Analyse, transformation et reconnaissance des signaux sonores

« Analysis, transformation and recognition of audio signals »

Gaël RICHARD
Professeur à Télécom Paris

Cérémonie Prix IMT-Académie des sciences, 1er décembre 2020

A research done in collaboration….

- **An immense thanks to all ….**

Analysis, transformation, recognition ... of audio signals

An audio signal

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_N]$$

ANALYSIS → **Signal representations**

Models

TRANSFORMATION → **Denoise, Separate, Time-shift,...**

RECOGNITION → **Classify, Identify,…**

« Hello, .. »

Text or symbols

« Hello, .. »

SYNTHESIS →

Audio signal

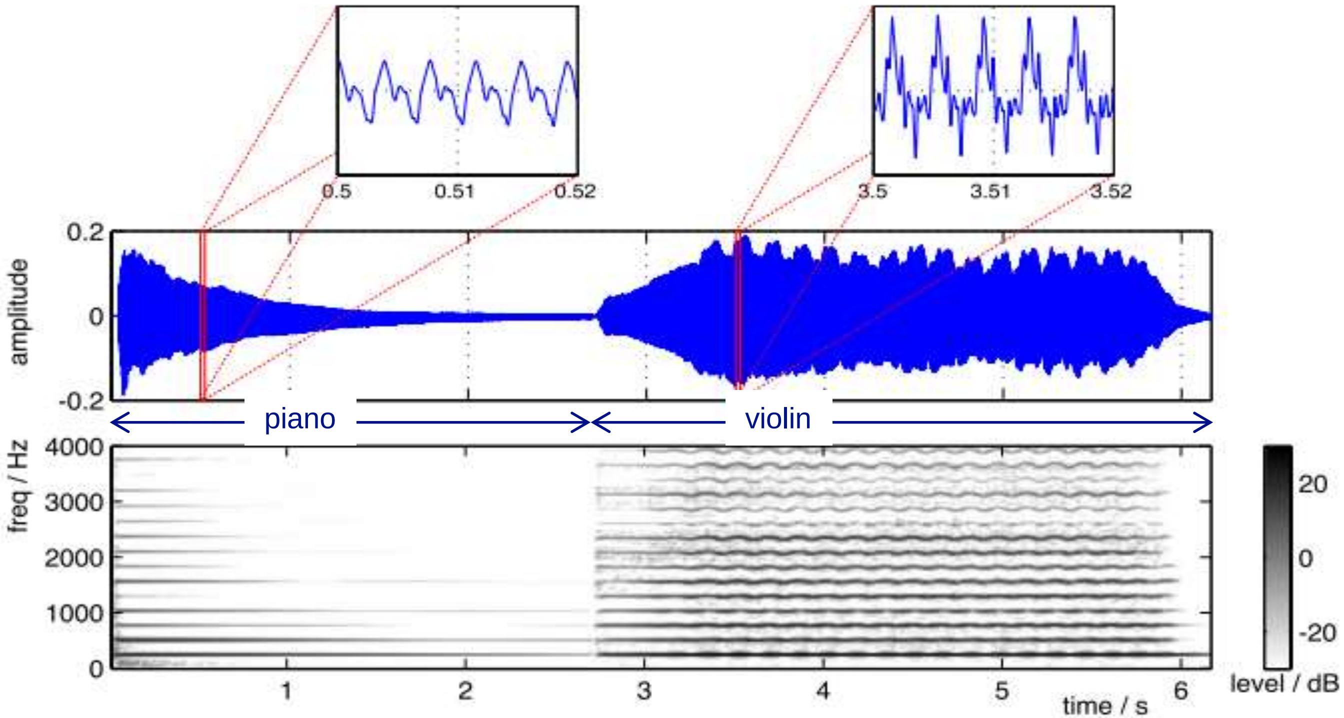
Analysis: Time-frequency representations

- **Example with a music signal :** note C (Do), with fundamental frequency of 262 Hz, played on a piano and then on a violin.

Time-domain signal

Spectrogram

(Short-term Fourier Transform)



Images from M. Mueller, D. Ellis, A. Klapuri, G. Richard « Signal Processing for Music Analysis, IEEE Trans. On Selected topics of Signal Processing, oct. 2011

Some generic signal models

- **Sum of sinusoïds + noise**

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

Some generic signal models

- **Sum of sinusoïds + noise**

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

- Modulated noise models for speech synthesis and modification [2]
- Damped sinusoids models for parametric audio coding [3]

[2] G. Richard, C. d'Alessandro, "Analysis/synthesis and modification of the speech aperiodic component", Speech Communication, Vol. 19, Issue 3, September 1996, Pages 221–244

[3] O.Derrien, R. Badeau, G. Richard, "A Parametric Audio Coding with Exponentially Damped Sinusoids, IEEE Trans. on Audio, Speech and Language Processing, Vol 21, N° 7, July 2013.

Some generic signal models

- **Sum of sinusoïds + noise**

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

  - Modulated noise models for speech synthesis and modification [2]
  - Damped sinusoids models for parametric audio coding [3]
  - Adaptive Signal subspace tracking [4]
  - Frequency estimation in Amplitude/Frequency modulated models [5]

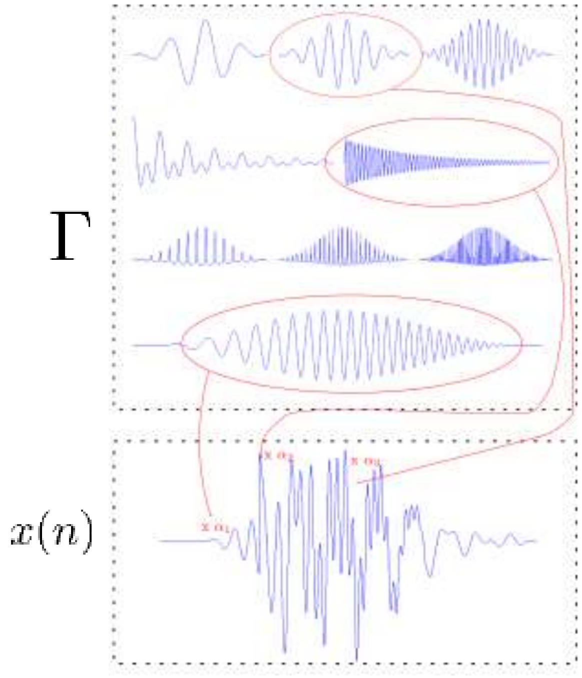
[2] G. Richard, C. d'Alessandro, “Analysis/synthesis and modification of the speech aperiodic component”, Speech Communication, Vol. 19, Issue 3, September 1996, Pages 221–244
[3] O.Derrien, R. Badeau, G. Richard, "A Parametric Audio Coding with Exponentially Damped Sinusoids, IEEE Trans. on Audio, Speech and Language Processing, Vol 21, N° 7, July 2013.
[4] R. Badeau, B. David and G. Richard, “Fast Approximated Power Iteration Subspace Tracking”, IEEE Trans. on Signal Processing, Vol. 53, Issue 8, Part 1, Aug. 2005 Page(s):2931 – 2941
[5] M. Betser, P. Collen, G. Richard and B. David « Estimation of frequency for AM/FM models using the phase vocoder framework», IEEE Trans. on Signal Processing, Vol. 56, N°. 2, February 2008., Page(s):505 - 517.

More specific models …

- **Signal « decomposition » methods**

$$x(n) = \sum_{\lambda \in \Gamma} \alpha_\lambda h_\lambda(n)$$

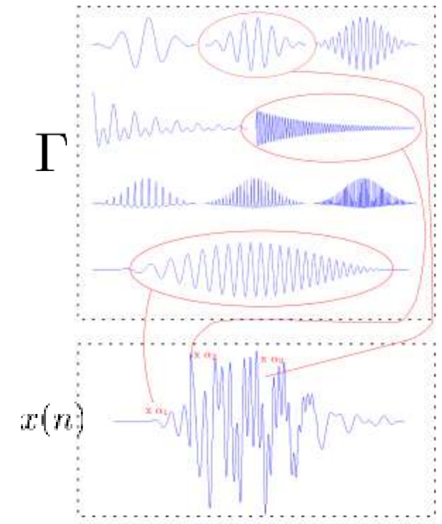
➡ The signal is a linear combination of atoms $h_\lambda(n)$ taken in a dictionary Γ



More specific models …

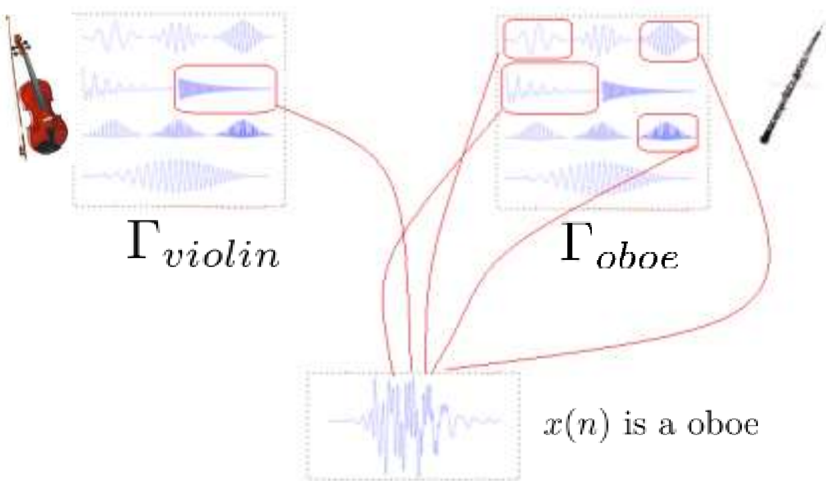
- **Signal « decomposition » methods**

$$x(n) = \sum_{\lambda \in \Gamma} \alpha_\lambda h_\lambda(n)$$



➡ With « informed » atoms :

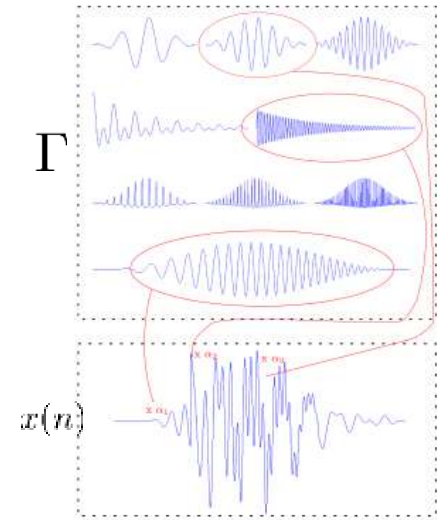
– *Source identification and separation* [6]



[6] P. Leveau, E. Vincent, G. Richard and L. Daudet, « Instrument-Specific Harmonic Atoms for Mid-Level Musical Audio Representation » IEEE Trans. on ASLP, Volume 16, N°1 Jan. 2008 Page(s):116 – 128

More specific models …

■ Signal « decomposition » methods



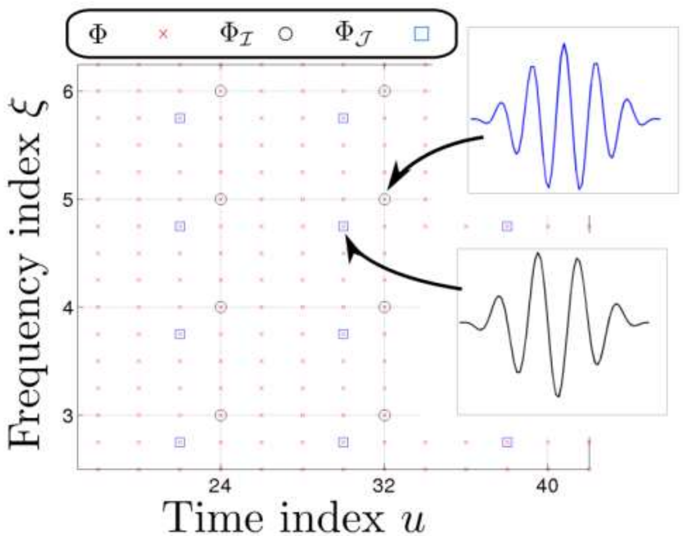
$$x(n) = \sum_{\lambda \in \Gamma} \alpha_\lambda h_\lambda(n)$$

➡ With multi-resolution time-frequency atoms :

– *Music signal compression* [7]

With sequence of random dictionaries :

– *Efficiency, denoising (audio, EEG signals)* [8]



[6] P. Leveau, E. Vincent, G. Richard and L. Daudet, « Instrument-Specific Harmonic Atoms for Mid-Level Musical Audio Representation » IEEE Trans. on ASLP, Volume 16, N°1 Jan. 2008 Page(s):116 – 128

[7] E. Ravelli, G. Richard, L. Daudet, Union of MDCT bases for audio coding, IEEE Trans. on Audio, Speech and Language Processing, Vol. 16, Issue 8, pp 1361-1372, Nov. 2008.

[8] M. Moussallam, L. Daudet, G. Richard, "Matching pursuits with random sequential subdictionaries", Signal Processing, 2012.

More specific models …

- **Non-negative Matrix factorization (NMF)[9]**

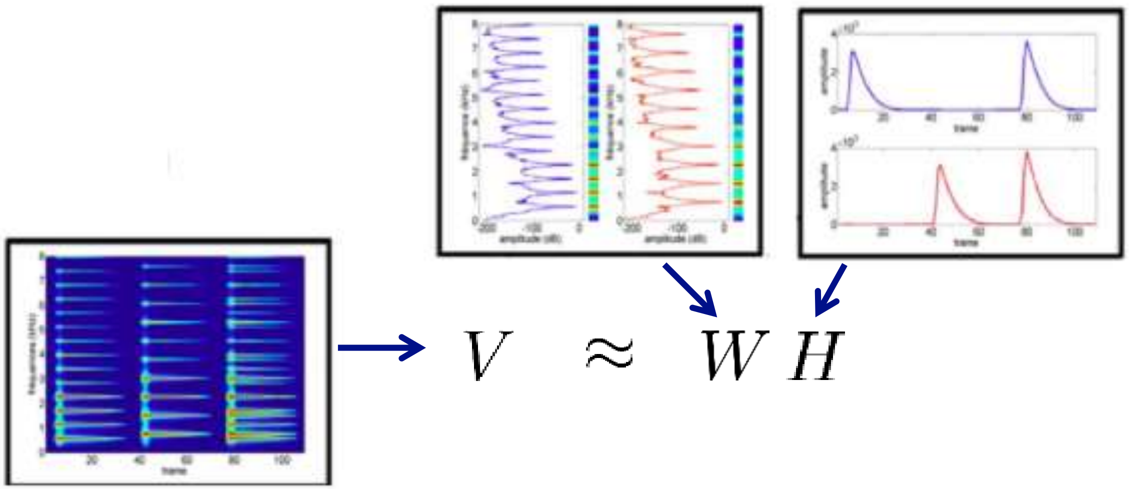
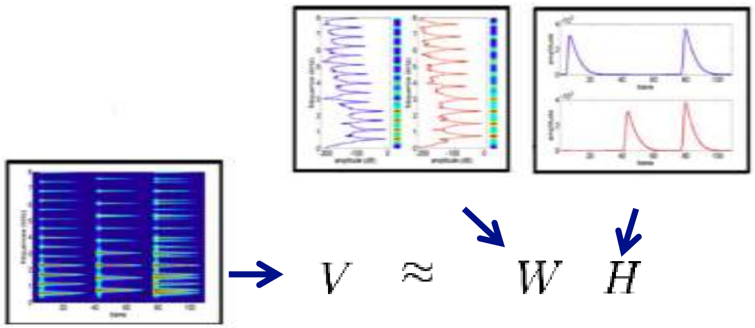


Figure from R. Hennequin

[9] D. Lee, H., Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).

More specific models …

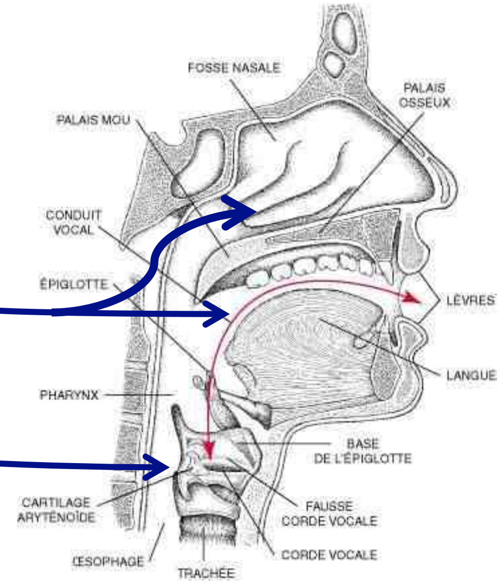
- **Non-negative Matrix factorization (NMF)[9]**
  - Underdetermined source separation



  - Un example on singing voice separation

$$\underbrace{\mathbf{X}}_{\text{Recording}} = \underbrace{\mathbf{V}}_{\text{Voice}} + \underbrace{(\mathbf{W}^M\mathbf{H}^M)}_{\text{music}},$$

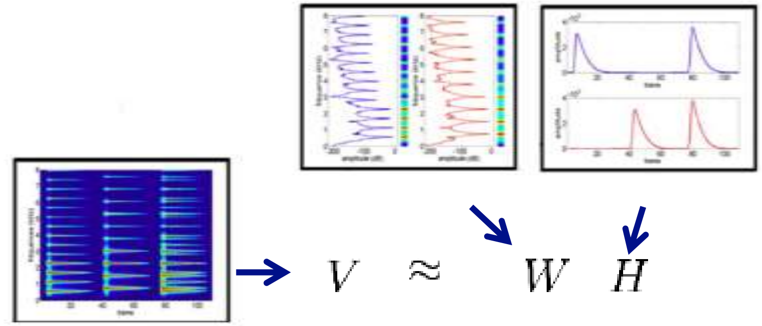
$$\underbrace{\mathbf{V}}_{\text{Voice}} = \underbrace{\mathbf{S}}_{\text{source}} \bullet \underbrace{\mathbf{F}}_{\text{filter}}$$



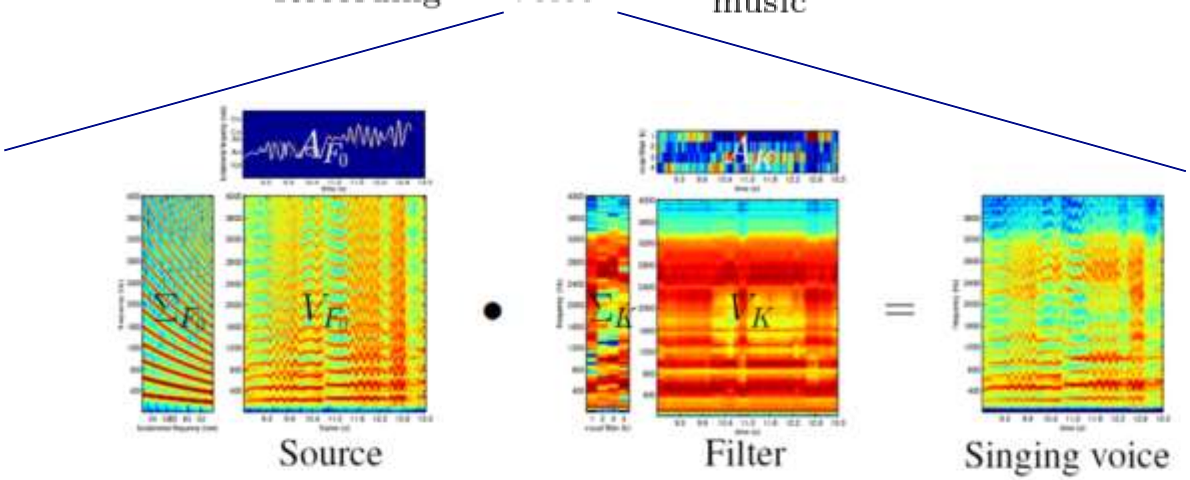
[9] D. Lee, H., Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).
[10] J-L Durrieu, B. David, G. Richard, A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, October 2011.

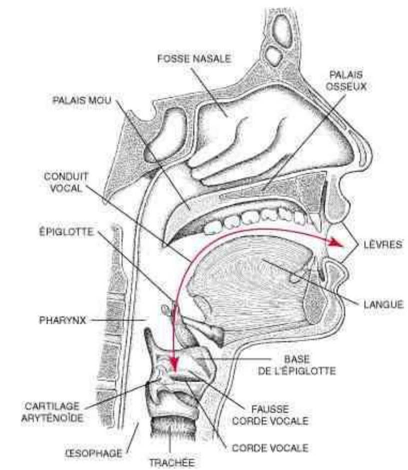
More specific models …

- **Non-negative Matrix factorization (NMF)[9]**
  - Underdetermined source separation
  - Un example on singing voice separation



$$\underbrace{\mathbf{X}}_{\text{Recording}} = \underbrace{\mathbf{V}}_{\text{Voice}} + \underbrace{(\mathbf{W}^M\mathbf{H}^M)}_{\text{music}},$$

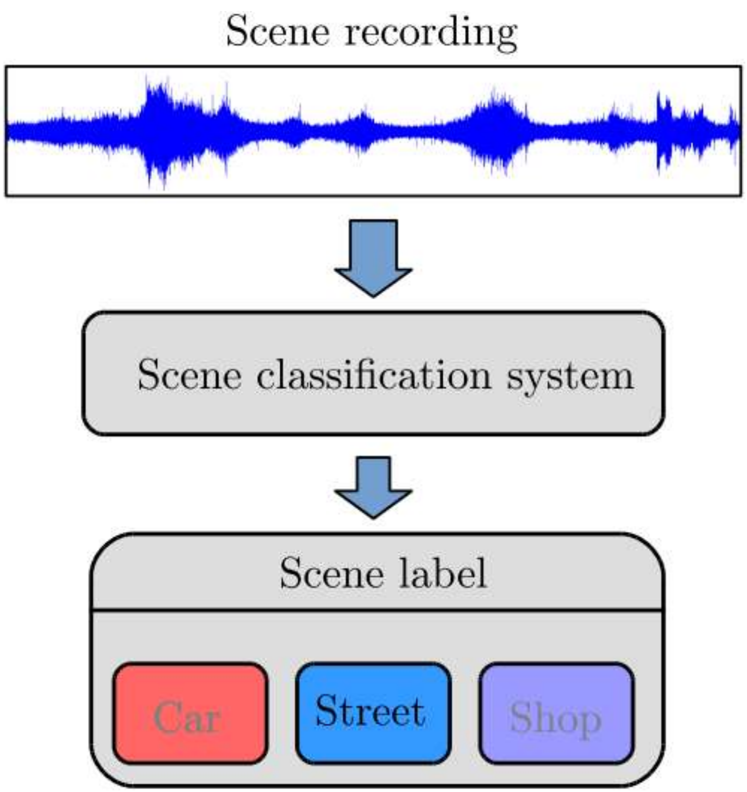


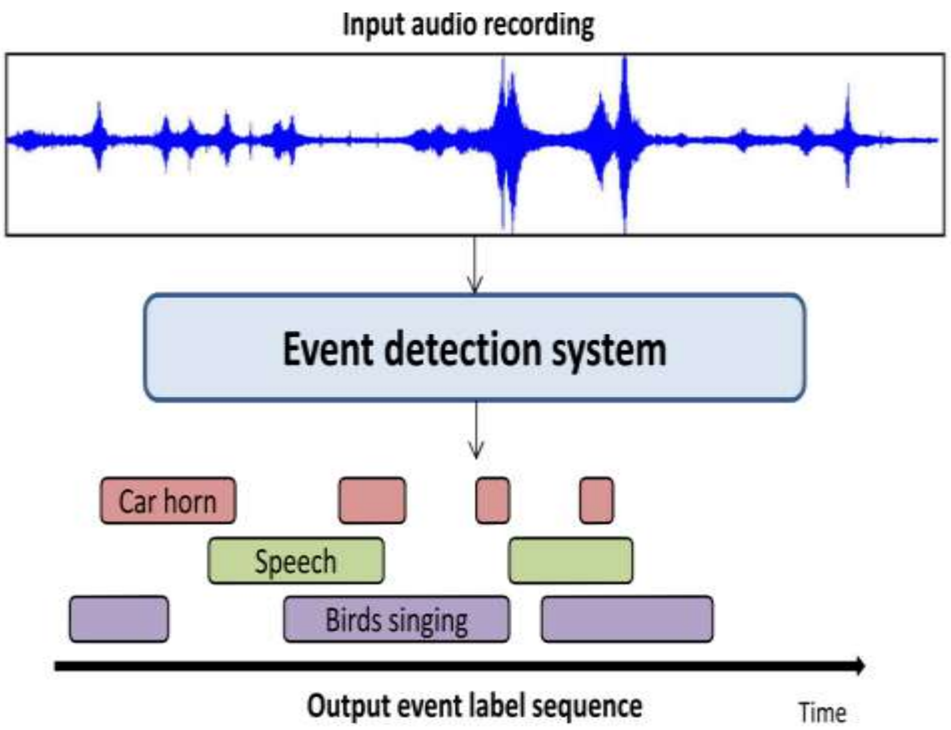


[9] D. Lee, H., Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401,** 788–791 (1999).
[10] J-L Durrieu, B. David, G. Richard, A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, October 2011.

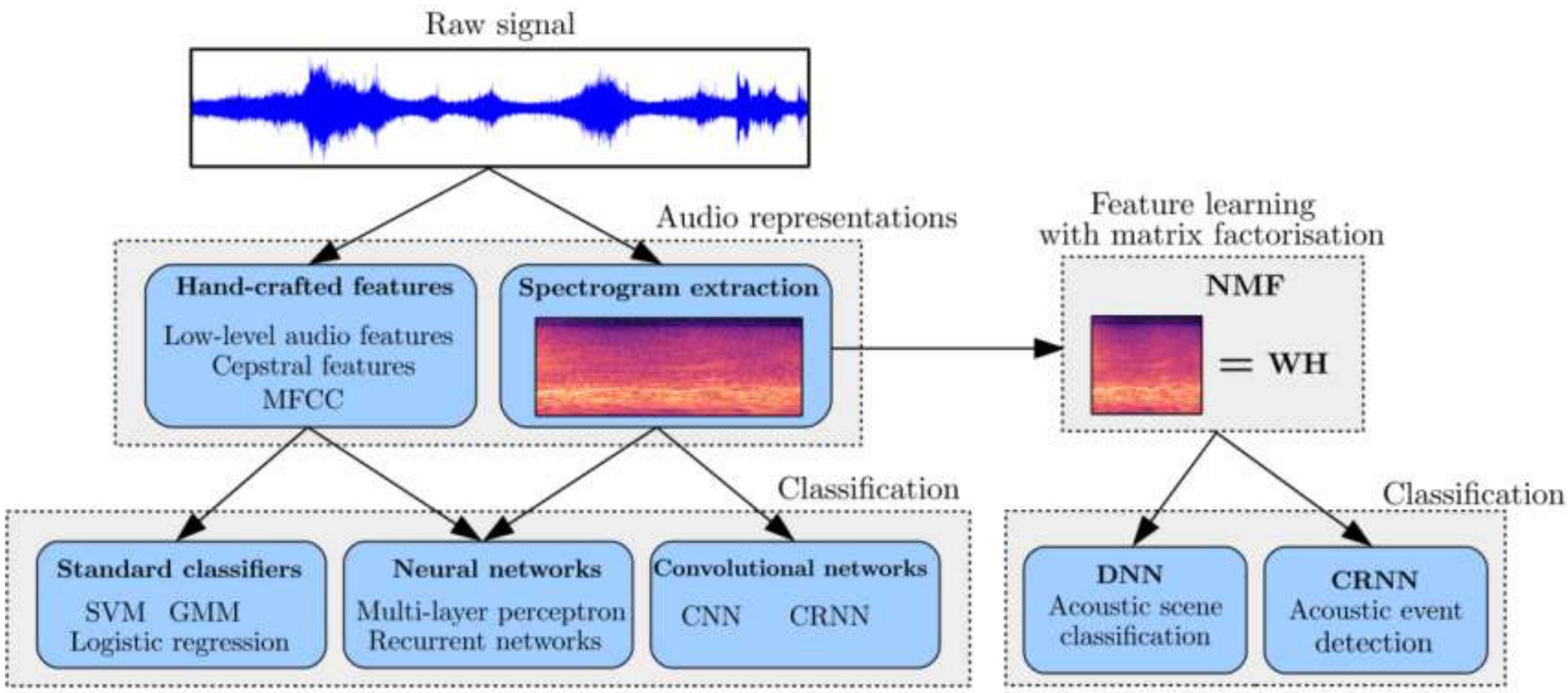
To recognize

Sound events and acoustic scene recognition





Associating signal models and deep learning



[11] V. Bisot, R. Serizel, S. Essid, G. Richard, "Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (2017), Special Issue on Sound Scene and Event Analysis.

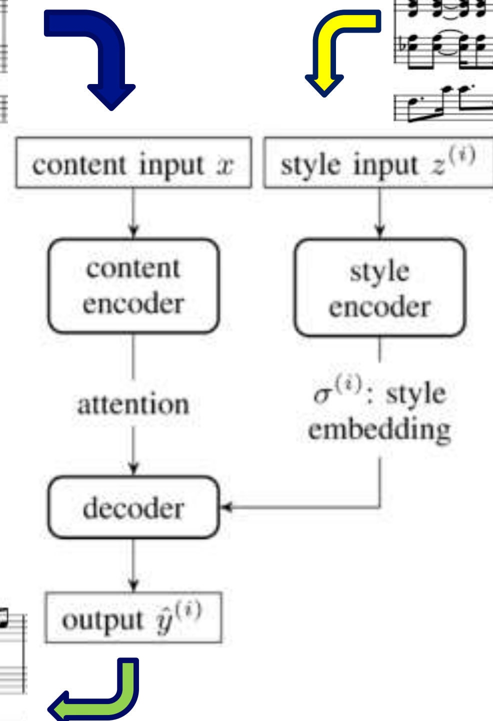
Recognize, Transform, Synthetize …

Symbolic music style transfer

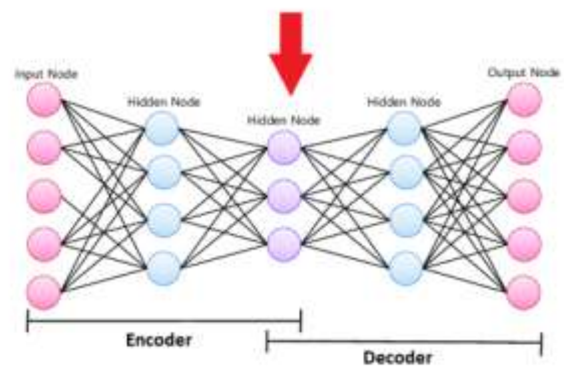
- … Or playing a given music file in the style of another music excerpt.

Swing jazz

Samba



Signal representation



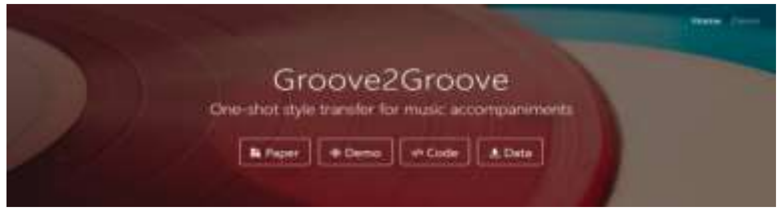
[13] Ondrej Cifka, Umut Simsekli, Gaël Richard, "Groove2Groove: One-Shot Music Style Transfer with Supervision from Synthetic Data", IEEE/ACM Transactions on Audio, Speech, and Language Processing, (preprint) accepted for publication, 2020

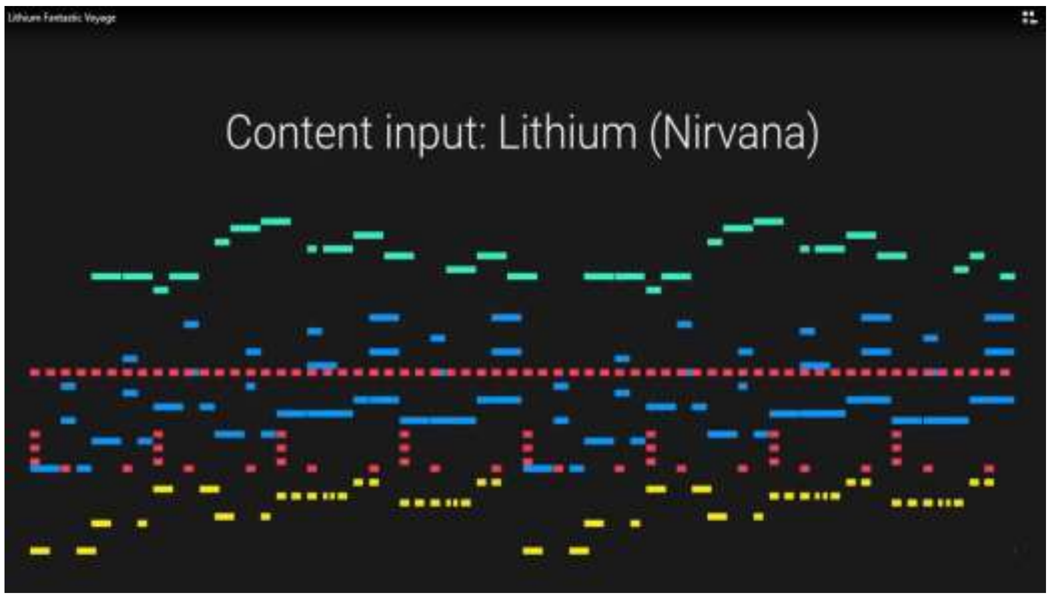
Sound examples at : *https://groove2groove.telecom-paris.fr*

Recognize, Transform, Synthetize …
Symbolic music style transfer

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

A short demo





[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*

Conclusion

- **… towards hybrid models**

  - Models are **parameter efficient** and **interpretable**,
  - Deep Neural networks are **very powerful** but needs **huge amount of data** and **computing power**, and are not always easily interpretable.

  - Hybrid models: promise for more explanability, frugality and efficiency