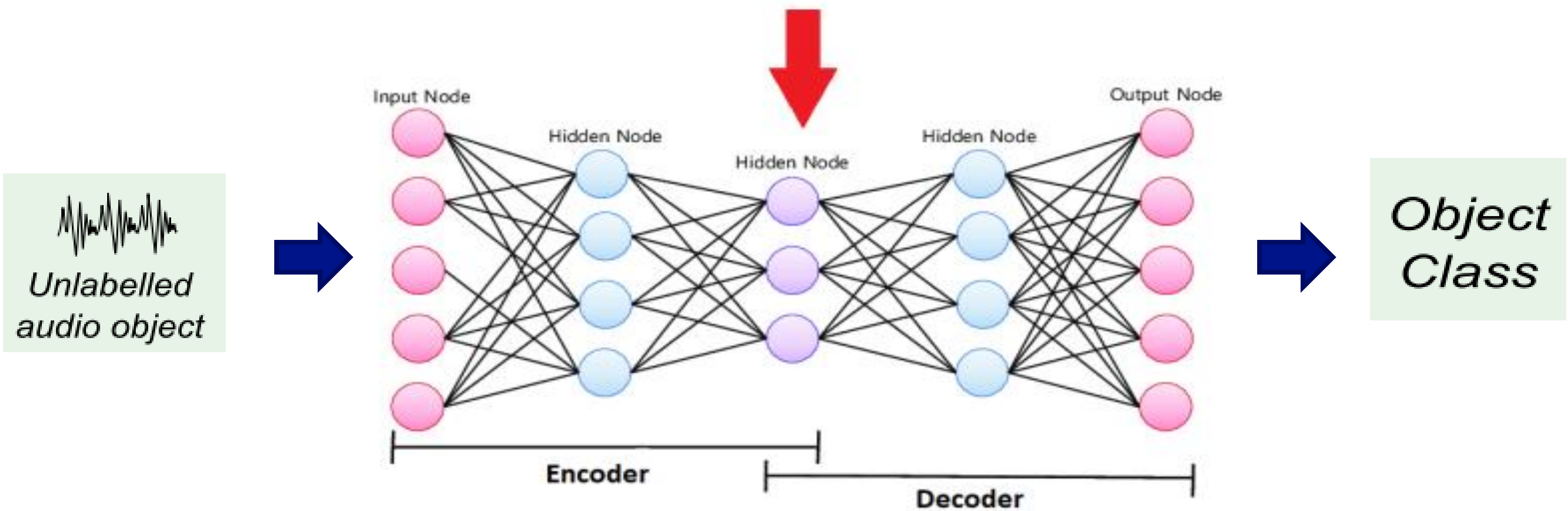
Traditional Classification system



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



DCASE:Detection and Classification of Acoustic Scenes and Events

- **A recent domain:**

  - A (very) brief historical view of
    - speech recognition
    - Music instrument recognition
    - DCASE

An overview of speech recognition

1952: Analog Digit Recognition, 1 speaker
Features: ZCR in 2 bands
Davis, Biddulph, Balashek

1956: Analog 10 syllable recognition
1 speaker
Features: Filterbank (10 filt.)

1962: Digital vowel Recognition, N speakers
Taxonomy consonant/ vowel
Features: Filterbank (40 filt.)
Schotlz, Bakis

1971: Isolated word Recognition,
Few speakers, DTW
Features: Filterbank
Vintsjuk,…

1975-1985: Rule-based Expert systems
1000 words, few speakers
Features: Many…Filterbanks, LPC, V/U detection, Formant center frequencies, energy, « frication » ….
Decision trees, probabilistic labelling
Woods, Zue, Lamel,…

1980: MFCC
Davis, Mermelstein

1980 - : HMM, GMM,
Baker, Jelinek, Rabiner ,…

2009 - :
Mel spectrogram
DNN
Hilton , Dahl…

An overview of music genre/instrument recognition

1964 - : musical timbre perception
Clarke, Fletcher, Kendall…..

1995 - : Music instrument recognition on isolated notes
Kaminskyj, Martin, Peeters ,..

2000 - : First use of MFCC for music modelling
Logan

2001 - : **Genre recognition**
Multiple musically motivated features + GMM
Tzanetakis,…

2004 - : **Instrument recognition (polyphonic music)**
Multiple timbre features + GMM, SVM, …
Eggink, Essid,…

2007 - : **Instrument recognition : exploiting source separation, dictionary learning**
NMF, Matching pursuit,…
Cont, Kitahara,Heittola, Leveau, Gillet, …

2009 - : instrument recognition
DNN, …
Hamel, Lee …

An overview of Acoustic scene/Events recognition

1980 - : HMM, GMM in speech/speaker recognition,
Baker, Jelinek, Rabiner ,…

1983,1990 Auditory Sound Analysis (Perception/Psychology):
Scheffer, Bregman, …

1993 Computational ASA
(Audio stream segregation)
Use of auditory periphery model
Blackboard model ('IA)
M. Cook & al.

1997 Acoustic scenes recognition
5 classes of sound
PLP + filter bank features,
RNN or K-NN
Sahwney & al.

1998 Acoustic scene recognition
Use of HMM
Clarksson &al.

2003: Acoustic scene recognition
MFCC+HMM+GMM
Eronen & al.

2005: Event recognition
MFCC+ other feat.
Feature reduction by PCA
GMM
Clavel & al.

From 2009: Scene/Event recognition
More specific methods exploiting sparsity, NMF, image features …
Chu & al, Cauchy & al,…

2014 - :
DNN for acoustic event recognition
Gencoglu & al, ...

DCASE:Detection and Classification of Acoustic Scenes and Events

- **A domain of growing interest: https://dcase.community/**

DCASE2022 WORKSHOP

November 2022, Nancy, France

- A yearly workshop

DCASE2022 CHALLENGE

Tasks

- Low-Complexity Acoustic Scene Classification
- Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques
- Sound Event Localization and Detection Evaluated in Real Spatial Sound Scenes
- Sound Event Detection in Domestic Environments
- Few-shot Bioacoustic Event Detection
- Automated Audio Captioning and Language-Based Audio Retrieval

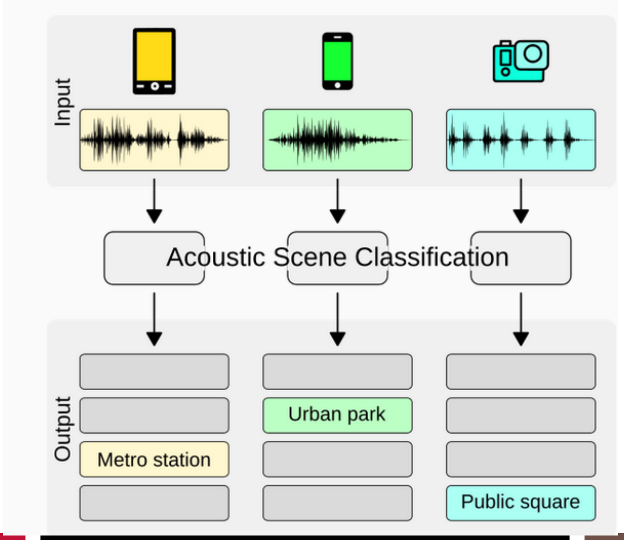
DCASE
Acoustic scene classification (ASC)

- **Goal**: to classify a test recording into one of the provided predefined classes that characterizes the recording environment

- **Two subtasks in the challence DCASE 2021 (1/2)**

Devices A Task 1

ASC with Multiple Devices (10 classes)
Classification of data from multiple devices (real and simulated)



Dataset : TAU Urban Acoustic Scenes 2020 Mobile.
- recordings from 12 cities
- 10 different acoustic scenes
- 4 different devices.

+ synthetic data for 11 mobile devices was created based on the original recordings.

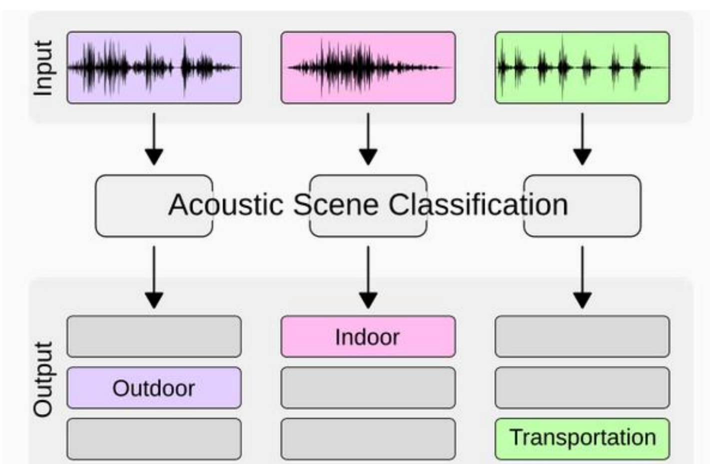
DCASE
Acoustic scene classification (ASC)

- **Goal**: to classify a test recording into one of the provided predefined classes that characterizes the recording environment

- **Two subtasks in the challence DCASE 2021 (2/2)**

Complexity
B
Task 1

low complexity ASC into three major classes: indoor, outdoor, and transportation.



Dataset : TAU Urban Acoustic Scenes 2020 3Class

- recordings from 12 cities
- 10 different acoustic scenes (*but 3 meta classes*)
- 1 device.

+ synthetic data for 11 mobile devices was created based on the original recordings.

DCASE: *Acoustic scene classification (ASC)*
Task 1.B: low complexity

System complexity requirements

- Classifier complexity limited to :

    - 500KB size for the **non-zero parameters**

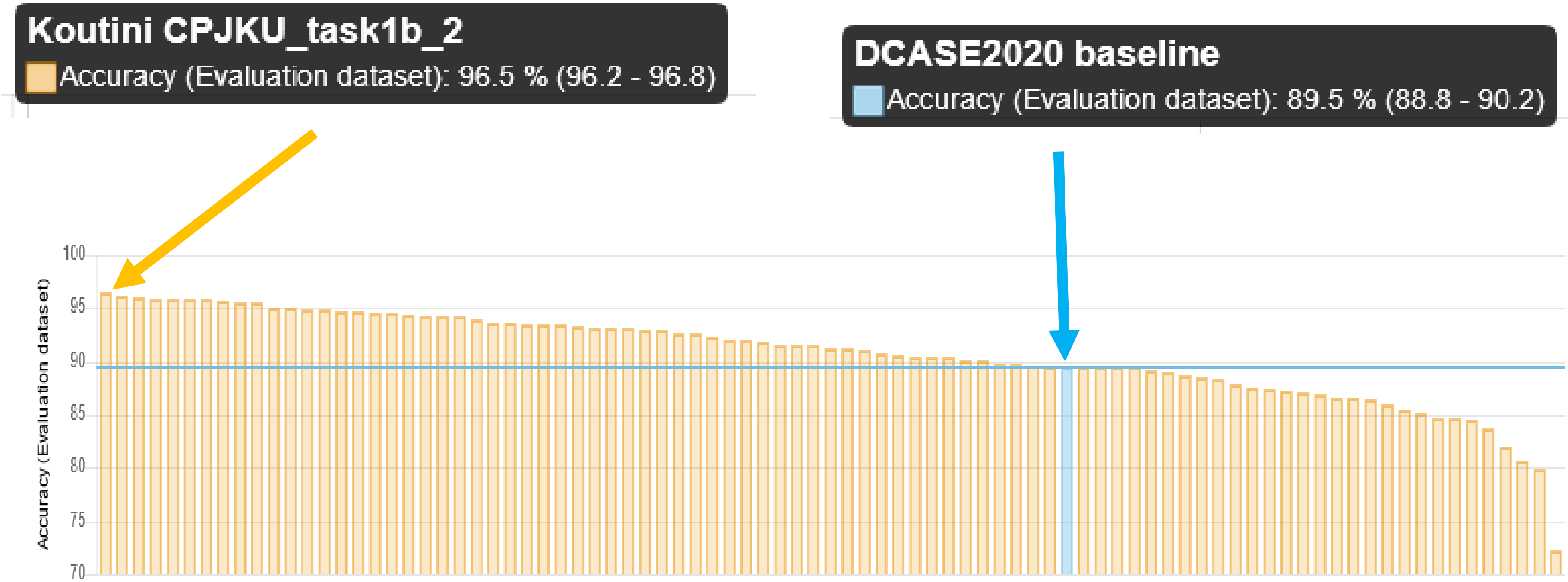
    (excluding layer 1 if it is a feature extraction layer, and batch normalization layers).
    but including the parameters of the network generating the embeddings if used (e.g VGGish, OpenL3, or EdgeL3),

Evaluation:

- macro-average accuracy (average of the class-wise accuracies)

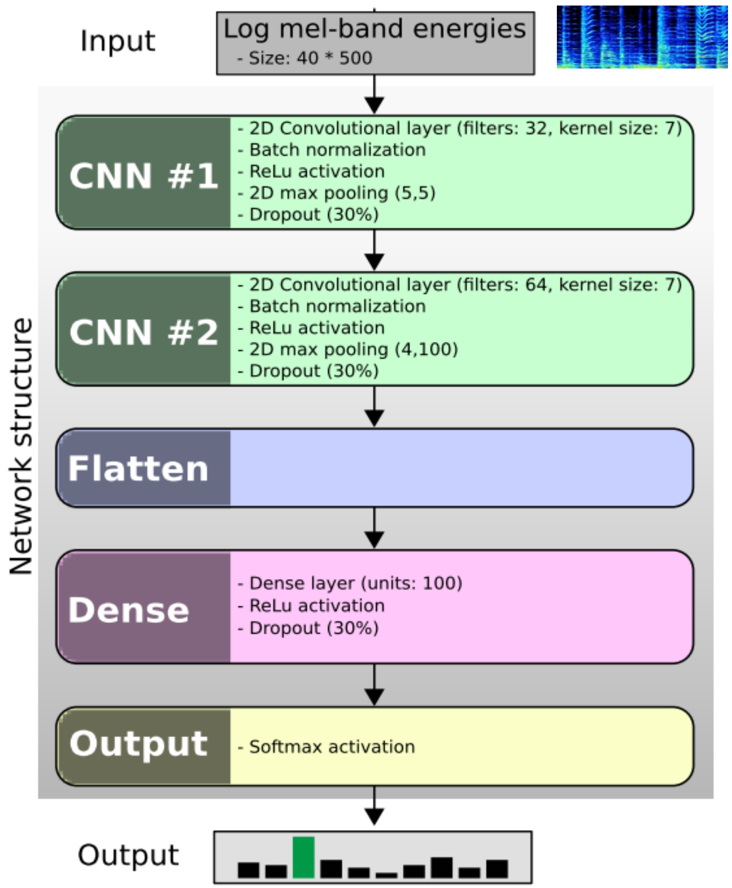
DCASE: *Acoustic scene classification (ASC) Task 1.B: low complexity*

- **Performances (DCASE 2020)**



DCASE: *Task 1.B: low complexity Baseline 2020 system*

- **Parameters (model size = 450 kB)**
- **Audio features:**
  - Log mel-band energies (40 bands), analysis frame 40 ms (50% hop size)
- **Neural network:**
  - Input shape: 40 * 500 (10 seconds)
  - Architecture:
    - CNN layer #1
      - 2D Convolutional layer (filters: 32, kernel size: 7) + Batch normalization + ReLu activation
      - 2D max pooling (pool size: (5, 5)) + Dropout (rate: 30%)
    - CNN layer #2
      - 2D Convolutional layer (filters: 64, kernel size: 7) + Batch normalization + ReLu activation
      - 2D max pooling (pool size: (4, 100)) + Dropout (rate: 30%)
    - Flatten
    - Dense layer #1
      - Dense layer (units: 100, activation: ReLu)
      - Dropout (rate: 30%)
    - Output layer (activation: softmax)
  - Learning: 200 epochs (batch size 16), data shuffling between epochs
  - Optimizer: Adam (learning rate 0.001)



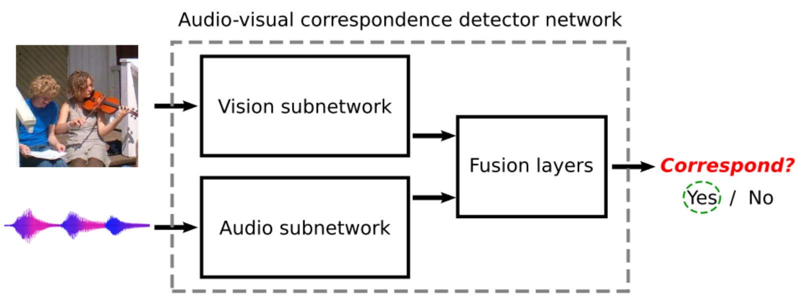
A. Mesaros, T. Heittola, and T. Virtanen. *A multi-device dataset for urban acoustic scene classification.* In Proc. of DCASE 2018.
T. Heittola & al. *Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions.* In Proc. of the DCASE 2020 Workshop

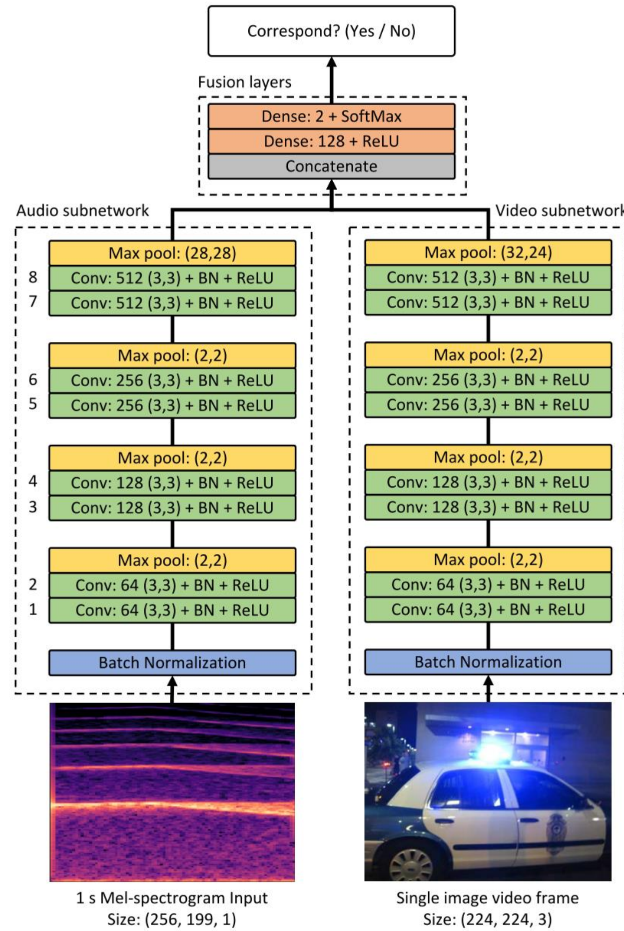
Comparasion with other baselines

System	Accuracy	Log loss	Audio embedding	Acoustic model	Total size
DCASE2020 Task 1 Baseline, Subtask A *OpenL3 + MLP (2 layers, 512 and 128 units)*	89.8 % (± 0.3)	0.266 (± 0.006)	17.87 MB	145.2 KB	19.12 MB
Modified DCASE2020 Task 1 Baseline, Subtask A *EdgeL3 + MLP (2 layers, 64 units each)*	88.9 % (± 0.3)	0.298 (± 0.003)	840.6 KB	145.2 KB	985.8 KB
DCASE2020 Task 1 Baseline, Subtask B *Log mel-band energies + CNN (2 CNN layers and 1 fully-connected)*	87.3 % (± 0.7)	0.437 (± 0.045)	-	450.1 KB	450 KB

DCASE: Audio Scene classification

DCASE2020 Task 1 Baseline, Subtask A ***OpenL3 + MLP (2 layers, 512 and 128 units)***





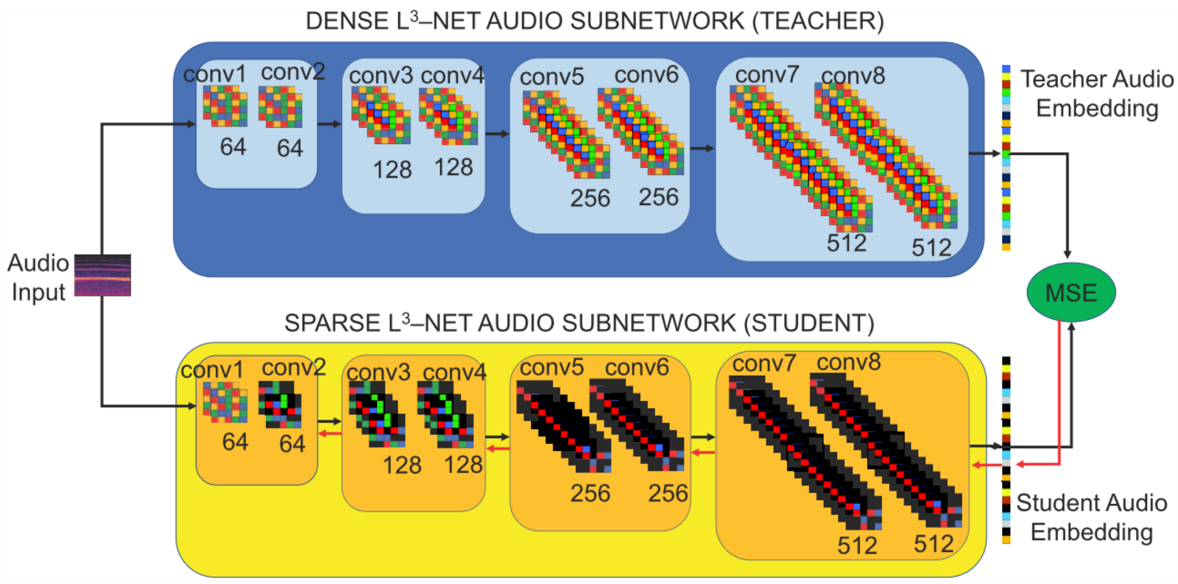
R. Arandjelovi´c and A. Zisserman, *“Look, listen and learn,”* in IEEE ICCV, 2017, pp. 609–617.
S. Kumari, D. Roy, M. Cartwright, J. P. Bello, and A. Arora. *EdgeI^3: compressing I^3-net for mote scale urban noise monitoring.* In 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW),

DCASE: Audio Scene classification

Modified DCASE2020 Task 1 Baseline, Subtask A

EdgeL3 + MLP (2 layers, 64 units each)

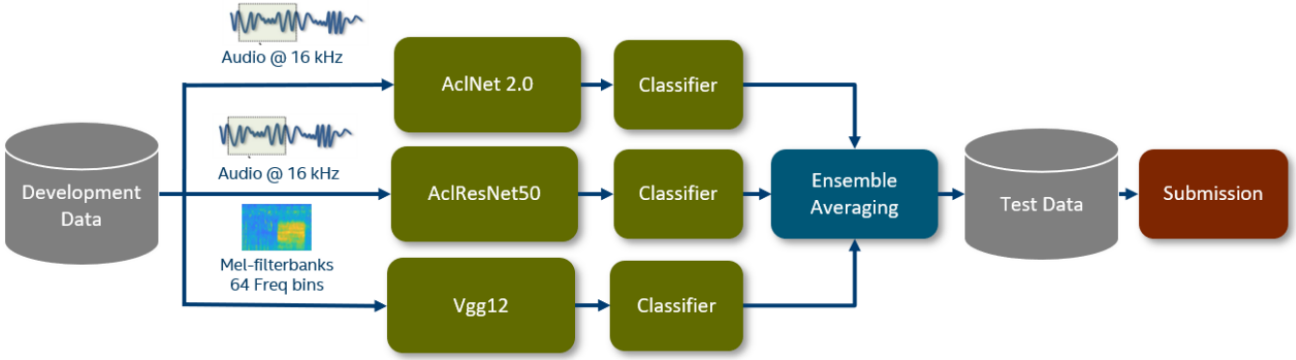
- **Sparsity**
  - Teacher-student
  - Different level of sparsity for each layer



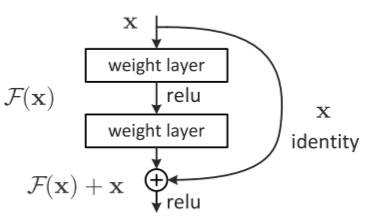
S. Kumari, D. Roy, M. Cartwright, J. P. Bello, and A. Arora. *Edgel^3: compressing l^3-net for mote scale urban noise monitoring.* In 2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW),

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

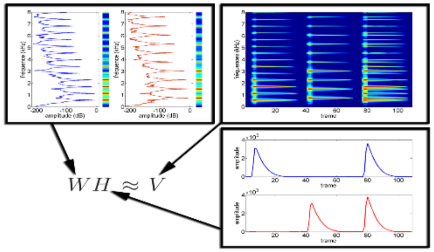


  - Use signal or audio models (NMF, ..)

P. Lopez & al. “Ensemble of Convolutional Neural Networks”, in DCASE 2020 Acoustic Scene Classification Challenge

Acoustic scene recognition:
Why using signal or perceptual models

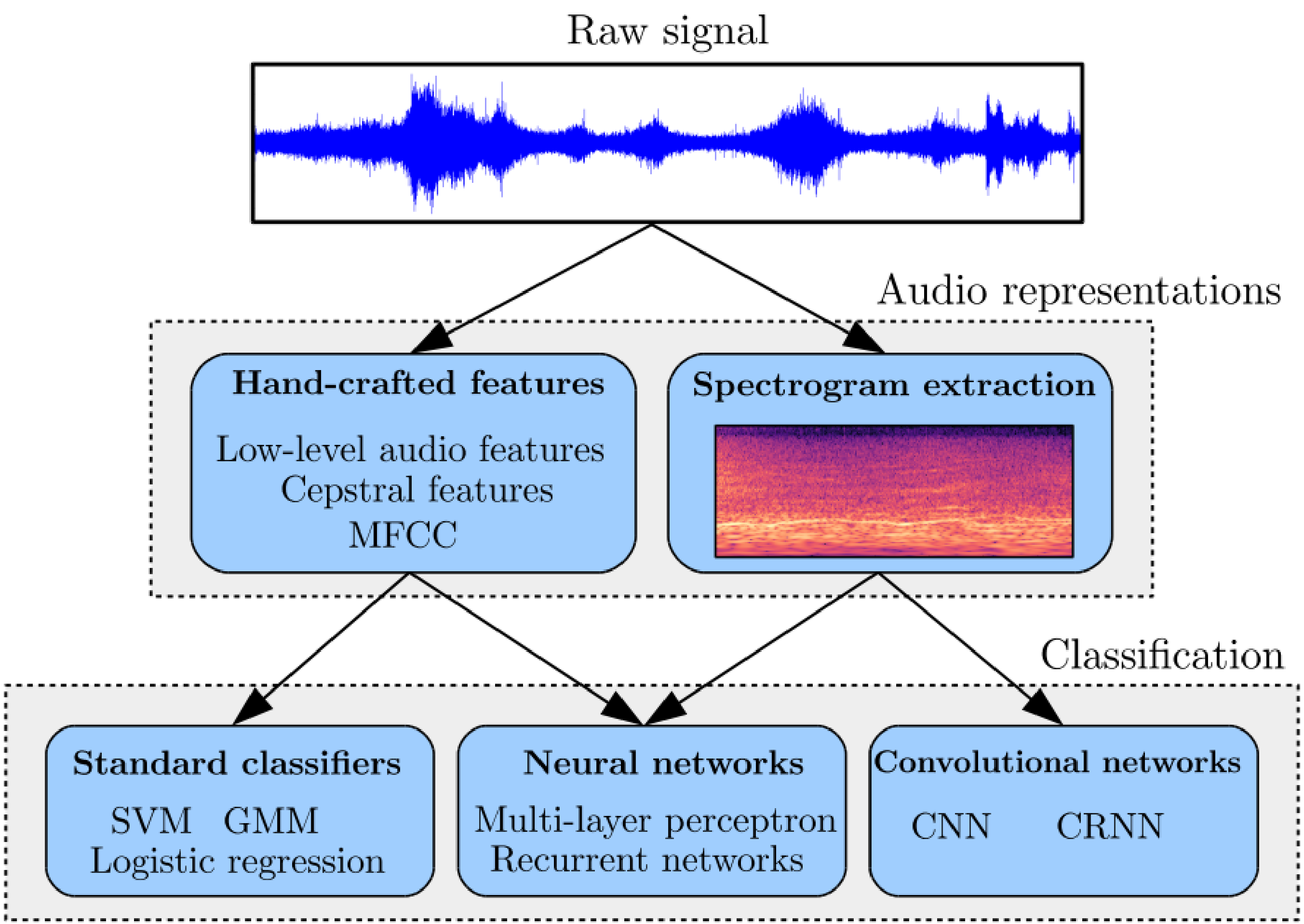
- **Using perceptual models**
  - Example: Mel specrogram, MFCC, CQT,..
  - The classifier does not learn what is not audible



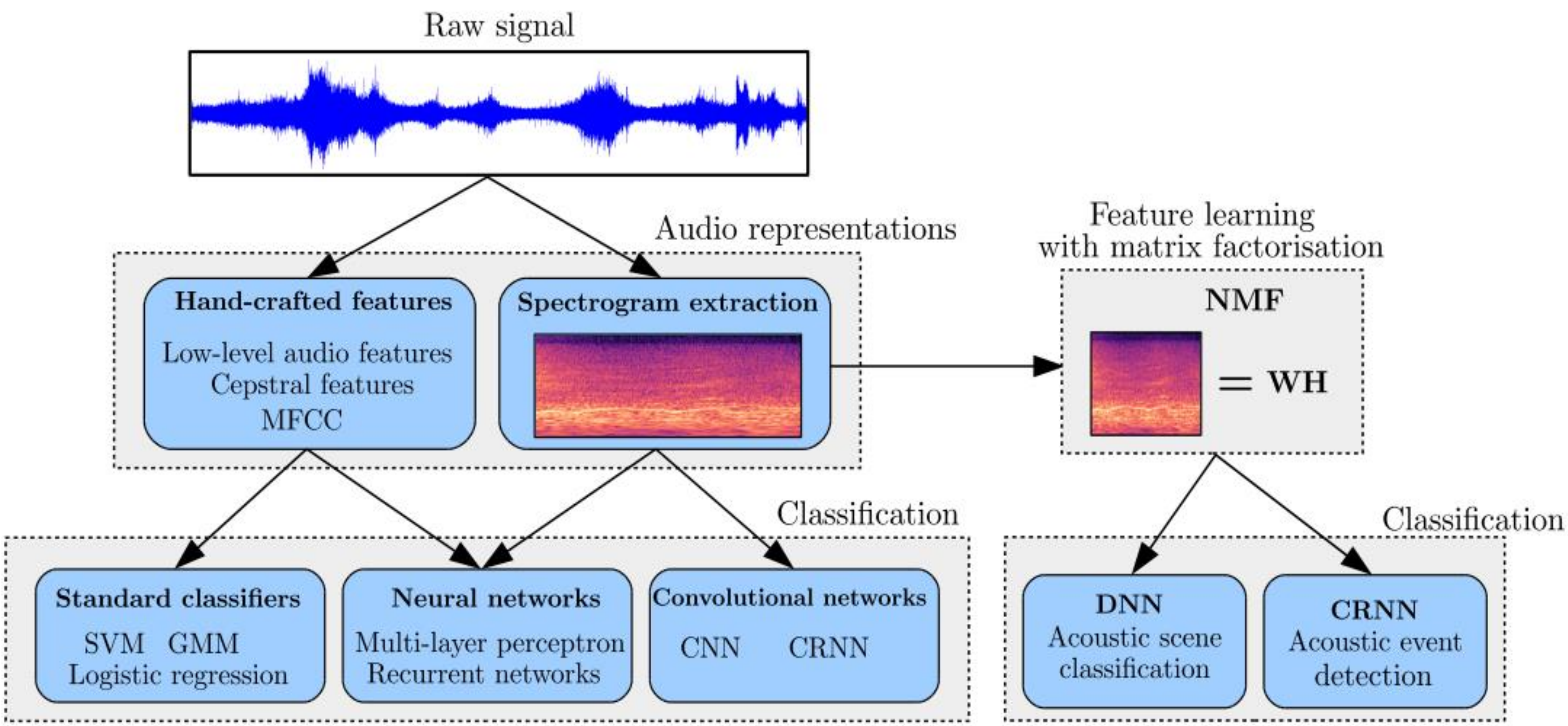
- **Using signal models**
  - Example: Harmonic + noise, Source filter, NMF, …
  - *e.g The classifier does not learn what is not typical of an audio signal*

- **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)

Recent approaches for Audio scene and event recognition



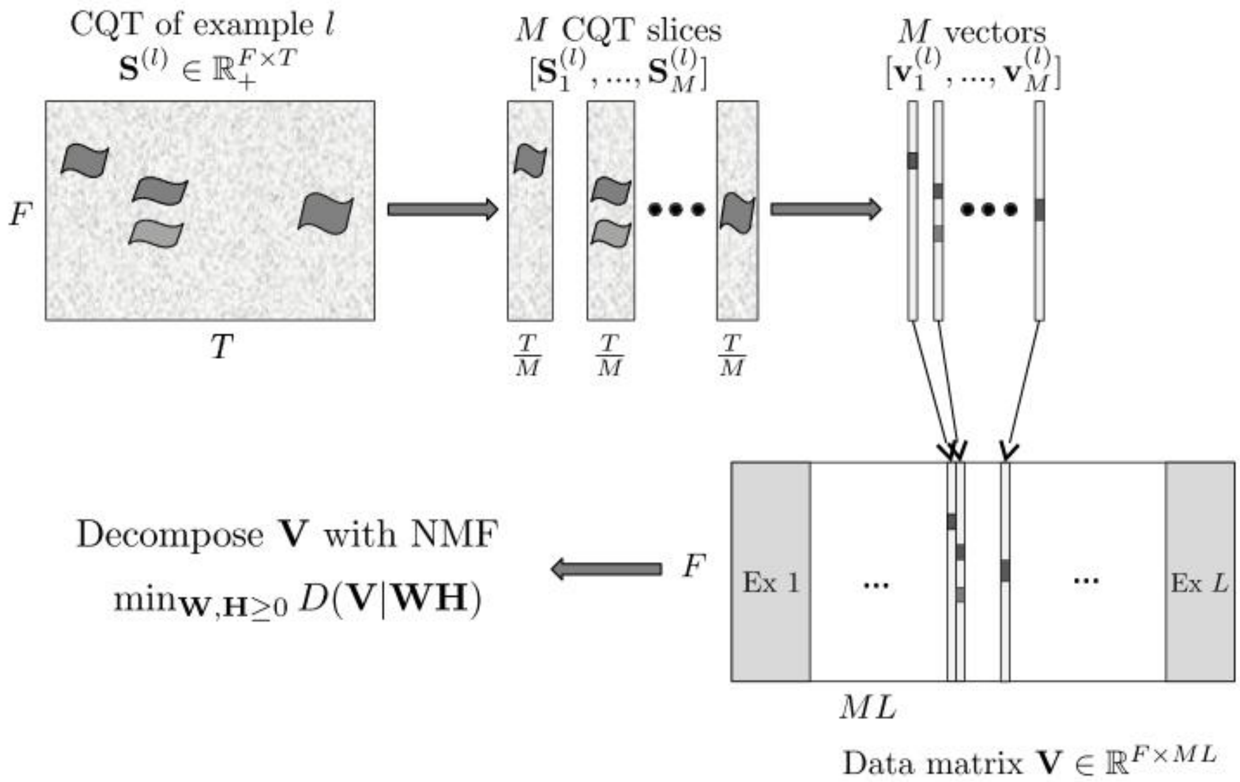
A recent framework for Audio scene and event recognition (Bisot & al. 2017)



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

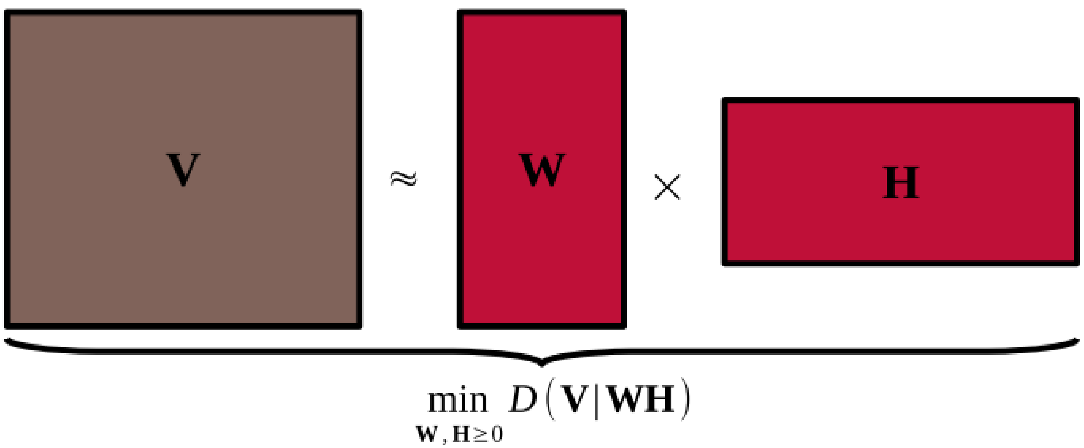


Unsupervised NMF for acoustic scene recognition

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

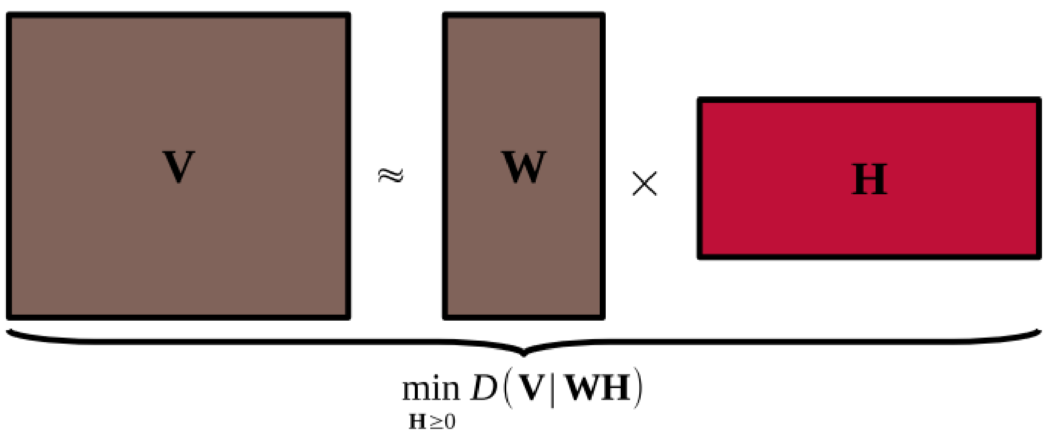


Unsupervised NMF for acoustic scene recognition

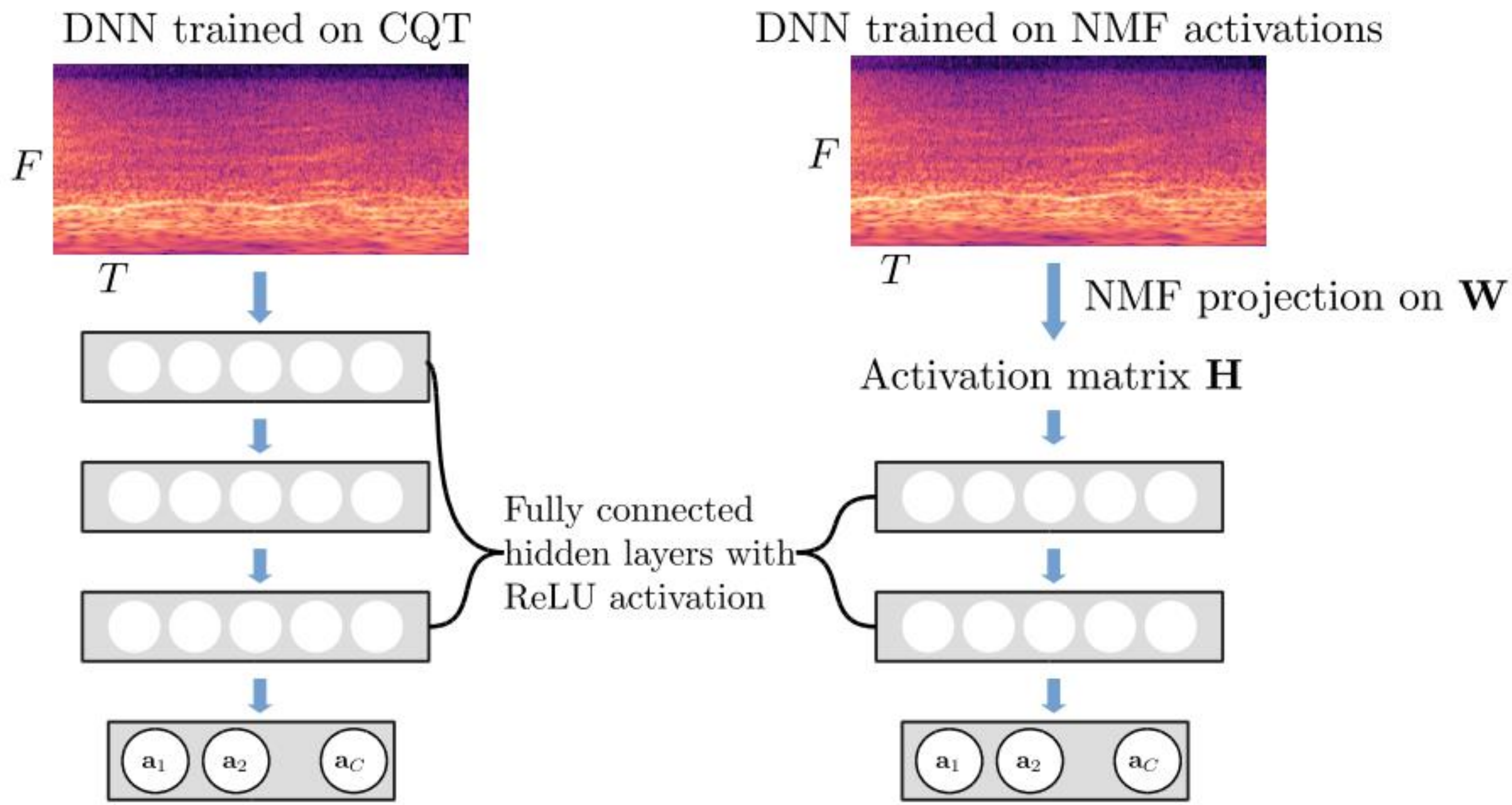
Nonnegative matrix factorization

$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{W}\mathbf{H})$ 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

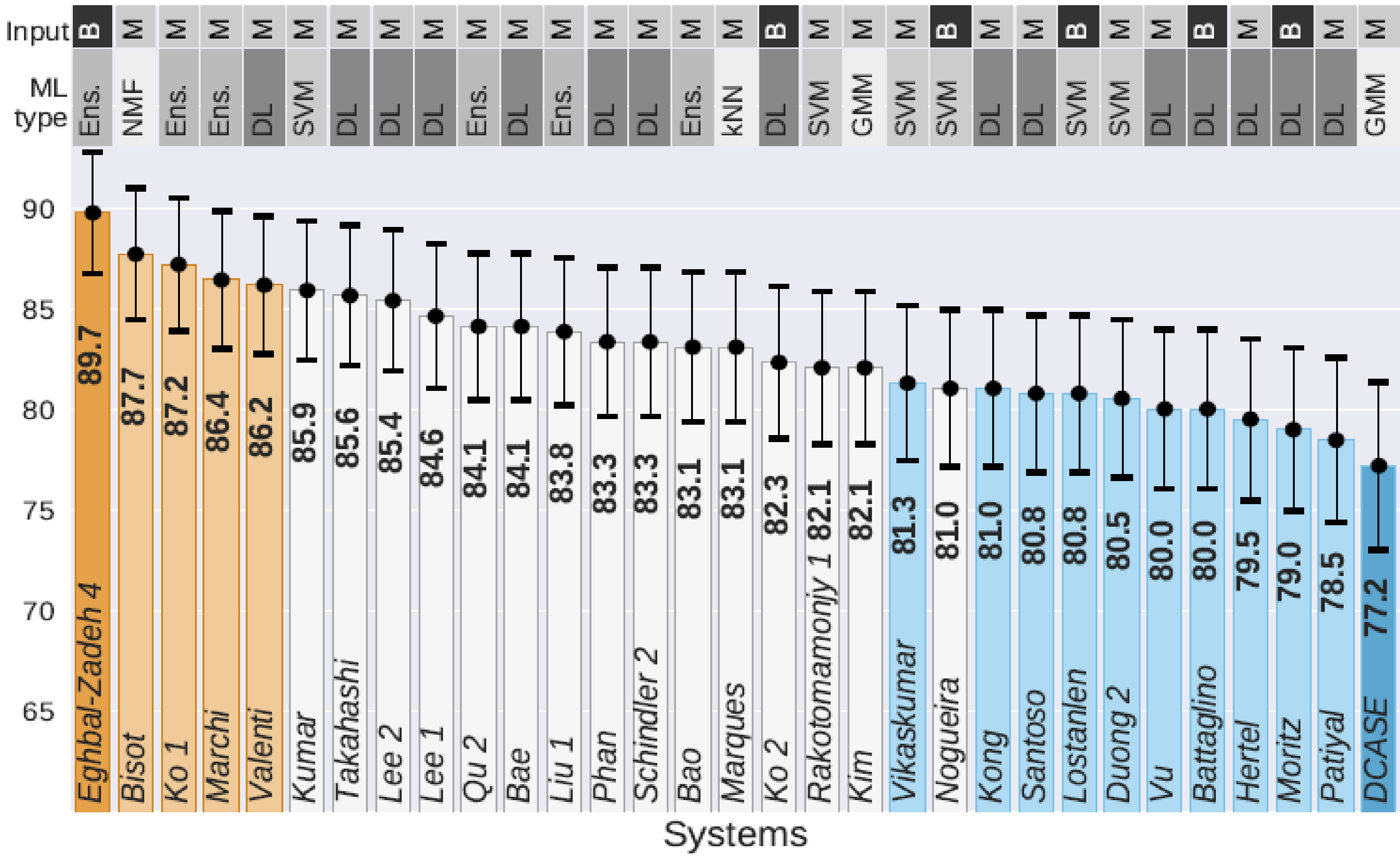


Example with DNN: acoustic scene recognition



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*

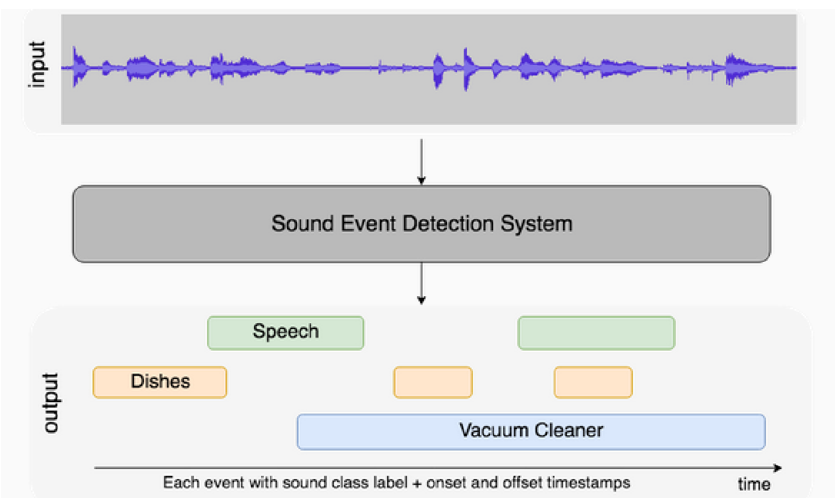
DCASE: Sound Event Detection and Separation in Domestic Environments

- **Goal**: the detection of sound events with their time localization using weakly labeled data (without timestamps).

- **Two subtasks in the challence DCASE 2021 (1/2)**

Domestic
Task 4

to provide the event class with event time localization given that multiple events can be present in an audio recording

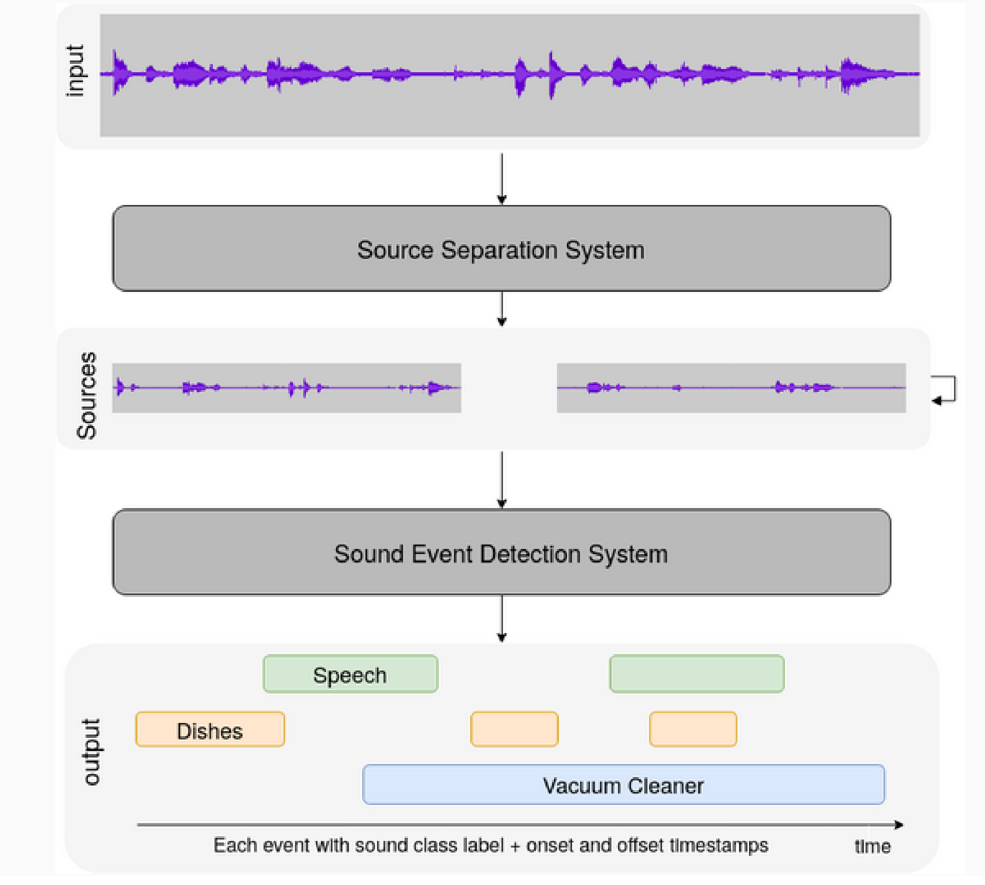


Dataset : many datasets *(see next slide)*
- DESED
- SINS
- TUT Acoustic scenes 2017
- FUSS
- FSD50K
- YFCC100M

DCASE: Sound Event Detection and Separation in Domestic Environments

- **Goal**: the detection of sound events with their time localization using weakly labeled data (without timestamps).

- **Possibility to use source separation** (until 2021)

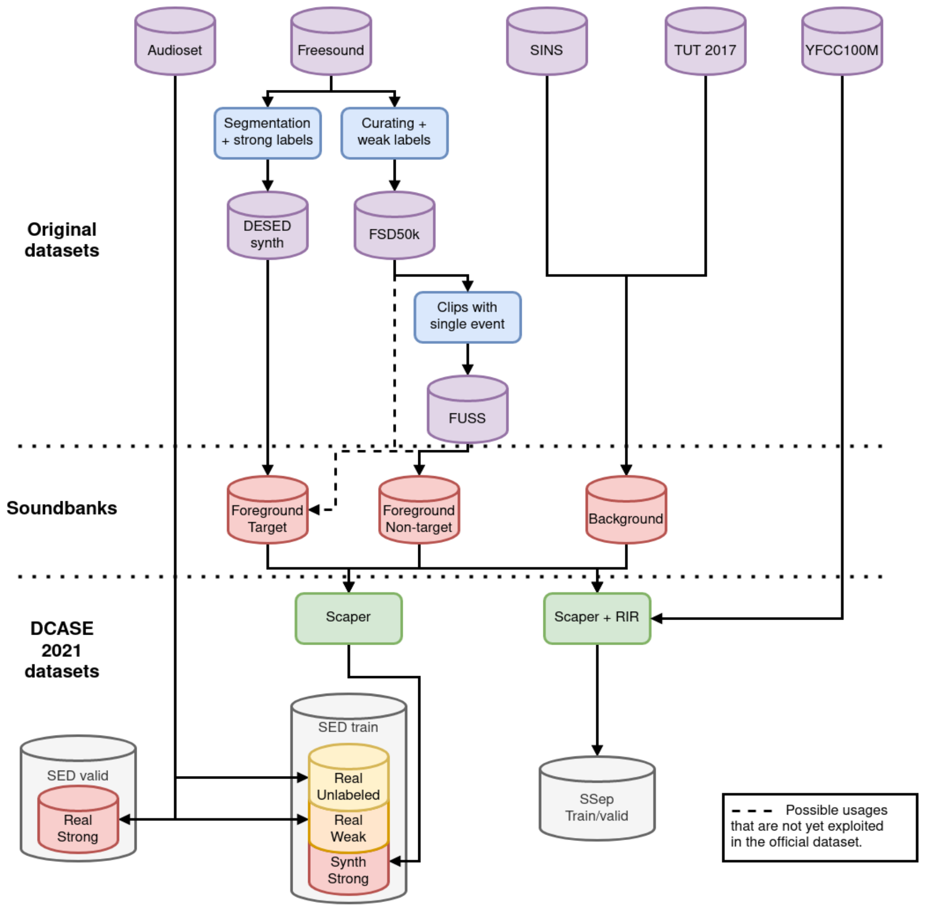


DCASE: task 4: datasets

Dataset	Subset	Type	Usage	Annotations	type	frequency
DESED	Real: weakly labeled	Recorded soundscapes	Training	Weak labels (no timestamps)	Target	44.1kHz
	Real: unlabeled	Recorded soundscapes	Training	No annotations	Target	44.1kHz
	Real: validation	Recorded soundscapes	Validation	Strong labels (with timestamps)	Target	44.1kHz
	Real: public evaluation	Recorded soundscapes	Evaluation **(do not use this subset to tune hyperparamters)**	Strong labels (with timestamps)	Target	44.1kHz
	Synthetic: training	Isolated events + synthetic soundscapes	Training/validation	Strong labels (with timestamps)	Target	16kHz
	Synthetic: evaluation	Isolated events + backgrounds	Evaluation **(do not use this subset to tune hyperparamters)**	Event level labels (no timestamps)	Target	16kHz
SINS		Background	Training/validation	No annotations	N/A	16kHz
TUT Acoustic scenes 2017, development dataset		Background	Training/validation	No annotations	N/A	44.1kHz
FUSS dataset		Isolated events + synthetic soundscapes	Training/validation	Weak annotations from FSD50K (no timestamps)	Target and non-target	16kHz
FSD50K dataset		Isolated events + recorded soundscapes	Training/validation	Weak annotations (no timestamps)	Target and non-target	44.1kHz
YFCC100M dataset		Recorded soundscapes	Training/validation	No annotations	Sound sources	44.1kHz

DCASE: sound event training set

- Weakly labeled training set : 1578 clips (2244 class occurrences)
- 14,412 unlabeled clips
- 10000 strongly labeled synthetic clips generated with Scaper.
- Non-target events from FUSS.
- Validation set (manually verified) with similar class distribution than the weakly labeled training set.



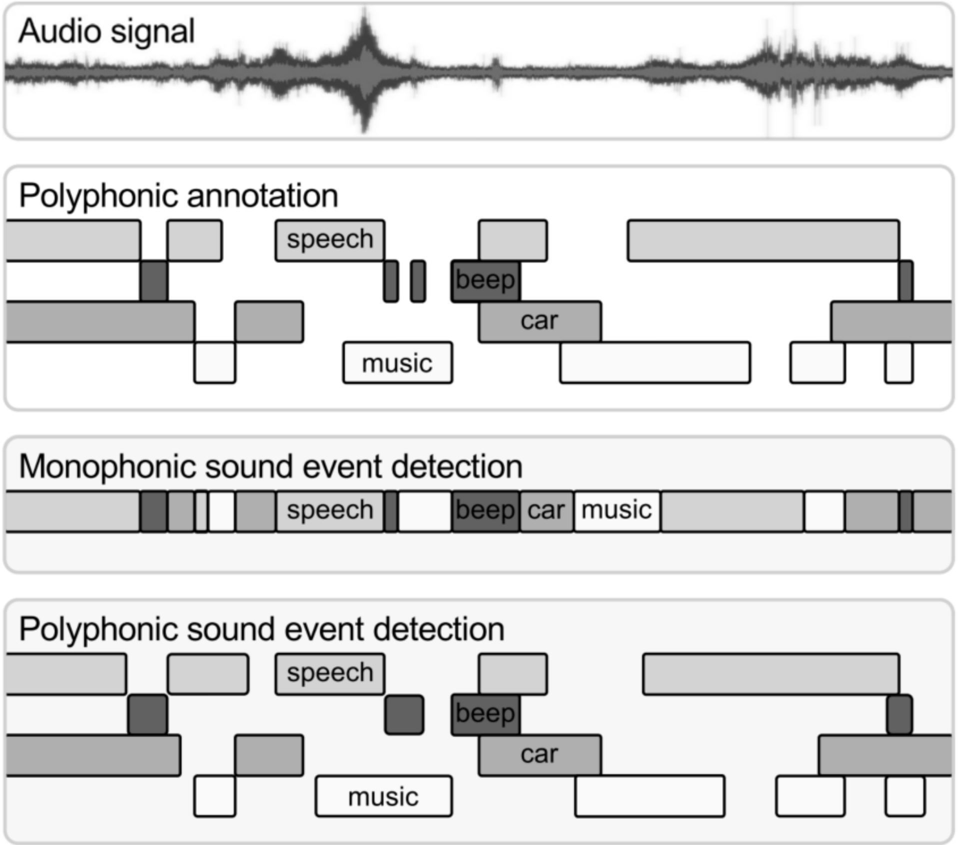
9

https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments

Salamon et. al. « Scaper: A Library for Soundscape Synthesis and Augmentation ». In *IEEE WASPAA 2017*
Wisdom et. al. « What's all the Fuss about Free Universal Sound Separation Data? » In IEEE *ICASSP 2021*

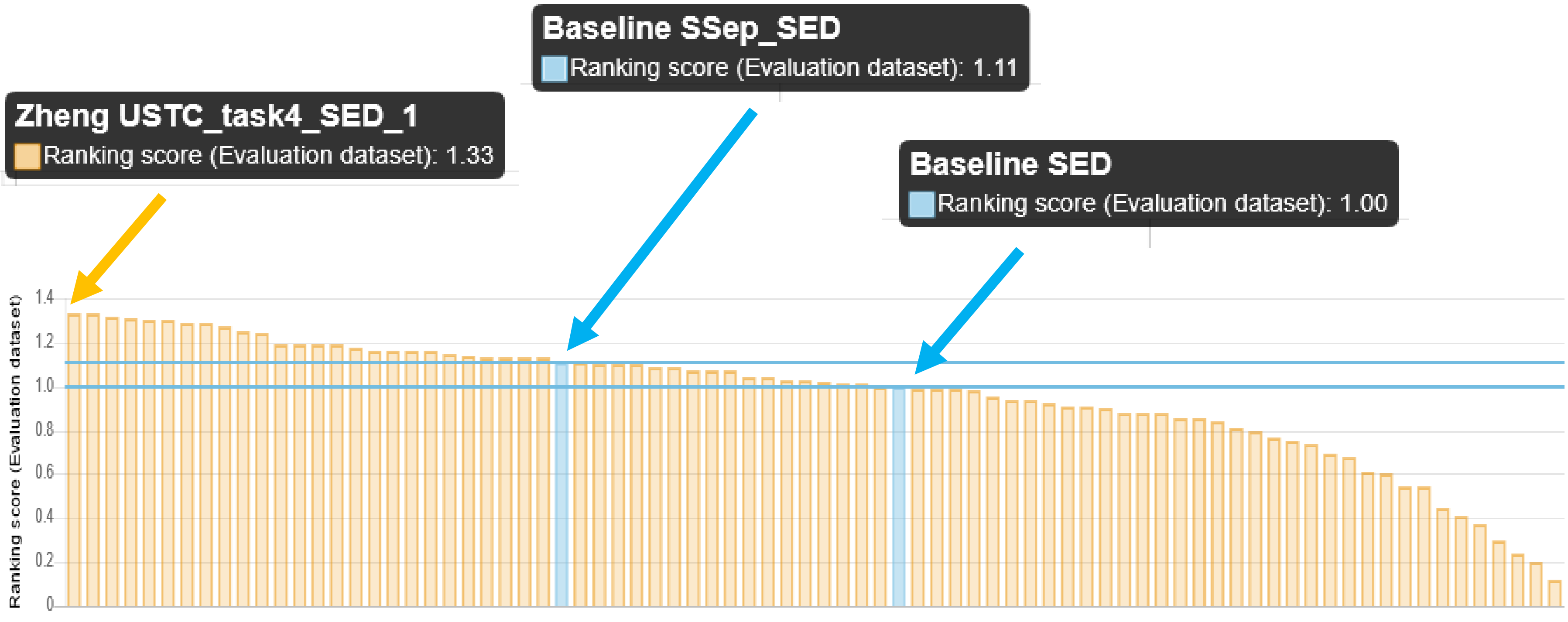
DCASE: Sound Event Detection and Separation in Domestic Environments

- **Evaluation: What is polyphonic event detection ?**



DCASE: Sound Event Detection and Separation in Domestic Environments

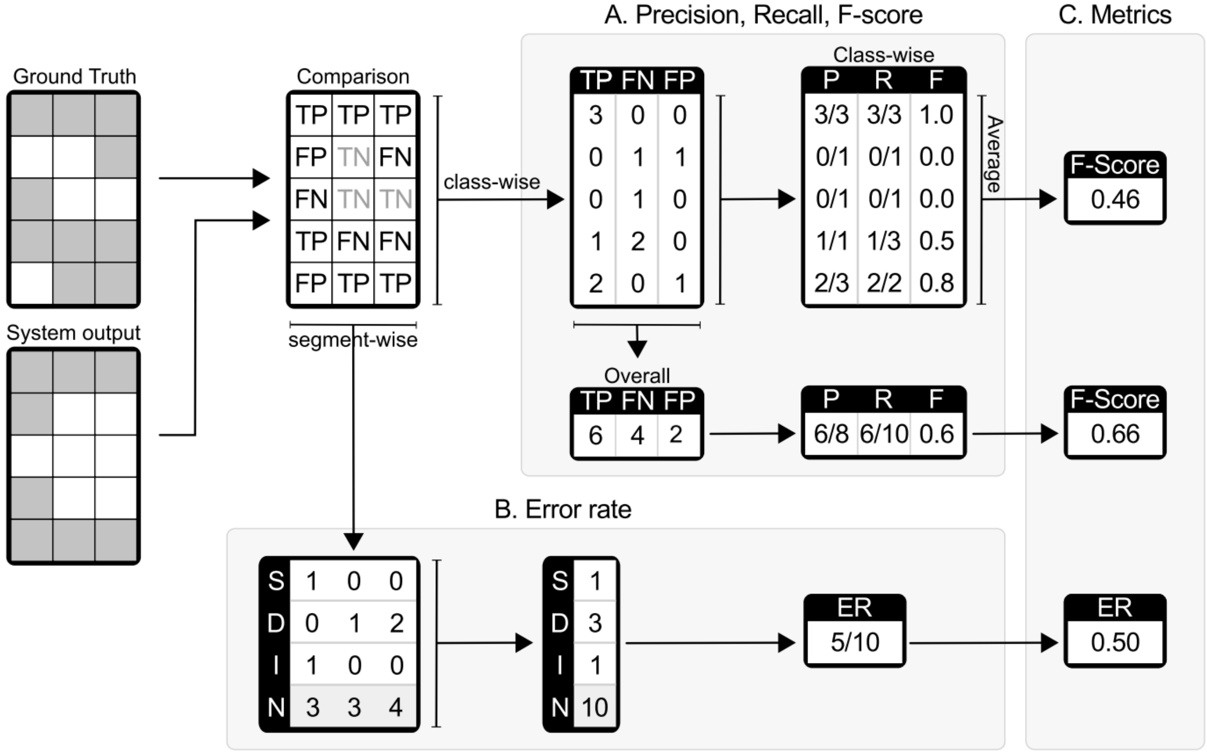
- **Performances**



Zheng, Xu and Chen, Han and Song, Zheng USTC Team's Submission For DCASE2021 Task4 – Semi-Supervised Sound Event Detection, DCASE2021 Challenge, Techn. Report

DCASE: Sound Event Detection and Separation in Domestic Environments

- How to evaluate Sound detection performances : **segment based metrics?**



TP/FP : True/False Positive
TN/FN: True/False Negative

P: Precision $= \frac{TP}{(TP+FP)}$

R: Recall $= \frac{TP}{(TP+FN)}$

F: F-measure $= \frac{2.P*R}{(P+R)}$

Error types:

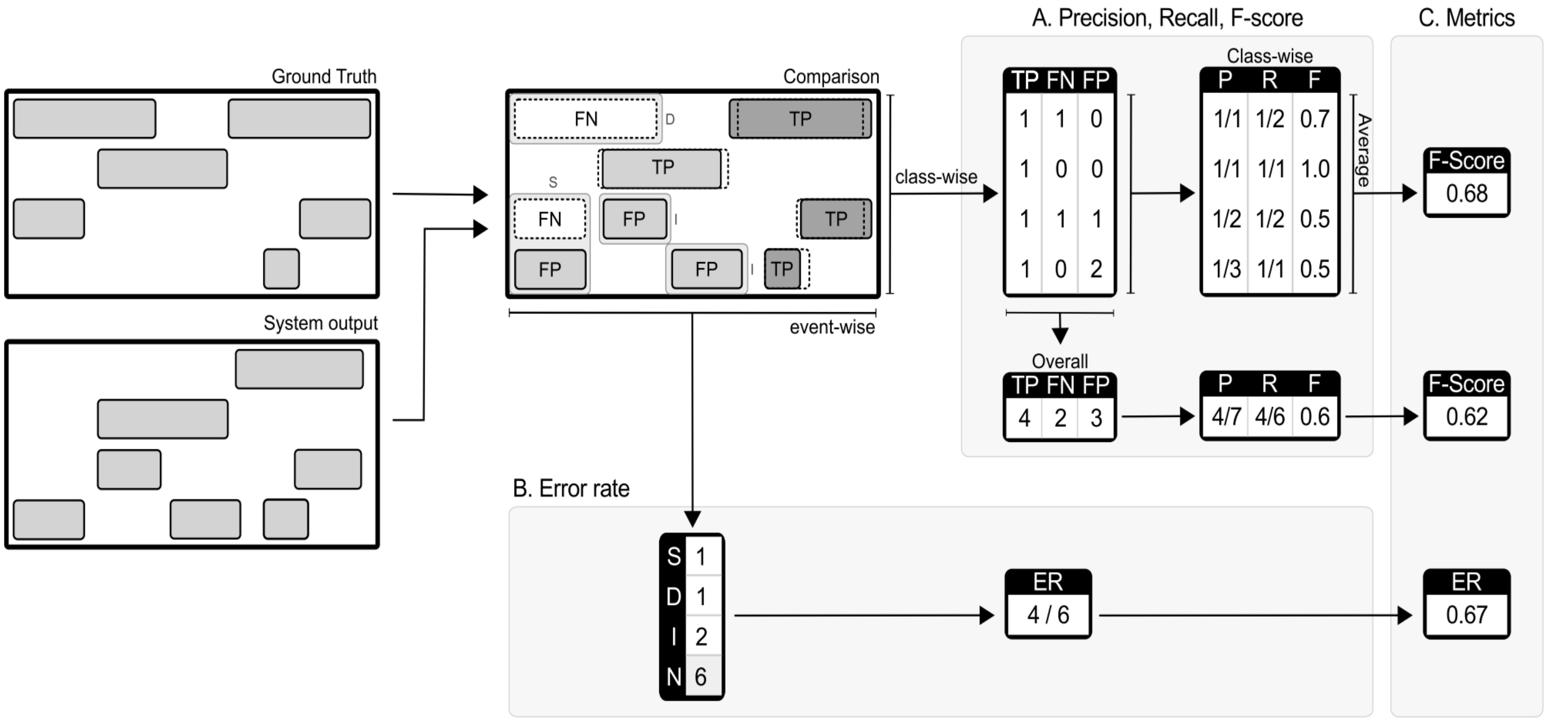
$$S(k) = min(FN(k), FP(k))$$
$$D(k) = max(0, FN(k) - FP(k))$$
$$I(k) = max(0, FP(k) - FN(k))$$

- S: Substitutions
- D: Deletions
- I: Insertions
- N: number of events active in a segment

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

DCASE: Sound Event Detection and Separation in Domestic Environments

- How to evaluate Sound detection performances : **Event-based metrics?**



Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

DCASE: Sound Event Detection and Separation in Domestic Environments

How to evaluate Sound detection performances ?

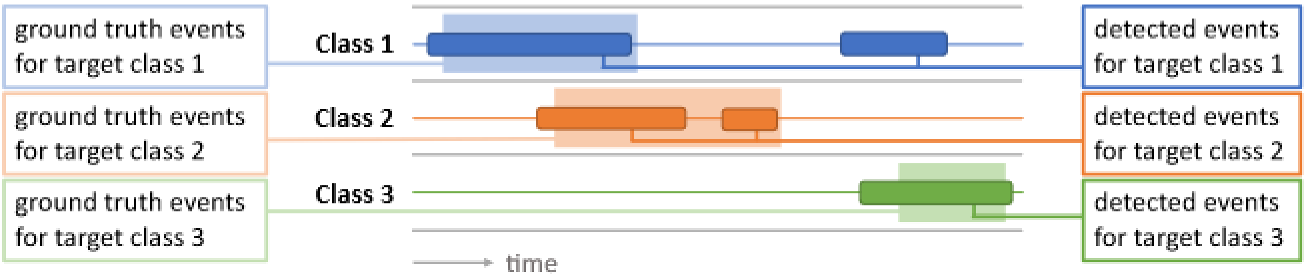
- Polyphonic Sound event Detection Scores (PSDS)
  - computed over the real recordings in the evaluation set
  - PSDS values are computed using 50 operating points (linearly distributed from 0.01 to 0.99)
  - Event-based metrics

- Many metrics « parameters »
  - Detection Tolerance criterion (DTC)
  - Ground Truth intersection criterion (GTC)
  - Cost of instability across class
  - Cross-Trigger Tolerance criterion
  - …

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Metrics for polyphonic sound event detection. Applied Sciences, 6(6):162, 2016. URL: http://www.mdpi.com/2076-3417/6/6/162, doi:10.3390/app6060162.

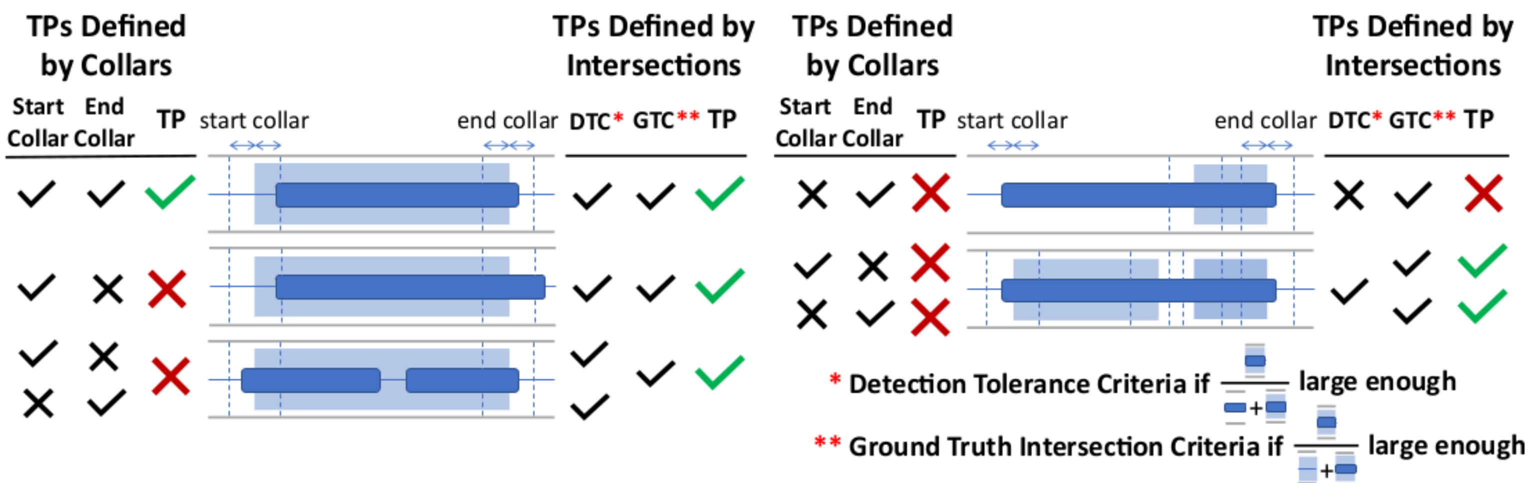
Evaluation of polyphonic sound event detection

- **Detected events vs Ground truth events**



Bilen et. al.. « A Framework for the Robust Evaluation of Sound Event Detection ». In *ICASSP 2020*

Metrics : Polyphonic sound event detection score (PSDS)



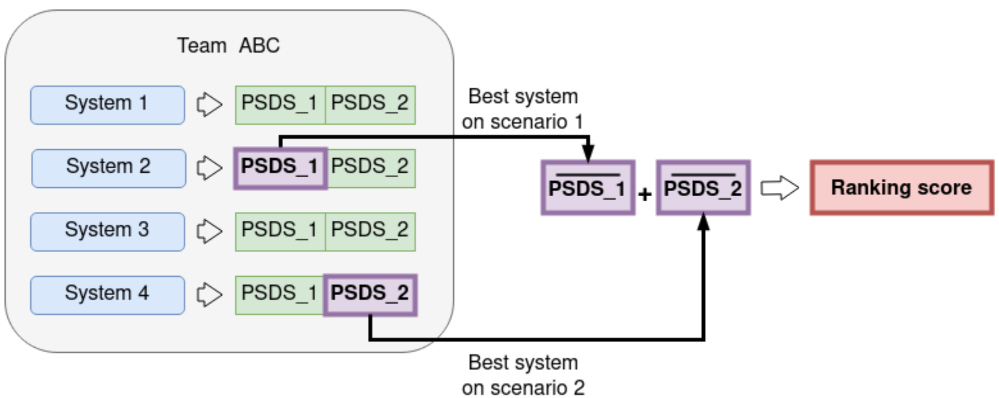
(a) TP decisions made by collars (left) vs. *DTC/GTC* (right).

- **Detection Tolerance Criteria:** controls how precise a system detection must be with respect to all the ground truths of the same class that it intersects.
- **Groudtruth Intersection Criteria:** defines the amount of minimum overlap necessary to count a ground truth as correctly detected.

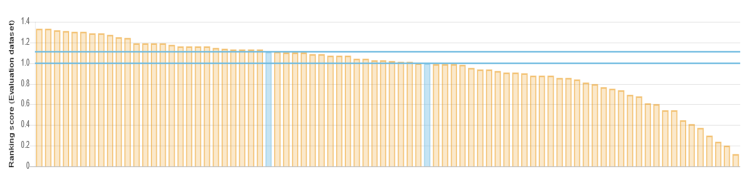
Bilen et. al.. « A Framework for the Robust Evaluation of Sound Event Detection ». In *ICASSP 2020*

Evaluation

- **Ranking teams with their two best systems on each scenario :**
  1. The system needs to react fast upon an event detection (e.g. to trigger an alarm, adapt home automation system...). The localization of the sound event is then really important.
  2. The system must avoid confusing between classes but the reaction time is less crucial than in the first scenario.

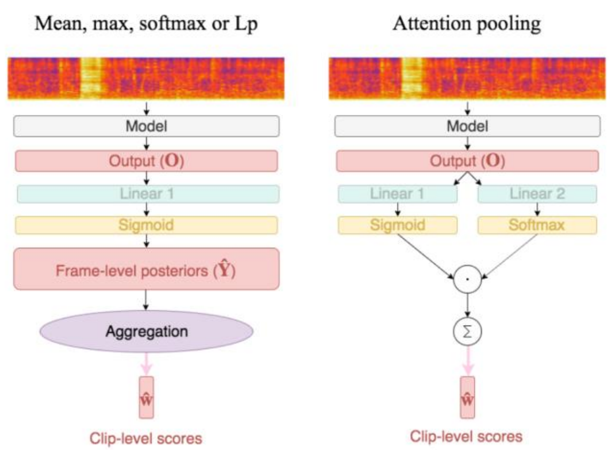


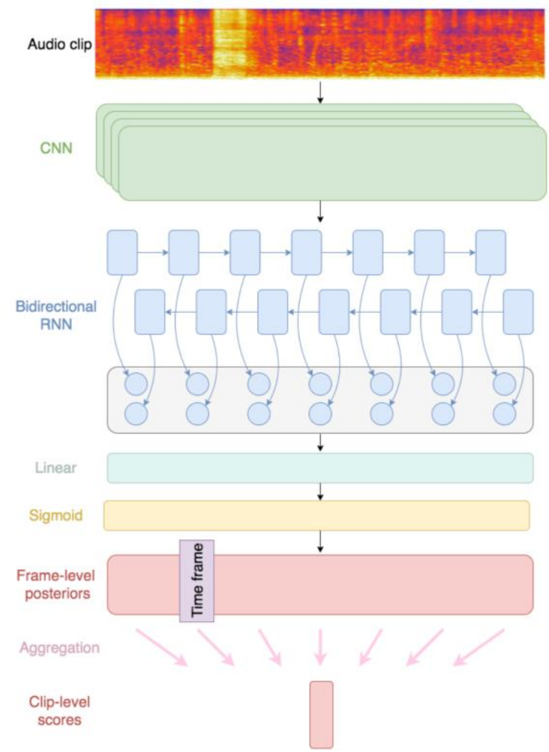
Ranking score



Baseline System : CRNN & Mean Teacher

- **Encoding frames with a CRNN**
- **Frame-level classification using dense layers**
- **Aggregation of frame-level output to get clip-level prediction**

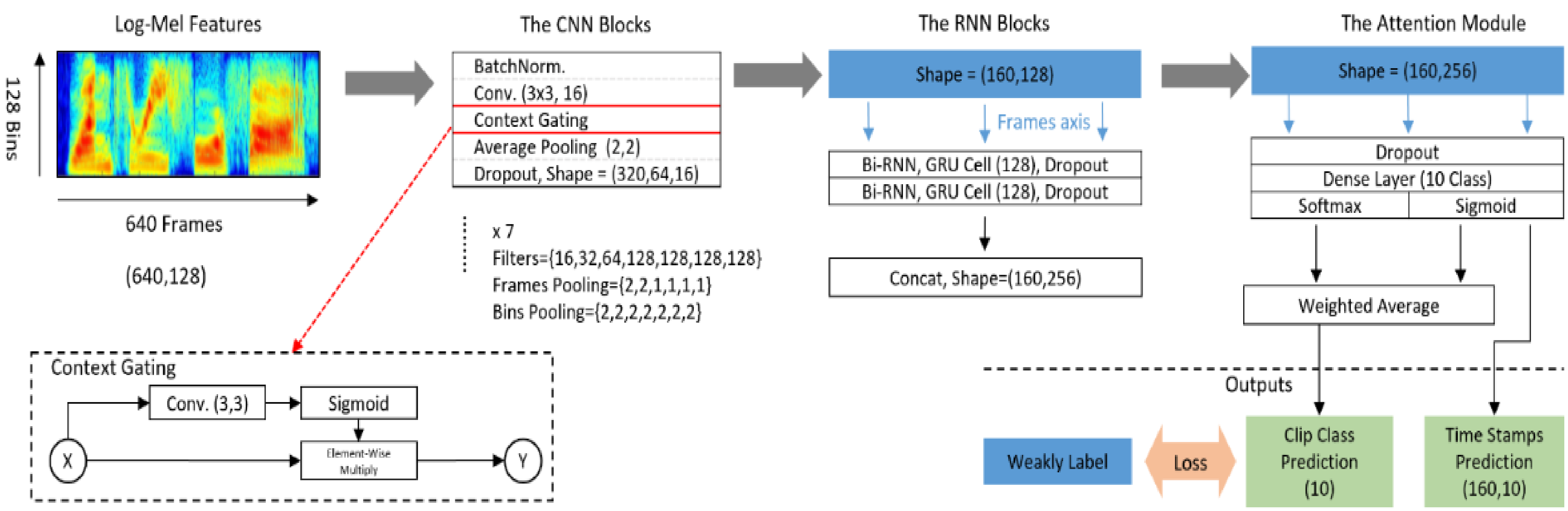




Turpault et. al. « Analysis of weak labels for sound event tagging». HAL-Inria 2021

DCASE: Sound Event Detection and Separation in Domestic Environments

- **Baseline system (another view..)**

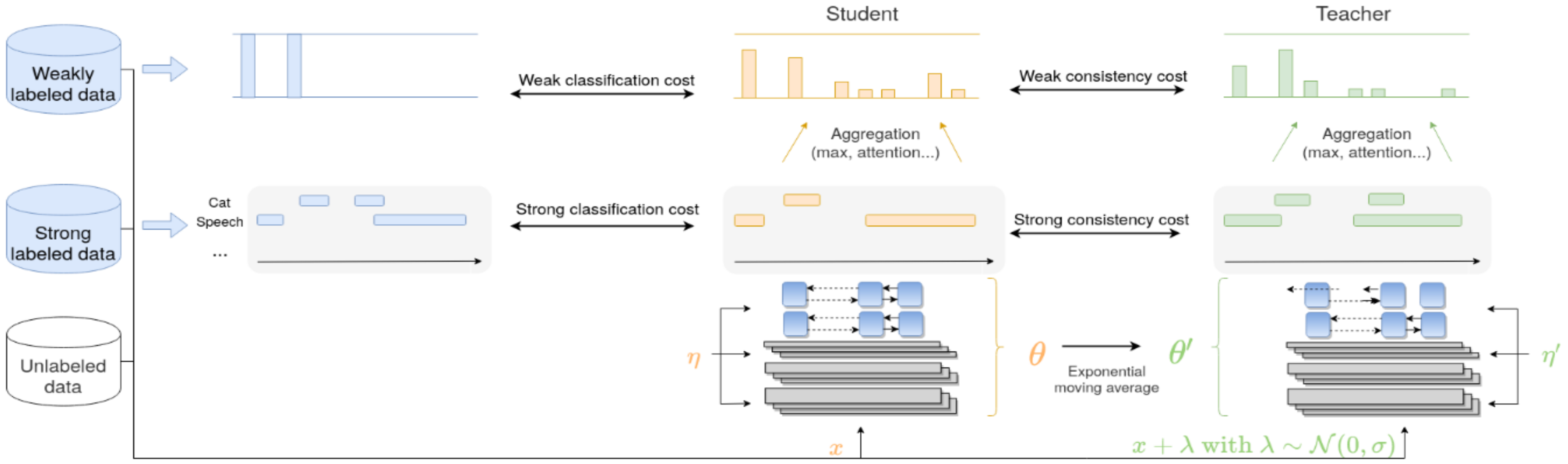


L. JiaKai, "Mean teacher convolution system for dcase 2018, task 4," DCASE2018 Challenge, Tech. Rep., September 2018

DCASE: Baseline System

- The student model parameters are updated based on a classification loss and a consistency loss between the student outputs and the teacher outputs.
- *The teacher model is not trained and is an average of consecutive student models*
- *The student model is used at inference time*

$$L(\theta) = L_{class_w}(\theta) + \sigma(\lambda) L_{cons_w}(\theta) + L_{class_s}(\theta_s) + \sigma(\lambda) L_{cons_s}(\theta_s)$$



Nicolas Turpault, Romain Serizel. Training Sound Event Detection On A Heterogeneous Dataset. DCASE Workshop, Nov 2020, Tokyo, Japan. hal-02891665v2
A. Tarvainen, H. Valpola. « Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results ». In Advances in Neural Information Processing Systems

Summary

- **Machine listening: a domain of growing interest**
- **… with many applications**

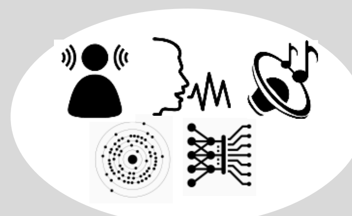
Audio surveillance, Audio scene analysis

Security, Health monitoring, bioacoustics



Transport & Communications

Autonomous cars, audio enhancement

Industry

Predictive maintenance

- **Some difficulties:**
  - Obtaining real-case annotated databases
  - Towards few-shot learning, unsupervised learning, …
  - … and distributed or sensor-based learning

A few additional references…

- ***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
  - R. Serizel, V. Bisot, S. Essid, G.Richard, Acoustic Features for Environmental sound Analysis, in Computational Analysis of Sound Scenes and Events, T. Virtanen, D. Ellis, M. Plumbley Eds., Springer International Publishing AG, pp 71-101, 2018

A few additional references…

- **Audio representation and models**

  - 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

- **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

- ***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.
  - R. Sonnleitner and G. Widmer, " Robust quad-based audio fingerprinting," IEEE Trans. Audio, Speech, Language Process. (2006–2013), vol. 24, no. 3, pp. 409–421, 2016.
  - J. Six and M. Leman, "Panako: A scalable acoustic fingerprinting system handling time-scale and pitch modification," in Proc. Int. Conf. Music Information Retrieval, 2014, pp. 259–264