INFORMED AUDIO SOURCE SEPARATION

Gaël Richard

Institut Telecom, Telecom ParisTech, CNRS LTCI, France

With help from A. Liutkus, A. Ozerov

AES 53rd Conference on Semantic Audio

Audio recordings

- **What is an audio recording ?**

Audio recordings

- **What is an audio recording ?**

  - It is composed of *audio objects* or *sources…*

    piano drums guitar …. (stop)

  - …. Which are mixed together into a *mixture (*i.e. the audio recording*)* which is possibly multichannel (stereo is the most common for music)

Audio recordings

- **What is an audio recording ?**

  - It is composed of *audio objects* or *sources*…

    piano drums guitar …. (stop)

  - …. Which are mixed together into a *mixture* *(*i.e. the audio recording*)* which is possibly multichannel (stereo is the most common for music)

- **In most cases only the mixture is available which limits *Active Listening* capabilities …**

Applications

- **What could we do if we had the separated audio objects ?**
  - Active listening
  - Karaoke
  - Remixing
  - Music information retrieval
    - Cover song detection,
    - Music transcription (audio-to-midi, instrument recognition,…)
  - ….

From Source separation to Informed Source Separation

- **How to recover the audio objects ?**
  - **Using blind source separation**
    - *Separation is only done using the audio mixture.*
    - *But…quality is often not sufficient for active listening applications.*

    - *Exemple of Blind leading voice extraction [Durrieu&al.2011]…*

	Original	Backgrounds	Leading voice
Singing voice			
Trumpet			

J-L Durrieu, & al. A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, *October 2011.*

From Source separation to Informed Source Separation

- **How to recover the audio objects ?**
  - **Or … relying on Informed Source Separation (ISS)**
    - Side information is transmitted to the separation module
    - Separation is done using the mixture and the side information

From Source separation to Informed Source Separation

- **How to recover the audio objects ?**
  - **Or … relying on Informed Source Separation (ISS)**
    - Side information is transmitted to the separation module
    - Separation is done using the mixture and the side information

    - *Side information can be:*
      - Information **about the sources** (e.g. MIDI scores, information extracted from cover versions, types of the sources, etc.…)
      - Directly **extracted from the source** signals in an encoding stage but with an additional constraint: this information needs to be small

Keynote content

- **Objective**
  - To provide an overview of major trends in Informed Source Separation (ISS)

- **Outline of the keynote**
  - Introduction on Informed Source Separation
  - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
  - Overview of three trends in ISS:
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*
    - *Coding-based informed source separation*
  - Conclusion

Source separation by filtering techniques

- **General principle :**
  - The sources are recovered by filtering the mixtures

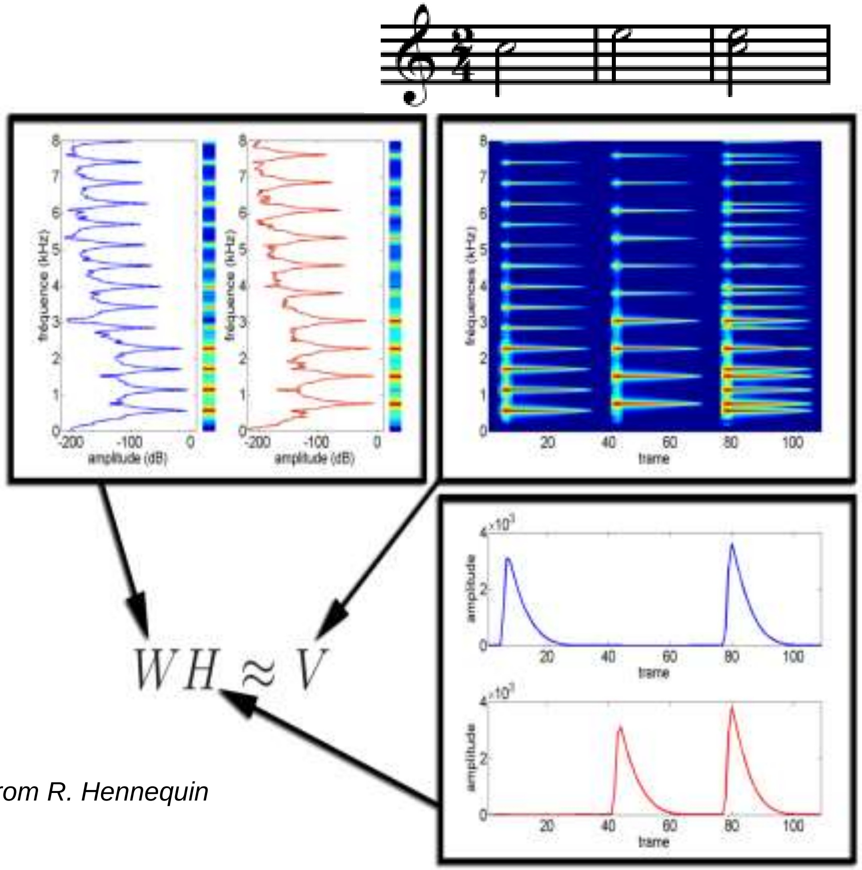
$$\underbrace{\hat{\mathbf{s}}}_{\text{sources}} = \underbrace{\mathcal{F}}_{\text{filtering technique}} \left\{ \underbrace{\mathbf{x}}_{\text{mixtures}} , \underbrace{\Theta}_{\text{parameters}} \right\}$$

A popular model for audio source separation : NMF

- **NMF = Non-negative Matrix Factorization**

“Templates or Atoms”

Original spectrogram



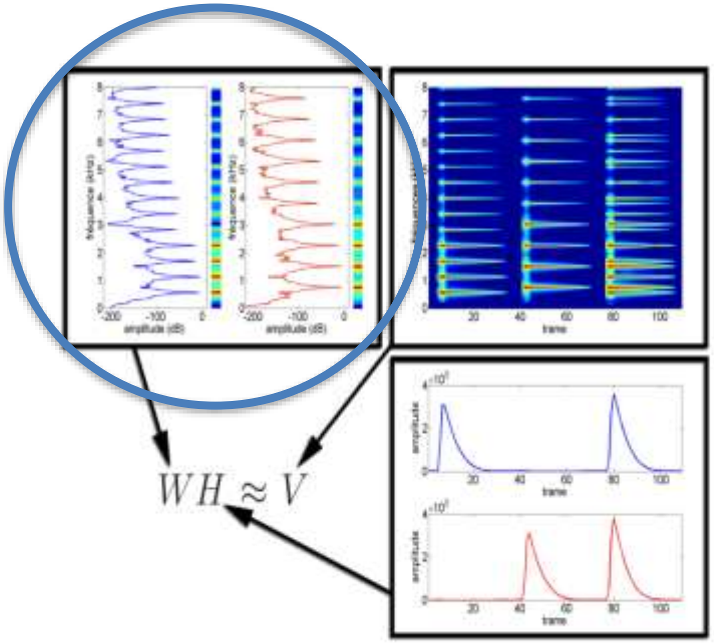
“Activations”

Image from R. Hennequin

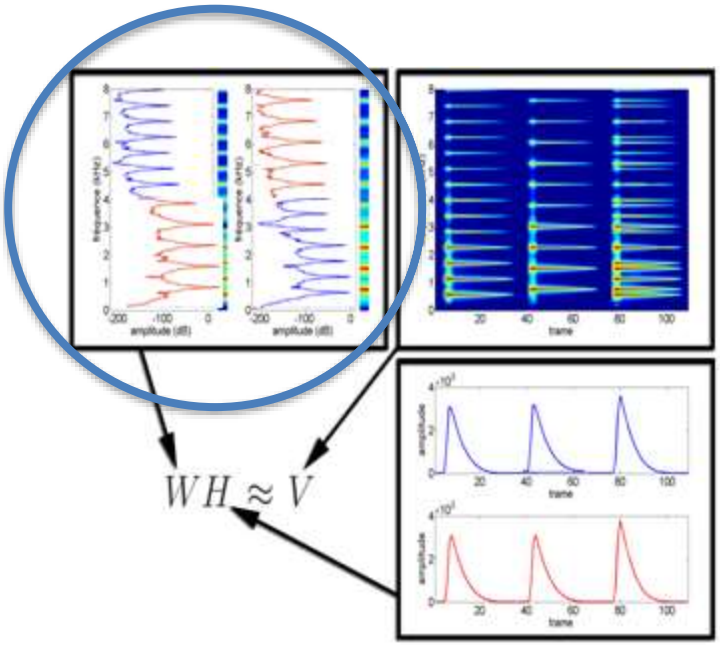
A popular model for audio source separation : NMF

- **NMF does not necessarily provides a semantically meaningful decomposition in absence of “constraints”**

Templates correspond to musical notes

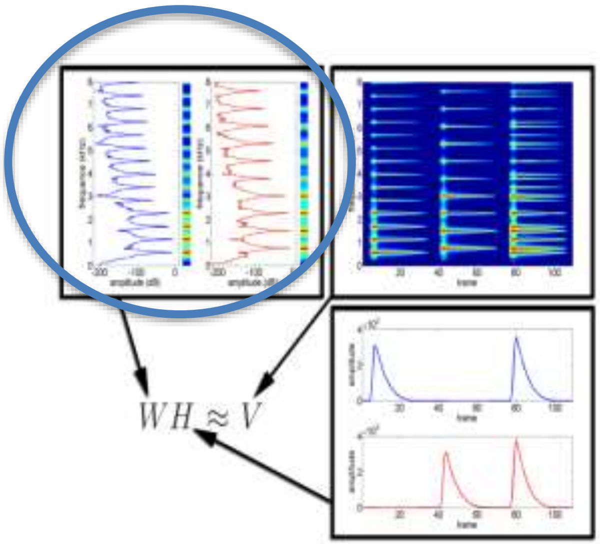


•Templates are built from half of each note and are less semantically meaningful
• Activations are less sparse



A popular model for audio source separation : NMF

- **How the template matrix W and activation matrix H are obtained [Lee&al. 1999]?**

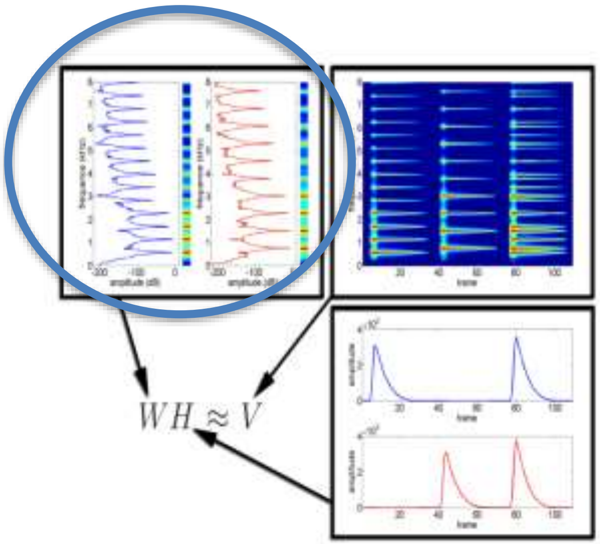


- **Minimization of D(V||WH)**
- **Problem separately convex in W and H** (for Euclidean and Kullback-leibler divergence)
- **Resolution leads to multiplicative update rules**

$$\begin{cases} \mathbf{H} \leftarrow \mathbf{H} \otimes \dfrac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T(\mathbf{WH})} \\ \mathbf{W} \leftarrow \mathbf{W} \otimes \dfrac{\mathbf{V}\mathbf{H}^T}{(\mathbf{WH})\mathbf{H}^T} \end{cases}$$

A popular model for audio source separation : NMF

- **What types of constraints can be used ?**



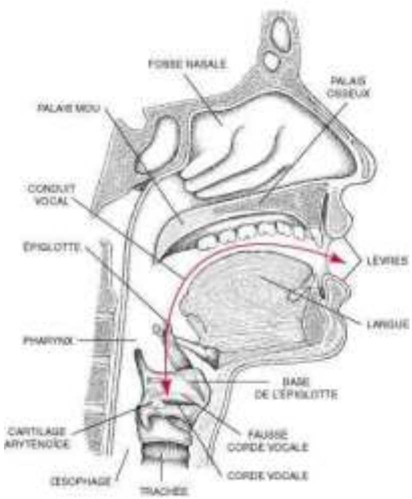
- **Harmonicity of the templates [Raczinsky&al.2007]**
  - To have a decomposition in “harmonic notes”
- **Spectral smoothness of the templates [Bertin&al.2010]**
  - To obtain realistic timbral notes
- **Temporal continuity of activation [Virtanen2007]**
  - To take into account that note activations are not erratic
- **Sparsity of the activations [Hoyer04][Smaragdis08]**
  - To take into account that not too many notes are played in a given time

A popular model for audio source separation : NMF

- **An example of model-based constraints for main melody separation:**

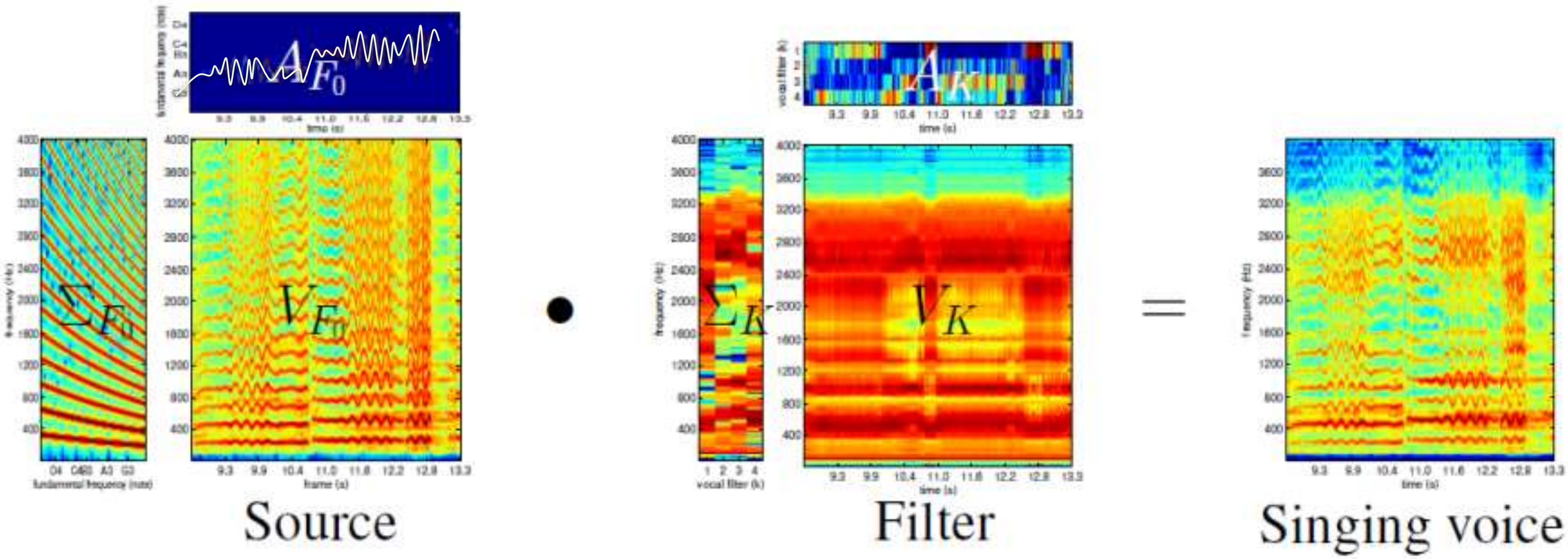
- The model: **A**udio **= V**oice **+ M**usic
  - The voice **V**oice follows a source filter production model : **V**oice = **S**ource * **F**ilter

  - Each component (Voice and Music) is represented by separate NMF



An example of model constrained NMF for singing voice extraction

- **Exploitation of a source/filter production model**



- **Exploitation of redundancy of the accompanying music**
  - Simple NMF model for background music ($\sum^m$ et A^m)

J-L Durrieu & al. G, Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals, IEEE Trans. On ASLP, March 2010.
J-L Durrieu, & al. A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, *October 2011*

Informed audio source separation

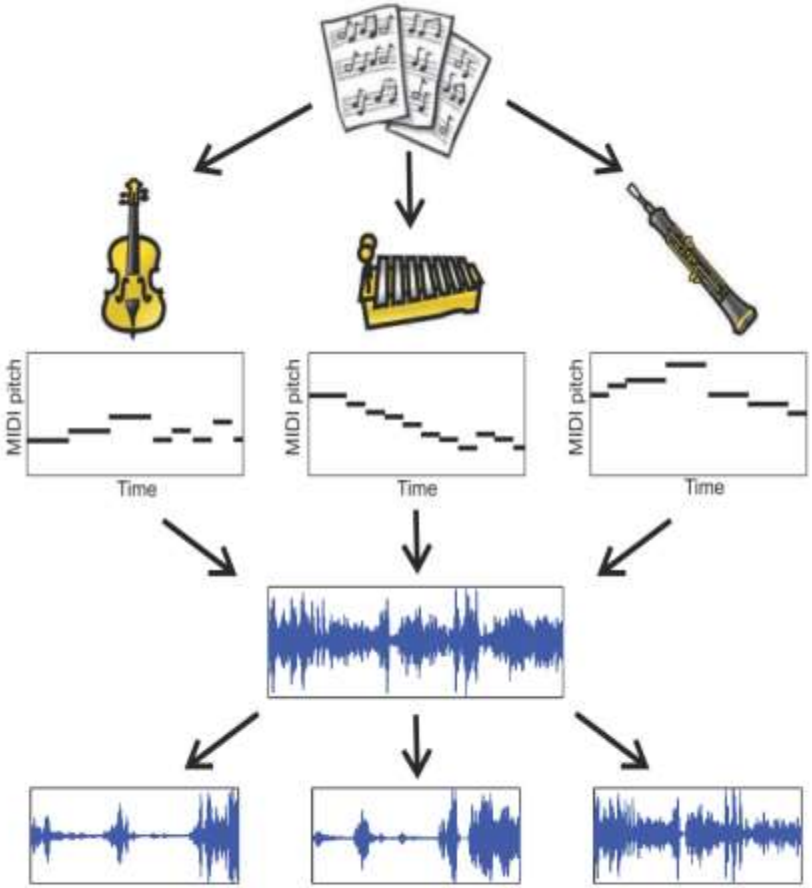
- **In Informed audio Source Separation (ISS), "a priori" constraints may be replaced (or completed) by specific "information"**
  - Overview of three trends in ISS:
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*
    - *Coding-based informed source separation*

Overview of three trends in ISS

Auxiliary data-informed source separation,
User-guided source separation,
Coding-based informed source separation

Auxiliary data-informed source separation

"Score-informed" source separation



Musical Score

Midi representation of each track (or source)

Use the MIDI information To guide audio separation

Separated tracks of Improved quality

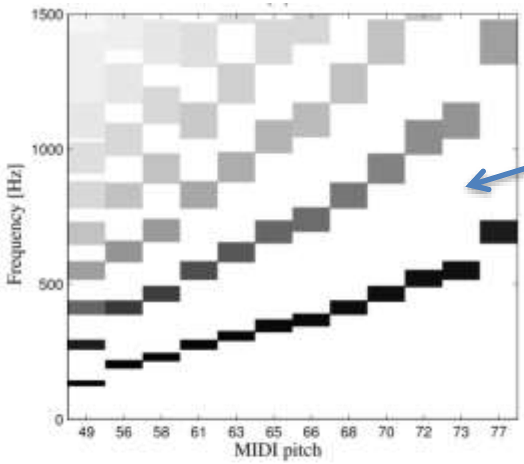
Figures from S. Ewert and M. Müller. Score informed source separation. In Multimodal Music Processing, Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

Auxiliary data-informed source separation
"Score-informed" source separation

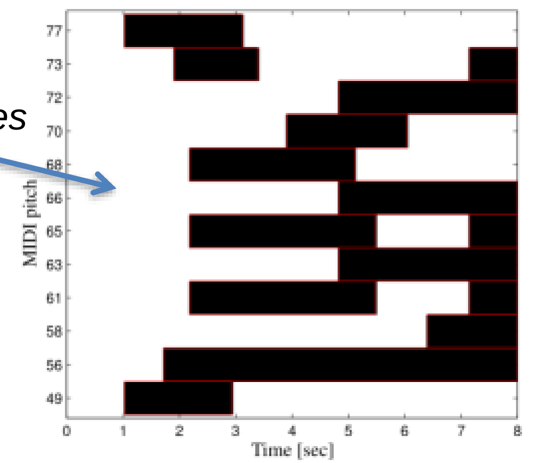
- **An example in the framework of NMF (V= W . H)**

Matrix W: synthetic harmonic templates are defined for each note

Matrix H: Idealized activations obtained from the MIDI score



White = Zero values



Due to multiplicative update rules, zero entries at the initialization stay at zero

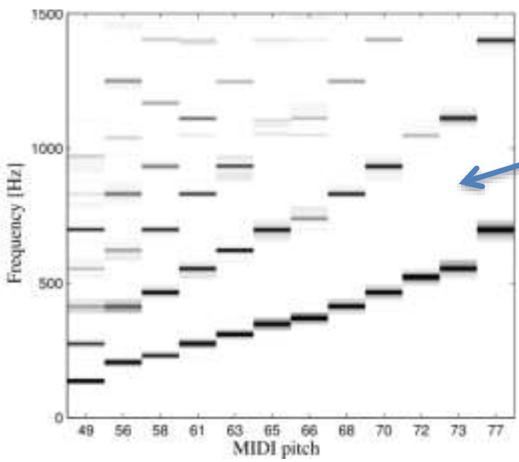
Auxiliary data-informed source separation

"Score-informed" source separation

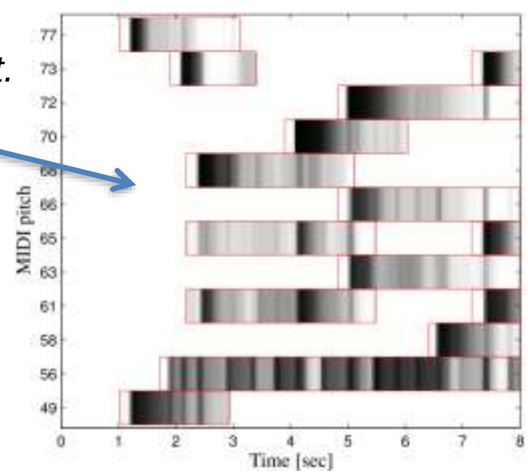
- **An example in the framework of NMF (V= W . H)**

Matrix W: obtained after convergence

Matrix H: obtained after convergence



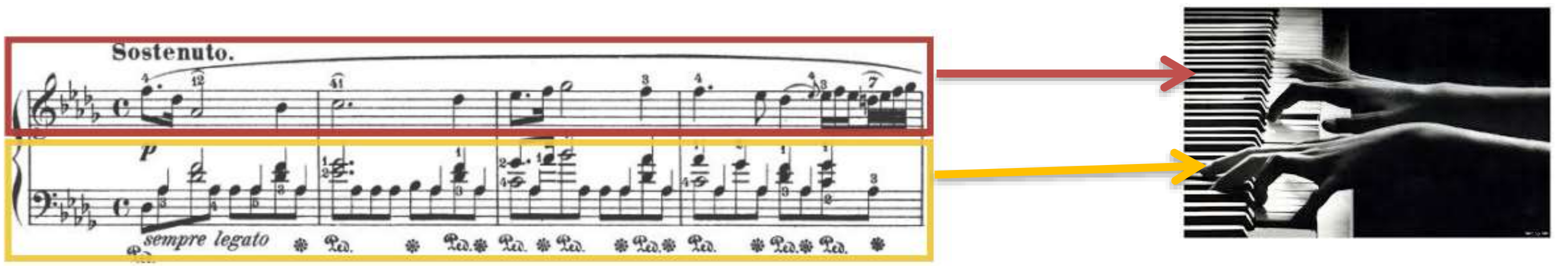
Null entries at init. remain null



Auxiliary data-informed source separation

"Score-informed" source separation

- **Demonstration: "left hand" – "right hand" separation**



Original recording (Chopin)

MIDI synthesis of the score

Left hand

Right hand

S. Ewert and M. Müller. Score informed source separation. In Multimodal Music Processing, Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

Auxiliary data-informed source separation

"Text-informed" speech separation

- **Extension of the source-filter model of Durrieu &al.**
  - Observed signal is described as "Speech + background"
    - X = S +B
  - The speech S is modeled as an Excitation-Filter-Channel signal:

$$\hat{\mathbf{V}}_S = \hat{\mathbf{V}}_S^e \odot \hat{\mathbf{V}}_S^\phi \odot \hat{\mathbf{V}}_S^c$$

Spectrogram of S

Spectrogram of Excitation (or source)

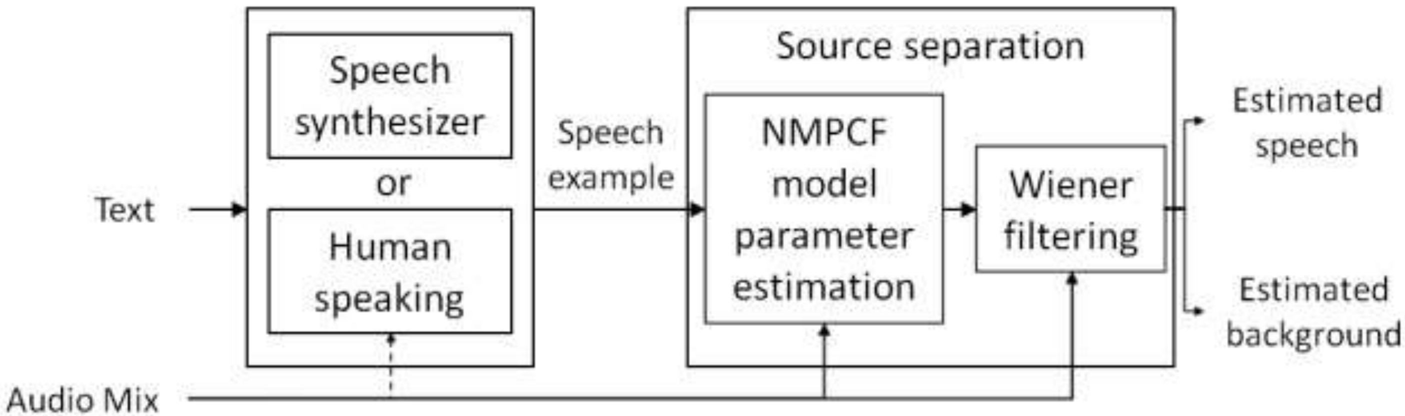
Spectrogram of "Filter" (e.g. formants)

Spectrogram of channel (e.g. microphone, reverberation,…)

Auxiliary data-informed source separation

"Text-informed" speech separation

- **How the text is used ?**



NMPCF = Non Negative Matrix partial co-factorization

L. Le Magoarou, A. Ozerov, N. Duong Text-Informed Audio Source Separation using Nonnegative Matrix Partial Co-Factorization, in Proc. of MLSP, 2013

Auxiliary data-informed source separation

"Text-informed" speech separation

- **Each component of the speech model is represented by a NMF**

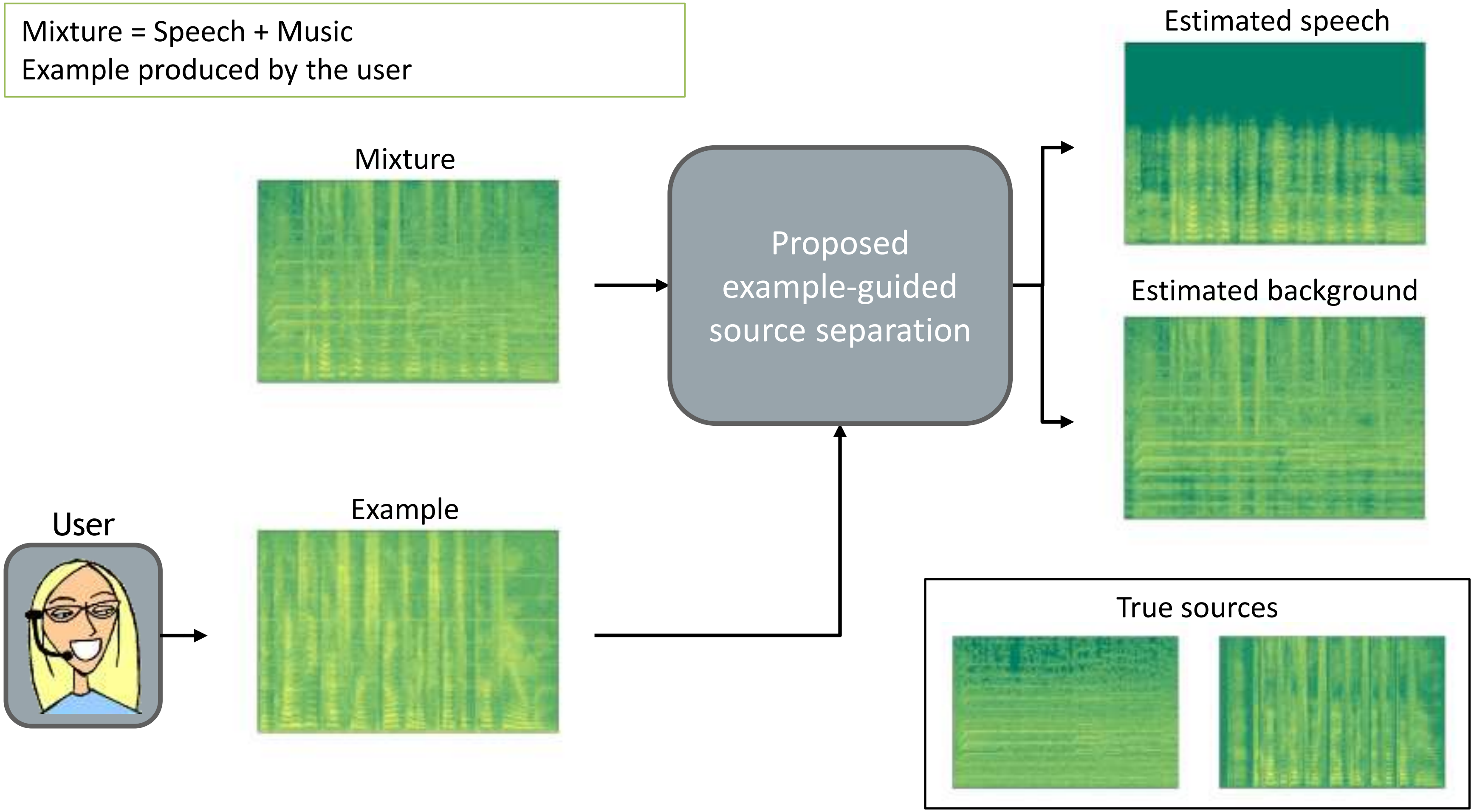
$$\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \overbrace{\underbrace{(\mathbf{W}^e\mathbf{H}_S^e)}_{\hat{\mathbf{V}}_S^e} \odot \underbrace{(\mathbf{W}_S^\phi\mathbf{H}_S^\phi)}_{\hat{\mathbf{V}}_S^\phi} \odot \underbrace{(\mathbf{w}_S^c\mathbf{i}_N^T)}_{\hat{\mathbf{V}}_S^c}}^{\hat{\mathbf{V}}_S} + \overbrace{\mathbf{W}_B\mathbf{H}_B}^{\hat{\mathbf{V}}_B}$$

- **In this representation the text (which gives phonetic information) will directly give information on the matrix linked to what is said, which is:** $\hat{\mathbf{V}}_S^\phi$

Auxiliary data-informed source separation

"Text-informed" speech separation : demonstration

Mixture = Speech + Music
Example produced by the user



Overview of three trends in ISS

Auxiliary data-informed source separation,
User-guided source separation,
Coding-based informed source separation

User-guided source separation

- **In this scenario, the user provides some partial information about the sources to be separated.**

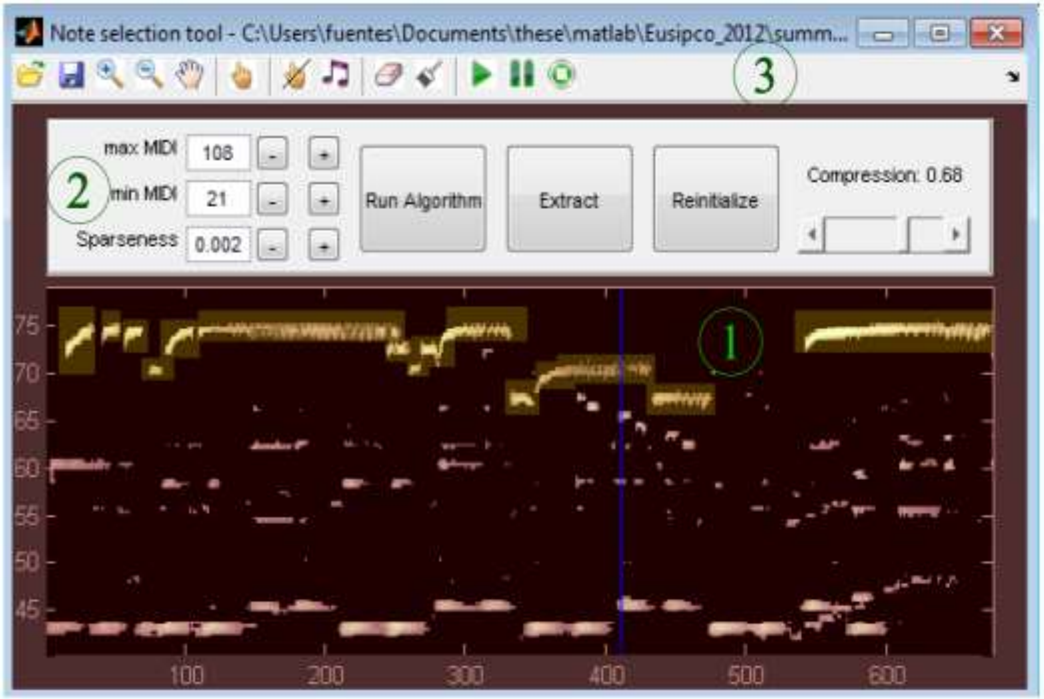
- **Two illustrative examples :**
  - Iterative source selection using a Graphical User-Interface (GUI)
  - Hummed-query for main melody extraction

  - Both examples are based on *Probabilistic Latent Component Analysis* models (which are probabilistic models similar to NMF)

User-guided source separation

User-selection using a GUI

- The user paints the parts corresponding to the melody in the GUI
- Algorithm is re-run but with many zero values in the initial decomposition for the melody part
- Several iterations are possible



B. Fuentes, R. Badeau et G. Richard : Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In Proc. of EUSIPCO, Bucarest, Romania, 2012.

User-guided source separation

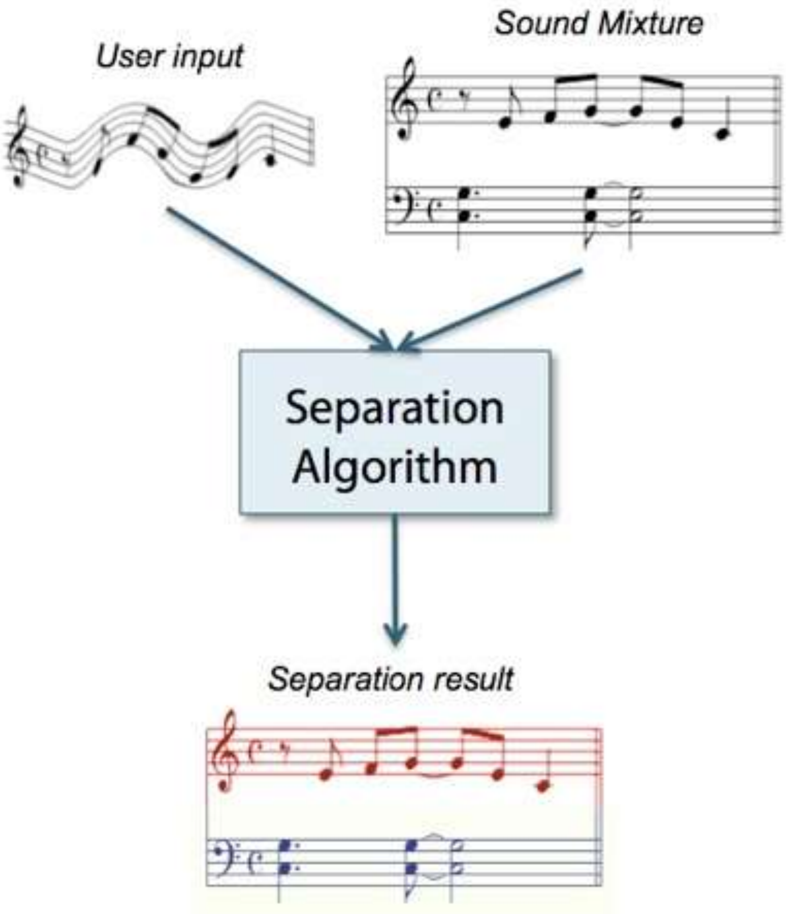
User-selection using a GUI

- Demo

B. Fuentes, R. Badeau et G. Richard : Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation. In Proc. of EUSIPCO, Bucarest, Romania, 2012.

User-guided source separation
Hummed melody input

- The user hums the melody of the instrument track that he wish to separate
- The melody produced is used as information for separating the melody in the mixture



From https://ccrma.stanford.edu/~gautham/Site/Humming.html

User-guided source separation

Hummed melody input

- **Demonstration: Video [Smaragdis &al. 2009]**

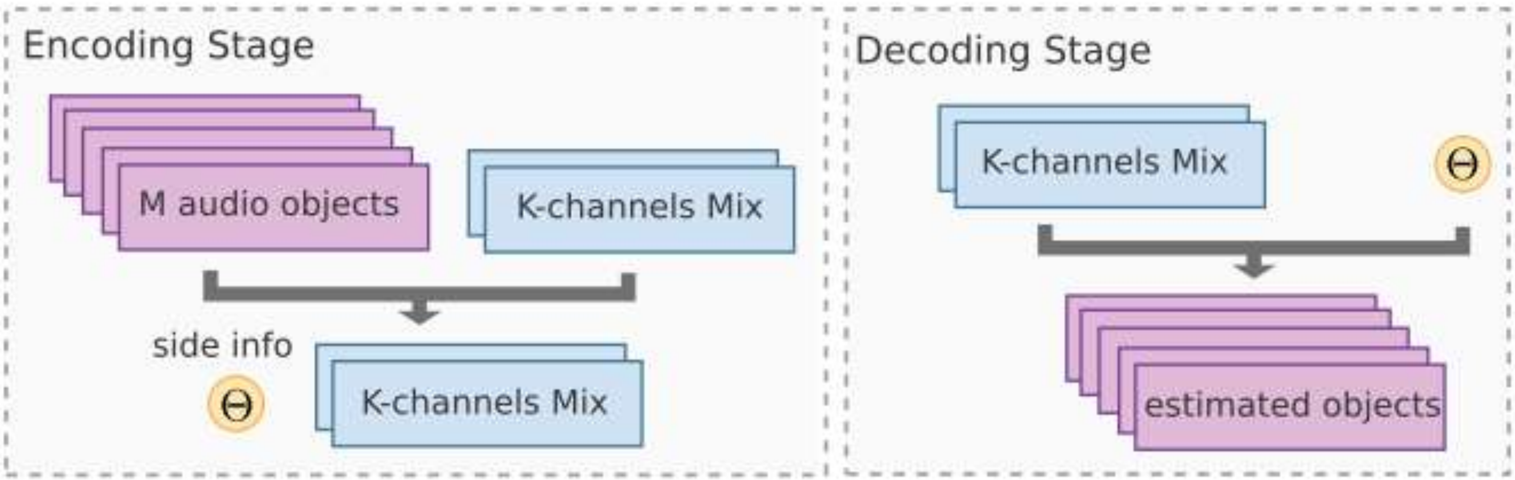
user-guide-short.mp4

P. Smaragdis, G. Mysore, "Separation by Humming": User Guided Sound Extraction from Monophonic Mixtures" in Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY. October 2009

Auxiliary data-informed source separation,
User-guided source separation,
Coding-based informed source separation

Coding-based informed source separation

- **Here, the information is obtained directly from the sources** (but the information needs to be well compressed to be useful)
- **Sources (or Audio objects) are known at a so-called encoding stage**



- Note that informed source separation in this case shares many similarities with Spatial Audio Object Coding approaches (see [Ozerov&al.11] for a discussion)

[Ozerov&al.11] A. Ozerov & al. Informed source separation: source coding meets source separation. In IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA'11), October 2011.

Coding-based informed source separation

- **What type of information is in the "side information"**

  - Could be the sources but then no point of source separation and huge bandwidth
  - Usually it is a partial information about the sources (obtained from the knowledge of the sources):

    - Time frequency activations of the two predominant sources [Parvaix & al.]
    - A compressed version of the source spectrograms (for example JPEG) [Liutkus & al.]

M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for informed source separation of audio signals with a single sensor. IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1464–1475, 2010.

A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard. Informed source separation through spectrogram coding and data embedding. Signal Processing, 92(8):1937 – 1949, 2012.

Coding-based informed source separation

- **What performances can be obtained ?**

Demo of CISS

- Original mix (7 sources)
- Demix signals (using 7 kbit/s per source for side info)

For comparison: AAC for a mono signal is around 32 – 64 kbis

Conclusion / Perspectives

- **Conclusion:**
  - Audio source separation is an extremely challenging task, especially when considering real-world stereophonic full-tracks.
  - Blind separation techniques do exist, but their performance may be greatly improved by using any available information apart from the mere mixture
  - The so-called Informed Source Separation was discussed with examples from three major trends, namely:
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*
    - *Coding-based informed source separation*

- **Some perspectives**
  - The type of information depends on the type of source separator and the application but how to limit the side-information to the minimum ?
  - How to exploit several informed source separators (e.g. separator fusion) in an optimal way ?
  - How to better exploit a multitrack cover version to perform source separation on the original recording ?
  - ….

Additional References

- [Hoyer04] P. Hoyer, “Non-negative Matrix Factorization with Sparseness Constraints”, Journal of Machine Learning Research 5 (2004) 1457–1469
- [Smaragdis08] P. Smaragdis , B. Raj et M.V. Shashanka : Sparse and shift-invariant feature extraction from non-negative data. In Proc. of ICASSP, pages 2069–2072, Las Vegas, Nevada, USA, 2008.
- [Virtanen2007] T. Virtanen : Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. on Audio, Speech and Language Processing, 15(3):1066–1074, 2007.
- [Bertin2010] N. Bertin , R. Badeau et E. Vincent : Enforcing Harmonicity and Smoothness in Bayesian Non-negative Matrix Factorization Applied to Polyphonic Music Transcription. IEEE Trans. on Audio, Speech and Language Processing, 18(3):538–549, 2010.
- [Raczinsky&al.2007] S. Raczinski, N. Ono, S. Sagayama, “Multipitch analysis with harmonic nonnegative matrix approximation”, in Proc. of ISMIR; Vienna, Austria, 2007