A General Modular Framework for Audio Source Separation

Alexey Ozerov, Emmanuel Vincent and Frederic Bimbot

VERSAMUS 2010 workshop / IRISA - 18/08/2010

Outline

- Introduction
- Framework presentation
- Experimental illustrations
- Conclusion and further work

Introduction

- Classical audio source separation methods are usually adapted to a particular scenario :
  - ***problem dimensionality*** ((over)determined, underdetermined, and single-channel case),
  - ***mixing process characteristics*** (synthetic instantaneous, anechoic, and convolutive mixtures, and live recorded mixtures),
  - ***source characteristics*** (speech, singing voice, drums, bass, noise, ...)

Introduction

- Limitations of classical approaches
  - No common formulation
    - Difficult to adapt a method to a different scenario, it was not originally conceived for
  - Developing a new method for a new scenario is time-consuming :
    - Modeling
    - Algorithm design
    - Programming
    - ...

Introduction

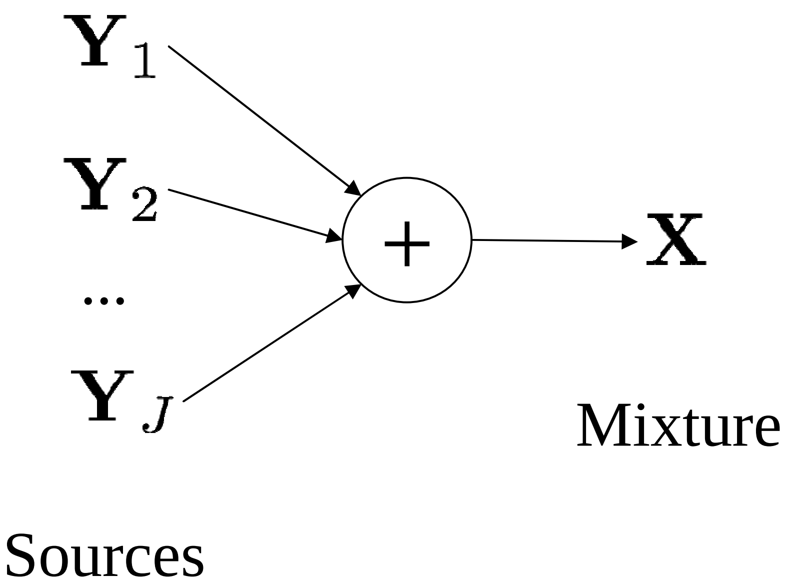
- To overcome these issues we would like to develop a new framework that should be
  - ***general***, generalizing existing methods and making it possible to combine them,
  - ***flexible***, allowing easy incorporation of the a priori information about a particular scenario considered,
  - ***modular***, allowing an implementation in terms of software blocks addressing the estimation of subsets of parameters,

Outline

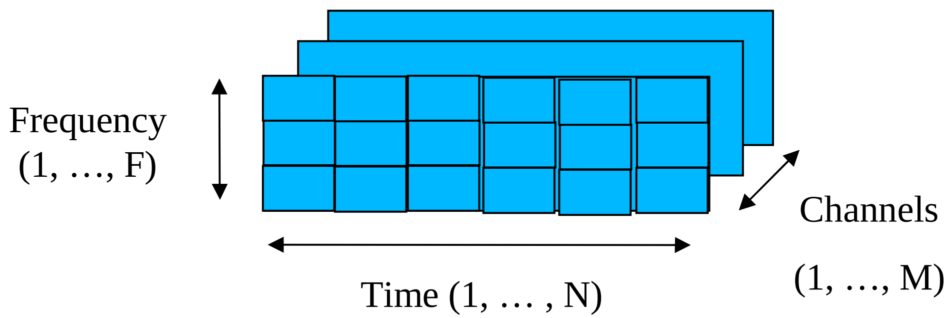
- Introduction
- Framework presentation
- Experimental illustrations
- Conclusion and further work

Audio source separation

$$\mathbf{Y}_j, \mathbf{X} \in \mathbb{C}^{F \times N \times M}$$



STFT



Flexible model

$$\mathbf{Y}_j = \{\mathbf{y}_{j,fn}\}_{f,n} \in \mathbb{C}^{F\times N\times M} \qquad \mathbf{y}_{j,fn} \in \mathbb{C}^{M}$$

$$\boxed{\mathbf{y}_{j,fn} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right)}$$

time-varying spatial covariance $\mathbf{R}_{j,fn} \in \mathbb{C}^{I\times I}$

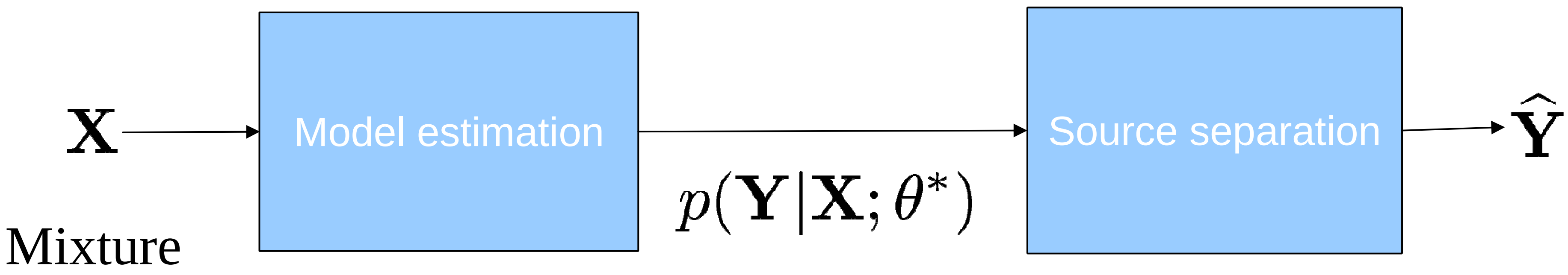
time-varying spectral power $v_{j,fn} \in \mathbb{R}_+$

$$\theta_j = \{v_{j,fn}, \mathbf{R}_{j,fn}\}_{f,n=1}^{F,N} \qquad \theta = \{\theta_j\}_{j=1}^{J}$$

source model

model

Global scheme



Source separation

$$\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n} \in \mathbb{C}^{F \times N \times M}$$

$$\boxed{\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\boldsymbol{\Sigma}_{\mathbf{x},fn}^{-1}(\theta)\mathbf{x}_{fn}}$$

$$\boldsymbol{\Sigma}_{\mathbf{x},fn}(\theta) \triangleq \sum_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn}$$

Maximum a posteriori (MAP) model estimation

$$\mathbf{X} = \{\mathbf{x}_{fn}\}_{f,n} \in \mathbb{C}^{F \times N \times M}$$

$$\boxed{\theta^* = \arg\min_{\theta \in \Theta} \sum_{f,n} \left[\mathrm{tr}\left(\mathbf{\Sigma}_{\mathbf{x},fn}^{-1}(\theta) \mathbf{x}_{fn} \mathbf{x}_{fn}^H \right) + \log\det \mathbf{\Sigma}_{\mathbf{x},fn}(\theta) \right] - \log p(\theta)}$$

Structure

Prior

Spatial Covariance Structures

- Time invariant $$\mathbf{R}_{j,fn} = \mathbf{R}_{j,f}$$
- Rank :
  - Rank-1 $$\mathbf{R}_{j,f} = \begin{bmatrix} h_{f,1}h^*_{f,1} & h_{f,1}h^*_{f,2} \\ h_{f,2}h^*_{f,1} & h_{f,2}h^*_{f,2} \end{bmatrix}$$
  - Full-rank $$\mathbf{R}_{j,f} = \begin{bmatrix} r_{f,11} & r_{f,12} \\ r^*_{f,12} & r_{f,22} \end{bmatrix}$$
- Mixing type
  - Linear instantaneous $$\mathbf{R}_{j,f} = \mathbf{R}_j$$
  - Convolutive
- Adaptive or fixed

Spectral Power Structures

- Excitation / Filter

$$v_{j,fn} = v_{j,fn}^{\text{excit}} \times v_{j,fn}^{\text{filt}}$$

$$v_{j,fn}^{\text{excit}} = \sum_{k=1}^{K_{\text{excit}}} p_{j,kn}^{\text{excit}} e_{j,fk}^{\text{excit}}$$

NMF

Activation coefficients

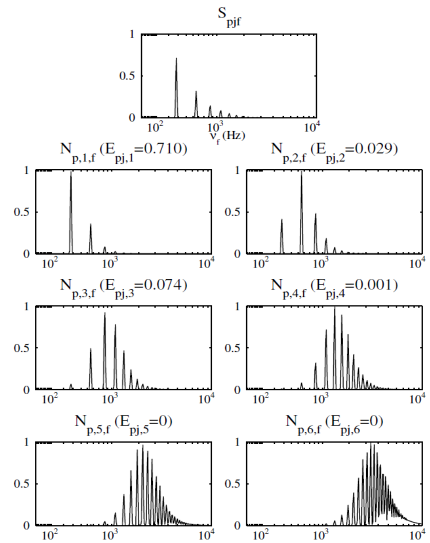
Characteristic spectral patterns

Spectral Power Structures

$$v_{j,fn}^{\text{excit}} = \sum_{k=1}^{K_{\text{excit}}} p_{j,kn}^{\text{excit}} e_{j,fk}^{\text{excit}}$$

$$v_{j,fn}^{\text{excit}} = \sum_{k=1}^{K_{\text{excit}}} \sum_{m=1}^{M_{\text{excit}}} h_{j,mn}^{\text{excit}} g_{j,km}^{\text{excit}} \sum_{l=1}^{L_{\text{excit}}} u_{j,lk}^{\text{excit}} w_{j,fl}^{\text{excit}}$$

$$\mathbf{V}_j^{\text{excit}} = \mathbf{W}_j^{\text{excit}} \, \mathbf{U}_j^{\text{excit}} \, \mathbf{G}_j^{\text{excit}} \, \mathbf{H}_j^{\text{excit}}$$

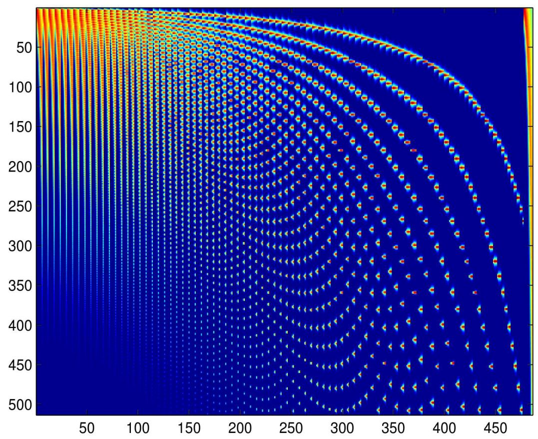


Spectral Power Structures

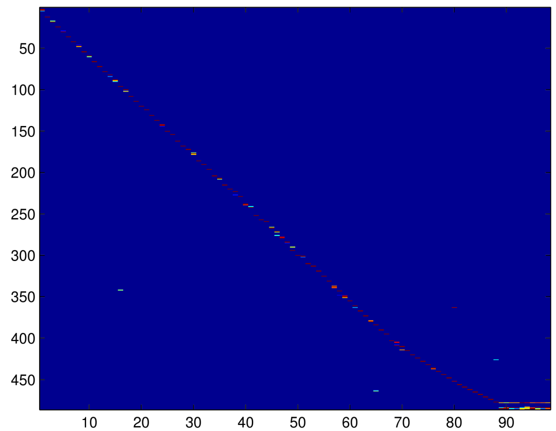
$$\mathbf{V}_j = \left(\mathbf{W}_j^{\text{excit}}\,\mathbf{U}_j^{\text{excit}}\,\mathbf{G}_j^{\text{excit}}\,\mathbf{H}_j^{\text{excit}}\right) \odot \left(\mathbf{W}_j^{\text{filt}}\,\mathbf{U}_j^{\text{filt}}\,\mathbf{G}^{\text{filt}}\,\mathbf{H}_j^{\text{filt}}\right)$$

- Each matrix can be fixed or adaptive
- Example

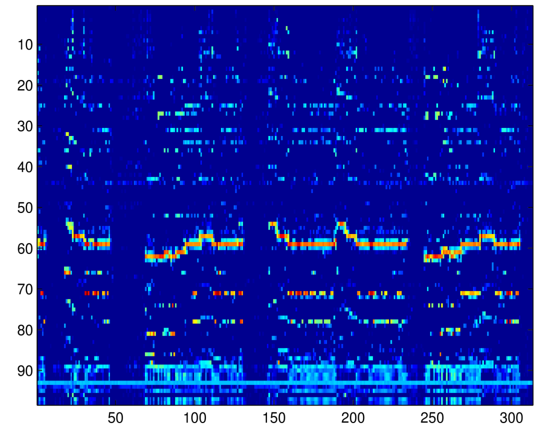
$$\mathbf{V}_j = \mathbf{W}_j^{\text{excit}}\,\mathbf{U}_j^{\text{excit}}\,\mathbf{H}_j^{\text{excit}}$$



Fixed



Adaptive

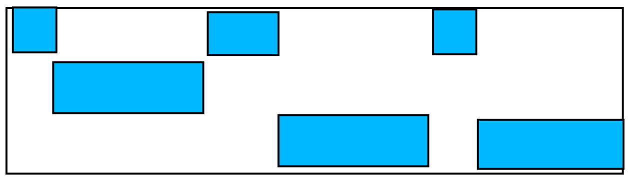


Adaptive

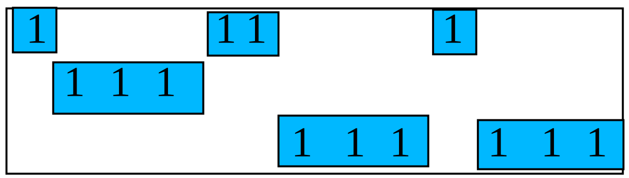
Spectral Power Structures

- Other structures on G or H matrix

NMF



SGMM / HSMM



GMM / HMM

Modular implementation

- Model

$$\theta = \{\theta_j\}_{j=1}^{J}$$

$$\theta_j = \{\theta_j^m\}_{m=1}^{9} =$$

$$= \{\mathbf{R}_j, \mathbf{W}_j^{\mathrm{excit}}, \mathbf{U}_j^{\mathrm{excit}}, \mathbf{G}_j^{\mathrm{excit}}, \mathbf{H}_j^{\mathrm{excit}}, \mathbf{W}_j^{\mathrm{filt}}, \mathbf{U}_j^{\mathrm{filt}}, \mathbf{G}_j^{\mathrm{filt}}, \mathbf{H}_j^{\mathrm{filt}}\}$$

- Generalized Expectation-Maximization algorithm with NMF updates
  - M-step : Loop over all (J x 9) parameters

Outline

- Introduction
- Framework presentation
- Experimental illustrations
- Conclusion and further work

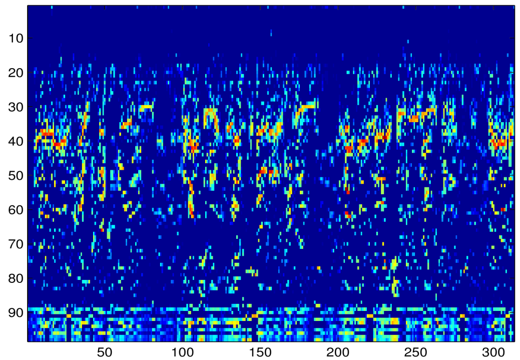
Experimental illustrations

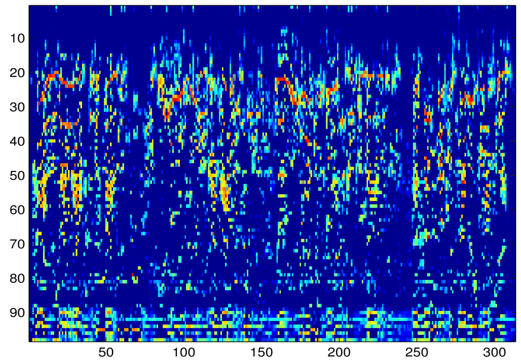
- SiSEC 2010 “Underdetermined speech and music mixtures task” data

Mixing	instantaneous		synth. convolutif				live recorded			
Sources	speech	music	speech		music		speech		music	
Microphone distance	-	-	5 cm	1 m	5 cm	1 m	5 cm	1 m	5 cm	1 m
baseline (l_0 min. or bin. mask.)	8.6	12.4	0.3	1.4	-0.8	-0.9	1.0	1.4	2.3	0.0
NMF / rank-1 [11]	9.6	**18.4**	1.0	2.3	-0.6	-0.6	2.0	2.4	**3.6**	0.3
NMF / full-rank [3]	8.7	17.9	1.2	2.9	-2.3	-0.5	2.2	2.9	3.3	**0.7**
harmonic NMF / rank-1	**10.6**	15.1	1.0	2.7	**-0.1**	**0.0**	2.2	3.4	2.2	0.6
harmonic NMF / full-rank	10.5	14.3	**1.5**	**3.5**	-1.8	-0.2	**2.5**	**3.9**	1.5	0.4

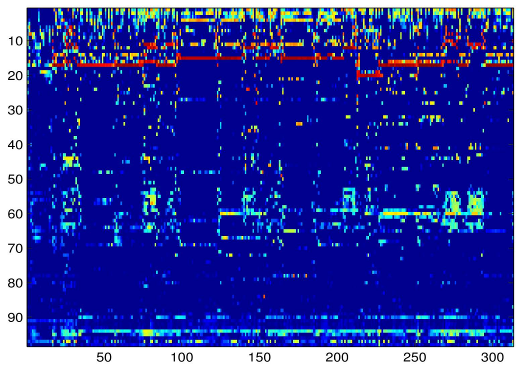
Experimental illustrations

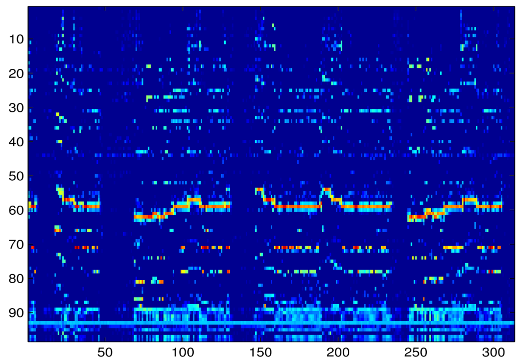
- Speech

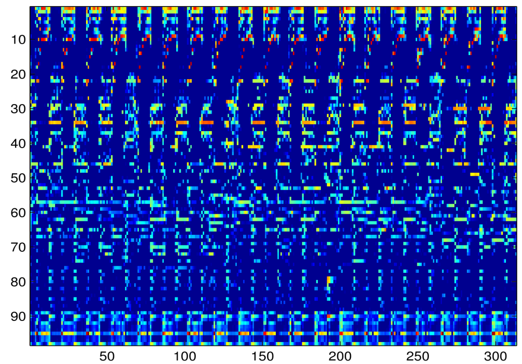




- Music



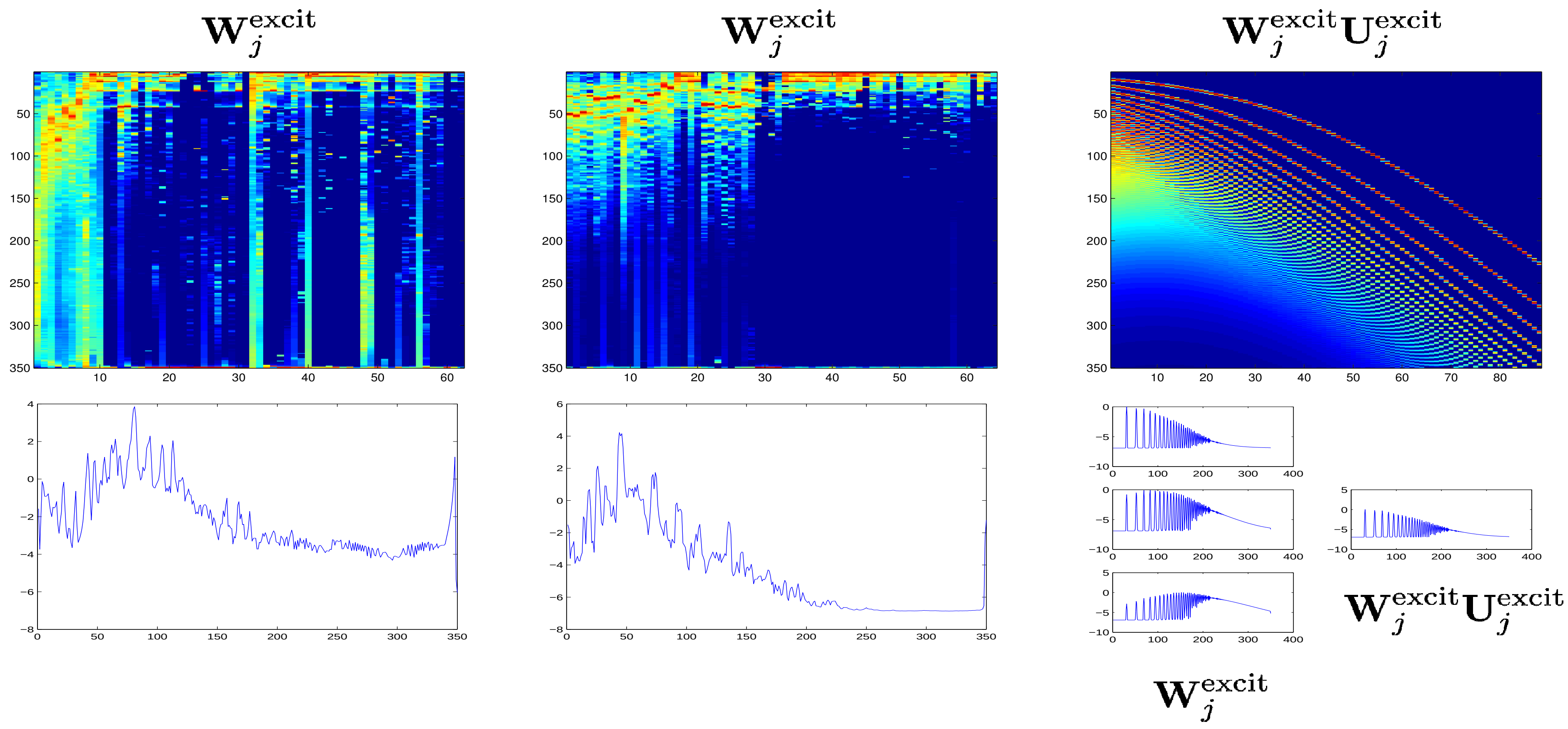




Experimental illustrations

- Drums and bass separation from professionally produced music recordings

Drums (fixed) | Bass (fixed) | Rest (semi-adaptive)



Outline

- Introduction
- Framework presentation
- Experimental illustrations
- Conclusion and further work

Conclusion

- General flexible framework
  - generalizes existing methods and brings them into a common framework
  - allows to imagine and implement new efficient methods for different audio source separation problems (as illustrated experimentally)
- A statistical implementation of CASA
  - primitive and learned grouping cues are used simultaneously (as opposed to sequentially)
  - primitive grouping cues: harmonicity, spectral smoothness, time continuity, common onset, common amplitude modulation, spectral similarity and spatial similarity

Further work

- Apply for separation of 4 components :
  - Melody, drums, bass, rest
- Add new features to the framework
  - Bayesian priors
  - Extension to more than 2 channels case
  - Time varying spectral covariances
- Make the framework implementation publicly available