Master MVA

Analyse des signaux Audiofréquences

Audio Signal Analysis, Indexing and Transformation

https://perso.telecom-paristech.fr/grichard/Enseignements/MVA/

Lecture on

Audio indexing or Machine Listening

Gaël RICHARD

Télécom Paris

Image, Data, Signal department

January 2026

Master MVA

Analyse des signaux Audiofréquences

Audio Signal Analysis, Indexing and Transformation

Registration to the course:
https://partage.imt.fr/index.php/s/XDFJ94EYenBPdTZ

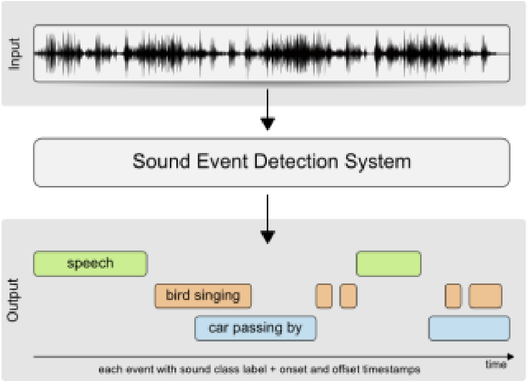
(important for communication/organisation)

Note: Labs will be done on your own computer except if your have have an account at Telecom Paris
(due to administration difficulties to rapidly open computer accounts using ecampus)

Aim of the course:

- To span several domains of audio signal processing including:
  - **Audio indexing/recognition or Machine listening**
  - **Audio models (**High-resolution spectral analysis)
  - **Sound rendering and transformation**

  (3D audio, audio effects, source separation)



Philosophy of the course:

- Lectures (15h) followed by Labs (TP, 7,5h) in Python (*or Matlab if preferred*)
- Course validation: papers reading/presentation + reports on Labs

Professors: Gaël Richard and Roland Badeau

Audio Signal Analysis, Indexing and Transformation
some details

- **Audio Indexing or Machine listening (3H lecture, 1,5H TP):**
  - audio signal analysis for content-based information retrieval (automatic music genre recognition, automatic musical instrument identification, tempo or downbeat estimation,...), Deep learning for audio.
- **High resolution methods (3H lecture, 3H TP)**
  - Beyond Fourier resolution, ESPRIT, MUSIC, sinusoidal models
- **Audio source separation (3H lecture; 1,5H TP):**
  - Audio source models, Mixing models (instantaneous, convolutive). Blind source separation methods, time vs Frequency domains methods, under-determined case, sparse models, DUET
- **3D audio rendering (3H lecture; 3H TP):**
  - Perceptual vs physical based approaches (binaural/transaural, holophony). Sound effects synthesis (artificial reverberation, distorsion, flanger,…)
- **Sound transformation (1,5H lecture, 1,5 TP)**
  - Pitch scaling, time scaling, phase vocoder..

Audio Signal Analysis, Indexing and Transformation
Planning

- **All lectures/TP @ Telecom Paris, 19 place M. Perey, Palaiseau, Wednesday afternoon from January 7th to March 18th (oral exam)**

Day	Time	Type	Title	Room	Professor
Wed 07/01/2026	13:30-15:00	Lecture	Audio , signal analysis and machine listening	0C03	Gaël RICHARD
	15:15-16h45	Lecture	Audio , signal analysis and machine listening	0C03	Gaël RICHARD
Wed 14/01/2026	13:30-15:00	Lecture	Deep learning for audio	1A260	Gaël RICHARD
	15:15-16h45	TP	Music signal analysis	1A260	Gaël RICHARD
Wed 21/01/2026	13:30-15:00	Lecture	Timbral, Scale, Pitch modifications	1A260	Roland BADEAU
	15:15-16h45	TP	Timbral, Scale, Pitch modifications	1A260	Roland BADEAU
Wed 28/01/2026	13:30-15:00	Lecture	Source Separation	1A242	Roland BADEAU
	15:15-16h45	Lecture	Source Separation	1A242	Roland BADEAU
Wed 04/02/2025	13:30-15:00	Lecture	High resolution methods	1D23	Roland BADEAU
	15:15-16h45	TP	Source Separation	1D23	Roland BADEAU
Wed 11/02/2026	13:30-15:00	Lecture	Sound effects and Reverberation	1A207	Gaël RICHARD
	15:15-16h45	TP	Sound effects and Reverberation	1A207	Gaël RICHARD
Wed 04/03/2026	13:30-15:00	Lecture	High resolution methods	1A207	Roland BADEAU
	15:15-16h45	TP	High resolution methods	1A207	Roland BADEAU
Wed 11/03/2026	13:30-15:00	Lecture	3D Sound Rendering	1A207	Gaël RICHARD
	15:15-16h45	TP	3D Sound Rendering	1A207	Gaël RICHARD
Wed 18/03/2026	13:30-16:45	Oral	Exam	1A340 1A344	Gaël RICHARD, Roland BADEAU

- More info on the dedicated web site:
  - https://perso.telecom-paristech.fr/grichard/Enseignements/MVA/
  - Documents: « polycopié » + slides + research papers

Objective of this lecture

Audio Indexing and machine listening

- Understanding what is an audio signal
- Understanding how to represent essential dimensions of the audio signal
- Illustrating specific machine learning tasks in audio with some examples
- A view of Deep learning for audio
- A Lab (TP) on « multiple frequency estimation »

Audio Indexing and machine listening : Content

- **Introduction**
  - Interest and some applications
  - A few dimensions of musical signals
  - Some basics in signal processing

- **Analysing the music signal**
  - Pitch and Harmony,…
    - *Pitch estimation, Chord recognition, Audio recognition*
  - Tempo and rhythm,…
  - Timbre and musical instruments,..

- **A view of Deep learning for audio**

- **Some other machine listening applications**
  - Audio scene recognition
  - Audio-based video search for music videos

Foreword….

- **Lecture largely based on :**
  - *M. Mueller, D. Ellis, A. Klapuri, G. Richard « Signal Processing for Music Analysis, IEEE Trans. on Selected topics of Signal Processing, Oct. 2011*

- ***With the help for some slides from :***
  - *O. Gillet,*
  - *A. Klapuri*
  - *M. Mueller*
  - *S. Fenet*
  - *V. Bisot*
  - *O. Cifka*
  - *S. Durand*
  - *S. Leglaive*

Machine listening

AI applied to Audio analysis, understanding and synthesis by a machine

A fast growing interdisciplinary field with many applications

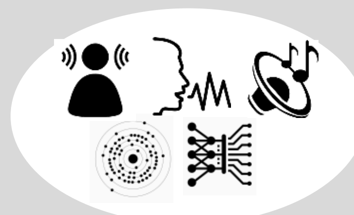
Audio surveillance, Audio scene analysis

Security, Health monitoring, bioacoustics



Industry

Predictive maintenance



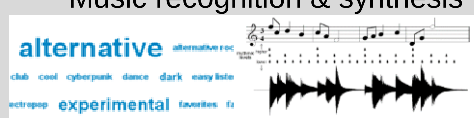
Transport & Communications

Autonomous cars, audio enhancement

Entertainment, Creativity

Music recommendation, sound design

Music recognition & synthesis



Search by content…..

Why analysing the music signal ?

Search by content

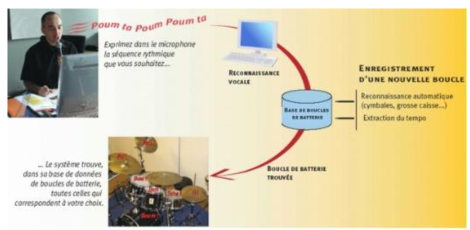
- From a music piece …
- From a hummed query…
- New music that I will like/love ….
- A cover version of my favorite title
- A video that matches a music piece..
- …

New applications

- Semantic playlist (play music pieces that are gradually faster …)
- « Smart » Karaoké (the music follows the singer…)
- Predict the potential success of a single
- Automatic mixing, Djing, music synthesis
- Active listening, style stransfer,…

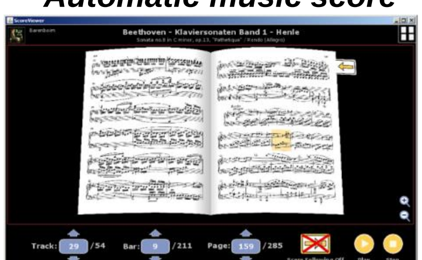
Music streaming services

Search by voice

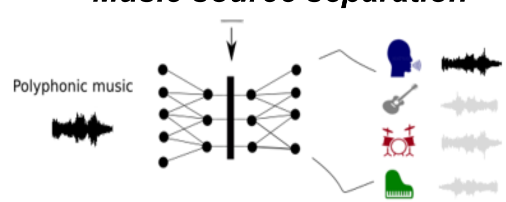


Musical Jogging

Automatic music score



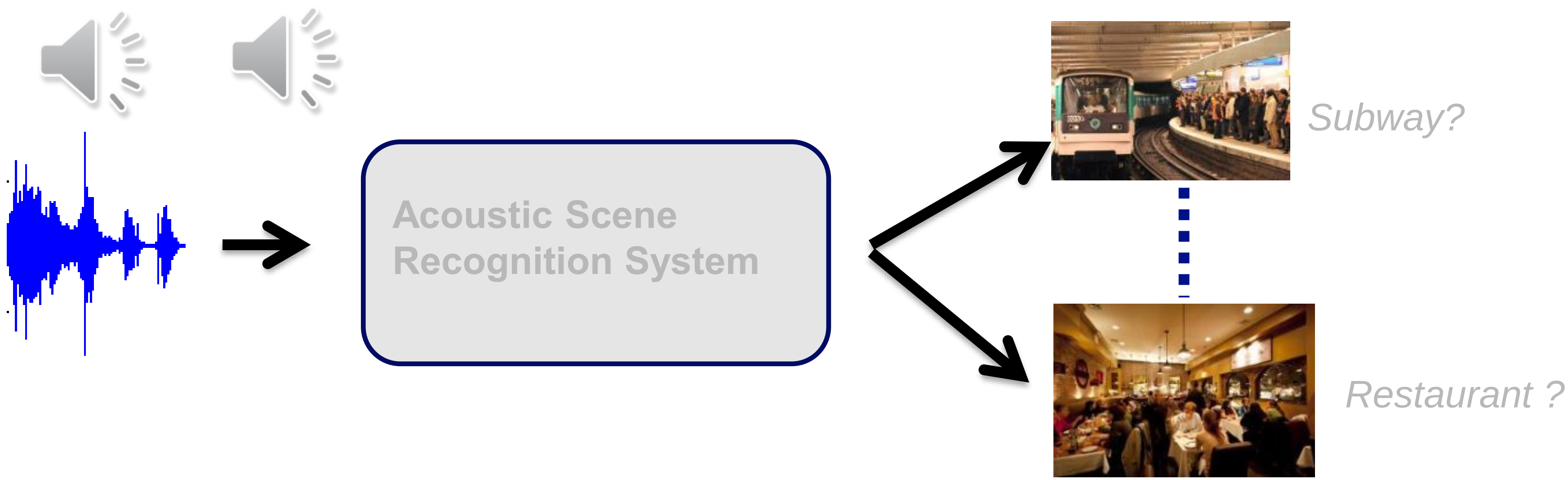
Music source separation



Music generation

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 »

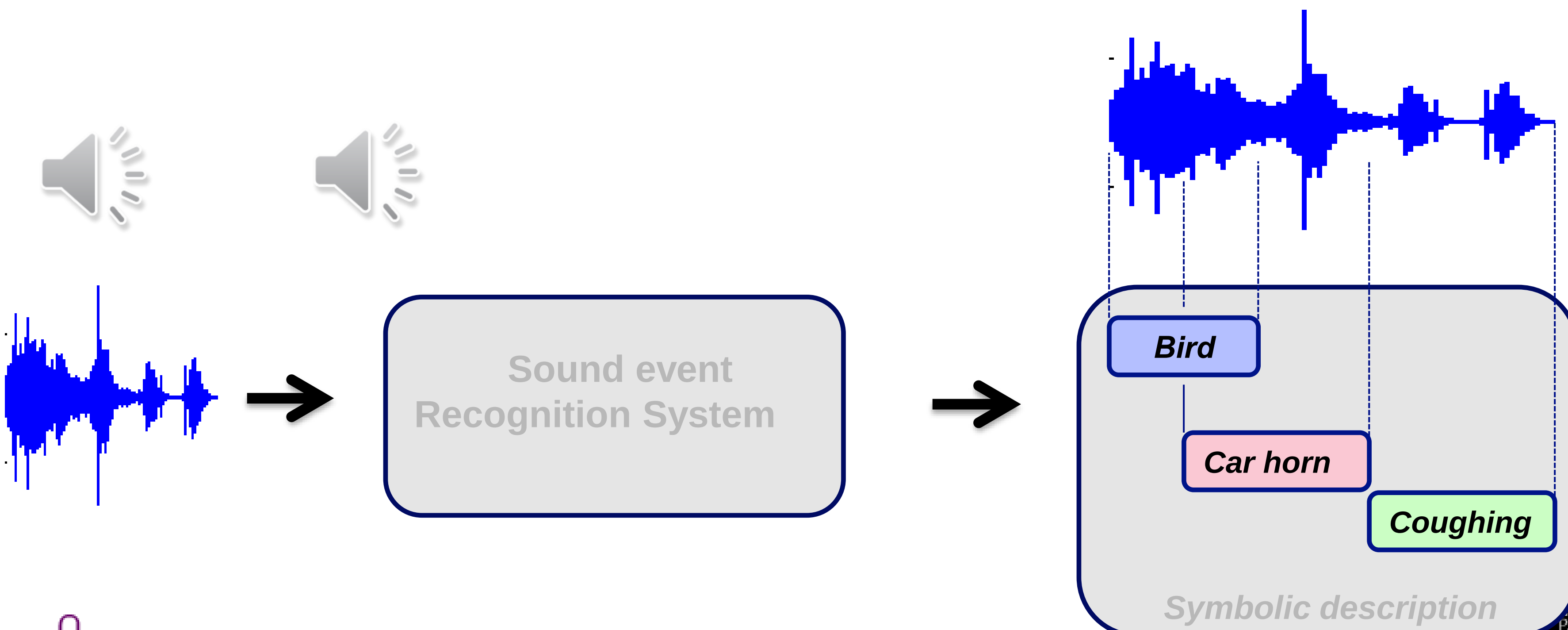


  - Related to CASA (*Computational* Auditory Scene Recognition) and SoundScape cognition (*psychoacoustics*)

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

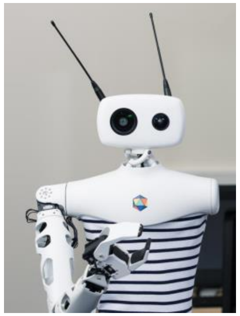
Acoustic scene and sound event recognition

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

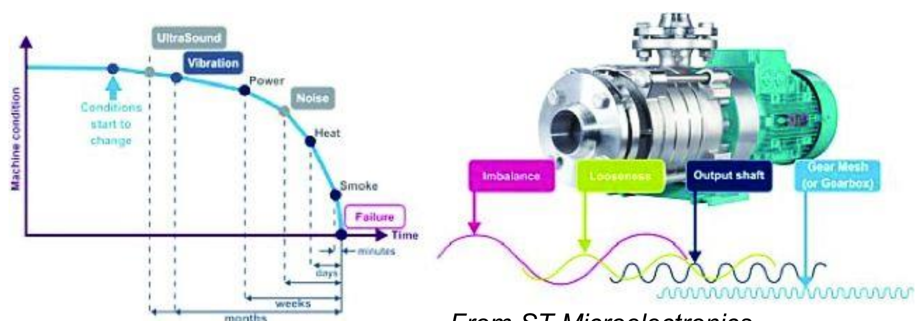


Applications of scene and events recognition

- **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



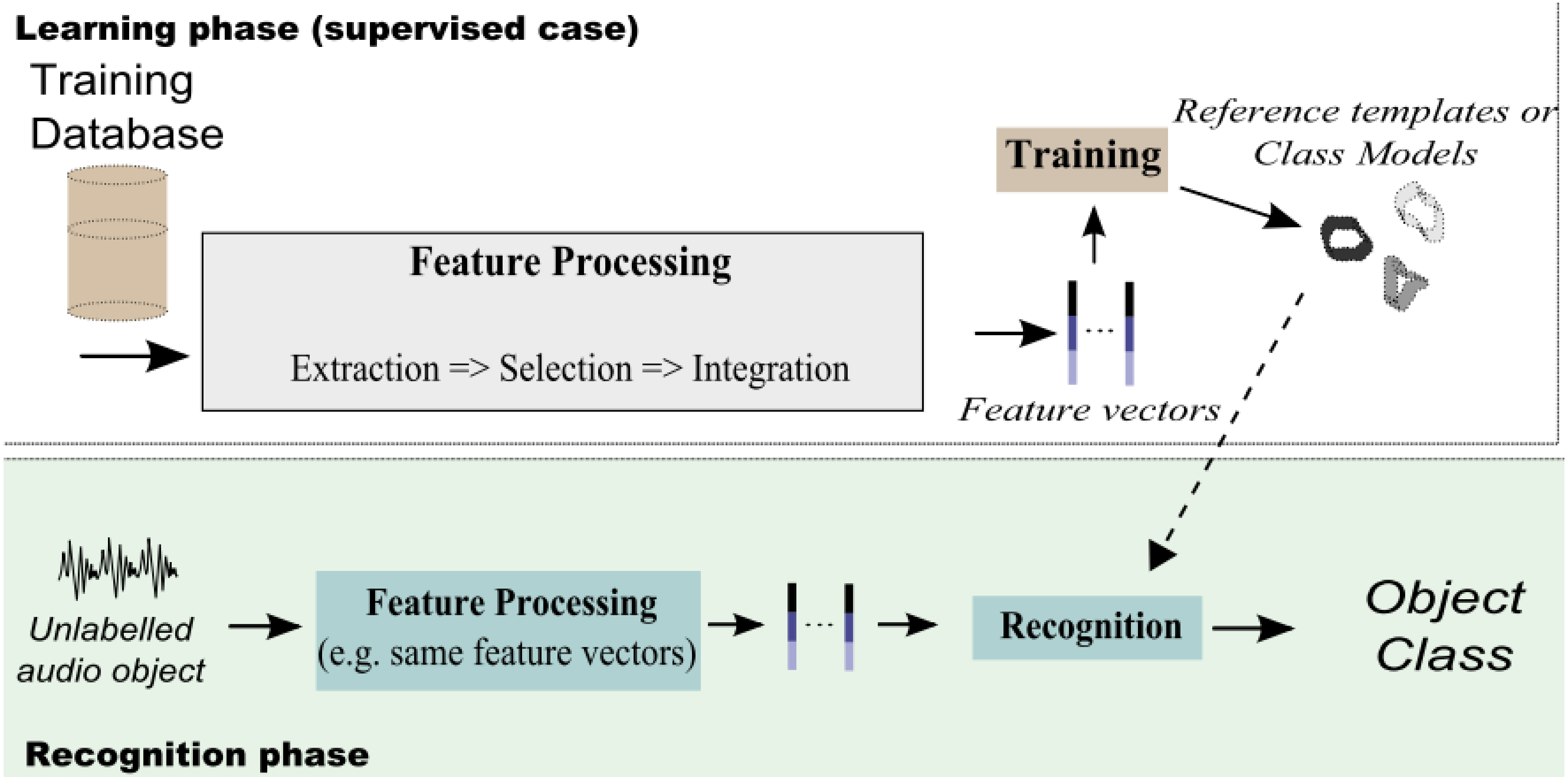
From ST Microelectronics

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...

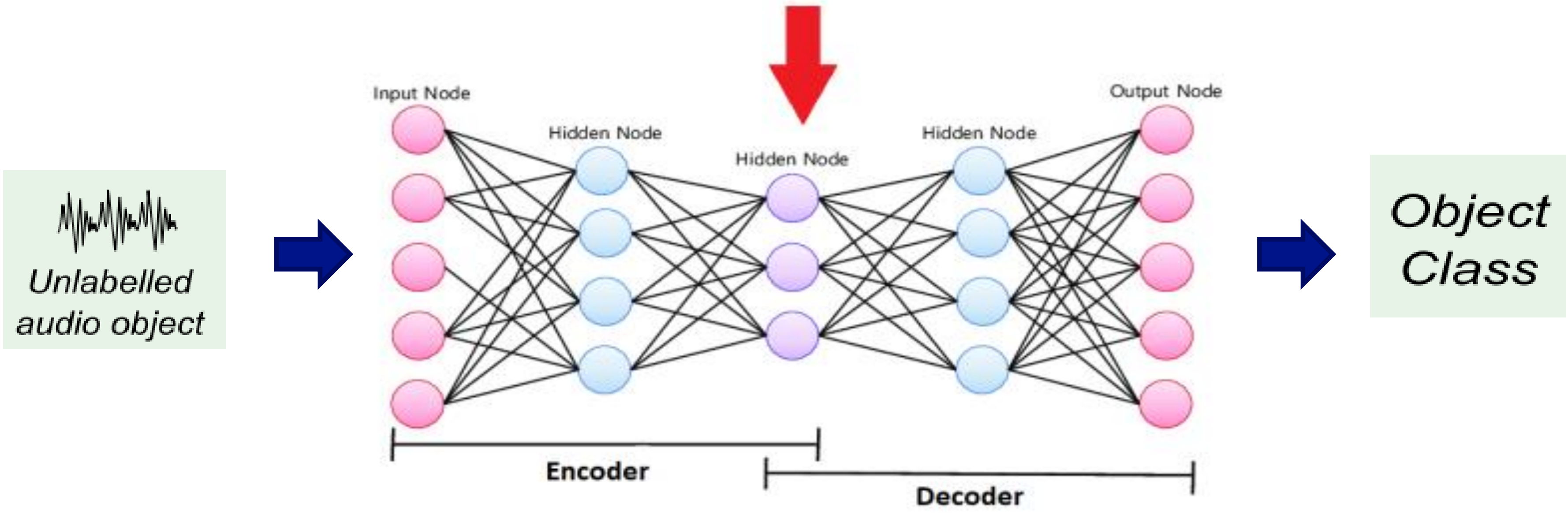
Traditional Classification system



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

Current trends in audio classification

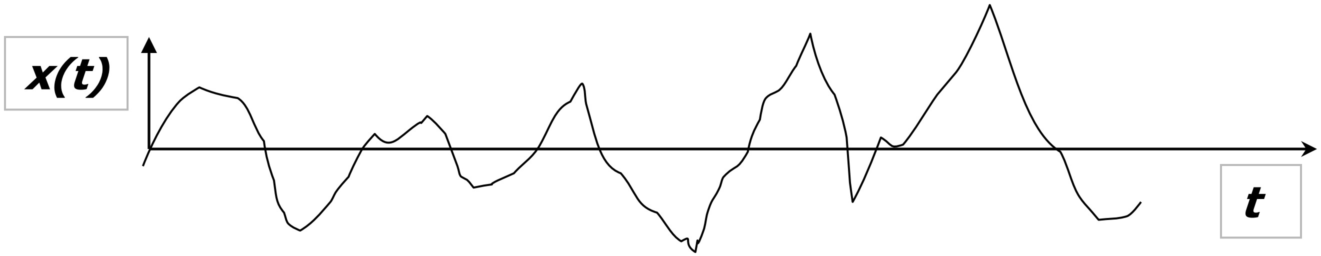
- **Deep learning now widely adopted**
  - For example under the form of encoder/decoder for representation learning



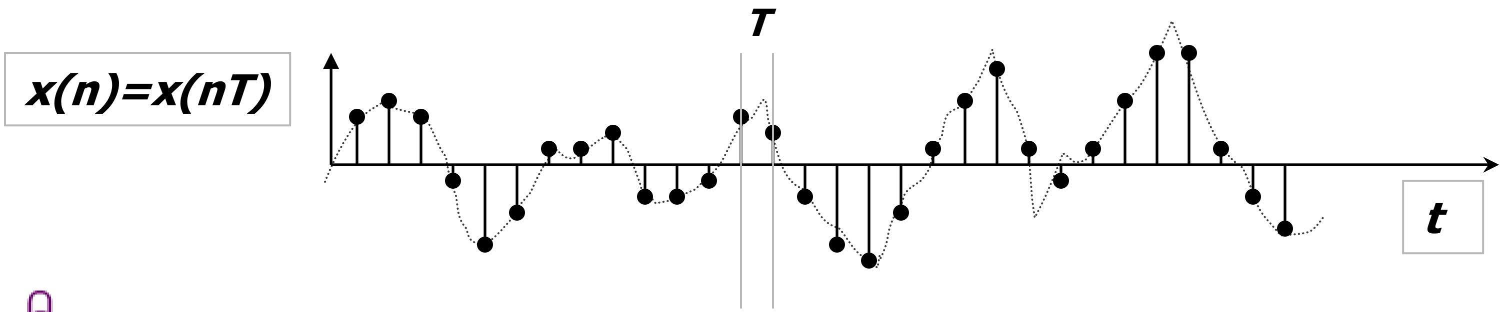
A little bit of signal processing

……A little bit of signal processing

- Let x(t) be a continuous signal (e.g. captured by a microphone):



- Let x(nT) be the discrete signal sampled at time *t=nT*



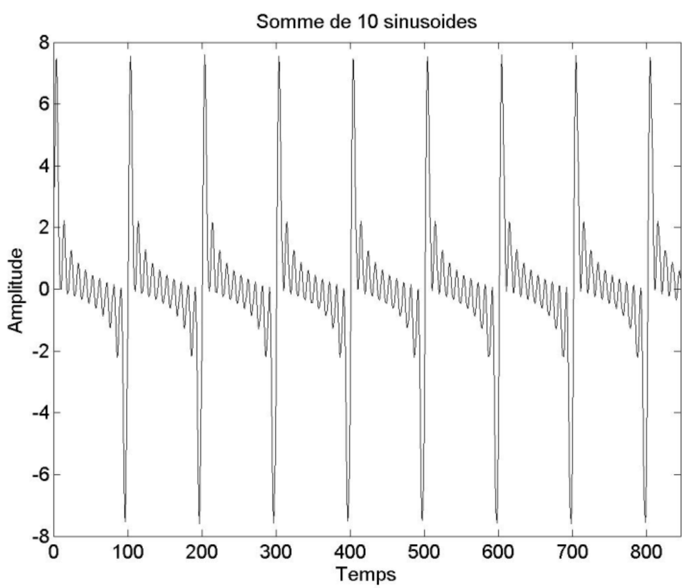
Time-Frequency representation

- **Fourier Transform**

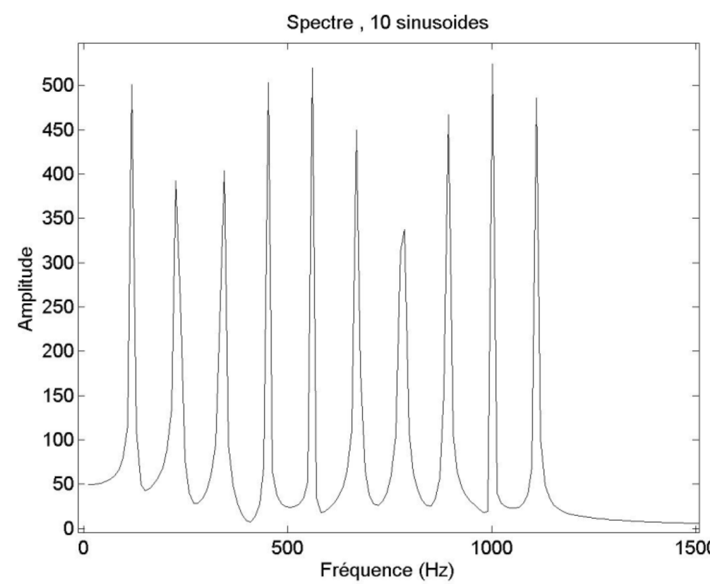
$$X_k = \sum_{n=0}^{N-1} x_n e^{-2j\pi nk/N}$$

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{2j\pi nk/N}$$

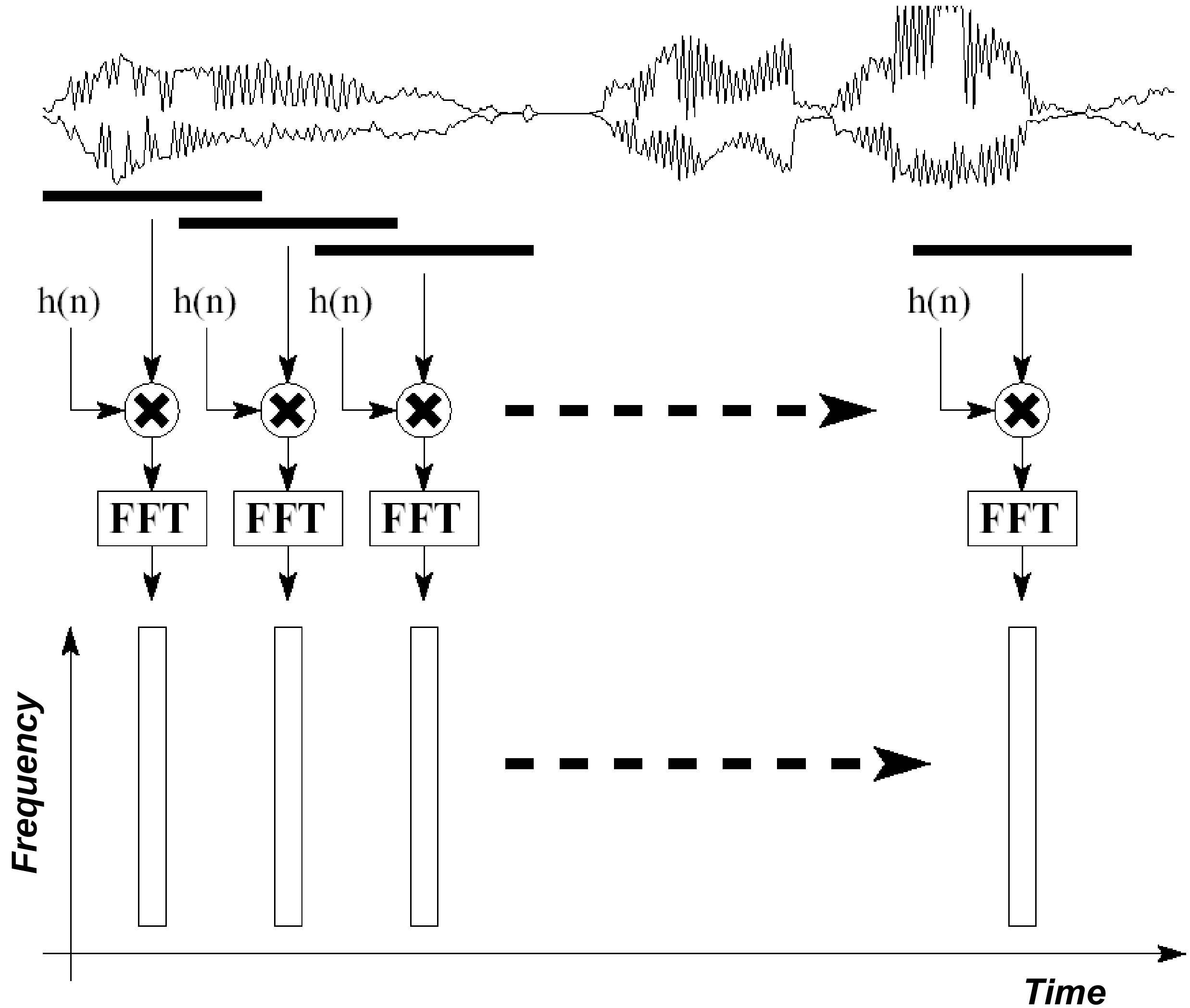
x_n



$|X_k|$



Spectral analysis of an audio signal (1)
(drawing from J. Laroche)

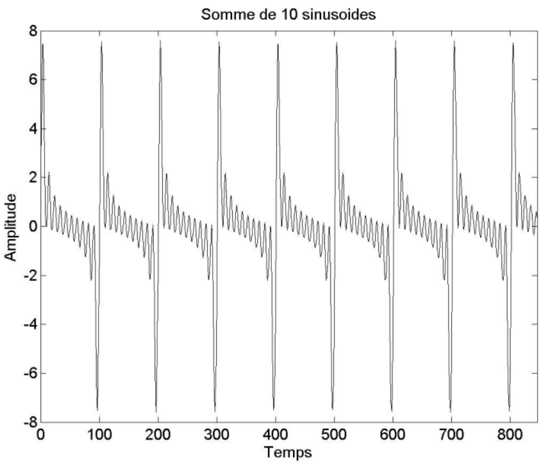


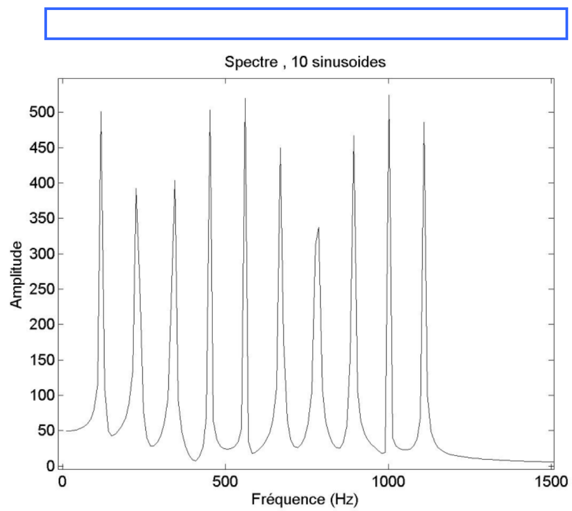
Spectral analysis of an audio signal (2)

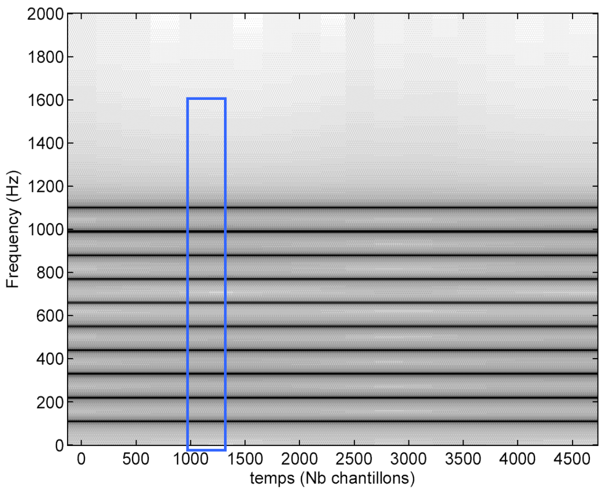
Spectrogram

$|X_k|$

x_n





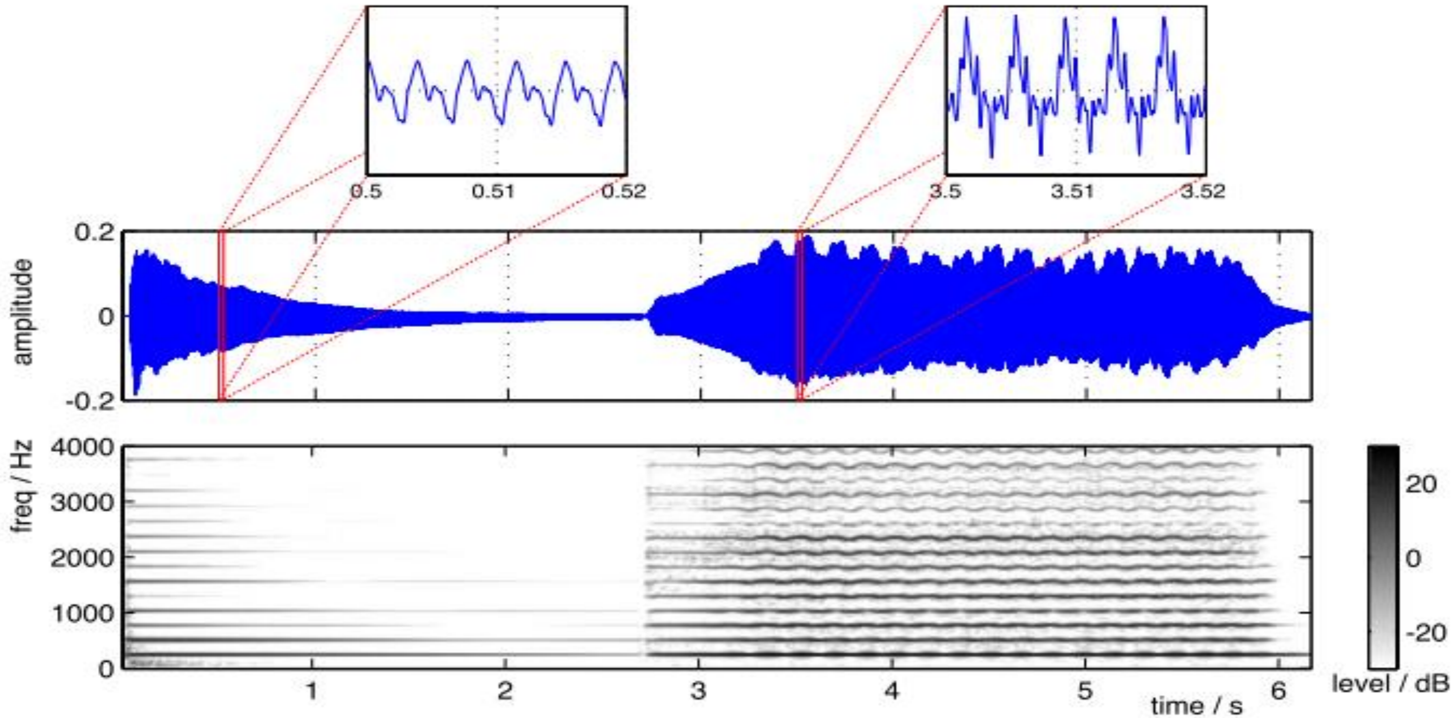


Audio signal representations

- **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

A bit more details on the Fourier analysis

- **Fourier transform and inverse Fourier transform**

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-2j\pi f t} dt$$

$$x(t) = \int_{-\infty}^{+\infty} X(f) e^{2j\pi f t} df$$

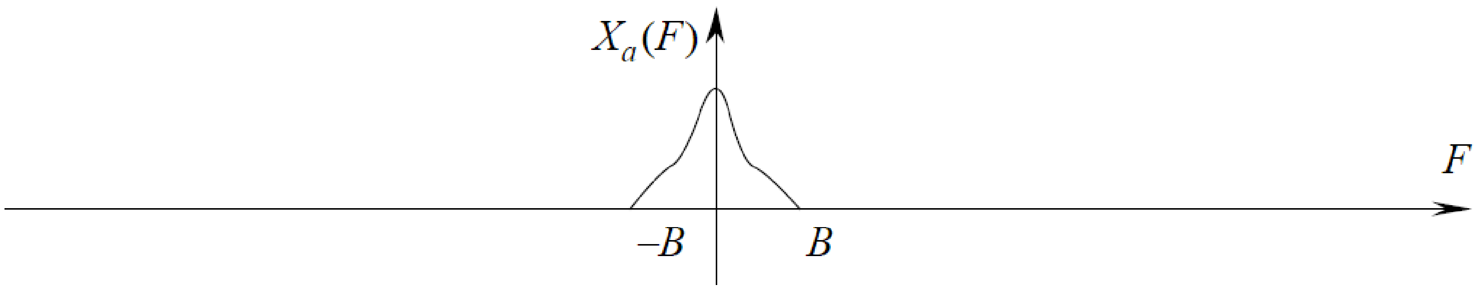
- **Some properties**

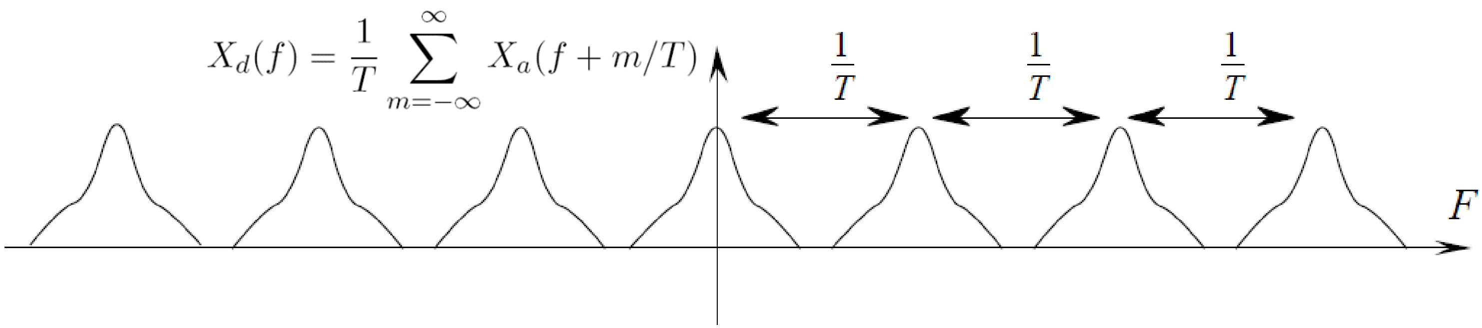
Properties	$x(t)$	$X(f)$
Convolution	$x(t) \star y(t)$	$X(f)Y(f)$
Similitude	$x(at)$	$\frac{1}{\|a\|} X(f/\|a\|)$
Translation	$x(t - t_0)$	$X(f) \exp(-2j\pi t_0 f)$
Modulation	$x(t) \exp(2j\pi f_0 t)$	$X(f - f_0)$
	real	$X(f) = X^*(-f)$

Effect of sampling: Poisson formula

- **Interpretation: Sampling ➡ Spectrum periodisation**

$$X_d(f) = \frac{1}{T}\sum_{m=-\infty}^{\infty} X_a(f+m/T) = \sum_{n=-\infty}^{\infty} x(n)e^{-2j\pi f n T}$$



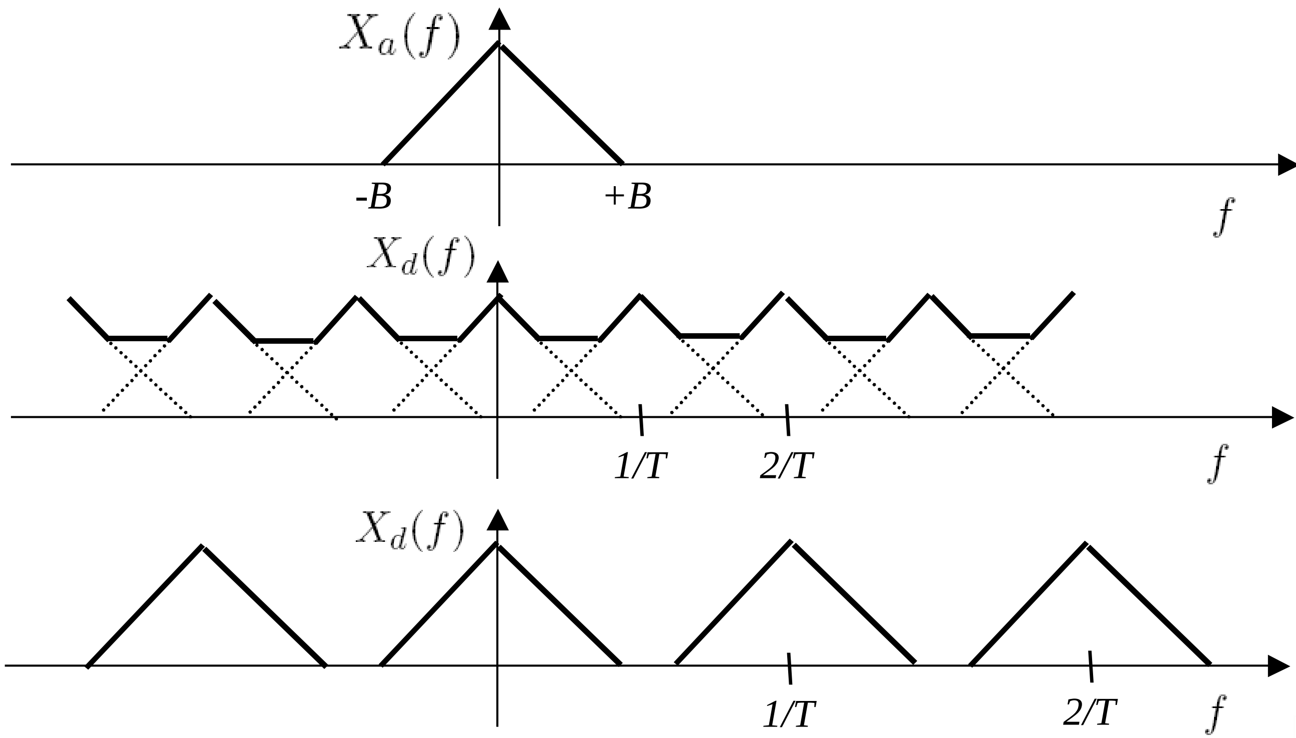


Towards reconstruction

- **2 situations:**

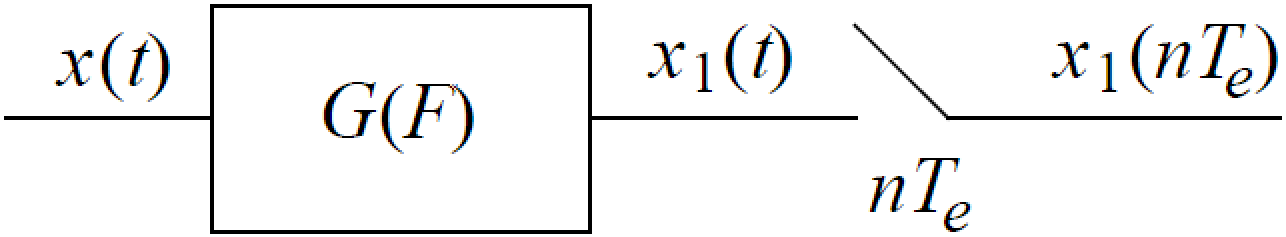
$$\frac{1}{T} < 2B$$

$$\frac{1}{T} > 2B$$



Sampling of an analog signal

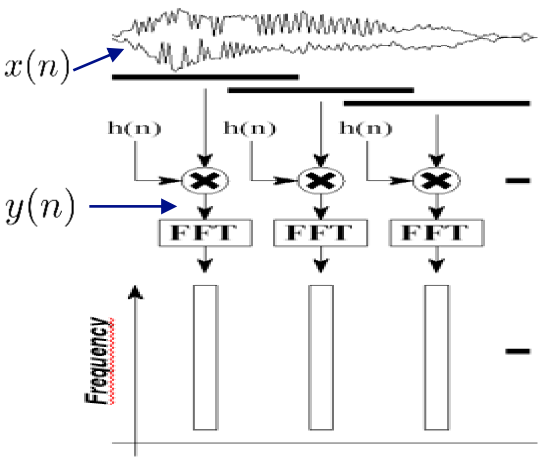
- **Important to filter the analog signal before sampling**



A bit more details on the Fourier analysis

- **Importance of the analysis window**

$$y(t) = h(t) \times x(t)$$



- **We recall that :**

Properties	$x(t)$	$X(f)$
Convolution	$x(t) \star y(t)$	$X(f)Y(f)$
Similitude	$x(at)$	$\frac{1}{\lvert a\rvert}X(f/\lvert a\rvert)$
Translation	$x(t-t_0)$	$X(f)\exp(-2j\pi t_0 f)$
Modulation	$x(t)\exp(2j\pi f_0 t)$	$X(f-f_0)$
	real	$X(f) = X^*(-f)$

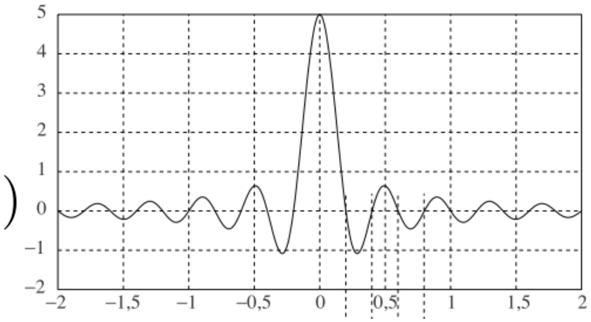
- **Then we have**

$$Y(f) = H(f) * X(f)$$

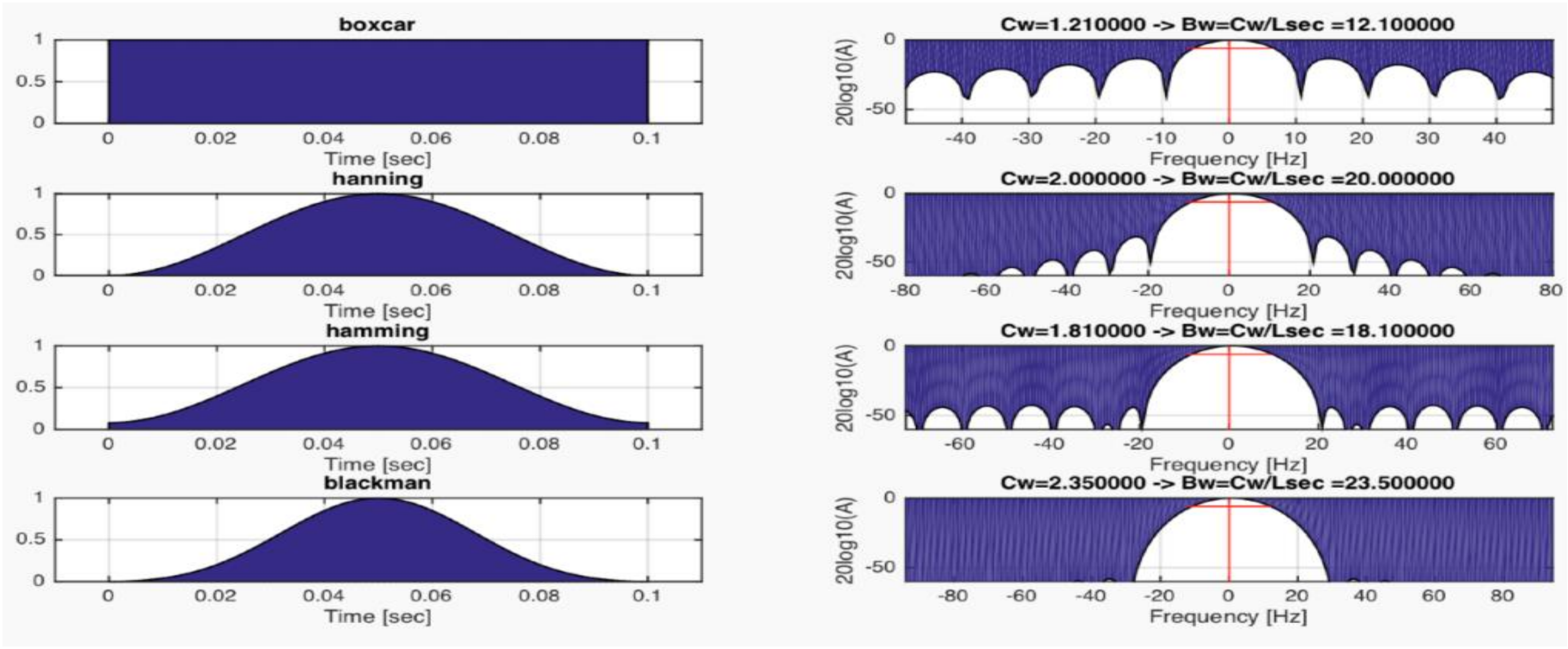
A bit more details on the Fourier analysis

- **Some examples of analysis windows**
  - Rectangular window: $h(t) = rect_{T_w}(t)$

$$H(f) = \frac{sin(\pi f T_w)}{\pi f} = T_w sinc(f T_w)$$



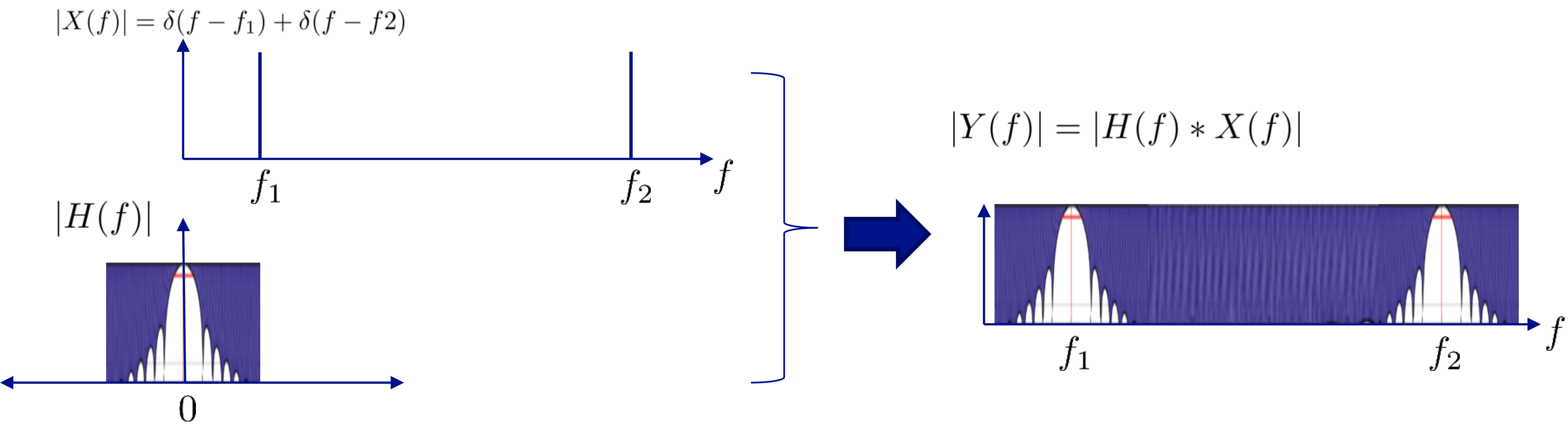
    - Width of the main lobe: $\frac{2}{T_w}$



A bit more details on the Fourier analysis

An example:

$$
\begin{aligned}
y(t) &= h(t) \times x(t) \\
&= h(t) \times (sin(2\pi f_1 t) + sin(2\pi f_2 t))
\end{aligned}
$$



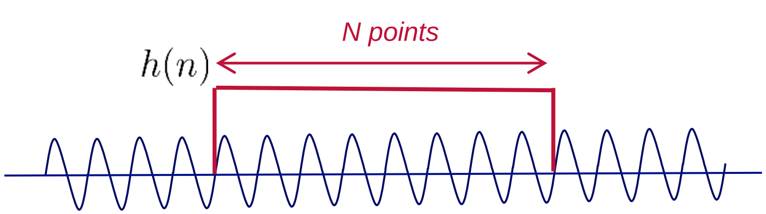
A bit more details on the Fourier analysis

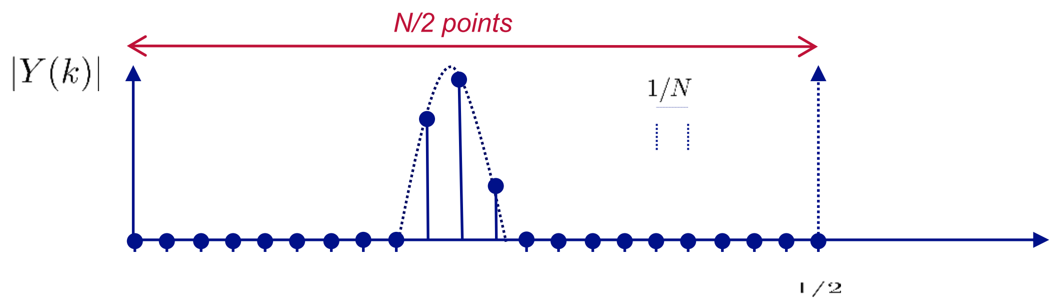
The notion of precision and resolution in discrete time:

$$
\begin{aligned}
y(t) &= h(t) \times x(t) \\
&= h(t) \times (sin(2\pi f_1 t))
\end{aligned}
$$

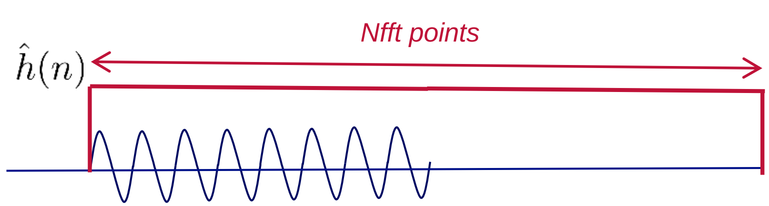
$$
\begin{aligned}
y(n) &= h(n) \times x(n) \\
&= h(n) \times (sin(2\pi f_1 .nT))
\end{aligned}
$$

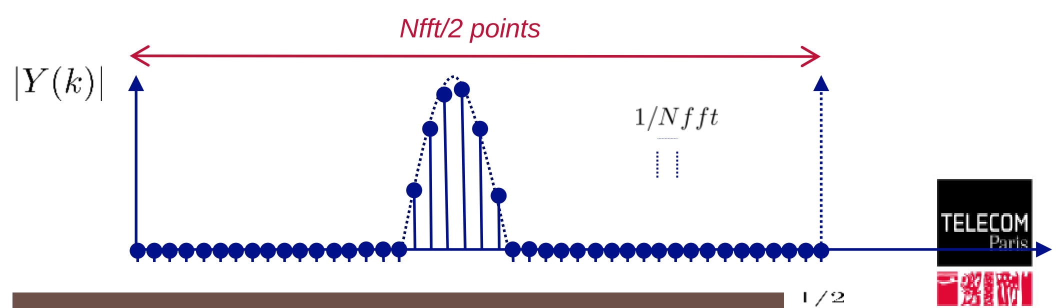
$$\Rightarrow \quad Y(k) = \sum_{n=0}^{N-1} y(n) e^{-2j\pi nk/N}$$





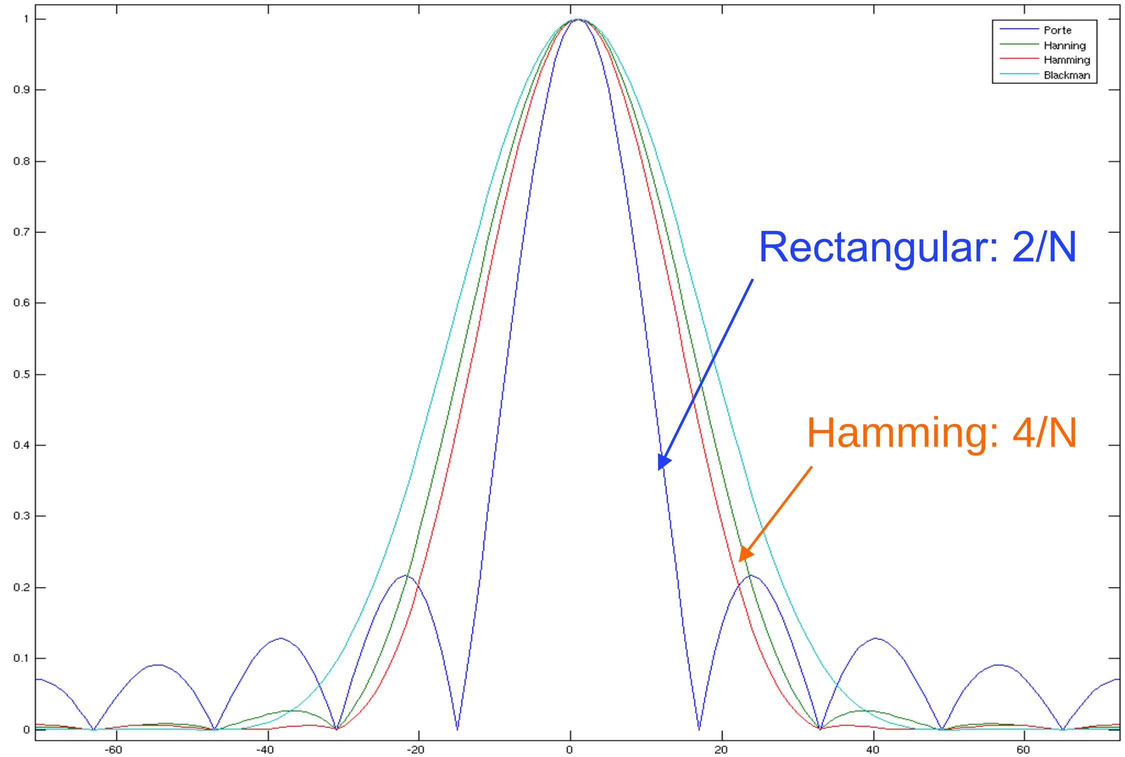
Zero padding





A bit more details on the Fourier analysis

- **Some examples of analysis windows (size N)**
  - Width of the main lobe:



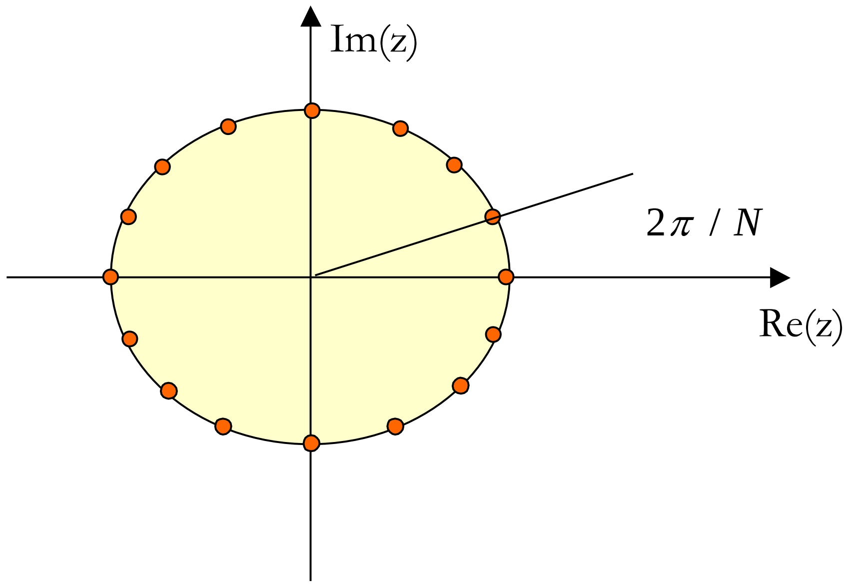
Z transform/ Discrete Fourier Trnasform

- **Z-transform of a signal x(n) is given by:**

$$X(z) = \sum_{n=-\infty}^{+\infty} x(n) z^{-n} \quad \textbf{with} \quad z \in \mathcal{C} = \{z \in \mathbb{C} : R_1 < |z| < R_2\}$$

- **Links Z-transform /DFT**

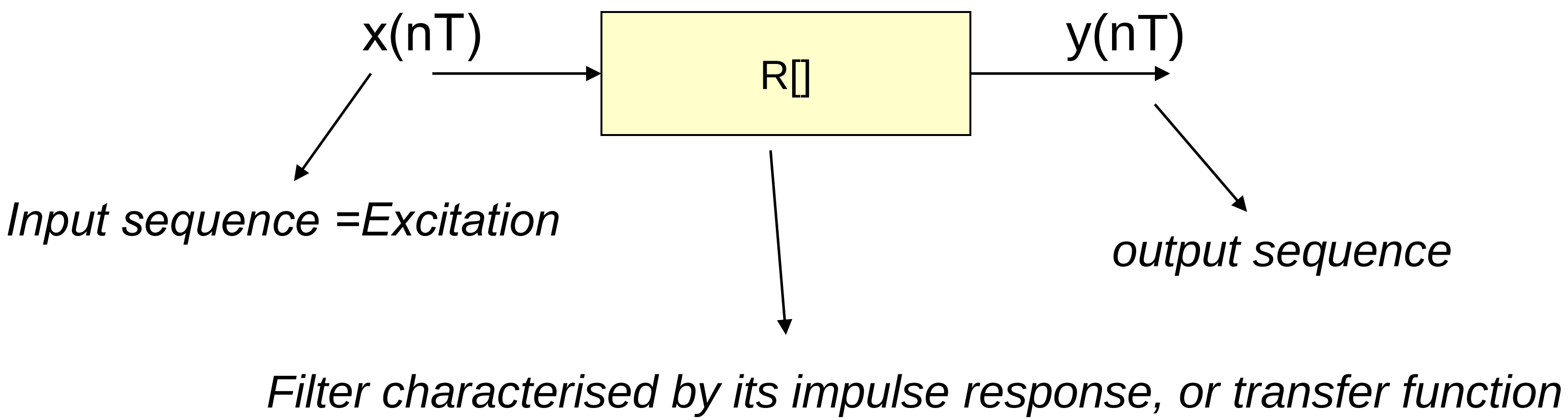
$$X(k) = X(z)\big|_{z=e^{2j\pi k/N}}$$



- This corresponds to a sampling of the Z-transform with N points regularly spaced on the unit circle.

Digital filtering

- **Linear shift invariant system**



$$Y(nT) = R[x(nT)]$$ where T is the sampling period.

By choosing T=1, we have: $Y(n) = R[x(n)]$

Digital filtering

- **Linear constant-coefficient Difference Equations (a sub class of shift invariant systems)**

$$y(n) = \sum_i a_i x(n-i) - \sum_j b_j y(n-j)$$

- **Causal recursive filters**

$$y(n) = \sum_{i=0}^{N-1} a_i x(n-i) - \sum_{j=1}^{M-1} b_j y(n-j)$$

- **Causal non-recursive filters**

$$y(n) = \sum_{k=0}^{N-1} a_i x(n-i)$$

Digital filtering: convolution

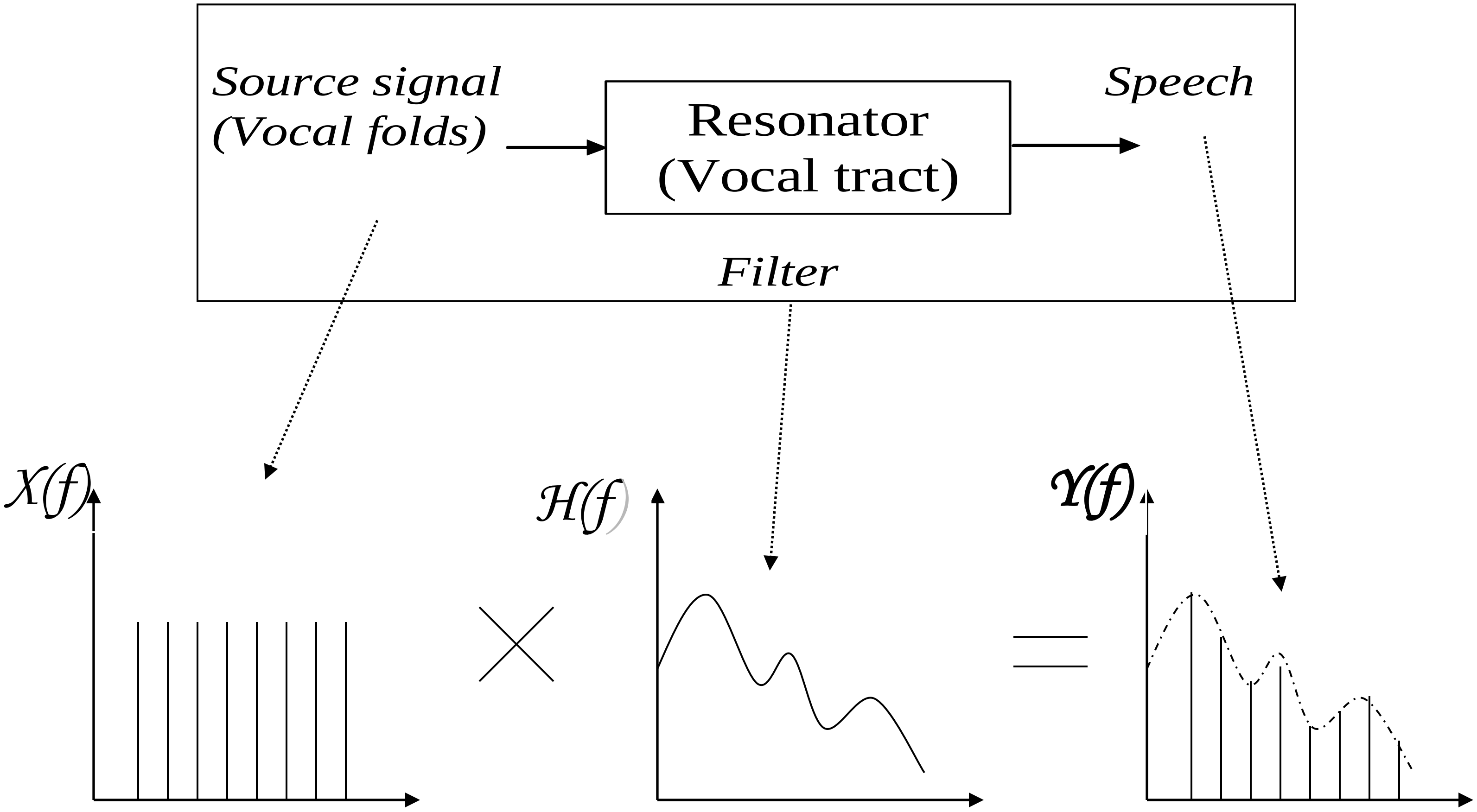
- **Convolution allows to represent the intput-output transformation realised by a linear shift-invariant filter**

$$
\begin{aligned}
y(n) &= \sum_{-\infty}^{\infty} x(k)h(n-k) = \sum_{-\infty}^{\infty} x(n-k)h(k) \\
y(n) &= x(n) * h(n)
\end{aligned}
$$

- The impulse response is also the response to $\delta(n)$ the unit sample at *n*=*k*:

$$
h(n) = \sum_{-\infty}^{\infty} h(k)\delta(n-k)
$$

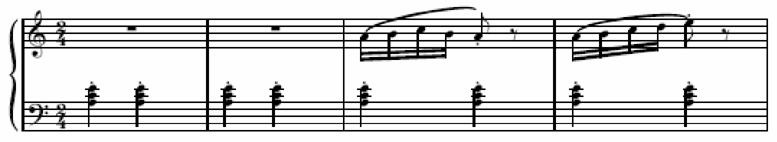
A widely used model: the source filter model



Some dimensions of the musical signal …

Pitch, Harmony..

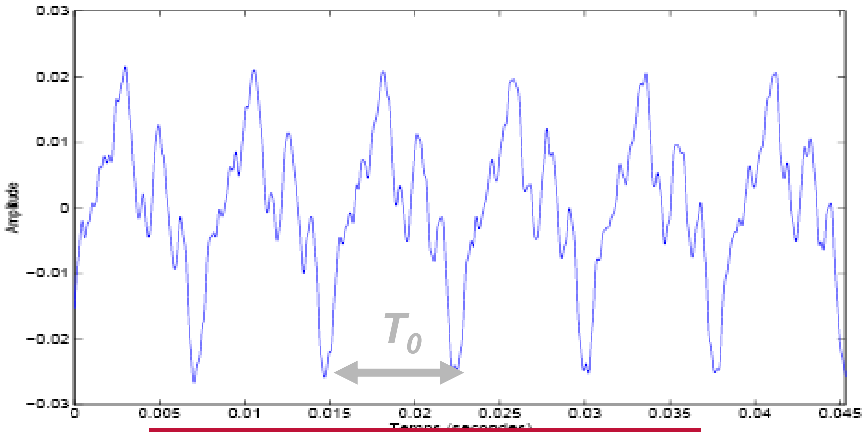
Tempo, rhythme,…



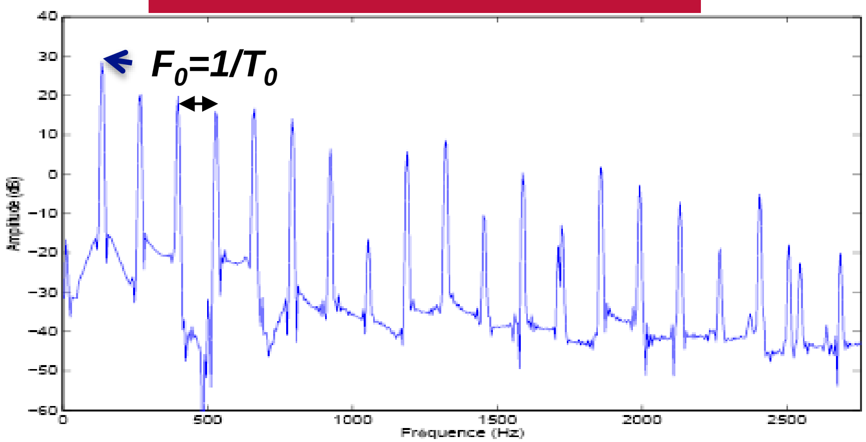
Timbre, instruments,…

Polyphony, melody, ….

A quasi-periodic sound



A piano sound (C3)



Spectrum of a piano sound

How can we estimate the height (pitch) of a note

or

How to estimate the **fundamental periode** (T_0)
or **frequency** (F_0) ?

Signal Model

$$x(n) = \sum_{k=1}^{H} 2A_k \cos(2\pi k f_0 n + \phi_k) + w(n)$$

$f_0 = \dfrac{1}{T_0}$ normalised fundamental frequency

- H is the number of harmonics
- Amplitudes $\{A_k\}$ are real numbers > 0
- Phases $\{\phi_k\}$ are independant r.v. uniform on $[0, 2\pi[$
- w is a centered white noise of variance σ^2, independent of phases $\{\phi_k\}$
- x(n) is a centered second order process with autocovariance

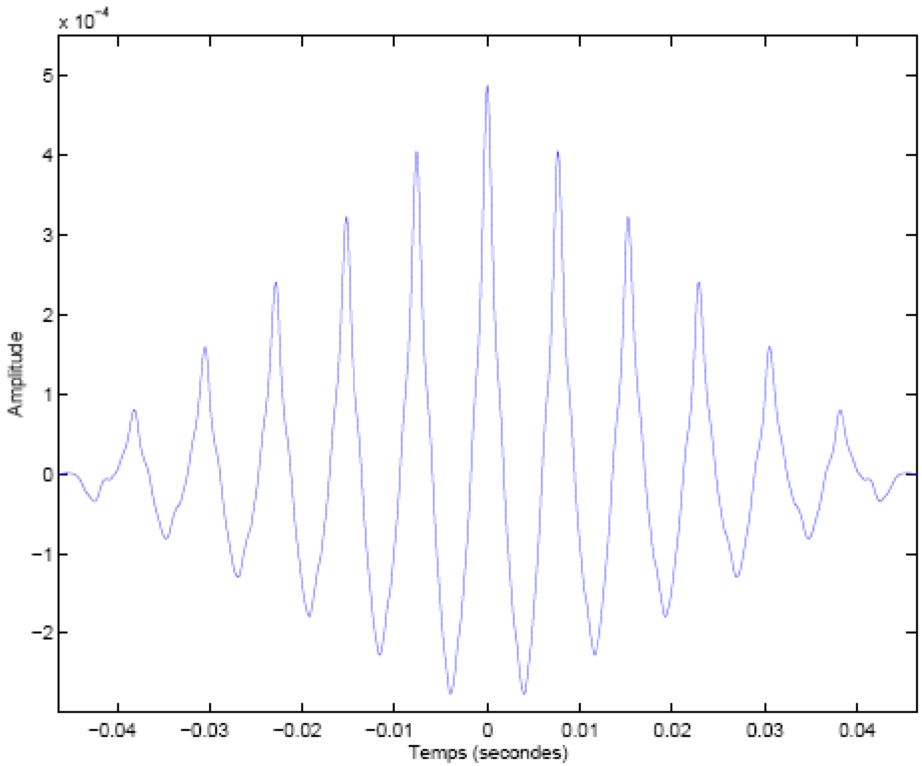
$$r_x(m) = \sum_{k=1}^{H} [2A_k^2 \cos(2\pi k f_0 m)] + \sigma^2 \delta[m]$$

Time domain methods

- **Autocovariance estimation (biased)**

$$\frac{1}{N}\sum_{n=0}^{N-1-m} x[n]\,x[n+m] \text{ si } m \geq 0$$

$$\mathbf{E}(\hat{r}_x[m]) = \frac{N-|m|}{N}\,r_x[m] \qquad\qquad |\hat{r}_x[m]| \leq \hat{r}_x[0]$$

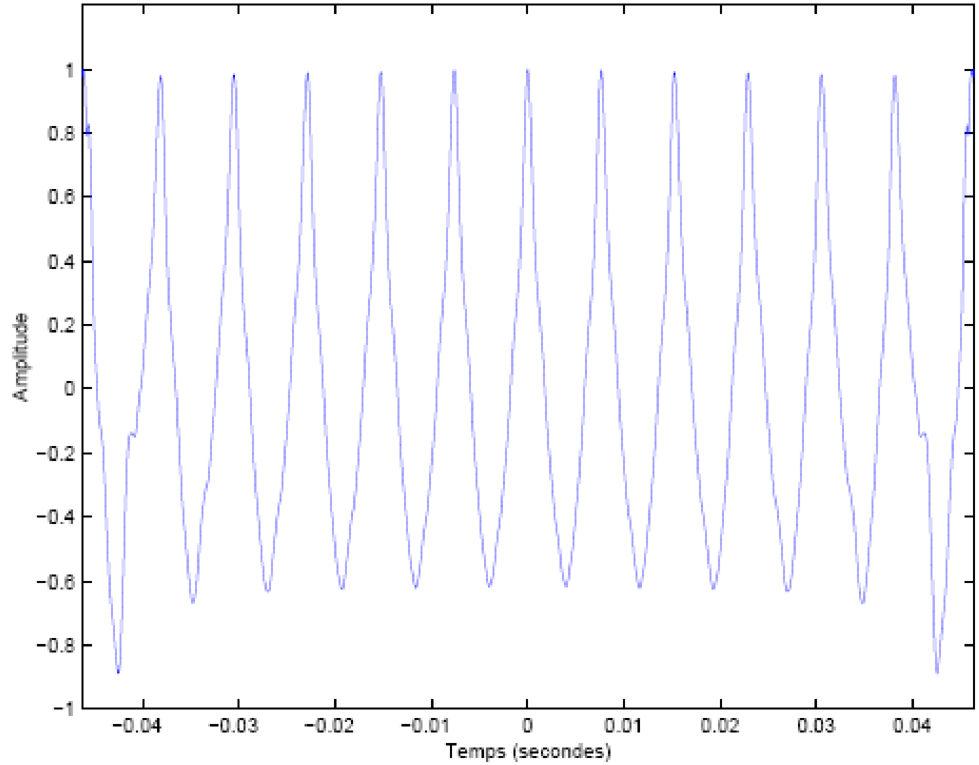


Time domain methods

- **Autocorrelation**

$$\bar{r}_x[m] = \frac{\sum_{n=0}^{N-1-m} x[n]\, x[n+m]}{\sqrt{\sum_{n=0}^{N-1-m} x[n]^2}\sqrt{\sum_{n=0}^{N-1-m} x[n+m]^2}} \text{ si } m \geq 0$$

$$|\bar{r}_x[m]| \leq \bar{r}_x[0] = 1 \qquad |\bar{r}_x[m]| = 1 \text{ ssi les vecteurs sont colinaires}$$



Maximum likelihood approach

- Signal model: $x(n) = a(n) + w(n)$
  - *a* is a deterministic signal of period T_0
  - *w* is white Gaussian noise of variance σ^2

- Observation likelihood

$$p(x|T_0, a, \sigma^2)) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n)-a(n))^2}$$

- Log-likelihood

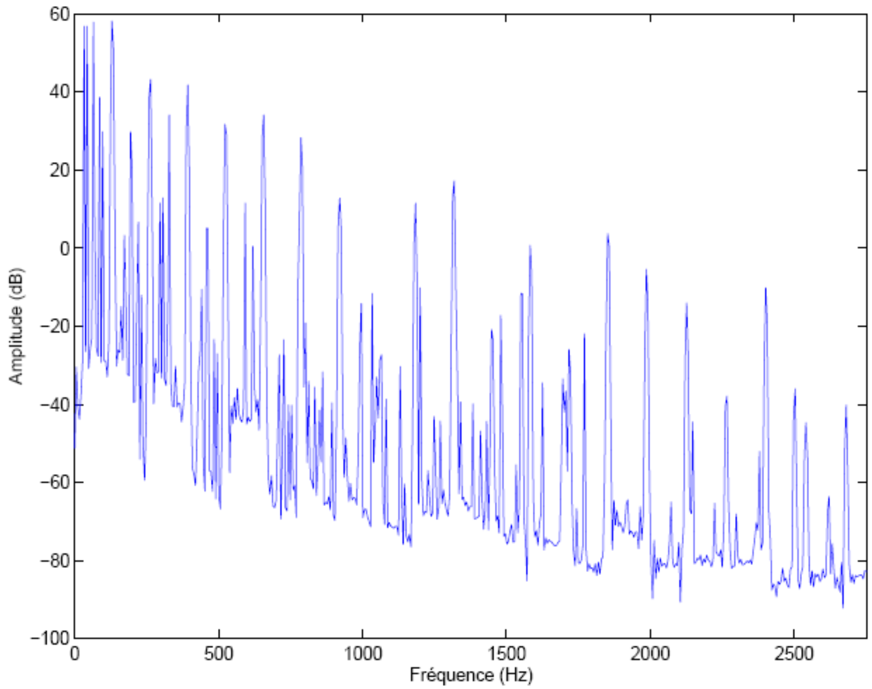
$$L(T_0, a, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x(n) - a(n))^2$$

- Method: maximise successively L with respect to a, then σ^2 and then T_0.

Maximum likelihood approach

- It can be shown that maximisation of L with respect to $F_0 = \frac{m}{N}$ is equivalent to maximise the spectral sum $S(k)$

- The spectral sum is $S(k) = \sum_{h=1}^{H} |X(h.k)|$



The spectral sum: a bit more explanation

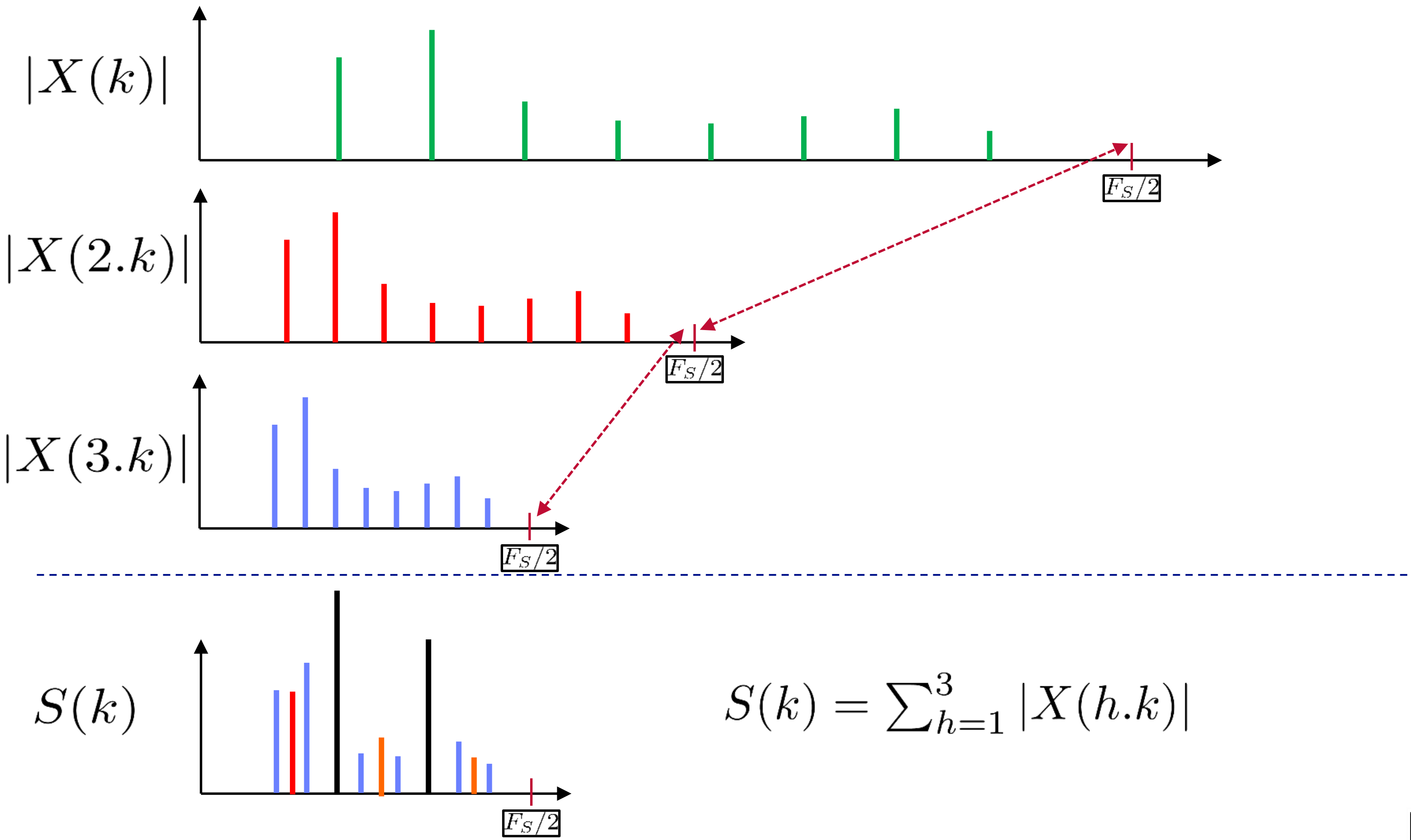
$$S(k) = \sum_{h=1}^{H} |X(h.k)|$$

- For a given k_i (e.g. frequency), $S(k_i)$ corresponds to the addition of the H spectral values : $|X(k_i)| + |X(2.k_i)| ... + |X(H.k_i)|$

- It can be seen as the scalar product of the original spectrum with a perfect comb of H teeth with a first tooth localised at k_i

- If k_i corresponds to a fundamental frequency, $S(k_i)$ will be the sum of the first H harmonics and leads to a maximum

A practical mean to compute the spectral sum (H=3)

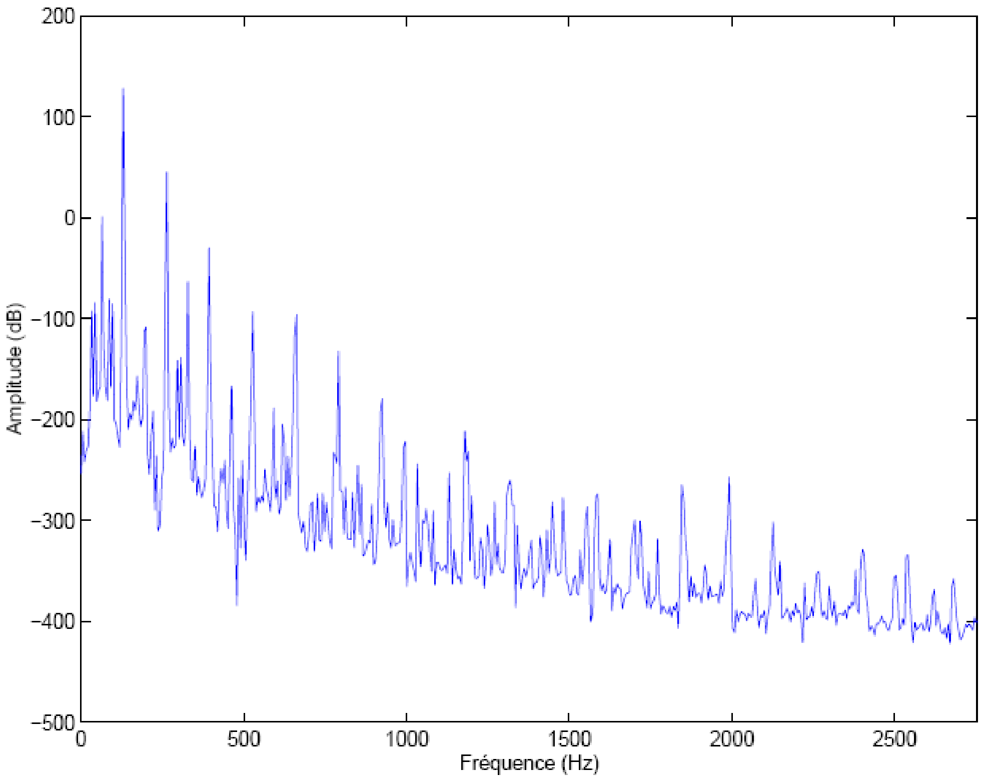


$$S(k) = \sum_{h=1}^{3} |X(h.k)|$$

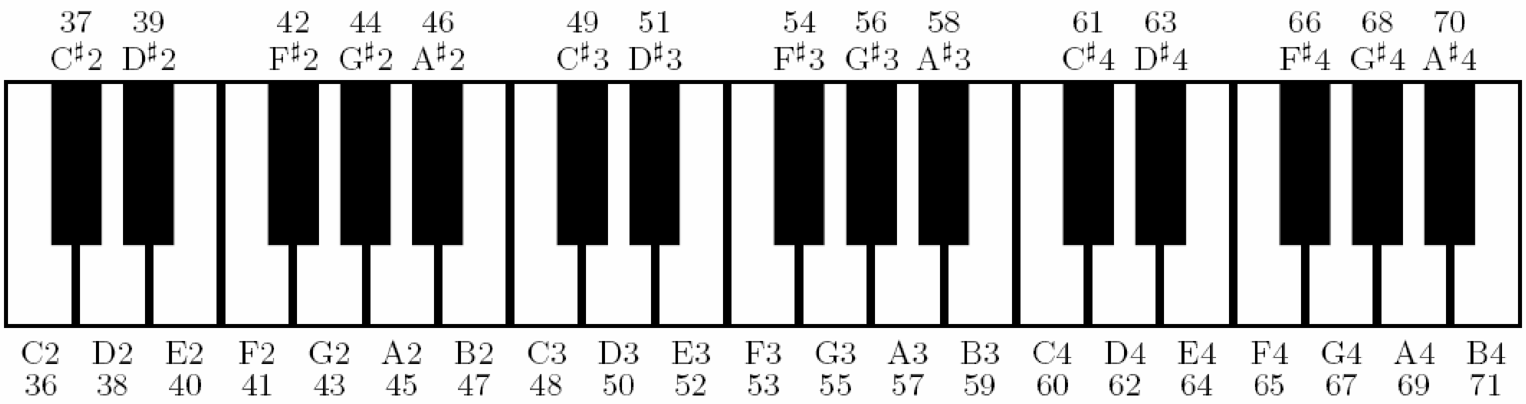
Spectral product

- By analogy to spectral sum (often more robust)

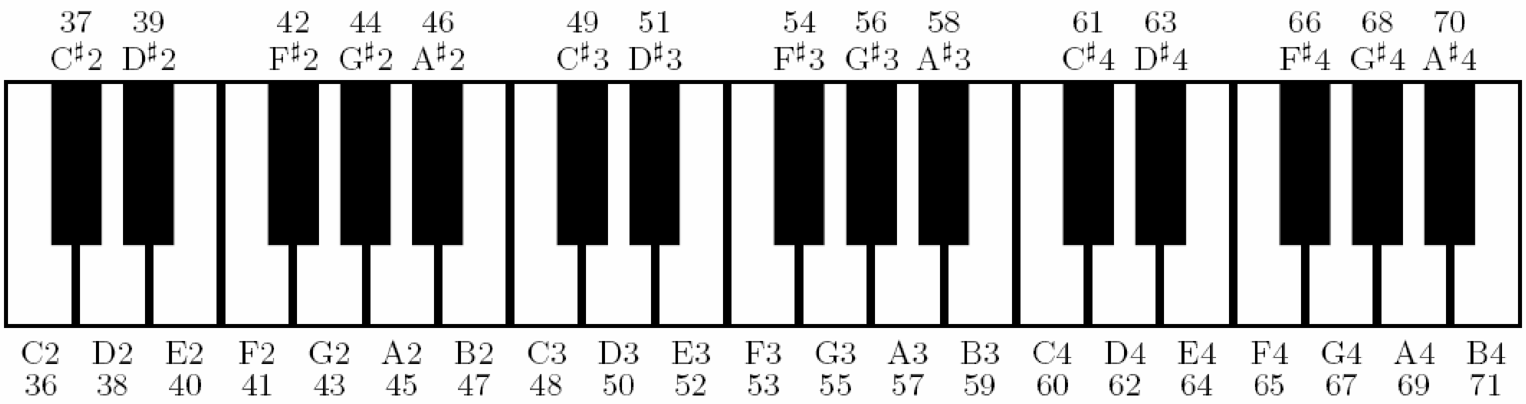
$$P(k) = \prod_{h=1}^{H} |X(h.k)|$$



Pitch Features



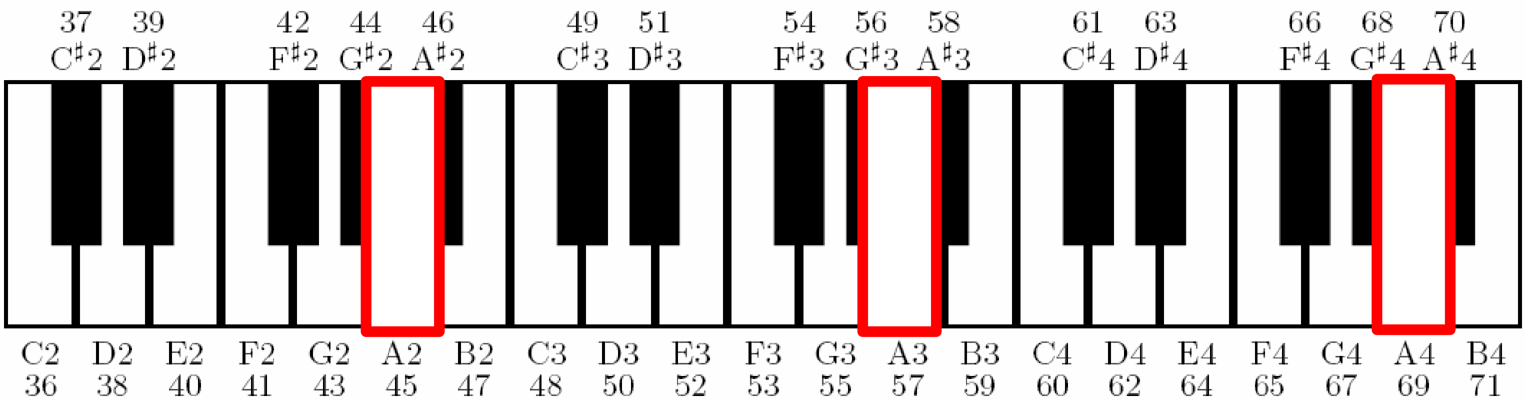
Pitch Features



Model assumption: Equal-tempered scale

- MIDI pitches: $p \in [1:128]$
- Piano notes: $p = 21\ (A0)\quad p = 128\ (C8)$
- Concert pitch: $p = 69\ (A4) = 440\ \ Hz$
- Center frequency: $f_{MIDI}(p) = 2^{\frac{p-69}{12}} \times 440\ \ Hz$

Pitch Features

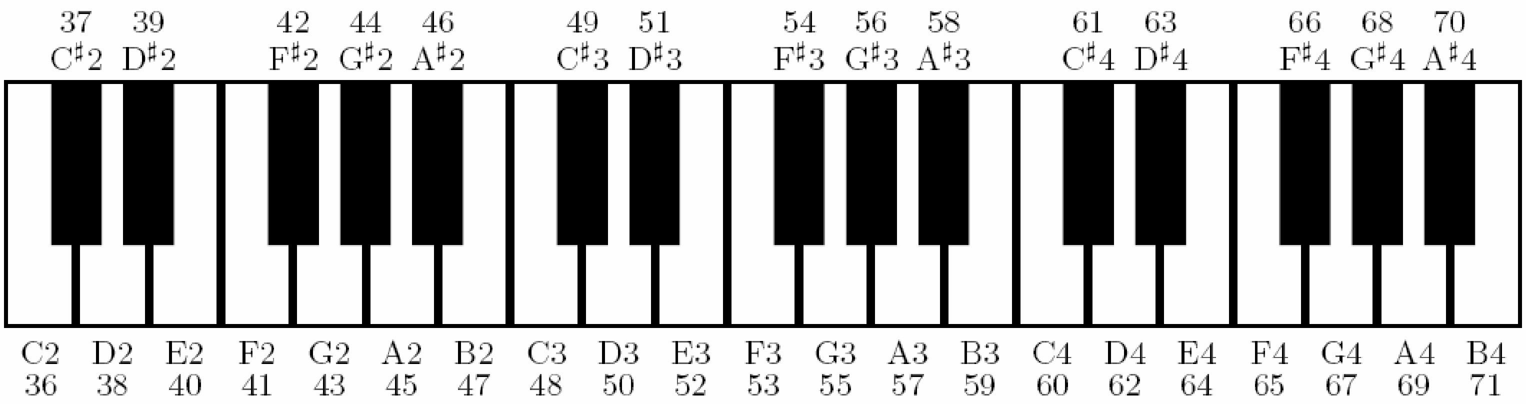


A2	A3	A4
110 Hz	220 Hz	440 Hz

Logarithmic frequency distribution

Octave: doubling of frequency

Towards a more specific representation

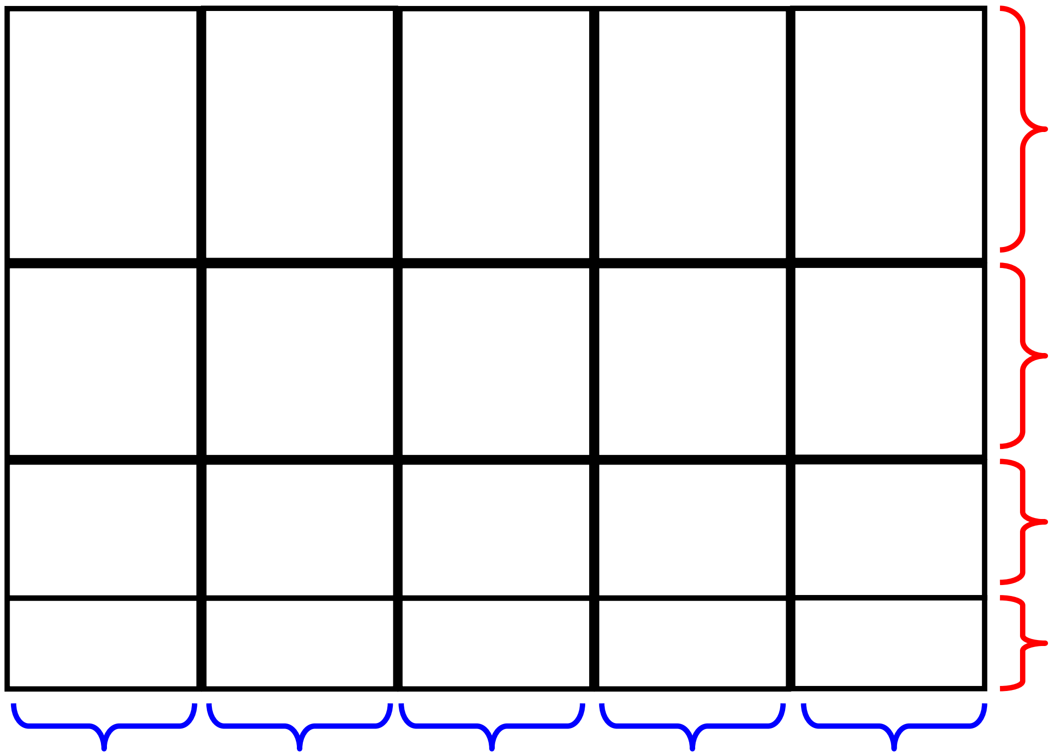


Idea: Binning of Fourier coefficients

- Divide up the frequency axis into logarithmically spaced "pitch regions"
- …and combine spectral coefficients (e.g. $|X_k|$) of each region to form a single pitch coefficient.

Towards a more specific representation

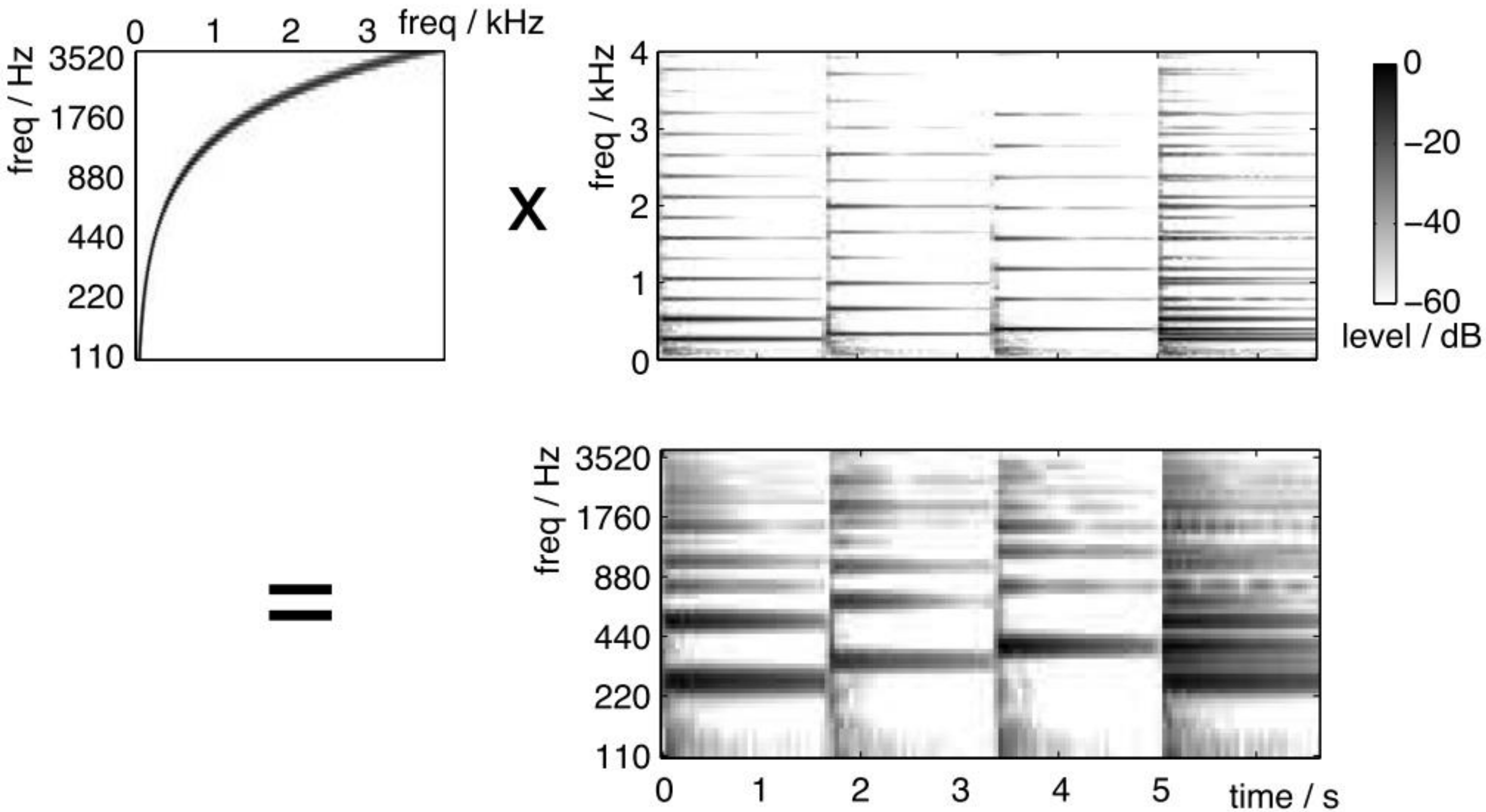
Towards a Constant-Q time-frequency transform: $R = \frac{f_k}{\Delta f_k} = cste$



Windowing in the frequency domain

Windowing in the time domain

Towards a more specific representation

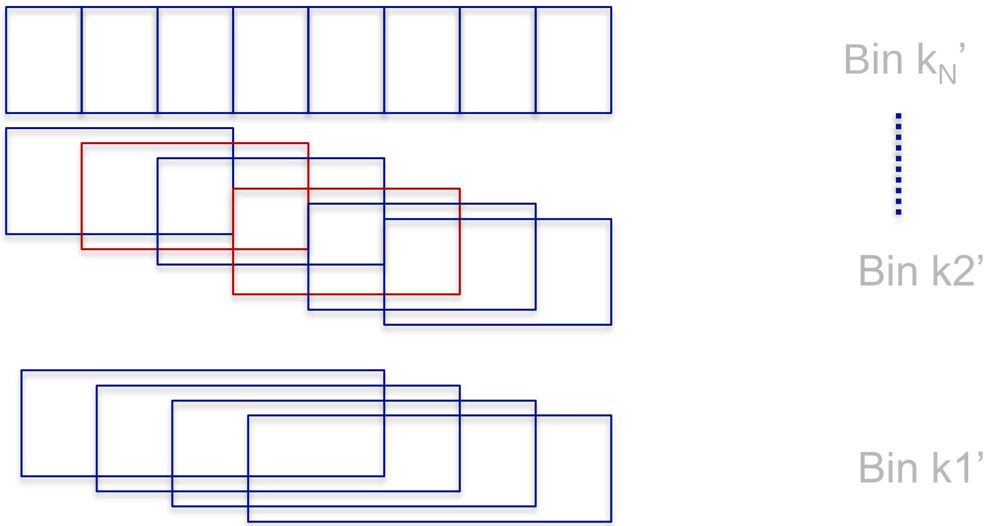


From M. Mueller & al. « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Towards a more specific representation

- **In practice:**
  - Solution is only partially satisfying

- **More appropriate solution: Use temporal windows of different size for each frequency bin k'**



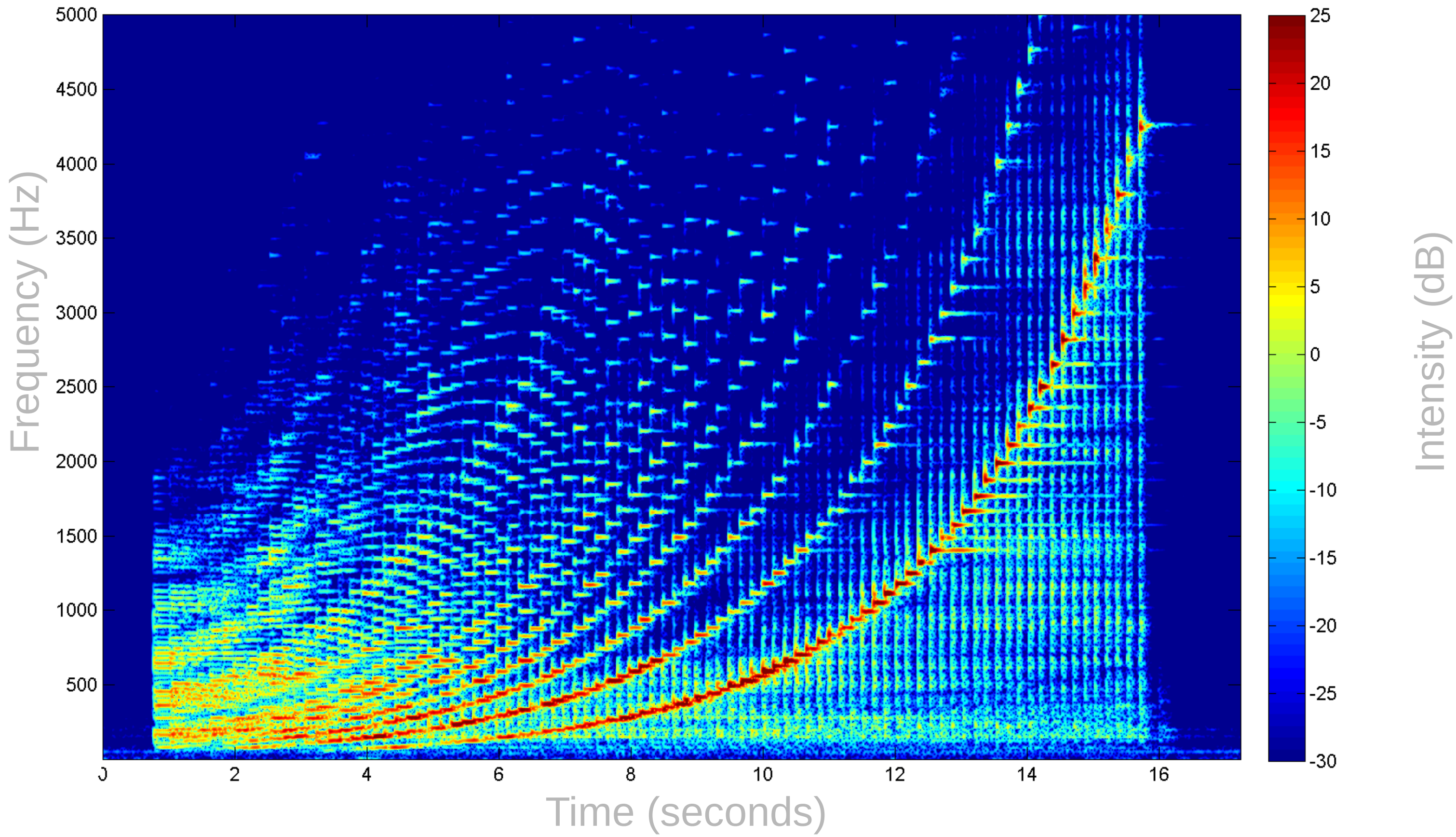
J. Brown and M. Puckette, An efficient algorithm for the calculation of a constant Q transform, JASA, 92(5):2698–2701, 1992.
J. Prado, Une inversion simple de la transformée à Q constant, technical report, 2011, (in French)
http://www.tsi.telecom-paristech.fr/aao/en/2011/06/06/inversible-cqt/

Towards a more specific representation

Example: Chromatic scale

(Credit M. Mueller)

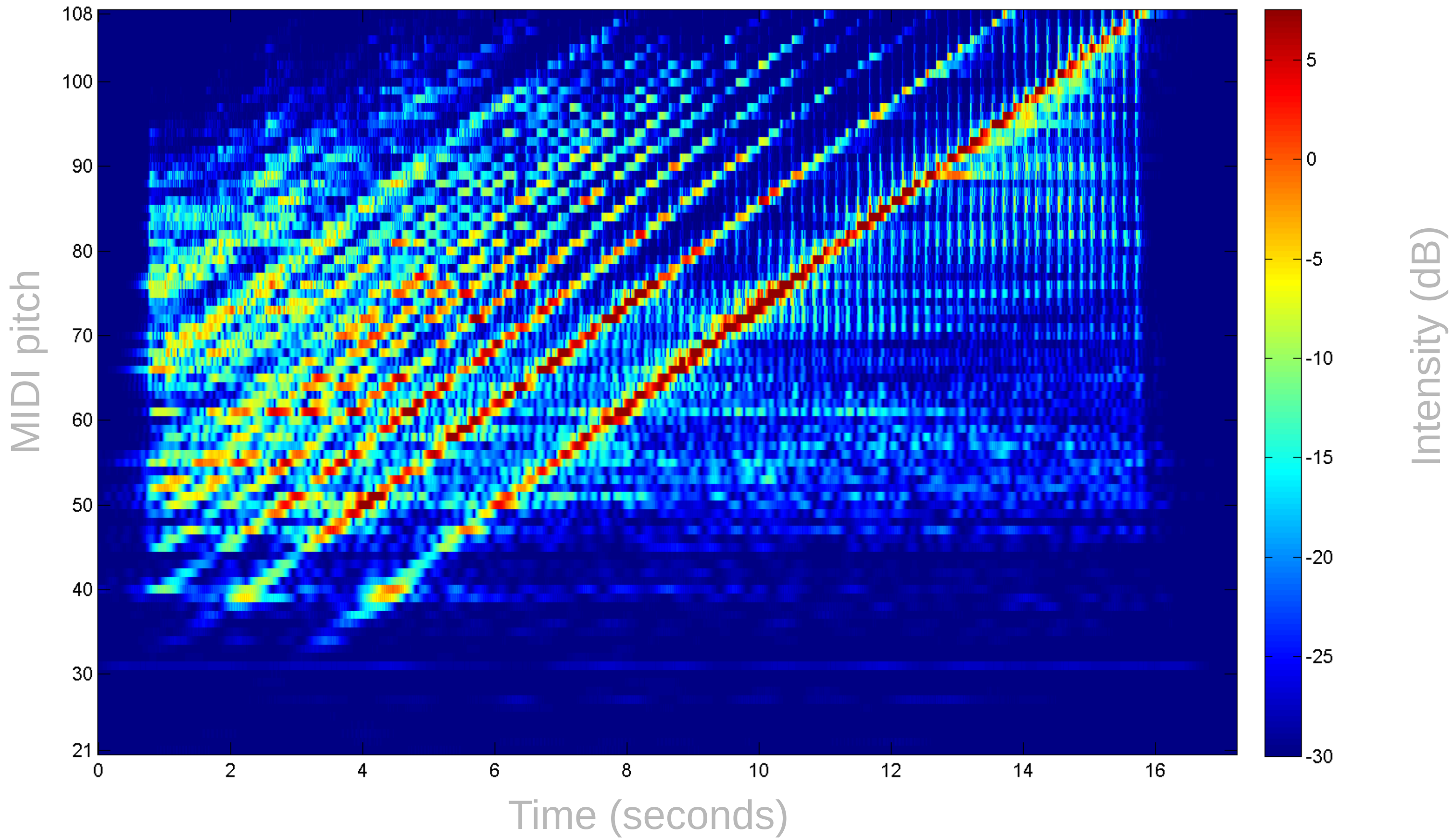
Spectrogram



Towards a more specific representation

Example: Chromatic scale

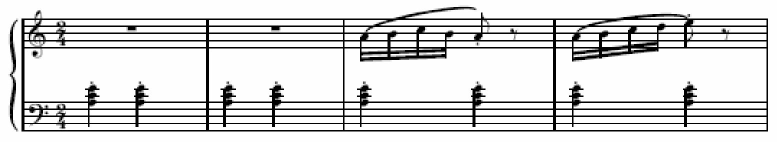
Log-frequency spectrogram



Some dimensions of the musical signal …

Pitch, **Harmony**..

Tempo, rhythme,…



Timbre, instruments,…

Polyphony, melody, ….

Detecting multiple notes
(e.g. multipitch estimation)

- **Why it is challenging ?**
- **How would you do it ?**

Detecting multiple notes
(e.g. multipitch estimation)

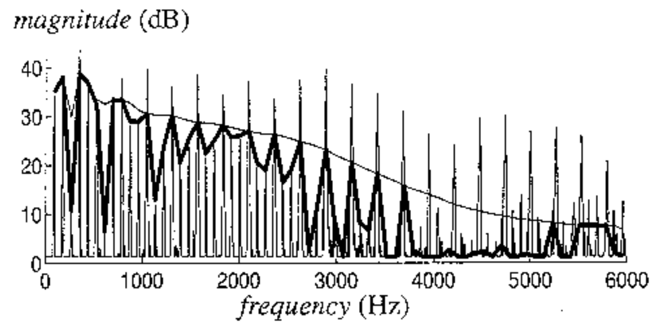
- **Why it is challenging ?**
- **How would you do it ?**
- **Different families of methods**
  - Time domain approaches
  - Frequency domain approaches
  - Statistical modelling, Decomposition models
  - Machine learning based (Bayesian models, classification models, deep neurla networks).

Exploiting basic iterative source separation principles

Iterative multi-pitch extraction …

- First, detect the most prominent note …
- Subtract this note from the polyphony
- Then, detect the next most prominent note
- Soustract this note from the polyphony
- Etc… until all notes are found

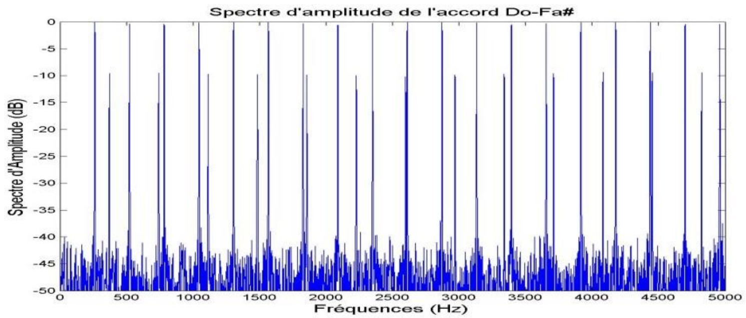
Spectral smoothness



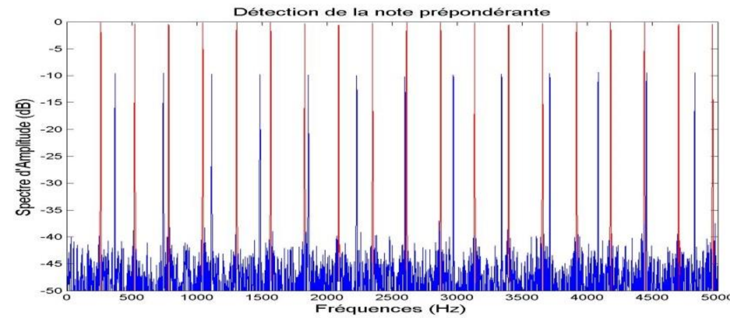
A. Klapuri, Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness, IEEE Trans. On Speech and Sig. Proc., 11(6), 2003
A. Klapuri "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model", IEEE Trans. On ASLP, Feb. 2008

Iterative multipitch estimation

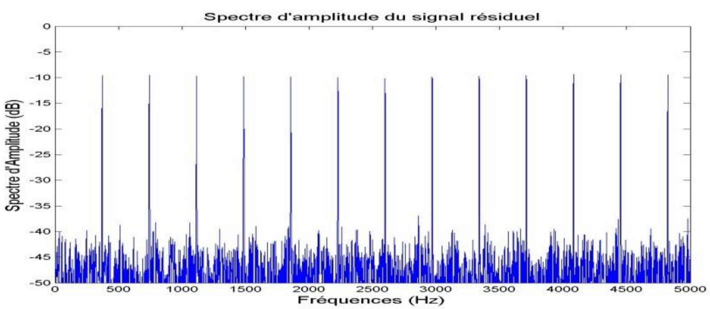
Chord of two synthetic notes C – F#



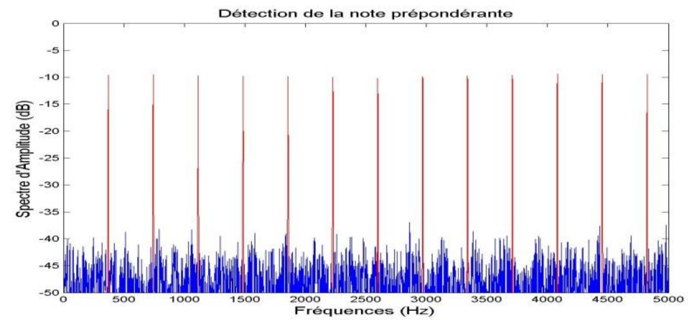
Detect the most prominent note (in red)



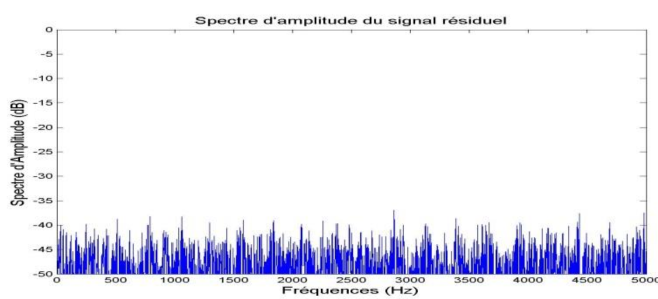
Subtract the detected note



Detect the next most prominent note



There is no more notes….chord C – F# is recognized



Harmony: the chroma features

- Pitches are perceived as related (or harmonically similar) if they differ by an octave (the notes have the same name)

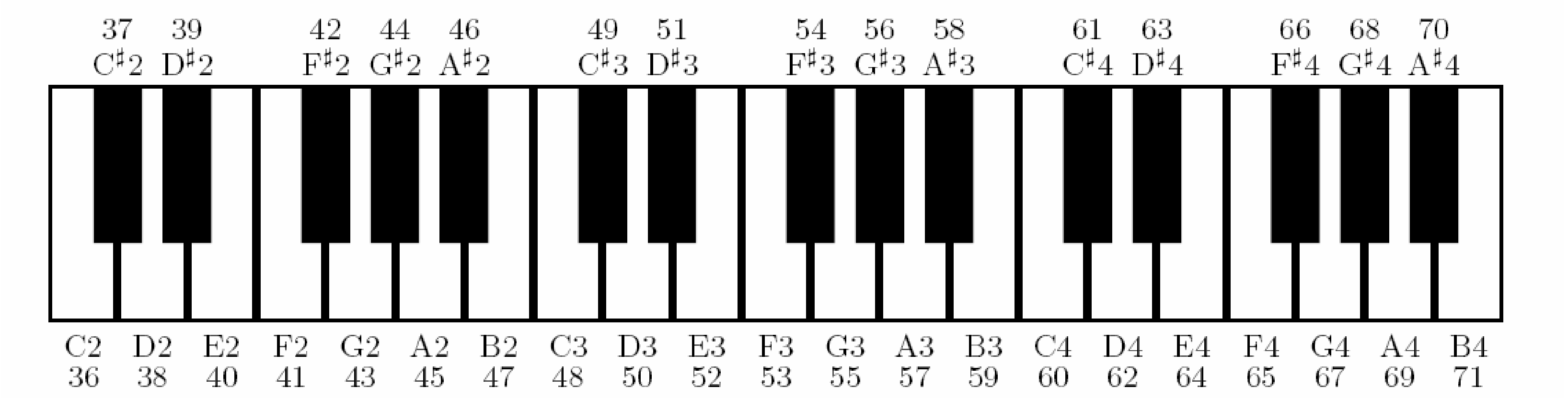
  ➡ idea: build parameters which gather this „similar“ information

- We consider the 12 traditionnal notes of the tempered scale

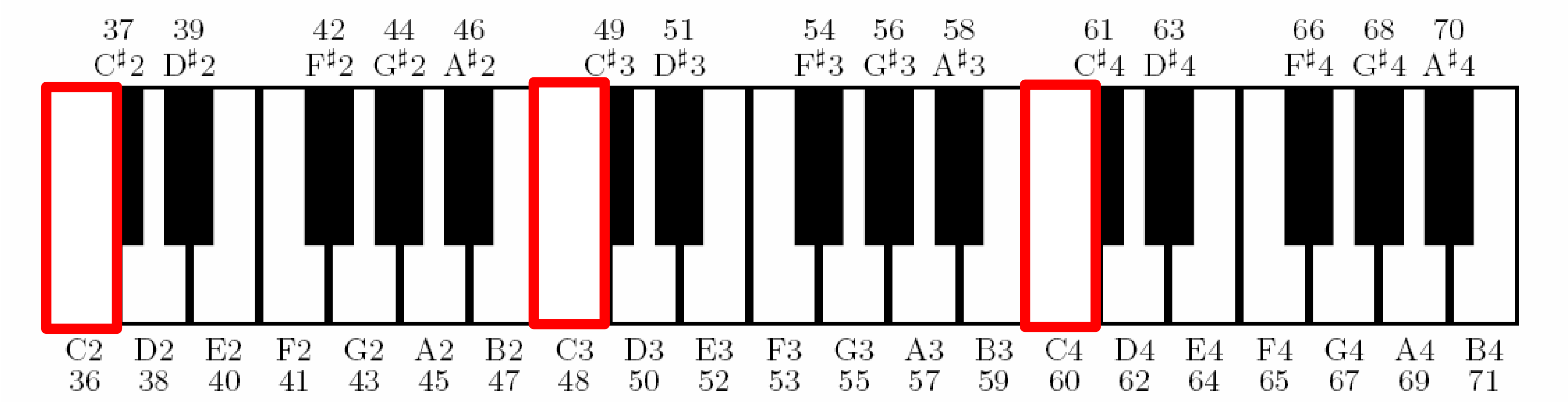
- Chromas are obtained, for a given note, by adding up contributions of all his octaves

  ➡ Obtention of a vector of dimension 12 (the „**chromas**“)

Chroma Features



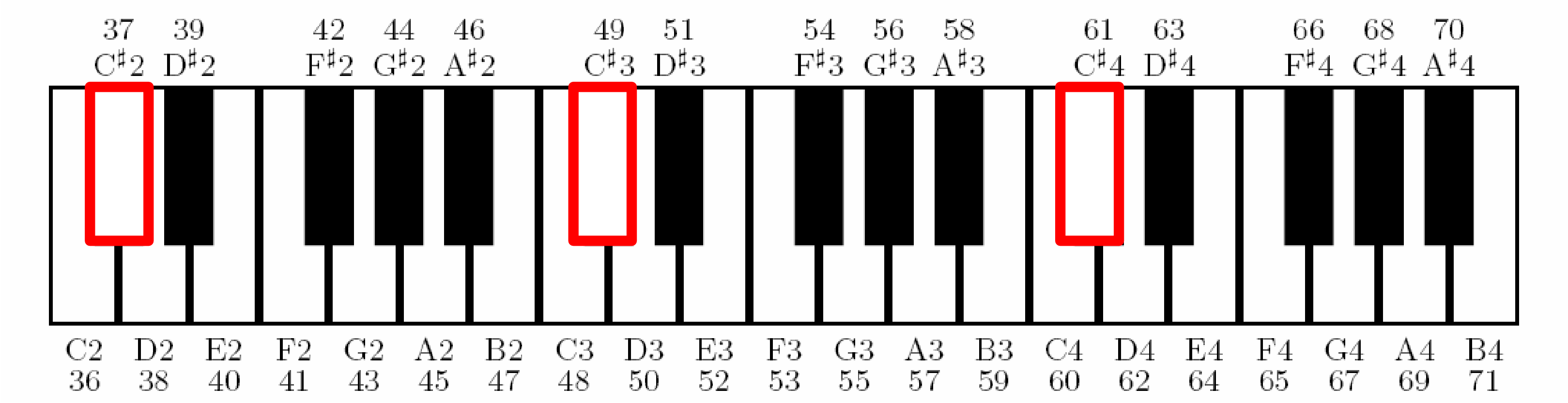
Chroma Features



C2 C3 C4

Chroma C

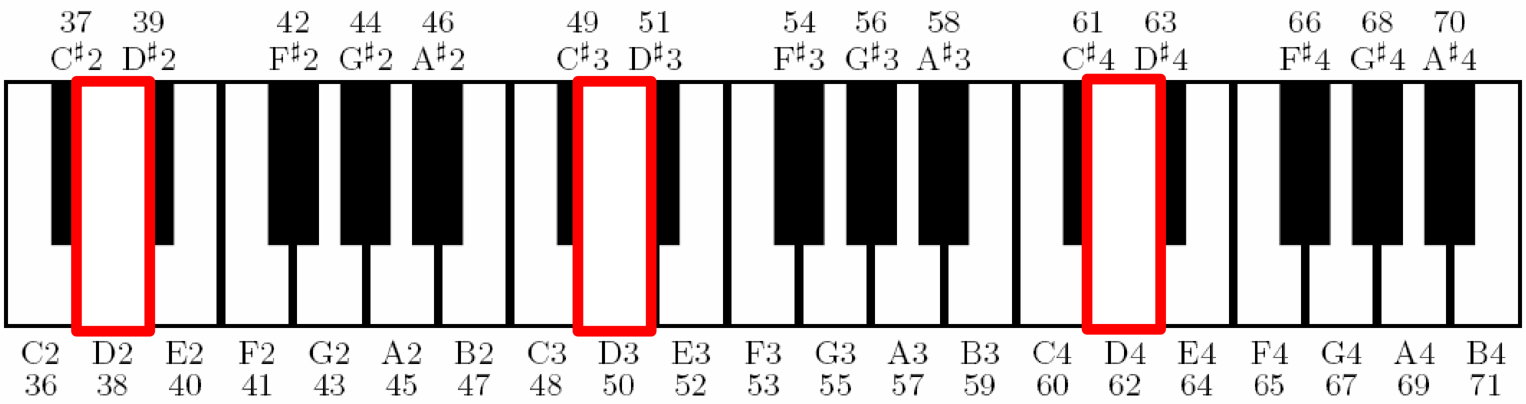
Chroma Features



C#2 C#3 C#4

Chroma C#

Chroma Features

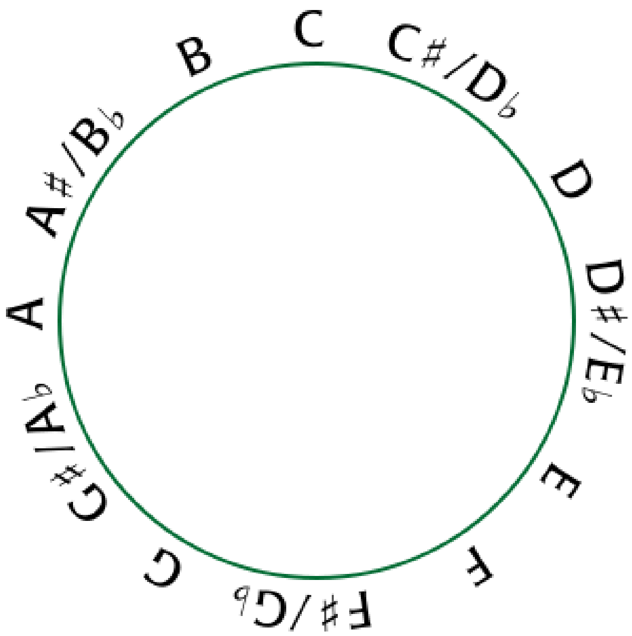


D2 D3 D4

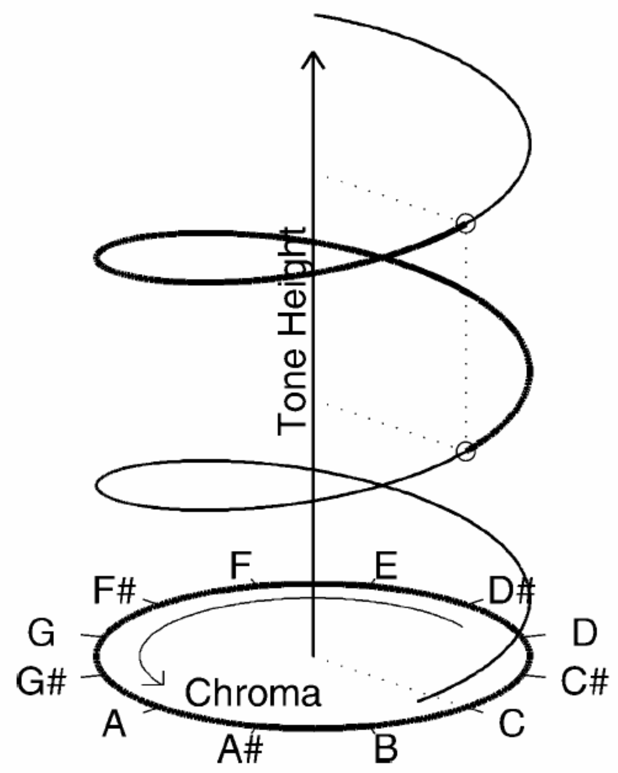
Chroma D

Chroma Features

Chromatic circle



Shepard‘s helix of pitch perception



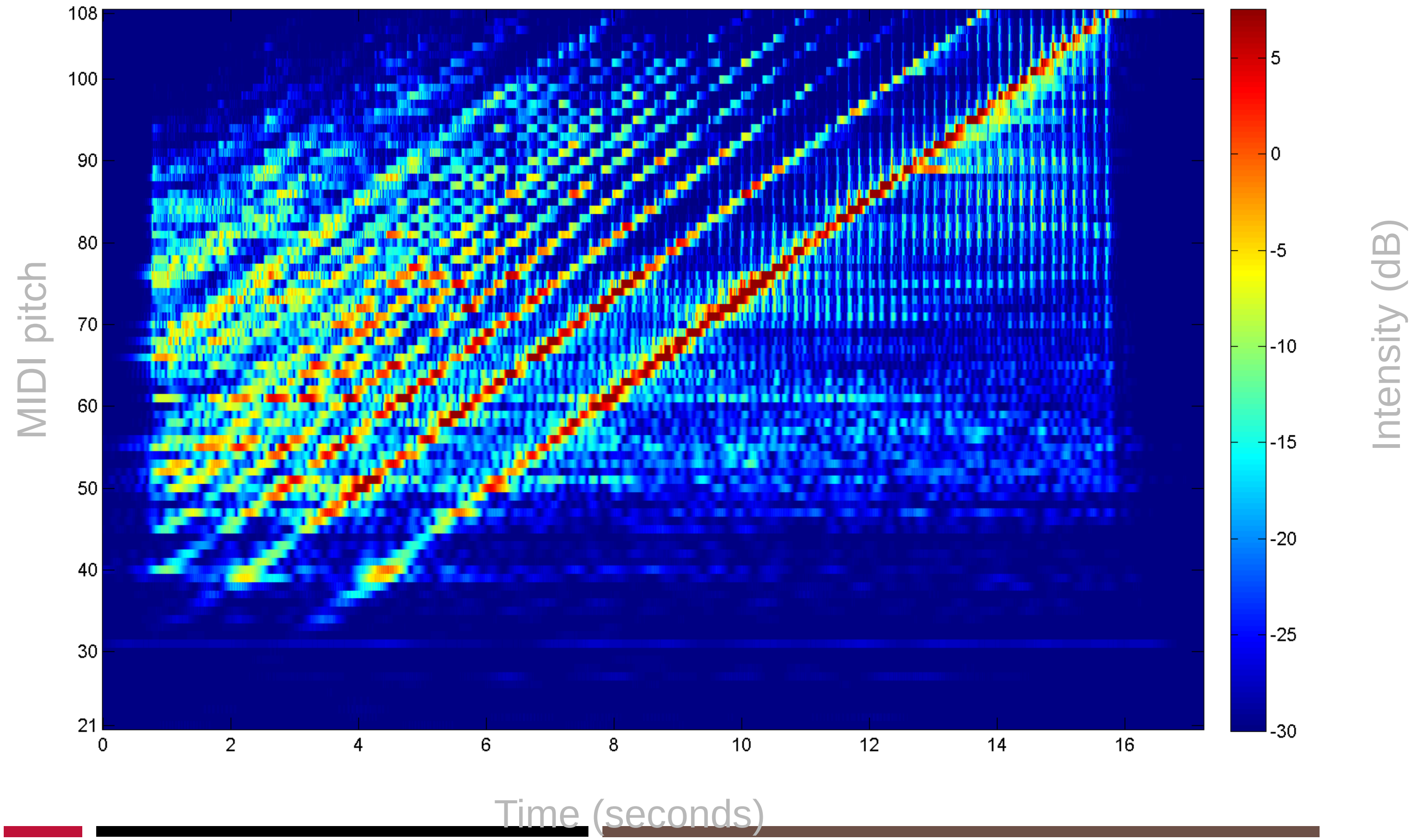
[Gómez, PhD 2006][Bartsch/Wakefield, IEEE-TMM 2005]

http://en.wikipedia.org/wiki/Pitch_class_space

Chroma Features

Example: Chromatic scale

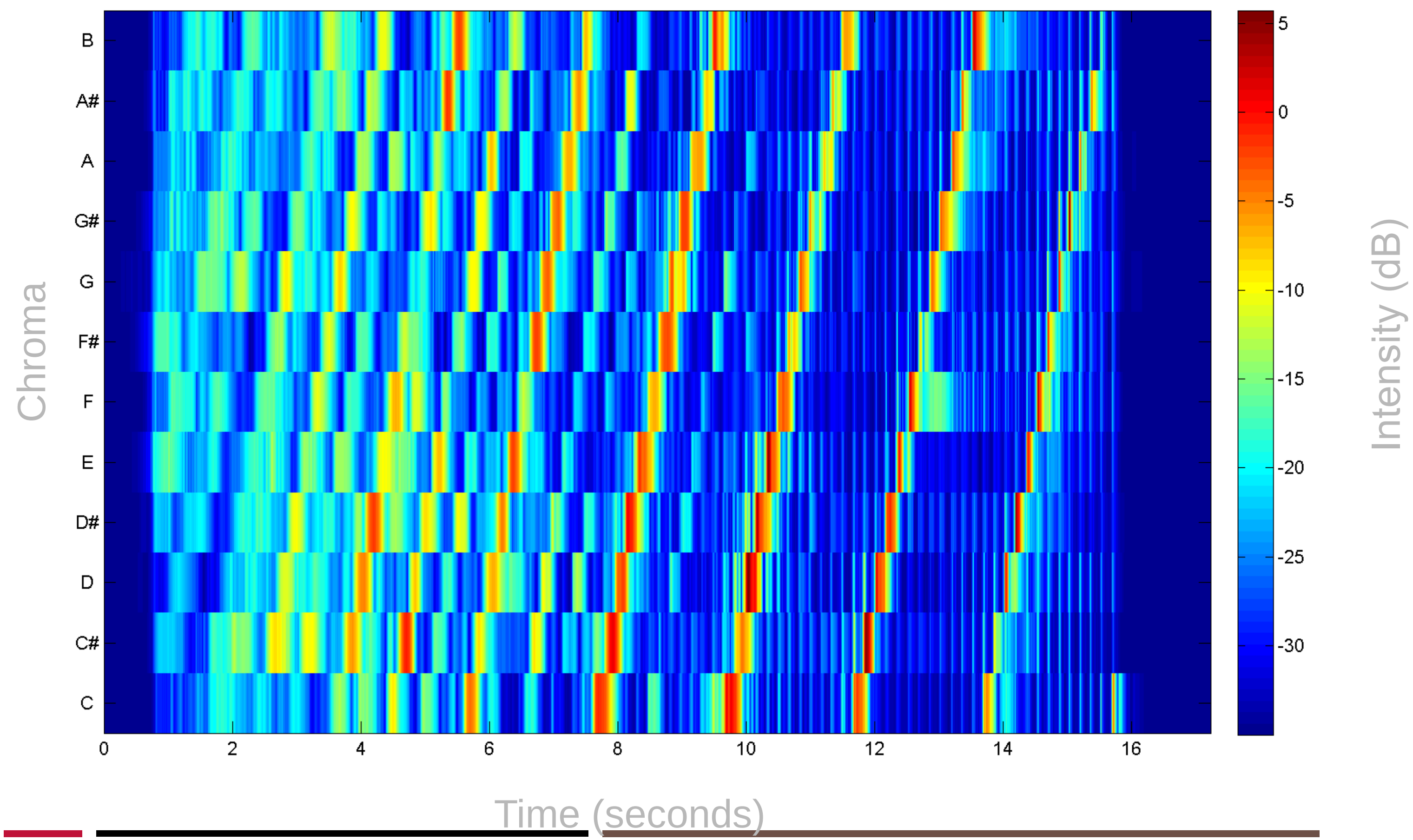
Log-frequency spectrogram



Chroma Features

Example: Chromatic scale

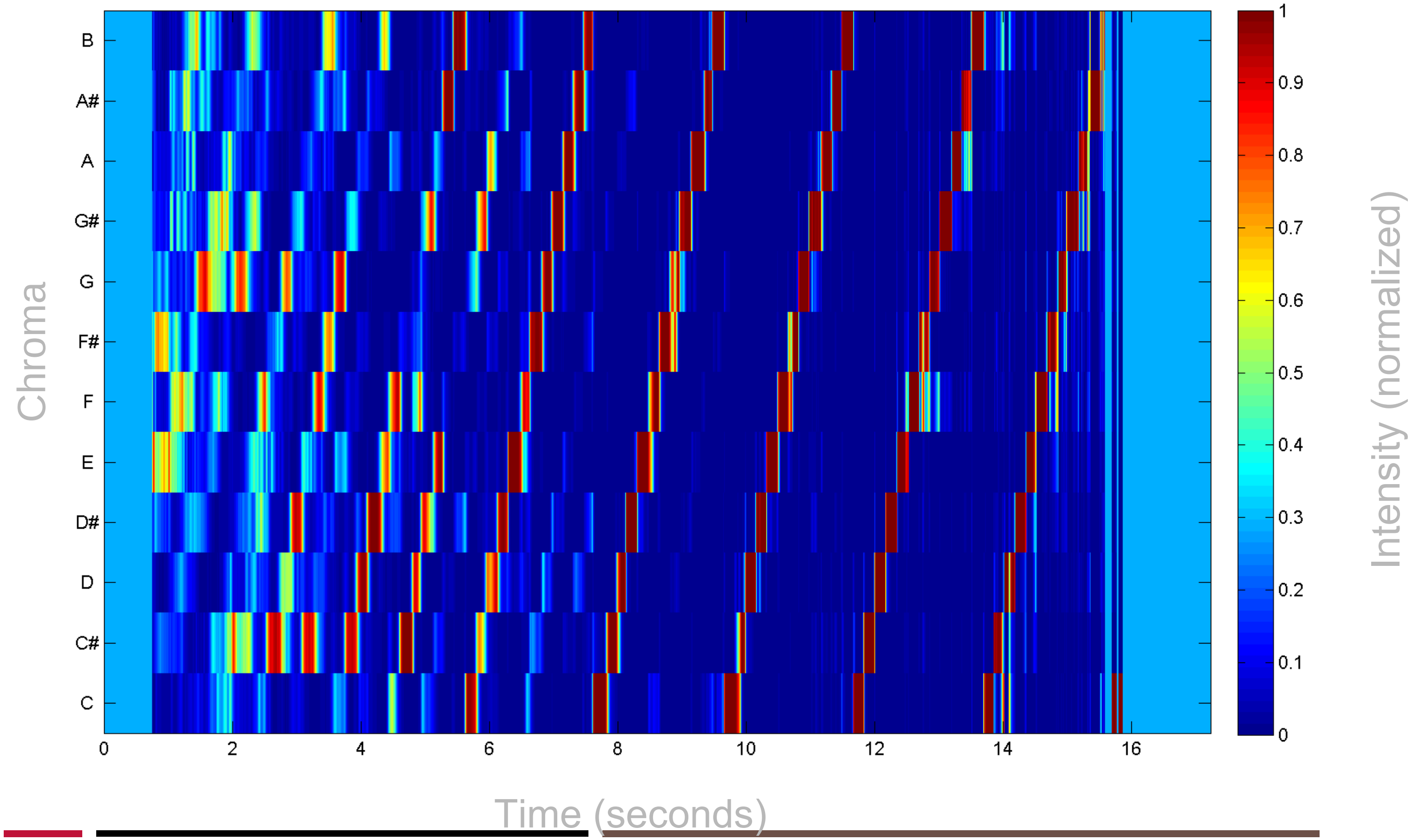
Chroma representation



Chroma Features

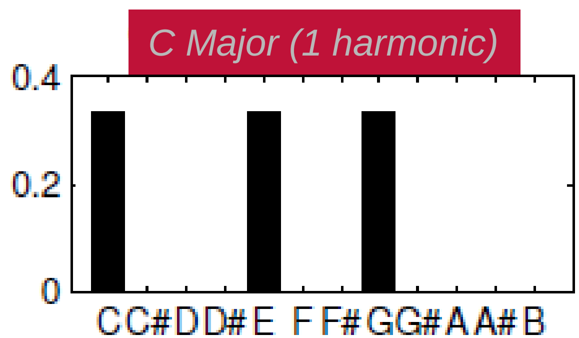
Example: Chromatic scale

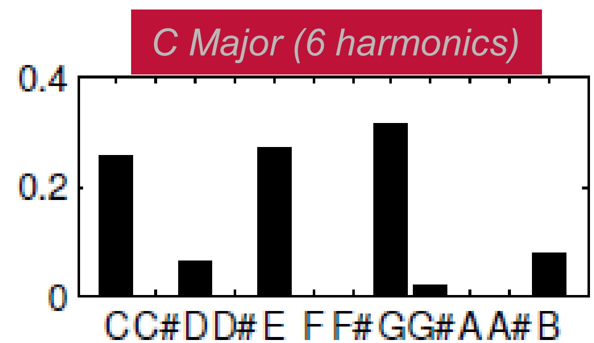
Chroma representation (normalized, Euclidean)



Application to Chord recognition …

- **Using theoretical chroma templates**
  - Examples of 2 chromas templates with or without integrating higher harmonics

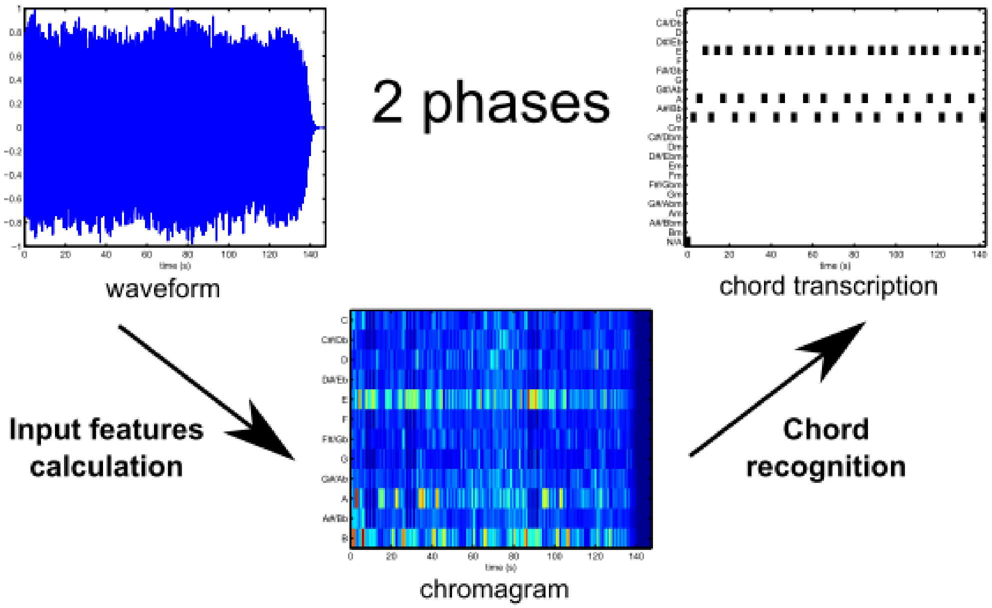




Application to Chord recognition …

- **Chords or/and tonality recognition ,…**

blah2_mono.mpg



From L.Oudre, PhD. Telecom ParisTech 2010

- Other applications:
  - Audio/Audio or Audio/Score alignment
  - Audiofingerprint, ….

Automatic chord recognition

- **A (historical) list of references**

as usual, the first systems <u>define the task</u>, the <u>performance measures</u>, and provide <u>a first test-set</u>;
later systems deals with scalability issues and create large test-set; current systems use this large dataset to train systems using deep-learning

– Frame-based/ template-based approach

• 1999 T. Fujishima. "Realtime chord recognition of musical sound: a system using common lisp music". In Proc. of ICMC,1999.

– Hidden-Markov-Model (HMM) based approaches

• 2003 A. Sheh and D. P. W. Ellis. "Chord segmentation and recognition using em-trained hidden Markov models". In Proc. of ISMIR, 2003

• 2007 H. Papadopoulos and G. Peeters. "Large-scale study of chord estimation algorithms based on chroma representation". In Proc. of IEEE CBMI, 2007

– Splitting into bass/middle/chroma

• 2012 Yizhao Ni, Matt McVicar, Raul Santos-Rodriguez, and Tijl De Bie. "An end-to-end machine learning system for harmonic analysis of music". IEEE TASLP, 2012.

–

Deep learning approaches

• 2013 Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. "Audio chord recognition with recurrent neural networks". In ISMIR, 2013

• 2016 Filip Korzeniowski and Gerhard Widmer. "Feature learning for chord recognition: the deep chroma extractor". In ISMIR, 2016.

• 2017 B. McFee and J. P. Bello. "Structured training for large-vocabulary chord recognition". In Proc. of ISMIR, 2017

• 2021 C. Weiß and G. Peeters. "Training deep pitch-class representations with a multi-label CTC loss". In Proc. of ISMIR, 2021

How to perform Music recognition or Audiofingerprint ?

Audio Identification ou AudioID

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

Audio identification

Information of the recording (e.g. fro music: title, artist, etc.., …)

- **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 .

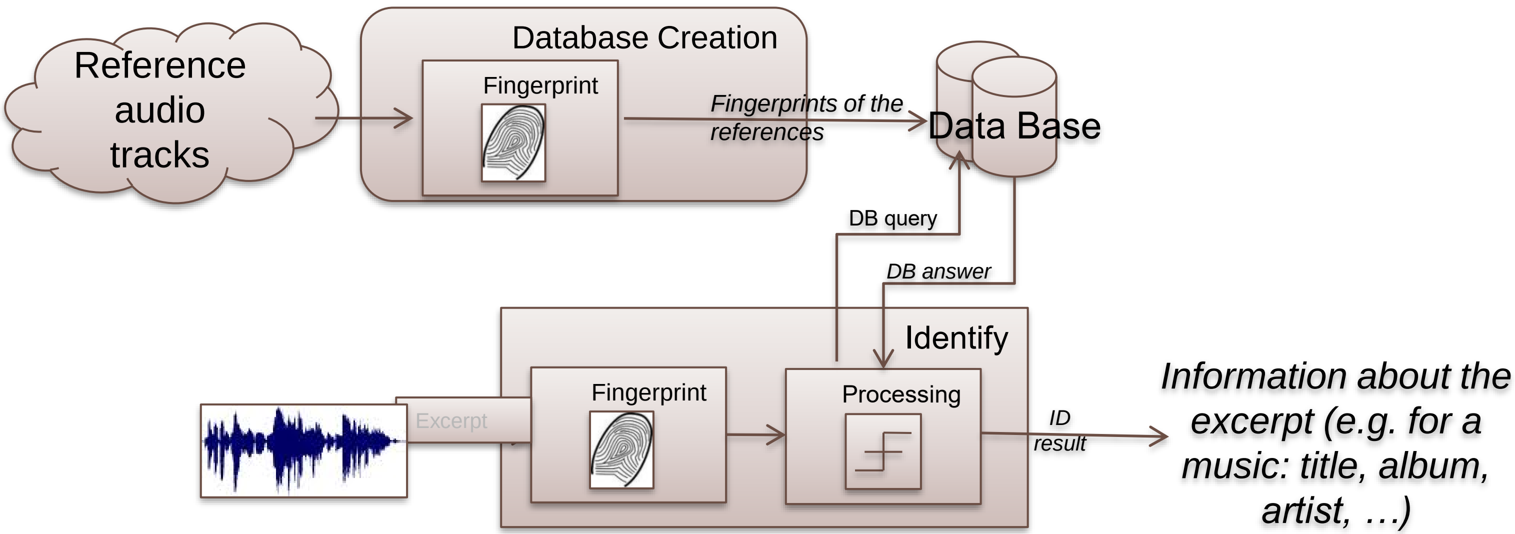
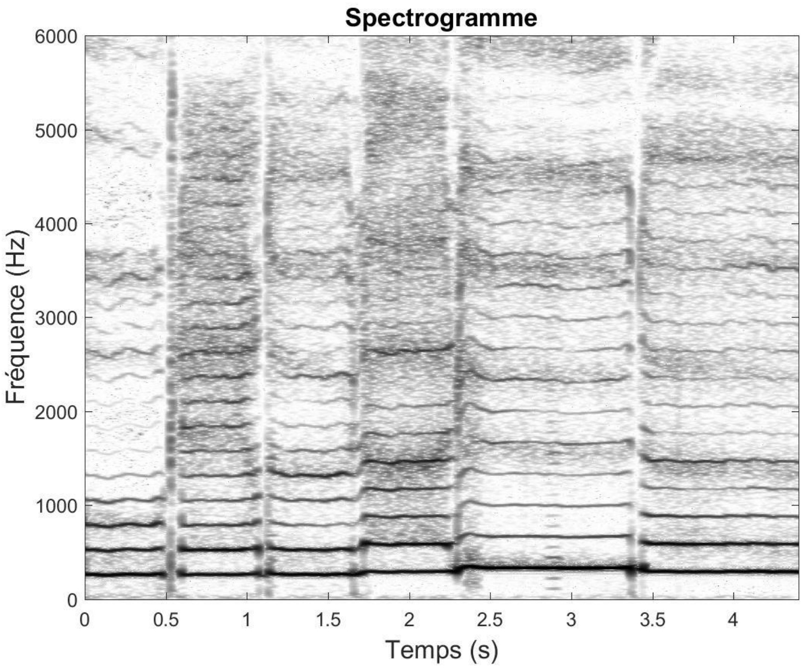
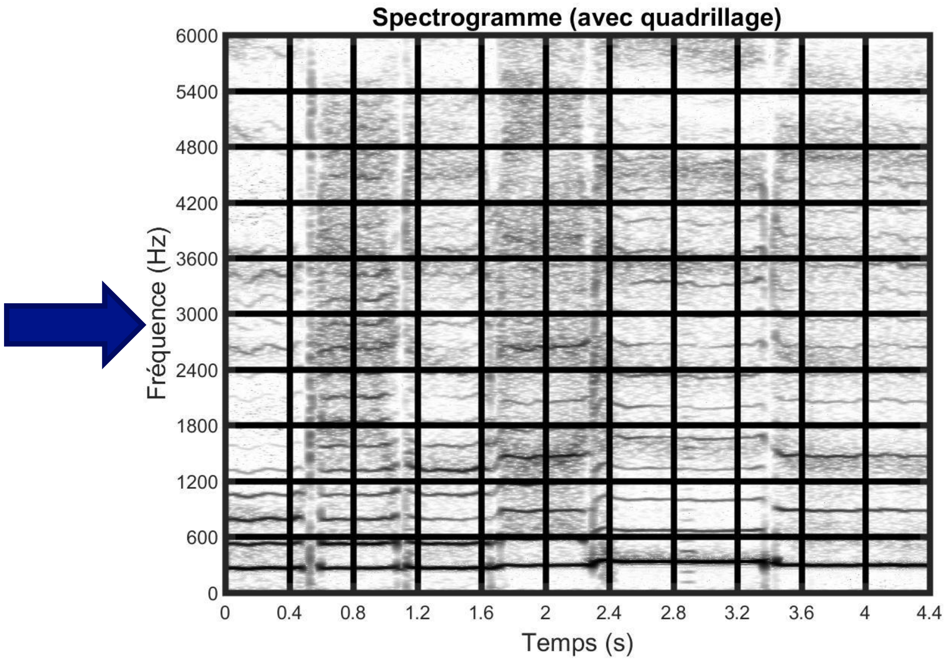


Figure from Sébastien Fenêt

Signal model : from spectrogram to “schematic binary spectrogram”

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

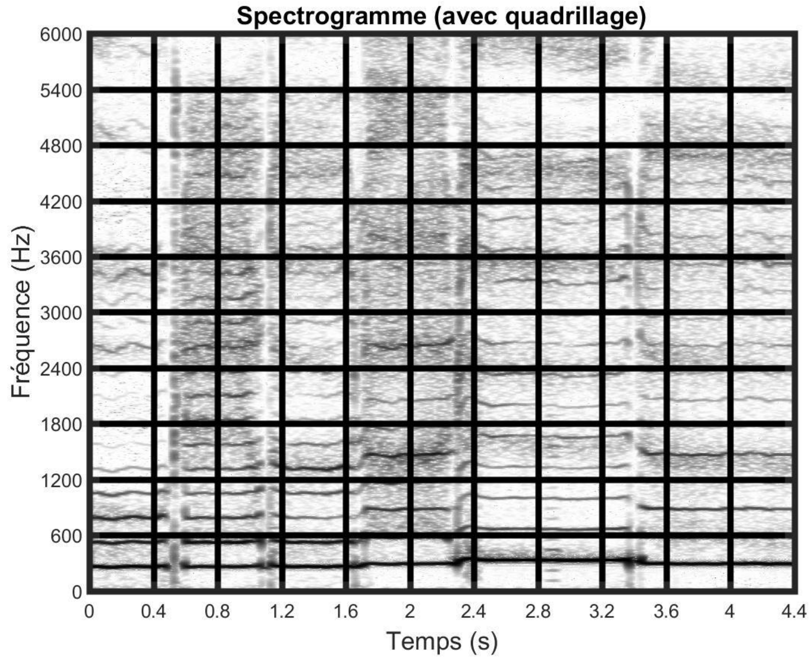


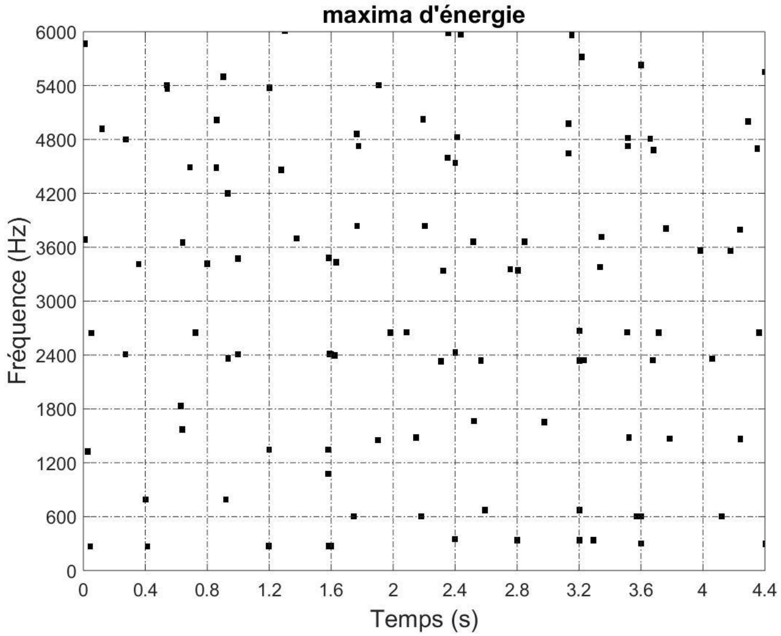


From A. Wang, “An industrial strength audio search algorithm,” in ISMIR, 2003. (The original Shazam algorithm)

Signal model : from spectrogram to “schematic binary spectrogram”

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

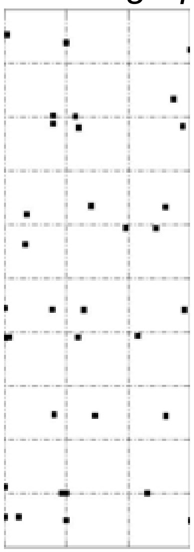




Efficient research strategy

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

Test fingerprint



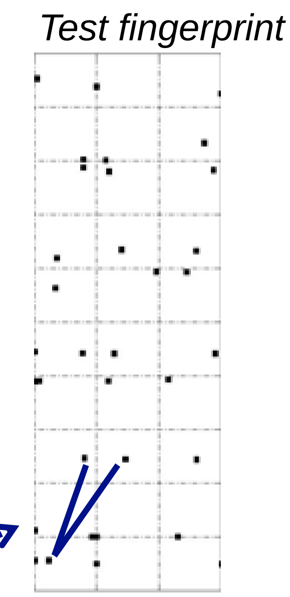
Potential strategies

- 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”

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.

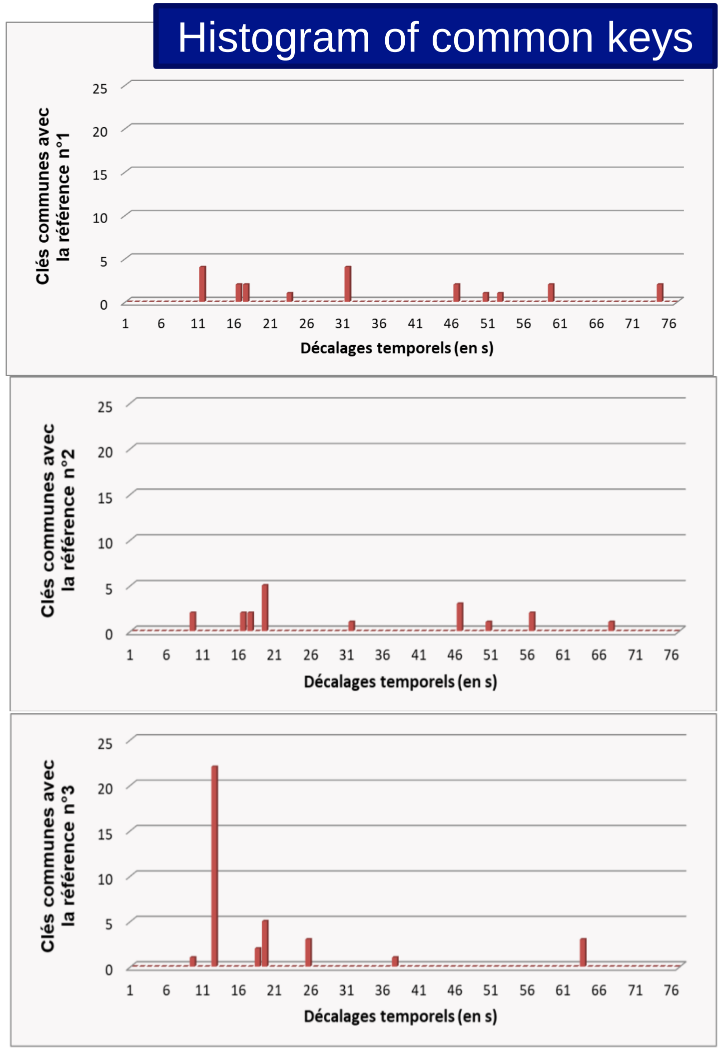
3. **Recognition:** The reference which has the most keys in common at a constant Δt is the recognized recording

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

Reference 1

Reference 2

Reference 3



Recognized recording

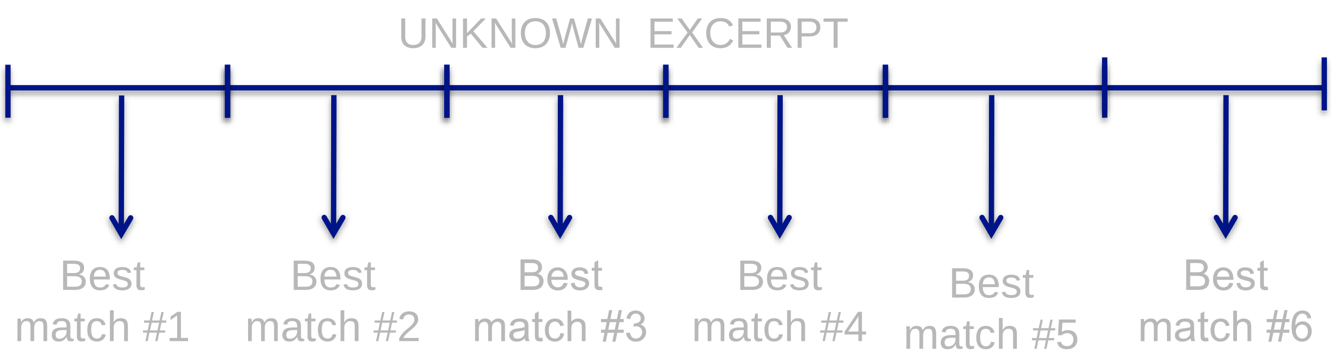
Droits d'usage autorisé

TELECOM Paris

IP PARIS

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%)

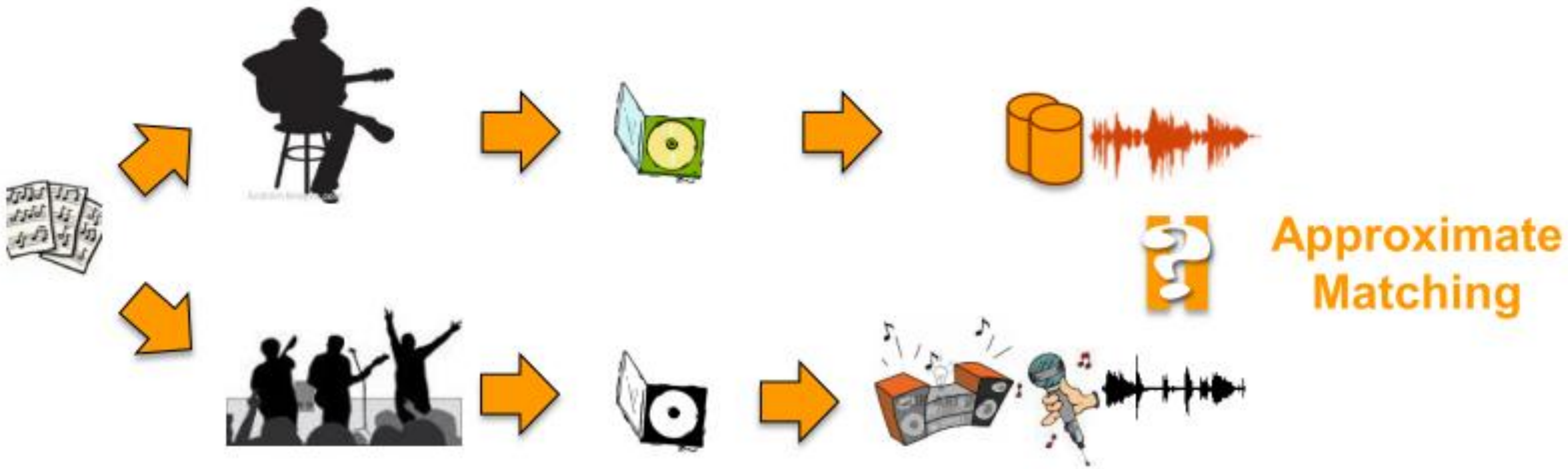
Limitations and other solutions

- **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)

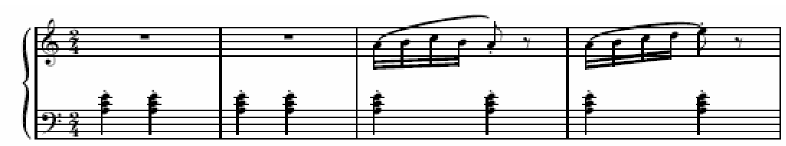
  - For music recommendation, second screen applications, …

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

Some dimensions of the musical signal …

Pitch, Harmony..

Tempo, rhythme,…



Timbre, instruments,…

Polyphony, melody, ….

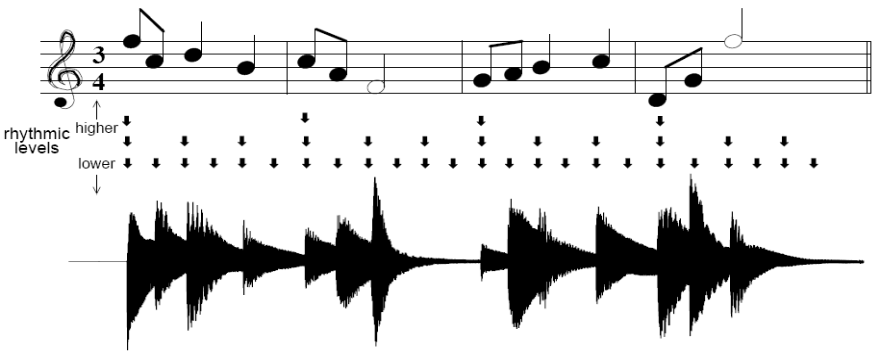
Interest of rhythmic information

- **Rhythm: is an essential component of the musical signal**

- **Numerous applications:**
  - Automatic mixing, DJing : synchronisation of tempo, rhythm,..
  - Smart Karaoké
  - Automatic playlists (podcast,…)…
  - Genre reconnaissance
  - Music/video synchronisation
  - Smart jogging shoes ? »
  - ..

Rhythm or Tempo estimation

- **Rythme: An intuitive concept easy to understand but difficult to define !!**

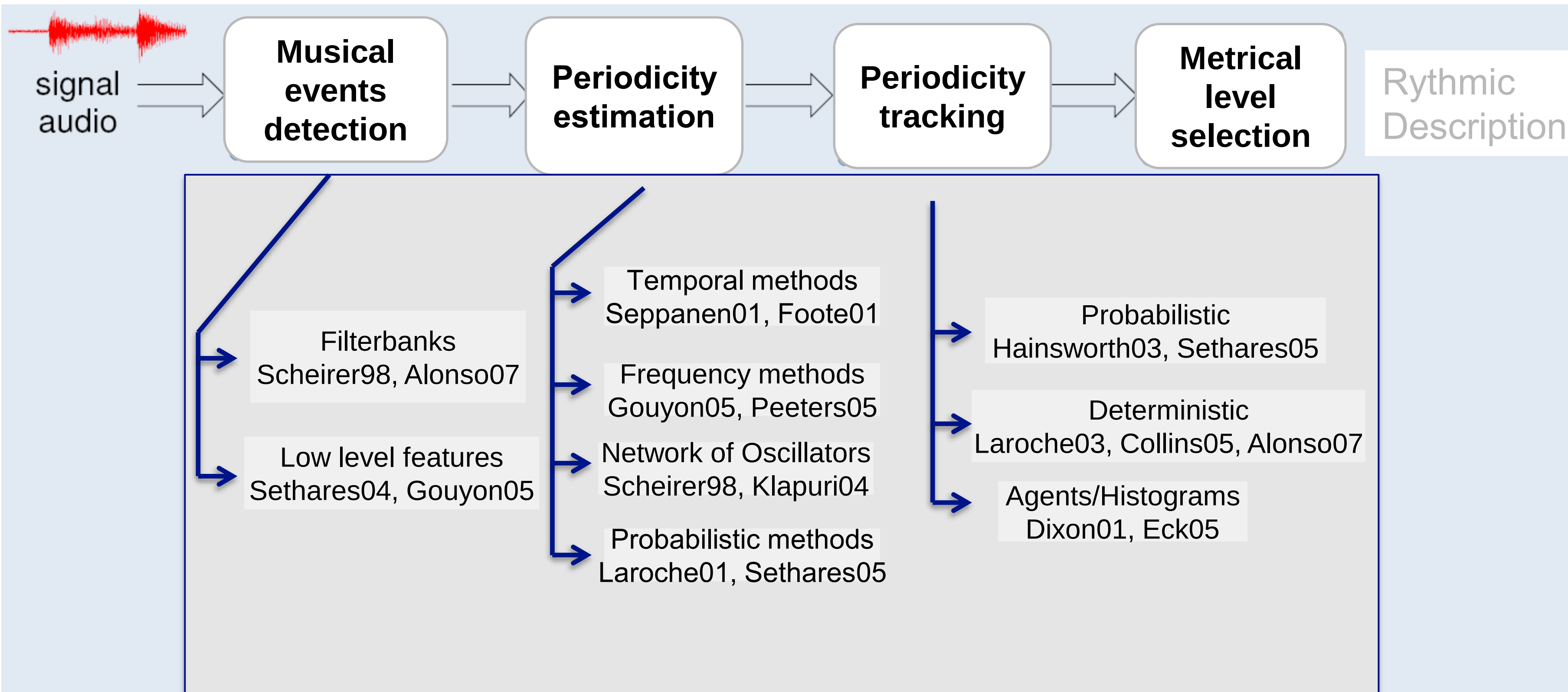


- **Handel (1989): *« The experience of rhythm involves movement regularity, grouping and yet accentuation and differentiation »***

- **There is not not a unique perception of rythm !**

Rhythm or "Tempo" Extraction

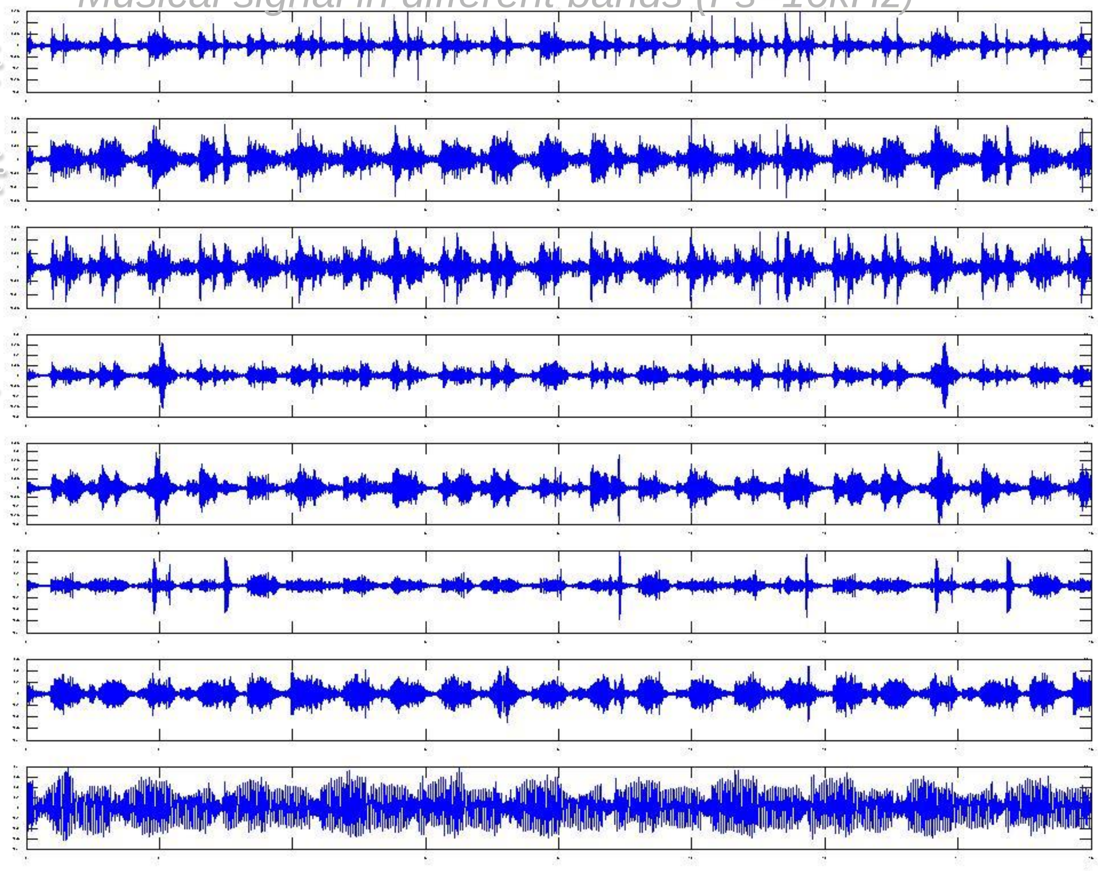
- **Principle**



Discovering the rhythmic information…

- **Use of filterbanks (e.g. separating the frequency information…)**

Musical signal in different bands (Fs=16kHz)



Bands 8-16 (3500 – 8000 Hz)

Band 7 (3000 – 3500 Hz)

Band 6 (2500 – 3000 Hz)

Band 5 (2000 – 2500 Hz)

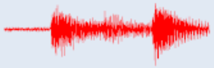
Band 4 (1500 – 2000 Hz)

Band 3 (1000-1500 Hz)

Band 2 (500 – 1000 Hz)

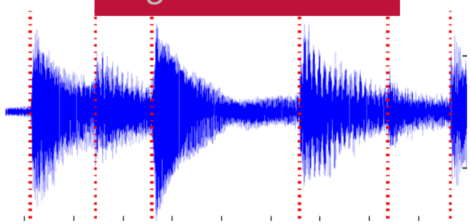
Band 1 (0 – 500 hz)

Rhythm or “Tempo” Extraction

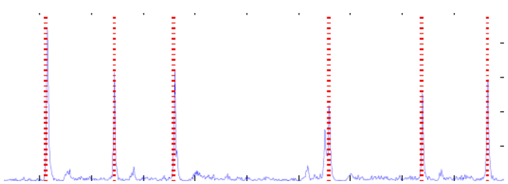


signal audio → **Musical events detection** → **Periodicity estimation** → **Periodicity tracking** → **Metrical level selection**

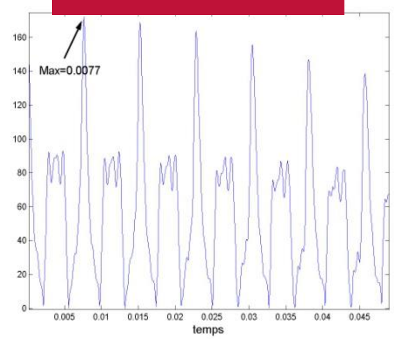
Signal + Onsets



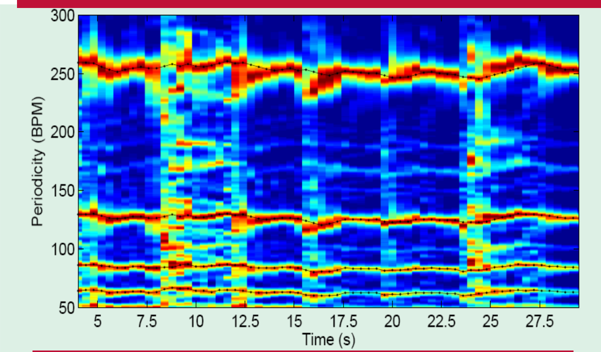
« Detection function »



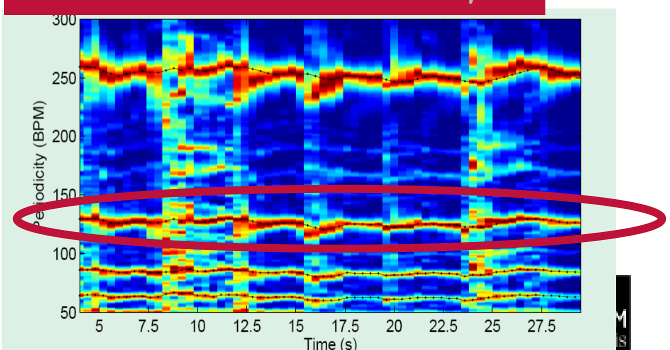
Autocorrelation



Periodicity tracking (« tempogramme»)

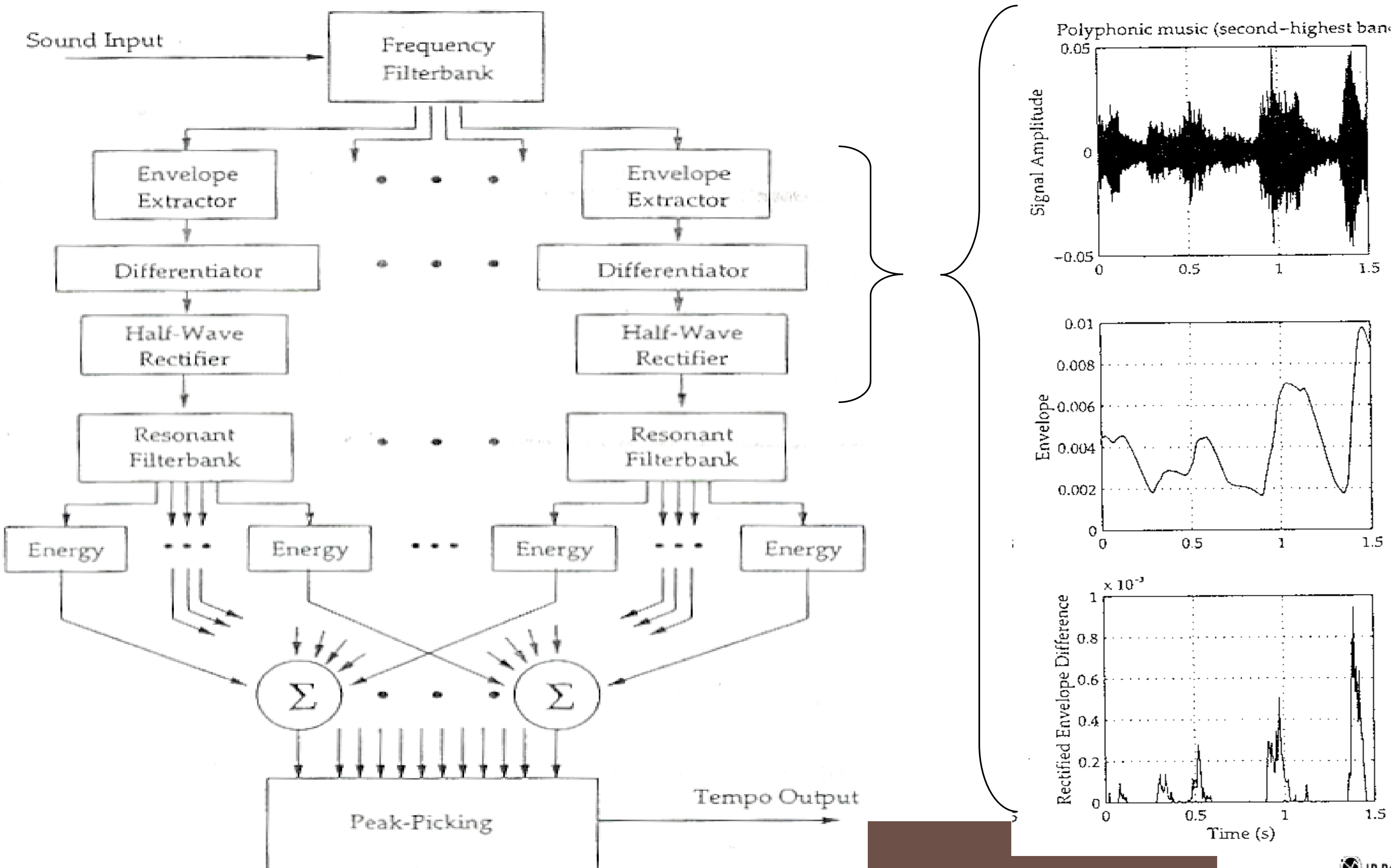


Metrical level selectionTempo



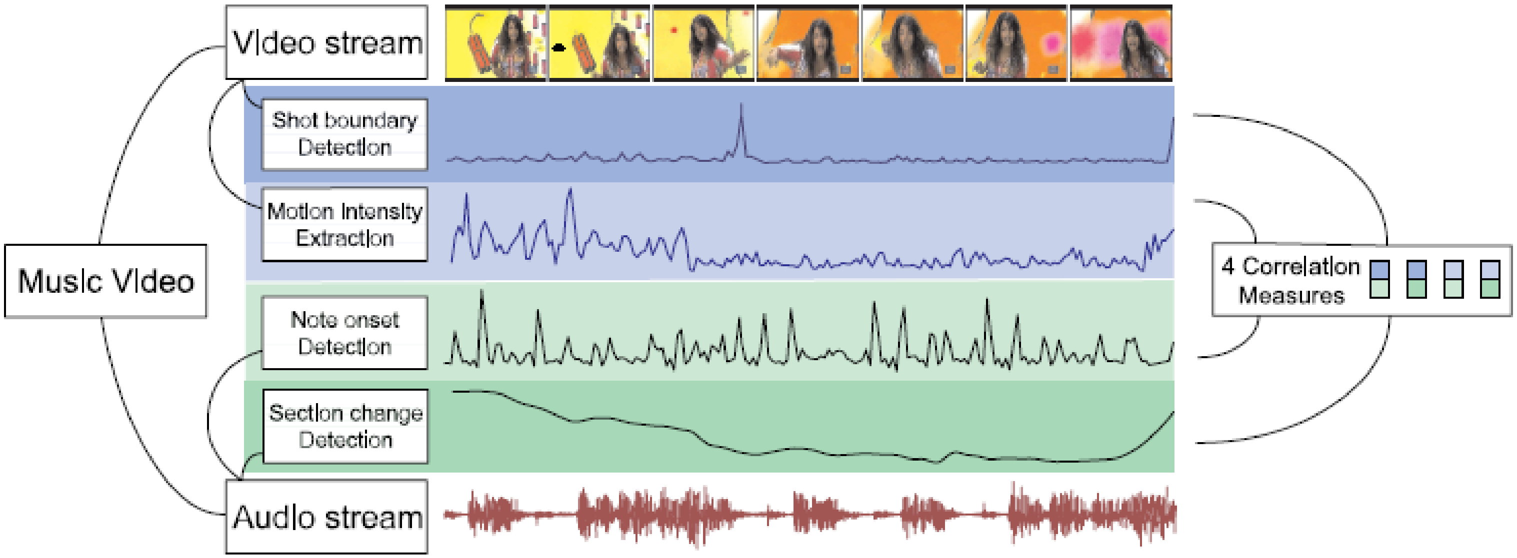
Tempo and beat extraction

- **A filterbank approach (Scheirer, 1998)**



Rhythm and tempo estimation : a feature a great interest

- **Audio-based video retrieval**
  - Exploit semantic correlations sémantiques between audio and vidéo
  - Application: search for audio that « fits » the video stream



O. Gillet, S. Essid and G. Richard, On the Correlation of Audio and Visual Segmentations of Music Videos. IEEE Transactions on Circuits and Systems for Video Technology, 17 (2), March 2007, pp 347-355.

Current trends …

- **Estimate rhytms (tatums,tempo) but also downbeat (but higher level semantic)**

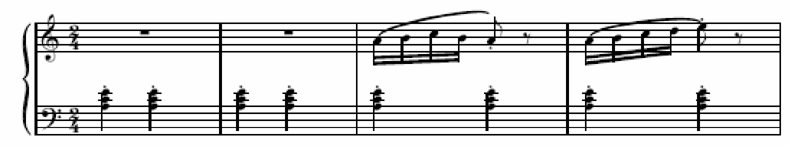
- To exploit machine learning (and deep learning in particular)

- Use and combine multiple representations
  - Rhythm is intrinsically multi-dimensionnal
  - Downbeat depends on melody, chords, bass, etc …

Some dimensions of the musical signal …

Pitch, Harmony..

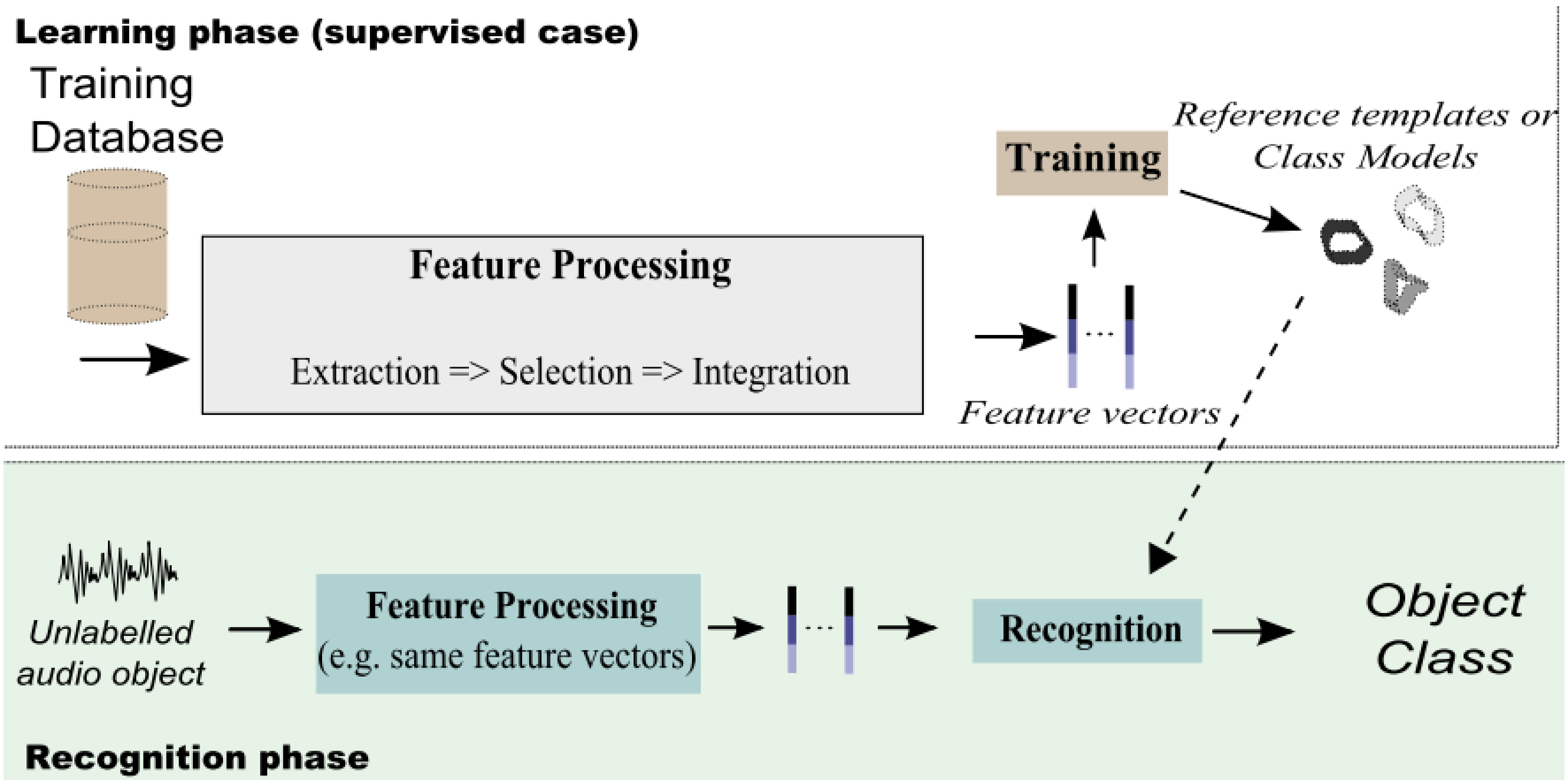
Tempo, rhythme,…



Timbre, instruments,…

Polyphony, melody, ….

Traditional Classification system



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

Timbre: What is this ?

- *A possible definition:* « The attribute of auditory perception that allows to differentiate 2 sounds of equal pitch and equal intensity.»

- Closely related to sound source identification and auditory organization

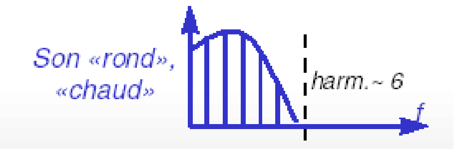
- Examples of sounds with the same pitch and root-mean-square (RMS) levels, but different timbre:

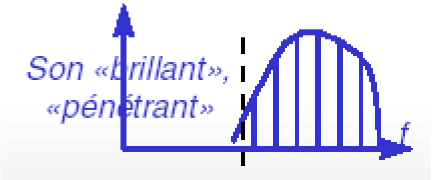
- Early work (*PhD theses*) addressing musical instrument recognition: [Essid06], [Kitahara-07], [Eronen-09]

Features for describing the timbre ?

- **Numerous feature were proposed:**
  - Spectral centroid

$$CGS = \frac{\sum_{k=1}^{N} k.|X_k|}{\sum_{k=1}^{N} |X_k|}$$





  - Spectral flux (*e.g derivative of spectrogram*)
  - Log attack time
  - Spectral irregularity
  - Spectral envelope
  - Perceptual model
  - Onset Spectral « Asynchrony »
  - Wavelet coefficient
  - Harmonic / noise separation
  - Entropy,
  - Entropy variation,
  - **Mel-Frequency Cepstral Coefficients (MFCC)**

Features for describing the timbre

- **Why it is interesting to rely on a filterbank analysis**

  - Allows to separate the information localised in specific frequency regions
  - Mimics (in a rudimentary way) the human auditory perception

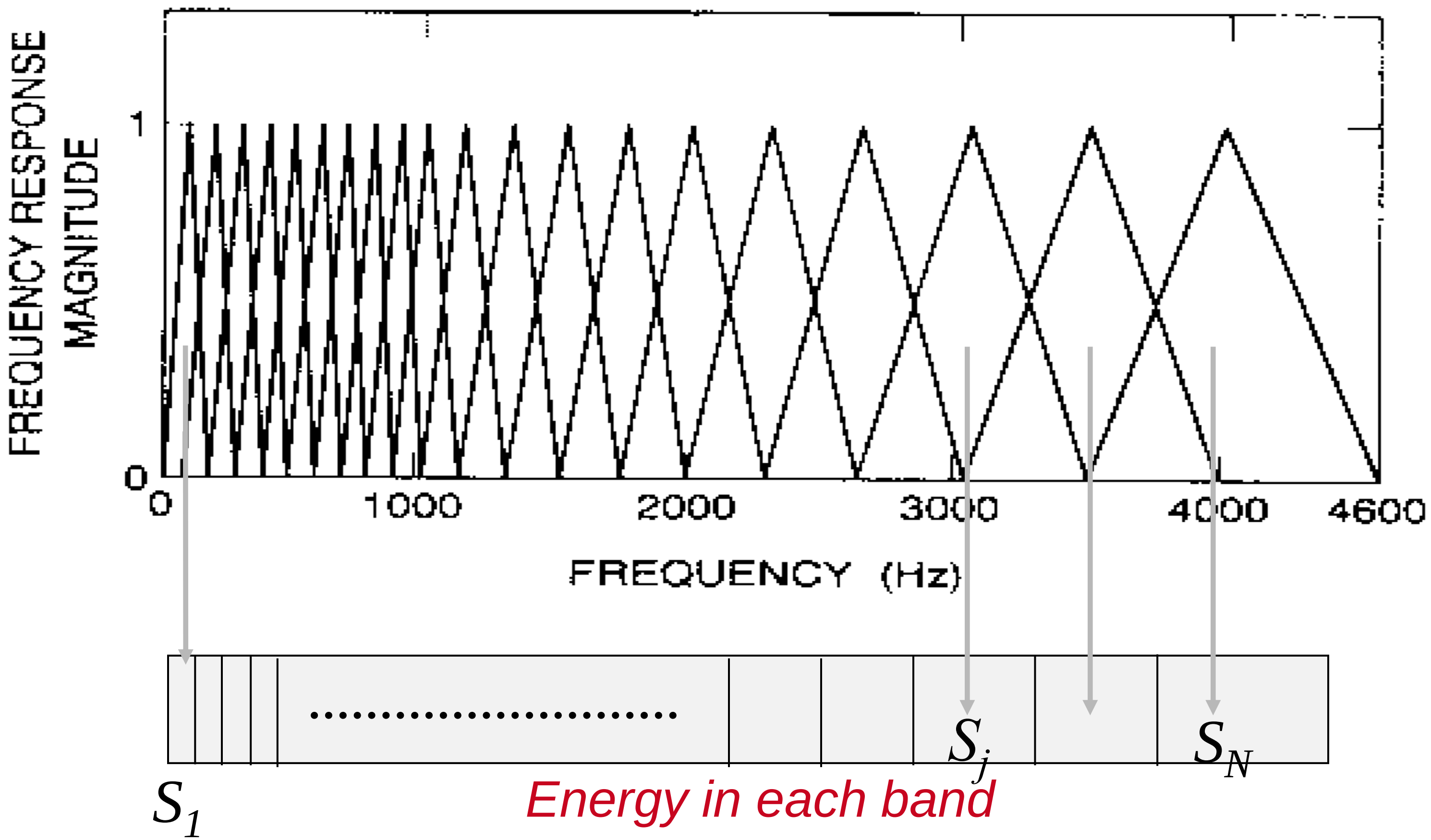
  - Possibility to use perceptual scales

    – Mel scale: corresponds to an approximation of perception of sound pitch (e.g. Tonie)

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

Filter banks distributed on a Mel Scale

- **Mel scale filtering (from Rabiner93)**



Energy in each band

Cepstral représentation

- **Interest**
  - Source/filter model of speech production

$$s(t) = g(t) * h(t)$$

✓ Source-filter model in the cepstral domain

$$S(\omega) = G(\omega)H(\omega)$$

✓ Cepstre (real): a sum of two almost non-overlapping terms

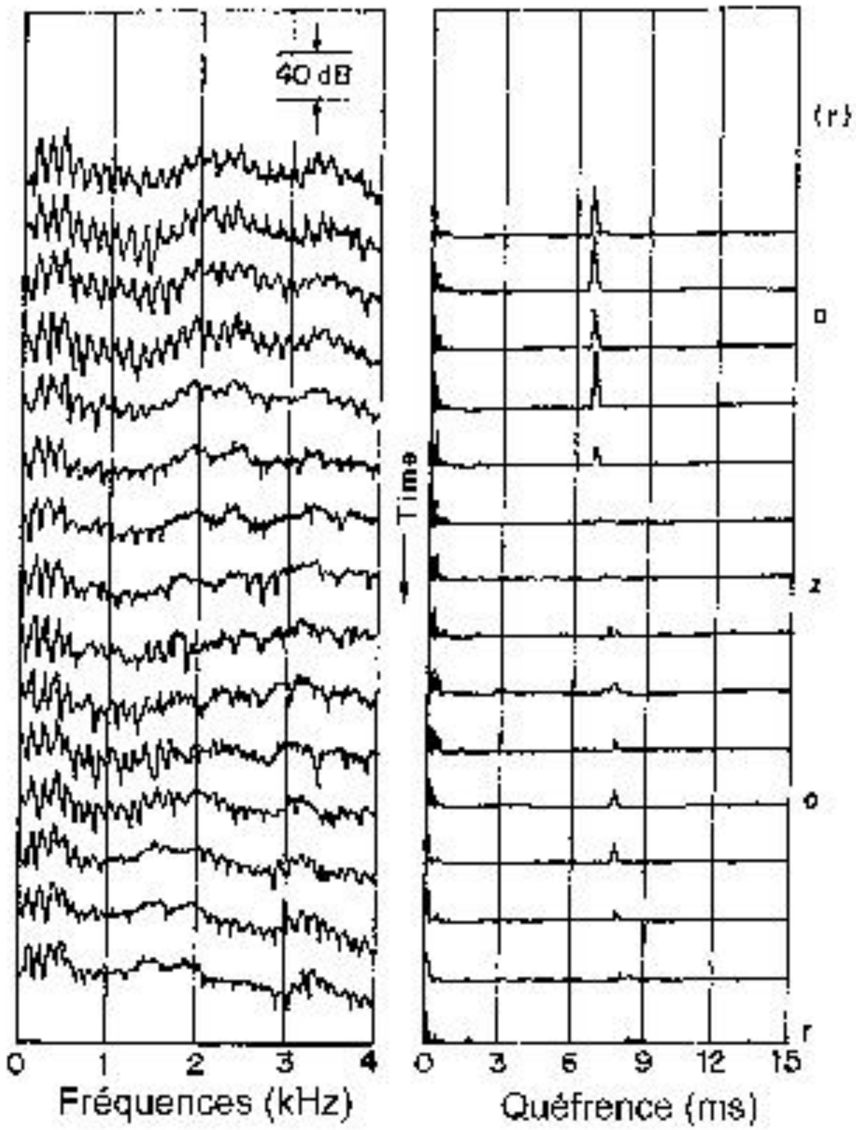
$$c(\tau) = FFT^{-1} \log|S(\omega)| = FFT^{-1} \log|G(\omega)| + FFT^{-1} \log|H(\omega)|$$

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log|X(k)| e^{2j(\pi)kn/N}$$

Cepstral Representation (from Furui2001)

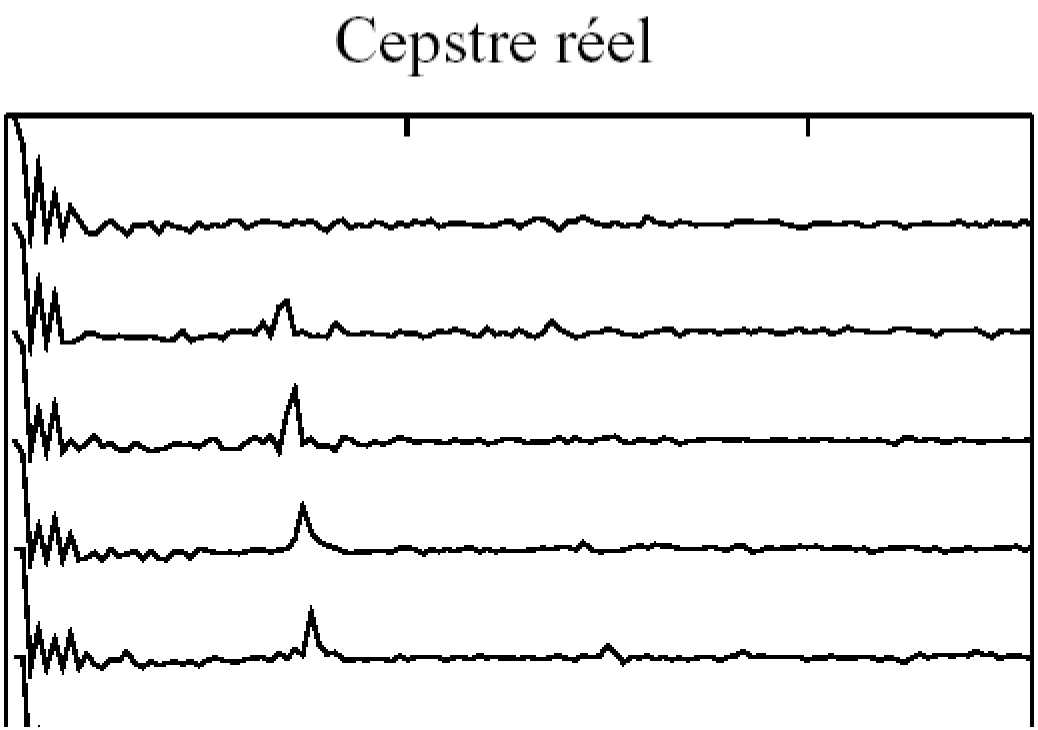
- **Examples:**
  - of Spectrum (left)
  - of Cepstrum c(τ) (right)

- **τ is homogeneous with a time and is called quefrency**



Cepstral Representation

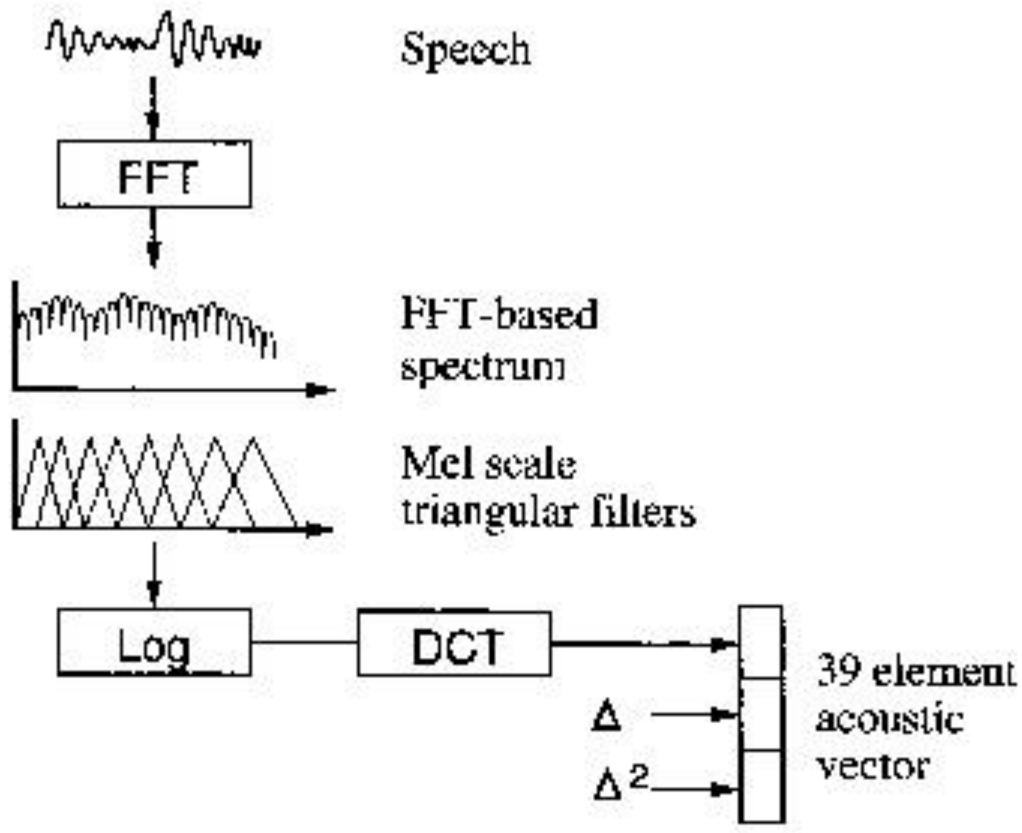
- **Separation of the vocal tract contribution and of the source contribution by liftering**



MFCC

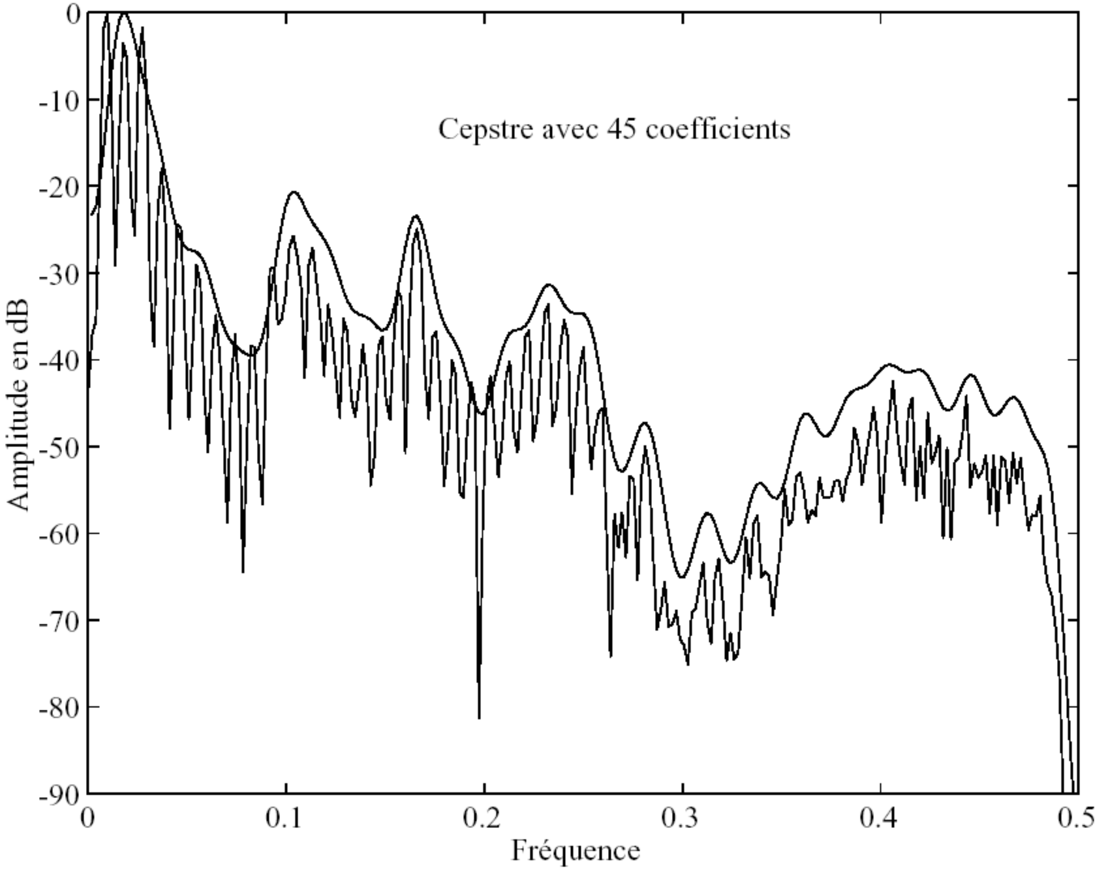
« Mel-Frequency Cepstral Coefficients »

- **The most common features (from Furui, 2001)**



Cepstral smoothing

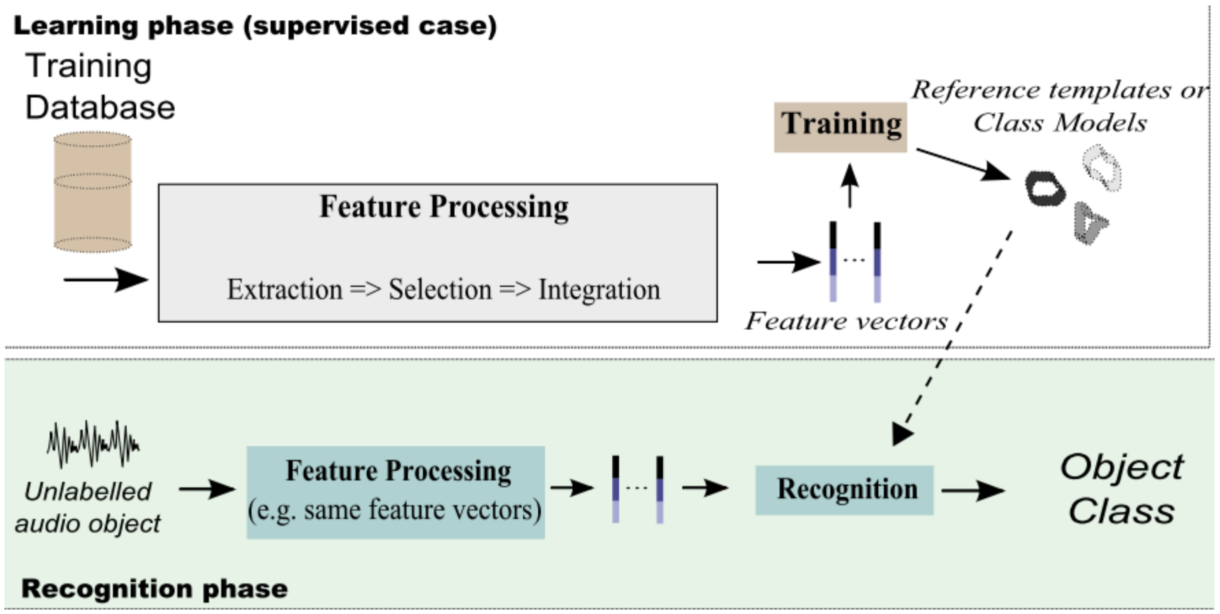
- **Envelope estimation by cepstrum:**
  - Compute real cesptrum $C_{n,}$, then low quefrency liftering
  - (log) Spectral envelope reconstruction $E = FFT(C_n)$



Classification

With the example of "automatic musical instrument recognition"

- ***Aim of classification:***
  - Find the class (i.e the instrument) from the features computed on the music signal



Some of the most common classifications schemes used in audio classifications

- **K-nearest neighbors (for simple problems)**
- **Gaussian Mixture Models (GMM)**
- **Support Vector machines**
- **Linear Regression**
- **Decision tree, Random forest**
- **…**

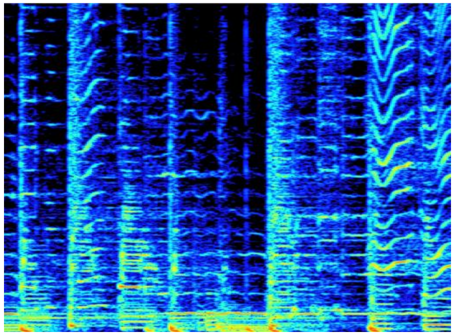
- **And more recently Deep neural networks**
  - Recurrent Neural networks (RNN) , Gated Recurrent Units (GRU)
  - Convolutional Neural Networks (CNN applied on spectrograms)
  - Long-Short Term Memory (LSTM)
  - Generative Adversarial Networks (GANs)

A view of Deep learning for audio

Deep learning for audio

Differences between an image and audio representation

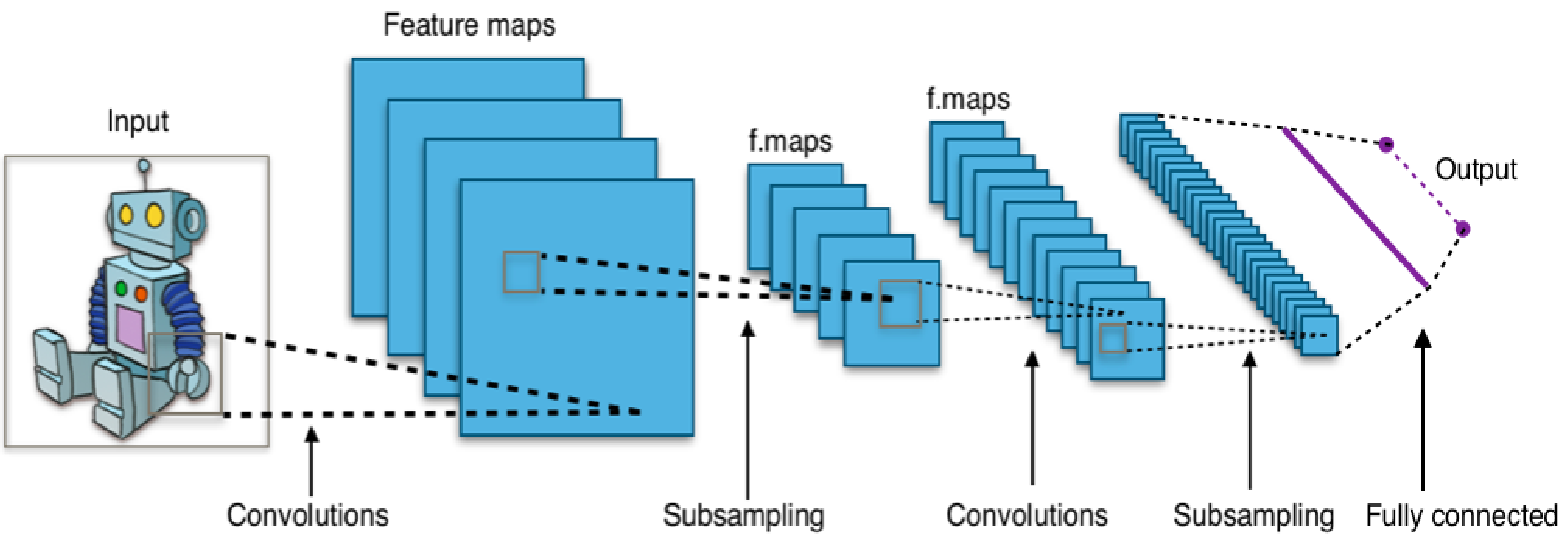
- x and y axes: **same concept** (spatial position).
- Image elements (cat's ear) : **same meaning** independently of their positions over x and y.
- **Neighbouring pixels** : often correlated, often belong to the same object
- **CNN are appropriate :**
  - Hidden neurons locally connected to the input 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**

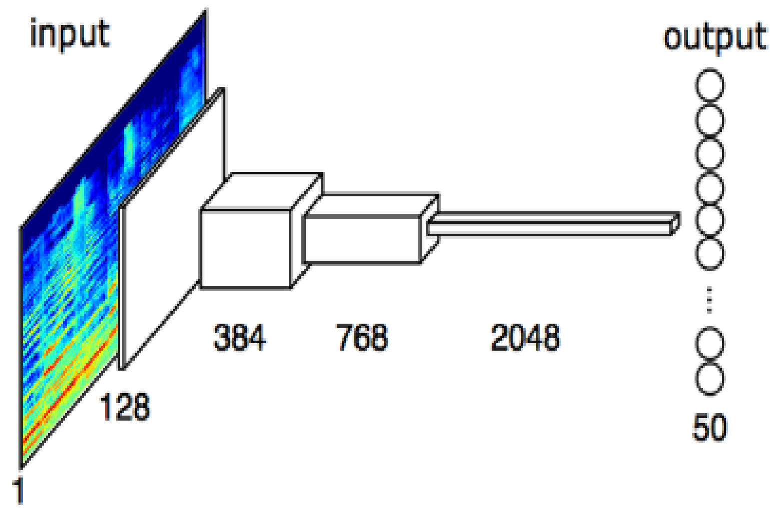
G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021

A typical CNN



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

Music automatic tagging with CNN



Tags are include:
- **emotion** (sad, anger, happy),
- **genre** (jazz, classical)
- **instrumentation** (guitar, strings, vocal, instrumental).

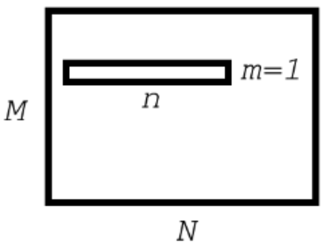
FCN-4
Mel-spectrogram *(input: 96×1366×1)*
Conv 3×3×128
MP (2, 4) *(output: 48×341×128)*
Conv 3×3×384
MP (4, 5) *(output: 24×85×384)*
Conv 3×3×768
MP (3, 8) *(output: 12×21×768)*
Conv 3×3×2048
MP (4, 8) *(output: 1×1×2048)*
Output 50×1 (sigmoid)

- Good results,…. despite the pure « image based » architecture (due to mel-spectrogram ?)
- **But can be improved…..**

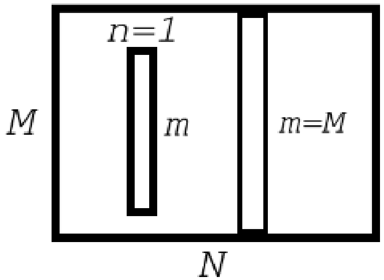
From: K. Choi & al. Automatic tagging usingdeep convolutional neural networks. InProc. of ISMIR (International Society for Music Information Retrieval), New York, USA, 2016.

An interesting idea: designing musically motivated convolutional neural networks

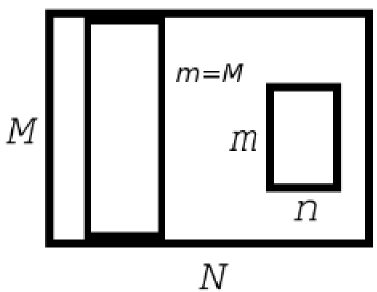
- **Using specific filters**
  - **Temporal features**
    - Filters can learn musical concepts at different time-scales
      - Onsets, attack-sustain-release: $n << N$
      - BPM and rhythm patterns: $n < N$



  - **Frequency filters**
    - Timbre + note: $m = M$
    - Timbre: $m < M$



  - **Rectangular filters**
    - Filters can learn different aspects depending on m and n

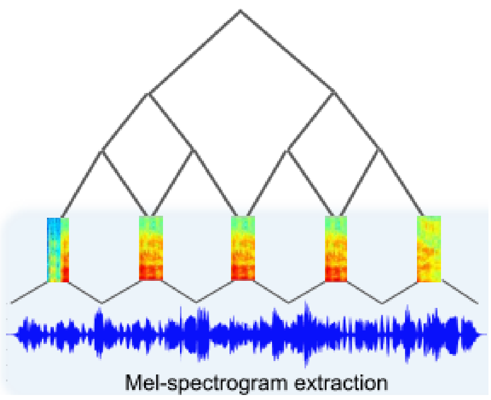


J.Pons & al.Experimenting with musically motivated convolutional neural networks. InProc. of IEEE CBMI, 2016

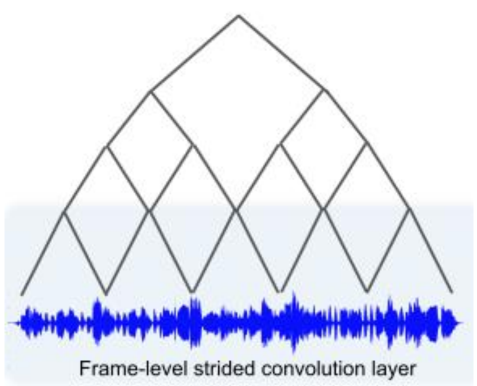
Using different input representations

- **Time domain waveform (end-to-end approaches)**

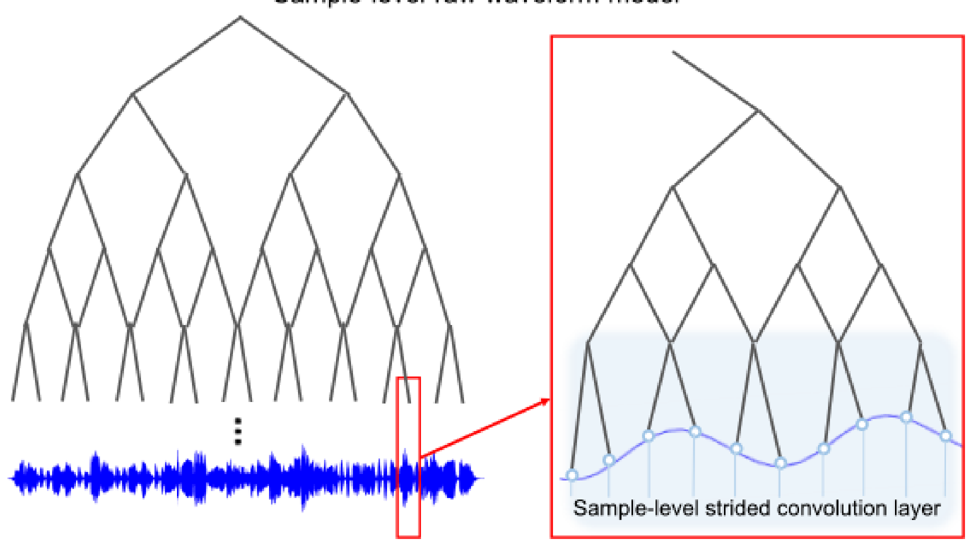
Frame-level mel-spectrogram model



Frame-level raw waveform model



Sample-level raw waveform model

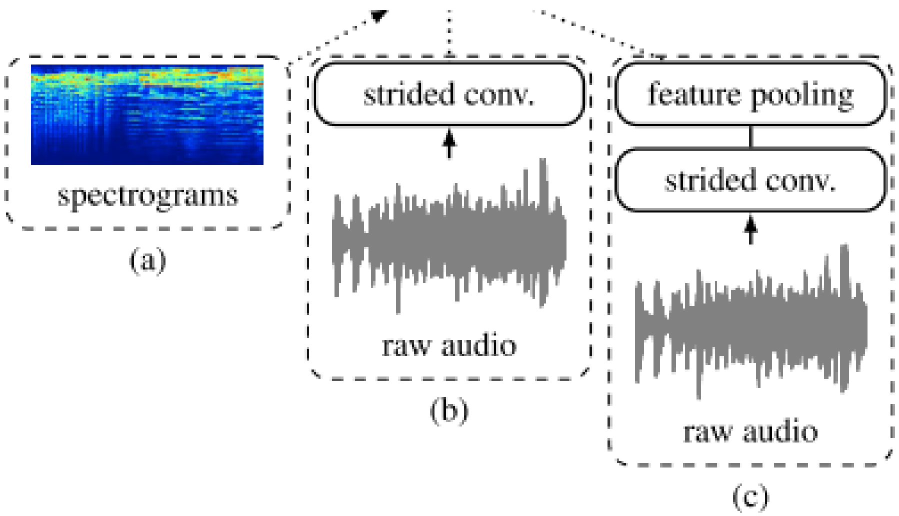


J. Lee & al. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms.arXiv preprint arXiv:1703.01789, 2017.

Popular architectures for Audio

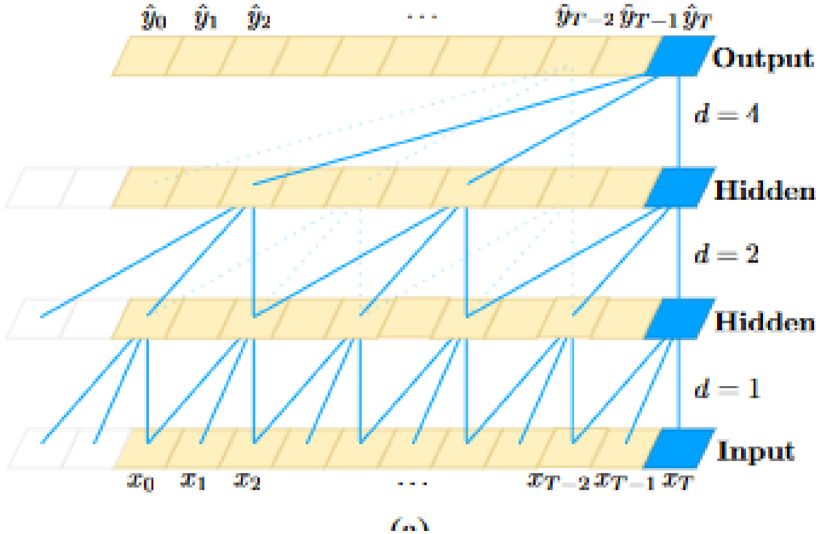
- **Temporal Neural Networks**
  - Main concept for tractable complexity: Dilated convolutions

Input to network



Strided convolutions

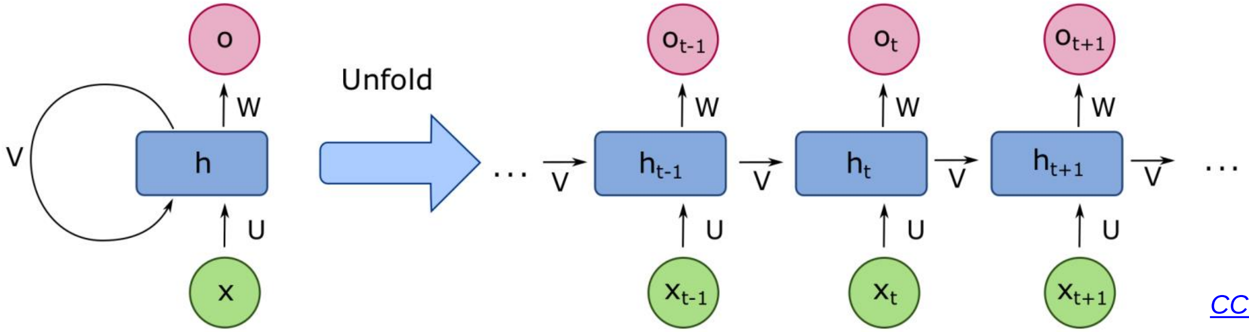
$$(x \circledast_d w)(n) = \sum_{i=0}^{l-1} w(i)x(n-(d\cdot i))$$



Popular architectures for Audio

Recurrent Neural Networks (RNN)

- CNN allows representing the spatial correlations of the data, but they do not allow to represent the sequential aspect of the data

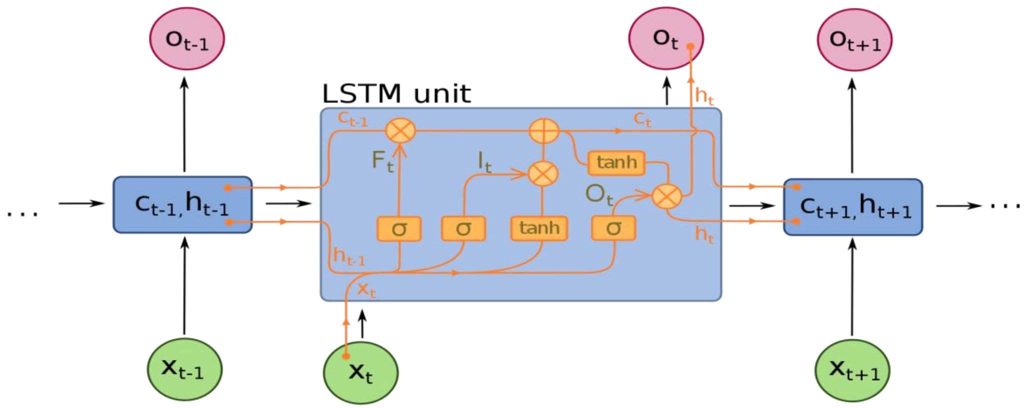


- Theoretically can represent long-term dependencies but suffer from the vanishing gradient problem

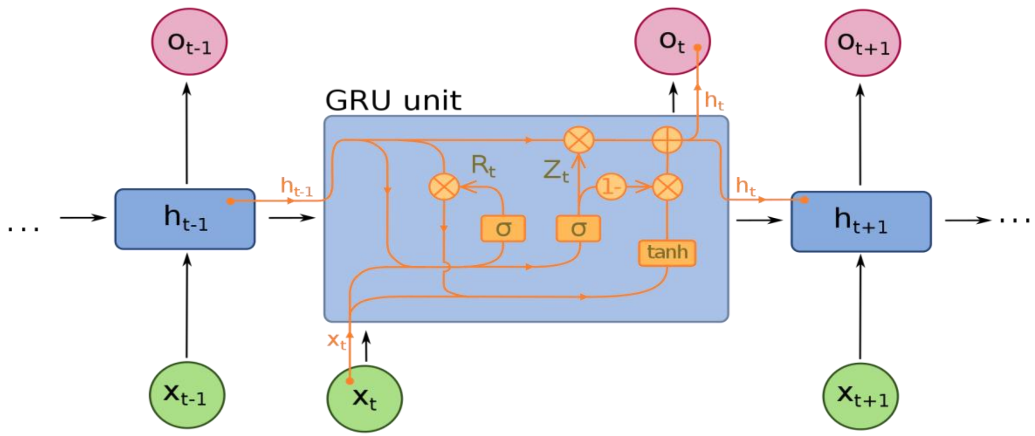
https://en.wikipedia.org/wiki/Recurrent_neural_network

Popular architectures for Audio

- **Recurrent Neural Networks (RNN)**
  - Long-Short-term (LSTM)



  - Gated recurrent unit (fewer parameters)

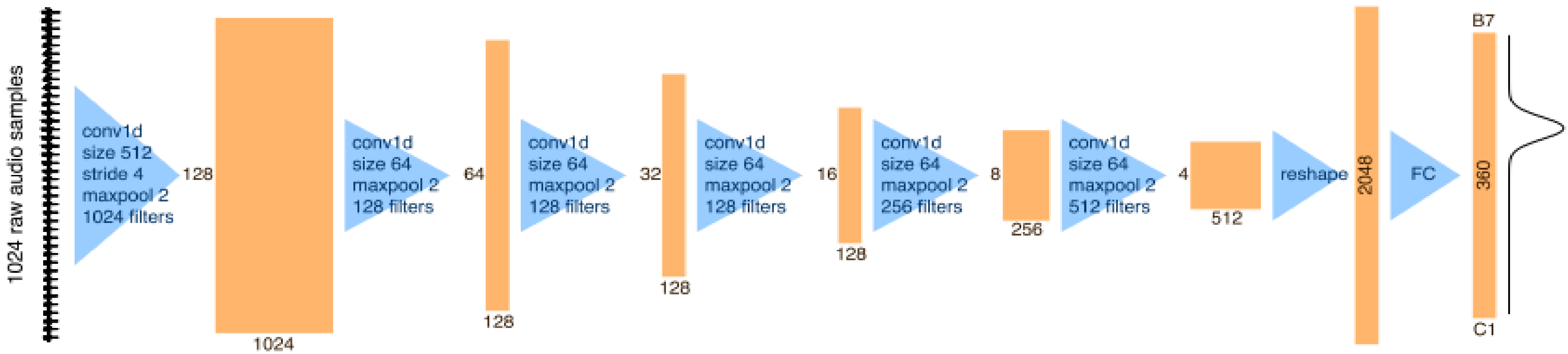


https://en.wikipedia.org/wiki/Recurrent_neural_network

Some examples of pitch estimation with Deep learning

CREPE: A deep learning model for monopitch estimation (1/2)

Exploiting deep learning for pitch estimation



Output:

- 360 nodes (20 cents apart (1/5th of a semitone) from C1 ou B7) $\text{¢}(f) = 1200 \cdot \log_2 \frac{f}{f_{\text{ref}}}$
- Pitch estimate is the weighted mean of the output: $$\hat{\text{¢}} = \frac{\sum_{i=1}^{360} \hat{y}_i \text{¢}_i}{\sum_{i=1}^{360} \hat{y}_i},$$
- Trained with binary cross entropy loss

$$\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{360} (-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)) \qquad y, \hat{y} \in \mathbb{R}_{[0-1]}$$

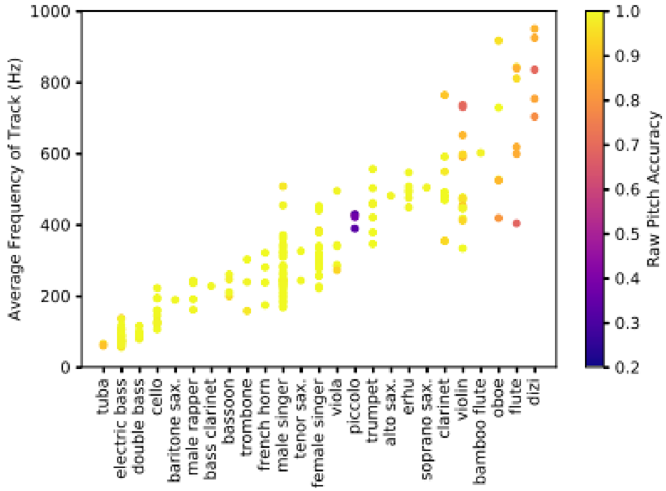
Kim, Jong Wook et al. "Crepe: A Convolutional Representation for Pitch Estimation." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018): 161-165.

CREPE: A deep learning model for monopitch estimation (2/2)

- **A few results**

Dataset	Threshold	CREPE	pYIN	SWIPE
RWC-synth	50 cents	**0.999±0.002**	0.990±0.006	0.963±0.023
	25 cents	**0.999±0.003**	0.972±0.012	0.949±0.026
	10 cents	**0.995±0.004**	0.908±0.032	0.833±0.055
MDB-stem-synth	50 cents	**0.967±0.091**	0.919±0.129	0.925±0.116
	25 cents	**0.953±0.103**	0.890±0.134	0.897±0.127
	10 cents	**0.909±0.126**	0.826±0.150	0.816±0.165

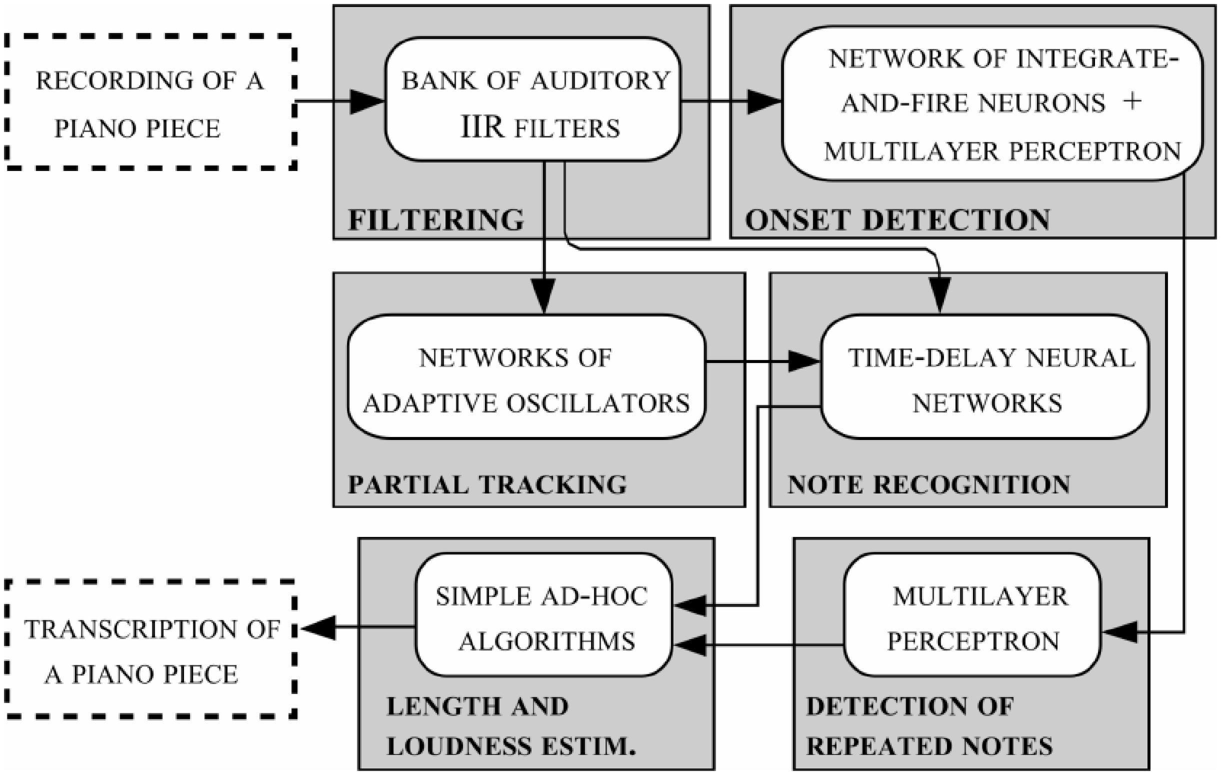
- **Better performances for low frequencies***



*: some errors due to small Numbers of sound exemples for some instruments

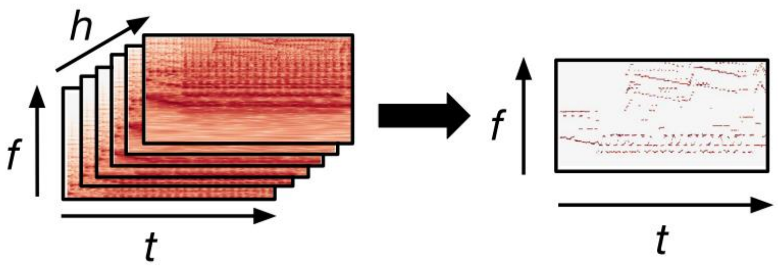
Multipitch estimation using neural networks

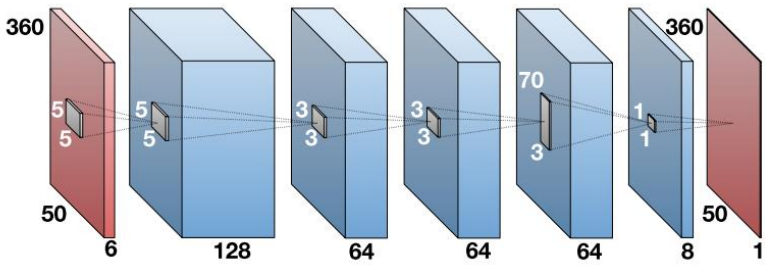
- An early example by M. Marolt (2004) for piano sounds



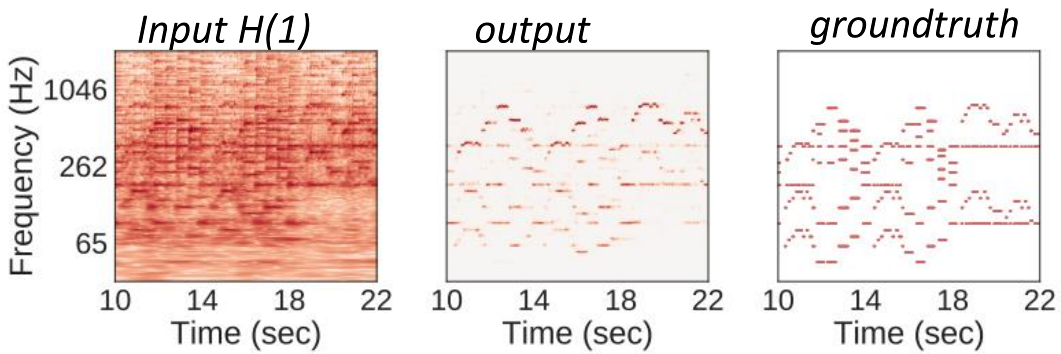
Marolt, Matija. (2004). A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music. Multimedia, IEEE Transactions on. 6. 439 - 449. 10.1109/TMM.2004.827507.

Multipitch estimation using neural networks



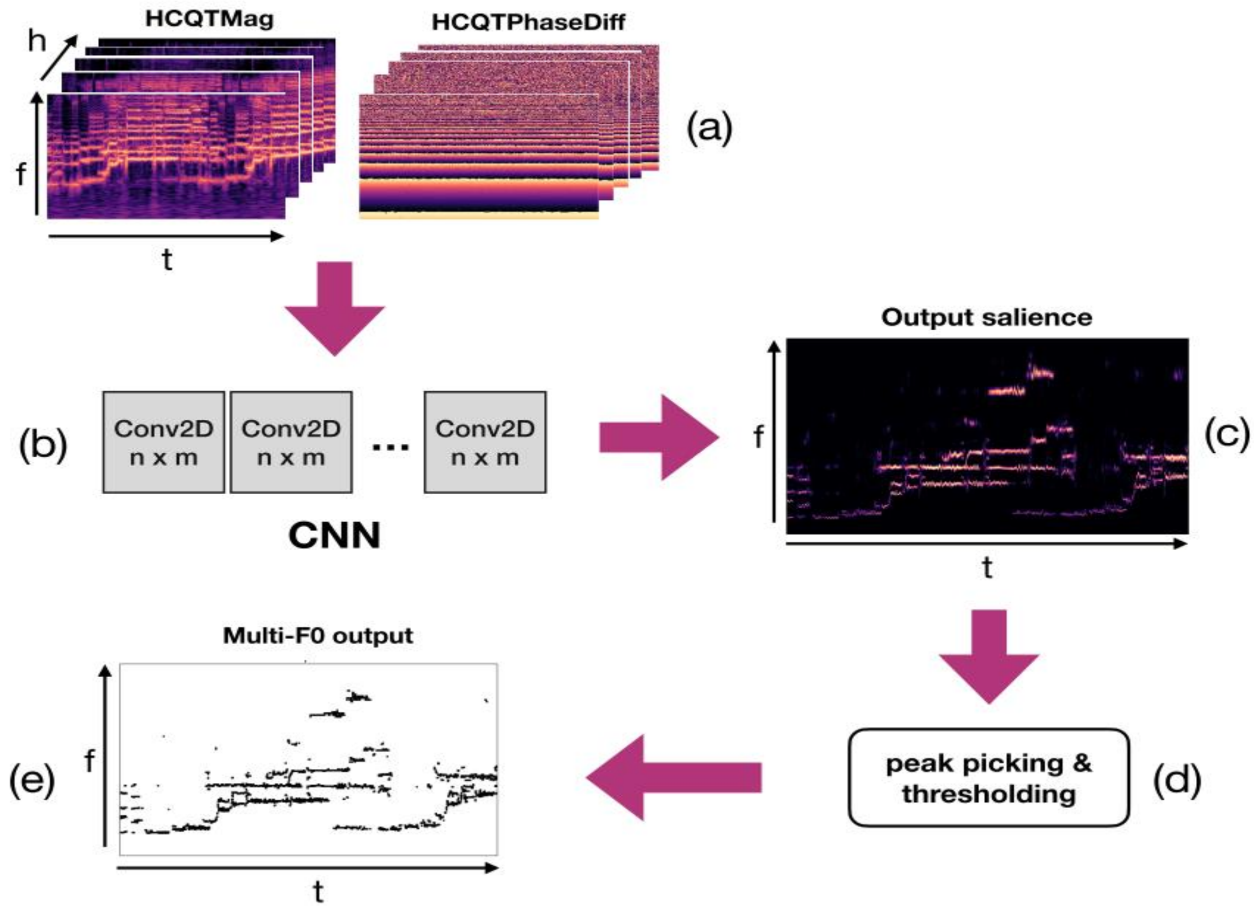


- Use of a specific input representation: the harmonic-CQT $f_k = h \cdot f_{\min} \cdot 2^{k/B}$
- CNN architecture with Relu ; Last layer with sigmoid
- The predicted saliency map can be interpreted as a likelihood score of each time-frequency bin belonging to an f0 contour.



Bittner, Rachel & McFee, Brian & Salamon, Justin & Li, Peter & Bello, Juan. (2017). Deep Salience Representations for f0 Estimation in Polyphonic Music. In proc ISMIR 2017

An extension with focus on singing voices

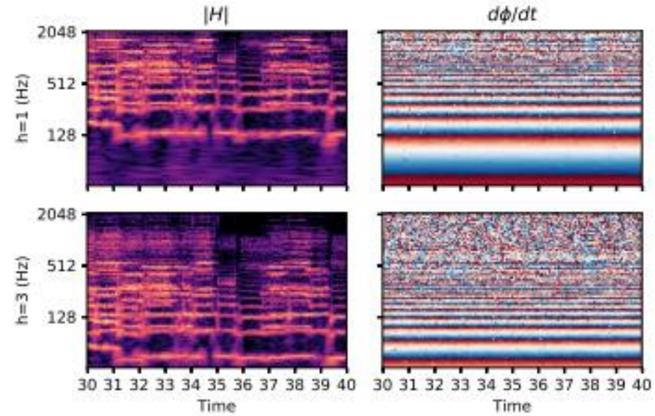


H. Cuesta, B. McFee, and E. Gomez, “Multiple f0 estimation in vocal ensembles using convolutional neural networks,” in Proc. ISMIR, 2020,

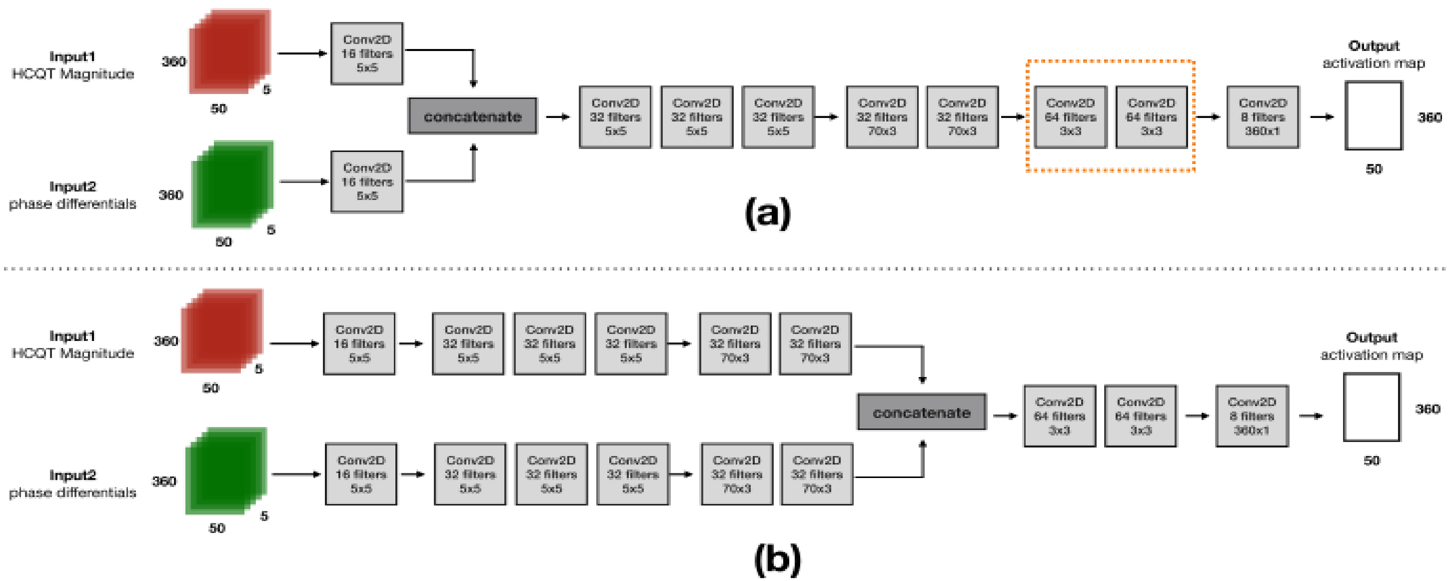
An extension focus on singing voices

- Extended input features with HCQT Phase (phase is directly linked to Instantaneous frequency)

$$\omega_{ins} = \frac{\delta \phi(t)}{\delta t} \rightarrow f_{ins} = \frac{1}{2\pi} \frac{\delta \phi(t)}{\delta t}$$

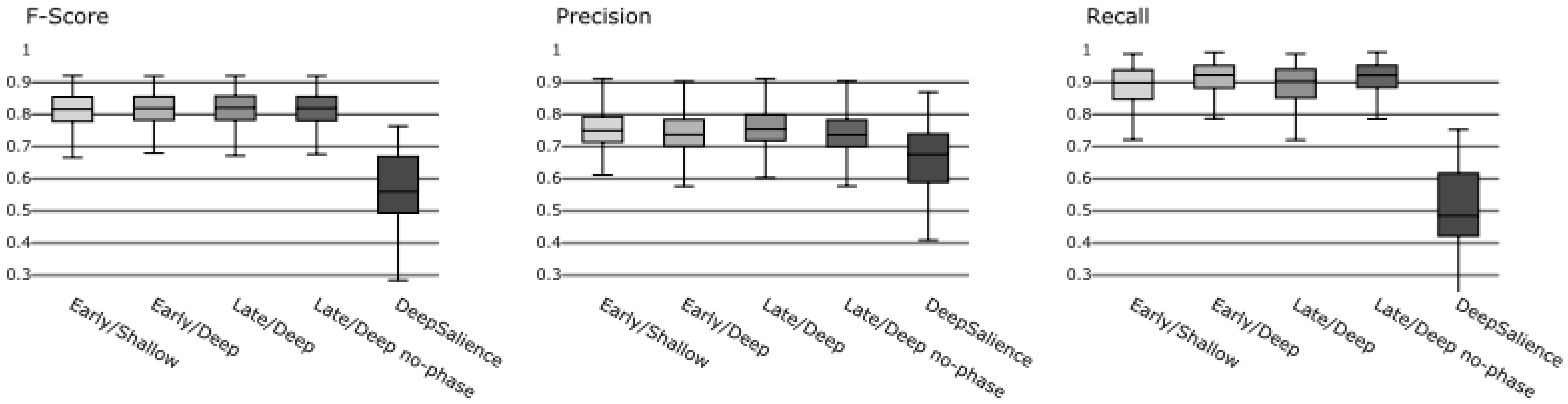


- New architectures (with fusion of input)



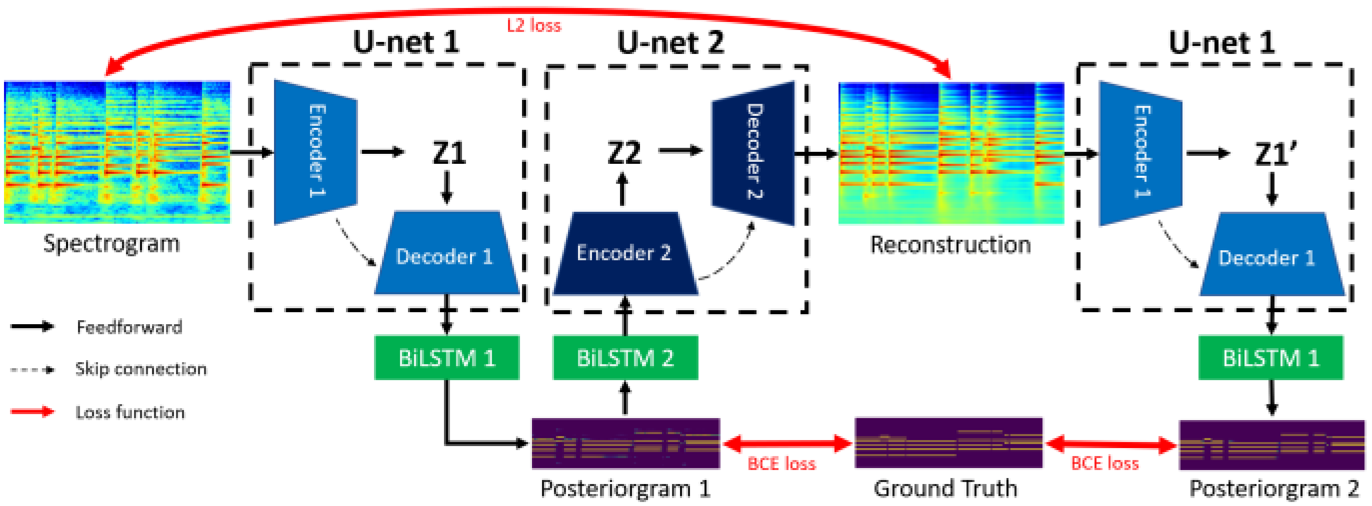
An extension with focus on singing voices

- **An idea of the performances (test sets > 3000 audio files)**



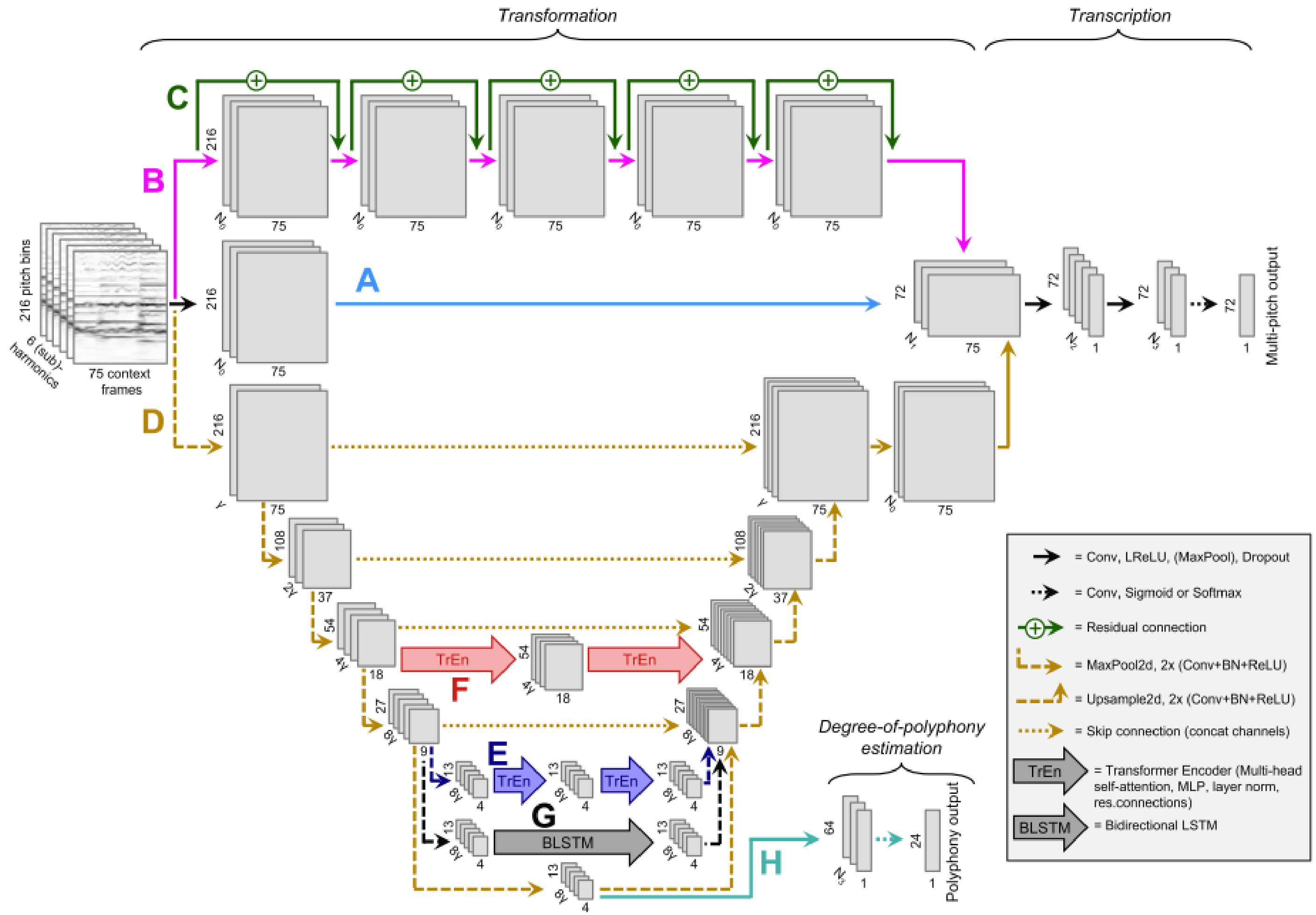
Multipith estimation using Unets (with spectrogram reconstruction)

- **Intuition: we mimic the human behaviour when evaluating a transcription:**
  - We « listen » to the transcription
  - We optimise the algorithm to reduce the errors



Cheuk, Kin Wai et al. "The Effect of Spectrogram Reconstruction on Automatic Music Transcription: An Alternative Approach to Improve Transcription Accuracy." 2020 25th International Conference on Pattern Recognition (ICPR) (2020): 9091-9098.

U-net architectures for multipitch estimation



C. Weiß and G. Peeters, "Comparing Deep Models and Evaluation Strategies for Multi-Pitch Estimation in Music Recordings," in *IEEE/ACM Trans. On AASP*, vol. 30, pp. 2814-2827, 2022, doi: 10.1109/TASLP.2022.3200547

Multipitch estimation using neural networks: other neural approaches

- Deep spiking networks [5]
- Multi-resolution spectrogram as input with LSTM networks [4]
- Use of a kind of “language model” in Neural Autoregressive Distribution Estimator, also known as NADE (*similar to wavenet architecture*) [3]
- A succession of 2 bi-LSTM networks (for note onset detection and note duration estimation), in [2]
- Unet networks (with self-attention [6], spectrogram reconstruction [7], varied architectures [8])
- Unsupervised approaches (but here only for monopitch estimation) [9]
- An interesting reading: [1]

« Yet, despite these [...] limitations, NMF-based methods remain competitive or even exceed the results achieved using NNs.”

[1] E. Benetos, S. Dixon, Z. Duan and S. Ewert, "Automatic Music Transcription: An Overview," in *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20-30, Jan. 2019, doi: 10.1109/MSP.2018.2869928.

[2] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. S. C. Raffel, J. Engel, S. Oore, and D. Eck, “Onsets and frames: Dual-objective piano transcription,” in Proc. Int. Society Music Information Retrieval Conf., 2018, pp. 50–57.

[3] S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 24, no. 5, pp. 927–939, 2016.

[4] S. Böck and M. Schedl, “Polyphonic piano note transcription with recurrent neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 2012, pp. 121–124.

[5] Qian, Hanxiao et al. “Robust Multipitch Estimation of Piano Sounds Using Deep Spiking Neural Networks.” 2019 IEEE Symposium Series on Computational Intelligence (SSCI) (2019): 2335-2341.

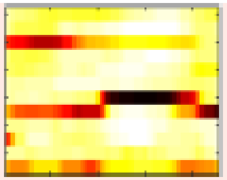
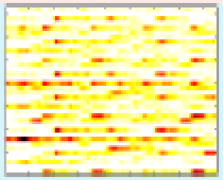
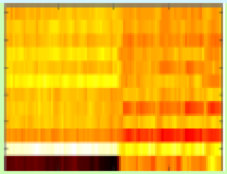
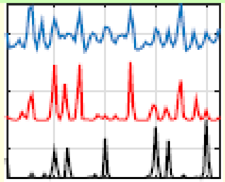
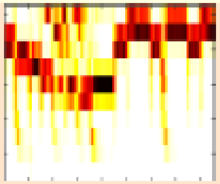
[6]Y. -T. Wu, B. Chen and L. Su, "Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2796-2809, 2020, doi:

[8] C. Weiß and G. Peeters, "Comparing Deep Models and Evaluation Strategies for Multi-Pitch Estimation in Music Recordings," in *IEEE/ACM Trans. On AASP*, vol. 30, pp. 2814-2827, 2022, doi: 10.1109/TASLP.2022.3200547.

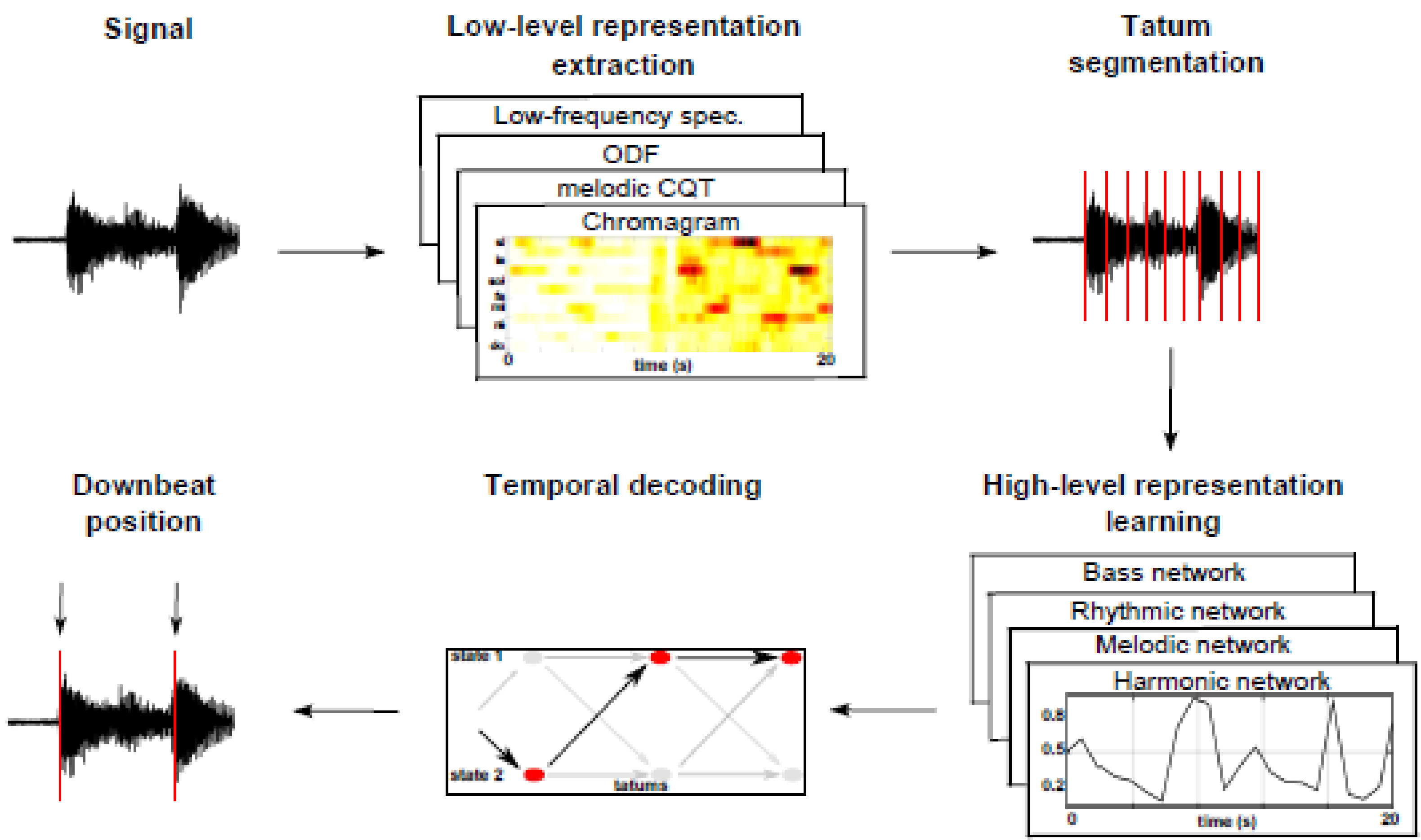
[9] A. Riou, B. Torres, B. Hayes, S. Lattner, G. Hadjeres, et al.. PESTO: Real-Time Pitch Estimation with Self-Supervised Transposition-Equivariant Objective. Transactions of the International Society for Music Information Retrieval (TISMIR), 2025, 8 (1),

An example in Downbeat estimation

Downbeat estimation
(Durand & al. 2017)

Cue	Examples	Input
Harmony	Chord change, Cadence	
Melody	Melodic pattern, pivot notes	
Timbre	Section change, new instrument	
Rhythm	Bar-length rhythm patterns	
Bass content	Bass, Double bass and kick drum highlight downbeats	

Downbeat estimation
(Durand & al. 2017)



S Durand & al., "Robust Downbeat Tracking Using an Ensemble of Convolutional Networks", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol 25, N°1, 2017

Downbeat estimation: démo

- **Examples at the output of each network**
  - https://simondurand.github.io/dnn_audio.html
- **Video example**
  - directory: Démos
- **Other audio example**

JBB (Tatum)

Exemple (Tatum)

JBB (Downbeat)

Exemple (Downbeat)

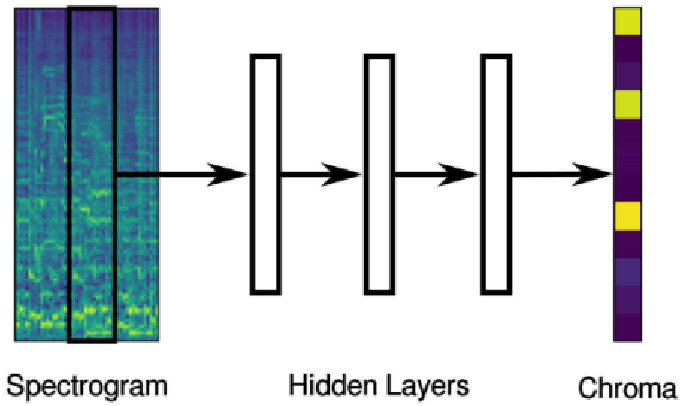
Some examples in Chords recognition

Slides from G. Peeters

Automatic Chords recognition with deep learning (1)

– **Goal**:
  - standard Chroma extractors = too noisy features
  - replace the Chroma front-end by learned features
    – encode harmonic information important for chord recognition, while being robust to irrelevant interferences
    – train a 3-layers MLP to output a ground-truth chroma representation
    – ground-truth ? Chroma corresponding to the notes of the chord)
    – feeding the network with an audio spectrum with context instead of a single frame as input
  - **Deep Chroma**

– **Evaluation**
  - plug the output to a simple logistic regression to estimate the chord (no post-processing, smoothing)



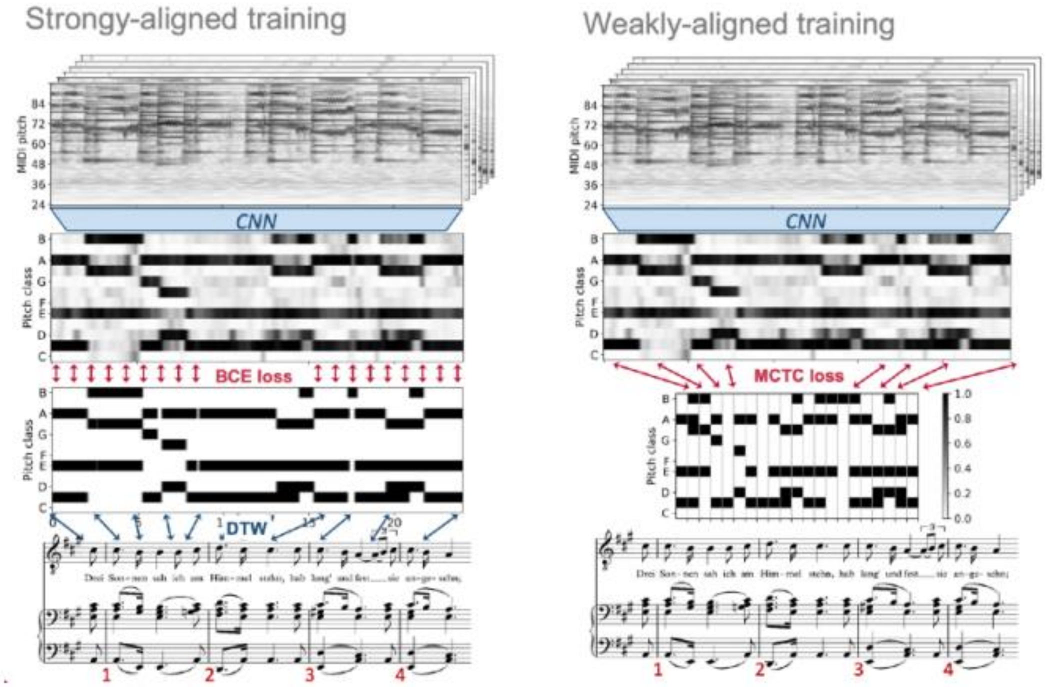
	Btls	Iso	RWC	RW	Total
C	71.0±0.1	69.5 ±0.1	67.4±0.2	71.1±0.1	69.2±0.1
C^W_{Log}	76.0±0.1	74.2 ±0.1	70.3±0.3	74.4±0.2	73.0±0.1
S_{Log}	78.0±0.2	76.5 ±0.2	74.4±0.4	77.8±0.4	76.1±0.2
C_D	**80.2**±0.1	**79.3**±0.1	**77.3**±0.1	**80.1**±0.1	**78.8**±0.1

C: standard chroma from CQT
C^W_{Log}: chromagram with frequency weighting and logarithmic compression
S_{Log}: quarter-tone spectrogram
C_D: deep-chroma

Korzeniowski and Gerhard Widmer. "Feature learning for chord recognition: the deep chroma extractor". In ISMIR, 2016.]

Automatic Chords recognition with deep learning (2)

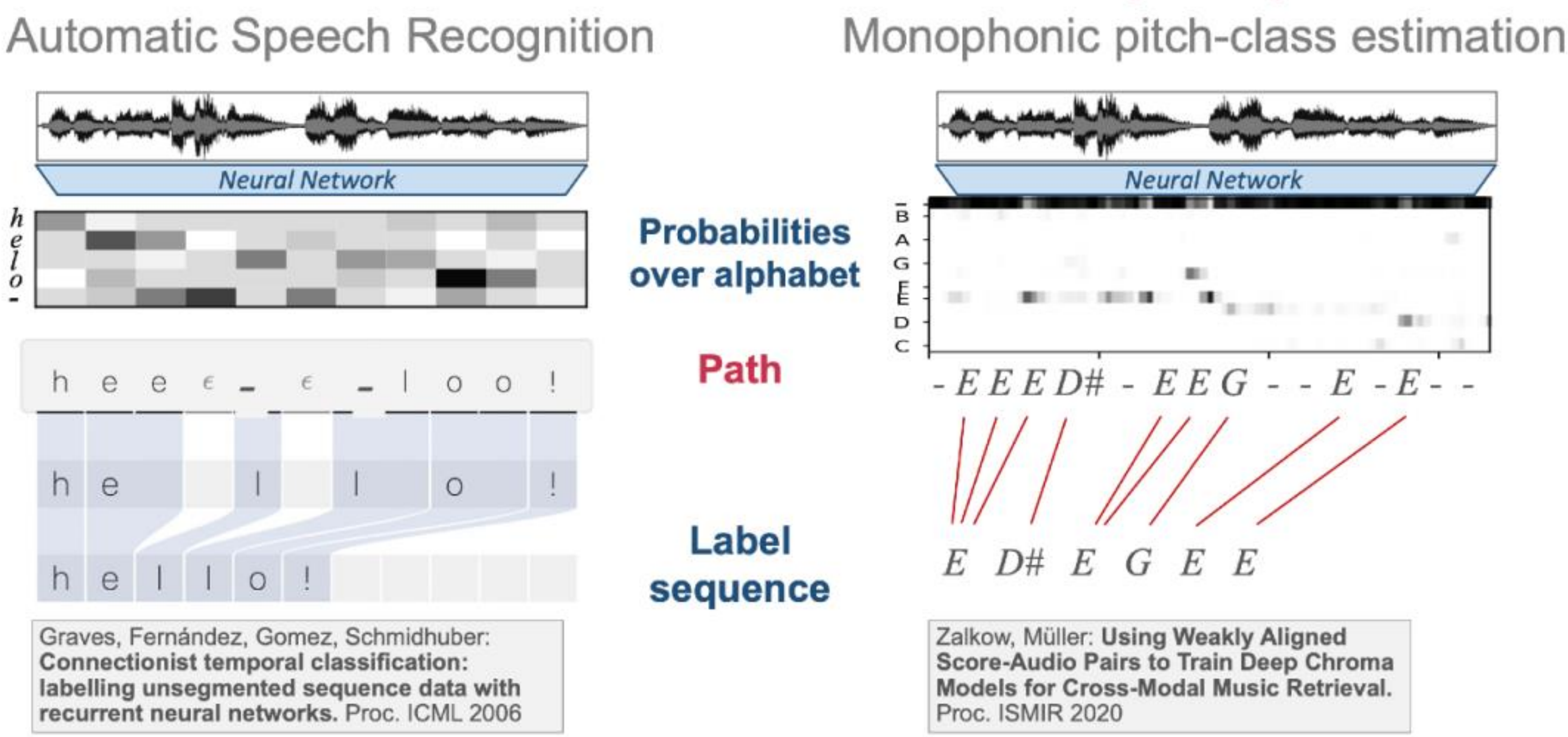
- **Goal:**
  - replace the Chroma/PCP front-end by learned features
  - Ground-truth ?
    - Aligned pitches (costly)
    - Non-aligned pitches (CTC)



[C. Weiß and G. Peeters. "Training deep pitch-class representations with a multi-label CTC loss". In Proc. of ISMIR, 2021]

Automatic Chords recognition with deep learning (2)

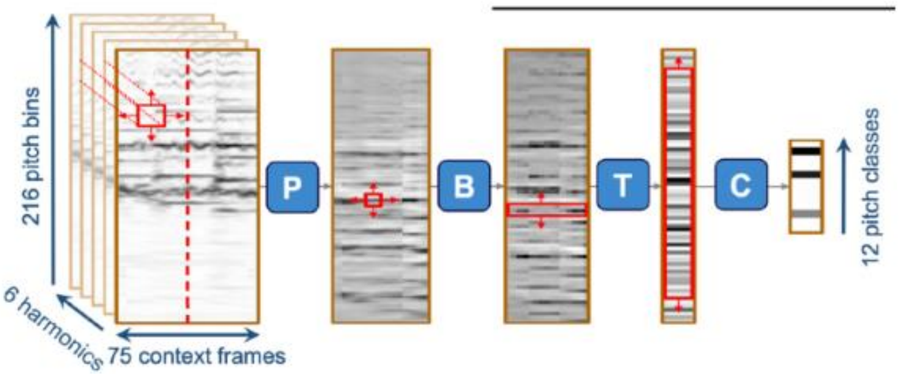
– **Connectionist Temporal Classification (CTC) Loss**



[C. Weiß and G. Peeters. "Training deep pitch-class representations with a multi-label CTC loss". In Proc. of ISMIR, 2021]

Automatic Chords recognition with deep learning (2)

– **CNN Architecture**
  - Input**: Harmonic-CQT**
  - Simple **5-layerCNN**
  - Roughly **48k parameters**
  - **P**re-filtering, **B**inning to midi-pitches (216 →72), **T**emporal reduction (75 →1), **C**hroma reduction (72 →12)
  - **Input**: Harmonic CQT



	Layer	Kernel size	Output shape	# Parameters
	Layer norm.		$(T+74, 216, 6)$	2592
P	Conv2D, MaxPool	15×15	$(T+74, 216, 20)$	27020
B	Conv2D, MaxPool	3×3	$(T+74, 72, 20)$	3620
T	Conv2D	75×1	$(T, 72, 10)$	15010
	Conv2D	1×1	$(T, 72, 1)$	11
C	Conv2D	1×61	$(T, 12+P, Q)$	$Q(62+73 \cdot P)$
	Total			48253 $+Q(62+73 \cdot P)$

[C. Weiß and G. Peeters. "Training deep pitch-class representations with a multi-label CTC loss". In Proc. of ISMIR, 2021]

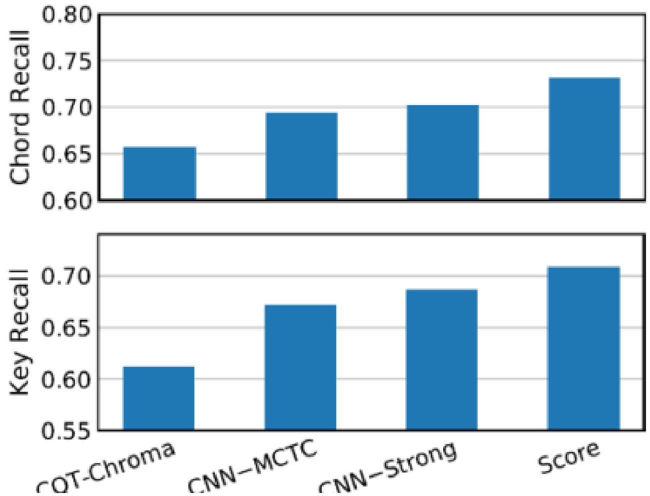
Automatic Chords recognition with deep learning (2)

– Evaluation

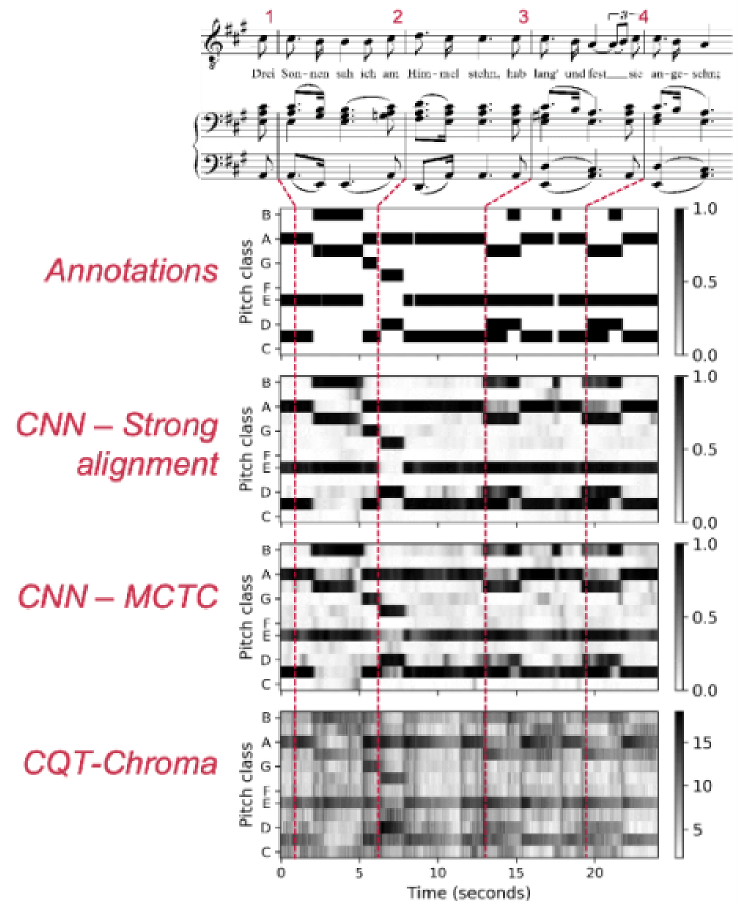
- Cosine similarity (CS), Average precision (AP)

Model/Loss	P	R	F	CS	AP
All-Zero	0	0	0	0.486	0.211
CQT-Chroma	0.512	0.681	0.579	0.701	0.594
CNN – SCTC	**0.850**	0.048	0.090	0.520	0.416
CNN – MCTC:NE	0.747	0.775	0.758	0.802	0.798
CNN – MCTC:WE	0.762	**0.853**	0.802	0.830	0.851
CNN – Strong alignment	**0.850**	0.790	**0.818**	**0.860**	**0.886**

– **Application**: Chord and Key estimation



Application: Visualization



[C. Weiß and G. Peeters. "Training deep pitch-class representations with a multi-label CTC loss". In Proc. of ISMIR, 2021]

Automatic Chords recognition with deep learning

Another approach

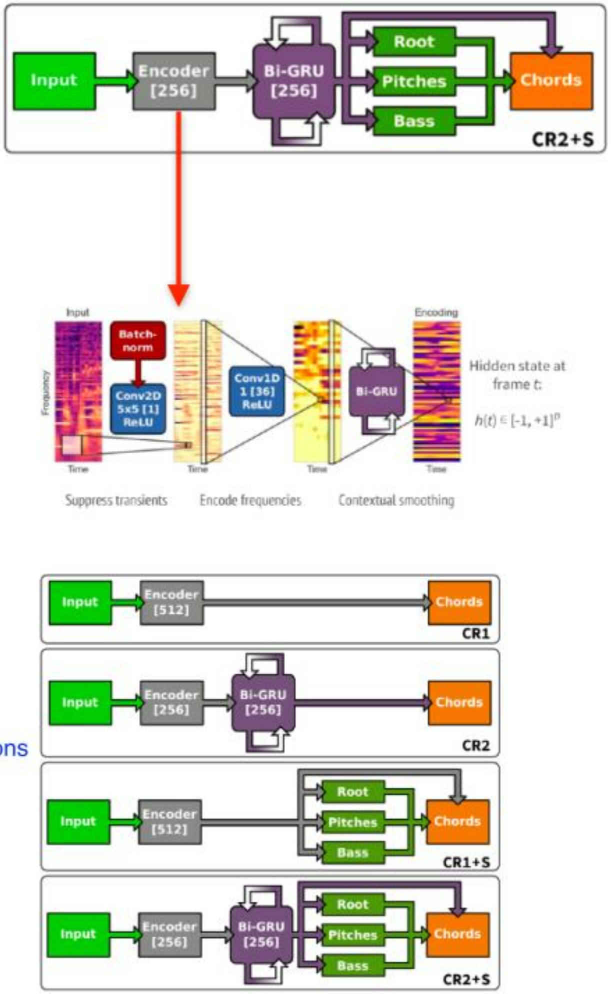
– **Goal 1:**
  - **End-to-end system**

– **Encoder:**
  - Input: T × F time-series of log-power constant-Q transform (CQT) spectra
  - First layer : can be interpreted as a harmonic saliency enhancer, as it tends to learn to suppress transients and vibrato while emphasizing sustained tones.
  - Second layer summarizes the pitch content of each frame, and can be interpreted as a local feature extractor

– **Decoder:**
  - 4 architectures

Encoder details

Decoder configurations

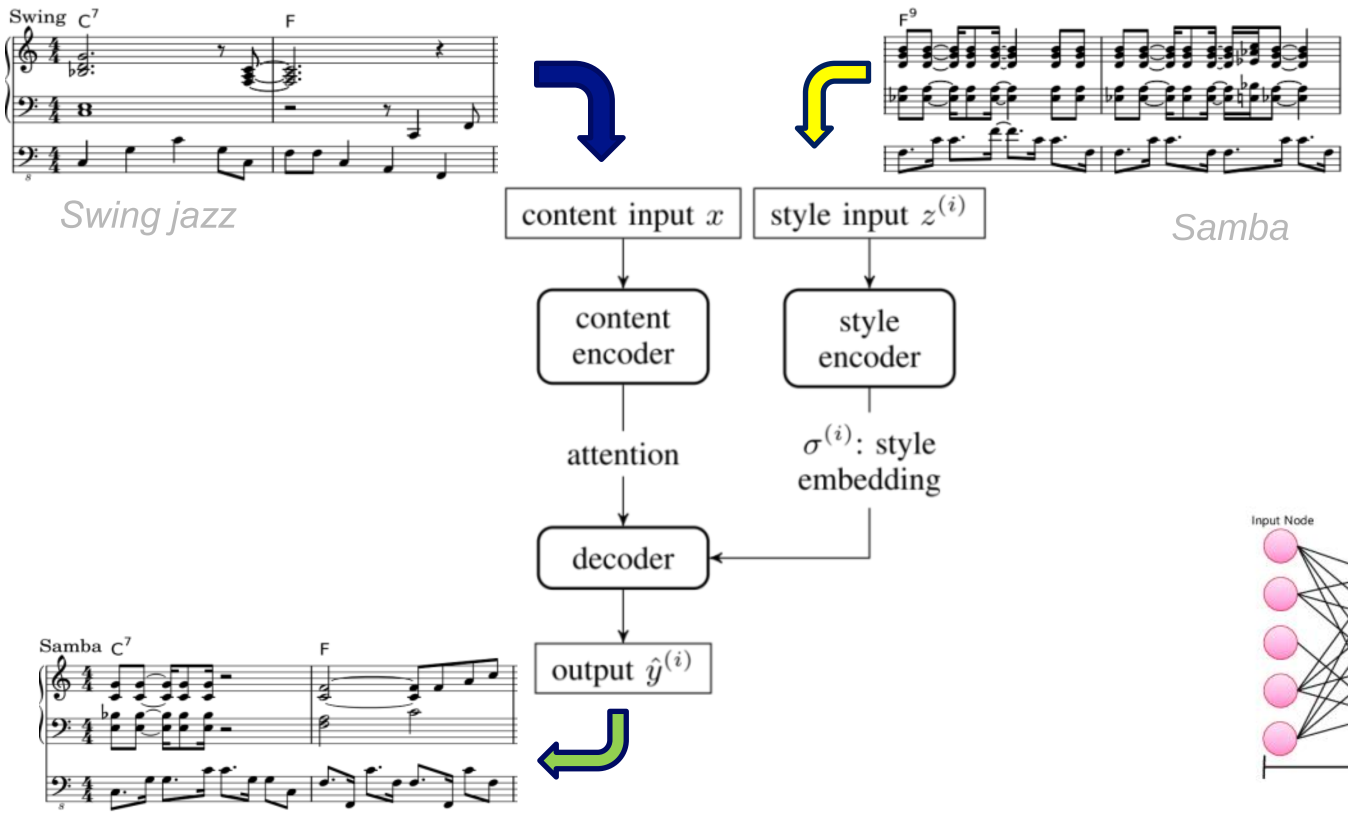


McFee and J. P. Bello. "Structured training for large-vocabulary chord recognition". In Proc. of ISMIR, 2017]

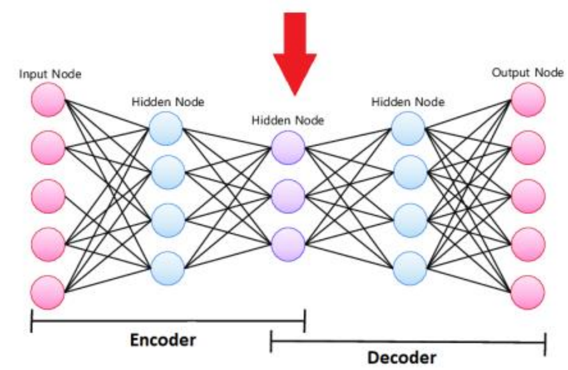
An example in Music style transfer

Symbolic music style transfer

- … Or playing a given music file in the style of another music excerpt.



Signal representation



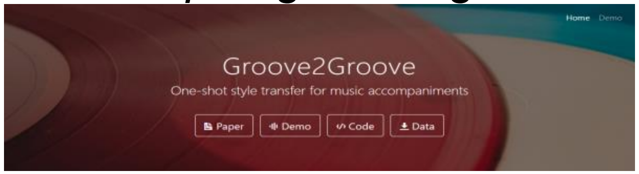
[13] Ondrej Cifka, Umut Simsekli, Gaël Richard, "Groove2Groove: One-Shot Music Style Transfer with Supervision from Synthetic Data", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (preprint) accepted for publication, 2020

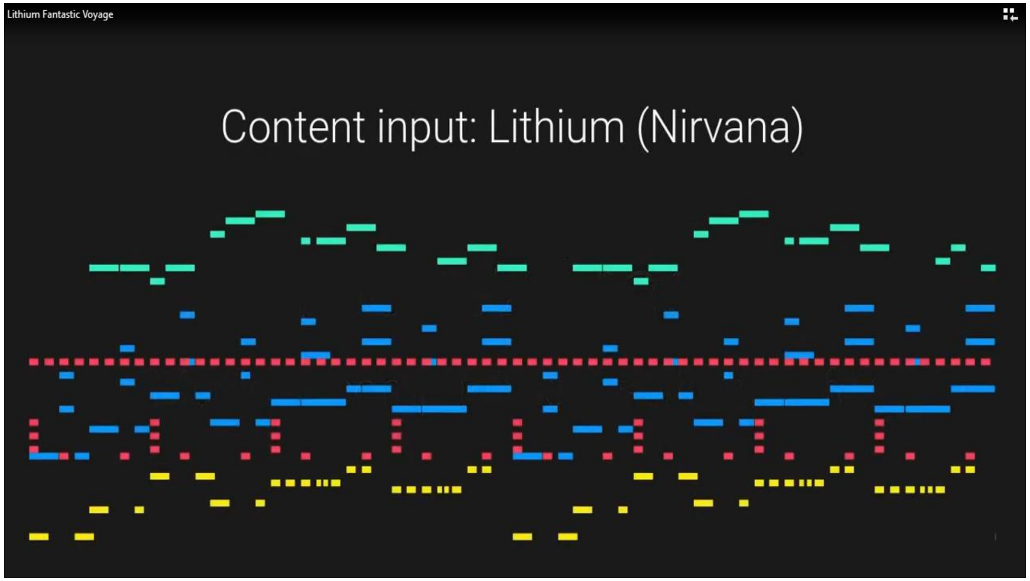
Sound examples at : *https://groove2groove.telecom-paris.fr*

Recognize, Transform, Synthetize …
Symbolic music style transfer

- **A short demo**

(more sound examples at : https://groove2groove.telecom-paris.fr)





[13] Ondrej Cifka, Umut Simsekli, Gaël Richard, "Groove2Groove: One-Shot Music Style Transfer with Supervision from Synthetic Data", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 2020

Sound examples at : *https://groove2groove.telecom-paris.fr*

Numerous « meta-structures »

- **Auto- encoders**
  - Variational Auto-encoders

- **Generative Adversarial Networks (GAN)**
- **Attention models**
- **Transformers**
- **…**

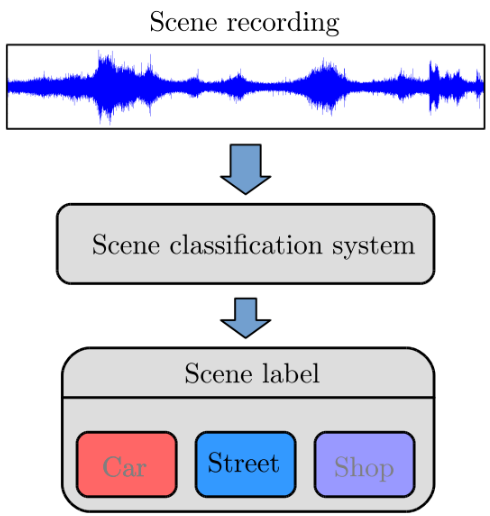
- **For more examples with applications to audio, see**

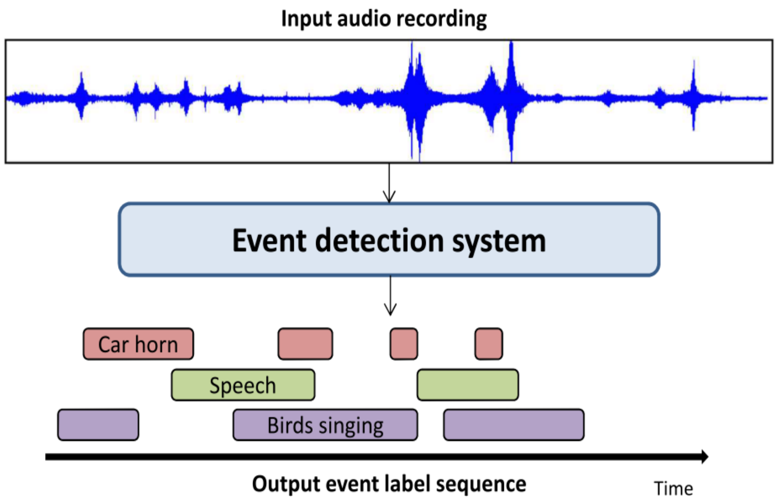
  - *G. Peeters, G. Richard, « Deep learning for audio» , Multi-faceted Deep Learning: Models and Data, Edited by Jenny Benois-Pineau, Akka Zemmari, Springer-Verlag, 2021*

Some examples in Audio scene and event recognition

Audio scene and event recognition

- **Acoustic scene recognition vs Acoustic event recognition**





- **DCASE: an active community (wokshop, challenges, …)**
  - https://dcase.community/

DCASE challenge tasks in 2025

- **DCASE challenges** *(from https://dcase.community/)*

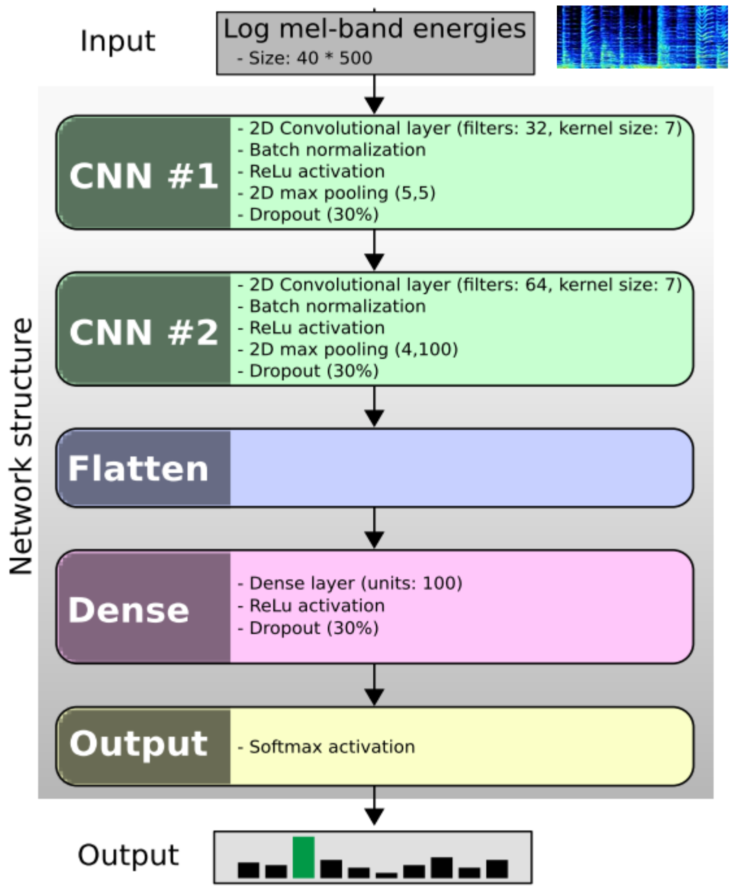
Challenge status

Task	Task description	Development dataset	Baseline system	Evaluation dataset	Results
Task 1, Low-Complexity Acoustic Scene Classification with Device Information	Released	Released	Released	Released	Released
Task 2, First-Shot Unsupervised Anomalous Sound Detection for Machine Condition Monitoring	Released	Released	Released	Released	Released
Task 3, Stereo Sound Event Localization and Detection in Regular Video Content	Released	Released	Released	Released	Released
Task 4, Spatial Semantic Segmentation of Sound Scenes	Released	Released	Released	Released	Released
Task 5, Audio Question Answering	Released	Released	Released	Released	Released
Task 6, Language-Based Audio Retrieval	Released	Released	Released	Released	Released

updated 2025/06/30

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.

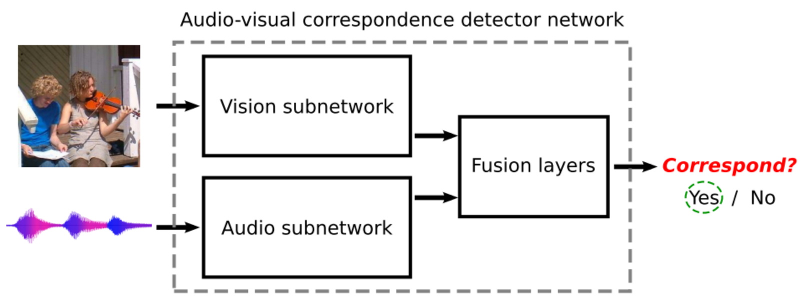
T. Heittola & al. *Acoustic scene classification in dcase 2020 challenge: generalization across devices and low complexity solutions.* In Proc. of the DCASE 2020 Workshop

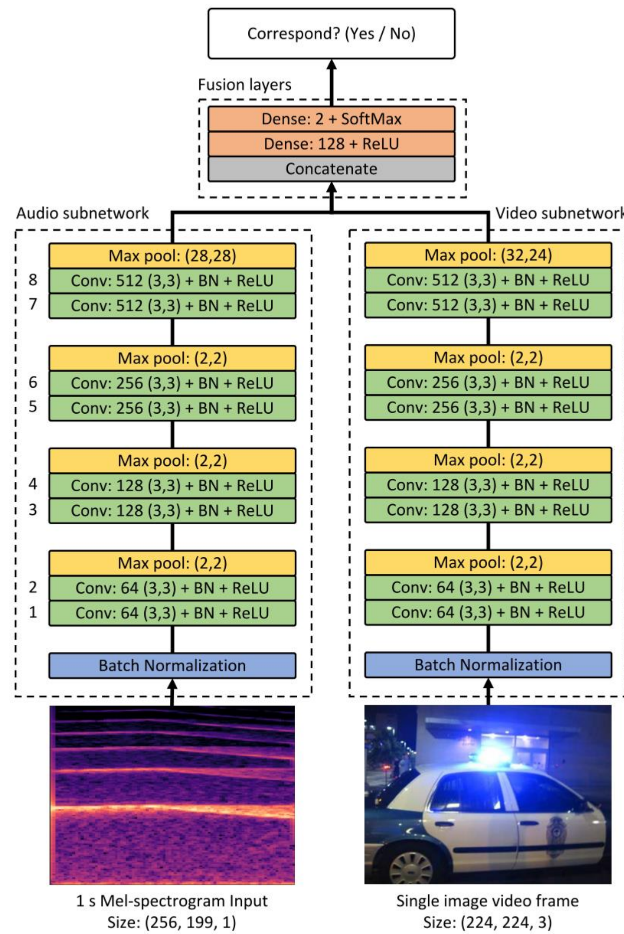
Comparasion with other baselines

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)*



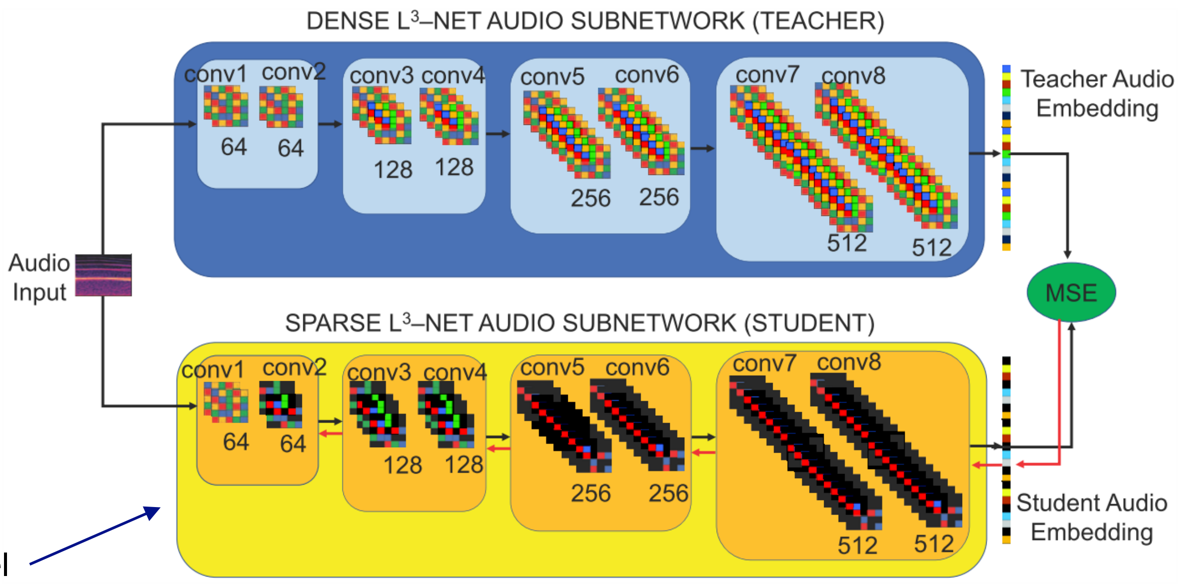


R. Arandjelovi´c and A. Zisserman, *“Look, listen and learn,”* in IEEE ICCV, 2017, pp. 609–617.
S. Kumari, D. Roy, M. Cartwright, J. P. Bello, and A. Arora. *EdgeI^3: compressing I^3-net for mote scale urban noise monitoring.* In 2019 IEEE International Parallel and Distributed Processing Symposium Workshops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

DCASE: Audio Scene classification

Modified DCASE2020 Task 1 Baseline, Subtask A
EdgeL3 + MLP (2 layers, 64 units each)

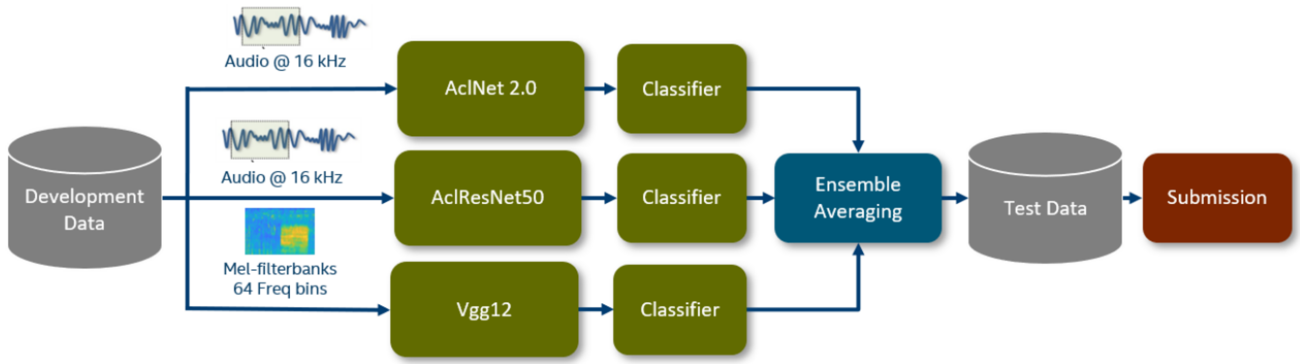
- **Sparsity**
  - Teacher-student
  - Different level of sparsity for each layer



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),

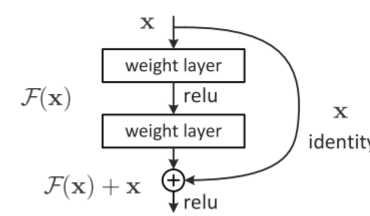
Acoustic scene recognition:
How to improve ?

- **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

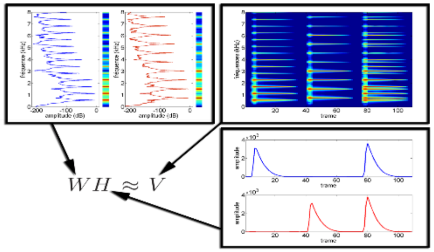


  - Use signal or audio models (NMF, ..)

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

Acoustic scene recognition:
Why using signal or perceptual models

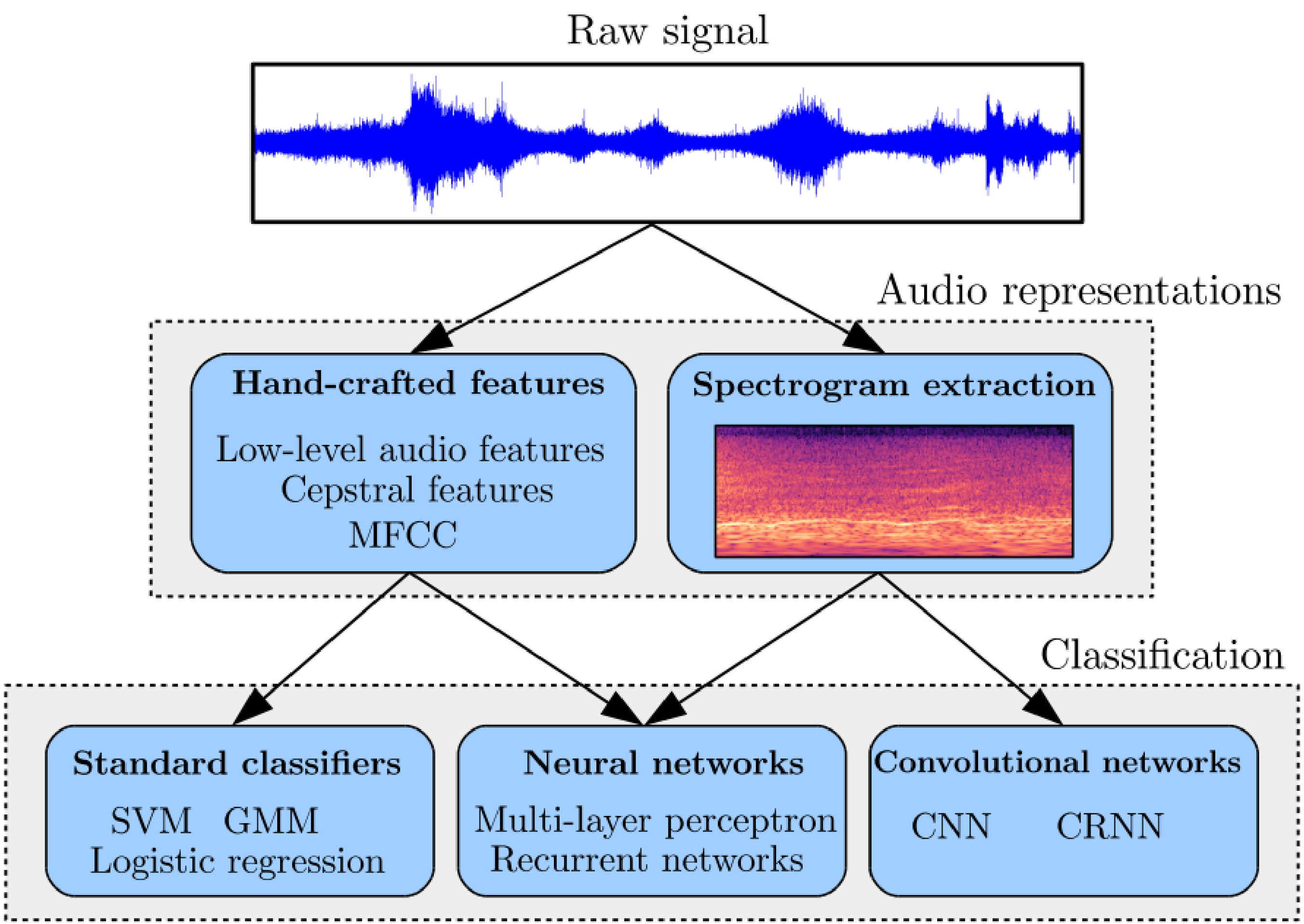
- **Using perceptual models**
  - Example: Mel spectrogram, MFCC, CQT,..
  - The classifier does not learn what is not audible



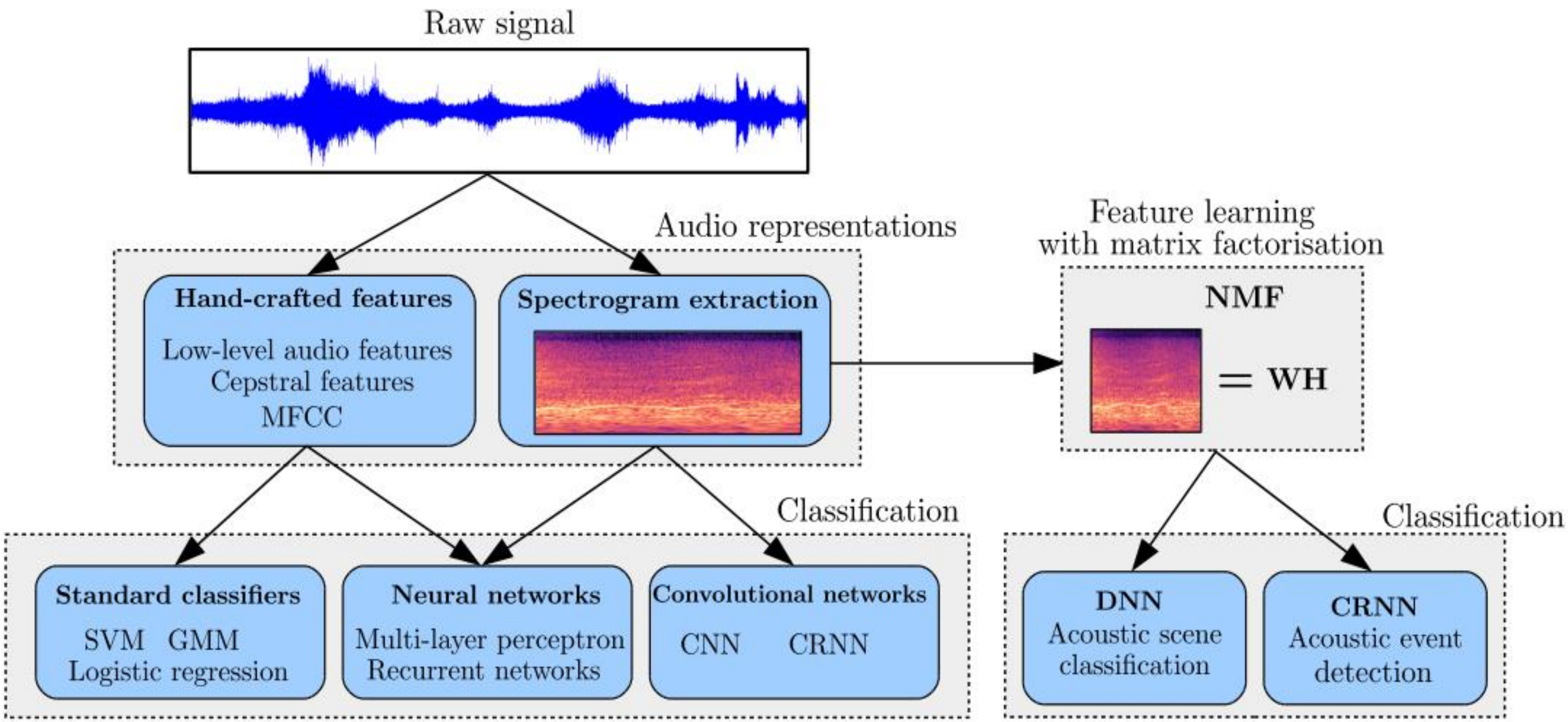
- **Using signal models**
  - Example: Harmonic + noise, Source filter, NMF, …
  - *e.g The classifier does not learn what is not typical of an audio signal*

- **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)

Audio scene and event recognition



Audio scene and event recognition usingNMF features(Bisot & al. 2017)



Why NMF ?

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

- **Principle of NMF :**

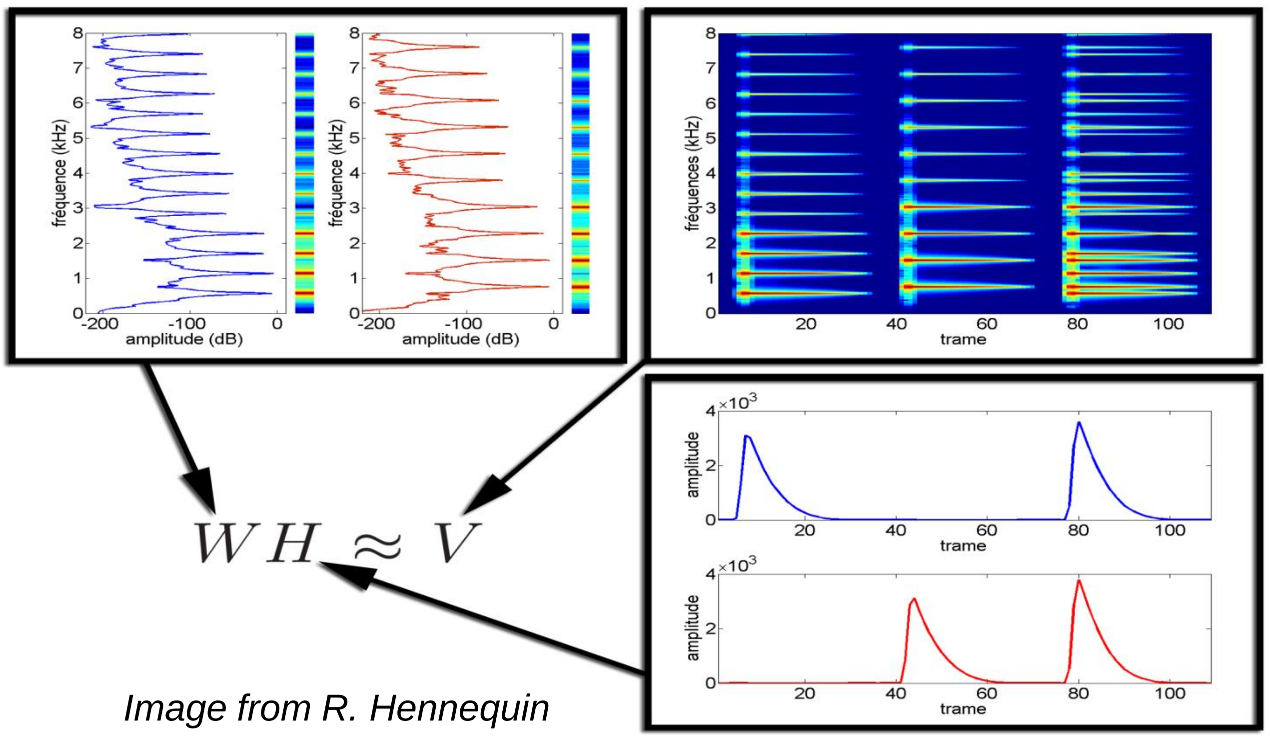
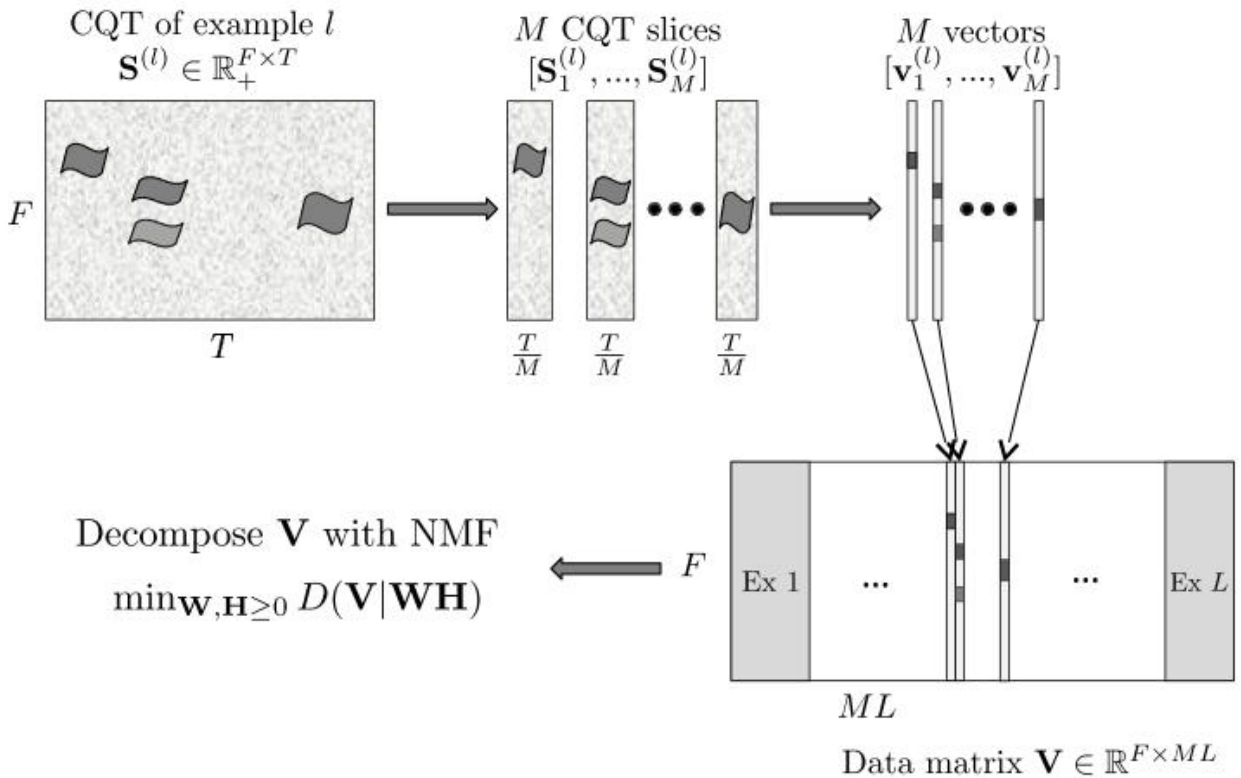


Image from R. Hennequin

Example for scene classification

From time-frequency representations to dictionary learning

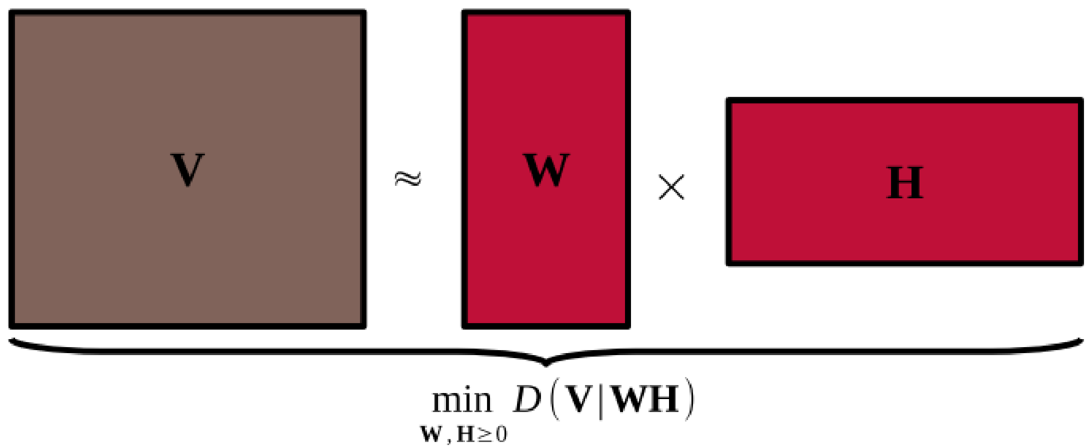


Unsupervised NMF for acoustic scene recognition

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

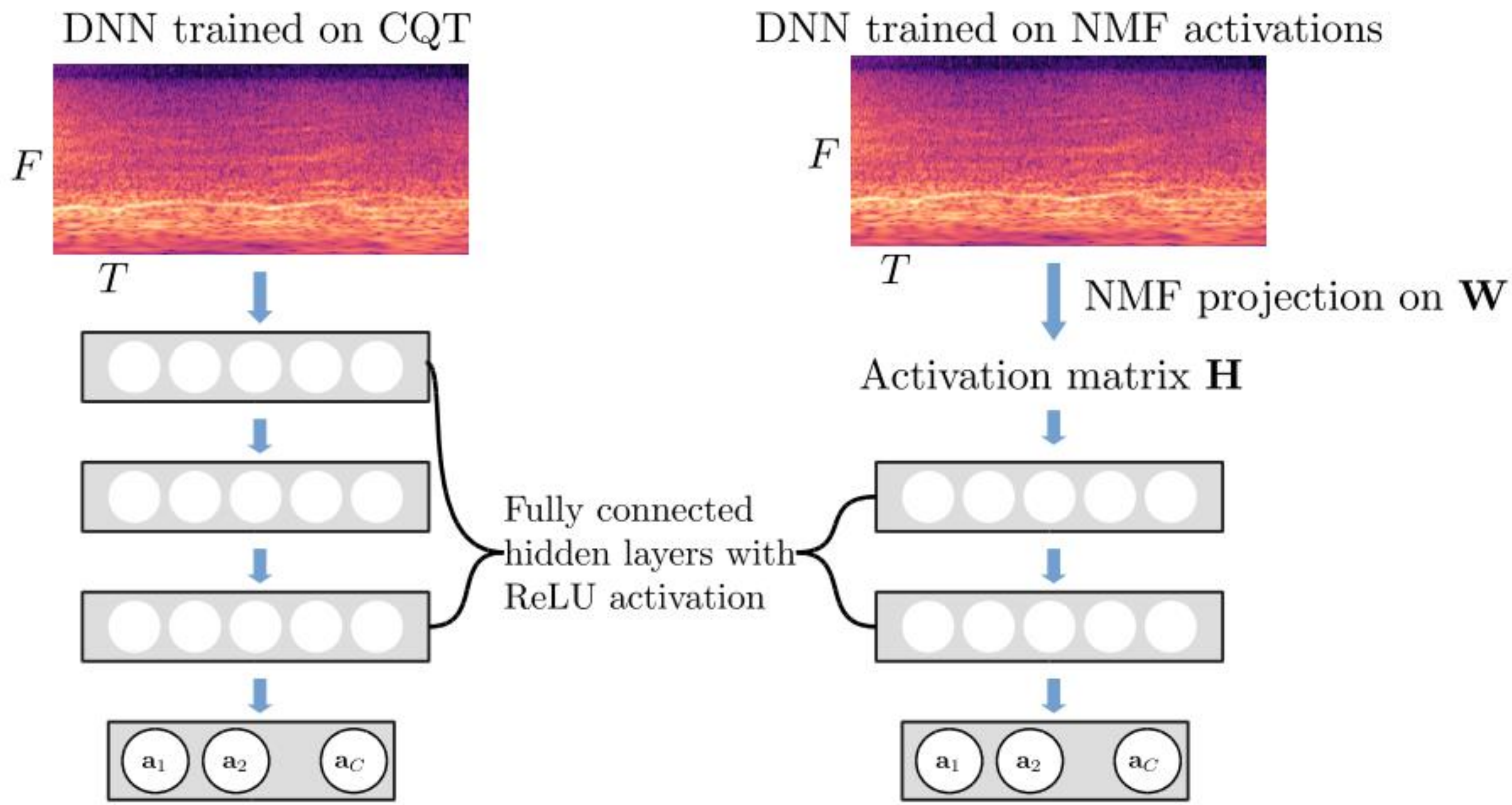
Nonnegative matrix factorization

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

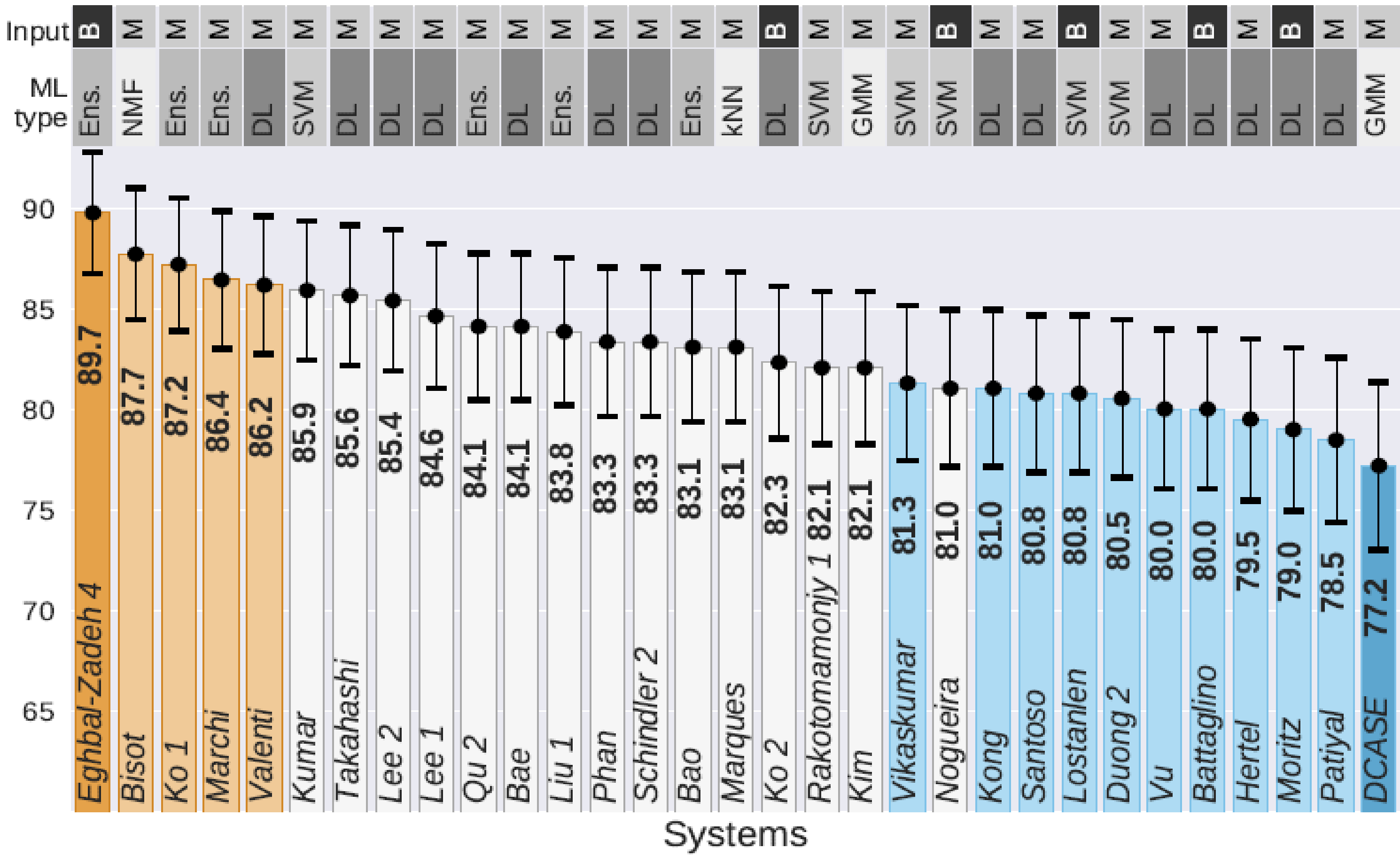
$$\underbrace{\mathbf{V} \approx \mathbf{W} \times \mathbf{H}}_{\min_{\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{W}\mathbf{H})}$$

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*

Summary : Machine listening
Audio scene and event recognition

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

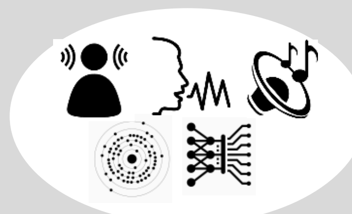
Audio surveillance, Audio scene analysis

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 classififcation / Music signal procesing**
  - *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*
  - 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. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM (2004)

- **Rhythm/tempo estimation**
  - M. Alonso, G. Richard, B. David, "*Accurate tempo estimation based on harmonic+noise decomposition", EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 82795, 14 pages, 2007.
  - Scheirer E., 1998, "*Tempo and Beat Analysis of Acoustic Musical Signals*", Journal of the Acoustical Society of America (1998), Vol. 103, No. 1, pp. 588-601. 50
  - Laroche, 2001] J. Laroche. Estimating Tempo, Swing, and Beat Locations in Audio Recordings. Dans Proc. of WASPAA'01, New York, NY, USA, octobre 2001
  - S Durand, J. Bello, S. Leglaive, B. David, G. Richard, "Robust Downbeat Tracking Using an Ensemble of Convolutional Networks", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol 25, N°1, 2017

- **Music instrument recognition**
  - S. Essid, G. Richard, B. David. *Instrument recognition in polyphonic music based on automatic taxonomies*. IEEE Trans. on Audio, Speech, and Language Proc. 14 (2006), no. 1
  - Eronen-09]A. Eronen, "Signal processing method for audio classification and music content analysis," Ph.D. dissertation, Tampere University of Technology, Finland, June 2009.
  - S. Essid, G. Richard, B. David. *Musical Instrument recognition by pairwise classification strategies*. IEEE Trans. on Audio, Speech and Language Proc. 14 (2006), no. 4
  - [Barbedo-11] J. Barbedo and G. Tzanetakis, "Musical instrument classification using individual partials," *IEEE Trans. Audio, Speech and language Processing, 19(1), 2011.*
  - [Leveau-08]: 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.
  - [Kitahara-07] T. Kitahara, "Computational musical instrument recognition and its application to content-based music information retrieval," Ph.D. dissertation,

A few references…

- **Chord Estimation,**
  - L. Oudre. *Template-based chord recognition from audio signals.* PhD thesis, TELECOM ParisTech, 2010.

- **Multipitch estimation**
  - A. Klapuri, “Multiple fundamental frequency estimation based on harmonicity and spectral smoothness,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 11, no. 6, pp. 804–816, 2003.
  - V. Emiya, PhD thesis. Telecom ParisTech.

- **Perception**
  - [Alluri-10] V. Alluri and P. Toiviainen, “Exploring perceptual and acoustical correlates of polyphonic timbre,” *Music Perception, vol. 27, no. 3, pp.* 223–241, 2010.
  - [Kendall-91] R. A. Kendall and E. C. Carterette, “Perceptual scaling of simultaneous wind instrument timbres,” *Music Perception, vol. 8, no. 4, pp. 369–404,* 1991.
  - [McAdams-95] McAdams, S., Winsberg, S., Donnadieu, S., DeSoete, G., and Krimphoff, J. “Perceptual Scaling of synthesized musical timbres: Common dimensions, specificities and latent subject classes,” *Psychological Research,* 1995.
  - Schouten’s [1968] J. F. Schouten, “The perception of timbre,” in *6th International Congress on Acoustics, Tokyo, Japan, 1968,*

- ***Source separation***
  - O. Gillet, G. Richard. *Transcription and separation of drum signals from polyphonic music*. IEEE Trans. on Audio, Speech and Language Proc. (2008)
  - M. Ryyn¨anen and A. Klapuri, “Automatic bass line transcription from streaming polyphonic audio,” in IEEE International
  - Conference on Acoustics, Speech and Signal Processing (ICASSP), Hawaii, USA, 2007.
  - S. Leglaive, R. Badeau, G. Richard, "Multichannel Audio Source Separation with Probabilistic Reverberation Priors", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, no. 12, December 2016
  - J-L Durrieu, B. David, G. Richard, A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, *October 2011.*

- ***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