A TUTORIAL ON INFORMED AUDIO SOURCE SEPARATION

Alexey Ozerov

Technicolor

Antoine Liutkus

INRIA

Gaël Richard

Institut Telecom, Telecom ParisTech, CNRS LTCI, France

ICASSP 2014, May 4-9, 2014, Florence, Italy

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 *"Blind"* 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			

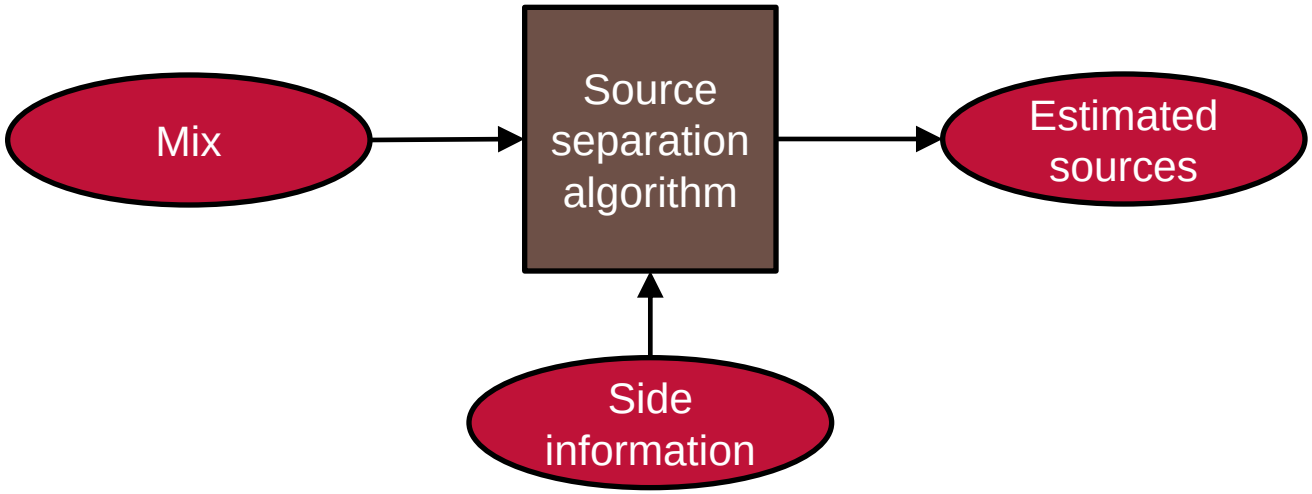
J-L Durrieu, & al. A musically motivated mid-level representation for pitch estimation and musical audio source separation, IEEE Journal on Selected Topics in Signal Processing, *October 2011.*

From *"Blind"* source separation to Informed Source Separation

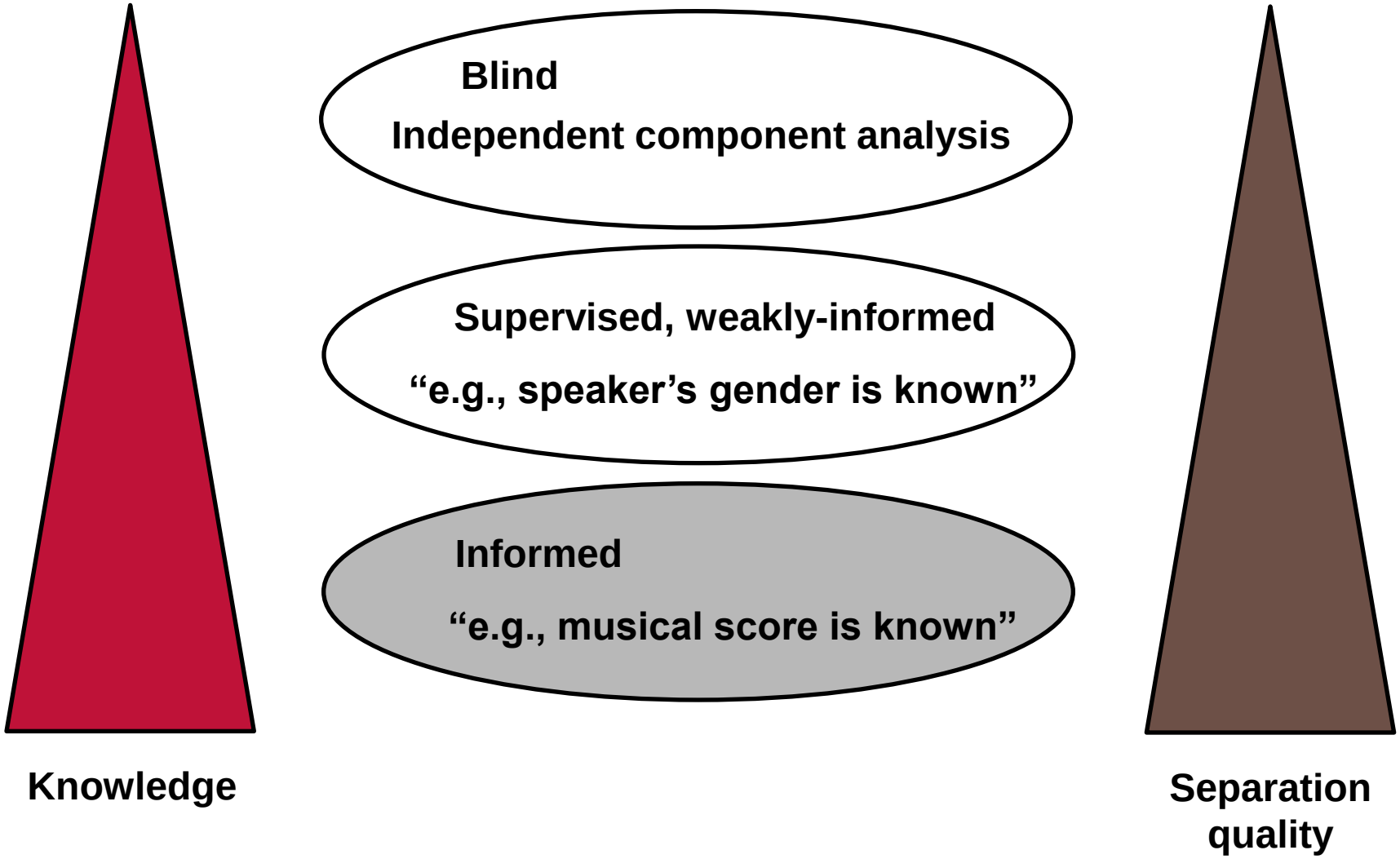
- **How to recover the audio objects ?**
  - **Or … relying on Informed Source Separation (ISS)**
    - Side information is transmitted to the separation module
    - Separation is done using the mixture and the side information

From *"Blind"* 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

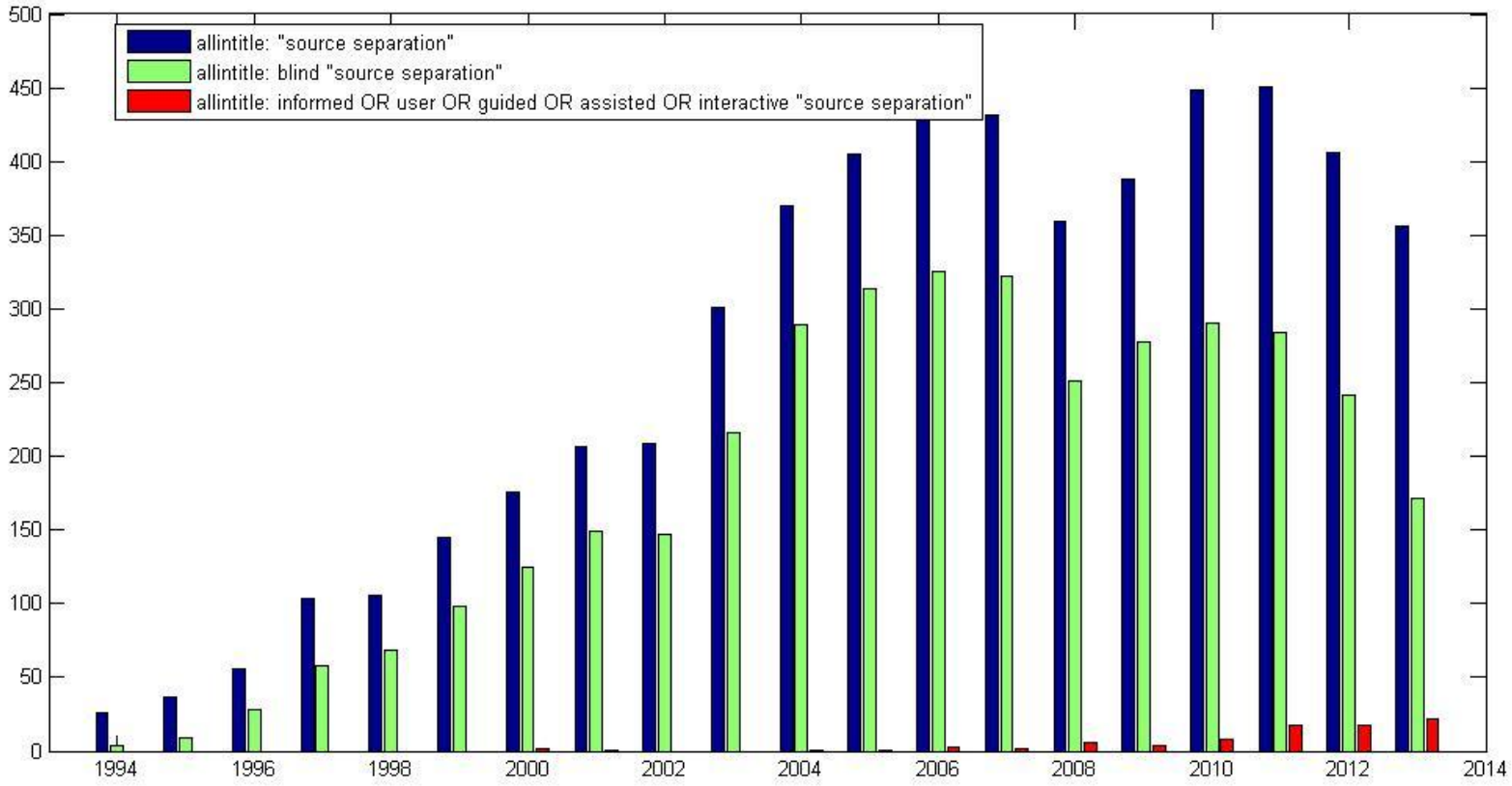


From *"Blind"* source separation to Informed Source Separation



Some figures about Informed Source Separation

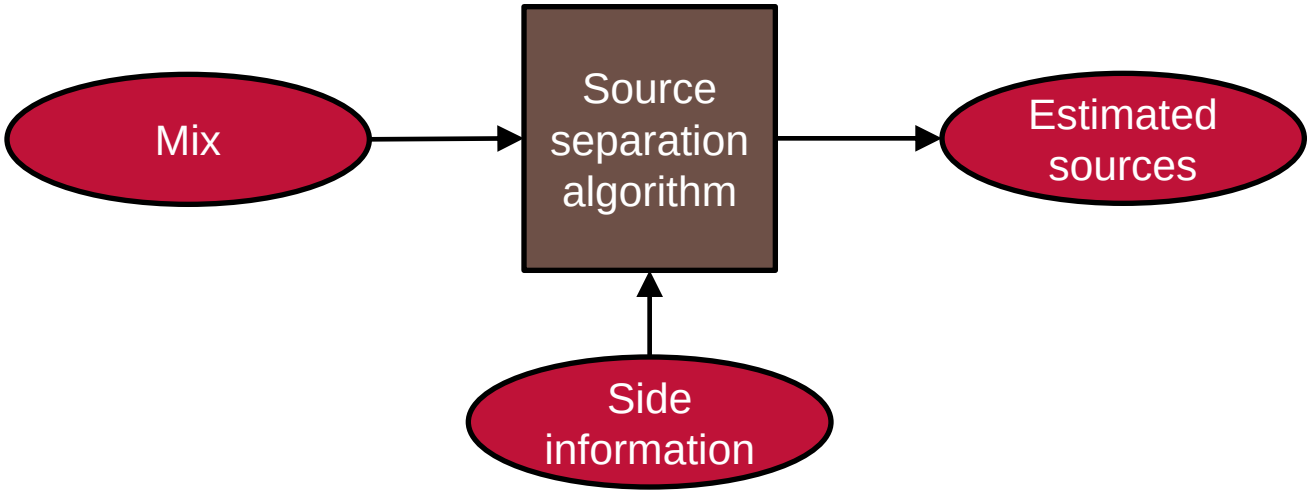
- ***Search results on Google scholar (March 2014)***



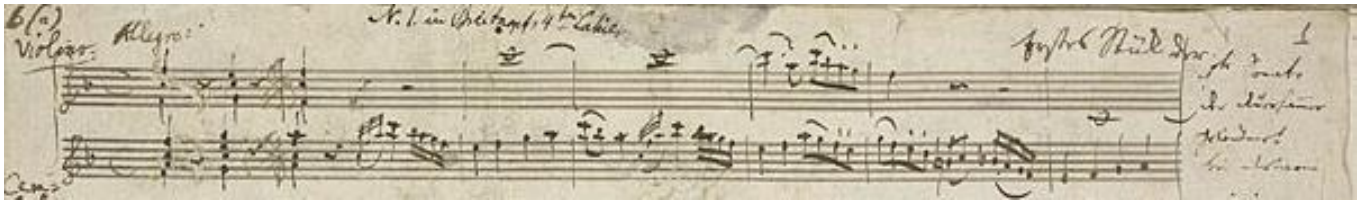
Informed source separation is growing … !

Trends in informed source separation

Trend 1: Auxiliary data-informed source separation

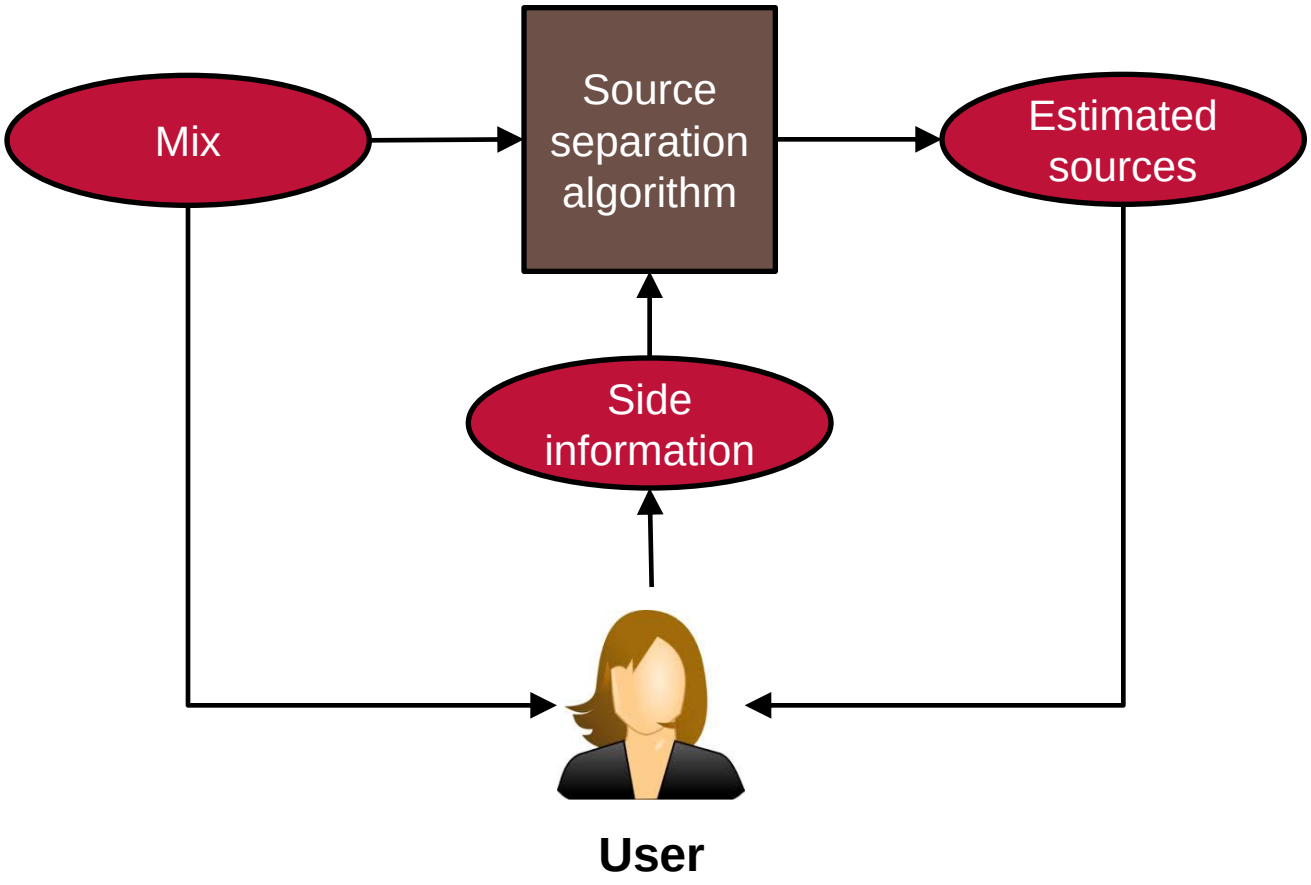


Existing side information, e.g., musical score



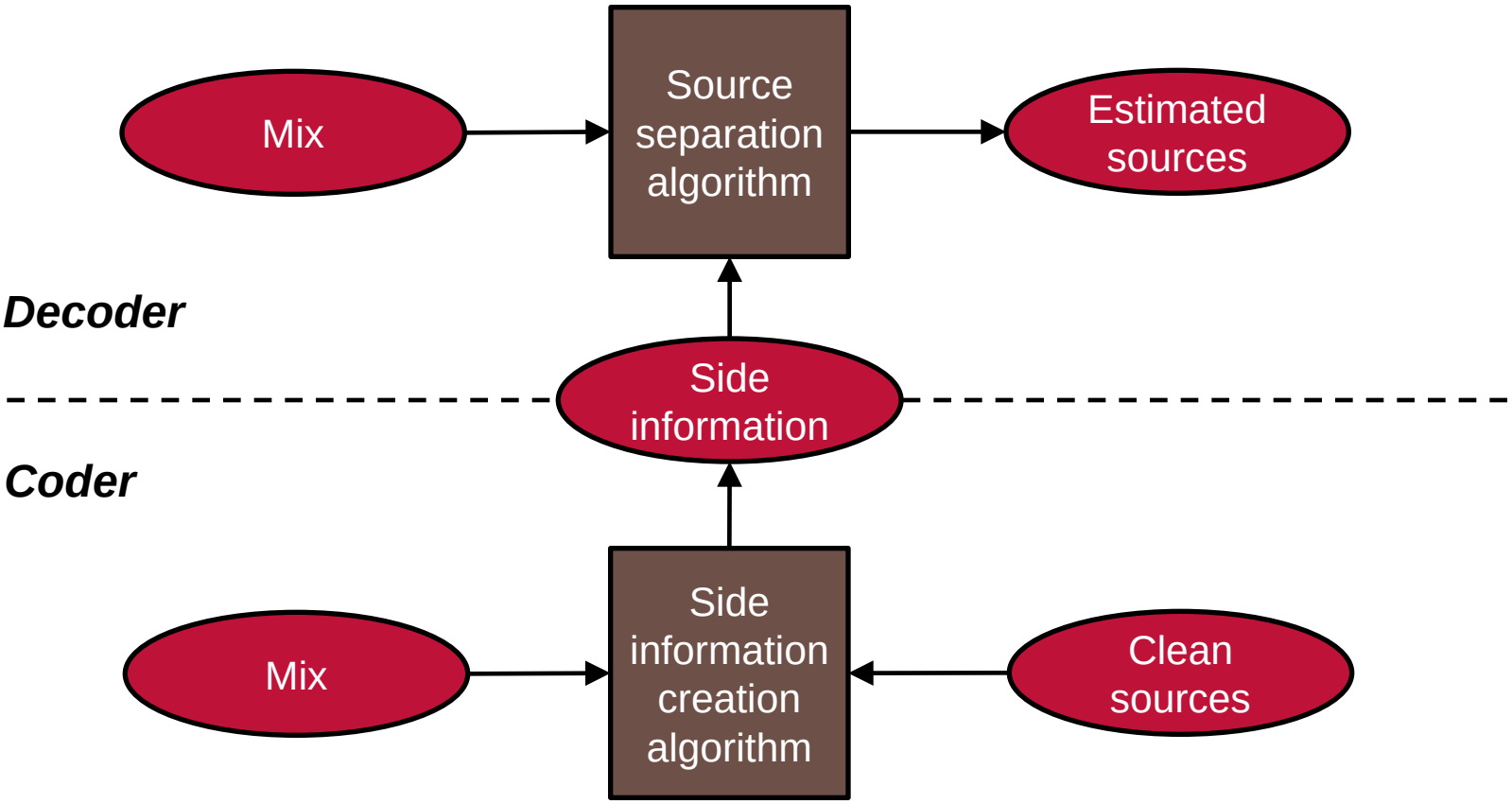
Trends in informed source separation

Trend 2: User-guided source separation



Trends in informed source separation

Trend 3: Coding-based informed source separation



Keynote content

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

- **Outline of the keynote**

  - *PART 1: Gaël RICHARD*
    - Introduction on Informed Source Separation
    - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).

  - *PART 2: Alexey OZEROV*
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*

  - *PART 3: Antoine LIUTKUS*
    - *Coding-based informed source separation*

  - *PART 4: Gaël RICHARD*
    - Conclusion

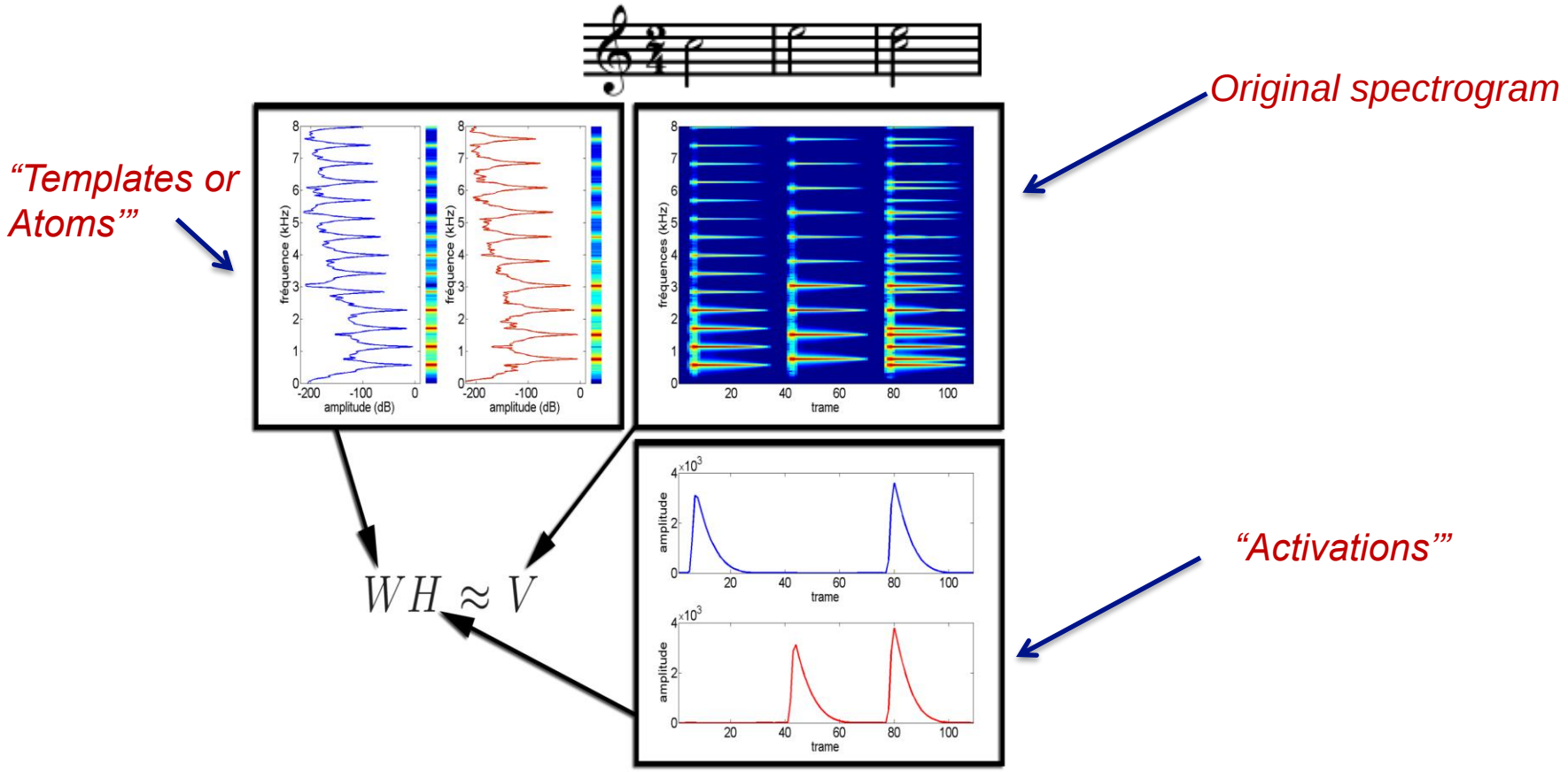
Source separation by filtering techniques

- **General principle :**
  - The sources are recovered by filtering the mixtures (for example by Wiener filtering)

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

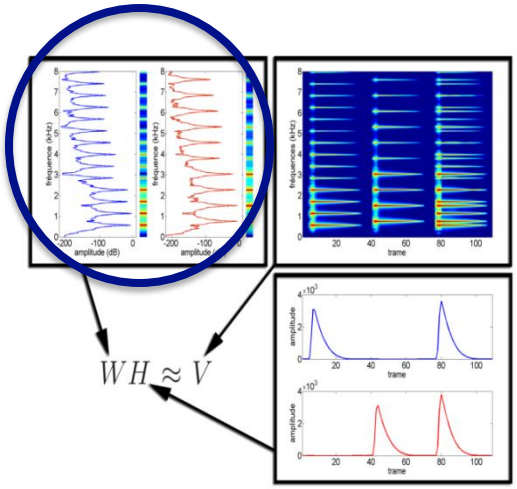
A popular model for audio source separation : NMF

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



A popular model for audio source separation : NMF

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



- **Minimization of**

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

- **Typical distances and divergences used:**

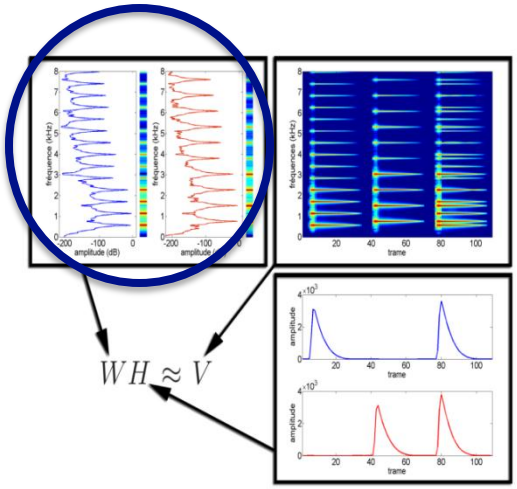
Euclidean	$d_{EUC}(a\|b) = (a-b)^2$
Kullback-Leibler divergence	$d_{KL}(a\|b) = a\log(\frac{a}{b}) - a + b$
Itakura-Saito divergence	$d_{IS}(a\|b) = \frac{a}{b} - \log(\frac{a}{b}) - 1$

β-divergence

$$d_\beta(a|b) = \begin{cases} \frac{1}{\beta(\beta-1)}(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1} & \beta \in \mathbb{R} \setminus \{0,1\} \\ a\log\frac{a}{b} + (b-a) & \beta = 1 \\ \frac{a}{b} - \log\frac{a}{b} - 1 & \beta = 0 \end{cases}$$

A popular model for audio source separation : NMF

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



- In general, the cost function is not convex in (**W,H**)…. However, it is **separately convex** in **W** and **H** (for Euclidean and Kullback-Leibler divergence)

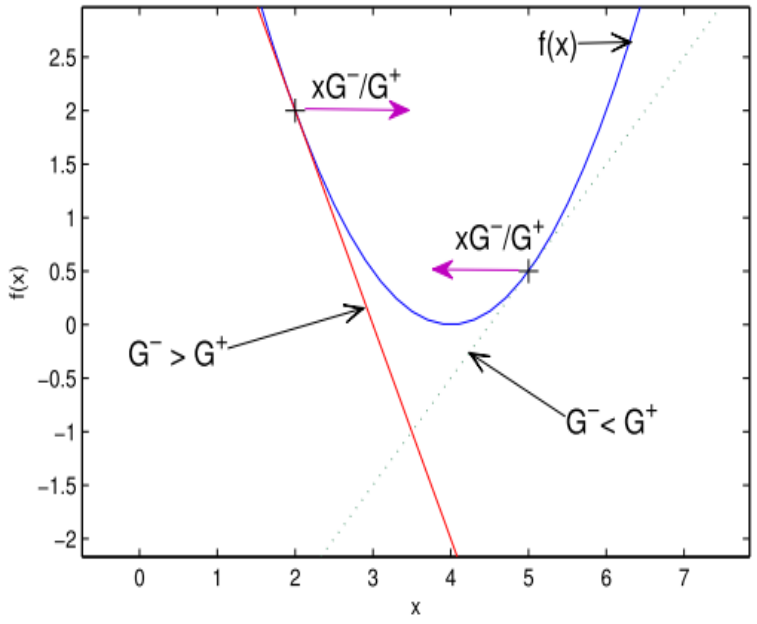
- **The solution is iteratively obtained by means of multiplicative update rules:**

  - **For example with the Euclidean distance:**

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

One way to obtain these update rules

- First, express the gradient of the cost function as $\frac{\delta f}{\delta x} = G^- - G^+$ where G^- and G^+ are positive terms

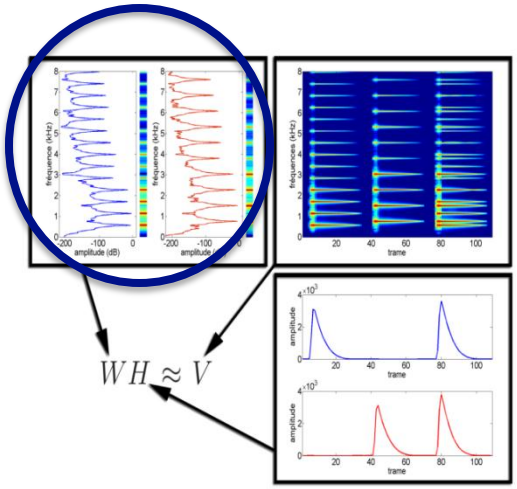


- Then the following update rules guarantees the decrease of the cost function *(under some restrictions and for some "distances" such as Euclidean and Kullback-Leibler)*

$$\begin{cases} \mathbf{W} \leftarrow \mathbf{W} \odot \frac{G^-_{\mathbf{W}} D(\mathbf{V}|\mathbf{WH})}{G^+_{\mathbf{W}} D(\mathbf{V}|\mathbf{WH})} \\ \mathbf{H} \leftarrow \mathbf{H} \odot \frac{G^-_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})}{G^+_{\mathbf{H}} D(\mathbf{V}|\mathbf{WH})} \end{cases}$$

A popular model for audio source separation : NMF

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

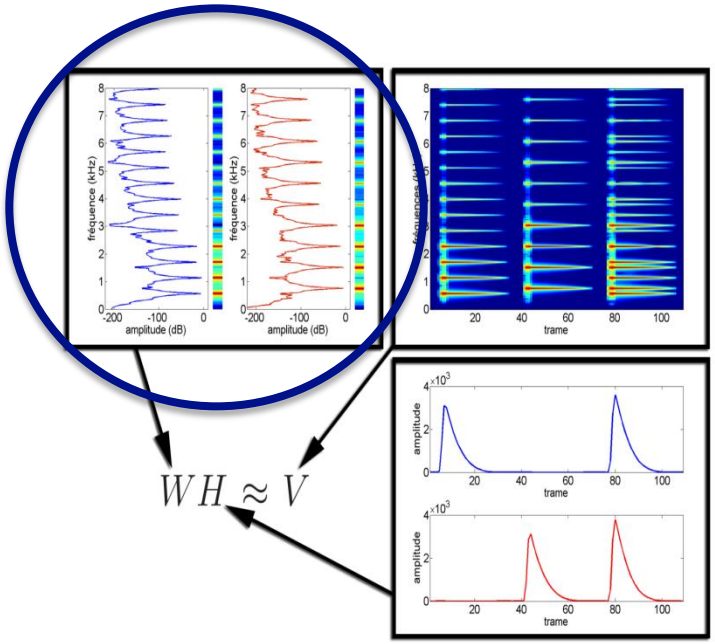


- Properties of such multiplicative update rules:
  - Associated cost function monotonously decreases along iterations
  - Non-negativity of the different coefficients is guaranteed

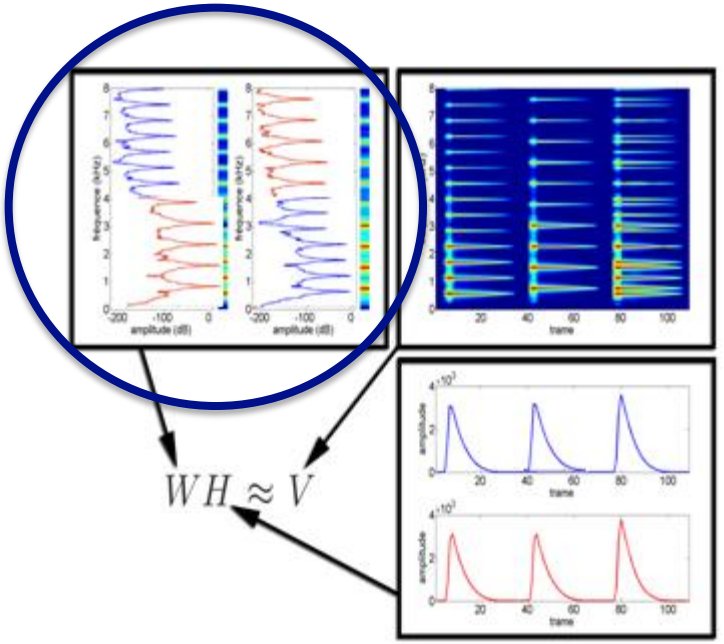
A popular model for audio source separation : NMF

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

Templates correspond to musical notes

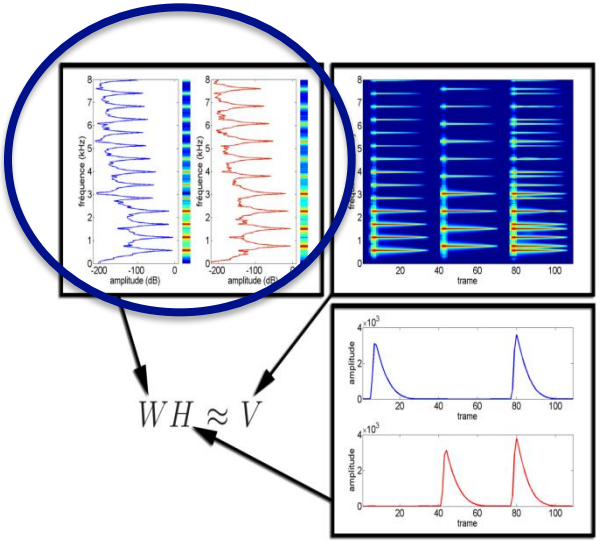


•Templates are built from half of each note and are less semantically meaningful
• Activations are less sparse
• Templates grouping for source recovery



A popular model for audio source separation : NMF

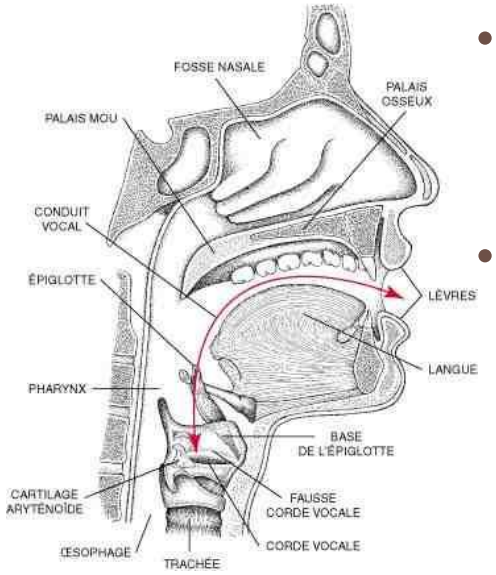
- **What types of constraints can be used ?**



- **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

An example of model-based constraints for main melody separation using NMF

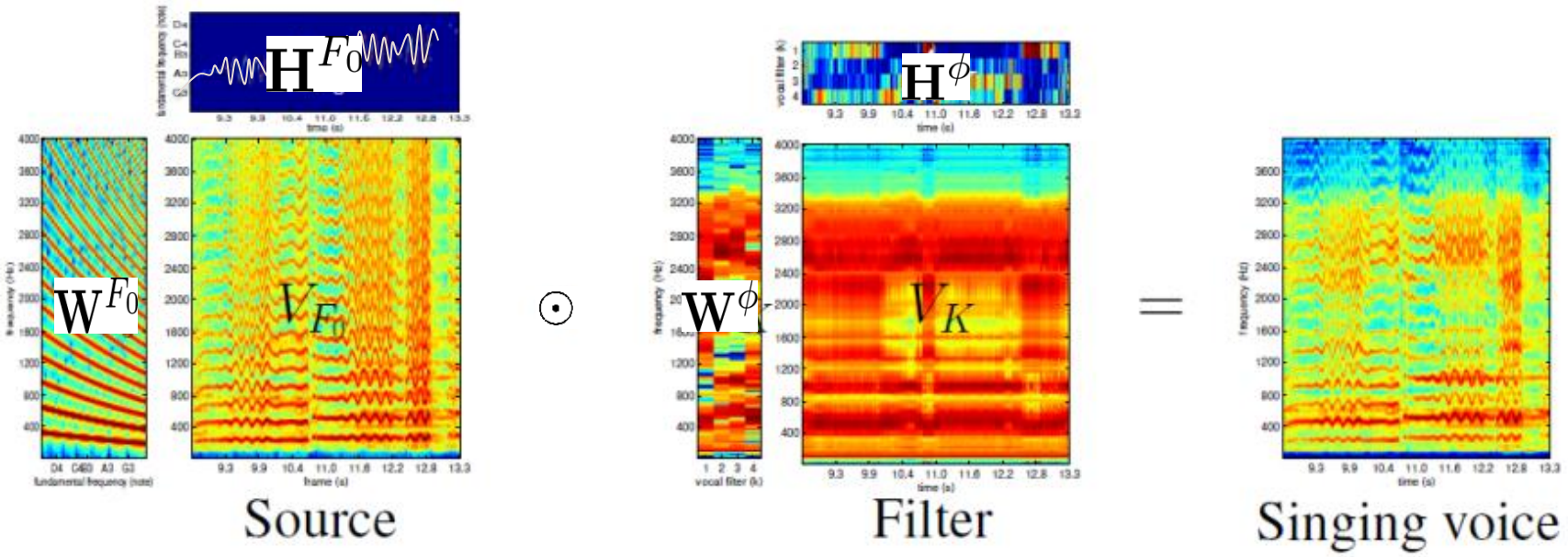
- The model: **Audio = Voice + Music**
  - The voice **V**oice follows a source filter production model : **V**oice = **S**ource * **F**ilter
  - Each component (Voice and Music) is represented by separate NMF



$$\underbrace{\mathbf{S}_{Audio}}_{\text{Spectrogram of the input audio signal}} = \underbrace{(\mathbf{W}^{F_0}\mathbf{H}^{F_0}) \odot (\mathbf{W}^{\phi}\mathbf{H}^{\phi})}_{\text{Spectrogram of the singing voice}} + \underbrace{(\mathbf{W}^{M}\mathbf{H}^{M})}_{\text{Spectrogram of the background music}}$$

An example of model-based constraints for main melody separation using NMF

- **Illustration of the source/filter model with NMF**



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 (Parts 2 & 3 of this tutorial):
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*
    - *Coding-based informed source separation*

Keynote content

- **Objective**
  - To provide an overview of major trends in Informed Source Separation (ISS)
- **Outline of the keynote**
  - *PART 1: Gaël RICHARD*
    - Introduction on Informed Source Separation
    - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
  - ***PART 2: Alexey OZEROV***
    - ***Auxiliary data-informed source separation,***
    - ***User-guided source separation,***
  - *PART 3: Antoine LIUTKUS*
    - *Coding-based informed source separation*
  - *PART 4: Gaël RICHARD*
    - Conclusion

Keynote content

- **Objective**
  - To provide an overview of major trends in Informed Source Separation (ISS)
- **Outline of the keynote**
  - *PART 1: Gaël RICHARD*
    - Introduction on Informed Source Separation
    - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
  - ***PART 2: Alexey OZEROV***
    - ***Auxiliary data-informed source separation,***
    - ***User-guided source separation,***
  - *PART 3: Antoine LIUTKUS*
    - *Coding-based informed source separation*
  - *PART 4: Gaël RICHARD*
    - Conclusion

Keynote content

- **Objective**
  - To provide an overview of major trends in Informed Source Separation (ISS)
- **Outline of the keynote**
  - *PART 1: Gaël RICHARD*
    - Introduction on Informed Source Separation
    - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
  - ***PART 2: Alexey OZEROV***
    - ***Auxiliary data-informed source separation,***
    - ***User-guided source separation,***

NMF-based

Single-channel case

  - *PART 3: Antoine LIUTKUS*
    - *Coding-based informed source separation*
  - *PART 4: Gaël RICHARD*
    - Conclusion

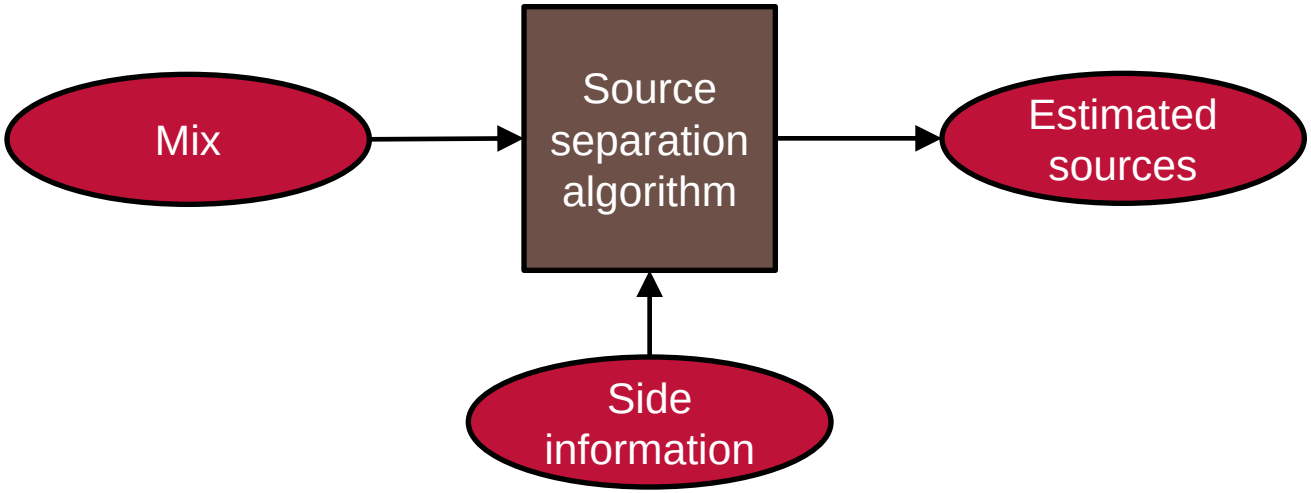
Overview of three trends in ISS

Auxiliary data-informed source separation,

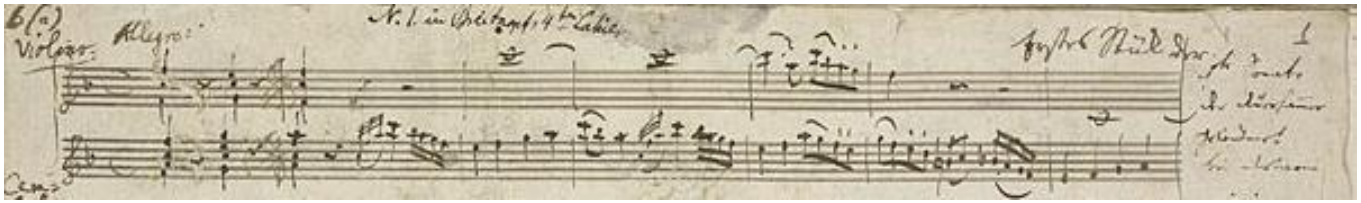
User-guided source separation,

Coding-based informed source separation

Auxiliary data-informed source separation



Existing side information, e.g., musical score

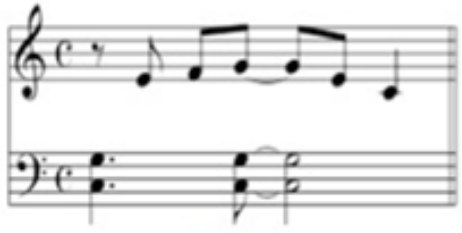


Auxiliary data-informed source separation
as multimodal source separation

Side information:

Different modalities

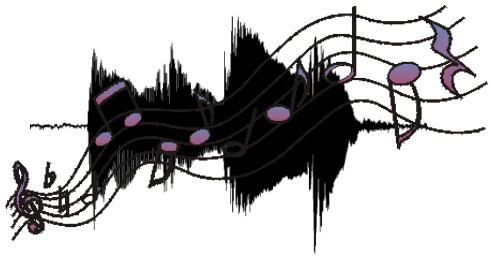
Musical score:



Text:

I remember him looking round the cover and whistling to himself as he did so, and then breaking out in that old sea-song that he sang so often afterwards:

Cover version:



Multilingual audio:

Video:

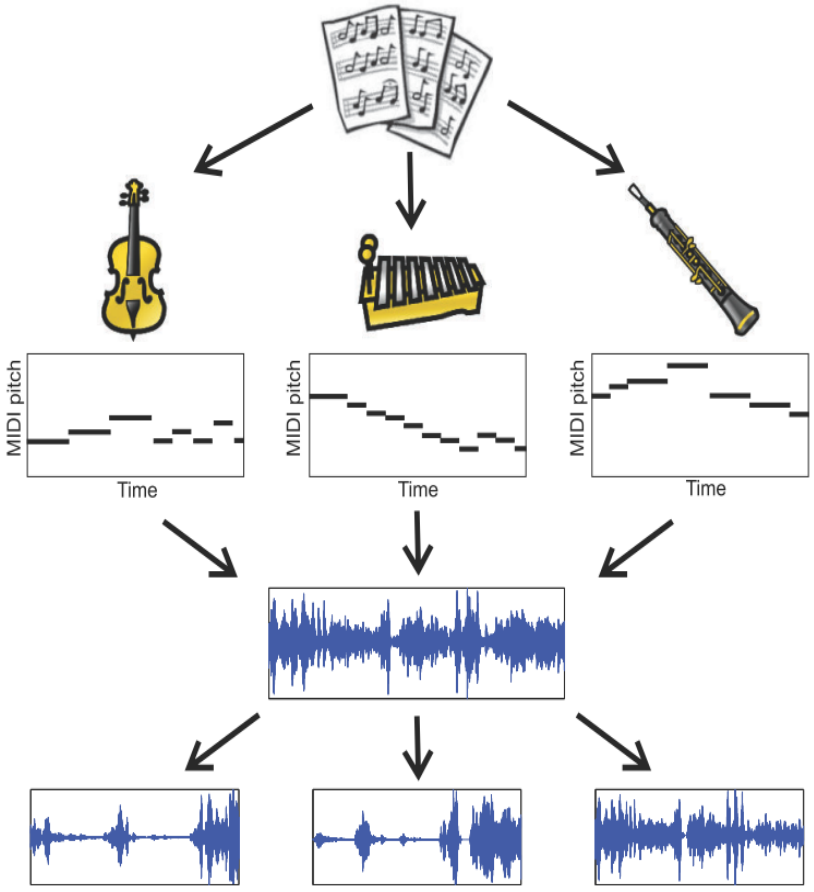
Auxiliary data-informed source separation
main approaches

- **Goal: use side information to constraint the NMF decomposition**

- **Main approaches**
  - **Enforced sparsity:** Setting to zero some NMF model values based on side information
  - **Synthesis-based:** Convert side information into the same modality (audio) via an appropriate synthesizer and then use it as an example to constraint the NMF
  - **Joint modeling:** Use the side information as it is within a joint framework

Auxiliary data-informed source separation

"Score-informed" source separation (enforced sparsity)



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 for music signals. In Multimodal Music Processing, Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

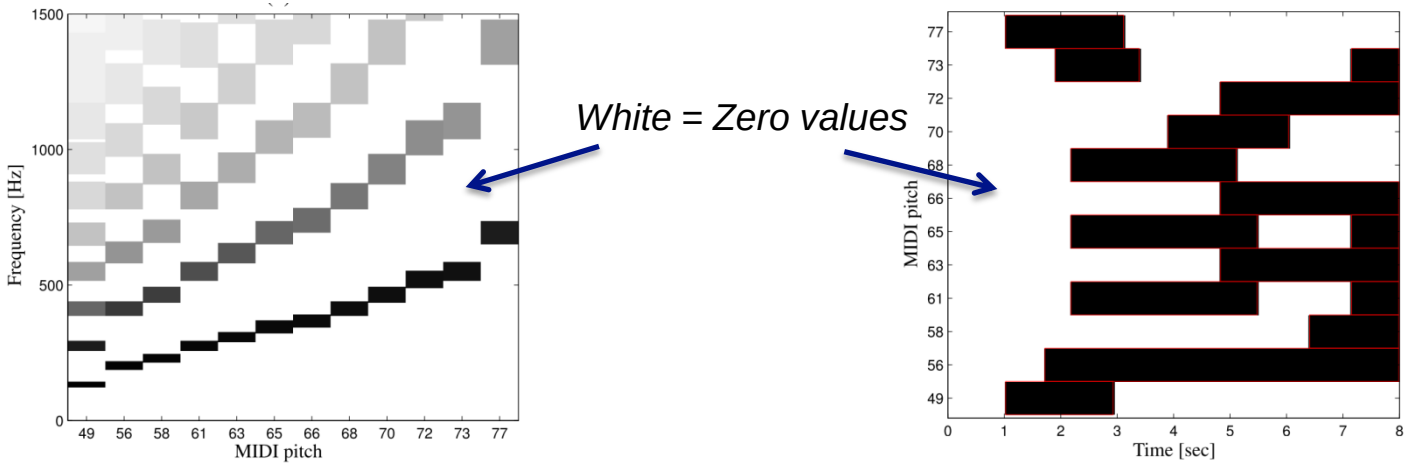
Auxiliary data-informed source separation

"Score-informed" source separation (enforced sparsity)

- **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



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

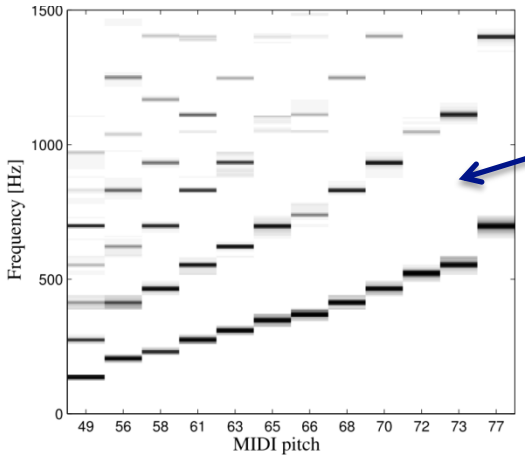
Auxiliary data-informed source separation

"Score-informed" source separation (enforced sparsity)

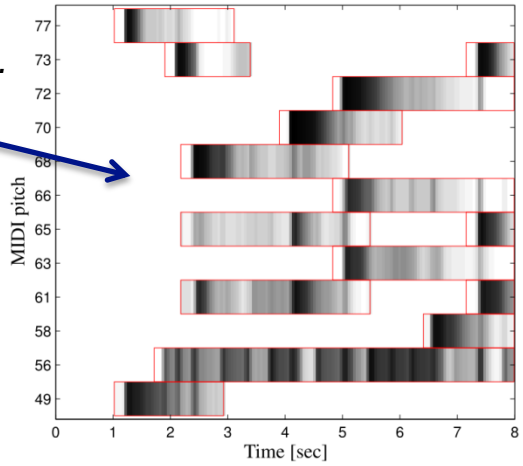
- **An example in the framework of NMF (V= W . H)**

Matrix W: obtained after convergence

Matrix H: obtained after convergence



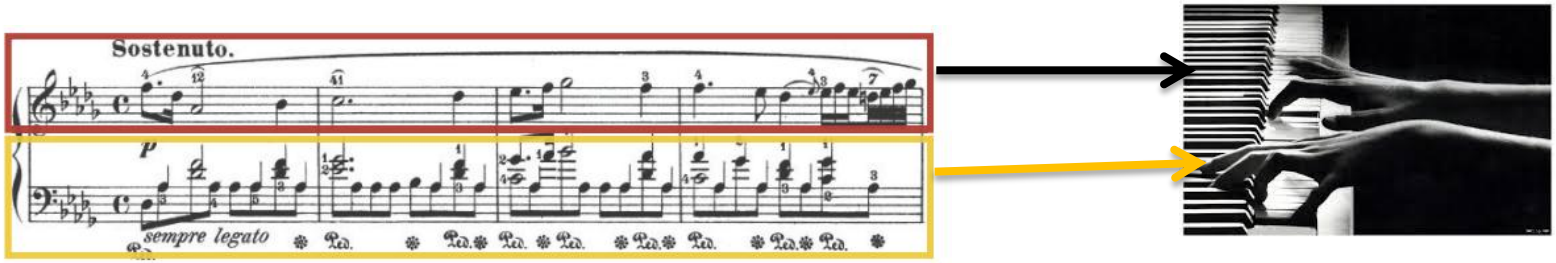
Null entries at init. remain null



Auxiliary data-informed source separation

"Score-informed" source separation (enforced sparsity)

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



Original recording (Chopin)

Left hand

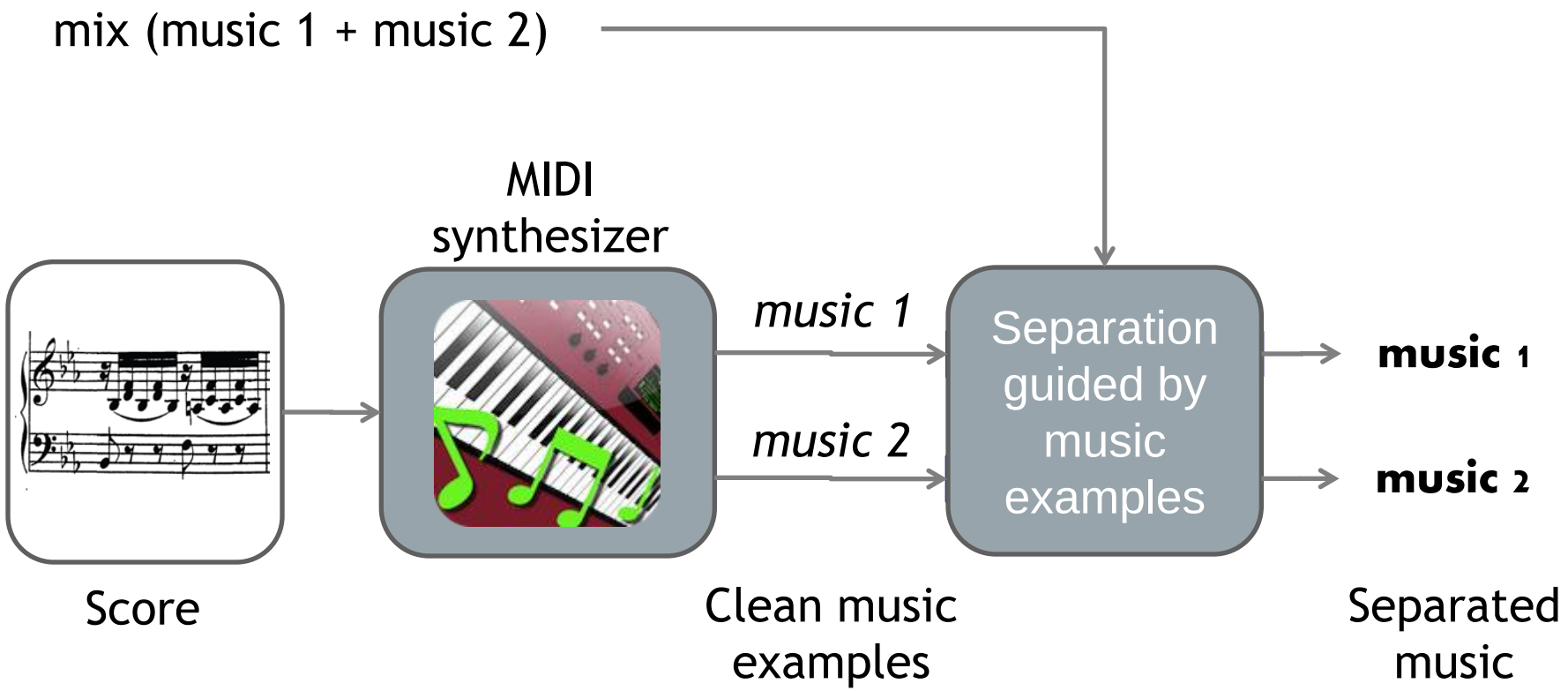
Right hand

Examples from S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in Proc. ICASSP, Kyoto, Japan, 2012, pp. 129–132.
http://www.mpi-inf.mpg.de/resources/MIR/ICASSP2012-ScoreInformedNMF/

Auxiliary data-informed source separation

"Score-informed" source separation (synthesis-based)

- **General scheme**

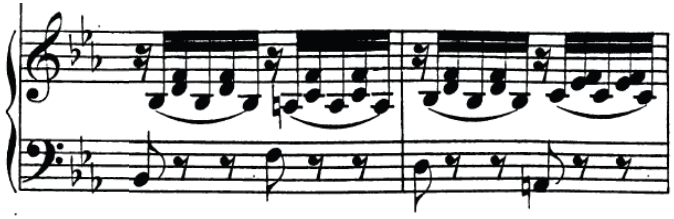


Auxiliary data-informed source separation

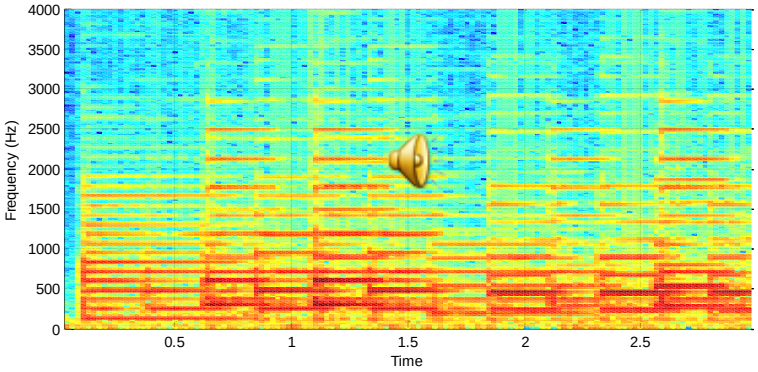
"Score-informed" source separation (synthesis-based)

- **Example of MIDI-synthesized piano example**

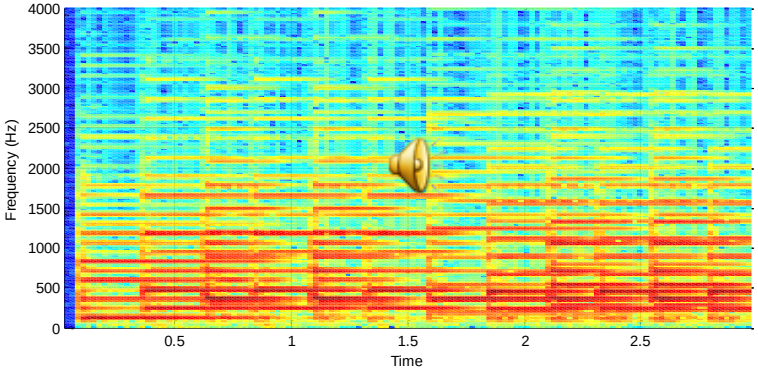
Score



Original music source



Synthesized music example

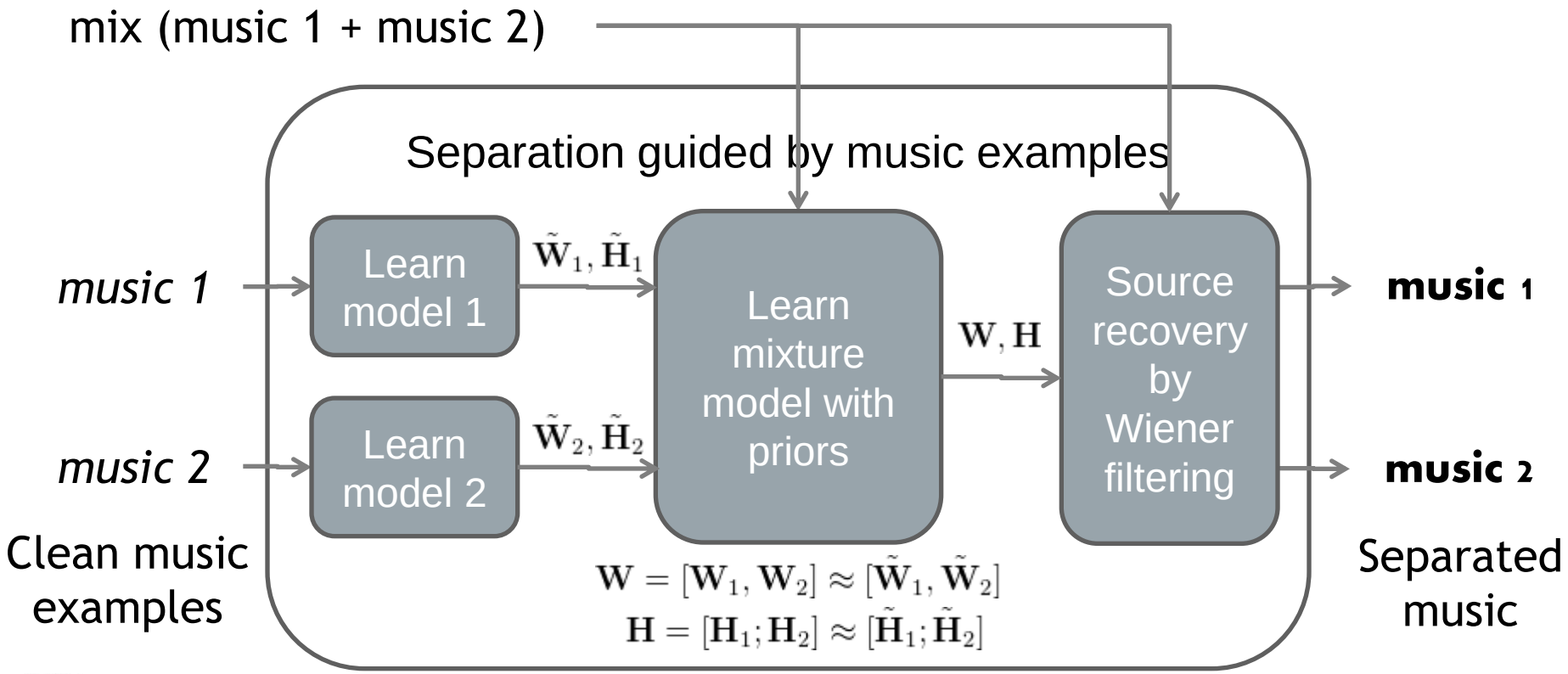


Examples from TRIOS dataset: J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," ICASSP 2013.

Auxiliary data-informed source separation

"Score-informed" source separation (synthesis-based)

- **An exemplar approach**
  - Examples are aligned to the mixture (e.g., by DTW)



J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders, "Source separation by score synthesis," in Proc. Int. Computer Music Conference (ICMC), New York, NY, June 2010, pp. 462–465.

Auxiliary data-informed source separation

"Score-informed" source separation (joint modeling)

- **Nonnegatve Matrix Partial Co-Factorization (NMPCF)**

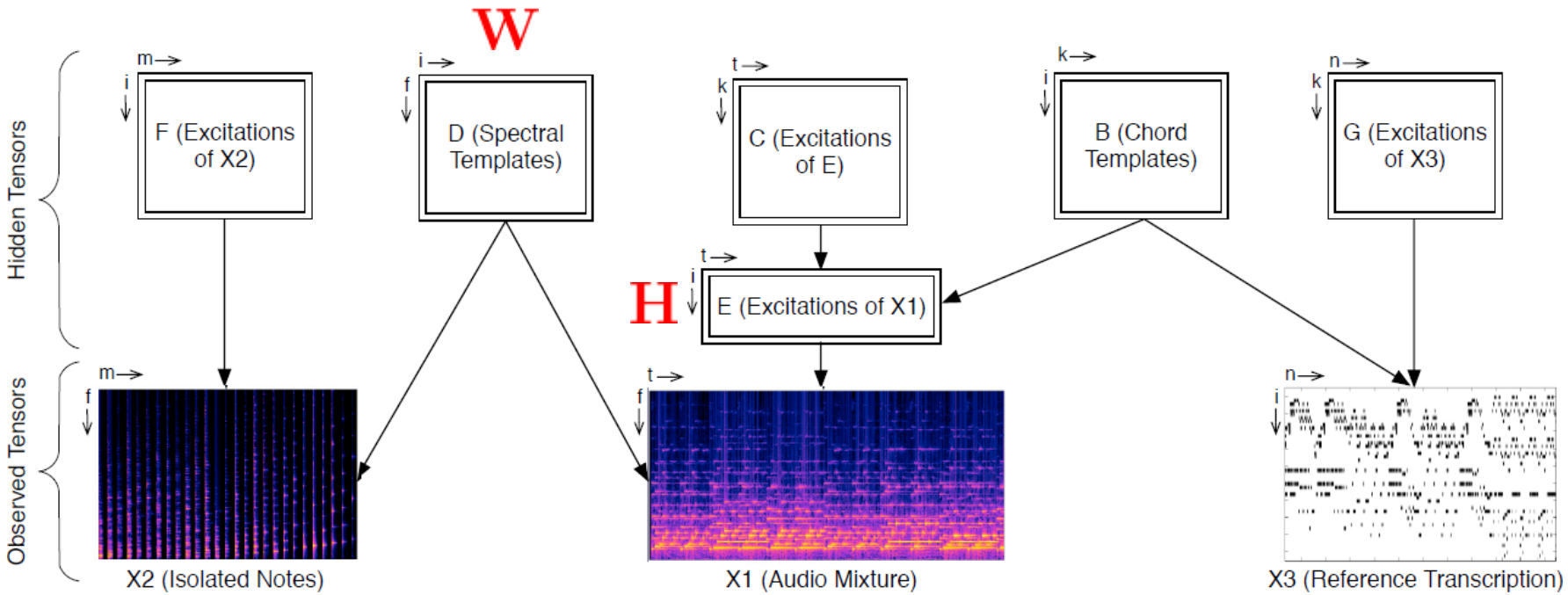
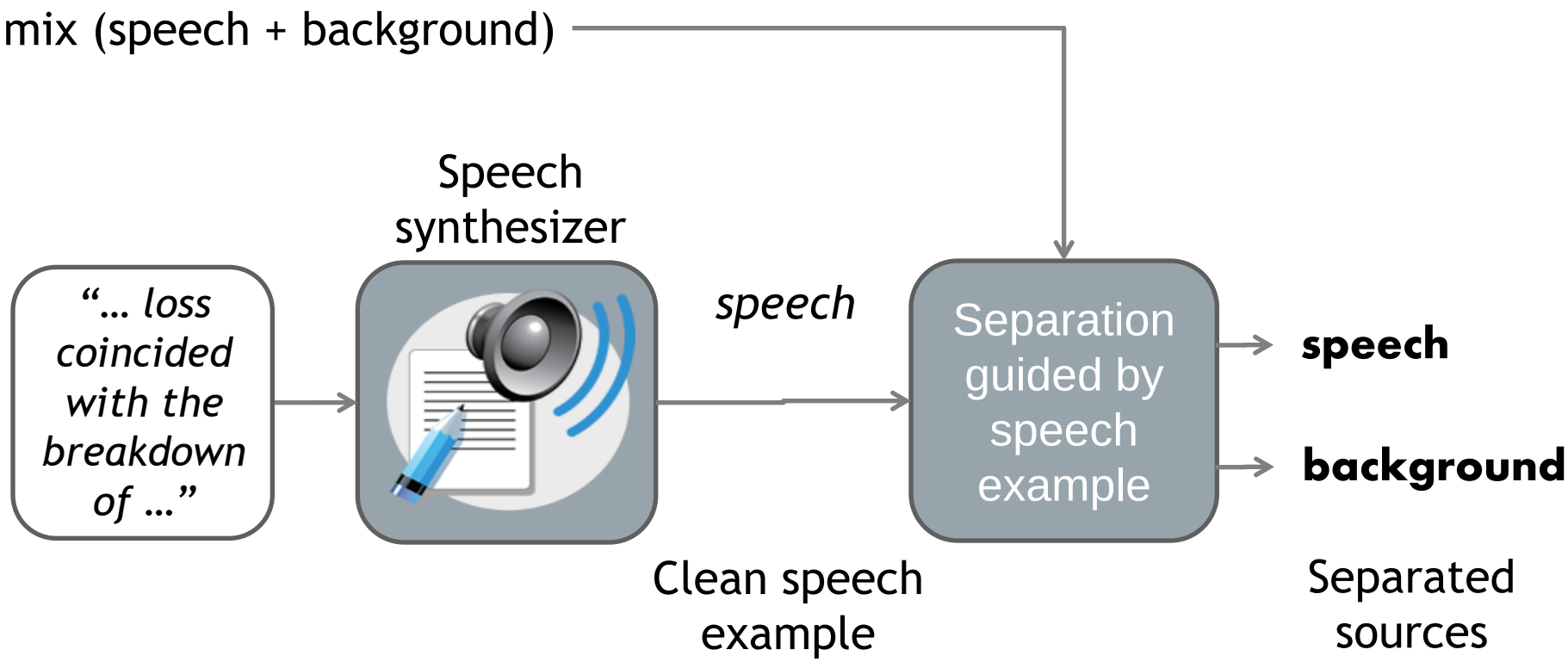


Figure from U. Simsekli and A. T. Cemgil, "Score Guided Musical Source Separation Using Generalized Coupled Tensor Factorization," in 20th European Signal Processing Conference (EUSIPCO), pp. 2639 - 2643, 2012.

Auxiliary data-informed source separation

"Text-informed" speech separation (synthesis-based)

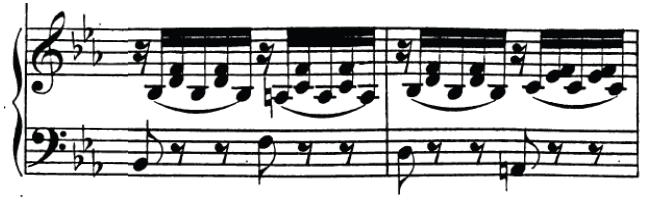
- **General scheme**



Auxiliary data-informed source separation

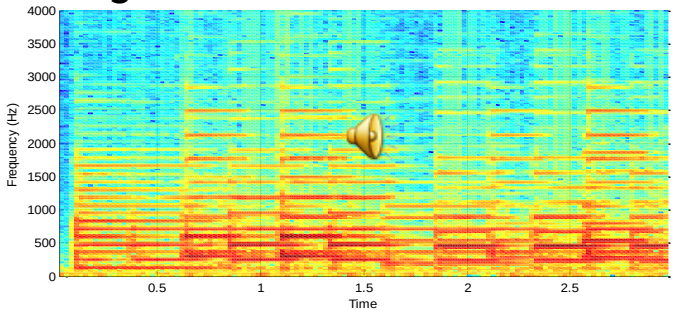
"Text-informed" speech separation (synthesis-based)

- **Speech is a particularly difficult case …**

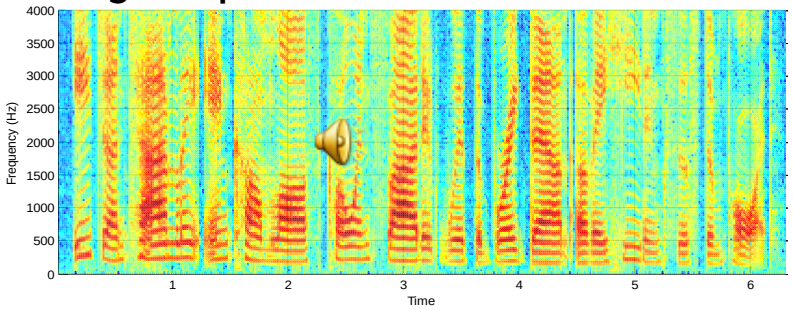


"Each untimely income loss coincided with the breakdown of a heating system part."

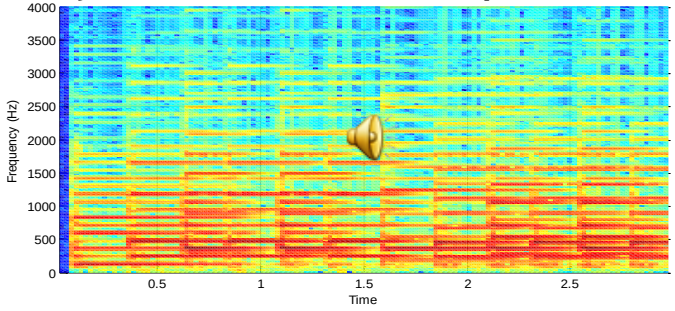
Original music source



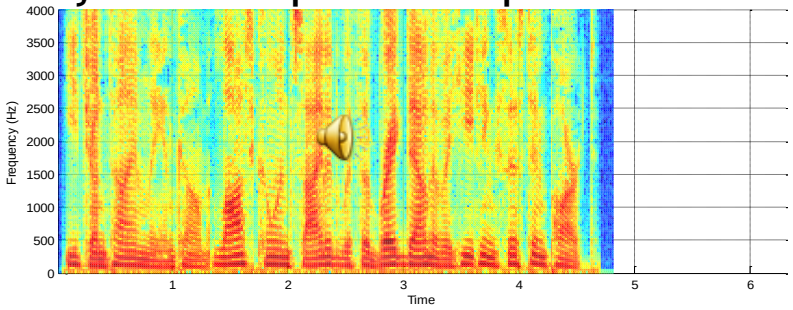
Original speech source



Synthesized music example



Synthesized speech example



Auxiliary data-informed source separation
"Text-informed" speech separation (synthesis-based)

- **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 decomposition:

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

Spectrogram of S

Spectrogram of Excitation (or source)

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

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

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 an NMF**

$$\mathbf{V}_X \approx \hat{\mathbf{V}}_X = \overbrace{\underbrace{\left(\mathbf{W}^e\mathbf{H}^e_S\right)}_{\hat{\mathbf{V}}^e_S} \odot \underbrace{\left(\mathbf{W}^\phi_S\mathbf{H}^\phi_S\right)}_{\hat{\mathbf{V}}^\phi_S} \odot \underbrace{\left(\mathbf{w}^c_S\mathbf{i}^T_N\right)}_{\hat{\mathbf{V}}^c_S}}^{\hat{\mathbf{V}}_S} + \overbrace{\mathbf{W}_B\mathbf{H}_B}^{\hat{\mathbf{V}}_B}$$

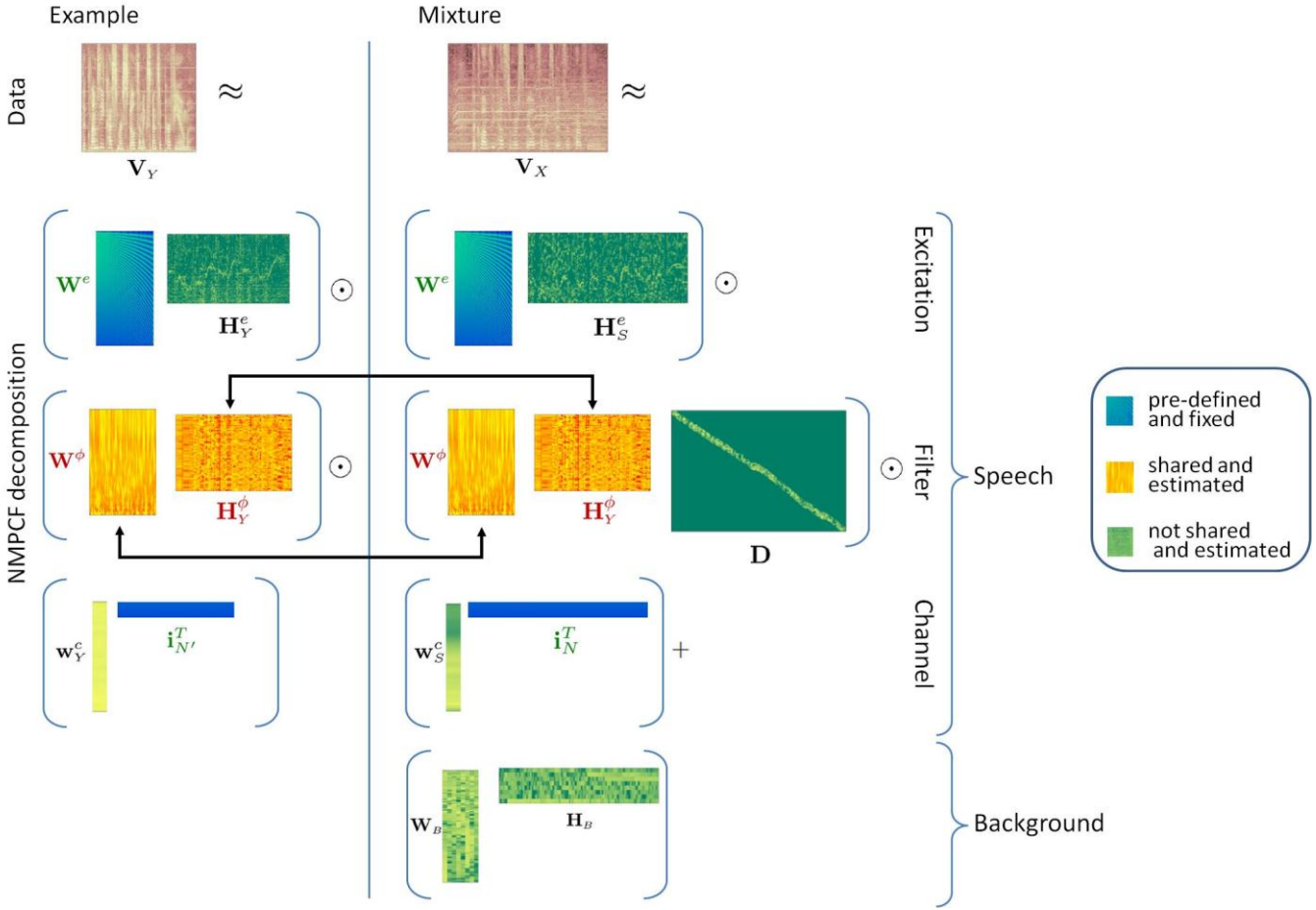
- **In this representation the text (which gives phonetic information) will directly give information on the matrix linked to what is said, which is:** $\hat{\mathbf{V}}^\phi_S$

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

- **Nonnegatve Matrix Partial Co-Factorization (NMPCF)**

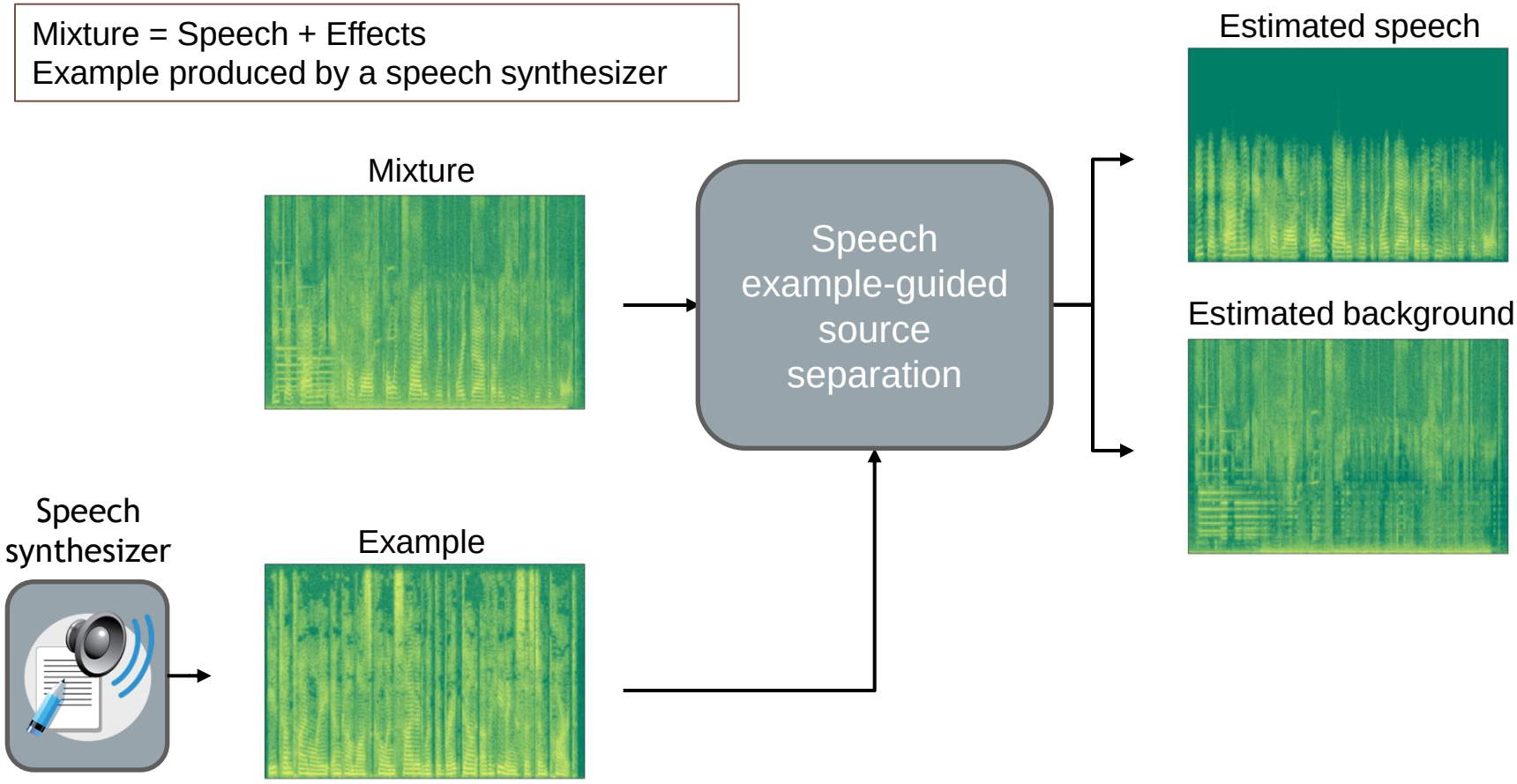


L. Le Magoarou, A. Ozerov, N. Duong Text-Informed Audio Source Separation using Nonnegative Matrix Partial Co-Factorization, in Proc. of MLSP, 2013

Auxiliary data-informed source separation

"Text-informed" speech separation : demonstration

Mixture = Speech + Effects
Example produced by a speech synthesizer

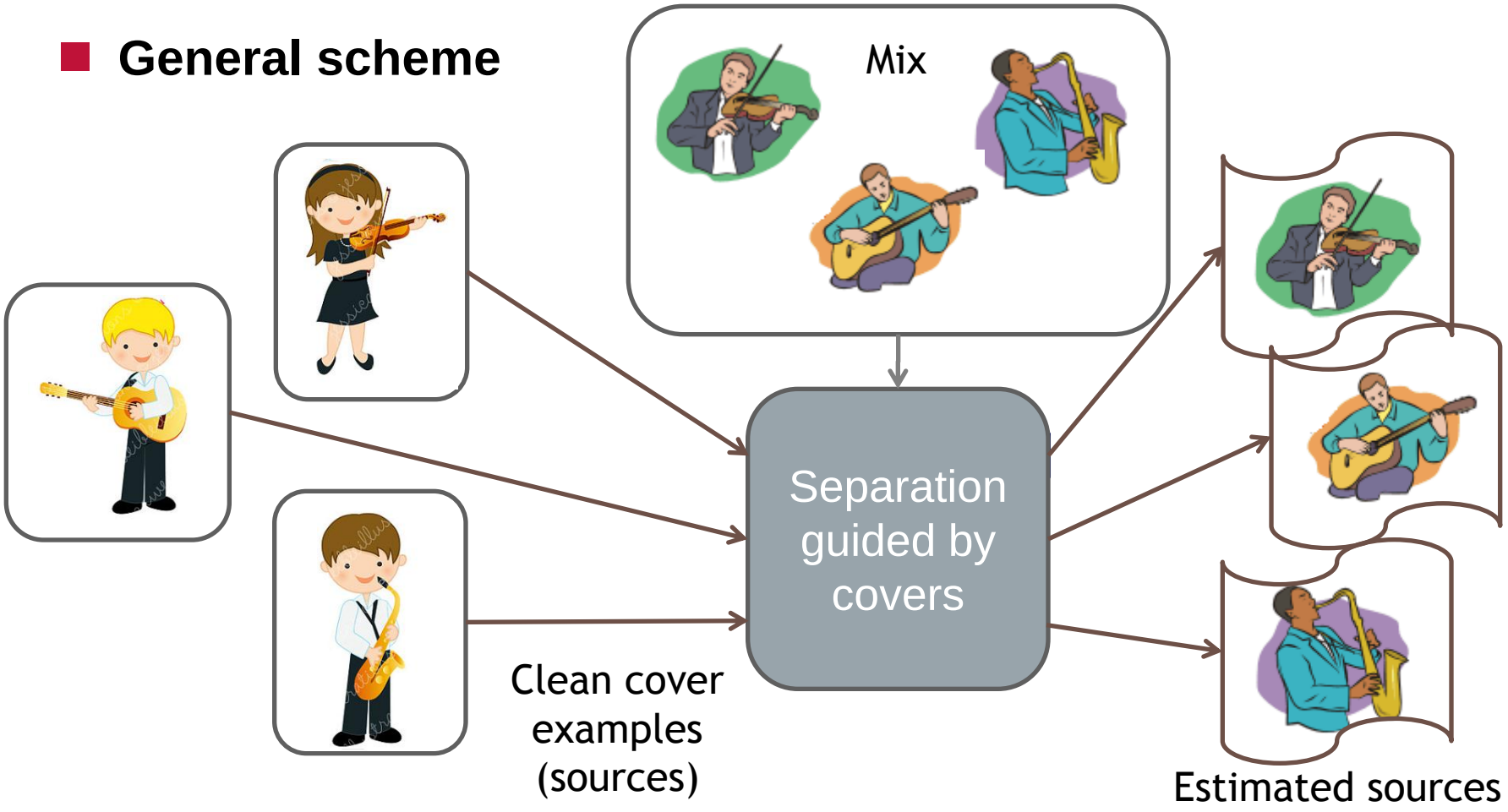


L. Le Magoarou, A. Ozerov and N. Duong "Text-Informed Audio Source Separation using Nonnegative Matrix Partial Co-Factorization", in Proc. of MLSP, 2013

Auxiliary data-informed source separation

"Cover-informed" source separation

- **General scheme**



T. Gerber, M. Dutasta, L. Girin and C. Févotte, "Professionally-produced music separation guided by covers," in Proc. 13th Int. Society for Music Information Retrieval Conf., 2012, pp. 85–90

Auxiliary data-informed source separation
"Cover-informed" source separation

- **Cover examples and source separation results**

  - Not these kind of covers (*but, why not ???*)

    Original by Gloria Gaynor

    Cover by Snuff

  - Rather these kind of covers: Rocket Man by Elton John

Original (mix)	Cover	Cover sources	Estimated sources
		Drums	Drums
		Piano	Piano

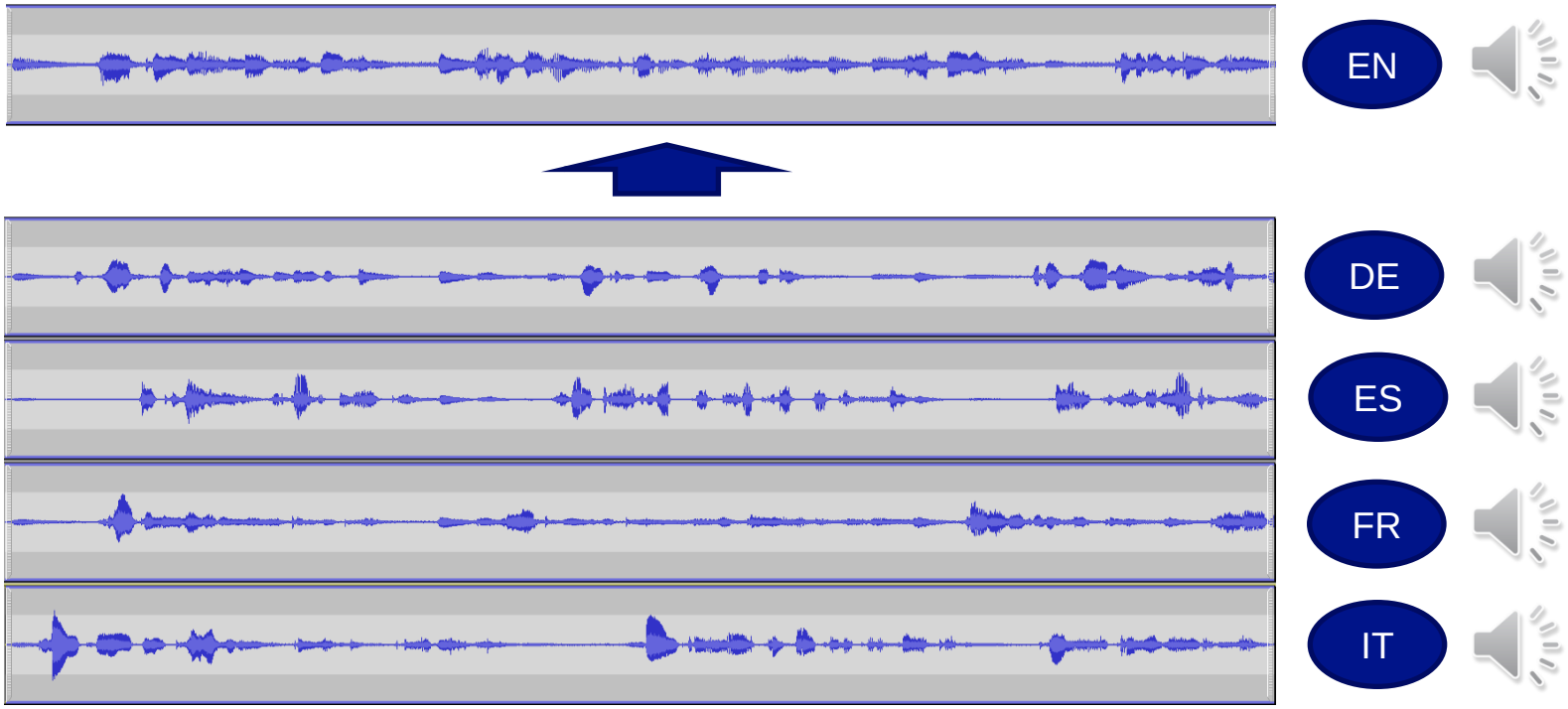
Examples below are from T. Gerber, M. Dutasta, L. Girin and C. Févotte, "Professionally-produced music separation guided by covers," in Proc. 13th Int. Society for Music Information Retrieval Conf., 2012, pp. 85–90
http://www.gipsa-lab.grenoble-inp.fr/~laurent.girin/demo/ismir2012.html

stop

Auxiliary data-informed source separation

"Multilingual audio-informed" source separation

- **Several international versions of the same movie**
  - Dialogs from "Hannah and her sisters" by Woody Allen

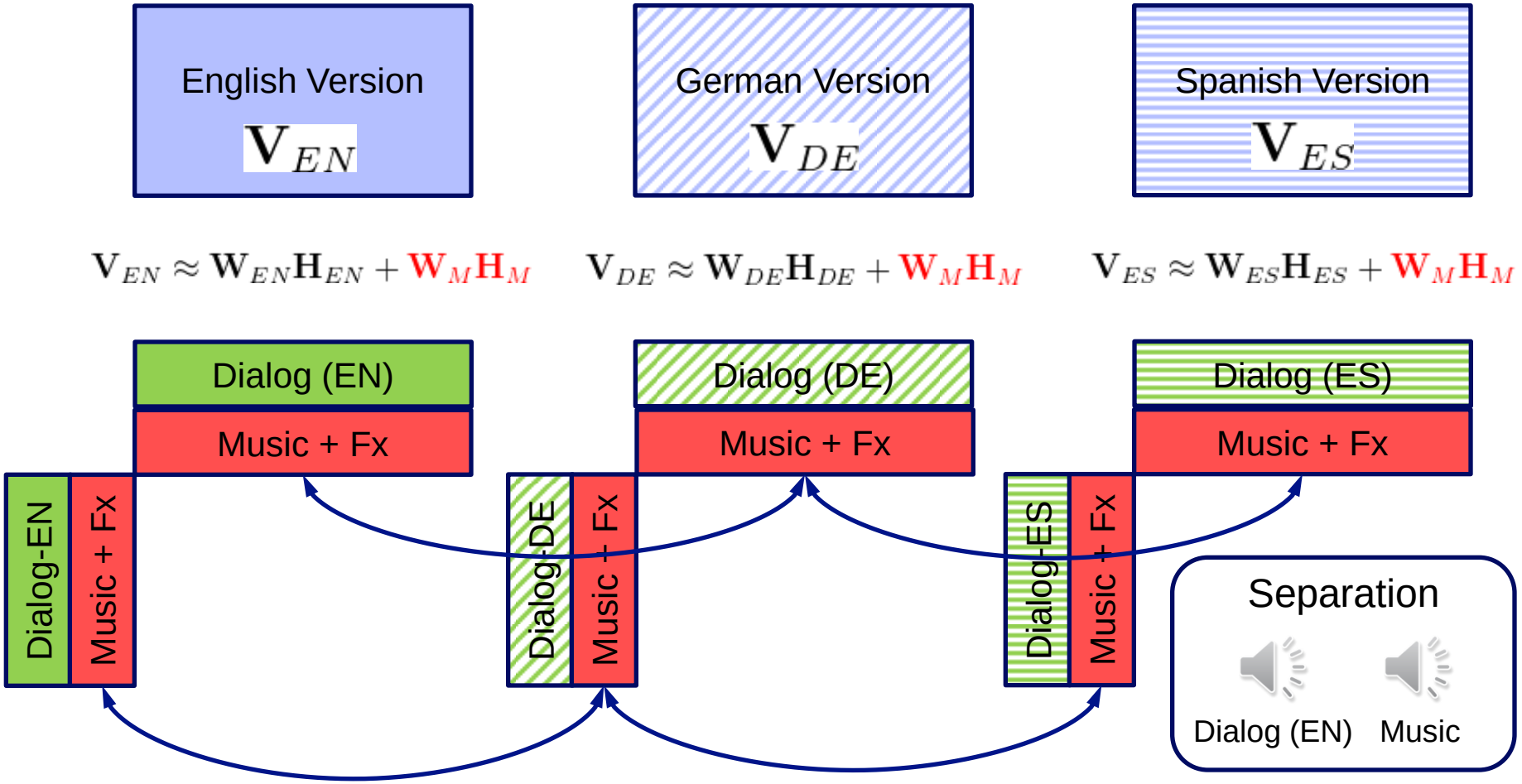


A. Liutkus and P. Leveau, "Separation of music+effects sound track from several international versions of the same movie," in Proc. 128th Audio Eng. Soc. Conv., London, U.K., May 22–25, 2010.

stop

Auxiliary data-informed source separation

"Multilingual audio-informed" source separation



P. Leveau, J. J. Burred, S. Maller and X. Jaureguiberry "Convolutive common audio signal separation," In Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011.

stop

Auxiliary data-informed source separation
"Video-informed" source separation

- **Synchronize audio-visual objects**
- **Identify regions of audio source activity**
- **Learn source models from "active-alone" regions**

Mix | **Estimated drums** | **Estimated guitar**

Results from A. Llagostera Casanovas, G. Monaci, P. Vandergheynst and R. Gribonval, Blind Audio-Visual Source Separation based on Sparse Redundant Representations, IEEE Transactions on Multimedia, Vol. 12, Nr. 5, pp. 358-371, 2010. http://www.eecs.qmul.ac.uk/~llagostera/research.htm

Overview of three trends in ISS

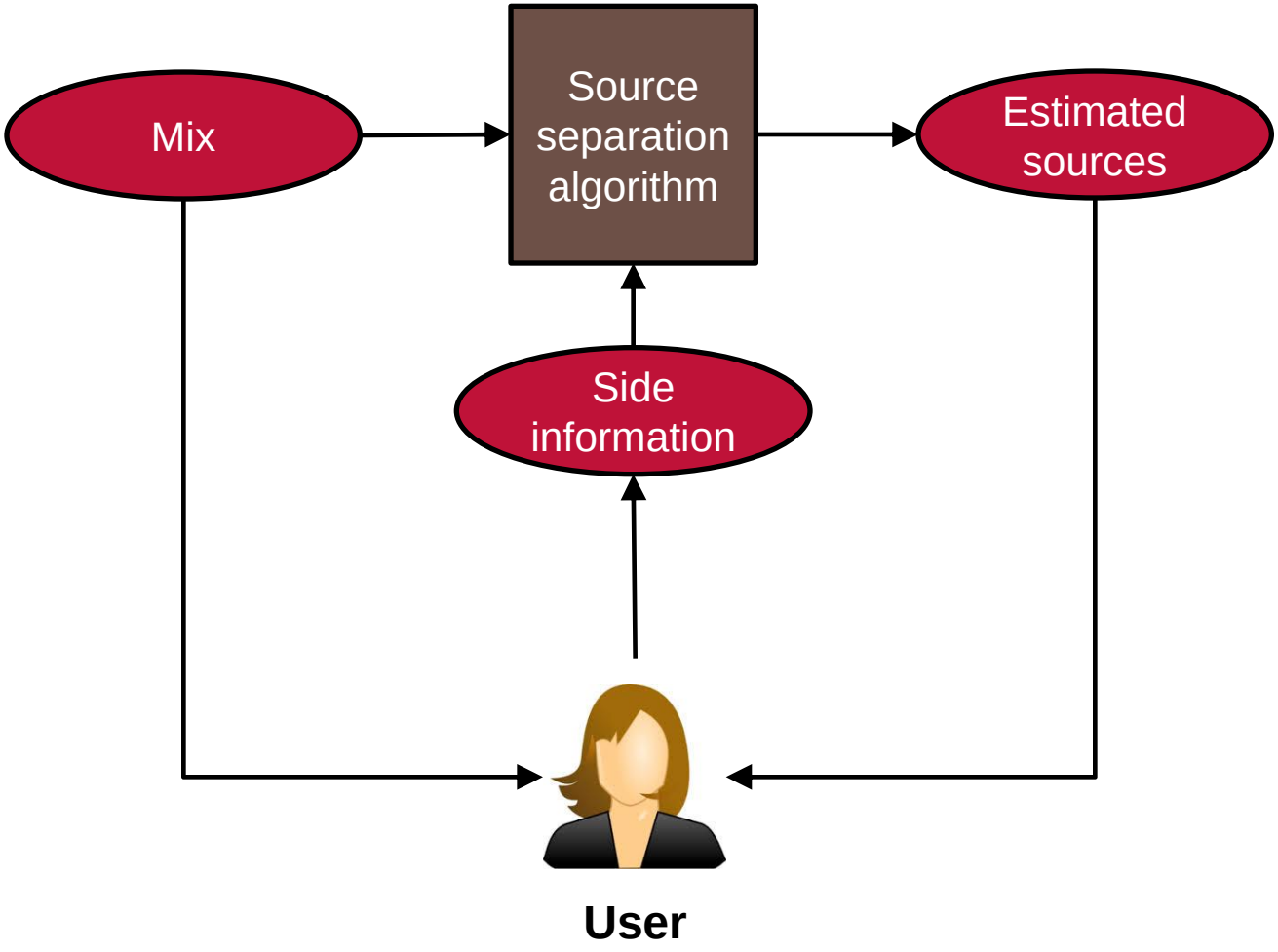
Institut
Mines-Télécom

Auxiliary data-informed source separation,

User-guided source separation,

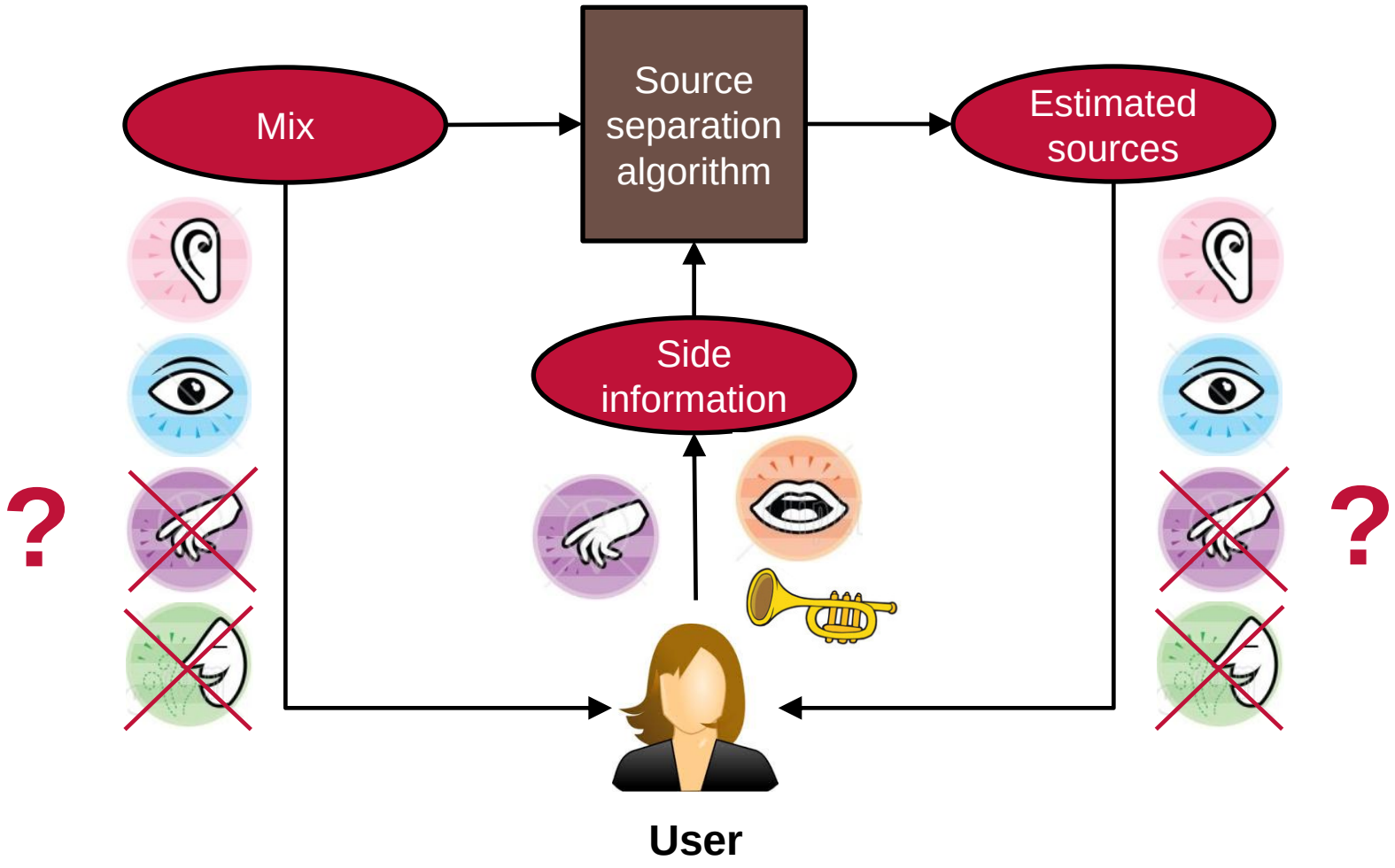
Coding-based informed source separation

User-guided source separation



User-guided source separation

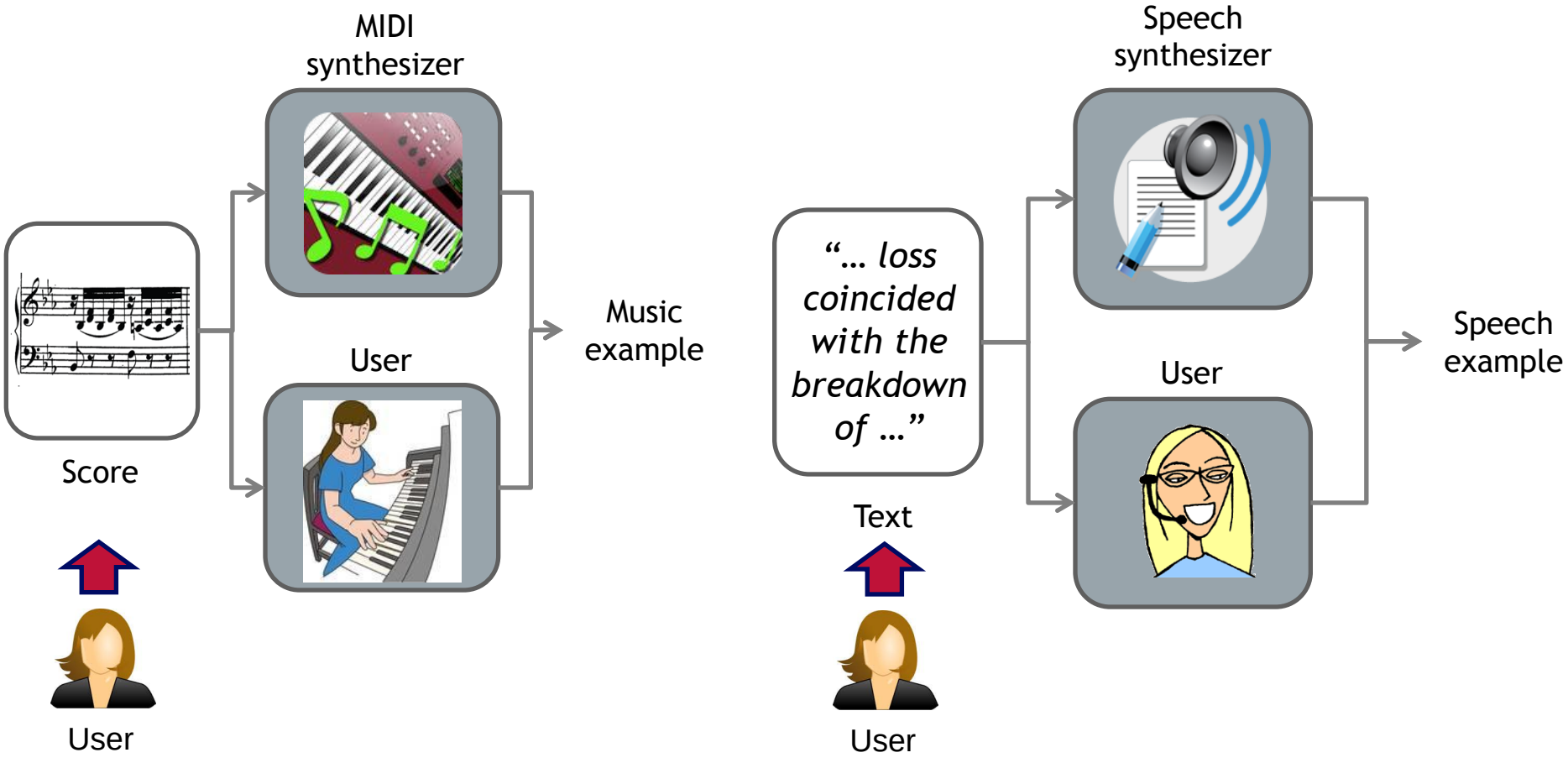
How does a user receive or transmit information?



User-guided source separation

The border between "auxiliary data-informed" and "user-guided" approaches is fuzzy

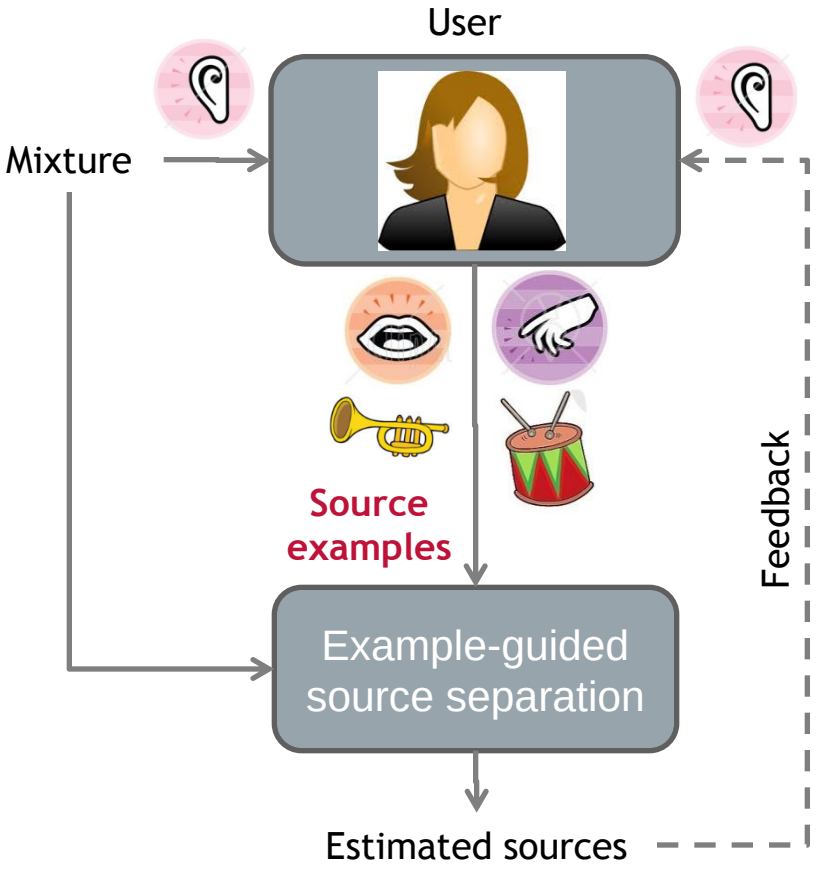
Synthesis-based vs. example-based approaches



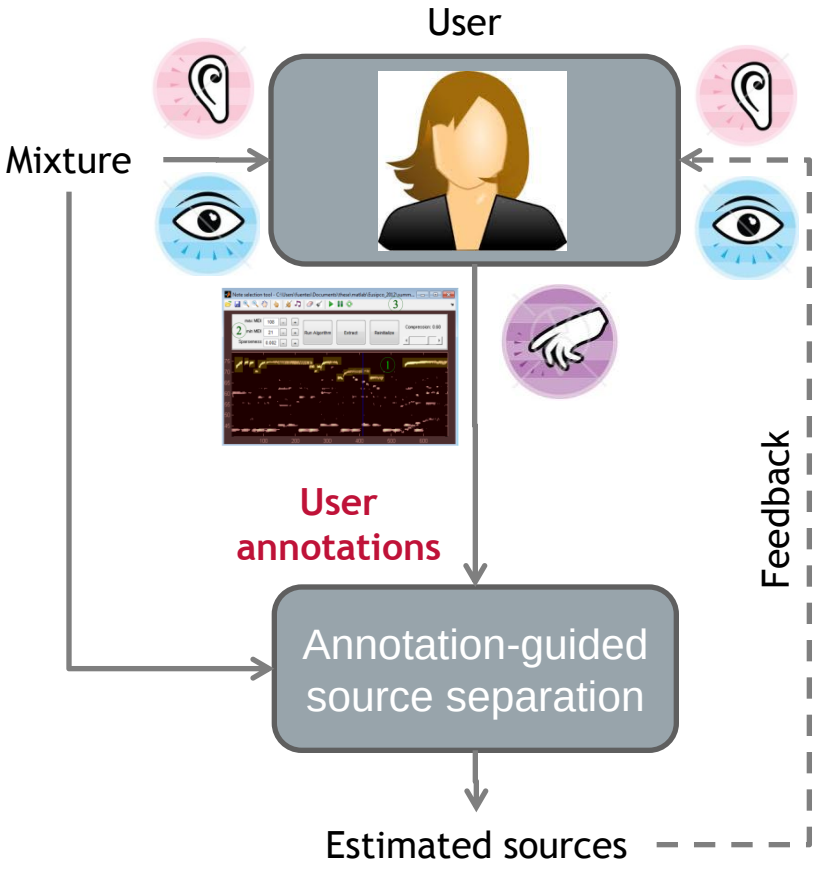
User-guided source separation

Main approaches

Example-based approaches



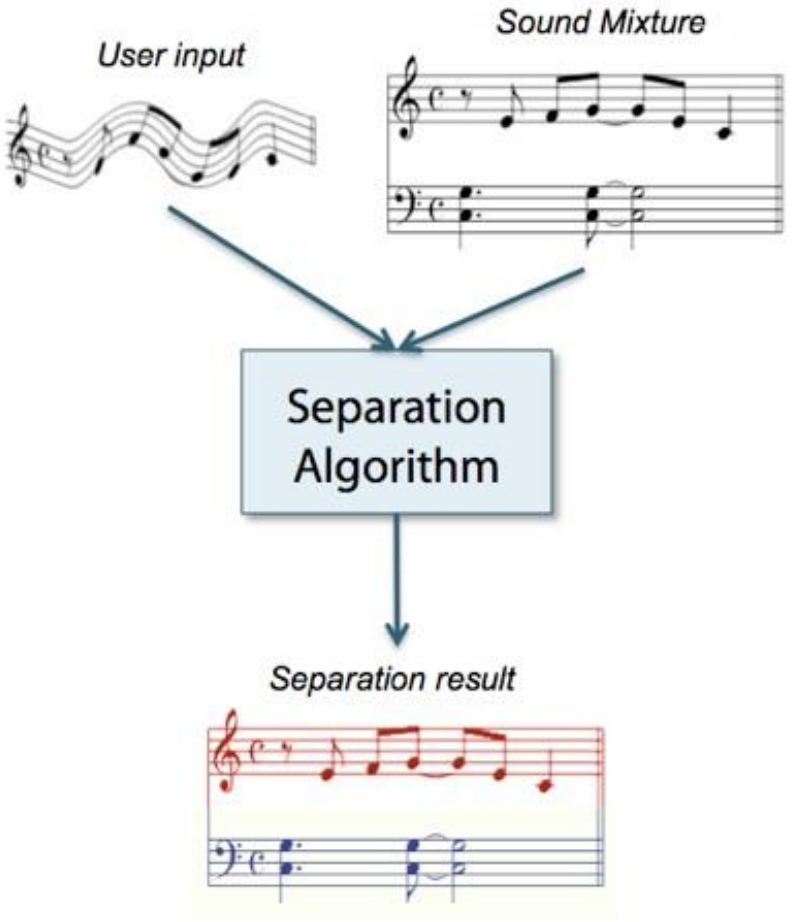
Annotation-based approaches



User-guided source separation
Example-based approaches

Separation by humming

- The user hums trying to mimic the target source that he/she wishes to separate, e.g., a music instrument or a cellphone ringing
- The hummed example is then used as information for separating the target source in the mixture

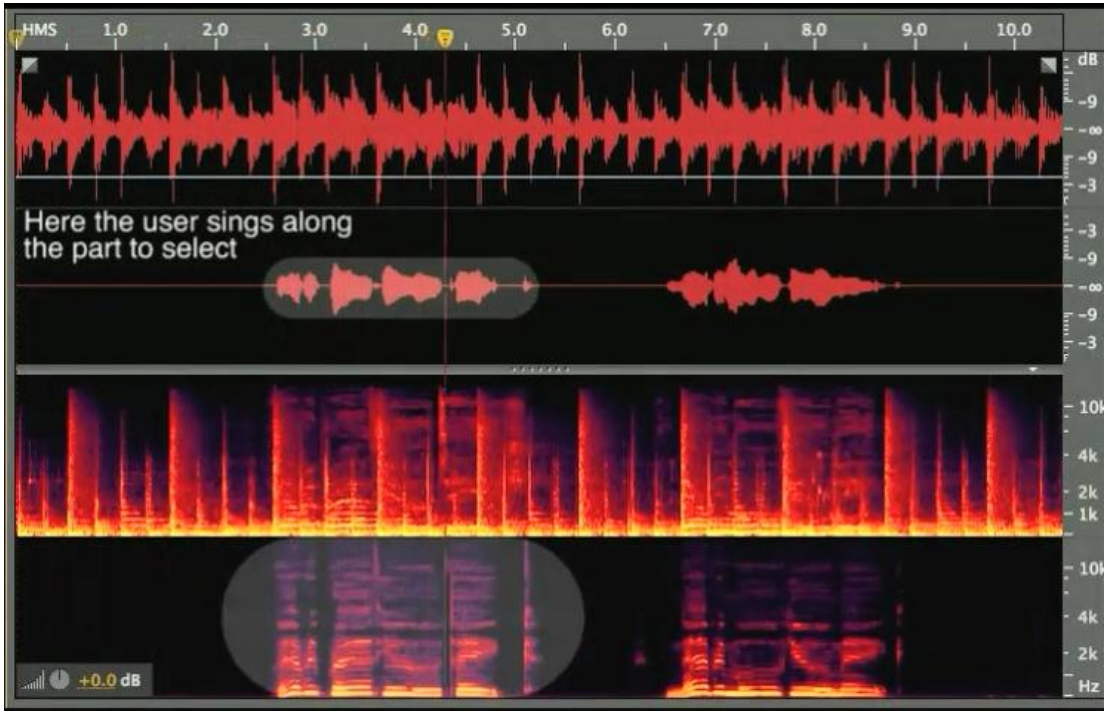


From https://ccrma.stanford.edu/~gautham/Site/Humming.html

User-guided source separation
Example-based approaches

Separation by humming

- **Demonstration video [Smaragdis & Mysore 2009]**



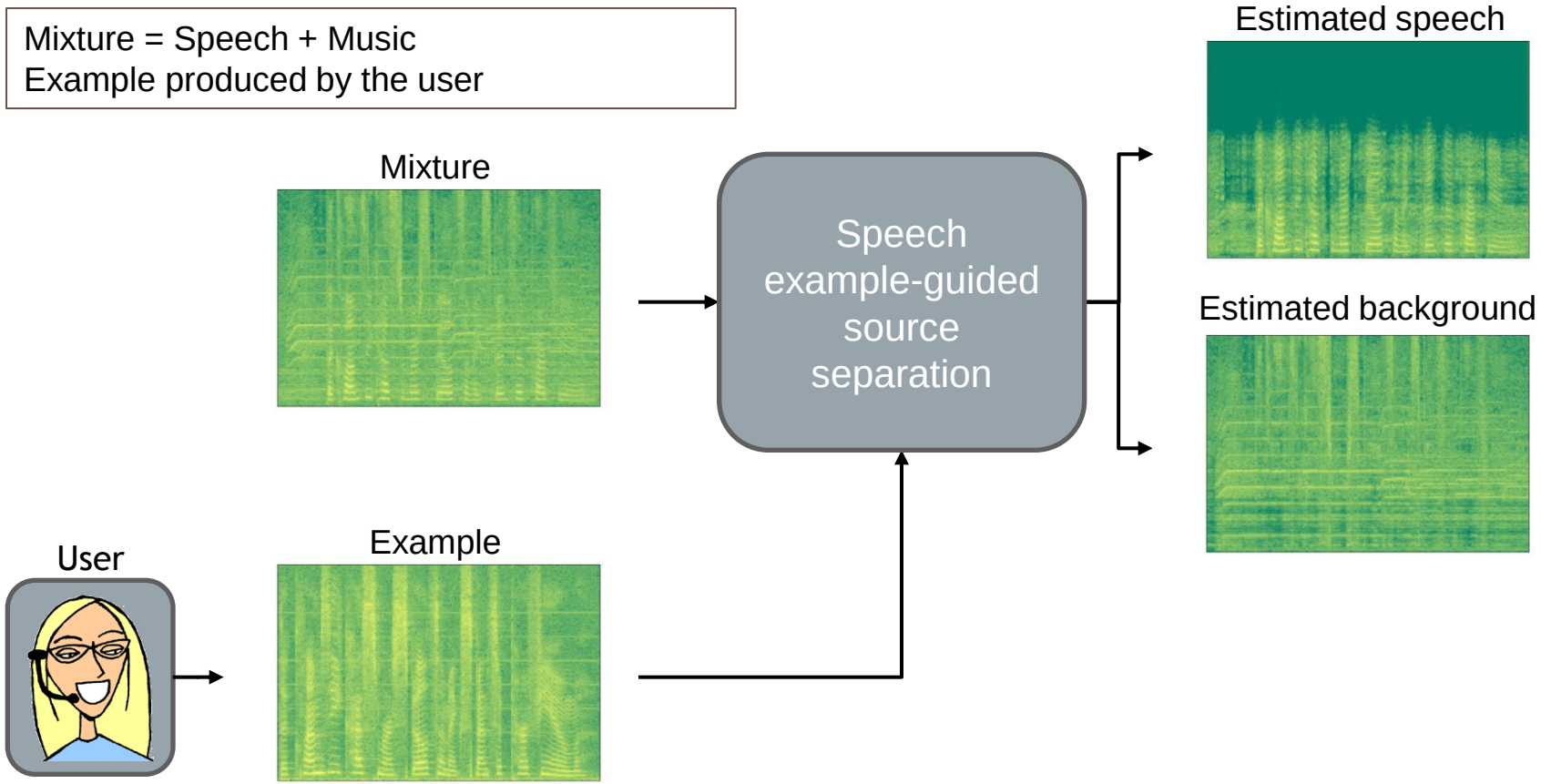
Video from P. Smaragdis and G. Mysore, “Separation by Humming”: User Guided Sound Extraction from Monophonic Mixtures” in Proc. WASPAA, New Paltz, NY. October 2009
http://www.cs.illinois.edu/~paris/demos/ai/user-guide.mp4

User-guided source separation

Example-based approaches

Separation by speaking

Mixture = Speech + Music
Example produced by the user



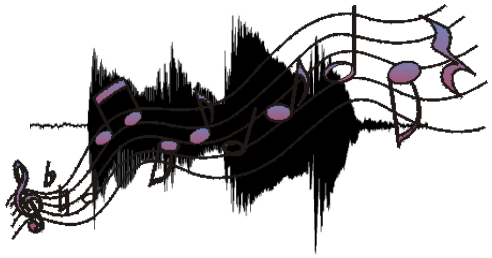
L. Le Magoarou, A. Ozerov and N. Duong “Text-Informed Audio Source Separation using Nonnegative Matrix Partial Co-Factorization”, in Proc. of MLSP, 2013

User-guided source separation
Example-based approaches

Separation by singing …

Separation by playing …

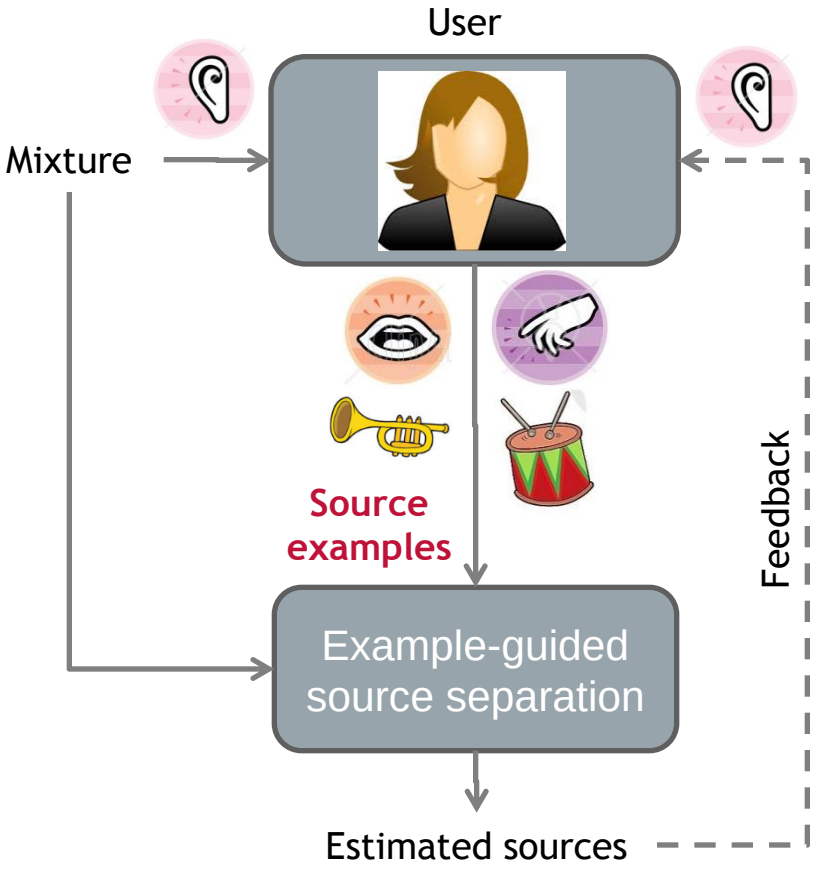
Related to cover-informed source separation



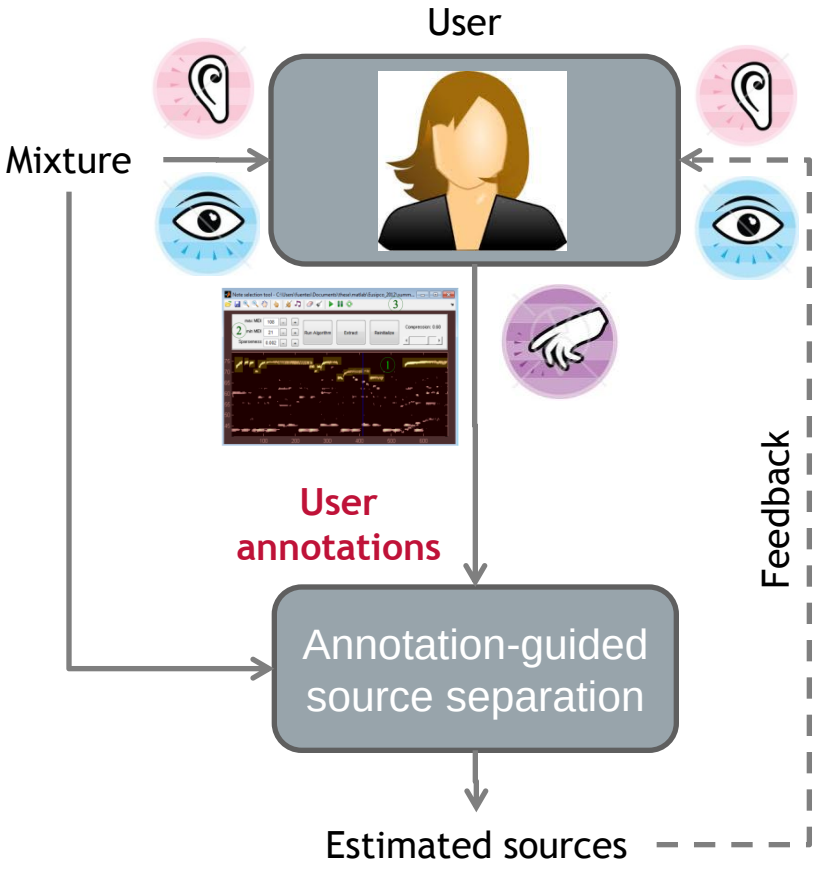
User-guided source separation

Main approaches

Example-based approaches

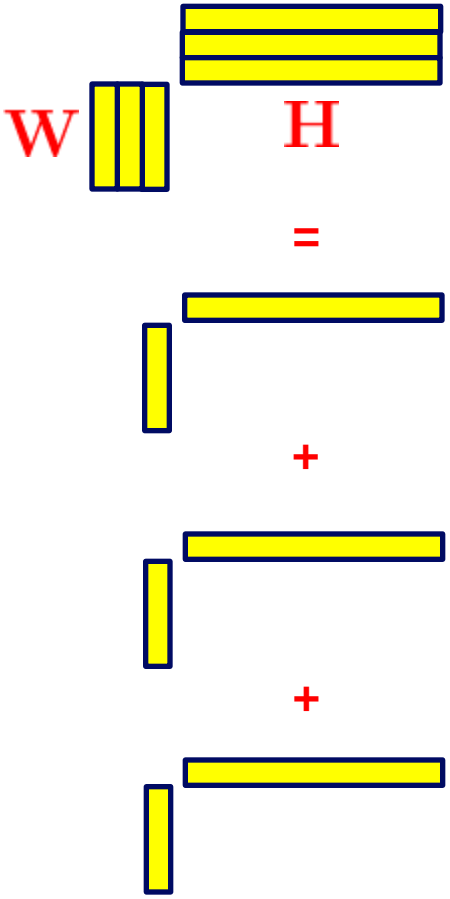


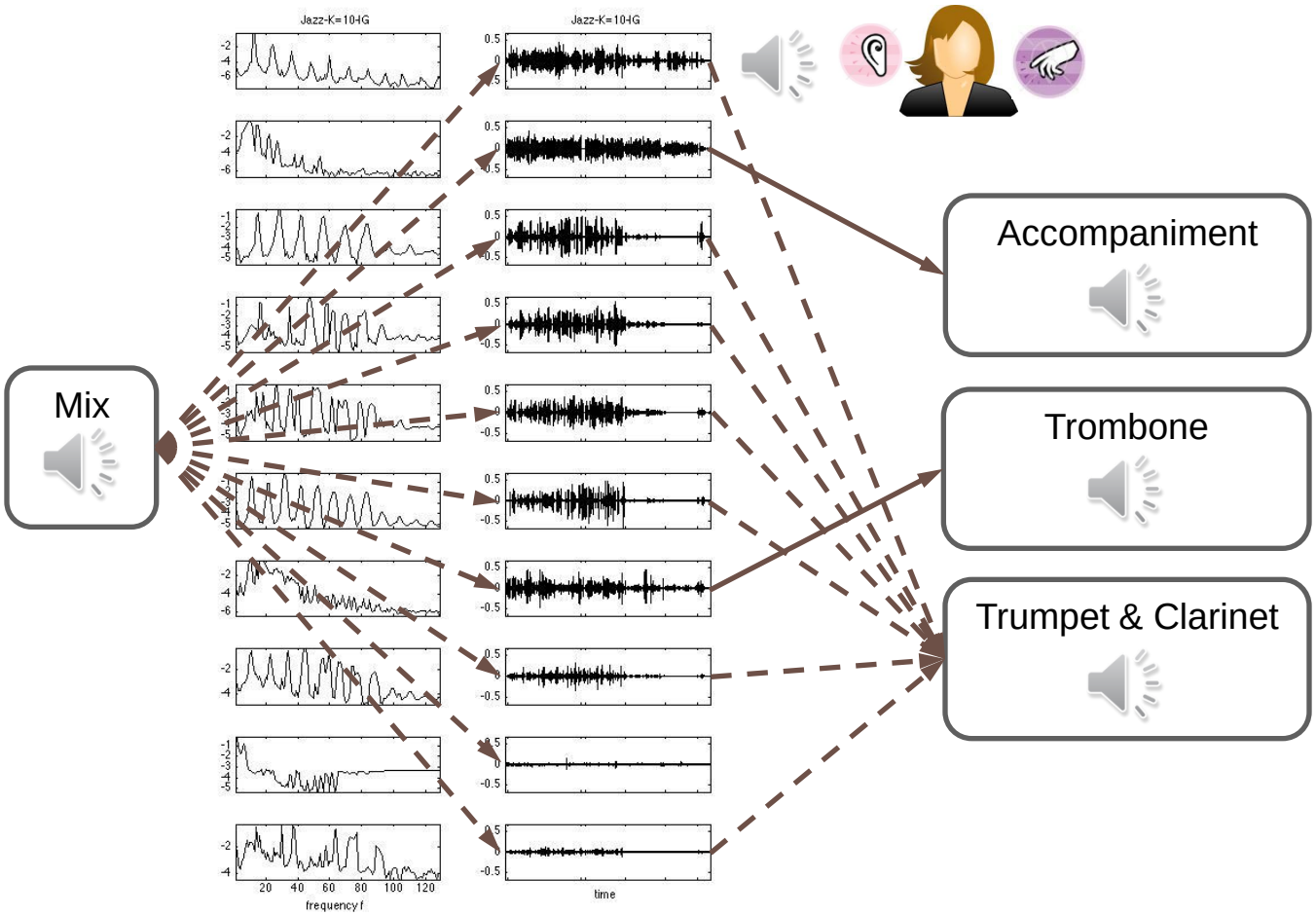
Annotation-based approaches



User-guided source separation

Manual NMF components grouping



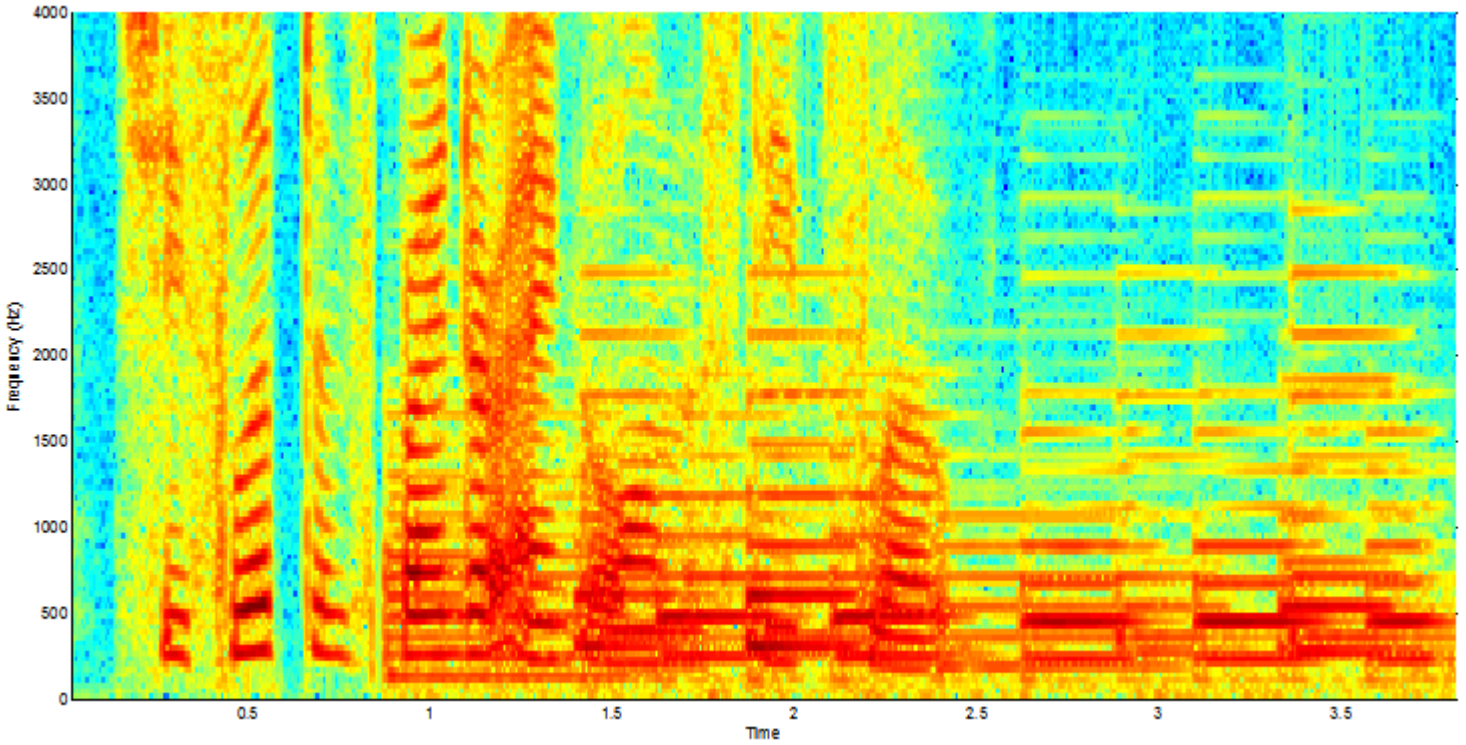


Examples from C. Févotte, N. Bertin, and J.-L. Durrieu. "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis." Neural Computation, 21(3):793–830, Mar. 2009. http://www.unice.fr/cfevotte/extras/neco09/

User-guided source separation

Temporal annotation-informed source separation

Mix = source 1 (speech) + source 2 (piano)

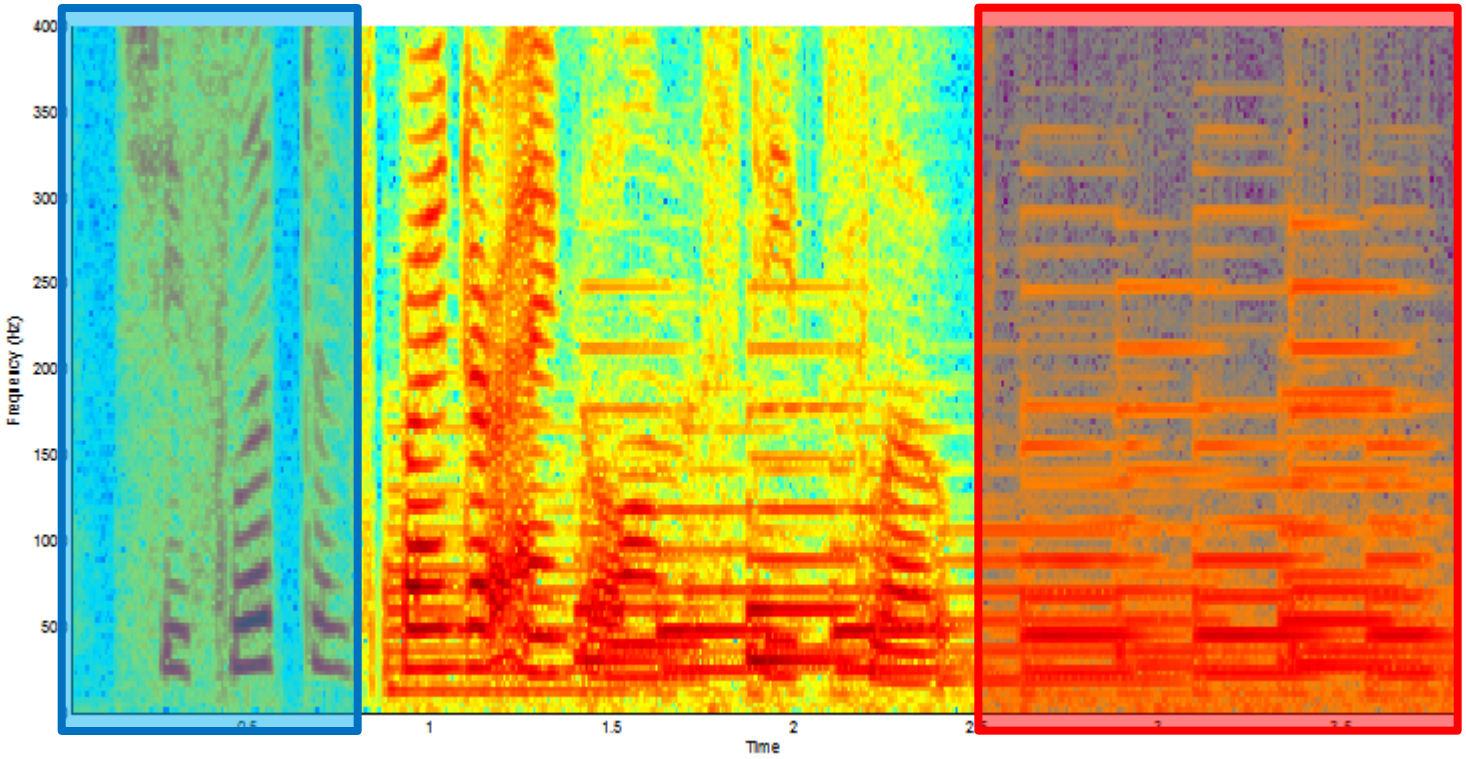


A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” ICASSP’11, Prague, 2011, pp. 257–260.

User-guided source separation

Temporal annotation-informed source separation

Mix = source 1 (speech) + source 2 (piano)

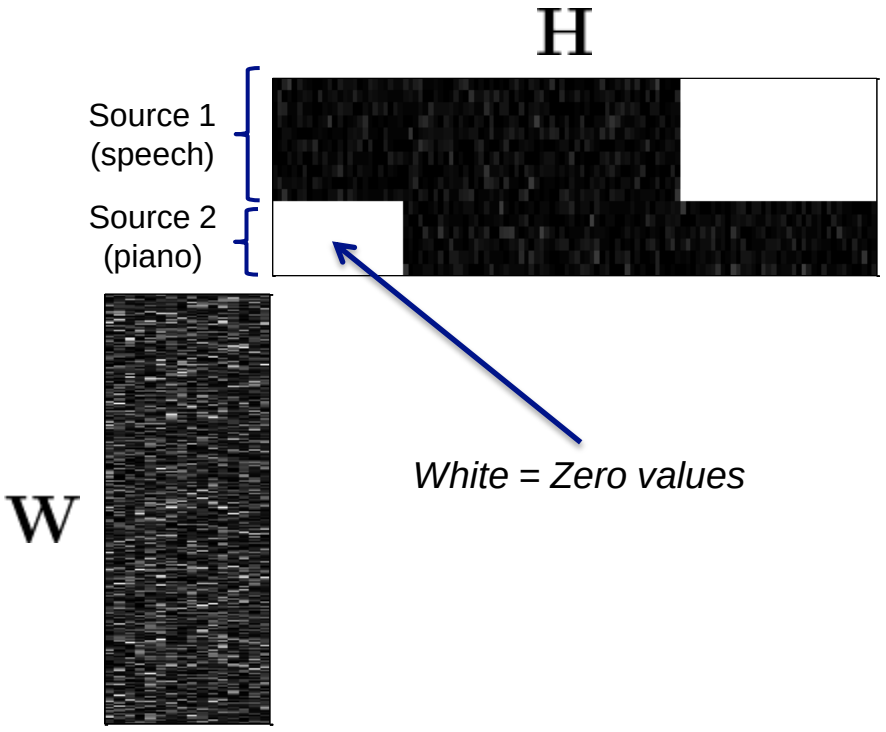


A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” ICASSP’11, Prague, 2011, pp. 257–260.

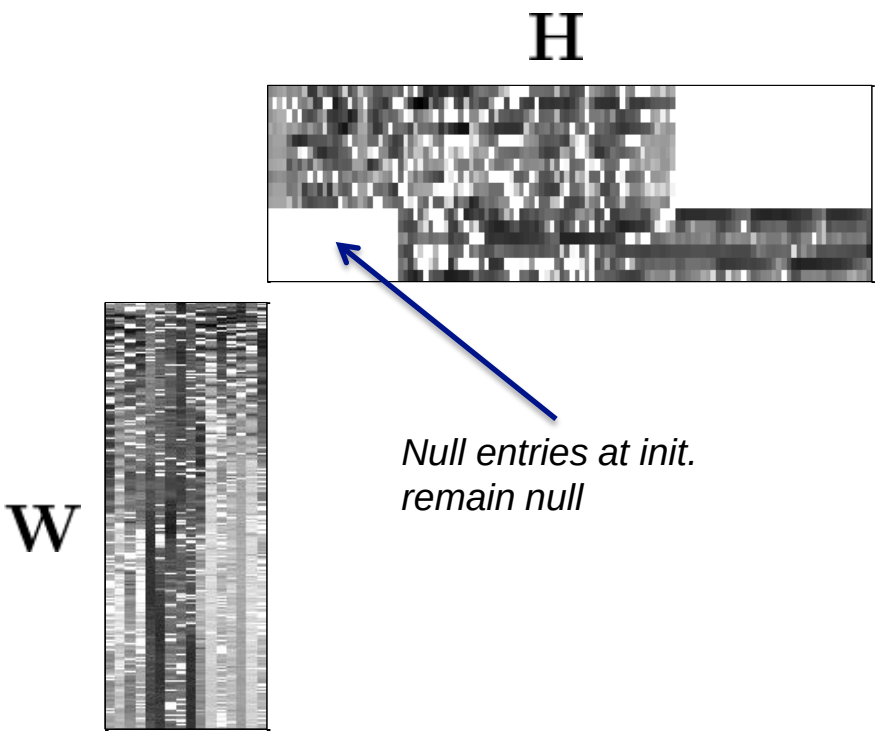
User-guided source separation

Temporal annotation-informed source separation

Initialization:



After convergence:



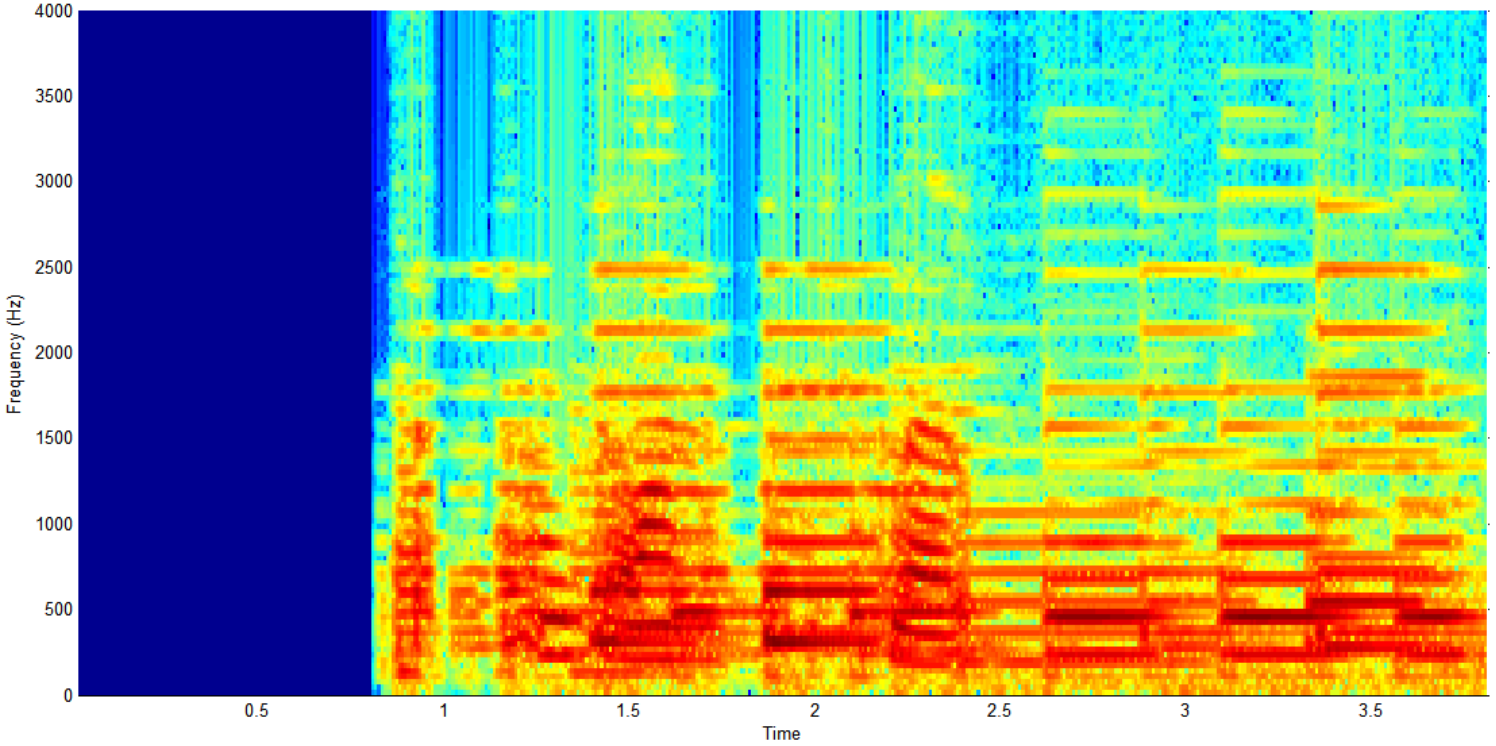
Due to multiplicative update rules, zero entries at the initialization stay at zero

A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” ICASSP’11, Prague, 2011, pp. 257–260.

User-guided source separation

Temporal annotation-informed source separation

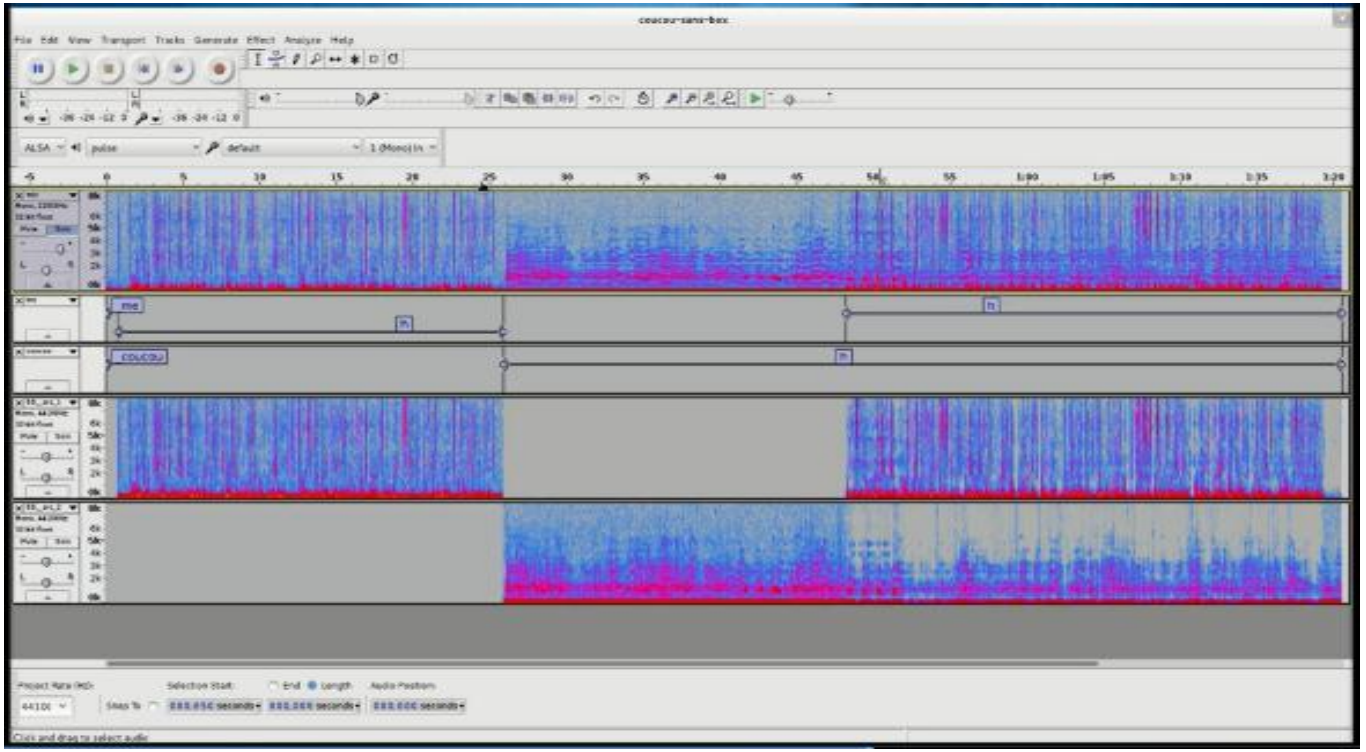
Source 2 (piano) estimate



A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” ICASSP’11, Prague, 2011, pp. 257–260.

User-guided source separation
Temporal annotation-informed source separation

Demo with GUI (simple example)

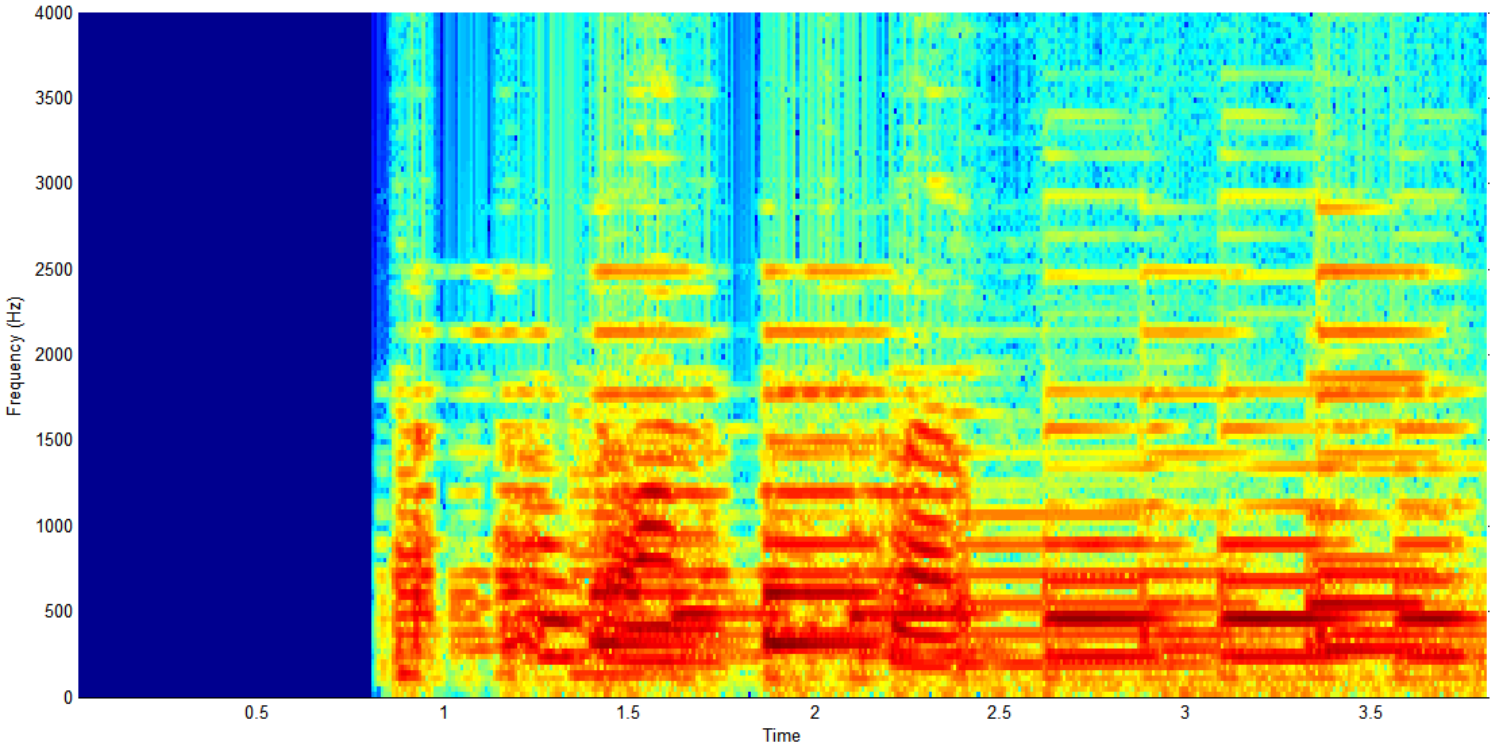


A. Ozerov, N. Q. K. Duong, and L. Chevallier, “Weighted nonnegative tensor factorization: on monotonicity of multiplicative update rules and application to user-guided audio source separation,” Tech. Rep., Technicolor, Oct. 2013, Available online: http://hal.inria.fr/hal-00878685
Demo video: https://www.youtube.com/watch?v=EjpLKvphpMo&feature=youtu.be

User-guided source separation

Temporal annotation-informed source separation

Source 2 (piano) estimate

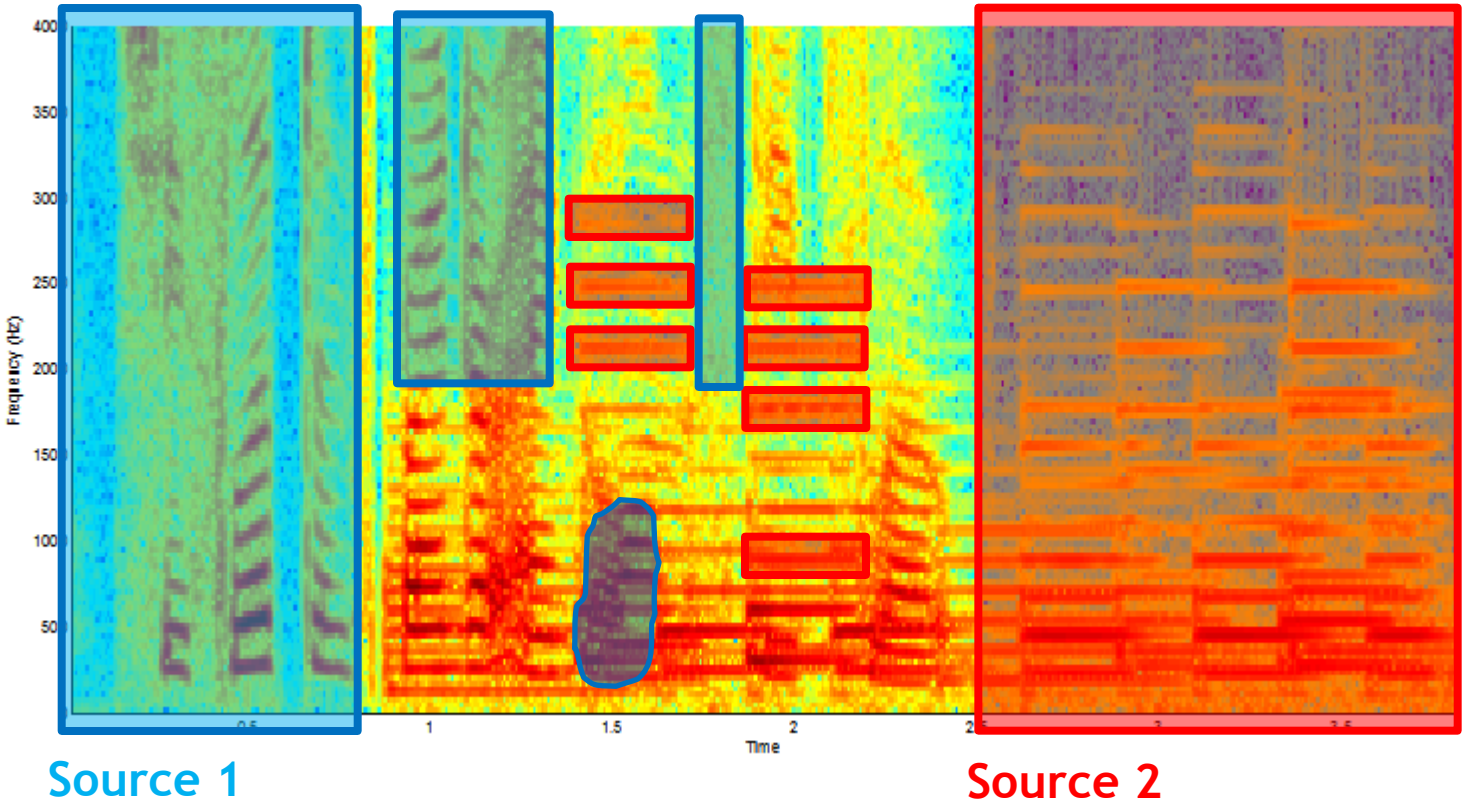


A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, “Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation,” ICASSP’11, Prague, 2011, pp. 257–260.

User-guided source separation

Time-frequency annotation-informed source separation

Mix = source 1 (speech) + source 2 (piano)



A. Lefèvre, F. Bach, and C. Févotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation in Proc. International Conference on Music Information Retrieval (ISMIR), 2012

User-guided source separation

Time-frequency annotation-informed source separation

- **How to supervise NMF with time-frequency annotations?**
- **One cannot any more just zeroing parts of W or H …**

  - Usual criterion: minimization of

$$D(\mathbf{V}|\mathbf{WH}) = \sum_{f,n=1}^{F,N} d\left(v_{fn}|[\mathbf{WH}]_{fn}\right)$$

  - Time-frequency annotation guided criterion: minimization of

$$\sum_{f,n=1}^{F,N} d\left(v_{fn}|[\mathbf{WH}]_{fn}\right) + \lambda \sum_{j,f,n=1}^{2,F,N} b_{j,fn} d\left(v_{fn}|[\mathbf{W}_j\mathbf{H}_j]_{fn}\right)$$

$$\mathbf{WH} = \mathbf{W}_1\mathbf{H}_1 + \mathbf{W}_2\mathbf{H}_2 \qquad b_{j,fn} \in \{0,1\}$$

Binary source annotation mask

A. Lefèvre, F. Bach, and C. Févotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation , in Proc. International Conference on Music Information Retrieval (ISMIR), 2012

User-guided source separation

Interactive time-frequency annotation-informed separation

Adding interactivity:

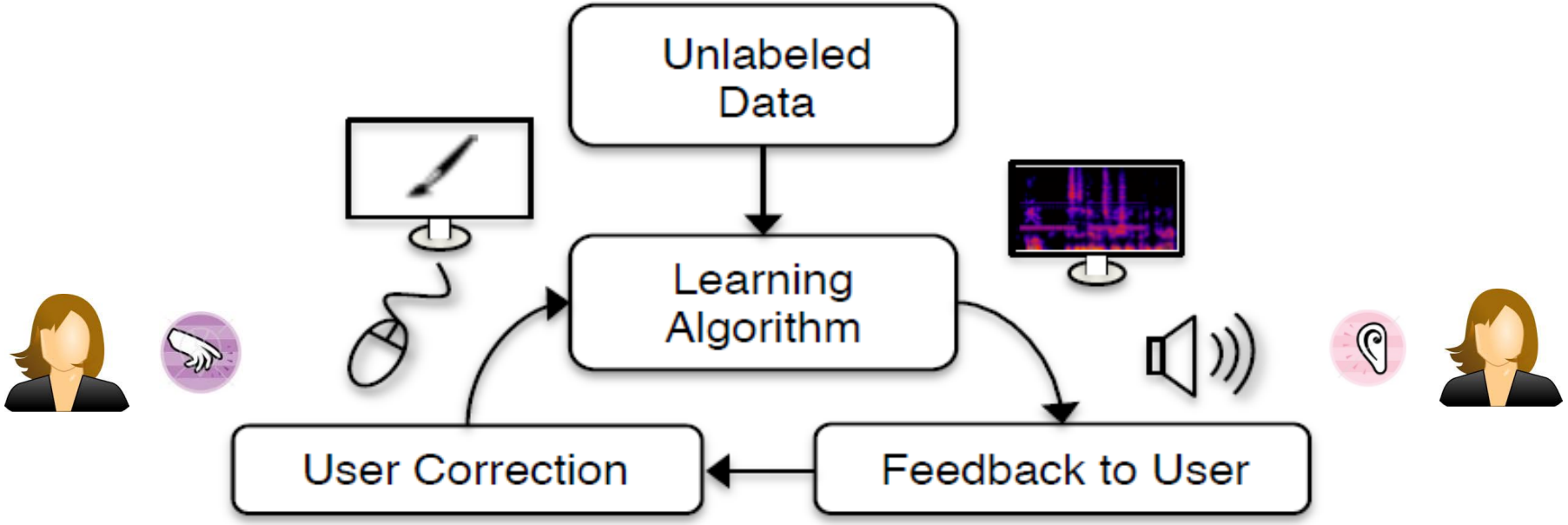
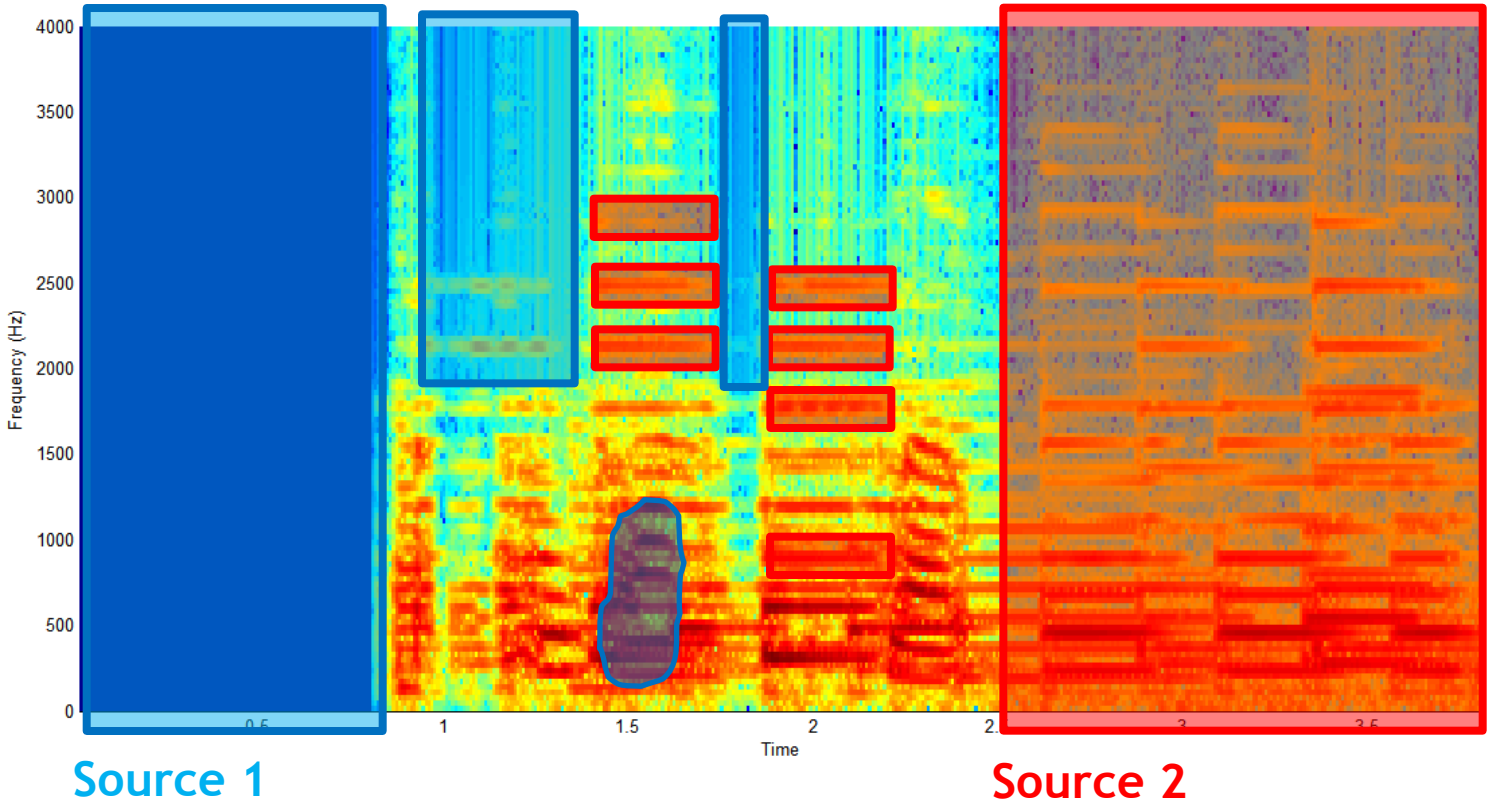


Figure from Nicholas J. Bryan, Gautham J. Mysore, Ge Wang, “ISSE: An Interactive Source Separation Editor”, to appear in the Proc. ACM Conference on Human Factors in Computing Systems (CHI), 2014

User-guided source separation

Interactive time-frequency annotation-informed separation

Source 2 (piano) estimate

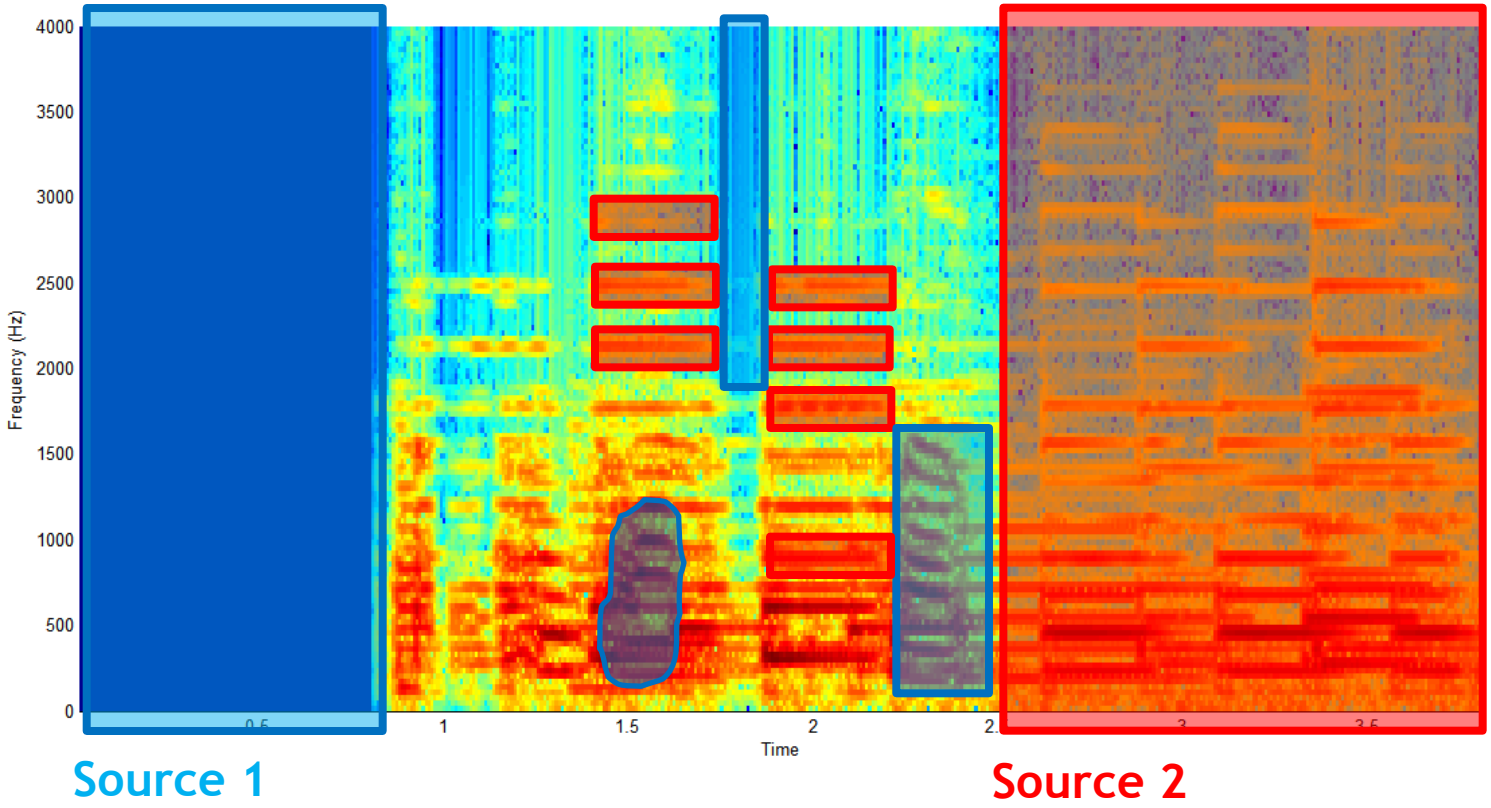


Nicholas J. Bryan, Gautham J. Mysore, ”Interactive Refinement of Supervised and Semi-supervised Sound Source Separation Estimates”, in ICASSP, Vancouver, Canada. May 2013

User-guided source separation

Interactive time-frequency annotation-informed separation

Source 2 (piano) estimate



Nicholas J. Bryan, Gautham J. Mysore, ”Interactive Refinement of Supervised and Semi-supervised Sound Source Separation Estimates”, in ICASSP, Vancouver, Canada. May 2013

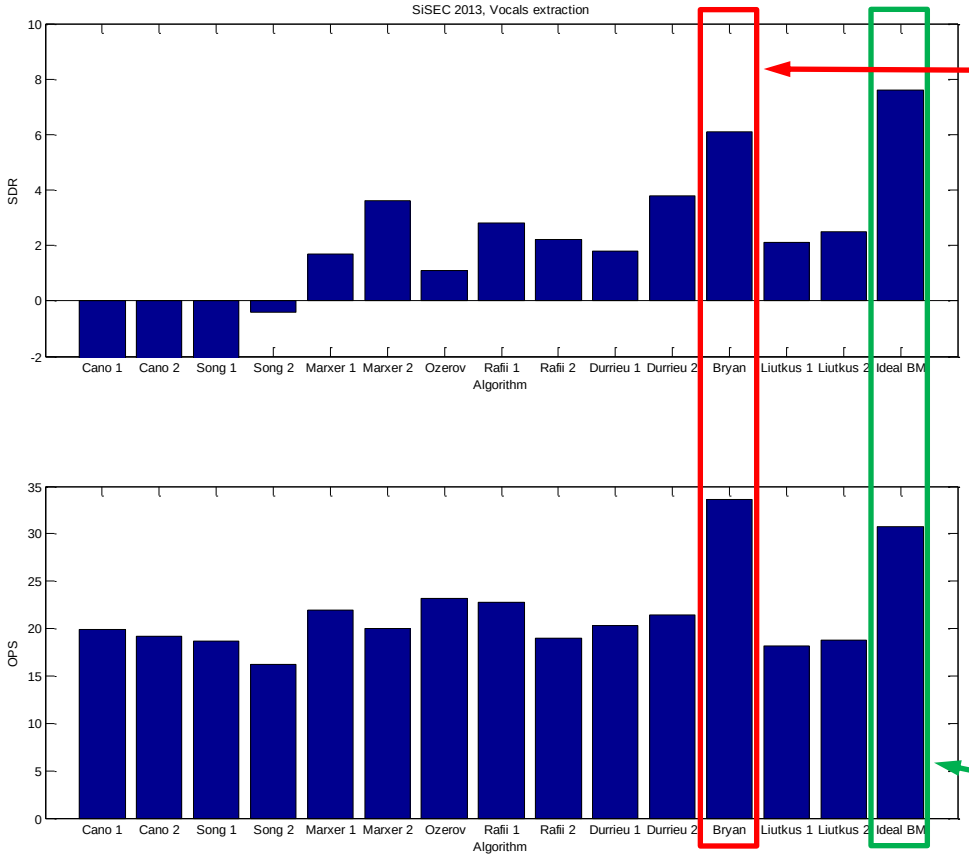
User-guided source separation

Interactive time-frequency annotation-informed separation

SiSEC 2013 Source Separation Evaluation Campaign (singing voice extraction)
http://www.onn.nii.ac.jp/sisec13/evaluation_result/MUS/testMUS2013.htm

SDR: usual measure

OPS: perceptual measure



[Bryan & Mysore 2013]
Interactive approach

User interaction:

15 to 60 minutes / source
Up to 100 iterations

Mix

Vocals **Drums**

Vocals **Drums**

[Oracle binary mask]

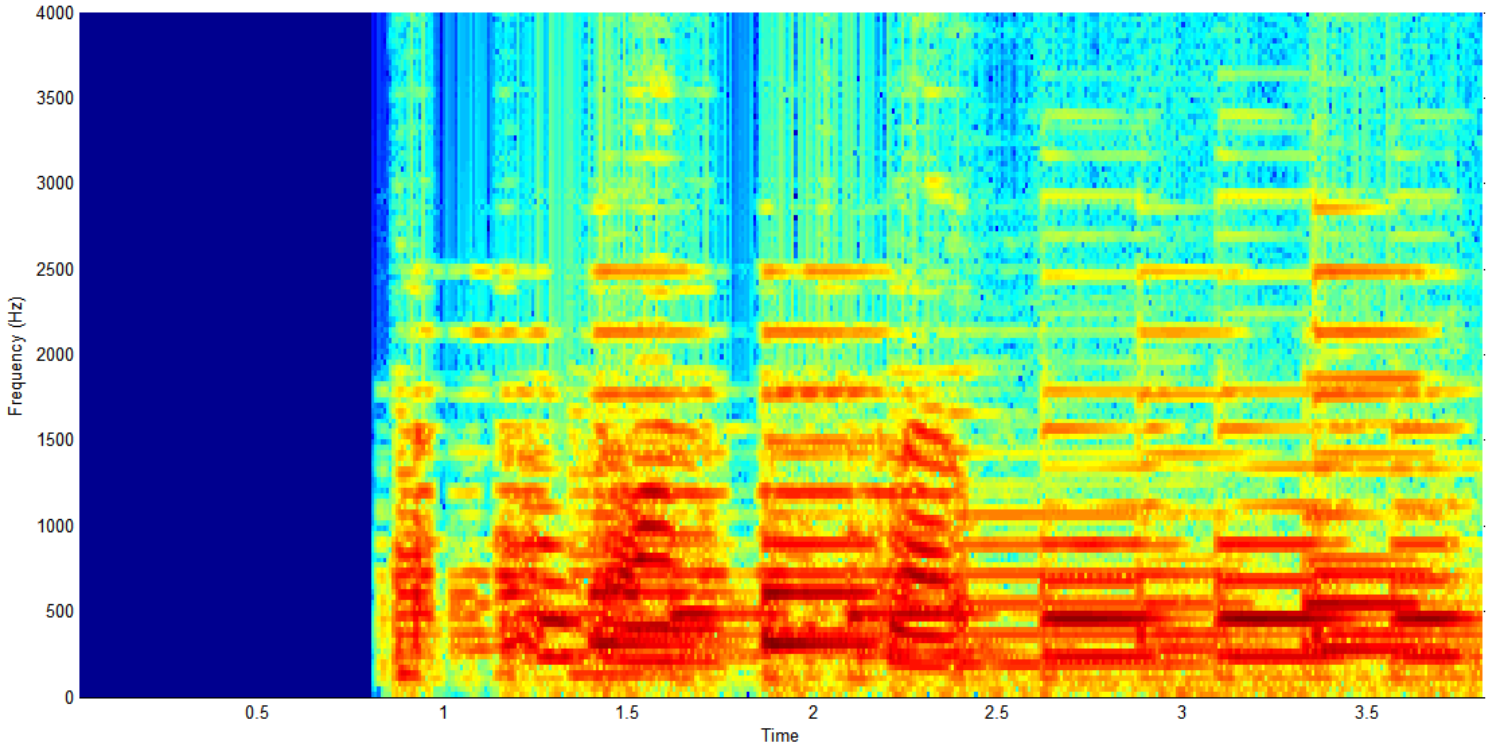
Nicholas J. Bryan, Gautham J. Mysore, "Interactive Refinement of Supervised and Semi-supervised Sound Source Separation Estimates", in ICASSP, Vancouver, Canada. May 2013

stop

User-guided source separation

Interactive time-frequency annotation-informed separation with "well-separated" annotations

Source 2 (piano) estimate

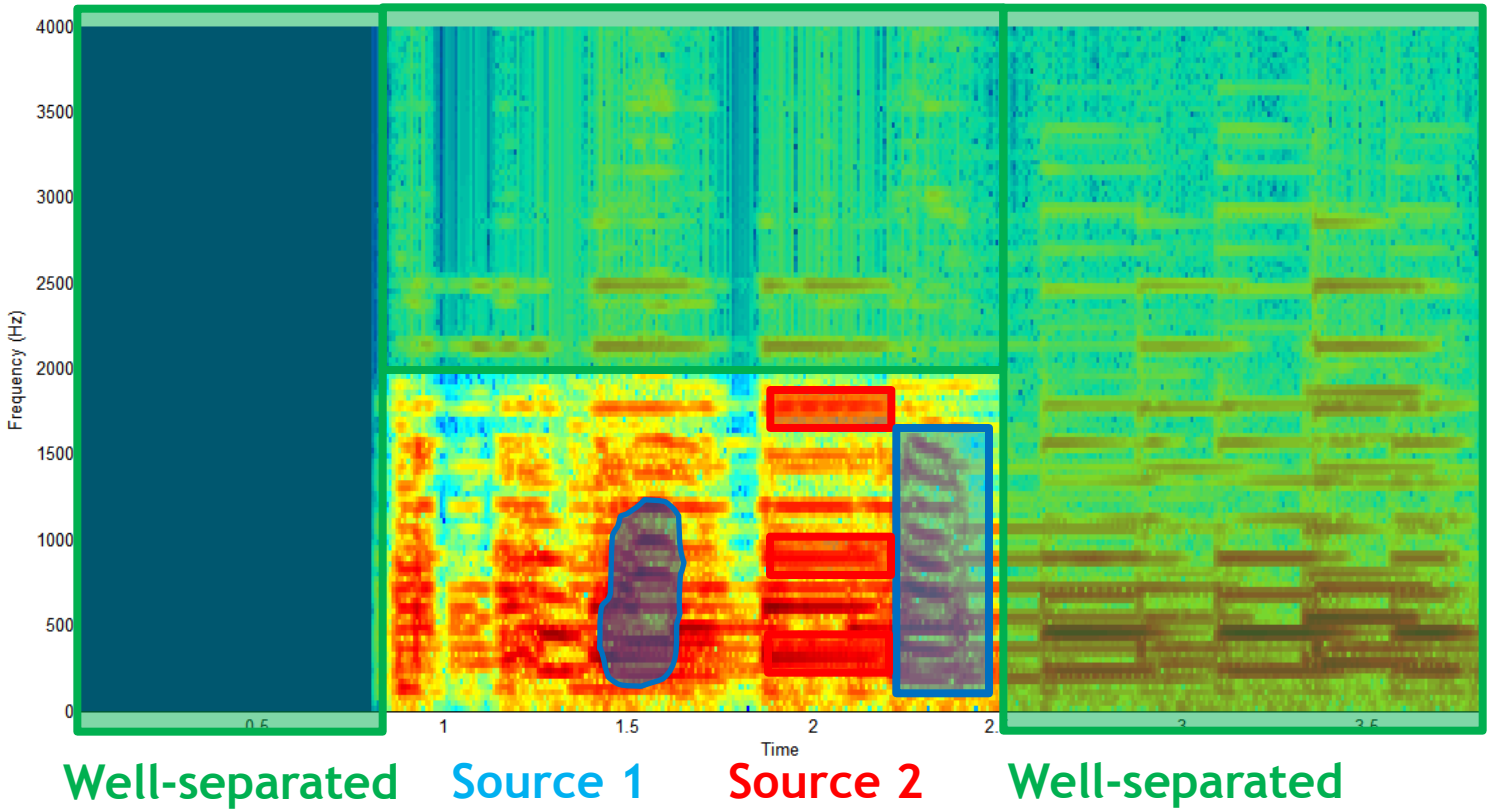


N. Q. K. Duong, A. Ozerov, L. Chevallier and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," ICASSP'14, Florence, Italy, May, 2014.

User-guided source separation

Interactive time-frequency annotation-informed separation with "well-separated" annotations

Source 2 (piano) estimate



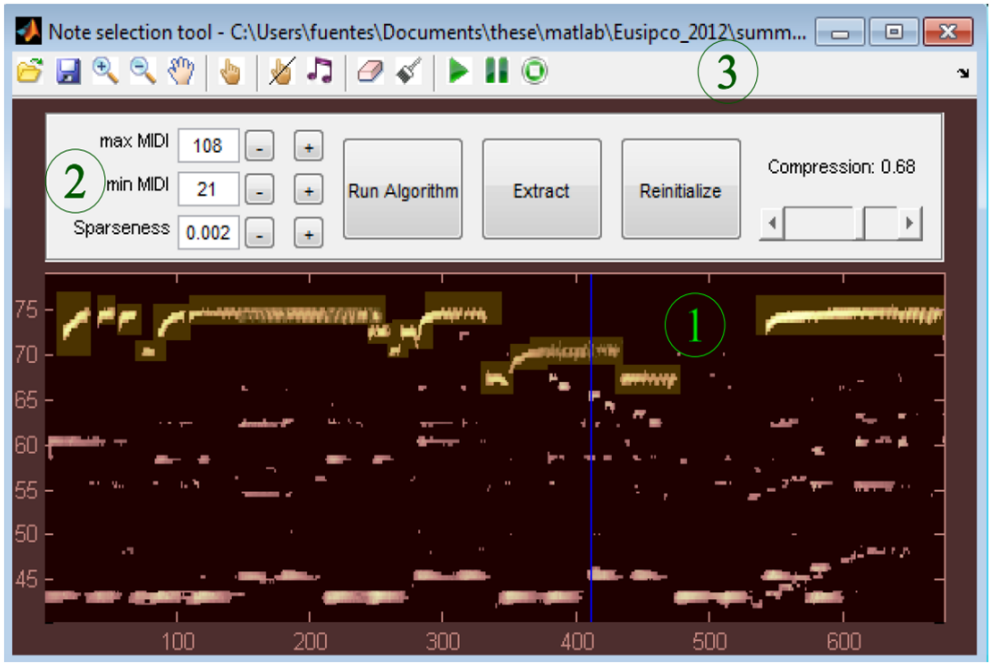
N. Q. K. Duong, A. Ozerov, L. Chevallier and J. Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," ICASSP'14, Florence, Italy, May, 2014.

User-guided source separation

Interactive time-pitch annotation-informed separation

- 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

Demo with a GUI



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

Keynote content

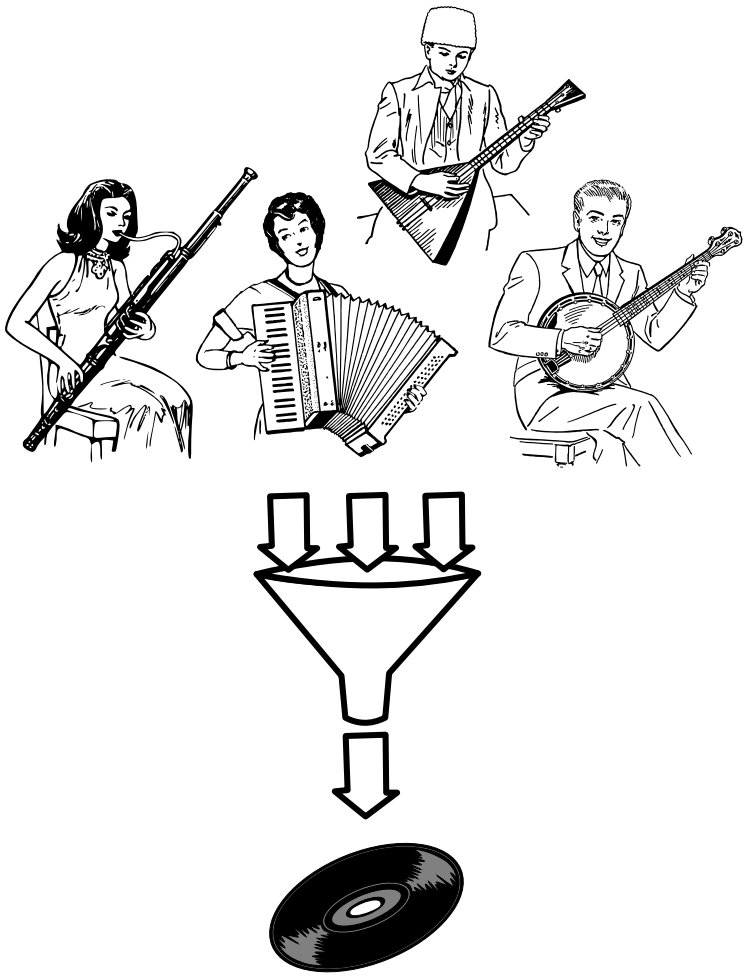
- **Objective**
  - To provide an overview of major trends in Informed Source Separation (ISS)
- **Outline of the keynote**
  - *PART 1: Gaël RICHARD*
    - Introduction on Informed Source Separation
    - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
  - *PART 2: Alexey OZEROV*
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*
  - ***PART 3: Antoine LIUTKUS***
    - ***Coding-based informed source separation***
  - *PART 4: Gaël RICHARD*
    - Conclusion

Coding-based Informed Source Separation

Antoine Liutkus
Inria, PAROLE, France

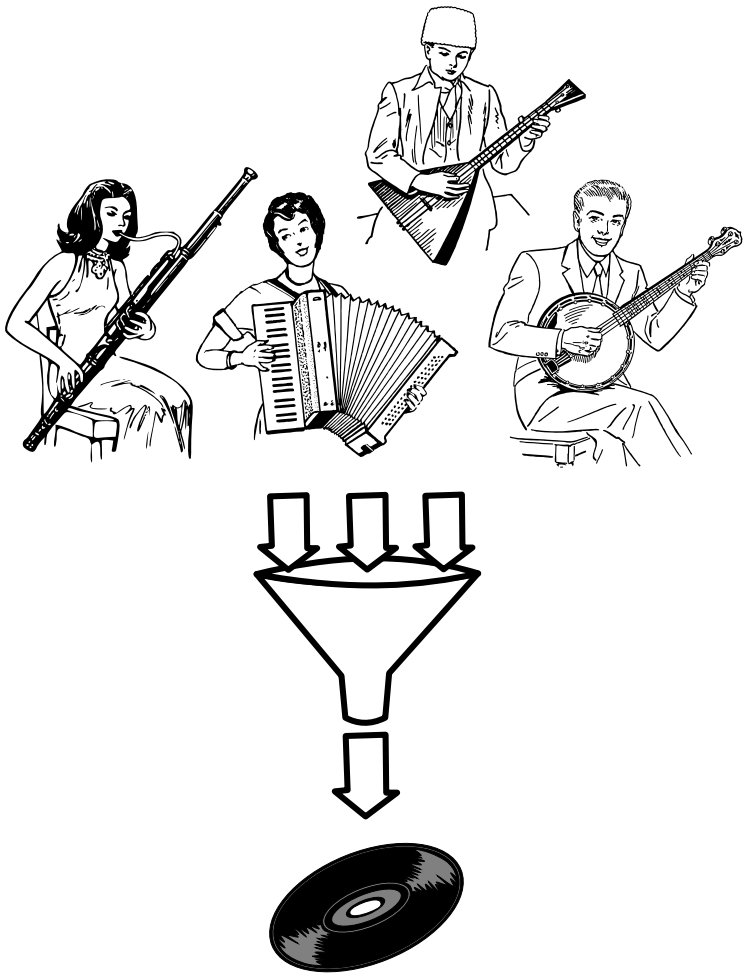
ICASSP, Florence, May 4th 2014

Active listening of music



A record is made of many instruments...

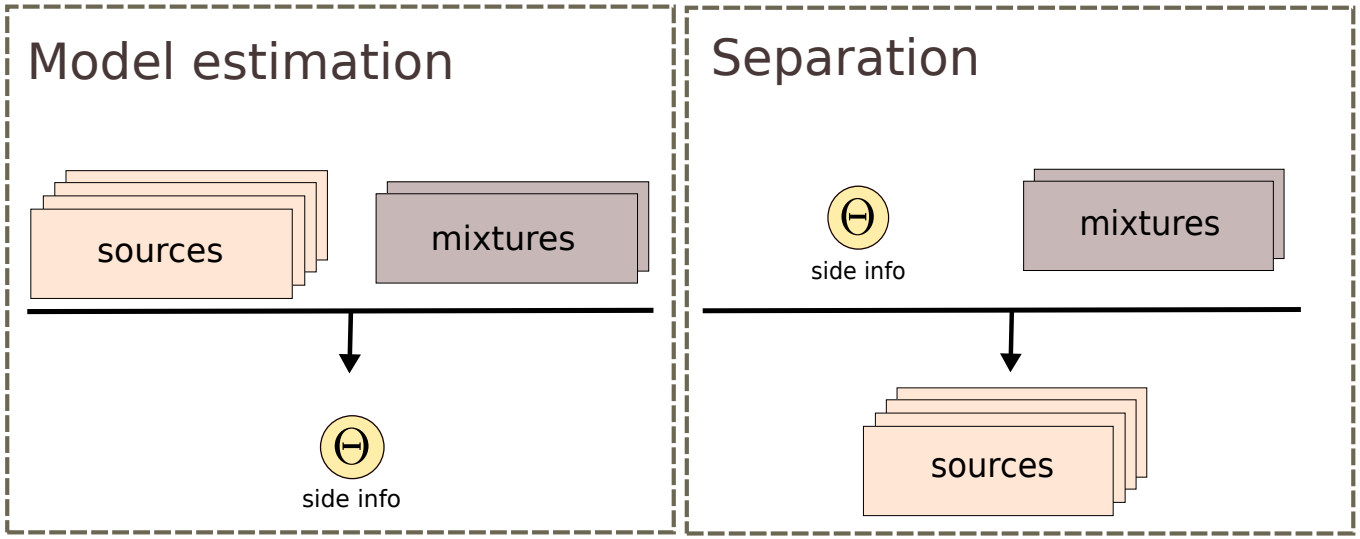
The case of music



A record is made of many instruments...

... What if I want to remove one of them ?

Coding-based informed separation

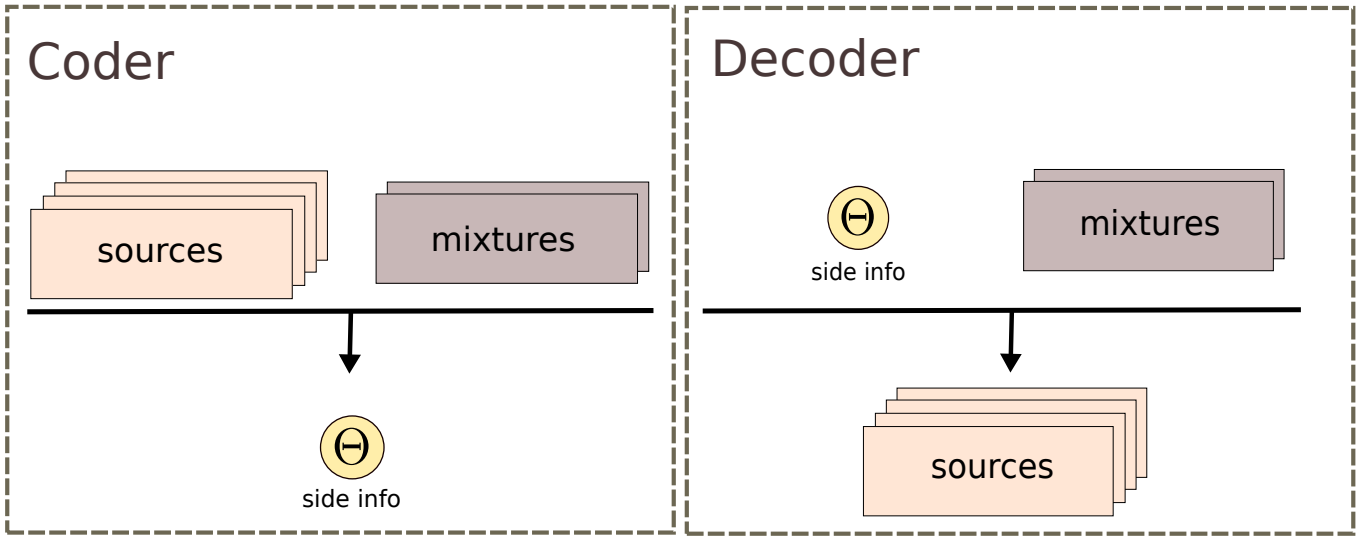


Interesting because Θ is small

M. Parvaix, L. Girin, and J.-M. Brossier. A watermarking-based method for single-channel audio source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 101–104, Taipei, Taiwan, 2009

M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1721 –1733, August 2011

Coding-based informed separation



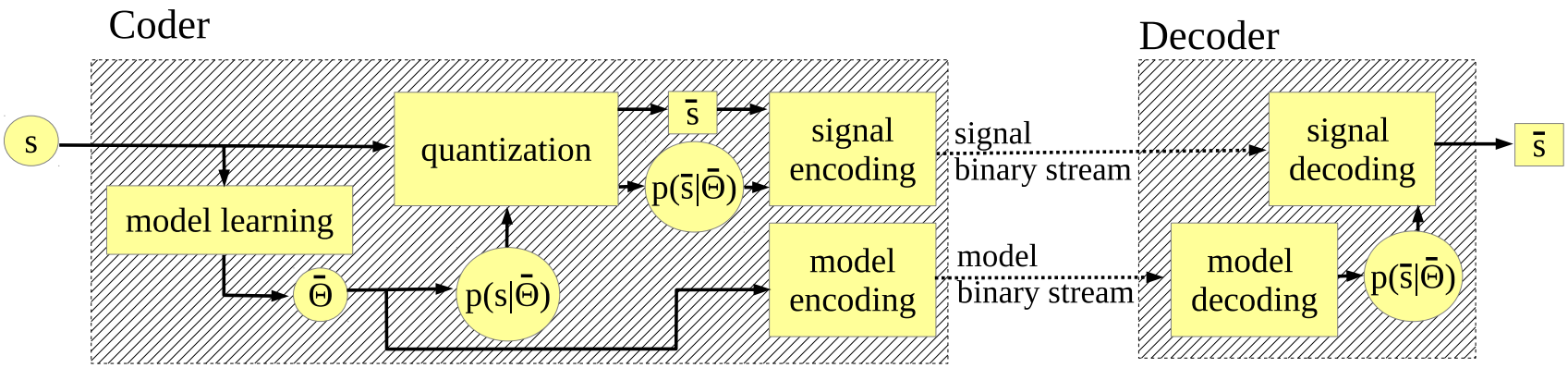
Strong connexions with audio coding

J. Herre and S. Disch. New concepts in parametric coding of spatial audio: From SAC to SAOC. In *IEEE International Conference on Multimedia and Expo (ICME 2007)*, pages 1894–1897, Beijing, China, July 2007

J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen. Spatial audio object coding (SAOC) - The upcoming MPEG standard on parametric object based audio coding. In *124th Audio Engineering Society Convention (AES 2008)*, Amsterdam, Netherlands, May 2008

Classical audio waveform coding

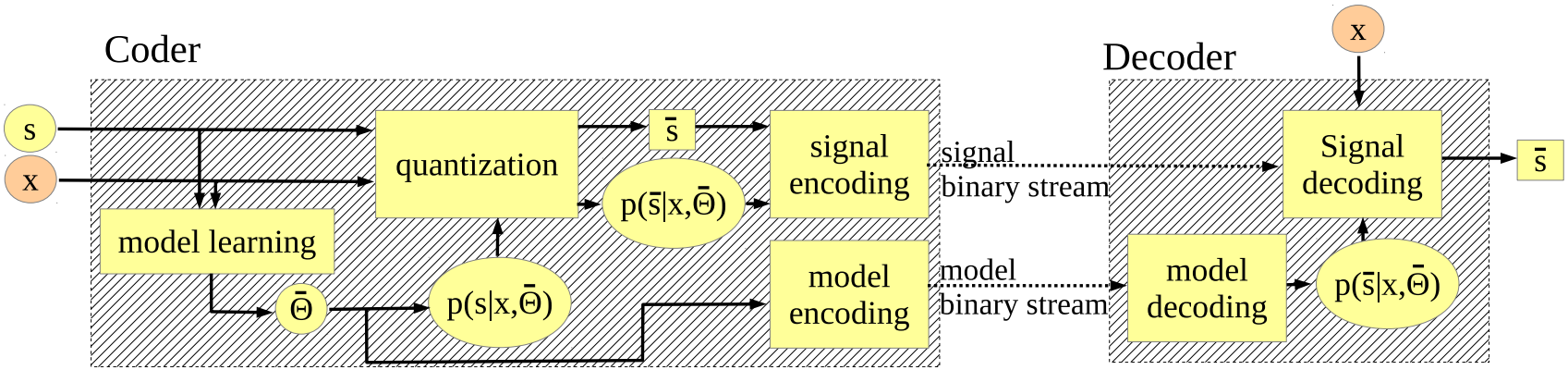
block-diagram



- Source : sequence of independent variables $\mathbf{s} = \mathbf{s}_1, \ldots, \mathbf{s}_L$
- Each $\mathbf{s}_l$:
  - vector of dimension $J \times 1$
  - Modelled by $p(\mathbf{s}_l \mid \Theta)$

Coding-based Informed Source Separation

block-diagram



- Same as source coding, but $\mathbf{x}$ is known at coder and decoder
  $\Rightarrow$ using $p(\mathbf{s}_l \mid \mathbf{x}, \Theta)$ instead of $p(\mathbf{s}_l \mid \Theta)$!
- Optimal exploitation of the mixture to convey the sources

connecting source separation and source coding

Outline

1 Introduction

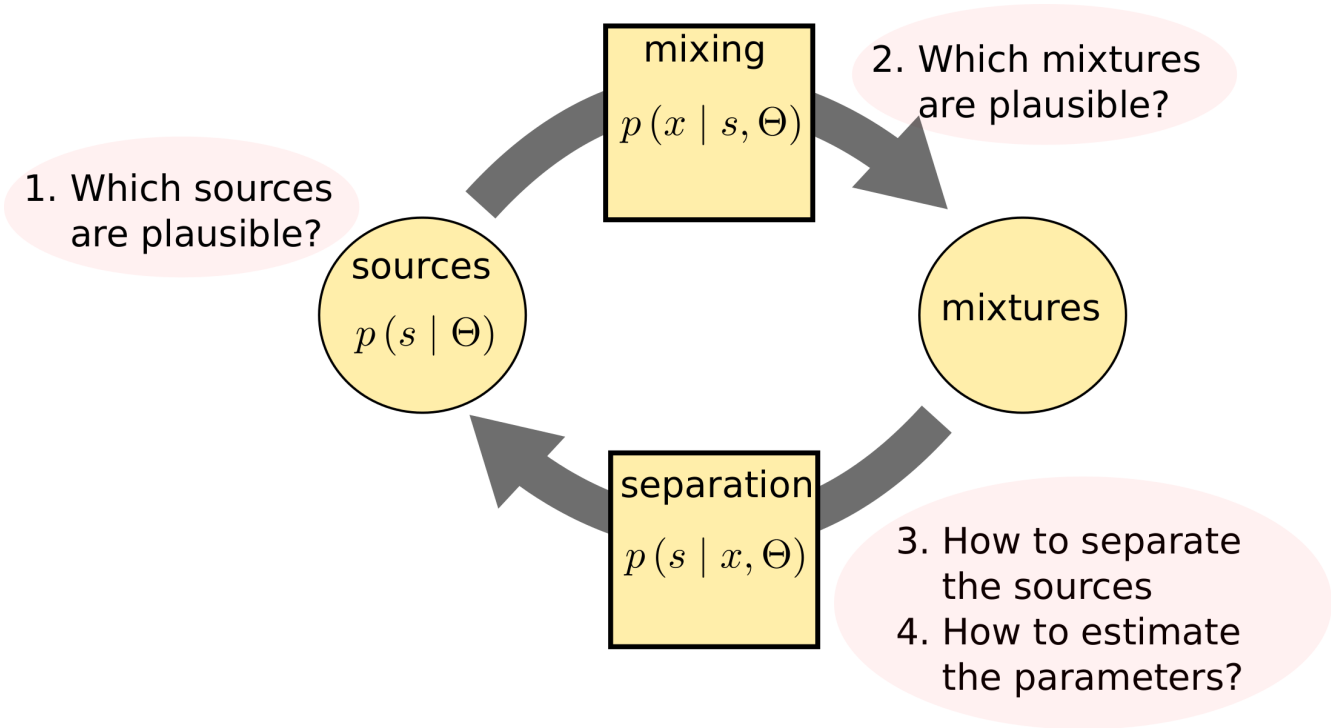
2 Probabilistic framework

3 Parametric Informed Source Separation

4 Posterior source coding

5 Conclusion

Probabilistic modelling



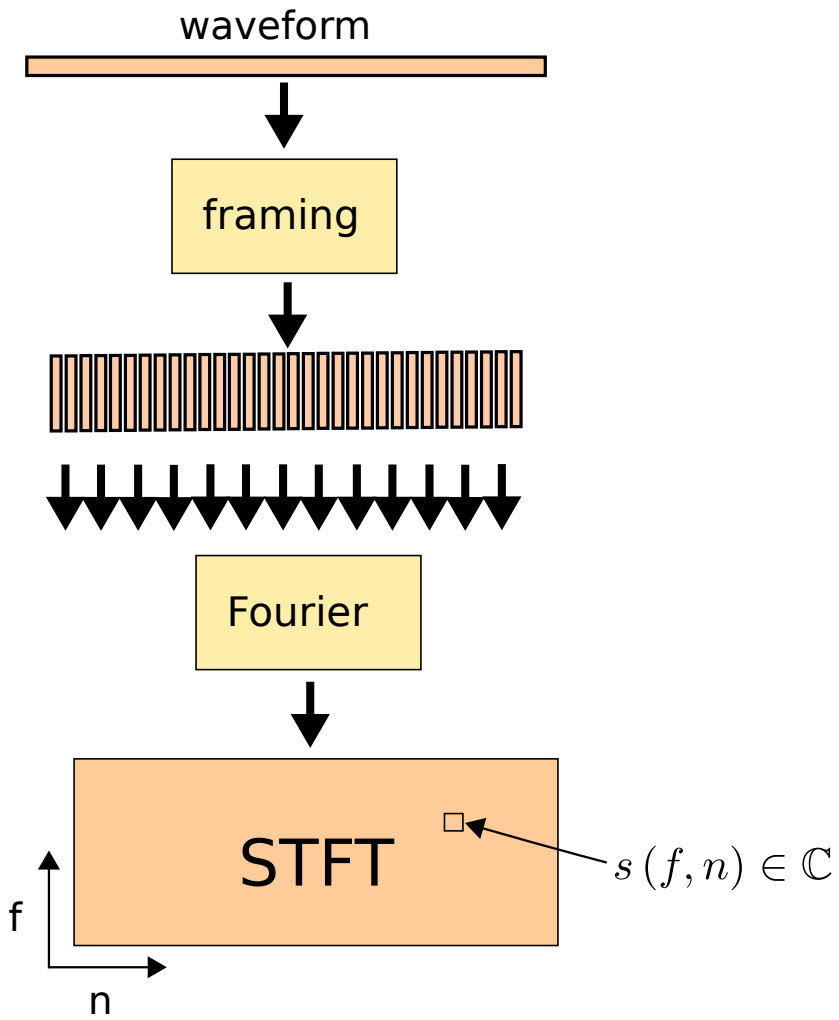
E. Vincent, G.M. Jafari, A.S Abdallah, D.M. Plumbley, and E.M. Davies. Probabilistic modeling paradigms for audio source separation. In W. Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, pages 162–185. IGI Global, 2010

A. T. Cemgil, S. J. Godsill, P. H. Peeling, and N. Whiteley. *The Oxford Handbook of Applied Bayesian Analysis*, chapter Bayesian Statistical Methods for Audio and Music Processing. Number ISBN13: 978-0-19-954890-3. Oxford University Press, 2010

Probability framework

1. *What kind of source signals s are plausible ?*
2. What pairs of (sources s,mixtures x) are plausible?
3. What can I say on s given x?

Short Term Fourier Transform



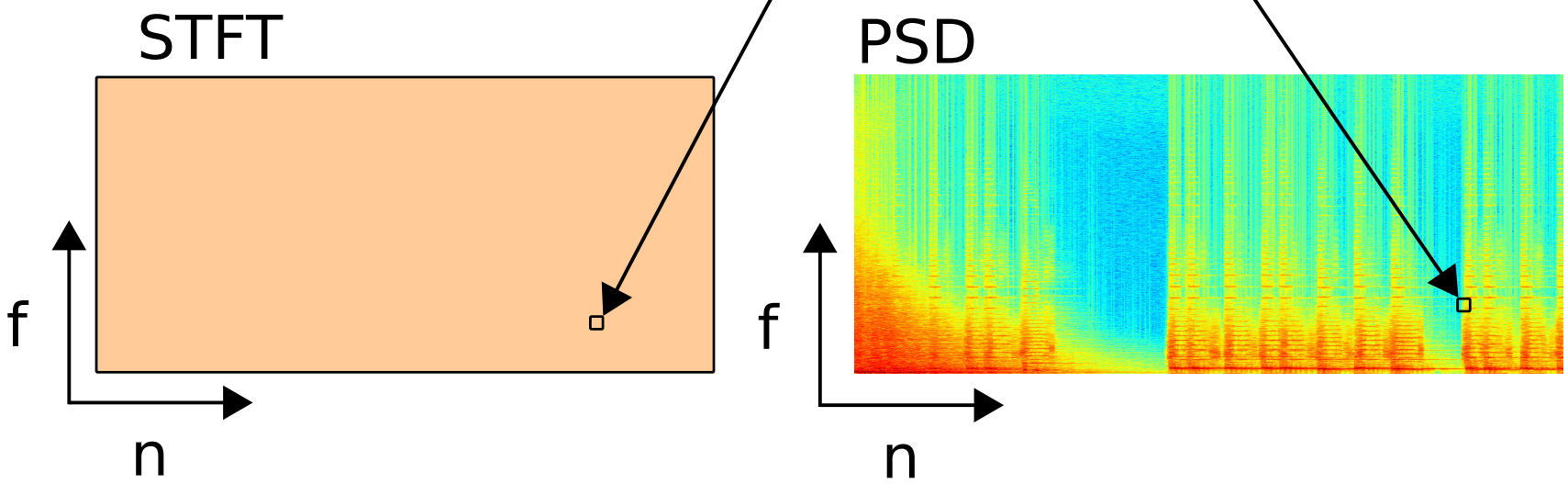
Locally Stationary Gaussian Processes

source model

- All sources are **independent** a priori
- All frames are **independent**, **Gaussian** and **stationary**

$\Rightarrow$ All coefficients of the STFT are Gaussian and independent

$$s(f,n,j) \sim \mathcal{N}_c(0, P(f,n,j))$$

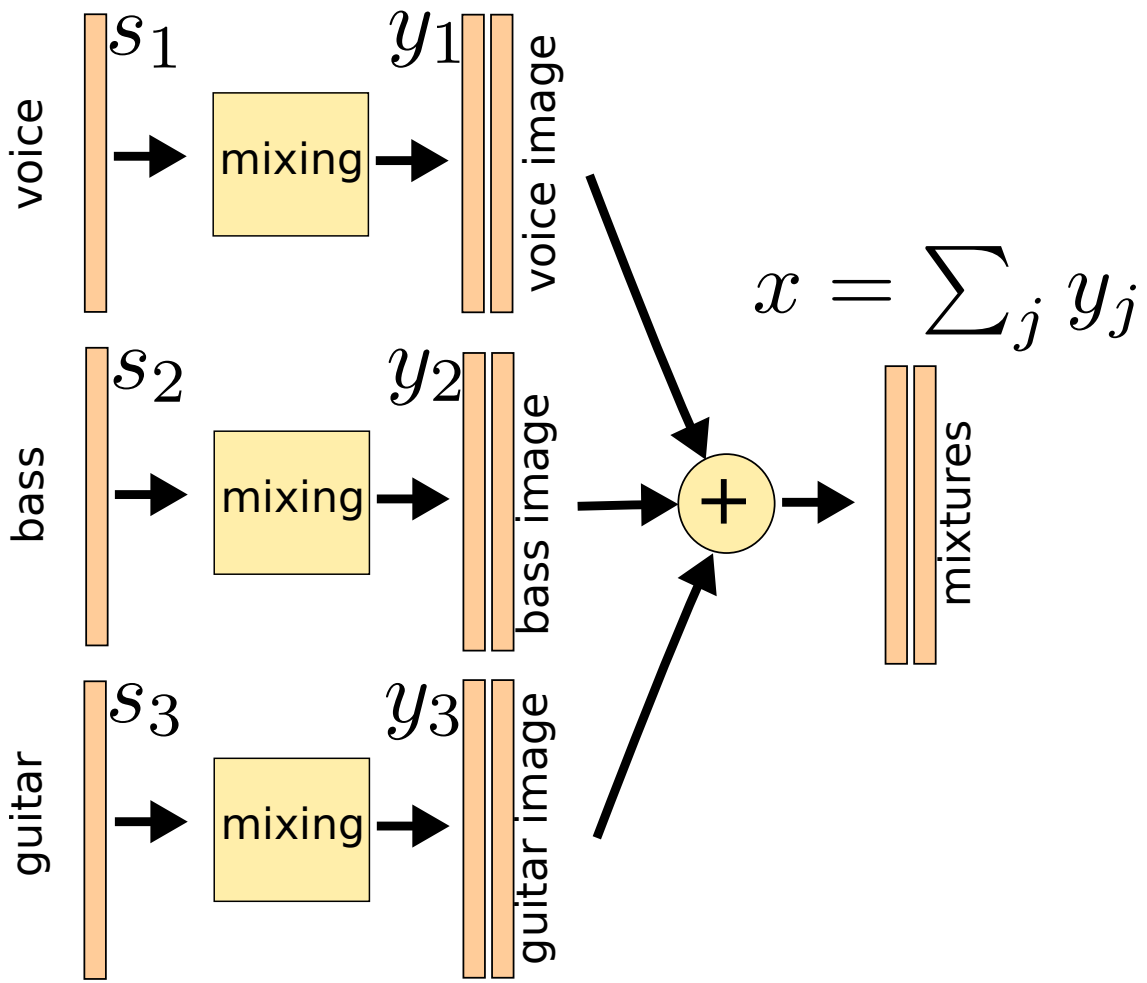


Probability framework

1. *What kind of source signals s are plausible ?*
2. What pairs of (sources s,mixtures x) are plausible?
3. What can I say on s given x?

Mixtures as a sum of spatial images

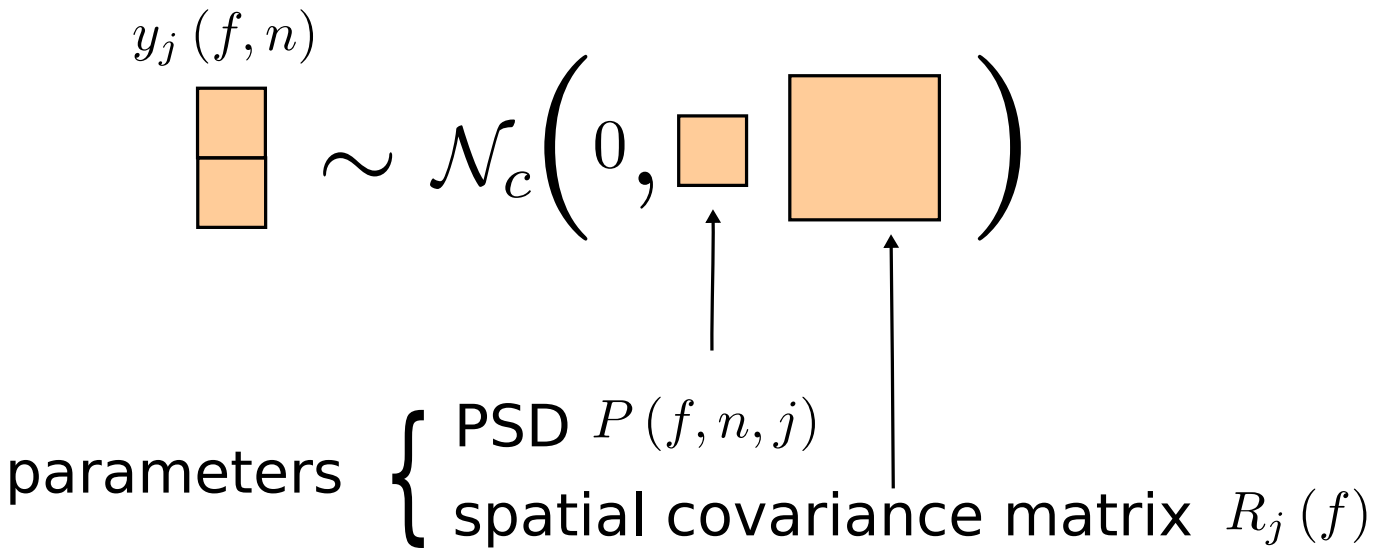
mixing model (1/2)



Diffuse mixing model

mixing model (2/2)

- Channels of an image y_j are supposed jointly Gaussian



- **Encodes inter-channel correlations.**
  Generalizes instantaneous & narrowband convolutive models

N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1830–1840, sept. 2010

Images/mix joint distribution (1/2)

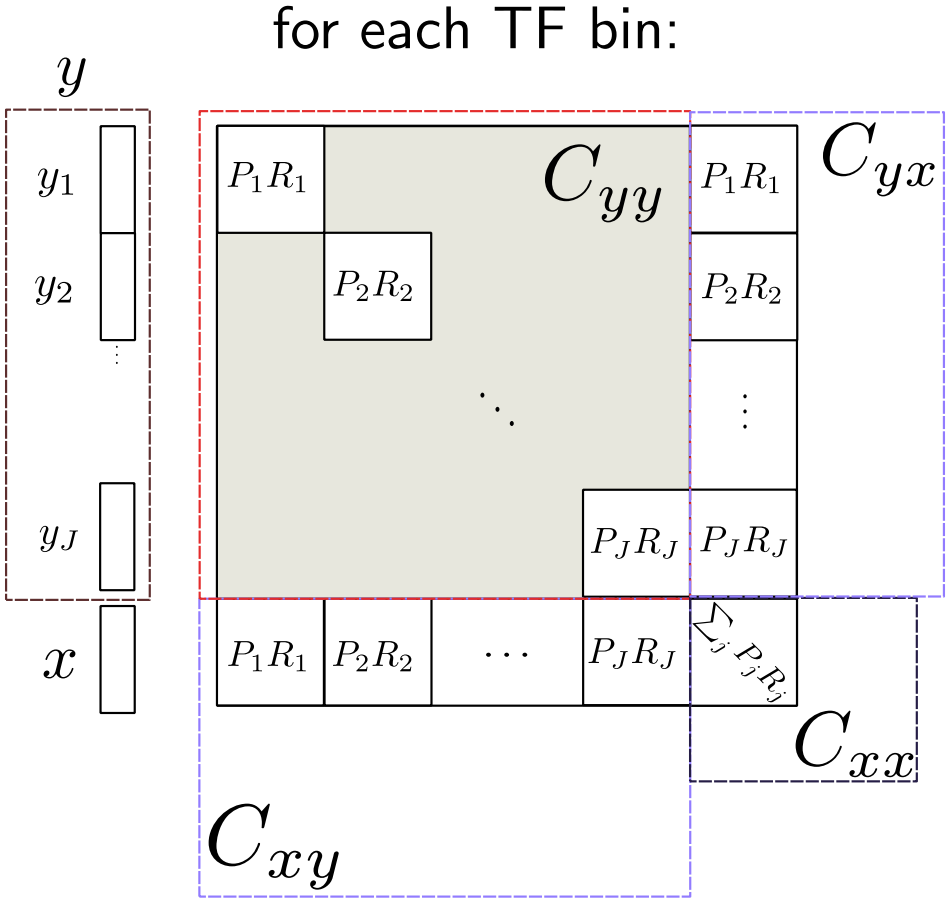
Mixture is sum of independent Gaussian images
$\Rightarrow$ Images and mixtures are jointly Gaussian :

$$\begin{bmatrix} y \\ x \end{bmatrix} \sim \mathcal{N}_c\left(0, \begin{bmatrix} A & C \\ C^H & B \end{bmatrix}\right)$$

A, B, C computed using
- power spectral density $P(f,n,j)$
- spatial covariance matrix $R_j(f)$

$\Rightarrow$ We can *easily* compute A, B and C with $\Theta = \{P(f,n,j), R_j(f)\}$

Images/mix joint distribution (2/2)



N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1830–1840, sept. 2010

Probability framework

1. *What kind of source signals s are plausible ?*
2. What pairs of (sources s,mixtures x) are plausible?
3. What can I say on s given x?

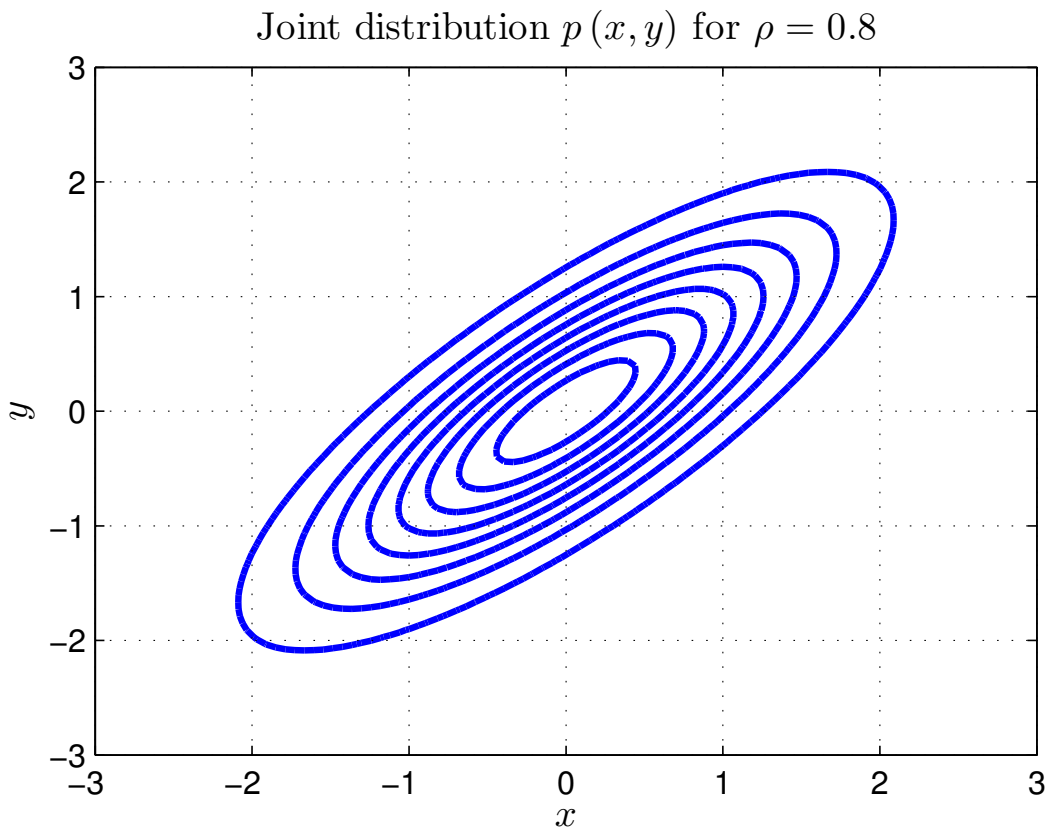
Recover images from mixtures

- We have $p(y, x \mid \Theta)$, which is **Gaussian** $\mathcal{N}_c\left(0, \begin{bmatrix} A & C \\ C^H & B \end{bmatrix}\right)$
- We want $p(y \mid x, \Theta)$
  Our knowledge about the images after observing the mixtures

Conditioning a Gaussian is **easy and yields a Gaussian.**

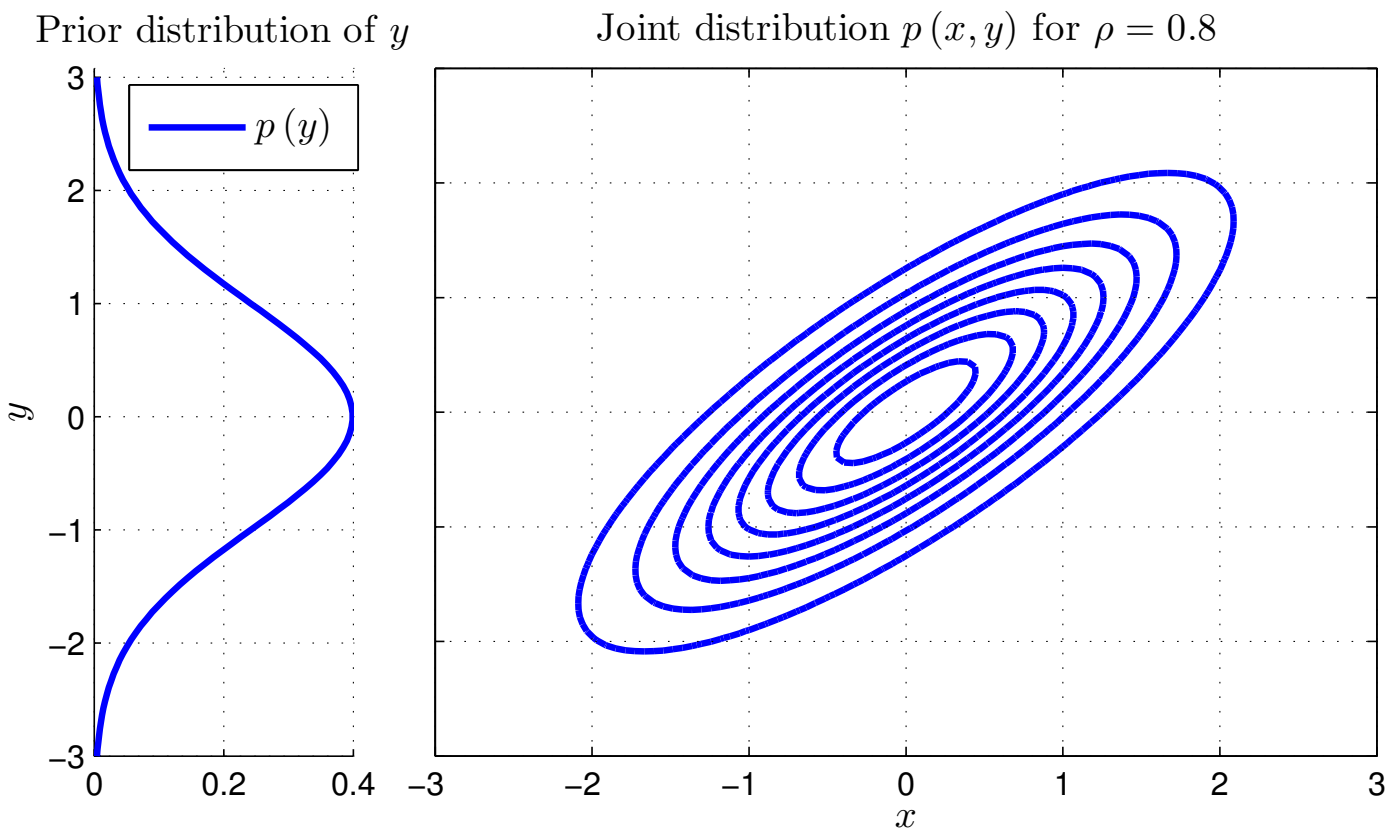
Gaussian conditioning

Example with $y, x \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$



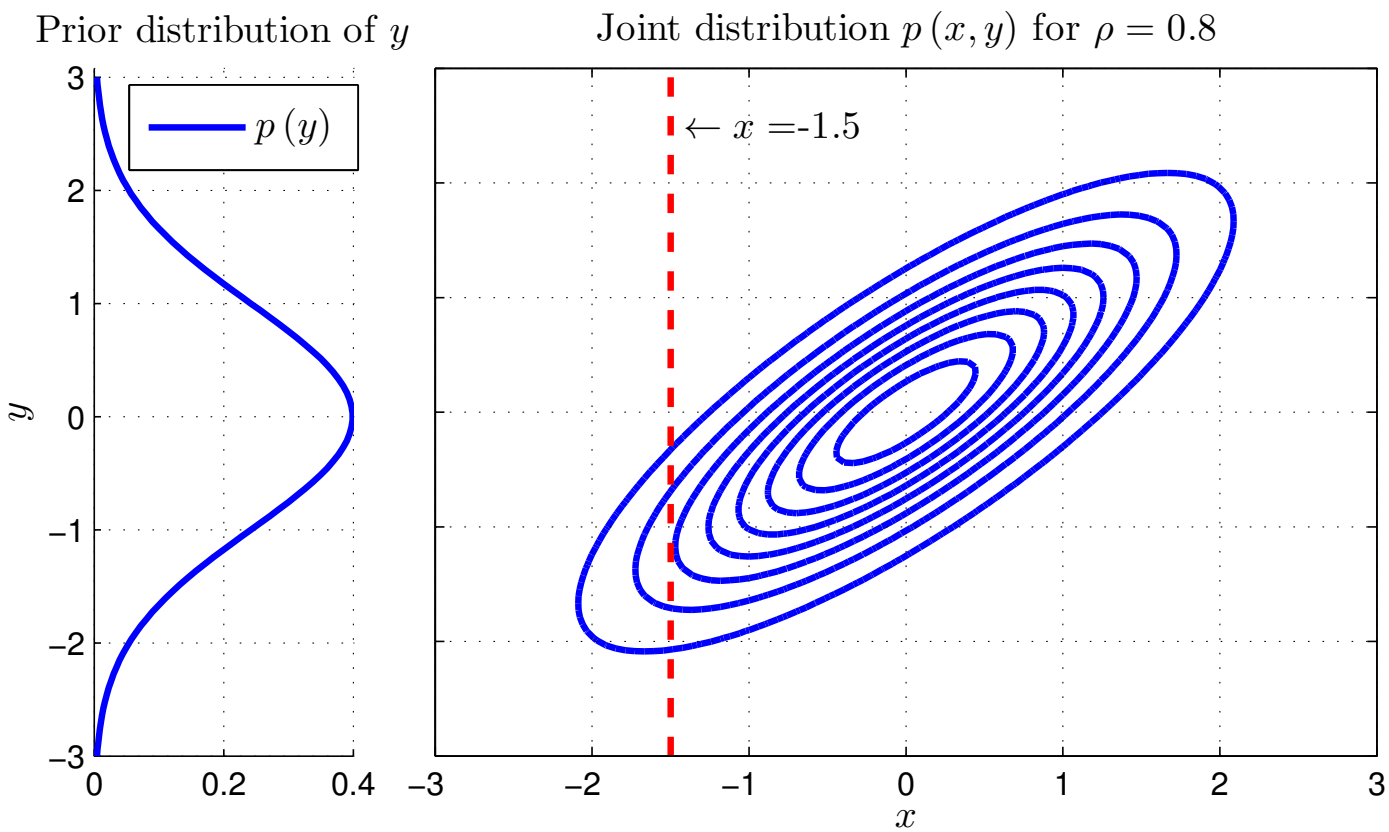
Gaussian conditioning

Example with $y, x \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$



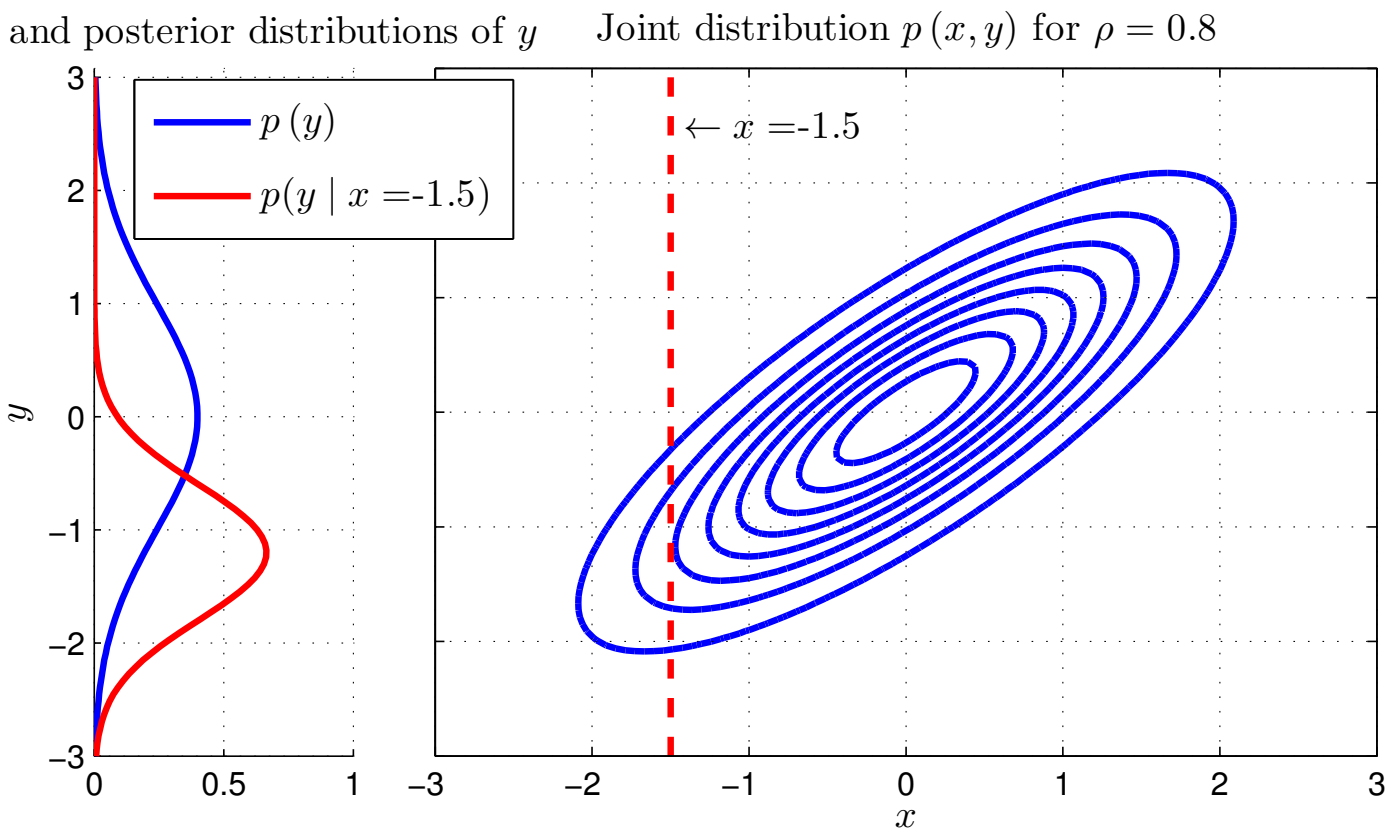
Gaussian conditioning

Example with $y, x \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$



Gaussian conditioning

Example with $y, x \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$



Gaussian conditioning

general case $y, x \sim \mathcal{N}\left(0, \begin{bmatrix} A & C \\ C^H & B \end{bmatrix}\right)$

we have:

$$y \mid x, \Theta \sim \mathcal{N}_c \left(\underbrace{CB^{-1}x}_{\mu_{y|x}}, A - CB^{-1}C^H \right).$$

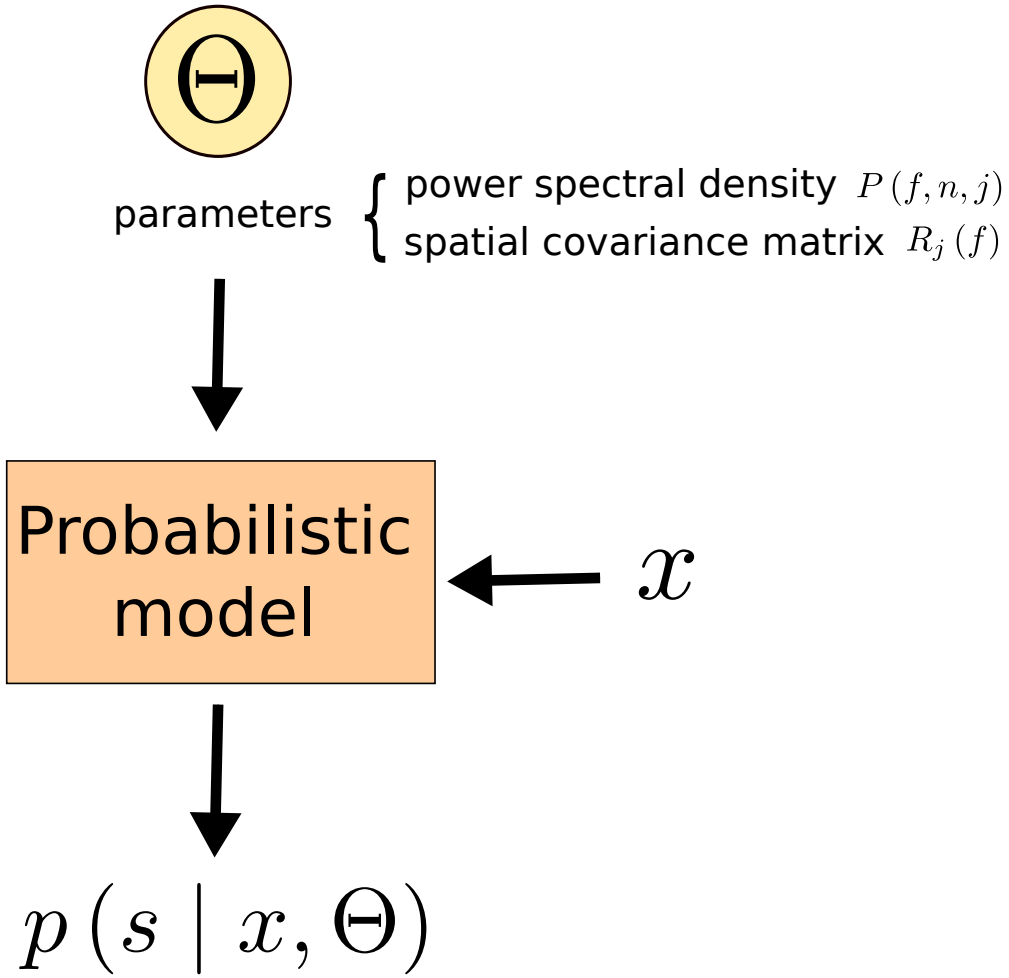
In the single channel case

- $x(f,n) = y_1(f,n) + y_2(f,n) + \cdots + y_J(f,n)$
- $\mu_{y_j|x}$ given by the classical Wiener filter:

$$\mu_{y_j|x}(f,n) = \frac{P_j(f,n)}{\sum_{j'=1}^{J} P_{j'}(f,n)} x(f,n),$$

extended easily to multichannel with P_j and R_j.

Probabilistic model : conclusions



Outline

1 Introduction

2 Probabilistic framework

3 Parametric Informed Source Separation

4 Posterior source coding

5 Conclusion

Parametric informed source separation

Proposed approach

1. At the coder: learn and quantize model Θ
   $\Rightarrow$ Source model: spectrograms P_j
   $\Rightarrow$ Mixing model: spatial covariances R_j
2. At the decoder
   $\Rightarrow$ Estimate $\hat{y} = \underset{y}{\operatorname{argmax}}\ p\left(y \mid x, \overline{\Theta}\right)$: Wiener filter

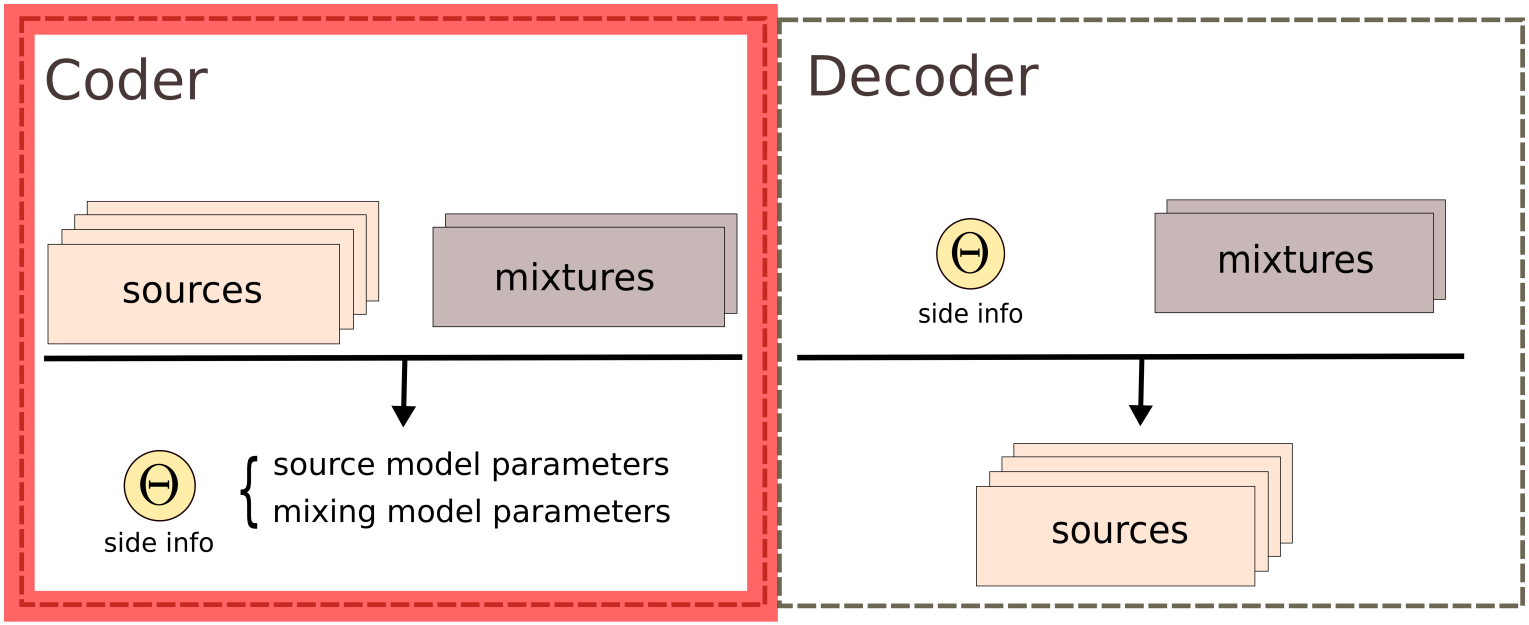
- Sources constructed as $\hat{s} = \mathcal{F}\left(x, \overline{\Theta}\right)$ at decoder.
- Analogy with audio parametric coding, but $\mathcal{F}$ is a filtering

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

A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard. Informed source separation : a comparative study. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, August 2012

A. Liutkus, R. Badeau, and G. Richard. Low bitrate informed source separation of realistic mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, may 2013

Objectives of the coder



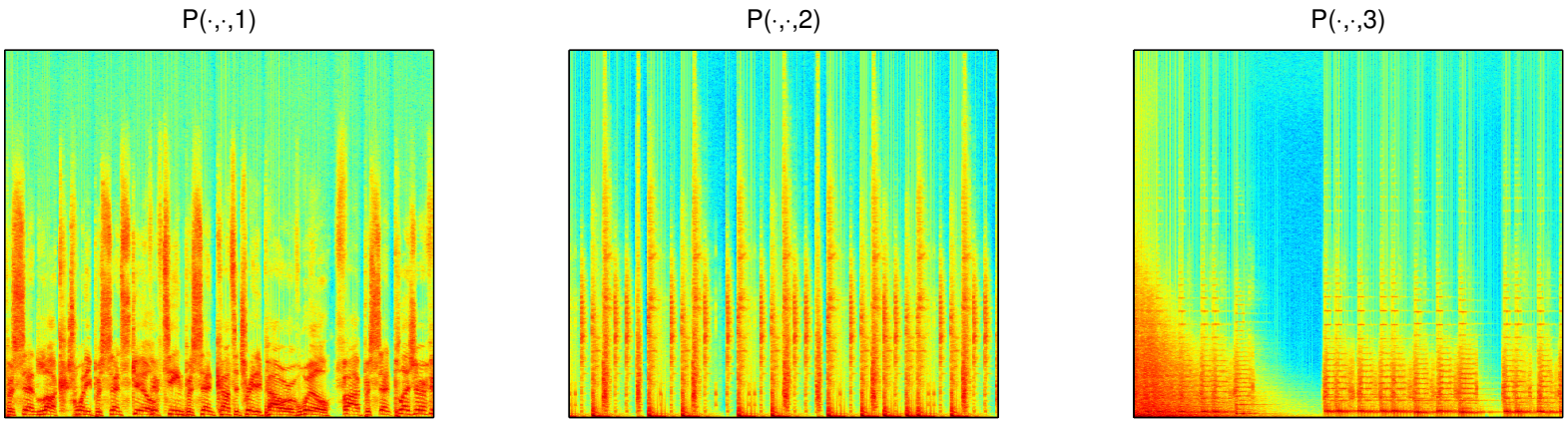
Model coder

1. Learning the source model
2. Learning the mixing model

Learning the PSD $P(f,n,j)$

Power spectrograms

- Coder observes $s(f,n,j) \sim \mathcal{N}_c(0, P(f,n,j))$
- Estimating the variance with a single realization?
  $\Rightarrow \hat{P}(f,n,j) = |s(f,n,j)|^2$: power spectrogram



- We have to transmit these PSD P to the decoder : **too big!**
- Two proposed techniques to compress them

Nonnegative Tensor Factorization

$$P(f,n,j) = \sum_{k=1}^{K} W(f,k)\, H(n,k)\, Q(j,k)$$

$(F+N+J)\,K$ parameters instead of FNJ

- K is the number of components
- W, H, Q are (small) nonnegative matrices

Maximum likelihood = minimization of Itakura-Saito divergence

$$\hat{W},\hat{H},\hat{Q} = \underset{W,H,Q}{\operatorname{argmin}} \sum_{f,n,j} d_{IS}\left(|s(f,n,j)|^2 \mid P(f,n,j)\right)$$

A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley Publishing, September 2009

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011

Nonnegative Tensor Factorization

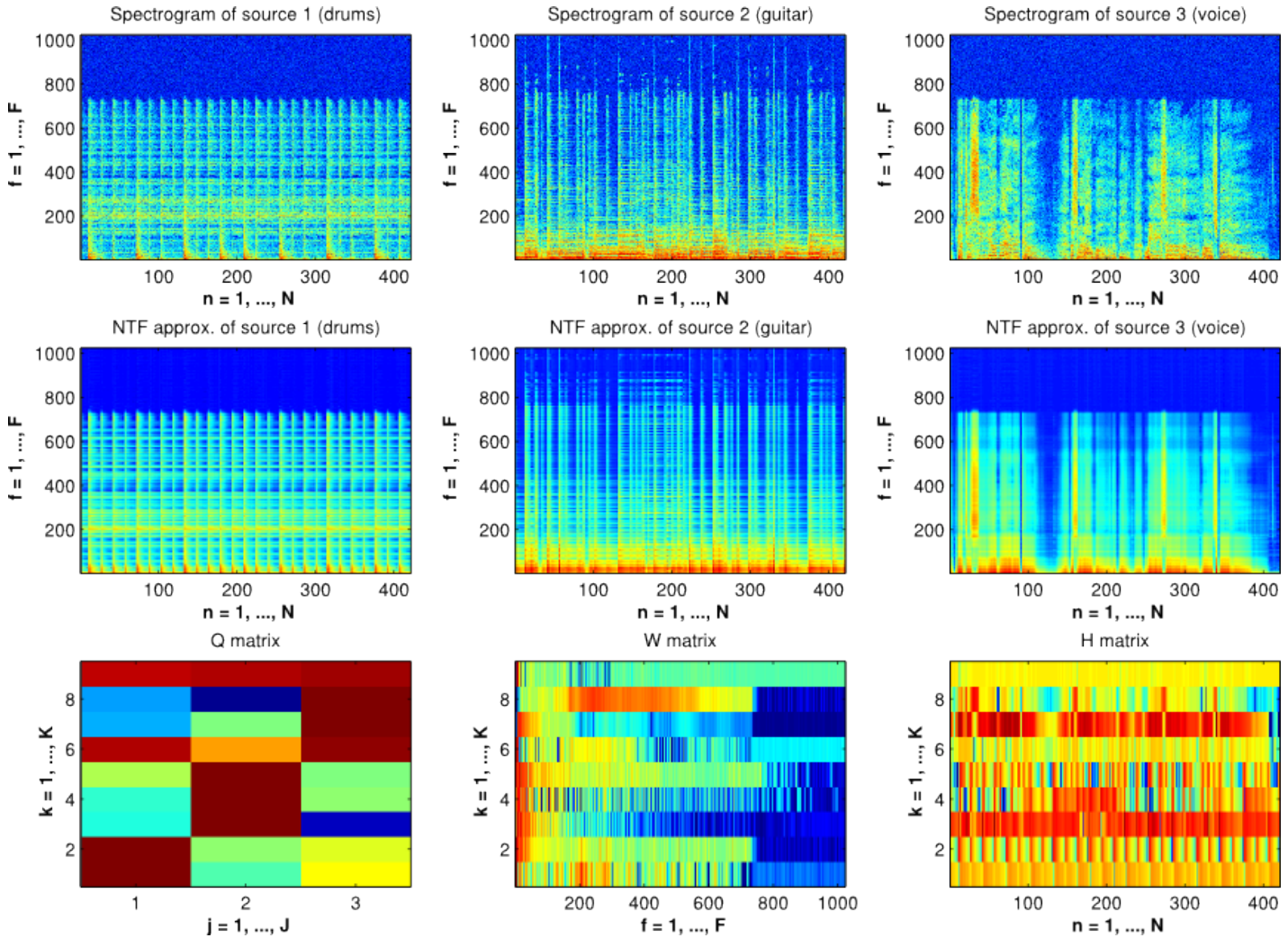
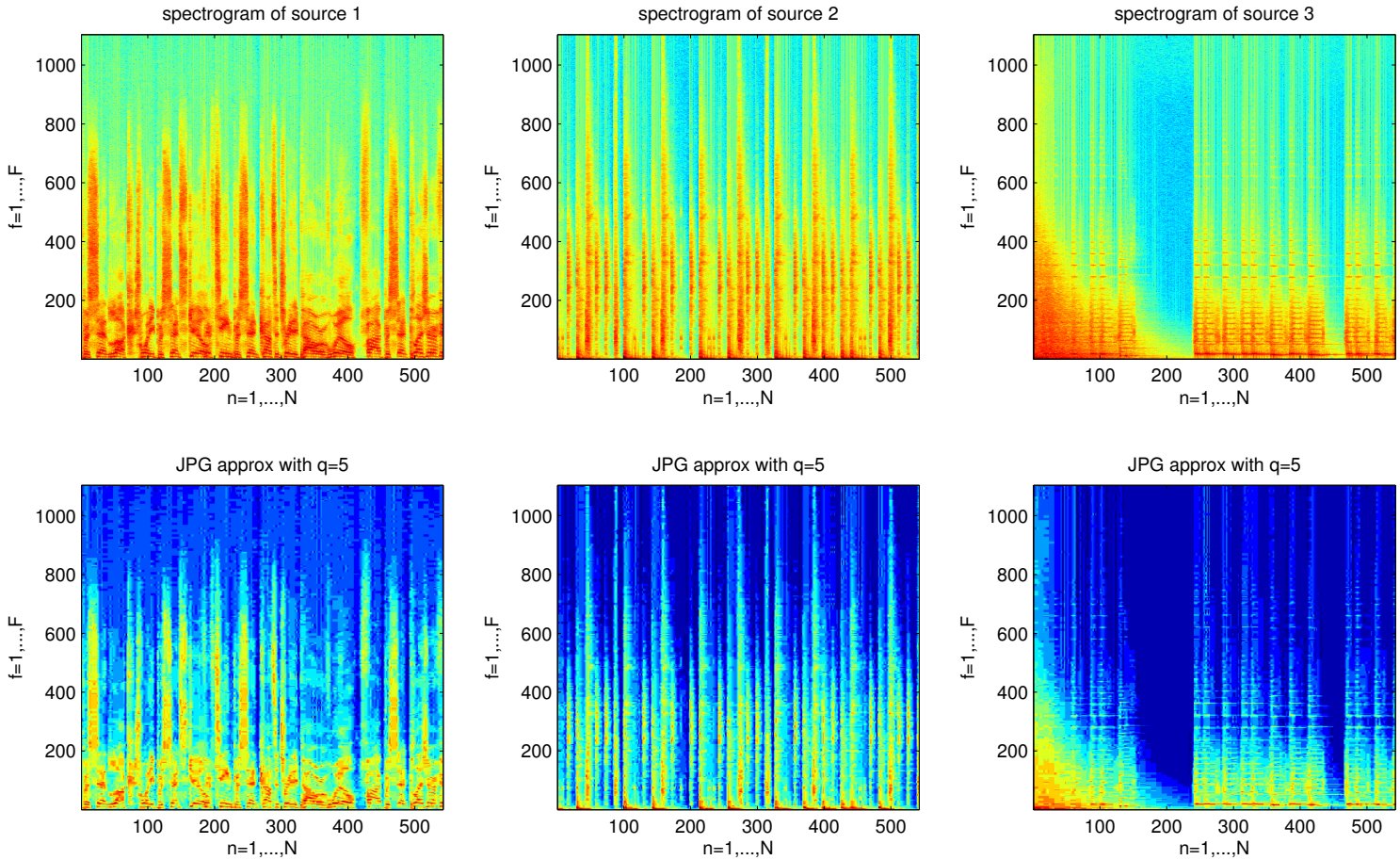


Image Compression



L. Girin, A. Liutkus, G. Richard, and R. Badeau. Procédé et dispositif de formation d'un signal mixé numérique audio, procédé et dispositif de séparation de signaux, et signal correspondant. Demande de brevet no. B10/3035FR / GBO, October 2010. Institut Polytechnique de Grenoble et Institut Télécom, Télécom ParisTech

Model coder

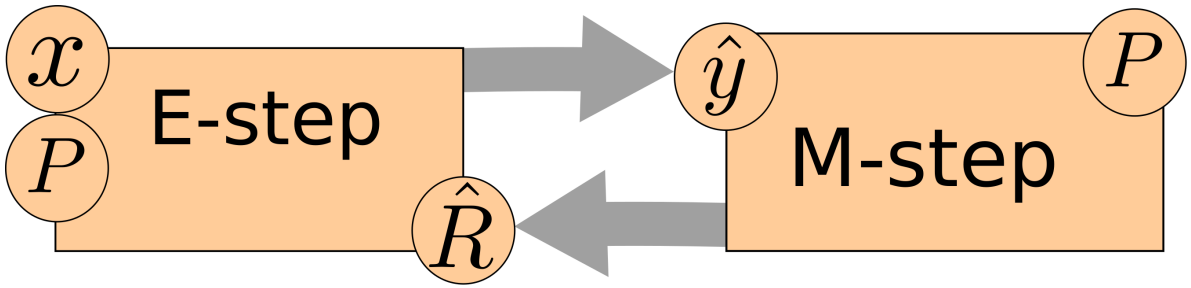
1. Learning the source model
2. Learning the mixing model

Mixing model estimation algorithm

- Mixture is a sum of independent Gaussian images

$$x(f,n,\cdot) \sim \mathcal{N}_c\left(0, \sum_{j=1}^{J} P(f,n,j)\, R_j(f)\right)$$

- **Expectation-Maximization (EM) algorithm**:
  we know P and x and we want $\hat{R} = \text{argmax}\; p(x \mid P, R)$



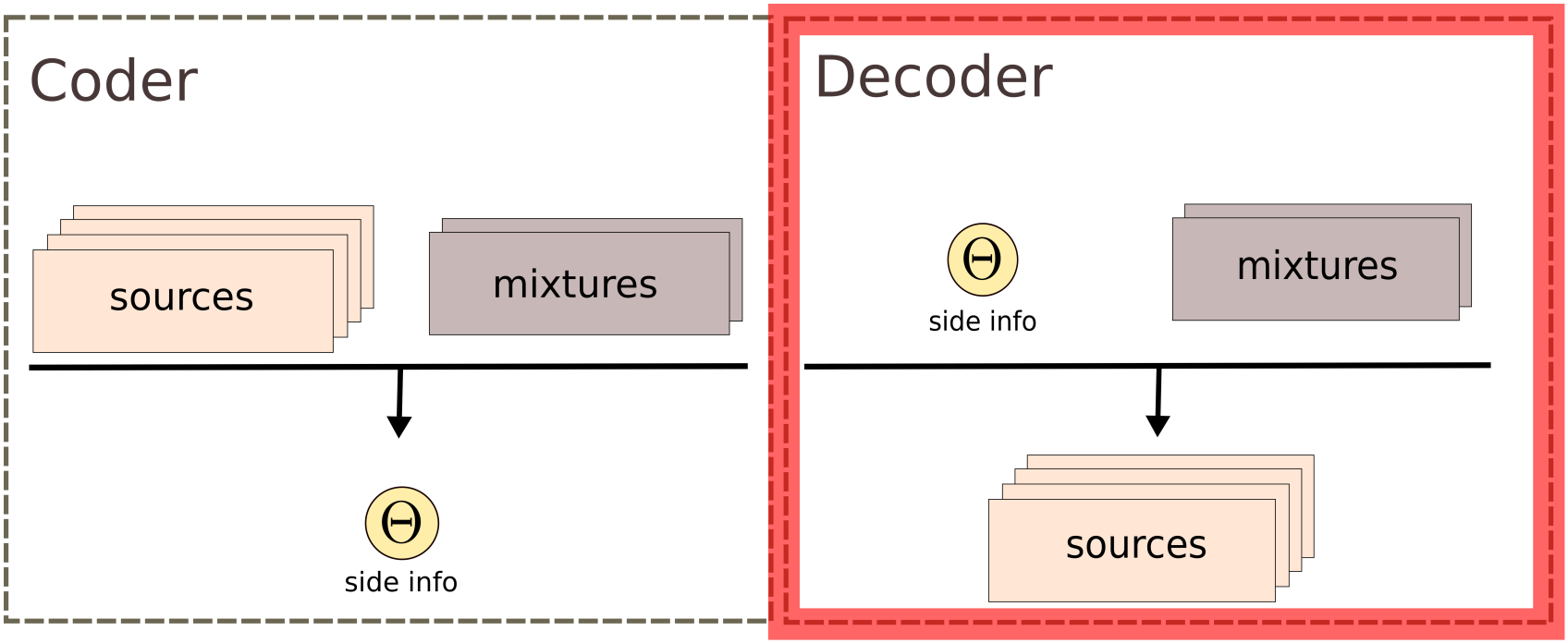
N.Q.K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7):1830–1840, sept. 2010

A. Liutkus, R. Badeau, and G. Richard. Low bitrate informed source separation of realistic mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, may 2013

Model coder : conclusion

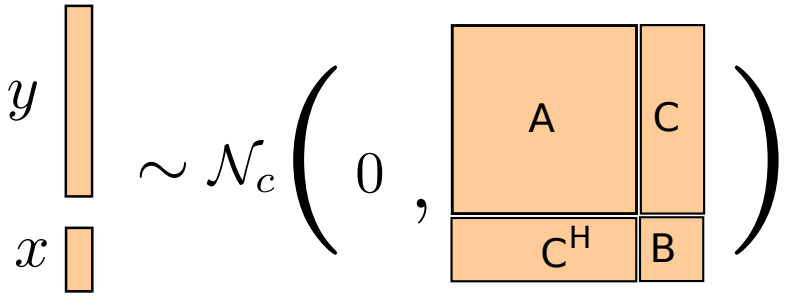
- Source model parameters: PSD $P(f, n, j)$
- Mixing model parameters: $R_j(f)$
- Typical bitrates : $1 - 15$kbps/source

Decoder: outlines



Recover images from mixtures

Images and mixtures are jointly Gaussian:



A, B, C computed using

- power spectral density $P(f,n,j)$
- spatial covariance matrix $R_j(f)$

Conditioning a Gaussian **yields a Gaussian:**

$$y \mid x, \Theta \sim \mathcal{N}_c\left(CB^{-1}x, A - CB^{-1}C^H\right)$$

Minimum mean squared error estimates of the sources are:

$$\hat{y} = CB^{-1}x$$

A. Liutkus, R. Badeau, and G. Richard. Informed source separation using latent components. In *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, St Malo, France, 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

Evaluation : Demonstration

- Professional mixture
- Low 4kbps/source bitrate
- Separation of the images

Evaluation : objective results

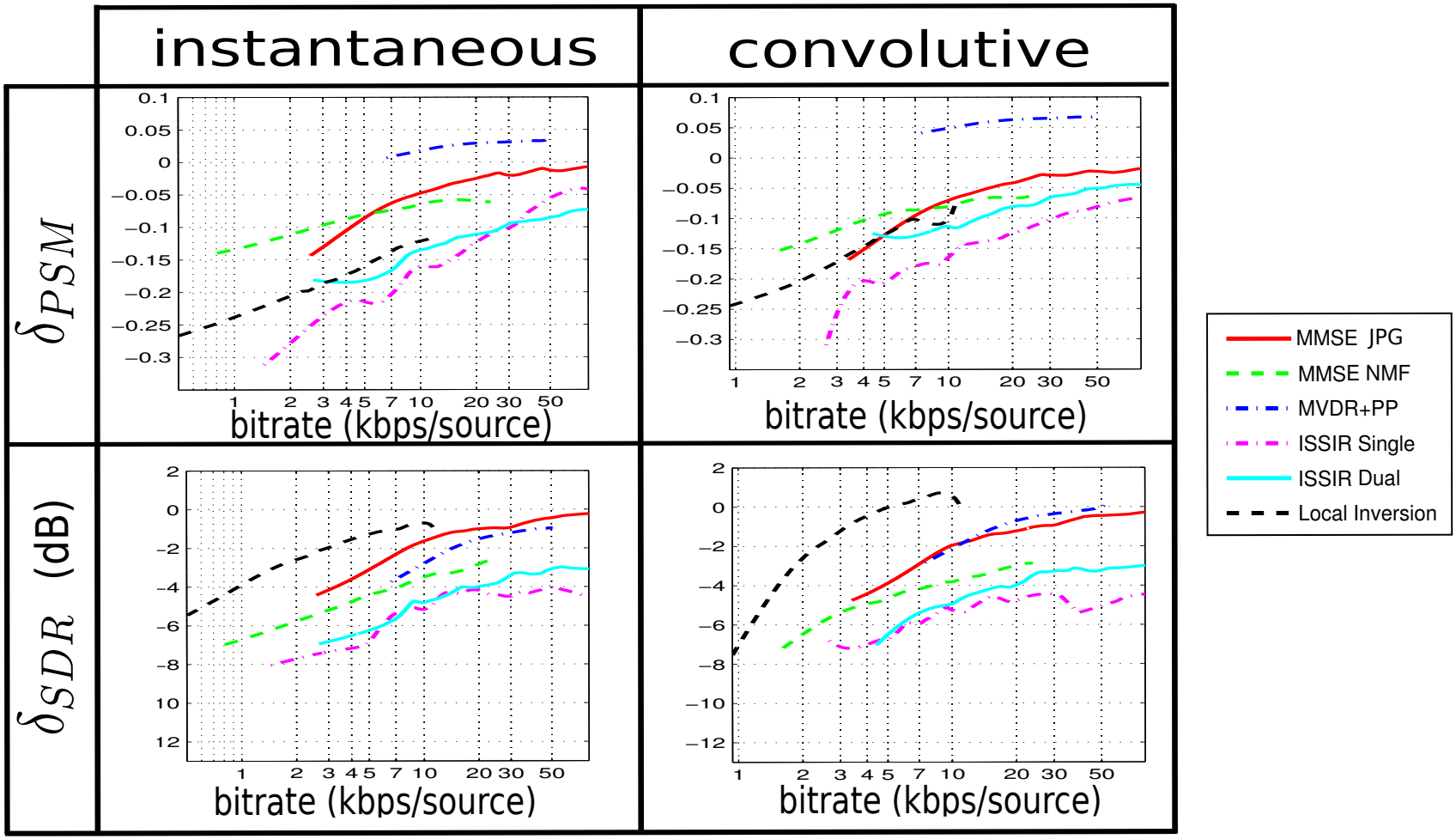
- Large evaluation
  - $\Rightarrow$ 15 excerpts with all constitutive tracks
  - $\Rightarrow$ 2 different mixing processes (instantaneous & convolutive)
  - $\Rightarrow$ Comparison of 6 informed source separation methods
  - $\Rightarrow$ 2 objective measures: PSM and SDR
  - $\Rightarrow$ 10 different quality per (excerpt,mix)
- Results displayed as averaged rate-quality curves

A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard. Informed source separation : a comparative study. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, August 2012

R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902 –1911, November 2006

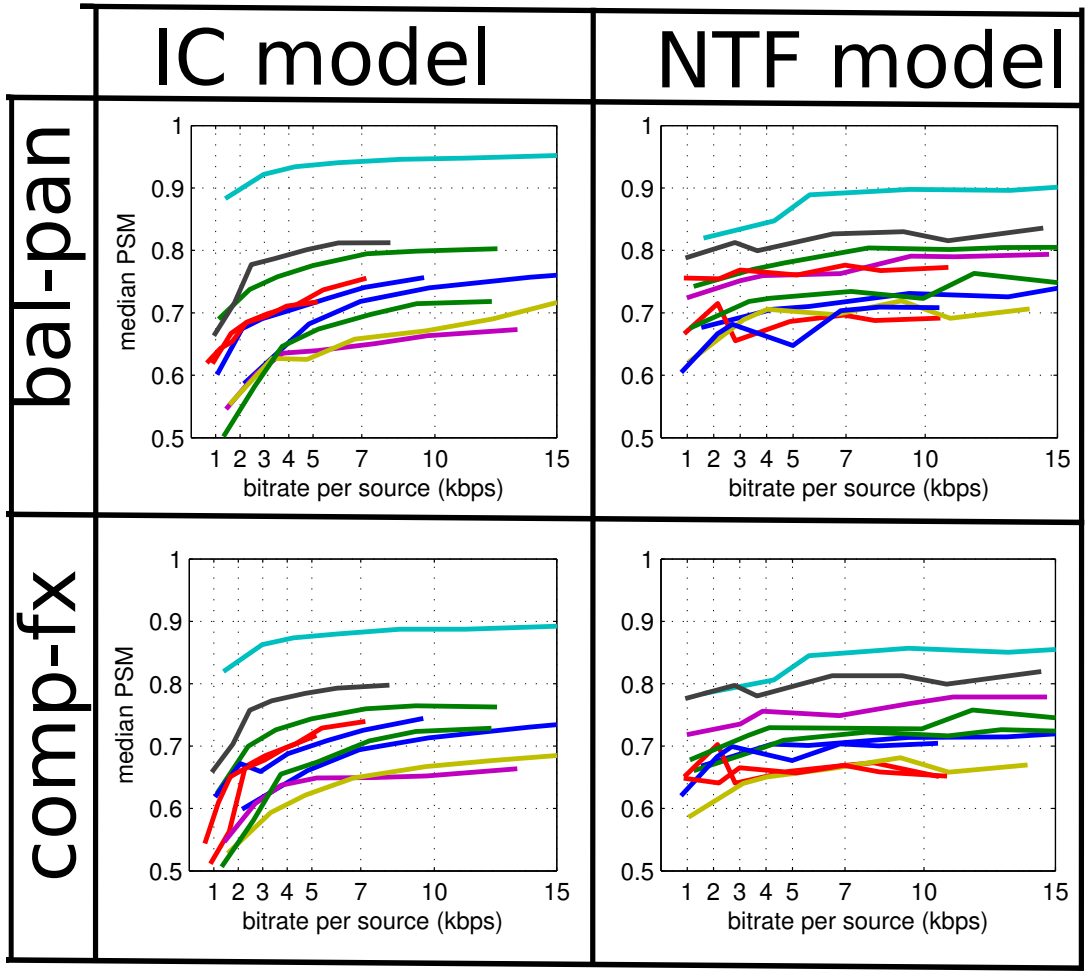
E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462 –1469, July 2006

Evaluation: comparative results



Evaluation: realistic mixtures

each line is a different excerpt



Evaluation : Conclusion

- Bitrate range: $1 - 15$kbps/source
  $\Rightarrow$Compare with 32 or 64 kbps AAC
- Professional mixtures well handled
  $\Rightarrow$ Not limited to instantaneous/convolutive laboratory mix

Better to separate rather than to independently encode the sources with low bitrate coders!

M. Parvaix and L. Girin. Informed source separation of linear instantaneous under-determined audio mixtures by source index embedding. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1721 –1733, August 2011

S. Gorlow and S. Marchand. Informed audio source separation using linearly constrained spatial filters. *IEEE Transactions on Audio, Speech and Language Processing*, 20(9), 2012

N. Sturmel and L. Daudet. Informed source separation using iterative reconstruction. arXiv:1202.2075v1

A. Liutkus, R. Badeau, and G. Richard. Low bitrate informed source separation of realistic mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, may 2013

Outline

1 Introduction

2 Probabilistic framework

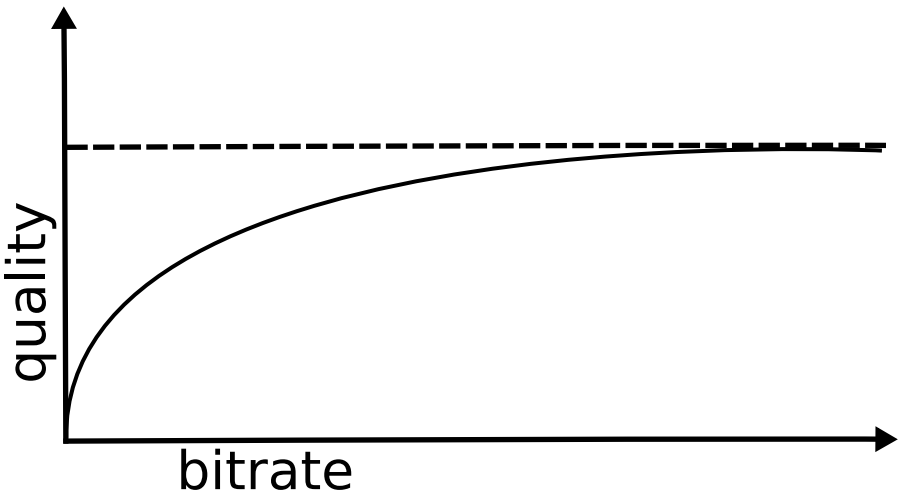
3 Parametric Informed Source Separation

4 Posterior source coding

5 Conclusion

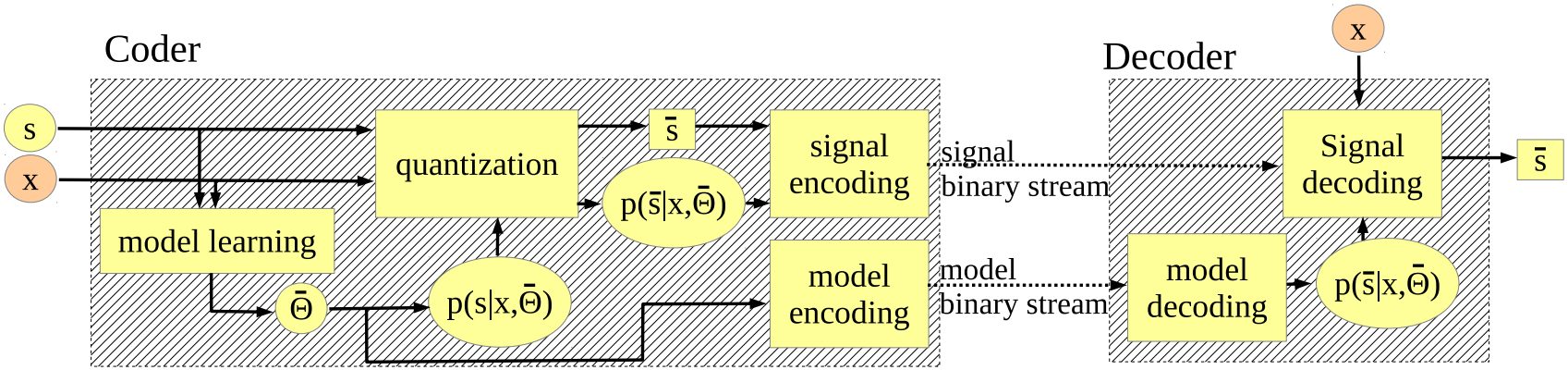
Bounded performances

- Problem with parametric decoder
  - Recovered sources are **never** the true ones
  - Performance is **bounded**



$\Rightarrow$ Higher bitrate does **not** yield higher performance !

Coding-based Informed Source Separation
block-diagram



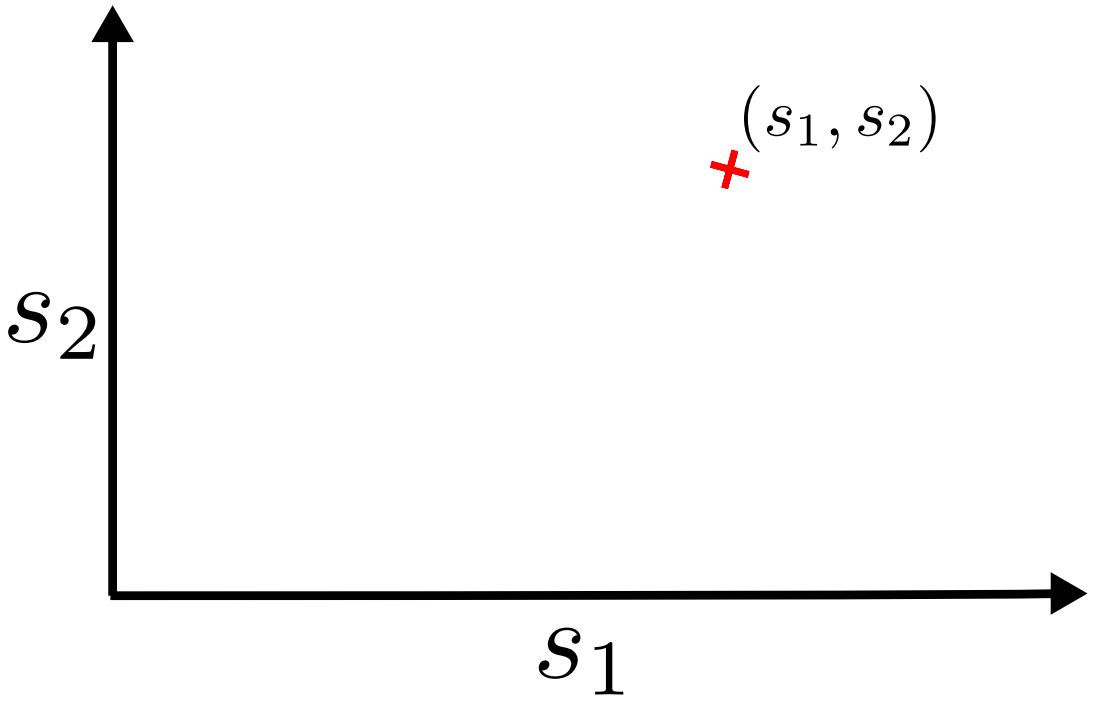
$$\text{Parametric ISS: } \hat{s} = \underset{s}{\operatorname{argmax}}\ p(s \mid x, \Theta)$$

Use $p(s \mid x, \Theta)$ for **source coding**

$\Rightarrow$ Unbounded distortion at minimal bitrate

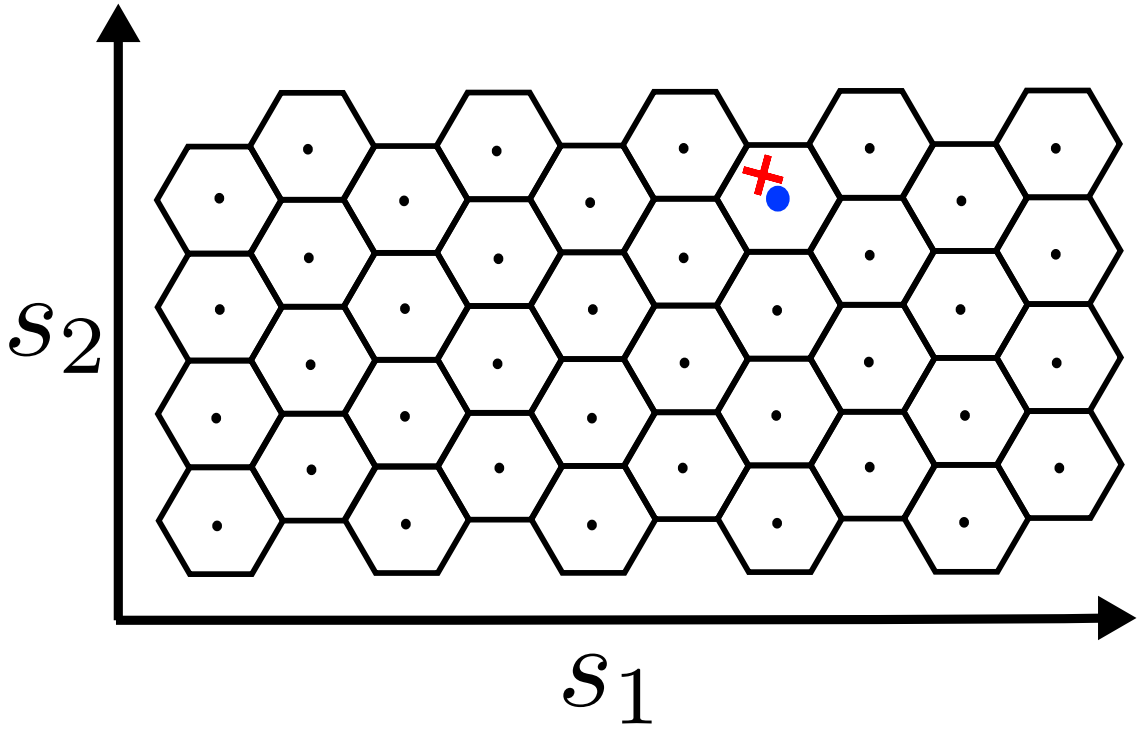
A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-based informed source separation: Nonnegative tensor factorization approach. *IEEE Trans. on Audio, Speech and Language Processing*, 2013

Quantization (1/2)



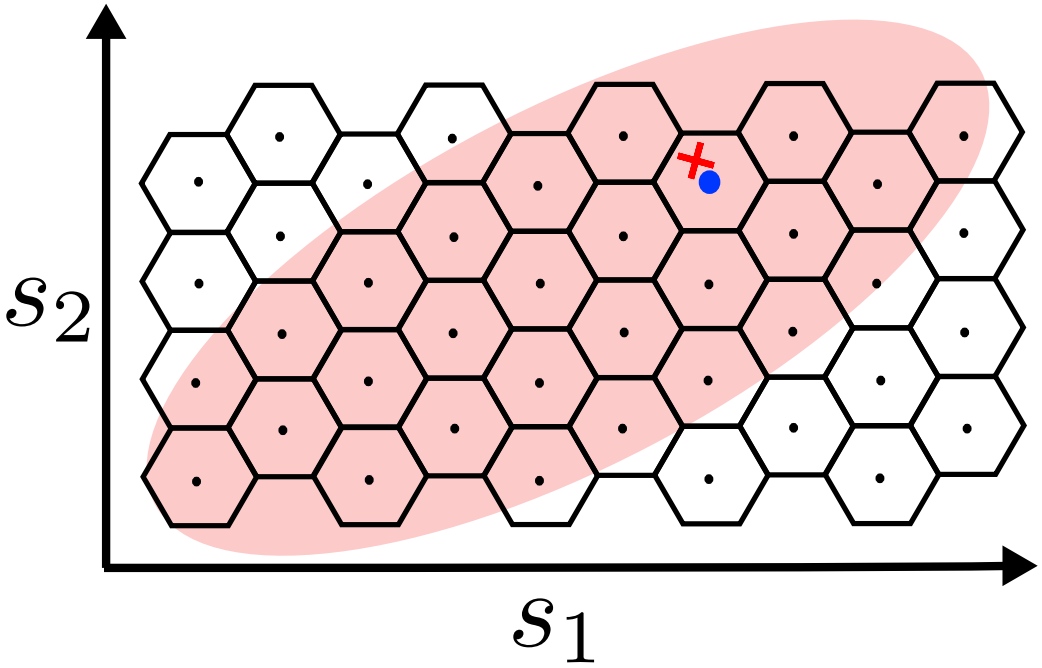
+ observed value

Quantization (2/2)



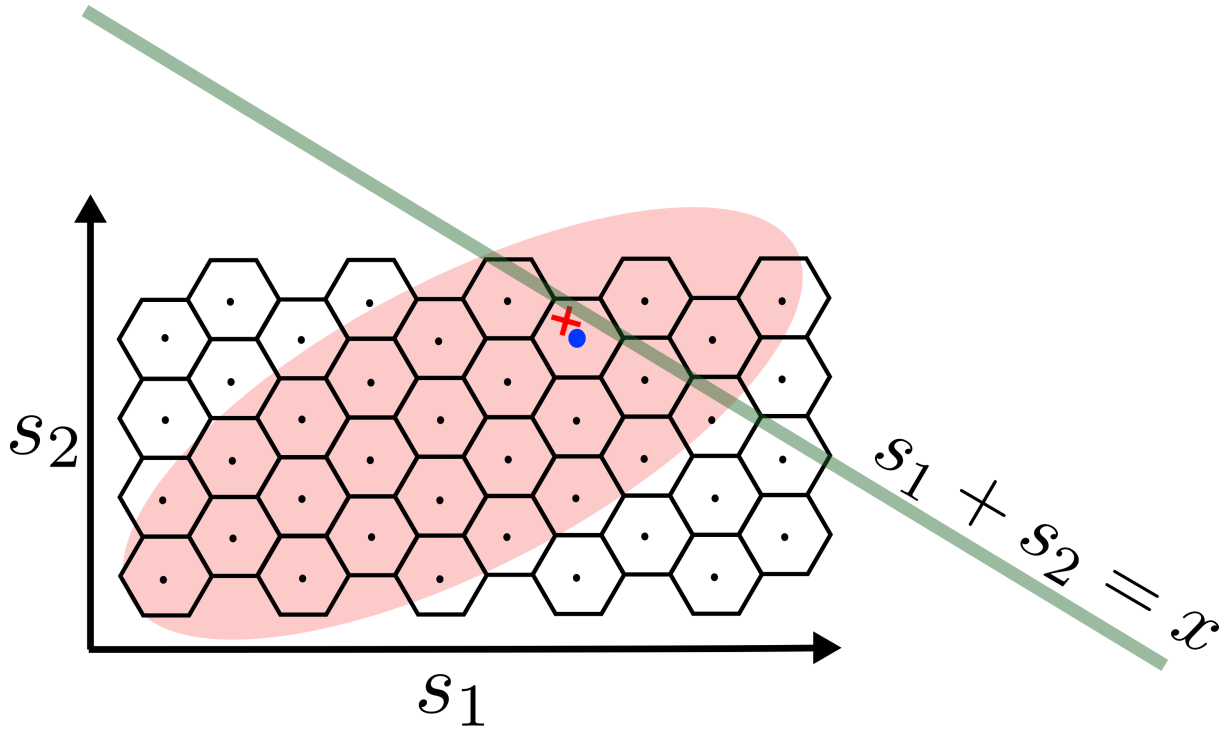
+ observed value
• quantized value

Source coding



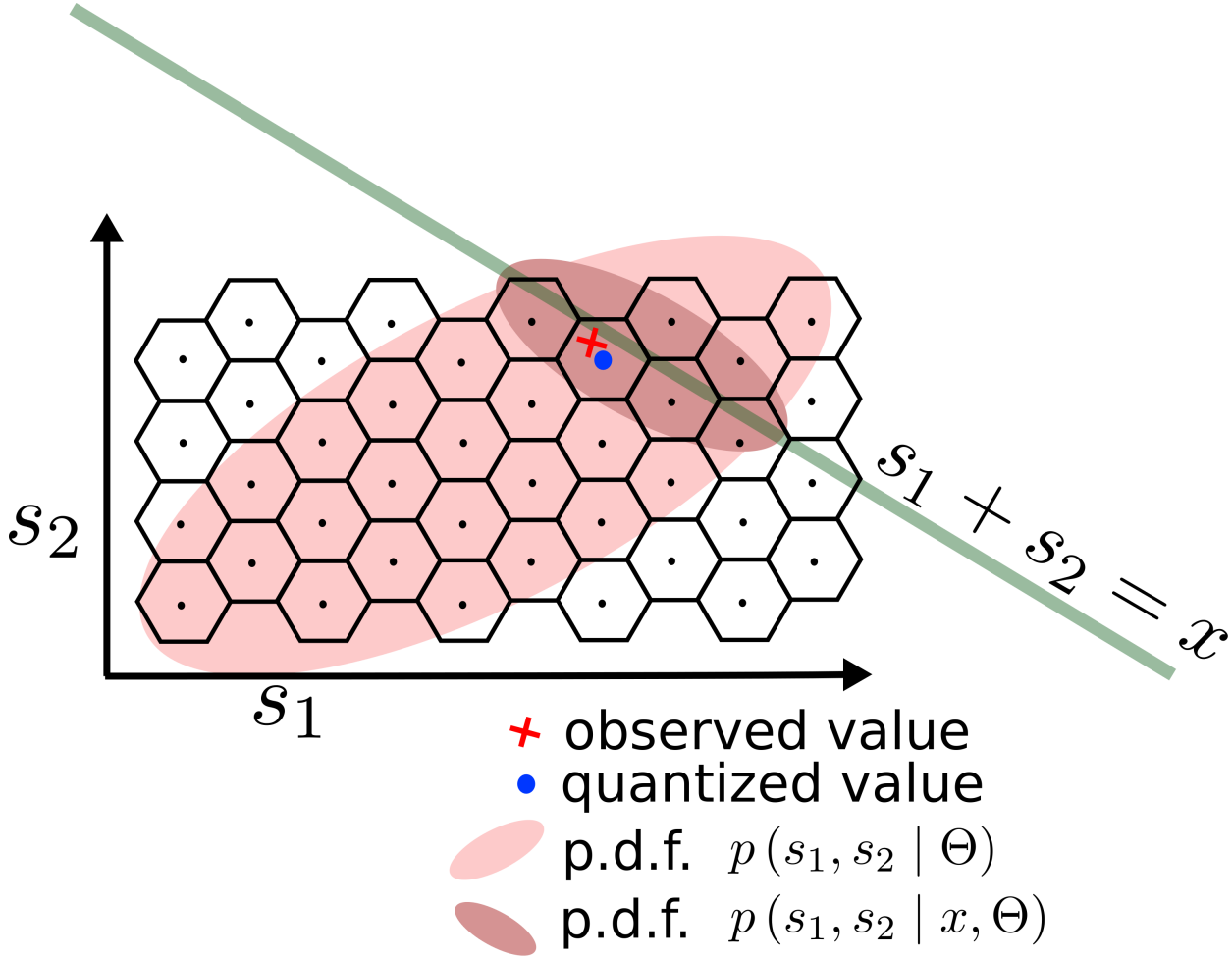
+ observed value
• quantized value
p.d.f $p(s_1, s_2 \mid \Theta)$

Introducing the mixture



observed value
quantized value
p.d.f $p(s_1, s_2 \mid \Theta)$

A posteriori source coding



Important results

- Practical coding/decoding algorithms
- Analytical operational rate-distortion function

$$\underbrace{R\left(D \mid \bar{\Theta}\right)}_{\text{total rate (bits)}} = \underbrace{R_{\bar{\Theta}}}_{\text{model bitrate}} \underbrace{- \log_2 p\left(s \mid x, \bar{\Theta}\right) - 2JFN \log_2 12D}_{\text{signal bitrate}}$$

- Bitrate gain compared to standard source coding:

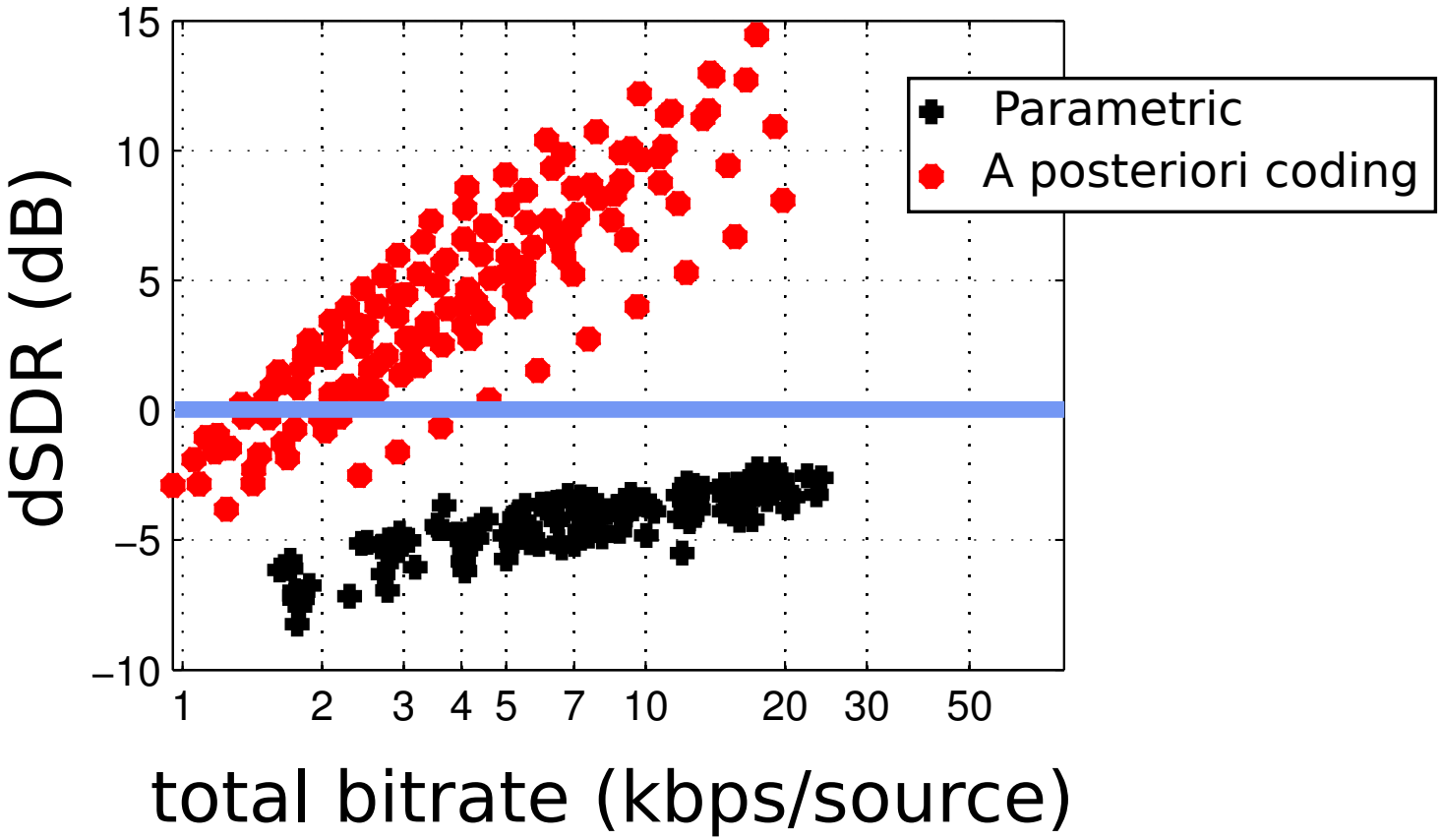
$$\sum_{f,n} \frac{1}{2}\left(\log\left|K\left(f,n\right)\right| - \log\left|K\left(f,n\right)_{\text{post}}\right|\right) \geq 0$$

A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Informed source separation: source coding meets source separation. In *IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA, October 2011

A. Liutkus, A. Ozerov, R. Badeau, and G. Richard. Spatial coding-based informed source separation. In *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, August 2012

A. Ozerov, A. Liutkus, R. Badeau, and G. Richard. Coding-based informed source separation: Nonnegative tensor factorization approach. *IEEE Trans. on Audio, Speech and Language Processing*, 2013

Evaluation results



Outline

1 Introduction

2 Probabilistic framework

3 Parametric Informed Source Separation

4 Posterior source coding

5 Conclusion

Coding-based informed source separation

- Source models
  - Powerful multichannel models for audio signals
  - Spectrograms may be compressed efficiently ($\approx 1-5$kbps/source)
- Mixing model
  - Diffuse models generalize instantaneous and convolutive
  - Easily estimated from data
- Informed source separation
  - A special-case of source coding
  - Good performance with small bitrates

Keynote content

- **Objective**
  - To provide an overview of major trends in Informed Source Separation (ISS)
- **Outline of the keynote**
  - *PART 1: Gaël RICHARD*
    - Introduction on Informed Source Separation
    - Outline of a popular (blind) source separation approach (based on Non-negative Matrix Factorization).
  - *PART 2: Alexey OZEROV*
    - *Auxiliary data-informed source separation,*
    - *User-guided source separation,*
  - *PART 3: Antoine LIUTKUS*
    - *Coding-based informed source separation*
  - ***PART 4: Gaël RICHARD***
    - **Conclusion**

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*

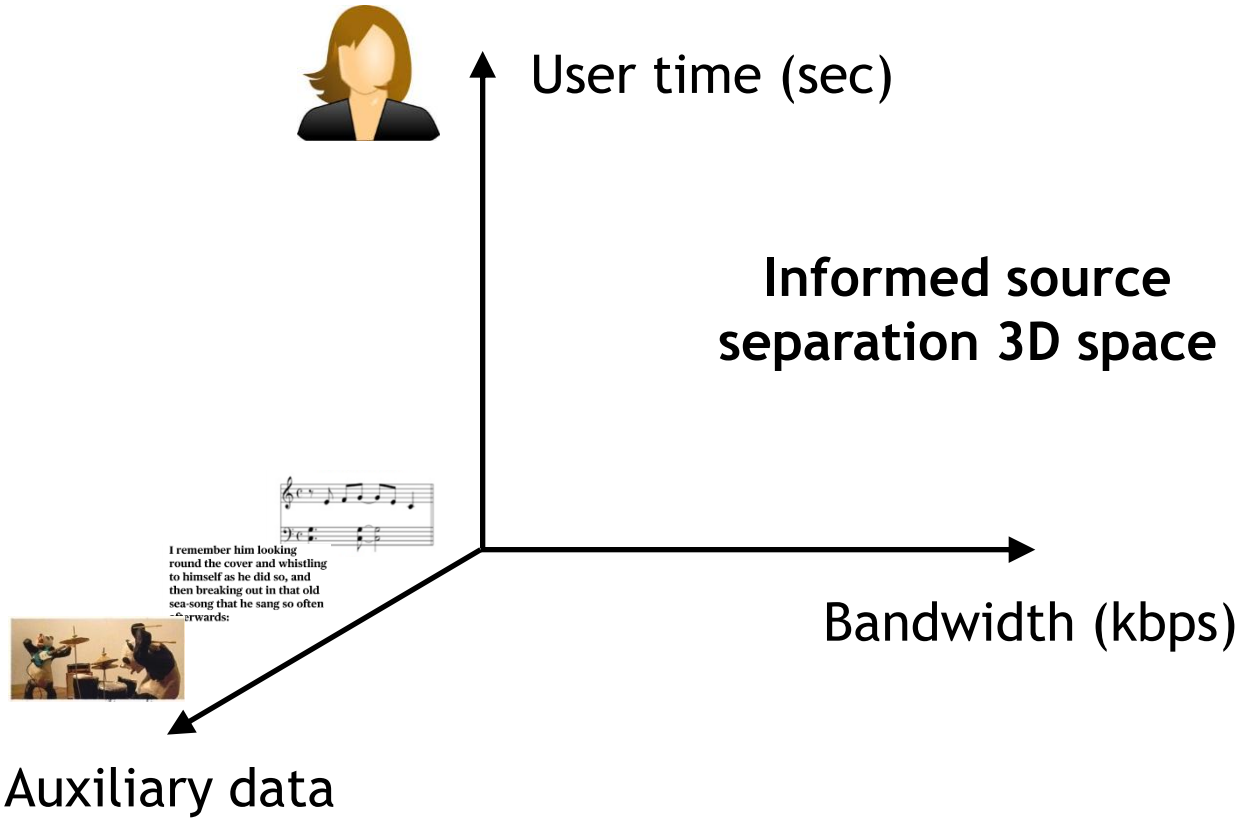
Conclusion

- **What could we expect from informed source separation and for which applications it is suitable?**

Trend	*Auxiliary data-informed source separation*	*User-guided source separation*	*Coding-based informed source separation*
Requirements	Auxiliary data availability	User intervention and user time	Clean sources and bandwidth
Expected separation quality	Depends on the auxiliary data, but usually not very high	Higher quality, given the user spends sufficient time	Any quality can be achieved, given a sufficiently large bandwidth
Applications	Any, given that some auxiliary data are available	When user intervention is possible, e.g., audio post-production	When clean sources are available, e.g., audio production

Conclusion

- **Could we mix up different informed source separation trends?**
- **Yes, when several conditions are met altogether**



Perspectives

- **Auxiliary data-informed source separation**
  - Which other sources of information could be exploited as auxiliary data?
  - How to better exploit a multitrack cover version to perform source separation on the original recording?
- **User-guided source separation**
  - Which other user-machine interactions could be used for source separation?
  - Could we attend a “perceptually perfect” source separation result, given that the user is allowed spending whatever time he/she needs?
- **Coding-based informed source separation**
  - How to introduce perceptual coding aspect into the coding-based informed source separation?
  - How to extend the concept of the coding-based informed source separation to the coding-based informed re-mixing, where the user is interested in the quality of the possible set of mixes, not the sources?
- **Altogether**
  - 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?

Some Software

FASST - Flexible Audio Source Separation Toolbox

http://bass-db.gforge.inria.fr/fasst/

BHADASSS - Blind Harmonic Adaptive Decomposition Applied to Supervised Source Separation

http://www.benoit-fuentes.fr/sharedCode/BHADASSS.zip

KAM - Kernel Additive Modelling for source separation

www.loria.fr/~aliutkus/kam/

ISSE - An Interactive Source Separation Editor

http://isse.sourceforge.net/

User-guided Source Separation

http://www.durrieu.ch/research/lvaica2012.html

Thank you!

Questions?

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