A General Modular Framework for Audio Source Separation

Alexey Ozerov [1], Emmanuel Vincent [1] and Frédéric Bimbot [2]

[1]INRIA, Rennes Bretagne Atlantique [2]IRISA, CNRS - UMR 6074, Rennes, France

e-mail: {alexey.ozerov,emmanuel.vincent,frederic.bimbot}@irisa.fr

Context

Classical audio source separation methods are usually adapted to a particular scenario:

- problem dimensionality ((over)determined, ...),
- mixing process characteristics (instantaneous, anechoic, ...),
- source characteristics (speech, singing voice, drums, bass, noise, ...)

Limitation:

- No common formulation $\Longrightarrow$ Difficult and time-consuming to adapt a method to a different scenario, it was not originally conceived for.

Goals

Design a new source separation 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.

Approach

Mixing in STFT (j, f and n being source, frequency and time indices):

$$\mathbf{x}_{fn} = \sum\nolimits_{j=1}^{J} \mathbf{y}_{j,fn}, \qquad (1)$$

where $\mathbf{x}_{fn}, \mathbf{y}_{j,fn} \in \mathbb{C}^I$ are the mixture and the source spatial images.

Use local Gaussian model as a basis of our framework:

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_c\left(\bar{0}, v_{j,fn}\mathbf{R}_{j,fn}\right), \qquad (2)$$

- $\mathbf{R}_{j,fn} \in \mathbb{C}^{I\times I}$: is called spatial covariance matrix,
- $v_{j,fn} \in \mathbb{R}_+$: is called spectral power

	Reference	Ozerov-07	Benaroya-06	Blouet-08	Durrieu-10	Abdallah-04	Vincent-10	Bertin-10	Fevotte-09	Fevotte-05	Arberet-09	Ozerov-10	Duong-10	Duong-10a	Arberet-10	Cardoso-08	Pham-03	Attias-03
Problem dimensionality	single channel	×	×	×	×	×	×	×	×									
	underdetermined									×	×	×	×	×	×	(×)		
	(over-)determined															(×)	×	×
Mixing type	linear instantaneous									×	×	(×)				×		
	convolutive											(×)	×	×	×		×	×
Spatial mixing model	rank-1 covariance									×	×	×				×	×	×
	full rank covariance												×	×	×	×		
Source power model	unconstrained												×	×				
	piecewise constant									×						×	×	
	GMM / HMM	×									×							×
	GSMM / S-HMM		×	×	(×)													
	NMF			×		×						×			×			
	harmonic NMF				(×)		×	×										
	temp. constr. NMF							×	×									
	source-filter	×			×													
Signal representation	STFT	×	×	×	×	×			×	×	×	×	×		×	×	×	×
	ERB						×	×						×				

Flexible model

Source separation via Wiener filtering:

$$\hat{\mathbf{y}}_{j,fn} = v_{j,fn}\mathbf{R}_{j,fn}\mathbf{\Sigma}^{-1}_{\mathbf{x},fn}(\theta)\mathbf{x}_{fn}, \qquad \mathbf{\Sigma}_{\mathbf{x},fn}(\theta) = \sum\nolimits_{j=1}^{J} v_{j,fn}\mathbf{R}_{j,fn} \qquad (3)$$

Model estimation using MAP criterion:

$$\theta^*, \eta^* = \arg\min_{\theta\in\Theta,\eta} \sum_{f,n}\left[\operatorname{tr}\left(\mathbf{\Sigma}^{-1}_{\mathbf{x},fn}(\theta)\mathbf{x}_{fn}\mathbf{x}^H_{fn}\right) + \log\left|\mathbf{\Sigma}_{\mathbf{x},fn}(\theta)\right|\right] - \log p(\theta|\eta). \qquad (4)$$

Spatial Covariance Structures

Spatial covariances are time invariant, i.e., $\mathbf{R}_{j,fn} = \mathbf{R}_{j,f}$, and can be:

- linear instantaneous (i.e., $\mathbf{R}_{j,f} = \mathbf{R}_j$) or convolutive,
- rank 1 (i.e., $\mathbf{R}_{j,f} = \mathbf{a}_{j,f}\mathbf{a}^H_{j,f}$) or full rank,
- fixed or adaptive.

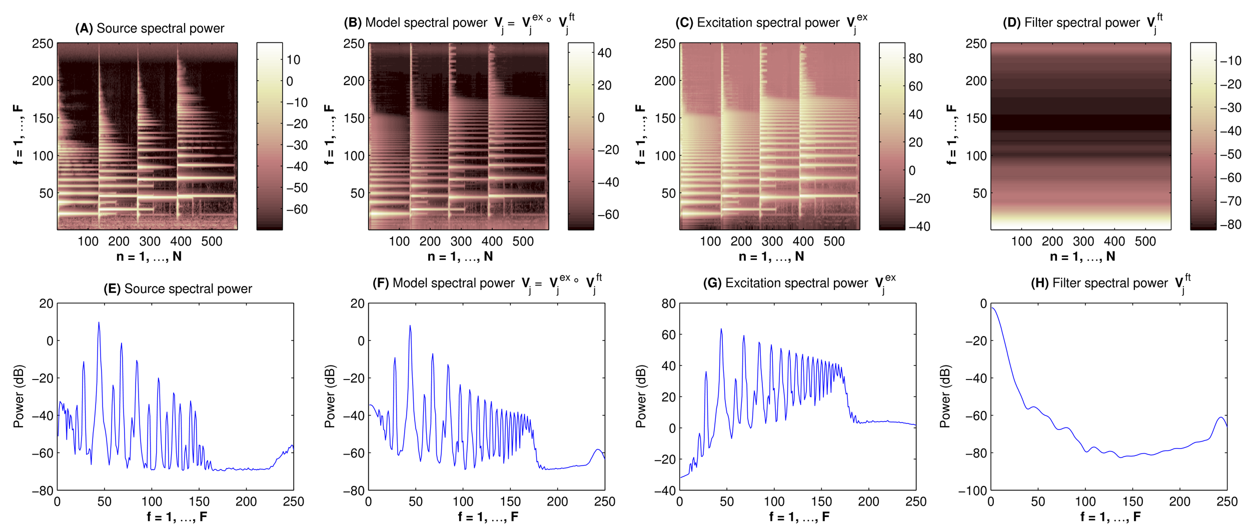
Spectral Power Structures

Excitation / filter NMF-like decomposition:

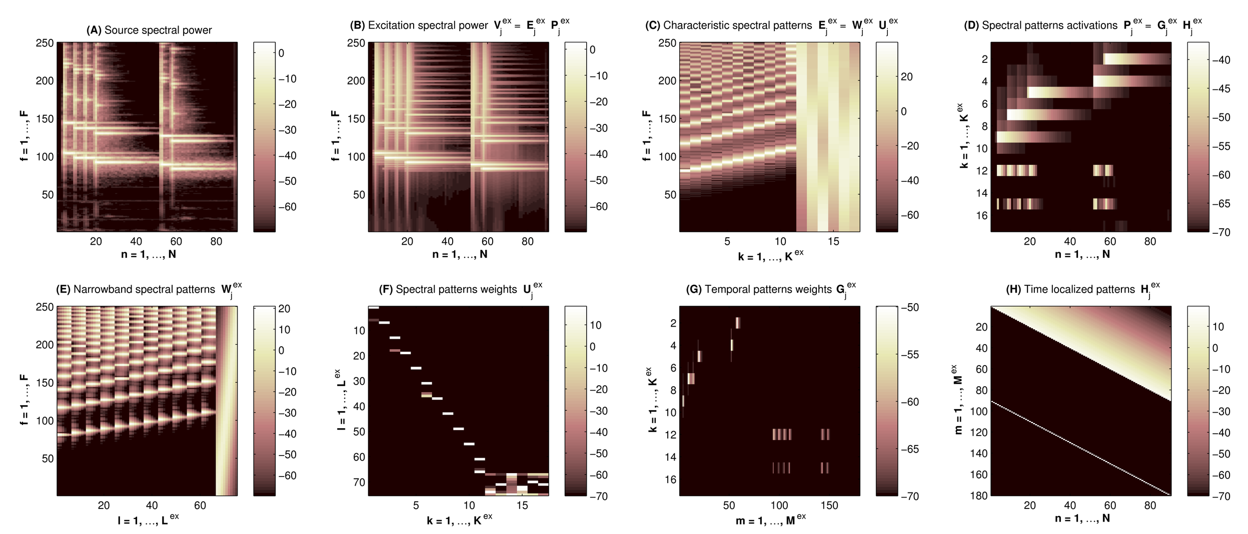
$$\mathbf{V}_j = \mathbf{V}^{\text{excit}}_j \odot \mathbf{V}^{\text{filt}}_j = \left(\mathbf{W}^{\text{excit}}_j\mathbf{U}^{\text{excit}}_j\mathbf{G}^{\text{excit}}_j\mathbf{H}^{\text{excit}}_j\right) \odot \left(\mathbf{W}^{\text{filt}}_j\mathbf{U}^{\text{filt}}_j\mathbf{G}^{\text{filt}}_j\mathbf{H}^{\text{filt}}_j\right) \qquad (5)$$

For example, with these matrices one can model spectral harmonicity, spectral smoothness, time continuity or other structure.

- each matrix can be either fixed or adaptive,
- each source can have additional GMM / HMM-like constraints.



Excitation structure example:



Modular Implementation

Model: $\theta = \{\theta_j\}_{j=1}^J$, where

$$\theta_j = \{\theta^m_j\}_{m=1}^9 = \{\mathbf{R}_j, \mathbf{W}^{\text{excit}}_j, \mathbf{U}^{\text{excit}}_j, \mathbf{G}^{\text{excit}}_j, \mathbf{H}^{\text{excit}}_j, \mathbf{W}^{\text{filt}}_j, \mathbf{U}^{\text{filt}}_j, \mathbf{G}^{\text{filt}}_j, \mathbf{H}^{\text{filt}}_j\}. \qquad (6)$$

Modular implementation is based on a Generalized Expectation-Maximization (GEM) algorithm and multiplicative NMF update rules:

- E-step: compute expectations of natural sufficient statistics.
- M-step: loop over all $J \times 9$ parameter subsets, and for every subset apply a specific update depending on its properties.

Experimental Illustrations

Underdetermined-Speech and Music Mixtures

Mixing	instantaneous		synth. convolutif		live recorded	
Sources	speech	music	speech	music	speech	music
baseline (l_0 min. or bin. mask.)	8.6	12.4	0.9	-0.8	1.2	1.2
NMF / rank-1 [Ozerov-10]	9.6	18.4	1.7	-0.6	2.2	2.0
NMF / full-rank [Arberet-10]	8.7	17.9	2.1	-1.4	2.6	2.0
harmonic NMF / rank-1 [NEW]	10.6	15.1	1.9	0.0	2.8	1.4
harmonic NMF / full-rank [NEW]	10.5	14.3	2.5	-1.9	3.2	1.0

Table 1: Average SDRs on subsets of SiSEC 2010 development data.

Professionally Produced Music Recordings

A fully automatic system (using pre-trained or adaptive spectral and adaptive spatial cues) that allows extracting the following three components from stereo recordings:

- bass, drums, and melody (e.g., singing voice).

Conclusions and Further Work

Proposed general flexible and modular source separation framework:

- generalizes several existing source separation methods,
- brings them into a common framework,
- allows to imagine and implement new efficient methods.

Proposed framework can also be seen as a statistical implementation of Computational Auditory Scene Analysis (CASA) principles.

Further work:

- Incorporate into the framework a possibility to use various prior distributions $p(\theta|\eta)$ (e.g., sparse priors).
- Make the framework implementation publicly available.