Master M2 - DataScience

Audio and music information retrieval

Lecture on
Signal Models, Decomposition models, Music recognition

Gaël RICHARD
Télécom Paris
March 2022

TELECOM Paris
IP PARIS

Content

- **Introduction**

- **A Sound production model**

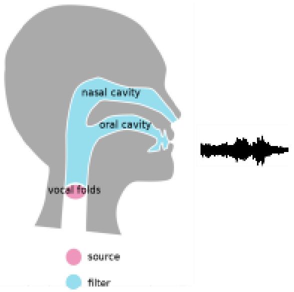
- **Elements of sound perception**
  - Basics of perception
  - Perception of pitch
  - Example of perception principles in models

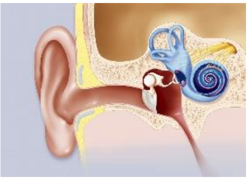
- **Signal decomposition models**
  - Sinusoidal models
  - Decomposition models (matching pursuit, NMF)
  - Exploitation of such models in scene analysis

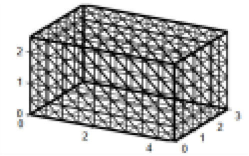
- **Audiofingerprint or Music recognition**

Audio Models

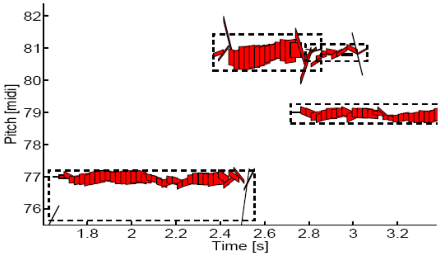
- **Audio models can represent the knowledge of**
  - How the sound is produced (sound production models)
  - How the sound is perceived (perception models)
  - How the sound propagates (sound rendering or reverberation models)
  - How the signal is structured (signal models, decomposition models)



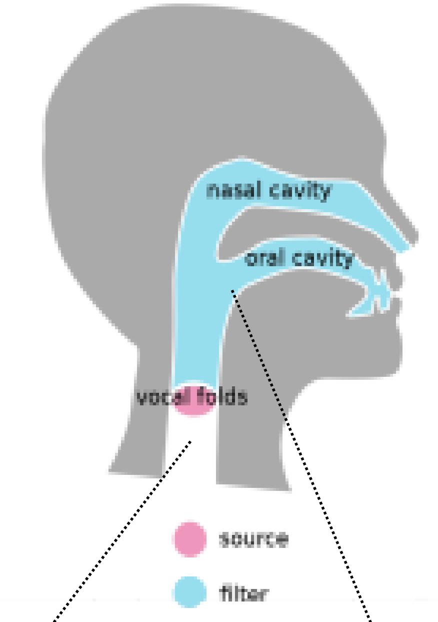




Discretized room



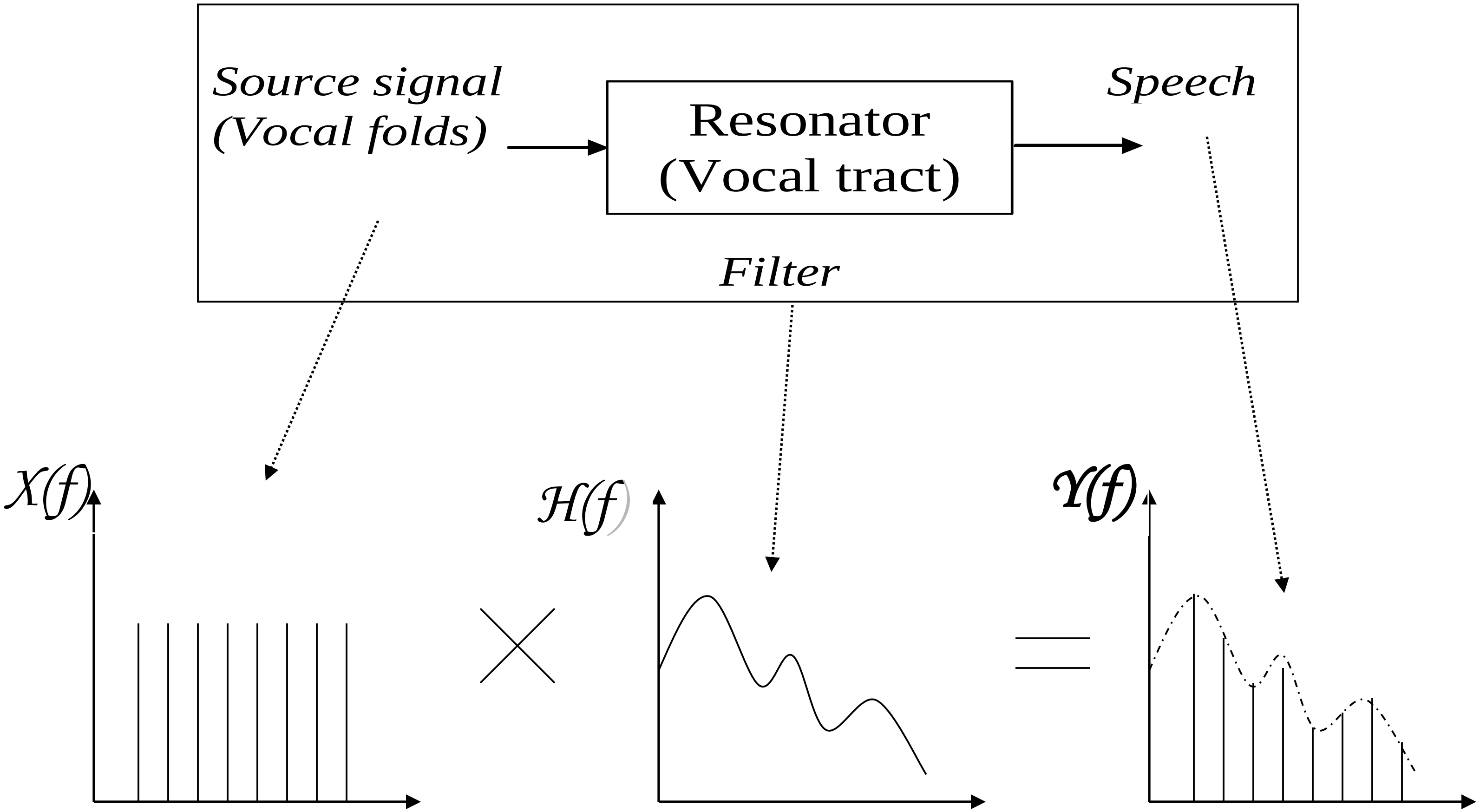
An example of a sound production model
the (speech) source filter model



Source signal (Vocal folds) → Resonator (Vocal tract) → *Speech*

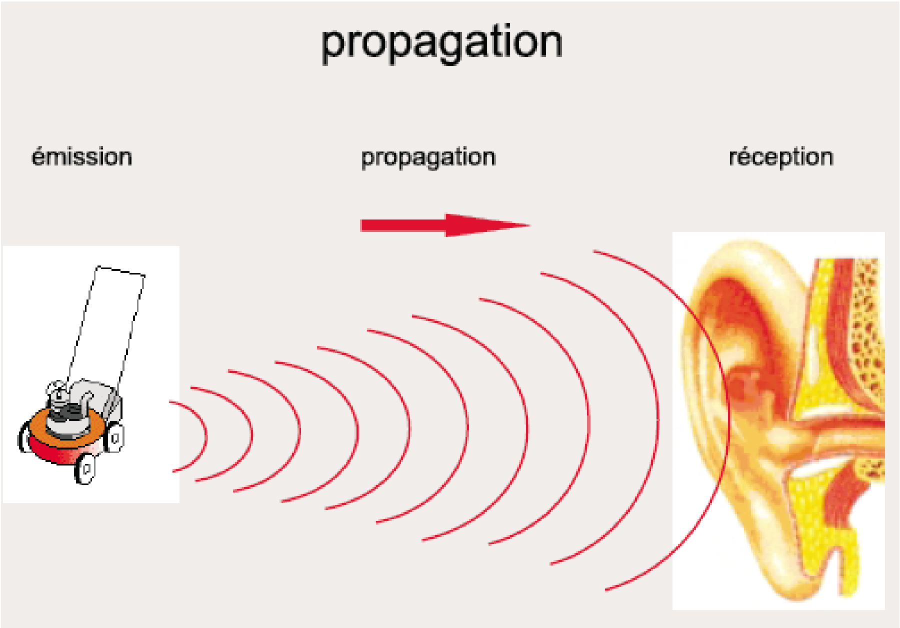
Filter

A widely used model: the source filter model



Perception and perception models

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



- **Decibel:**

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

$$I \propto P^2$$

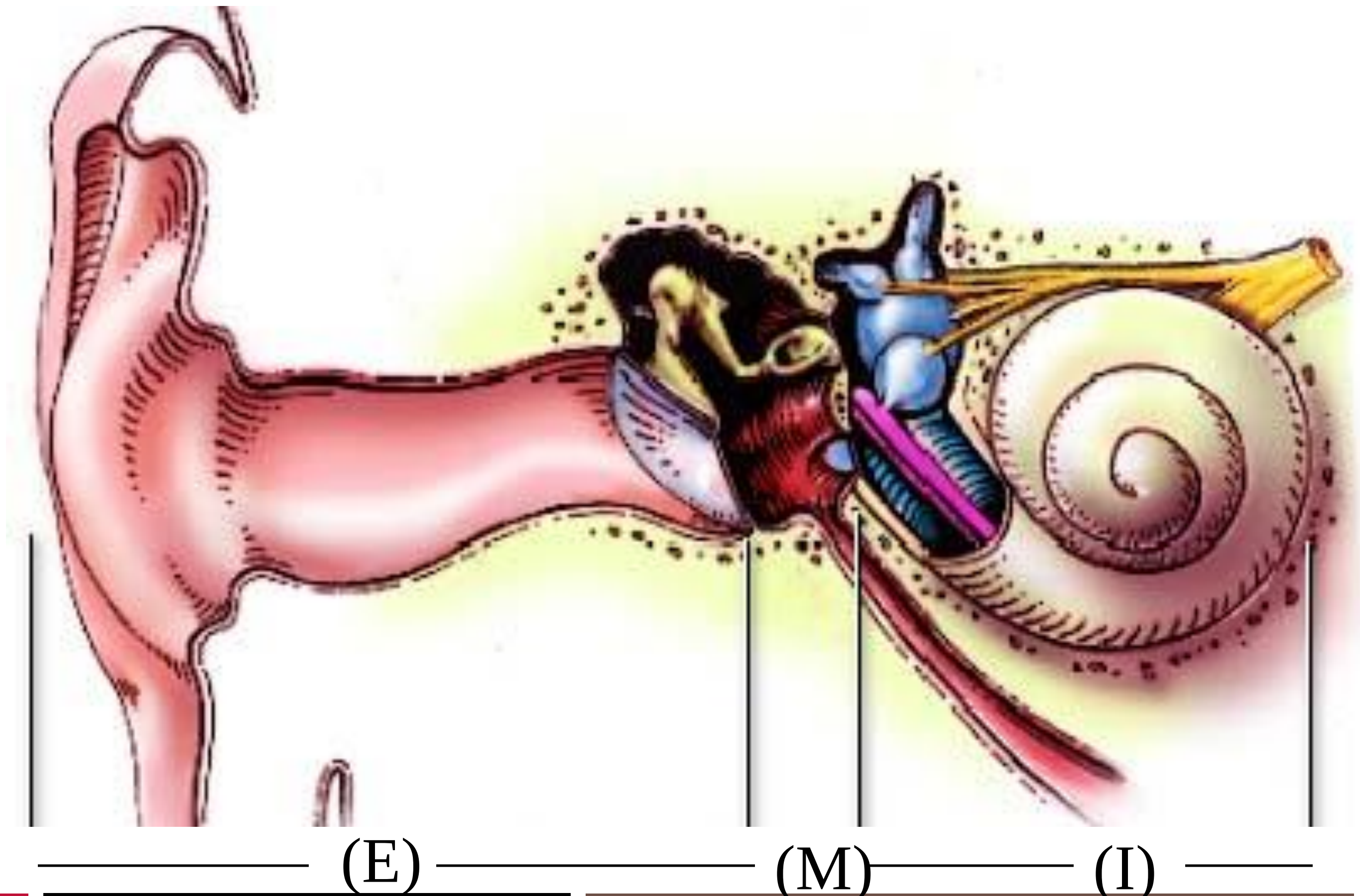
Perceptual scales

- **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

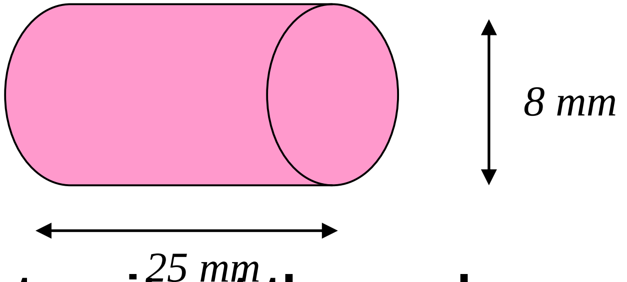
Outer ear (E), middle ear (M) and inner ear (I)



Outer ear

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

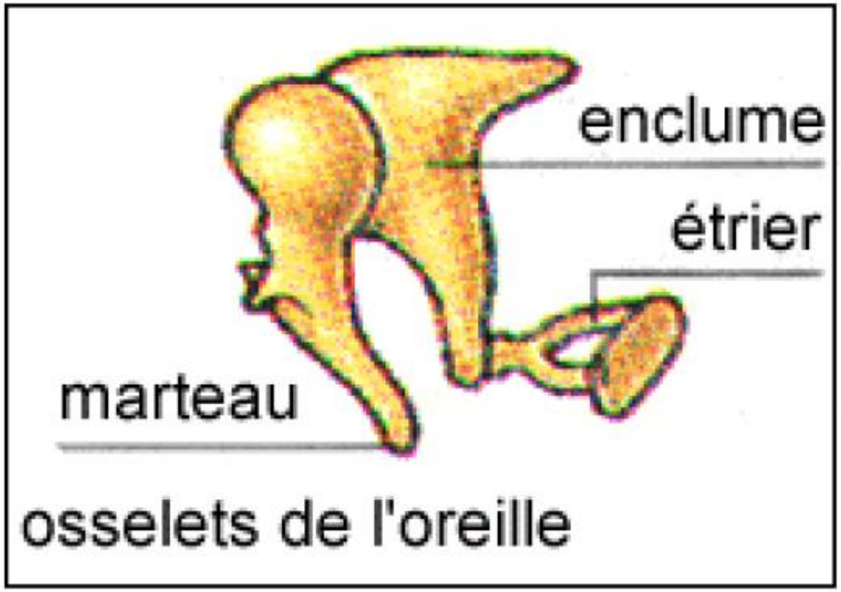
- **The External Auditory Canal** (E.A.C) = waveguide, to the eardrum



- **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)



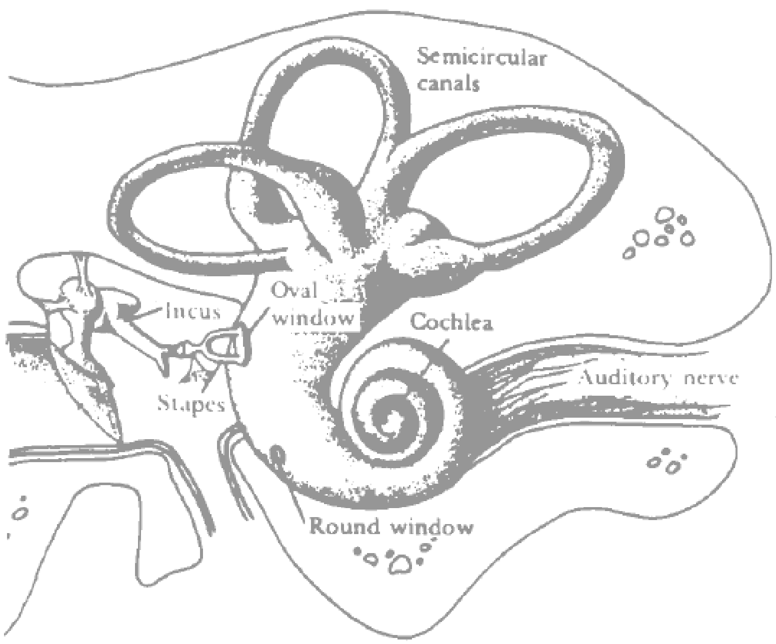
- **Hammer and Anvil attached with ligaments**

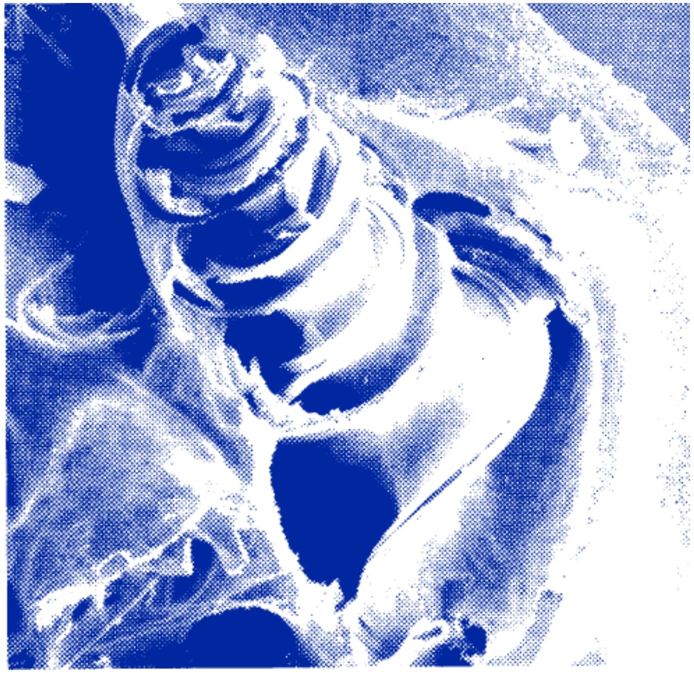
Middle ear: role

- **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)

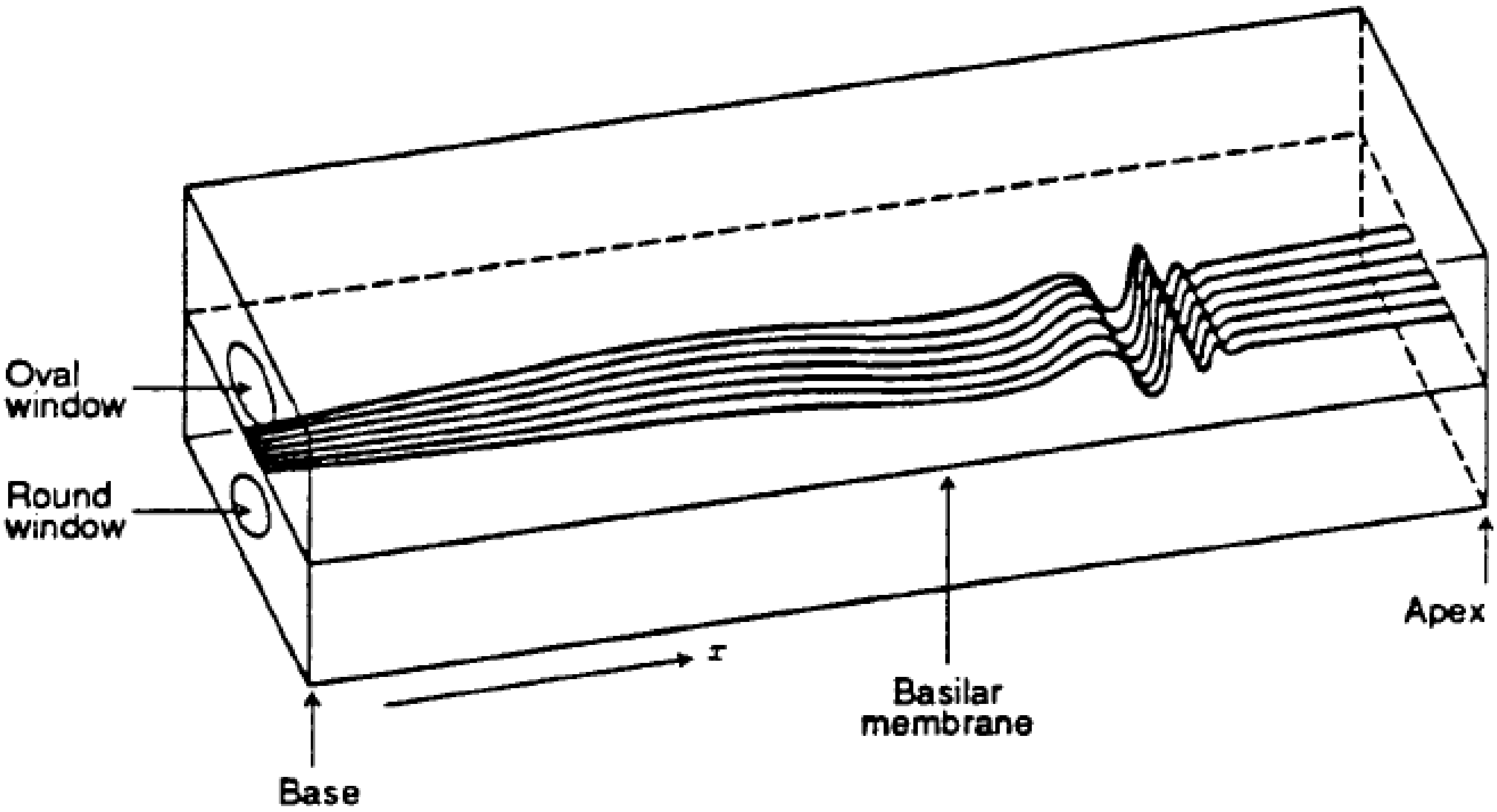
Inner ear

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

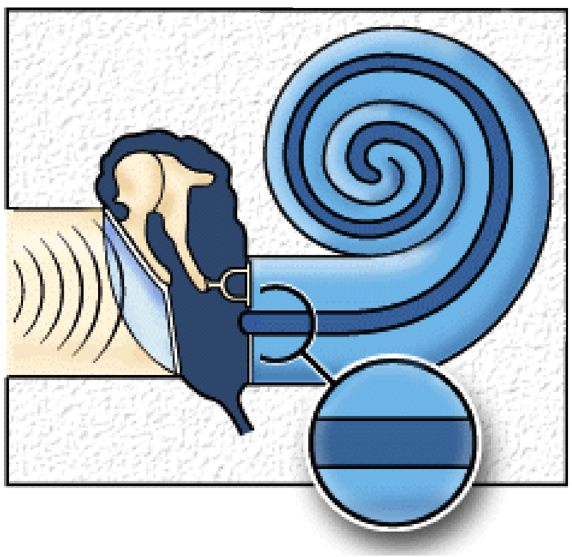


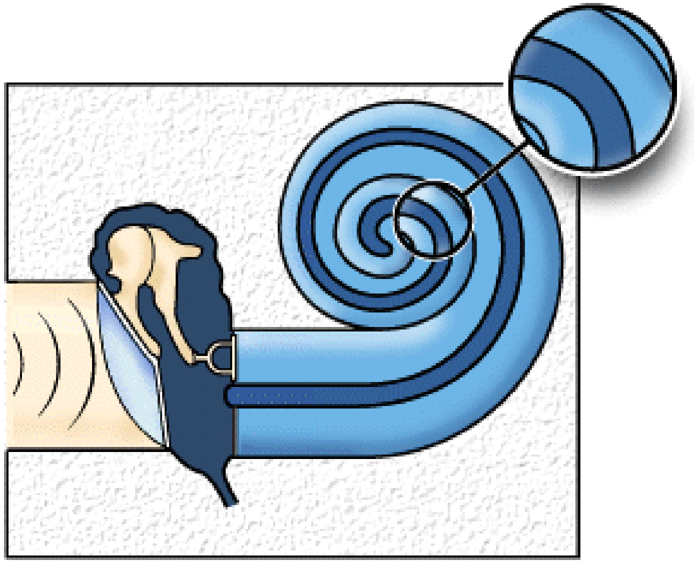


Cochlear canal



Ear: cochlear tonotopy



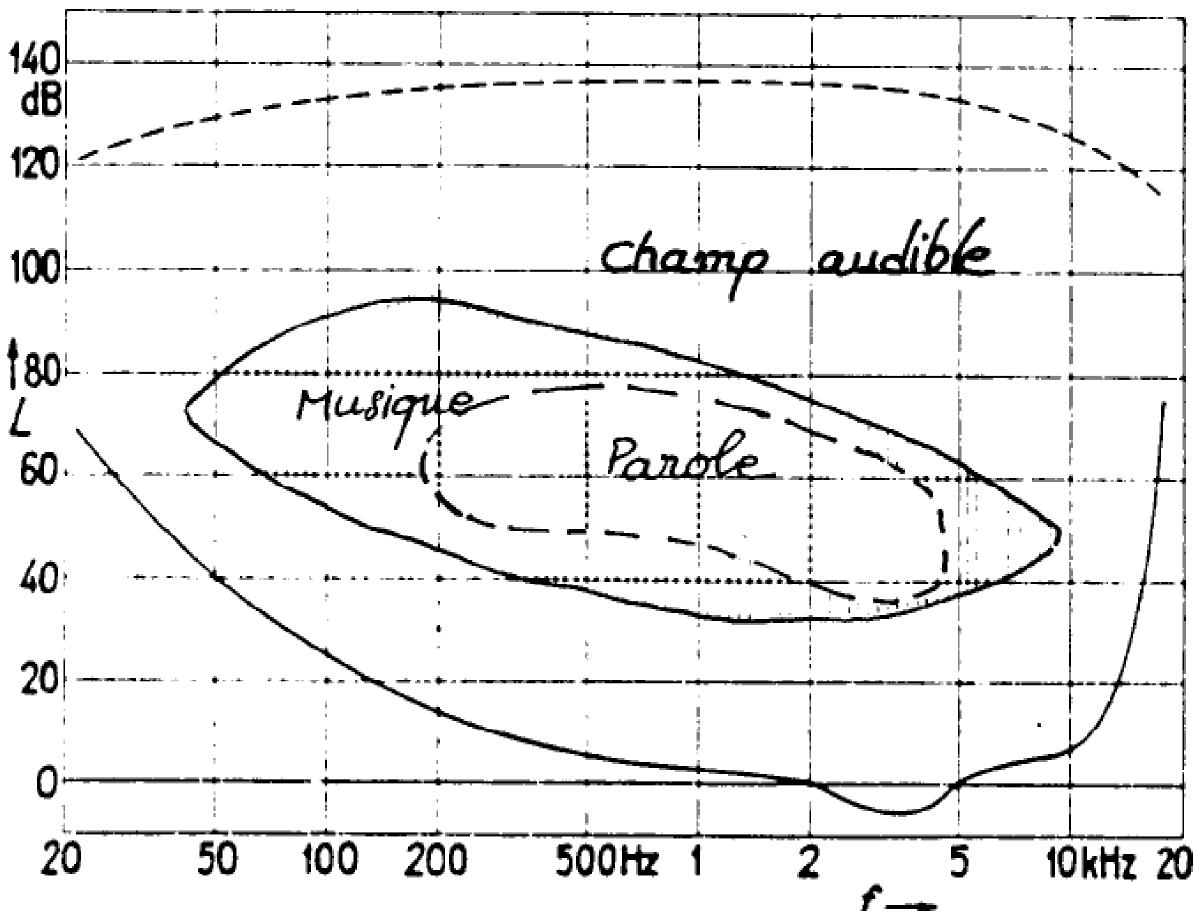


Trebble sound

Bass sound

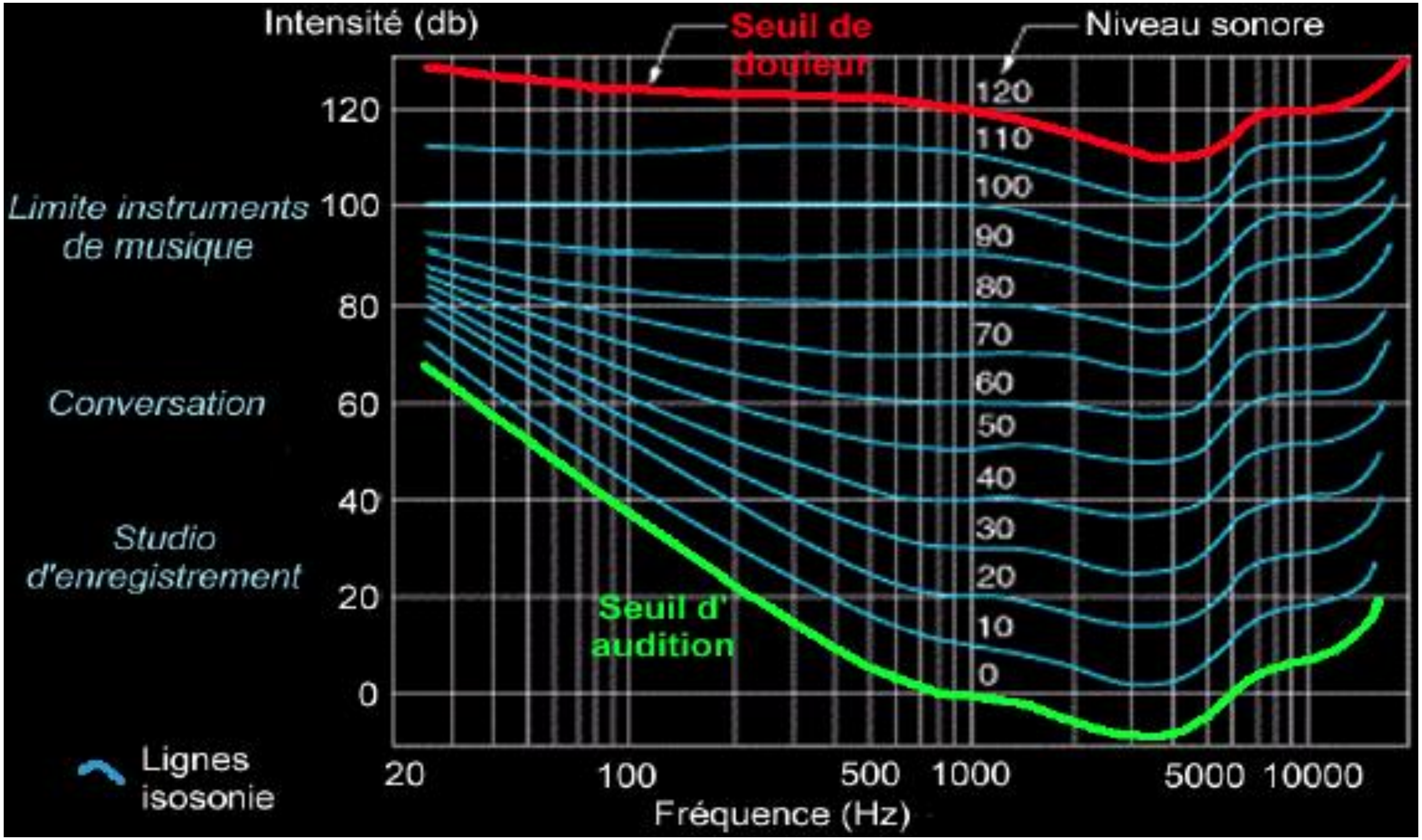
Audition

Dynamic of the ear: 120 dB!!



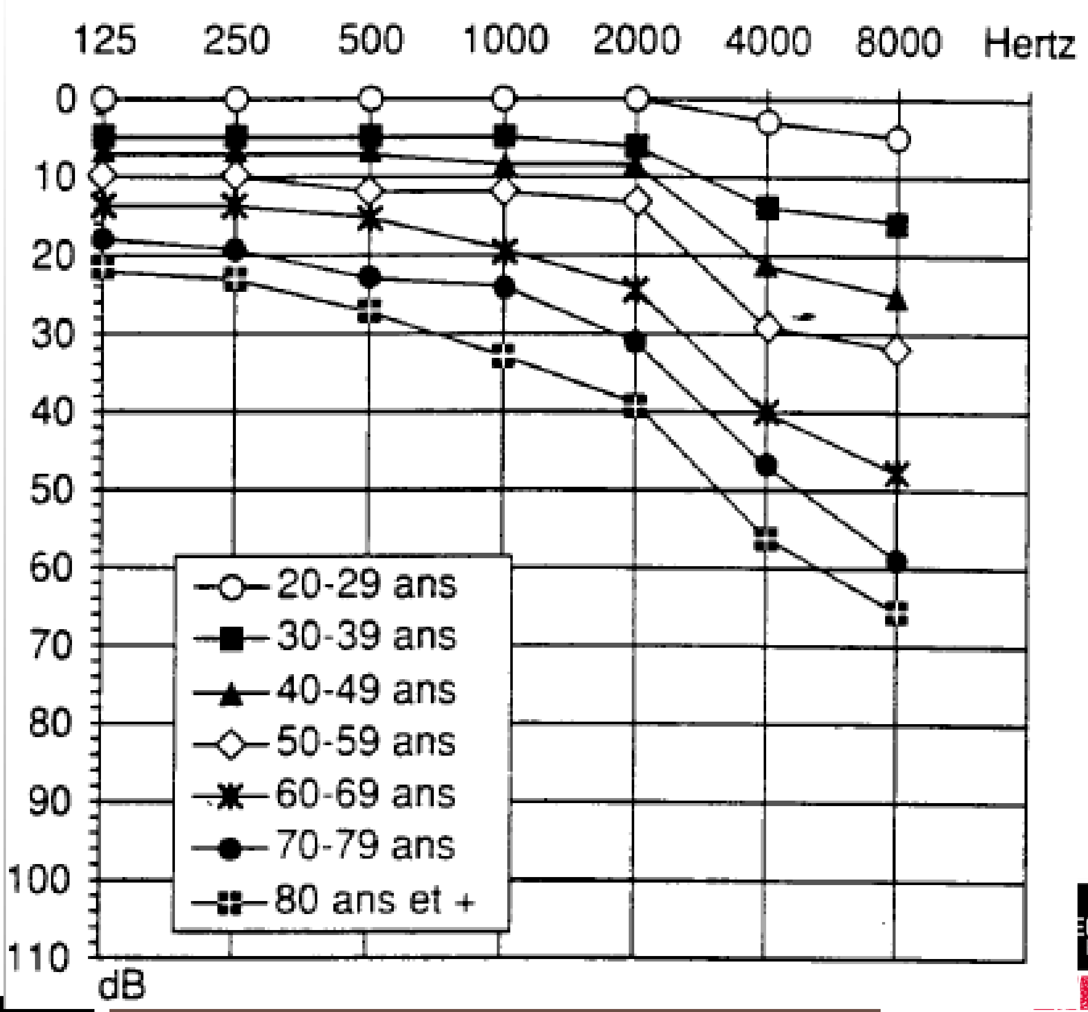
Isonosy : the phons

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

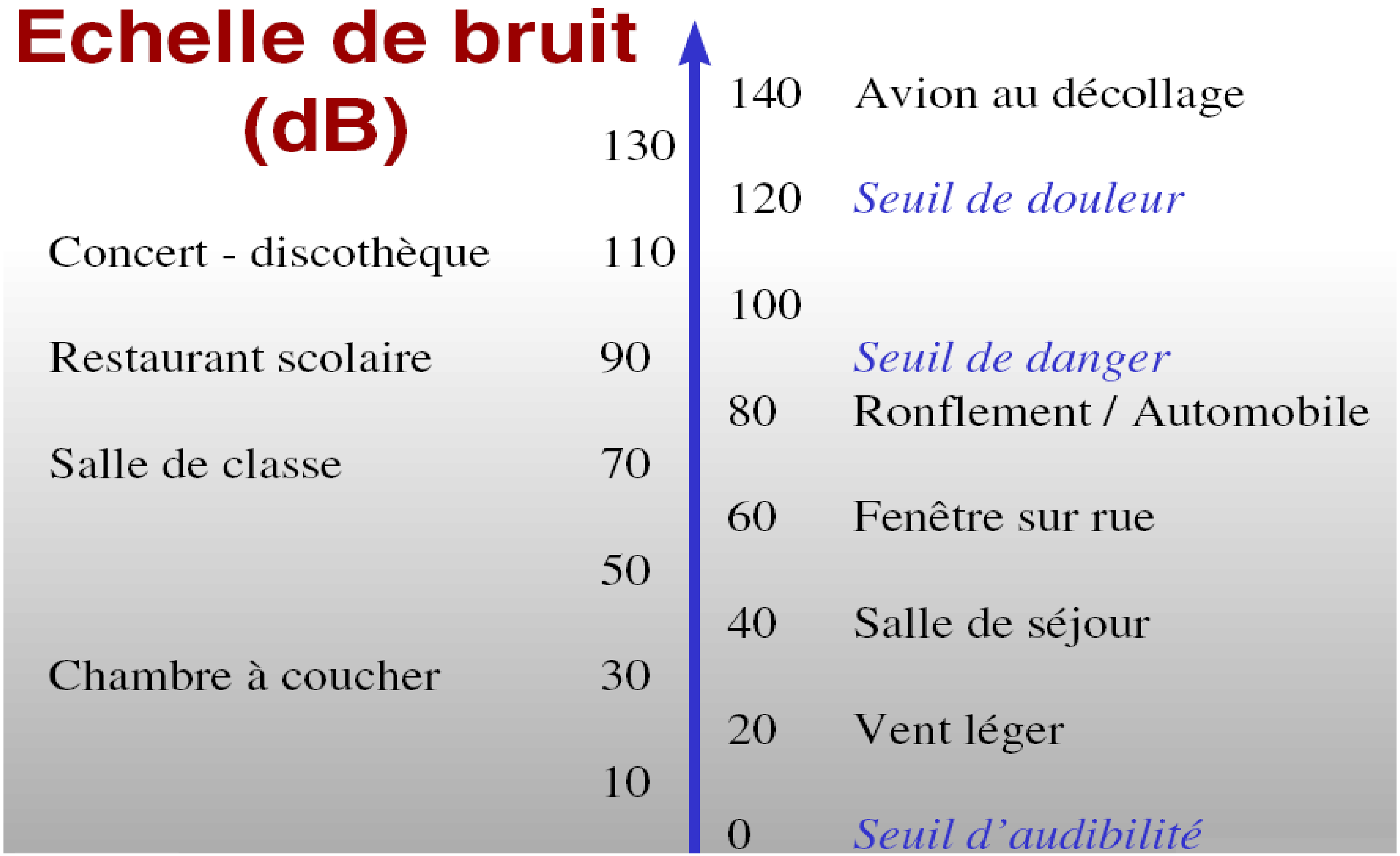


An audiogramme

- **Ratio of hearing threshold to the mean (normalised)**
- **Normal audition is the straight 0 dB line**



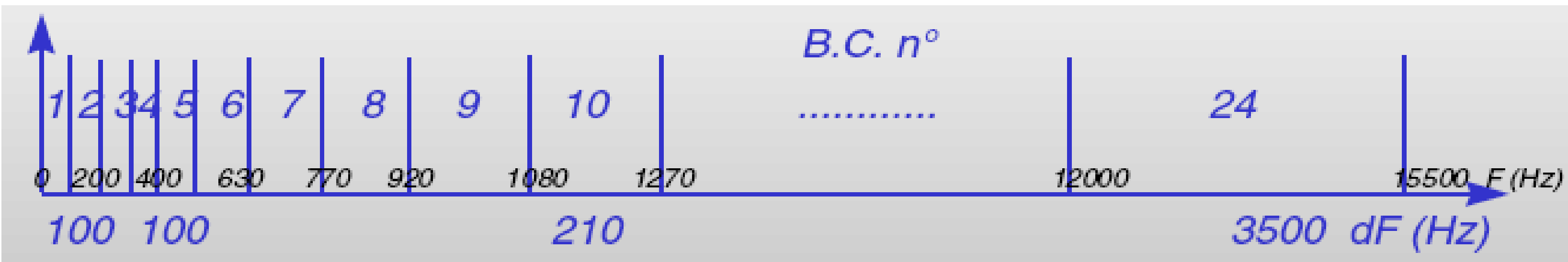
Noise scale


Echelle de bruit (dB)
140 Avion au décollage
130
120 Seuil de douleur
Concert - discothèque 110
100
Restaurant scolaire 90
Seuil de danger
80 Ronflement / Automobile
Salle de classe 70
60 Fenêtre sur rue
50
40 Salle de séjour
Chambre à coucher 30
20 Vent léger
10
0 Seuil d'audibilité

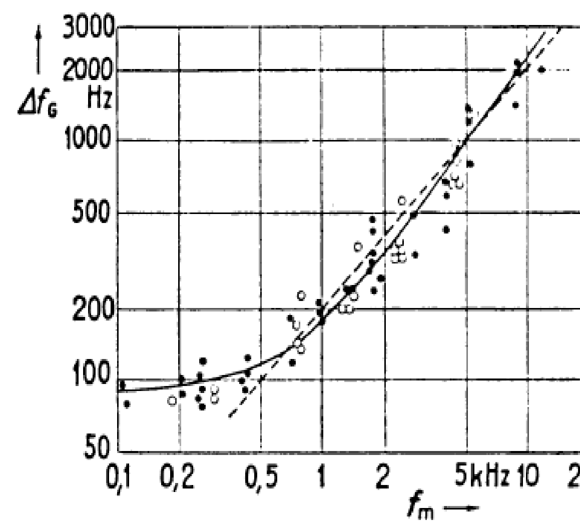
Critical bands

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

  at 1 kHz the filter has approx. 160 Hz bandwidth

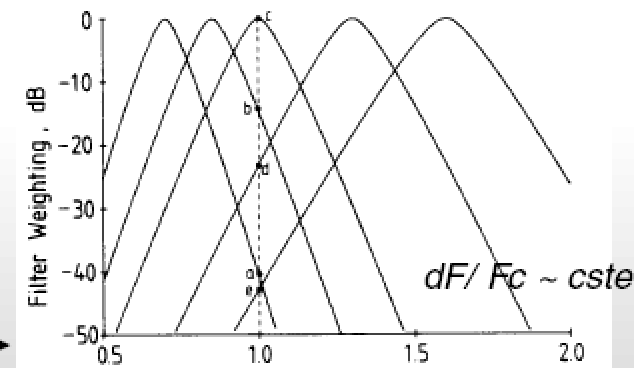


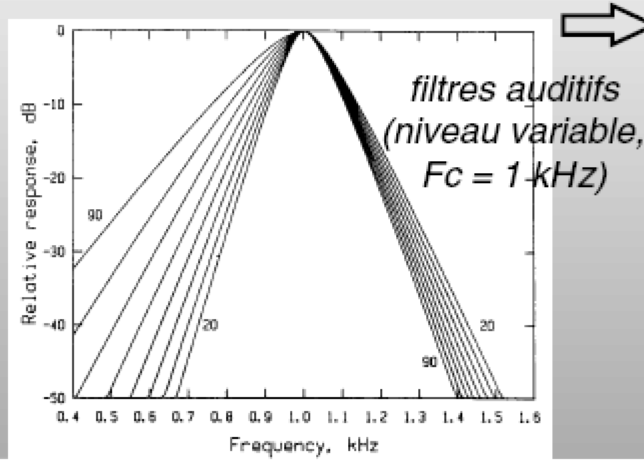
Log-variation of CB bandwidths

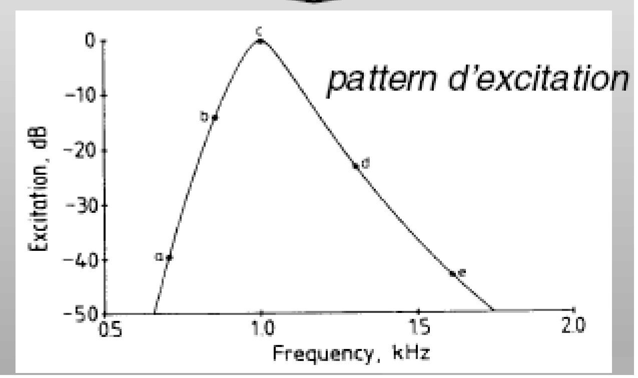


Critical bands : Excitation patterns

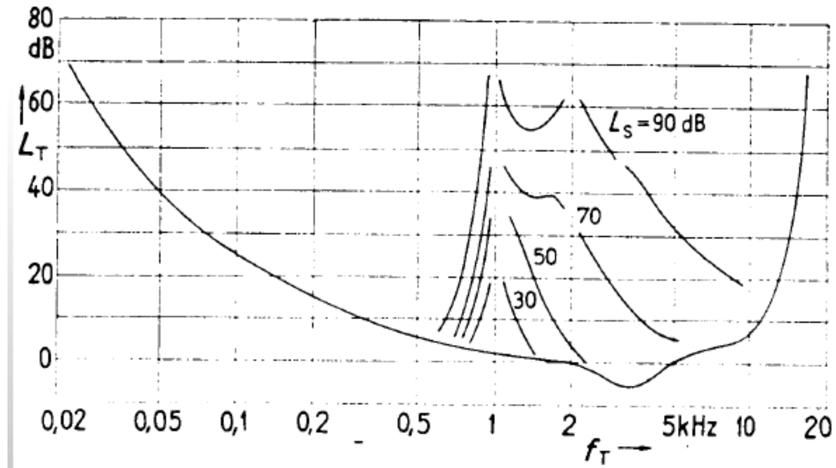
...s'explique (en partie seulement, cf phénomènes actifs et c.c. externes) par la *sélectivité en fréquence* de la membrane basilaire :



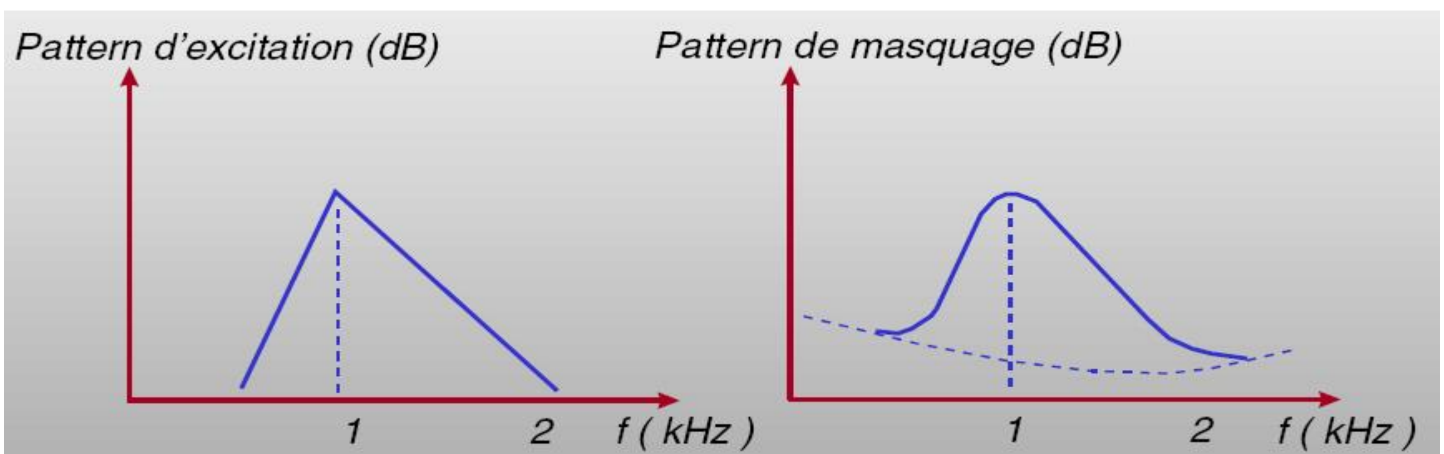




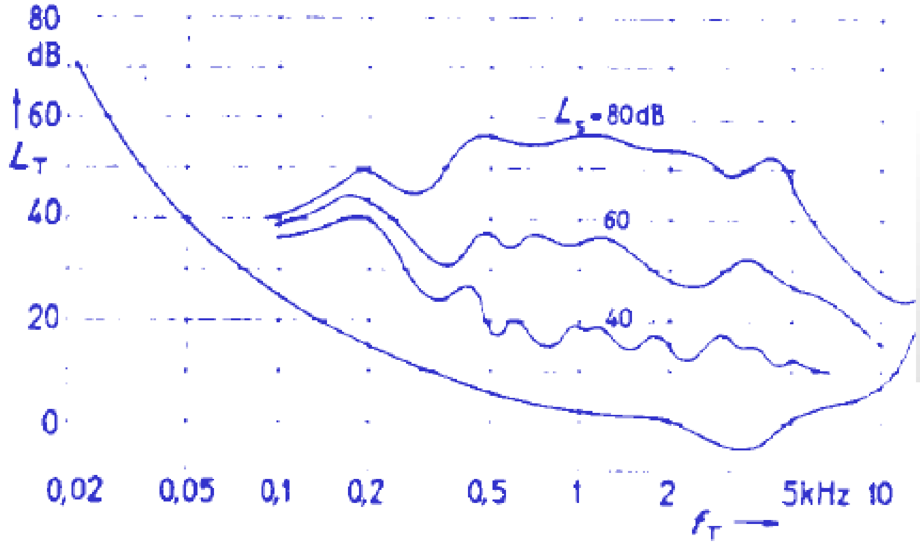
Masking properties of pure sinoidal sounds



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

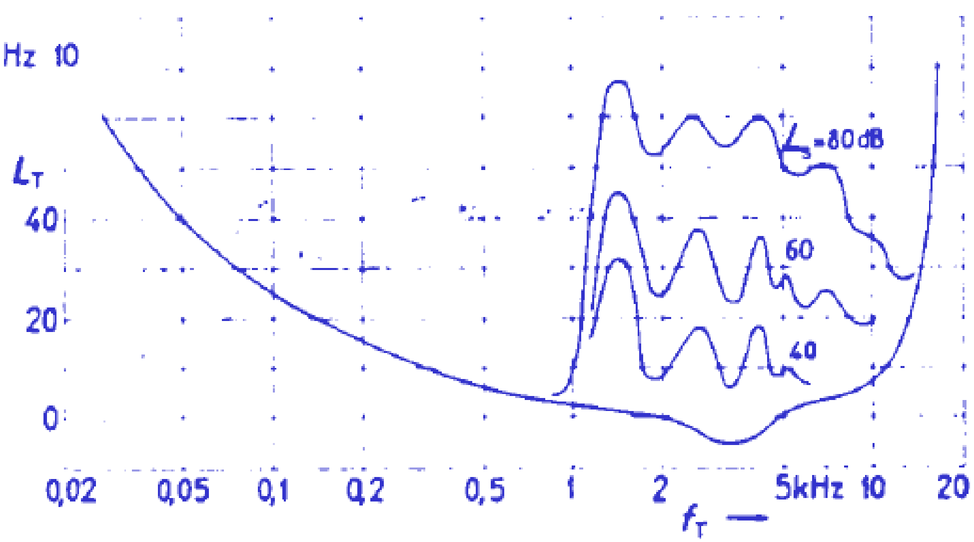


Masking by complex sounds



Courbes d'effet de masque de violons graves

Courbes d'effet de masque de violons aigus



Pitch perception (periodic sounds)

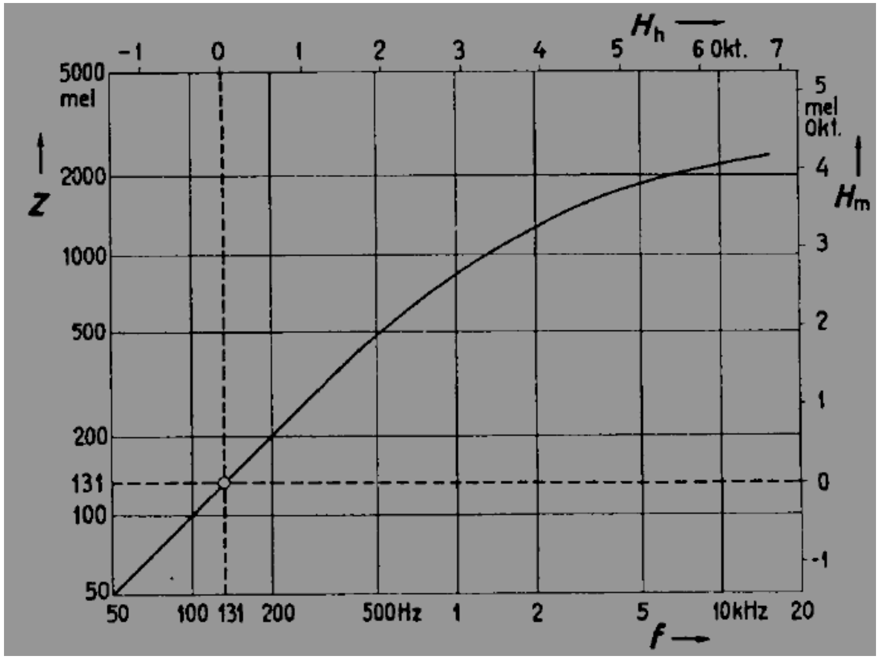
- **Perception of height (pure sinousoidal sounds)**

  - **Tons scale:** The « tonie » doubles if a sound is perceived twice as high as the previous one

  - « Tonie » is proportionnal to the frequency

  - The Mel scale

$$mel(f) = 1000 \log_2(1 + \frac{f}{1000})$$

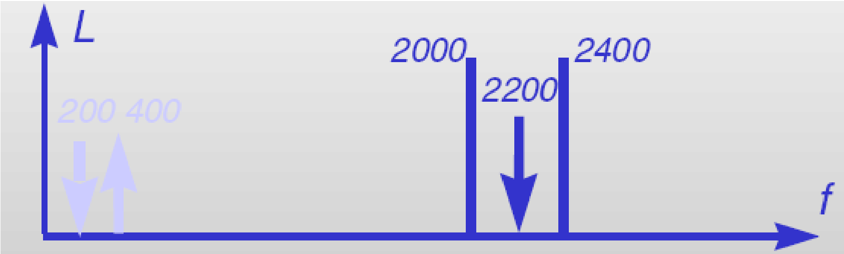


Pitch perception of complex sounds

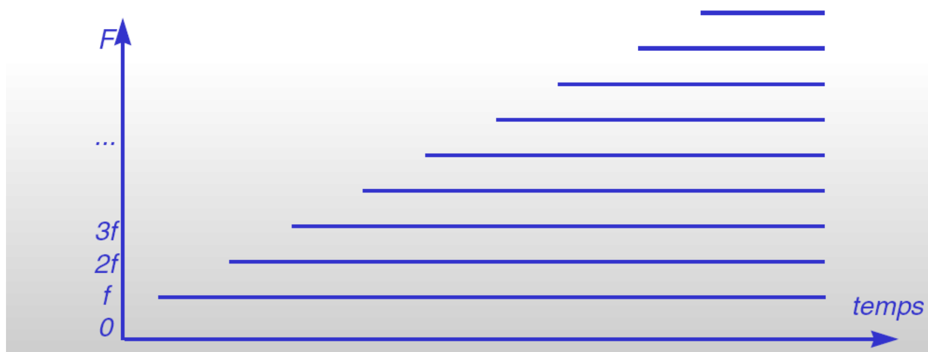
- For sounds composed of “partials”, the ear often synthesizes the perception of these partials to hear one or more pitches. *This is the case of harmonic sounds.*

- But the pitch of harmonic sounds is not dictated by the lowest frequency : in a complex harmonic sound we often still hear the pitch even if the fundamental is absent

- We hear the “Greatest common divisor (PGCD)



Analytic vs Global: Do we hear the individual partials ?



Complete sound

Sound built harmonic per harmonic

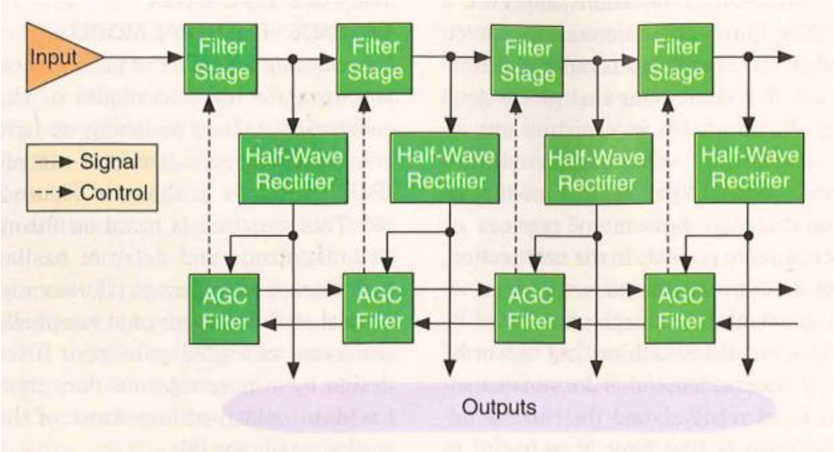
- **Illusion: a sound continuously going down… (JC Risset)**

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

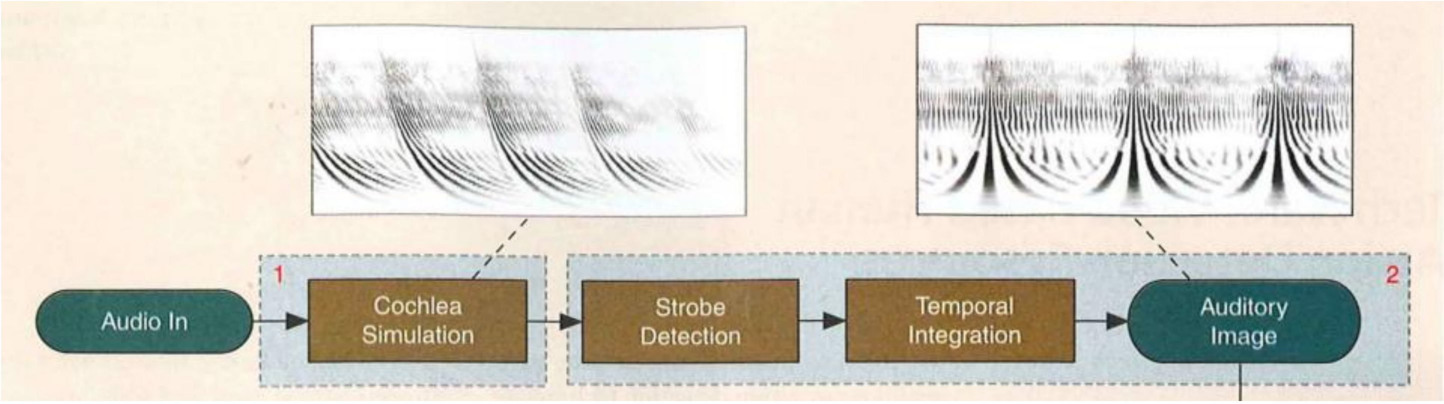
- **« Perceptual » time-frequency representations**
  - Mel-spectrograms
  - CQT (Constant Q transform)
  - Wavelets
  - Gammatone filterbanks
- **« Perceptual » features**
  - MFCC (Mel-frequency Cepstral Coefficients)
- **Psychoacoustics models**
  - In audio coding (e.g. masking patterns)

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

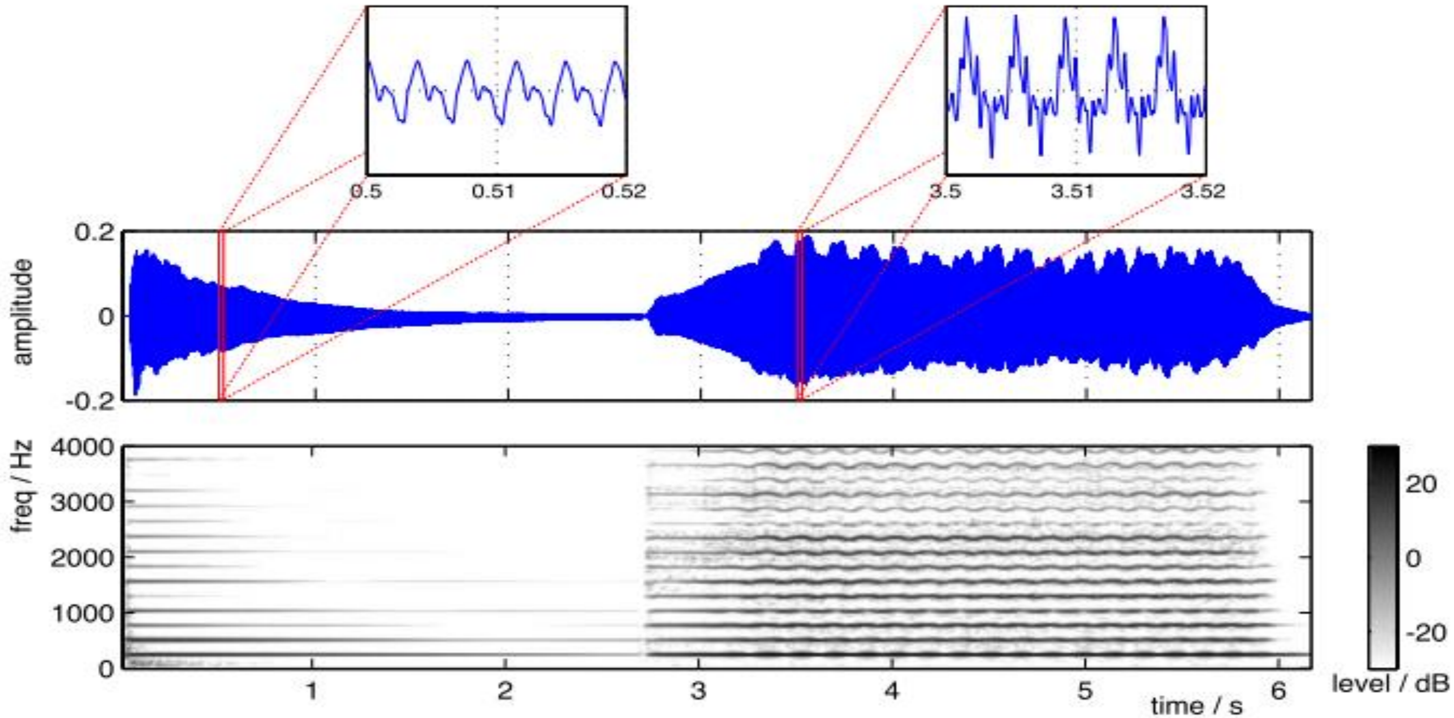
- **Sinusoidal models**
- **Harmonic + noise models**
- **Other « decomposition » models**
  - Sparse representations
  - Non-negative matrix factorization

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

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**

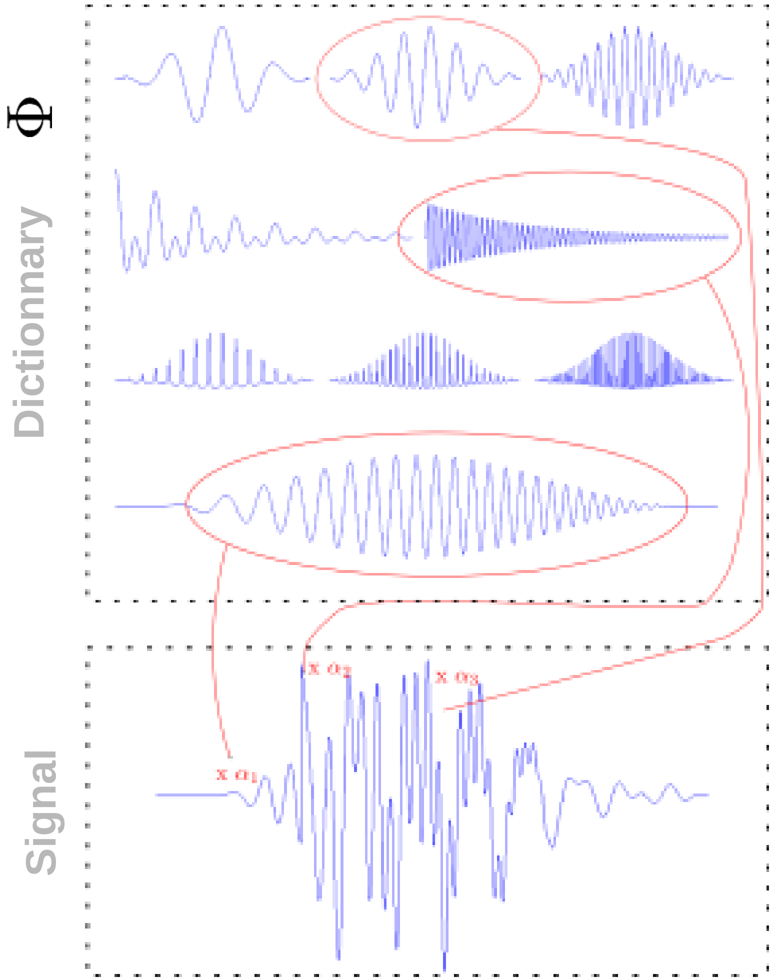
$$x(n) = \sum_{i=1}^{I} A_i(n).sin(2\pi\nu_i n + \phi_i) + m(n).b(n)$$

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

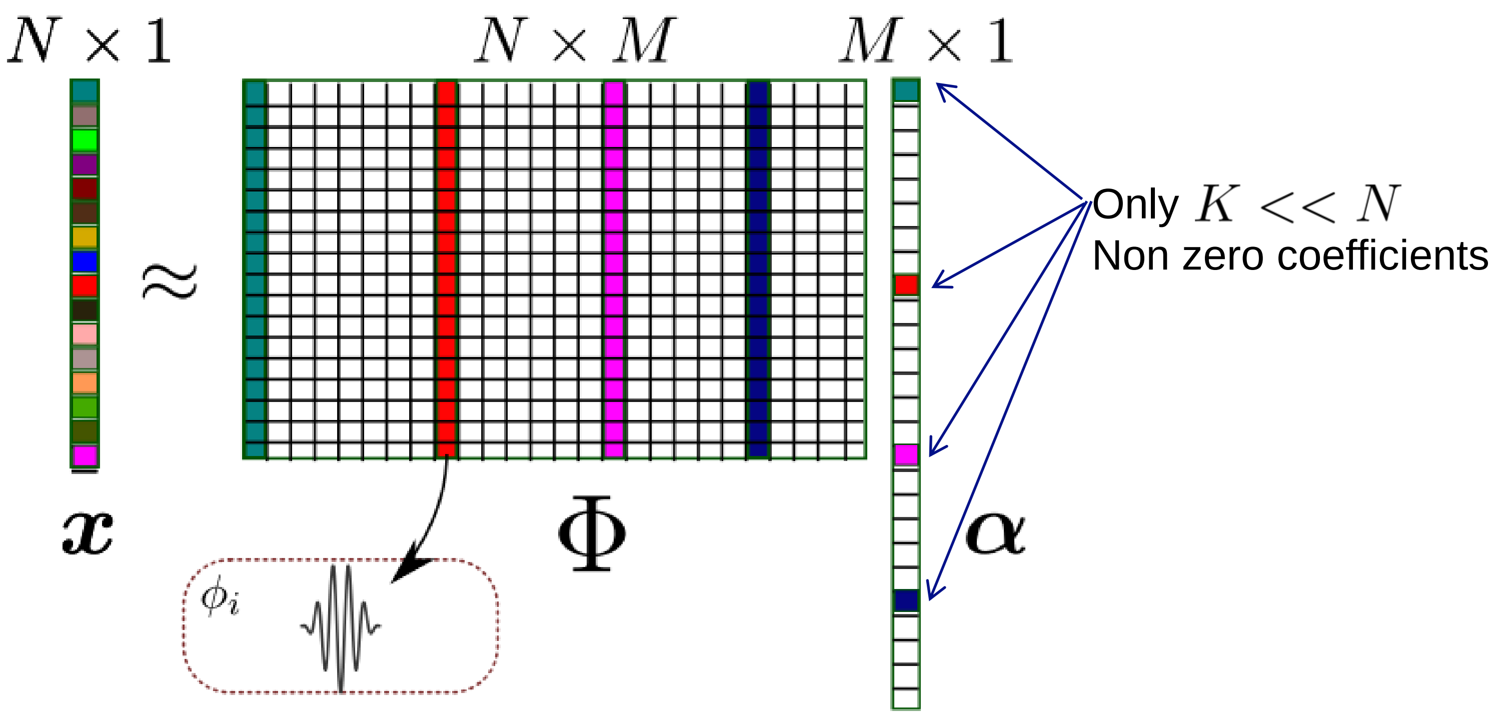
$$\min \|\alpha\|_0 \text{ s.t. } x = \Phi\alpha$$

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*)

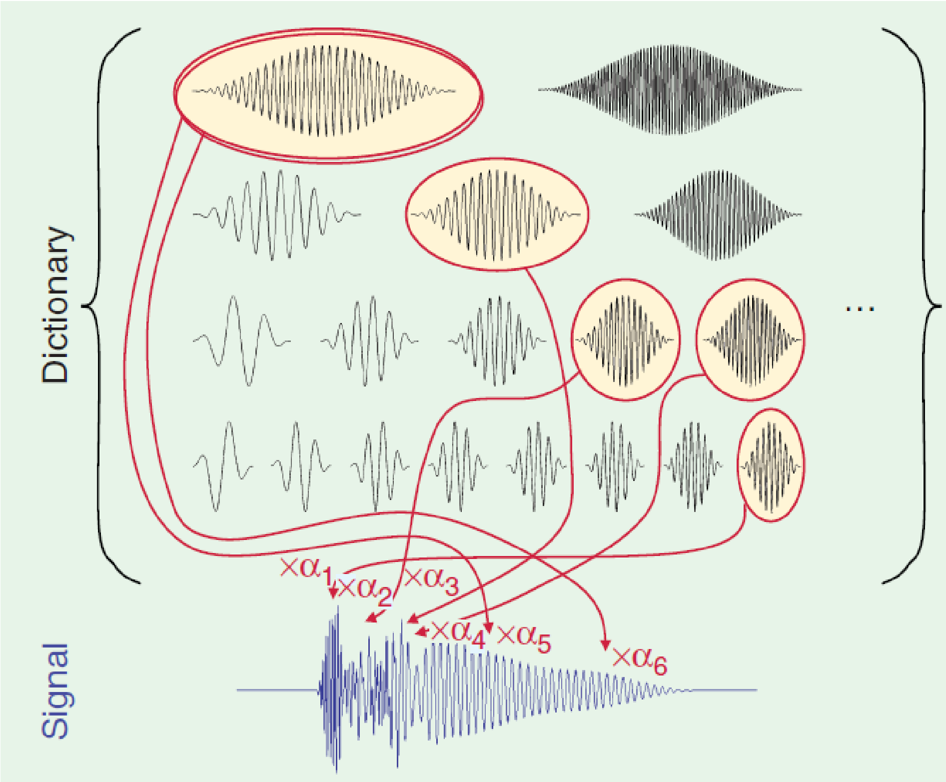
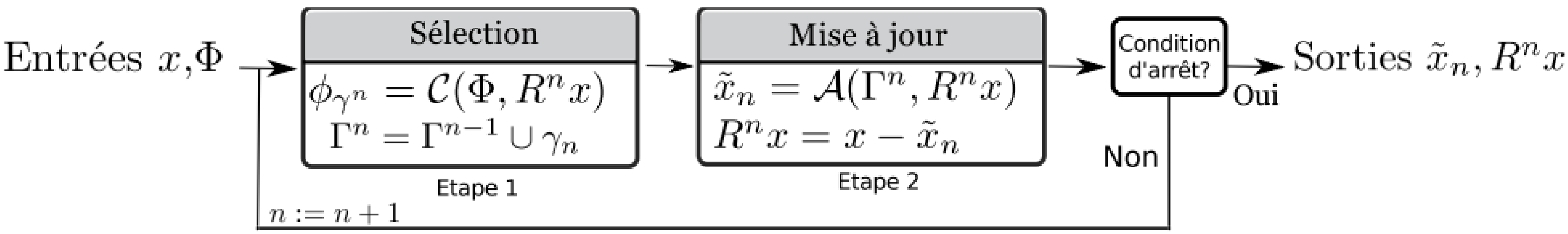


Figure from L. Daudet: *Audio Sparse Decompositions in Parallel,* IEEE Signal Processing Magazine, 2010

Standard Matching pursuit



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

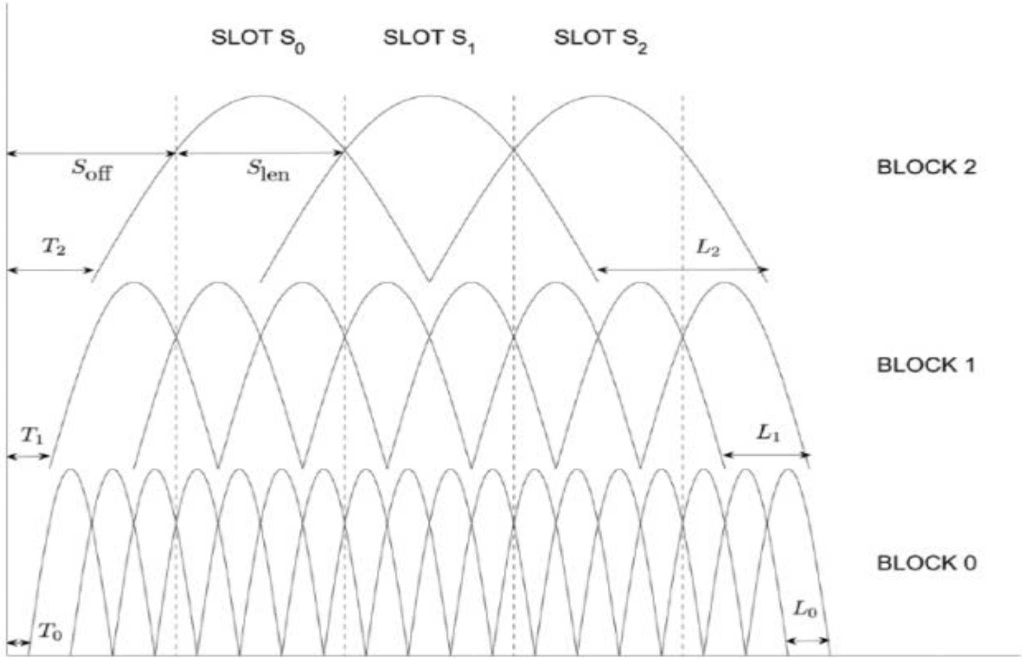
$$\phi_{\gamma^n} = \arg\max_{\phi_i \in \Phi} |\langle R^n x, \phi_i \rangle|$$

- **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)

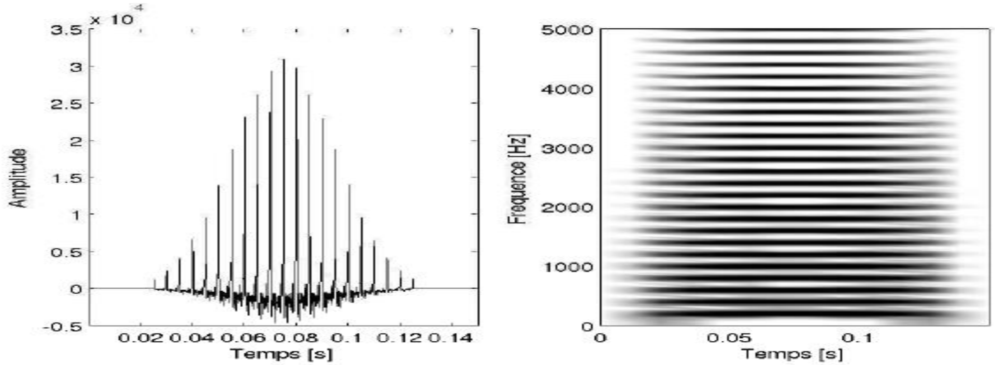
Use in music transcription

- **Harmonic atoms**

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

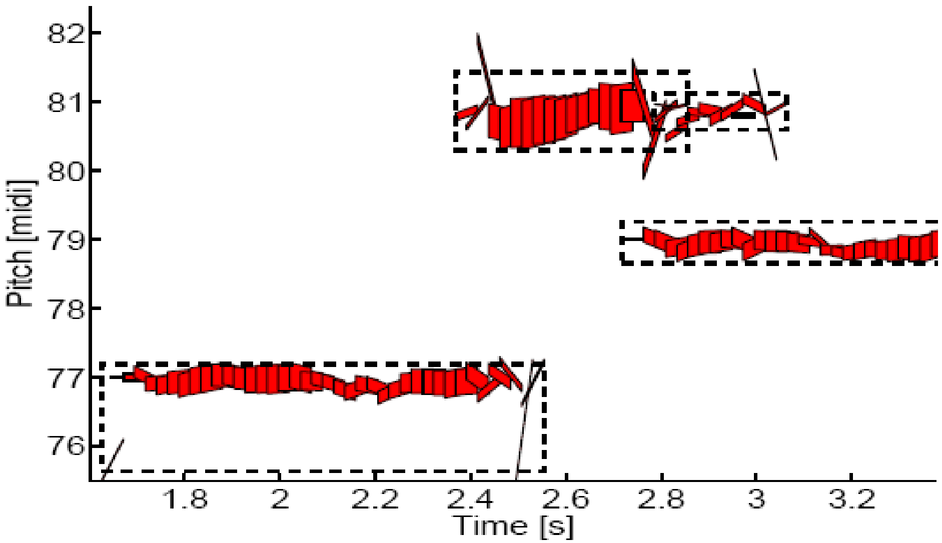
  - a_m (resp ϕ_m) amplitudes (resp. phases) des partiels
  - s paramètre d'échelle
  - u localisation temporelle
  - f_0 (resp c_0) fundamental frequency and chirp rate

(from P. Leveau & al.2008)

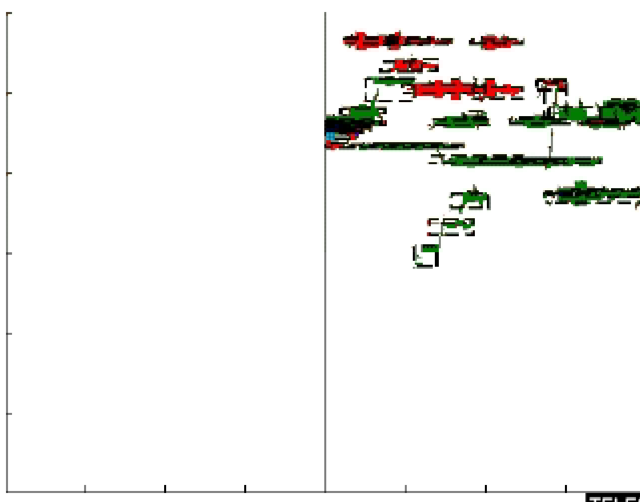


Use in music transcription

- **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 :**

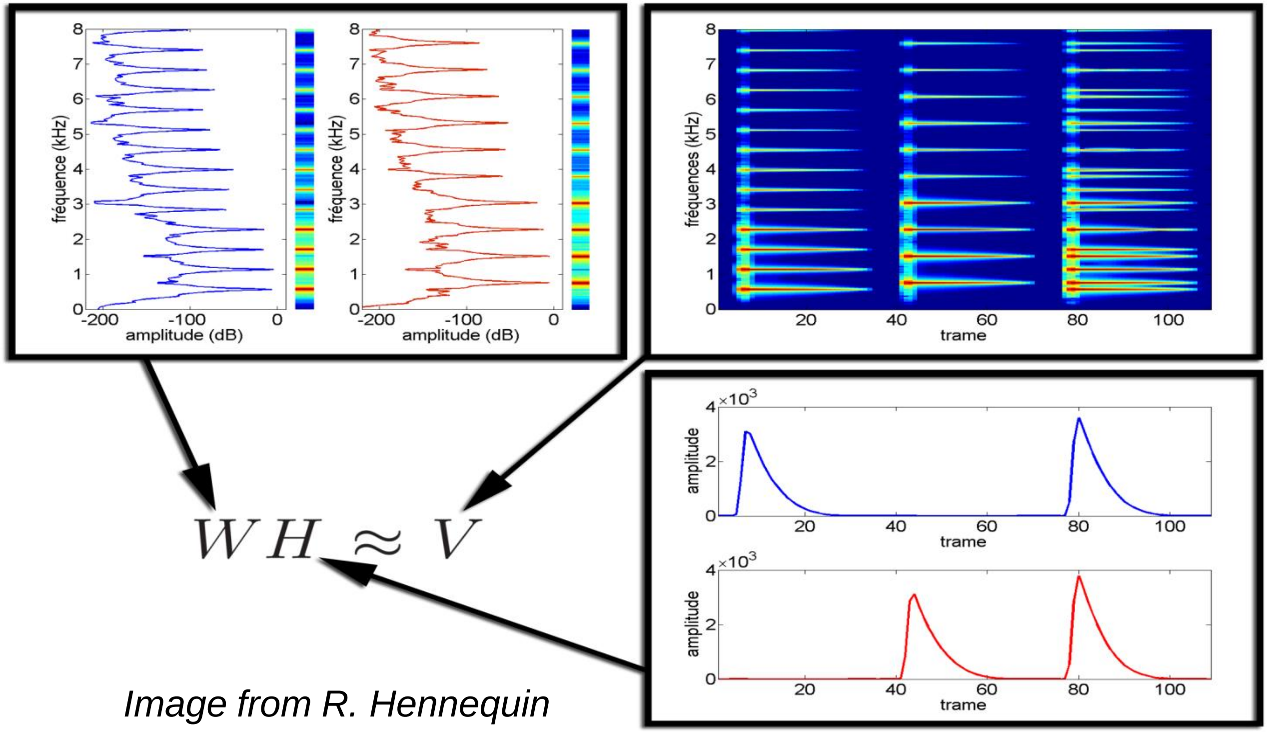


Image from R. Hennequin

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**

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

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

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 .

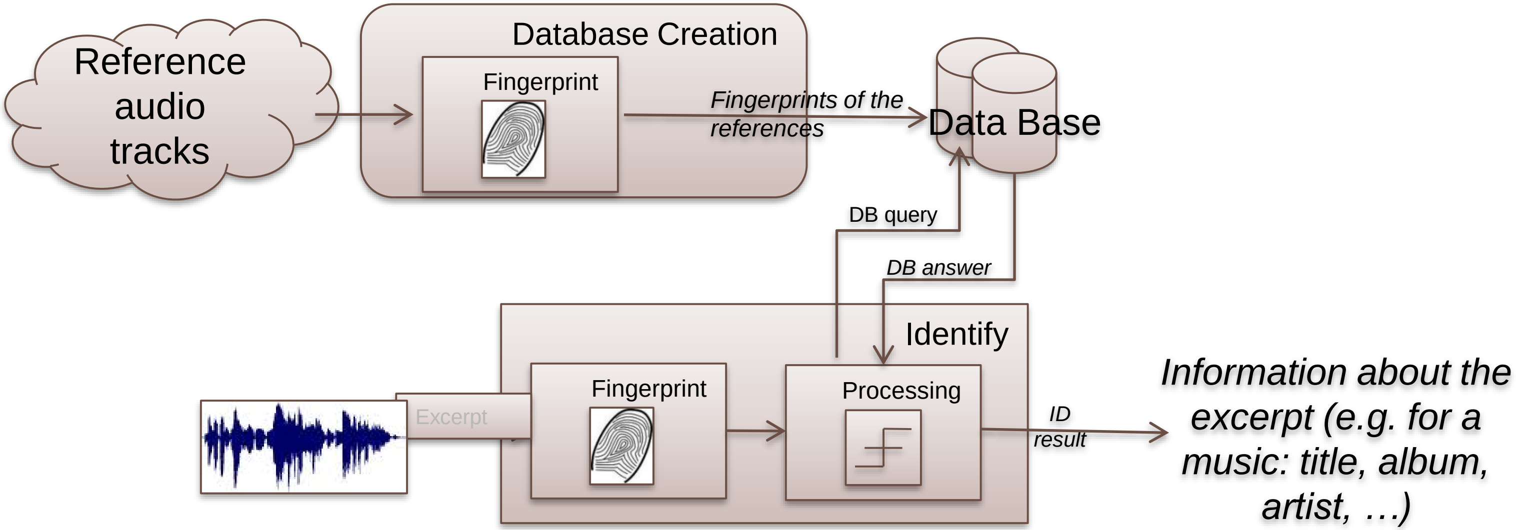
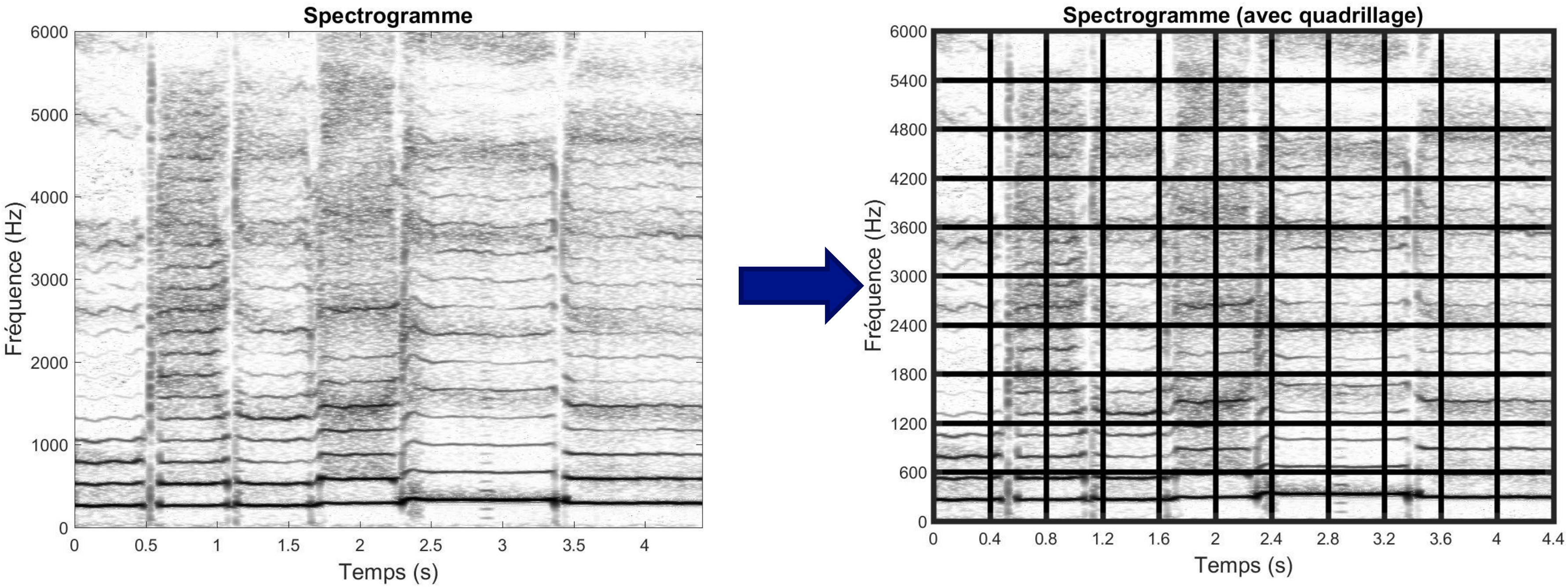


Figure from Sébastien Fenêt

TELECOM Paris
IP PARIS

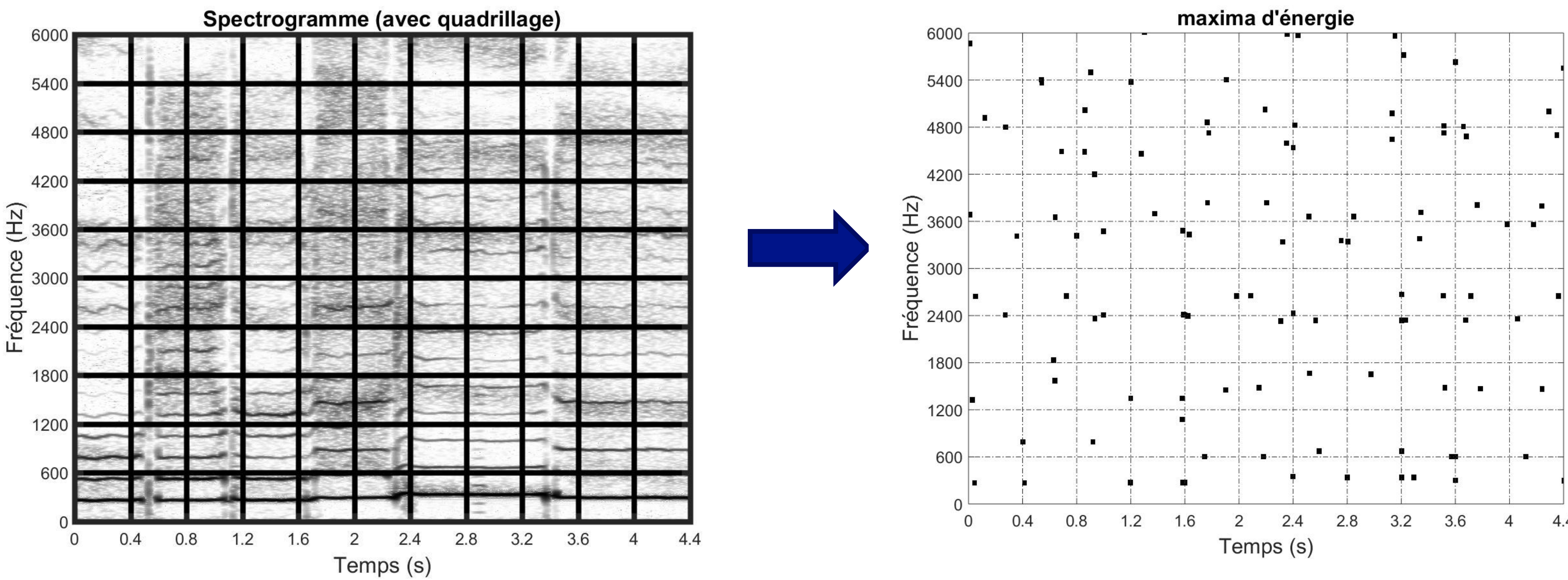
Signal model : from spectrogram to “schematic binary spectrogram”

- **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

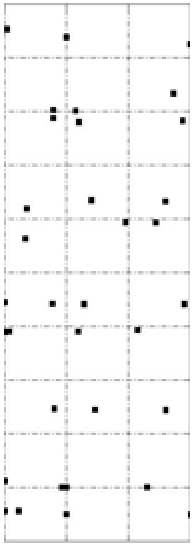
- **Towards idetifying an Unknown recording using a large database of known references**

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

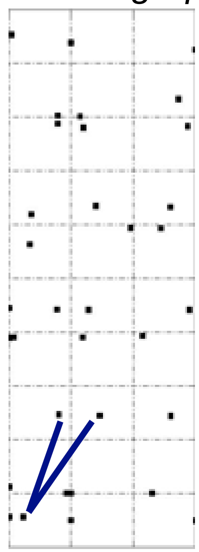
- **Towards idetifying an Unknown recording using a large database of known references**

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 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:**
  - 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.

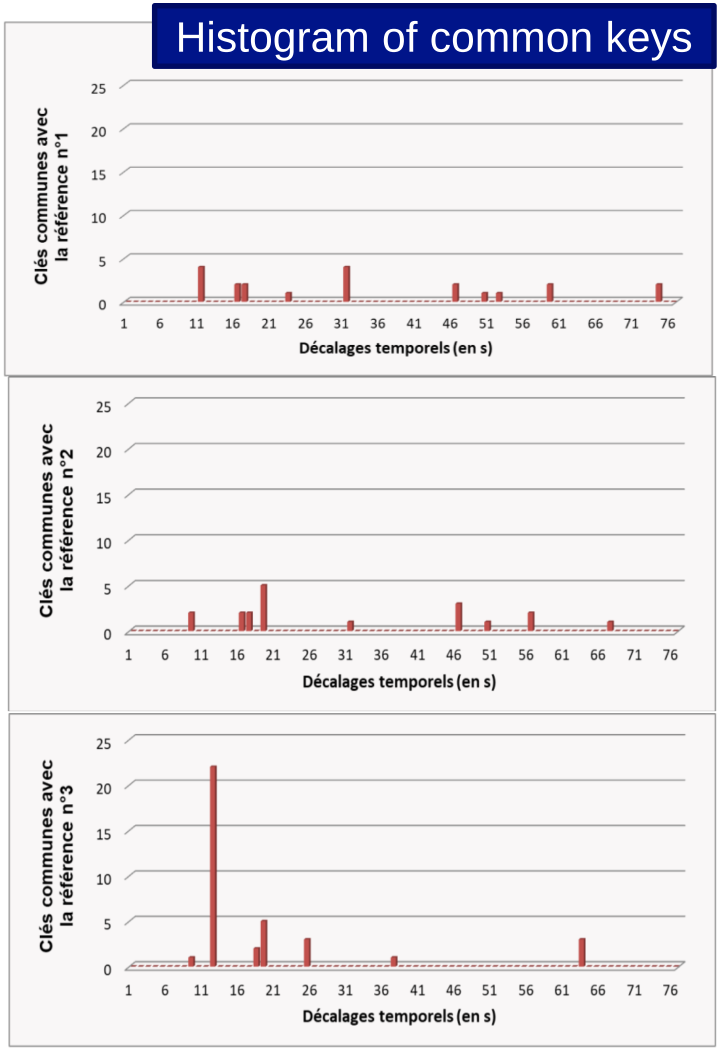
  3. **Recognition:** The reference which has the most keys in common at a constant Δt is the recognized recording

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

Reference 1

Reference 2

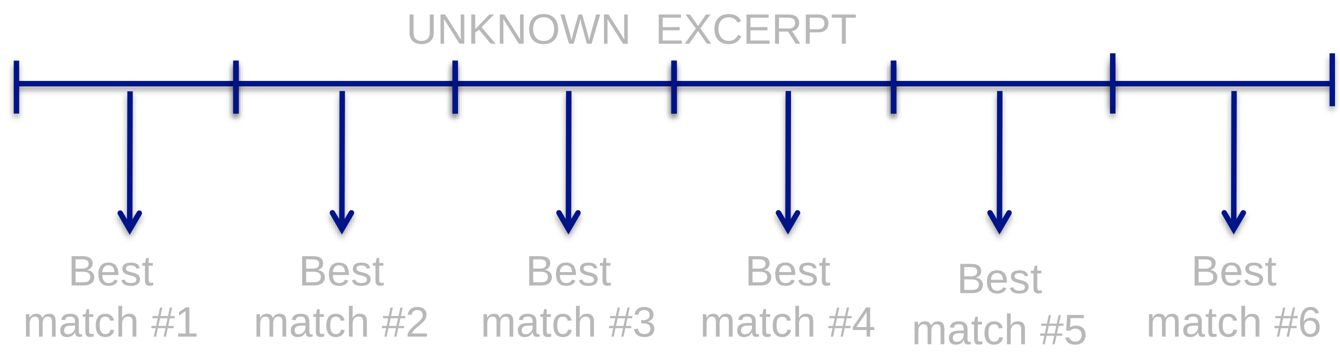
Reference 3



Recognized recording

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

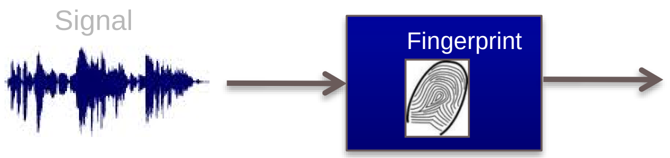
- The unknown recording is divised in sub-segments
- For each sub-segment, the algorithm gives back a best candidate



- 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**



$$\left.\begin{array}{c}(key_1, t_1)\\(key_2, t_2)\\\dots\\(key_n, t_n)\end{array}\right]$$

Signal fingerprints

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

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

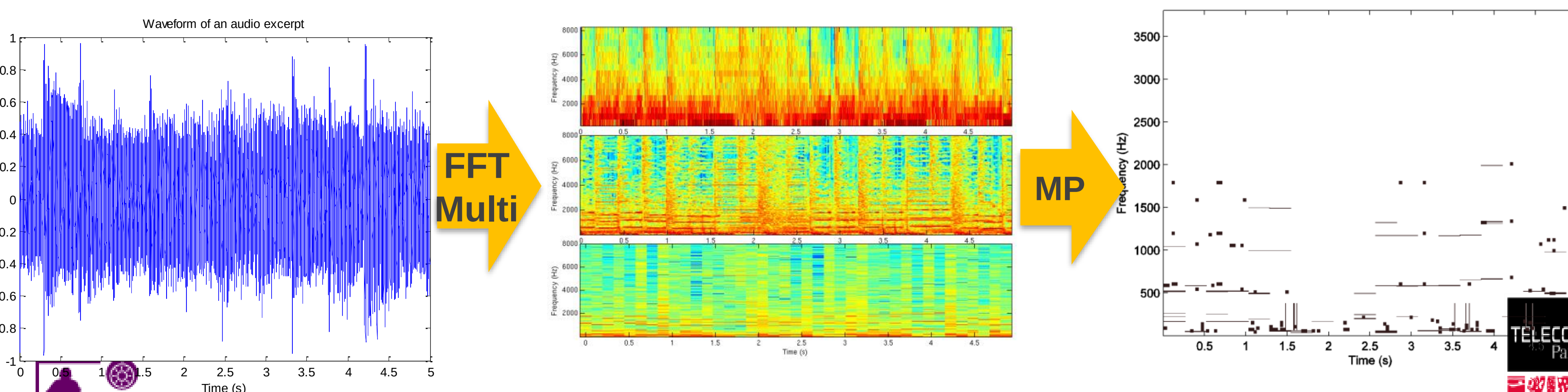
$$\mathcal{M}(\phi_i | \Gamma^n) = 1 - \max_{\gamma \in \Gamma^n} |\langle \phi_i, \phi_\gamma \rangle|$$

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

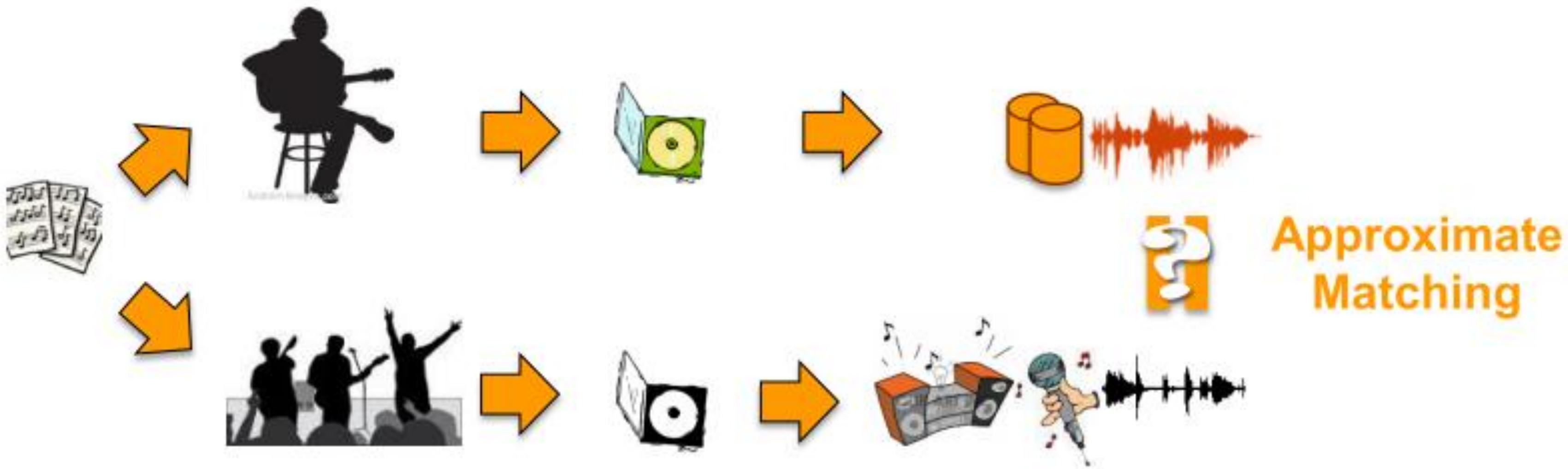
Limitations and other solutions

- **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

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