Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures With Application to Blind Audio Source Separation

Alexey Ozerov [1] and Cédric Févotte [2]

[1]Institut Télécom, Télécom ParisTech, CNRS LTCI, Paris, France [2]CNRS LTCI, Télécom ParisTech, Paris, France

e-mail: {ozerov,fevotte}@telecom-paristech.fr web: http://perso.telecom-paristech.fr/~ozerov/

Abstract

- Nonnegative Matrix Factorization (NMF) is usually used for single-channel audio signal power spectrogram decomposition.
- We propose a multichannel NMF framework in a general case of convolutive mixtures of sources.
- Possible applications: source separation (blind or supervised) and information retrieval from audio (e.g., music transcription).
- Here we apply multichannel NMF to different stereo audio source separation tasks, and obtain very promising results.

Introduction

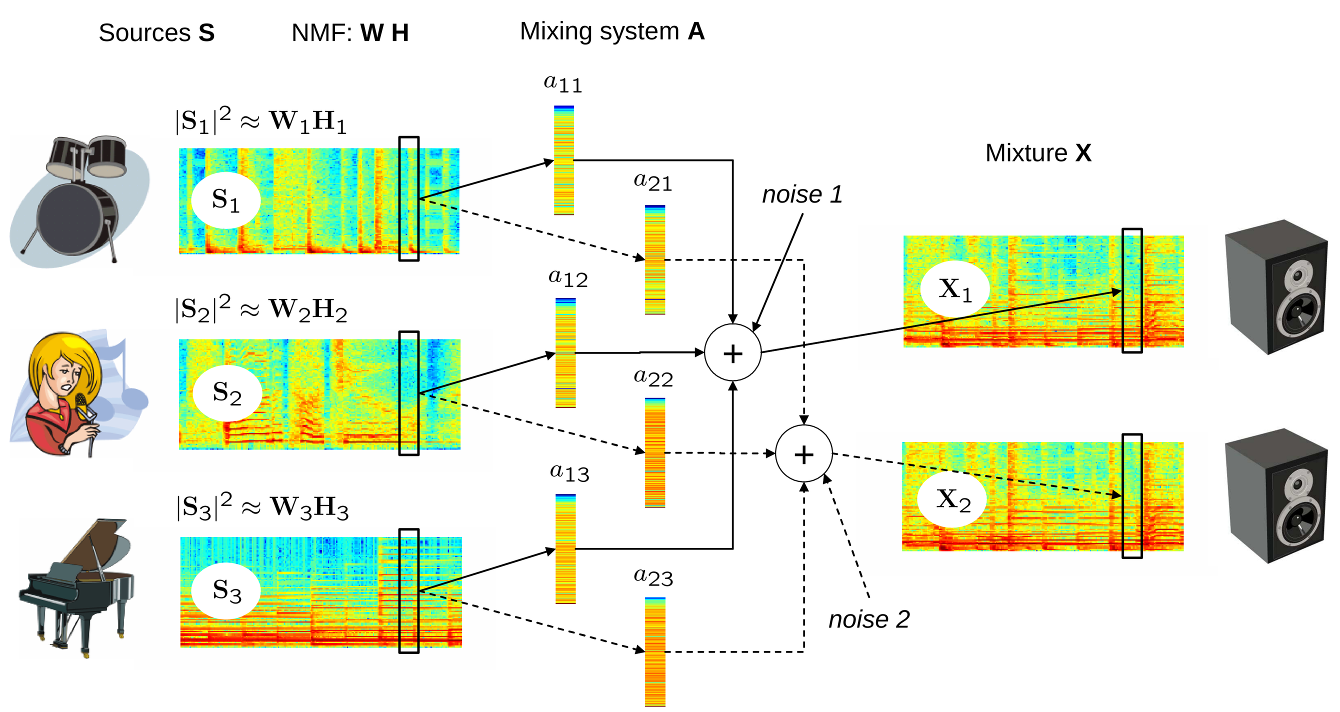
- Single-channel NMF for power spectrogram decomposition:

$$|\mathbf{X}|^2 \approx \mathbf{W}\mathbf{H}, \qquad \mathbf{X} \in \mathbb{C}^{F \times N}, \quad \mathbf{W} \in \mathbb{R}_+^{F \times K}, \quad \mathbf{H} \in \mathbb{R}_+^{K \times N}$$

- Convolutive mixing equation (in STFT domain):

$$\underbrace{x_{i,fn}}_{\text{mixture}} = \sum_{j=1}^{J} \underbrace{a_{ij,f}}_{\text{mix. coef}} \underbrace{s_{j,fn}}_{\text{source}} + \underbrace{b_{i,fn}}_{\text{noise}} \qquad i = 1, \ldots, I$$

- Multichannel NMF problem: estimate $\mathbf{A}$, $\mathbf{W}$ and $\mathbf{H}$, given $\mathbf{X}$.
- Our approach is based on a probabilistic NMF formulation.



Probabilistic NMF Model

- Source STFT is modeled as a sum of latent Gaussian components:

$$s_{j,fn} = \sum_{k=1}^{K_j} c_{j,k,fn} \qquad \text{with} \qquad c_{j,k,fn} \sim \mathcal{N}_c(0, w_{j,fk} h_{j,kn})$$

- Maximum Likelihood (ML) estimation of $\mathbf{W}_j = [w_{j,fk}]_{f,k}$ and $\mathbf{H}_j = [h_{j,kn}]_{k,n}$ given source STFT $\mathbf{S}_j = [s_{j,fn}]_{f,n}$ is equivalent to NMF decomposition $|\mathbf{S}_j|^2 \approx \mathbf{W}_j \mathbf{H}_j$ with Itakura-Saito (IS) divergence [1].

Proposed Multichannel NMF Methods

Exact Likelihood Maximization with EM Algorithm

- Criterion ($\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{W}, \mathbf{H}\}$, $\mathbf{x}_{fn} = [x_{1,fn}, \ldots, x_{I,fn}]^T$):

$$C_1(\boldsymbol{\theta}) = -\log p(\mathbf{X}|\boldsymbol{\theta}) = -\sum_{fn} \log p(\mathbf{x}_{fn}|\boldsymbol{\theta}).$$

- Expectation-Maximization (EM) Algorithm ($\mathbf{C} = [c_{j,k,fn}]_{j,k,f,n}$):

$$\begin{array}{lrcl} \text{E step:} & Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) & = & \int \log p(\mathbf{X}, \mathbf{C}|\boldsymbol{\theta})\, p(\mathbf{C}|\mathbf{X}, \boldsymbol{\theta}^{(l)})\, d\mathbf{C}, \\ \text{M step:} & \boldsymbol{\theta}^{(l+1)} & = & \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}). \end{array}$$

- $\mathbf{A}^{(l+1)} \approx \mathbf{A}^{(l)}$ for small noise, thus a "simulated annealing" strategy is used.
- Related to other model-based methods (e.g., GMM-based [2]).

Individual Likelihoods Maximization with MU rules

- Criterion (IS divergence $d_{IS}(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1$):

$$\begin{aligned} C_2(\boldsymbol{\theta}) &= -\sum_{i=1}^{I} \log p(\mathbf{X}_i|\boldsymbol{\theta}) = -\sum_{i,fn} \log p(x_{i,fn}|\boldsymbol{\theta}) \qquad (\neq \log p(\mathbf{X}|\boldsymbol{\theta})) \\ &= \sum_{i,fn} d_{IS}\left(|x_{i,fn}|^2 \Big| \sum_j |a_{ij,f}|^2 \sum_{k=1}^{K_j} w_{j,fk} h_{j,kn}\right). \end{aligned}$$

- Multiplicative Update (MU) Rules:

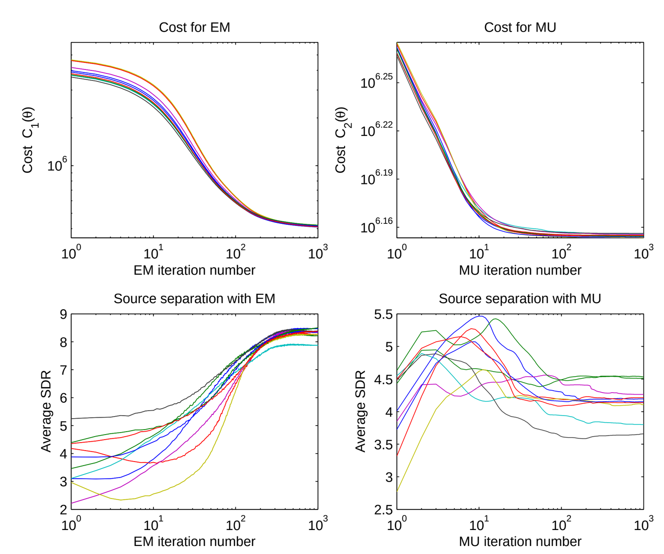
$$\theta_r \leftarrow \theta_r \, [\nabla_{\theta_r} C_2(\boldsymbol{\theta})]_- \, / \, [\nabla_{\theta_r} C_2(\boldsymbol{\theta})]_+,$$

$$\nabla_{\theta_r} C_2(\boldsymbol{\theta}) = [\nabla_{\theta_r} C_2(\boldsymbol{\theta})]_+ - [\nabla_{\theta_r} C_2(\boldsymbol{\theta})]_- \text{ and } [\nabla_{\theta_r} C_2(\boldsymbol{\theta})]_+, [\nabla_{\theta_r} C_2(\boldsymbol{\theta})]_- \geq 0.$$

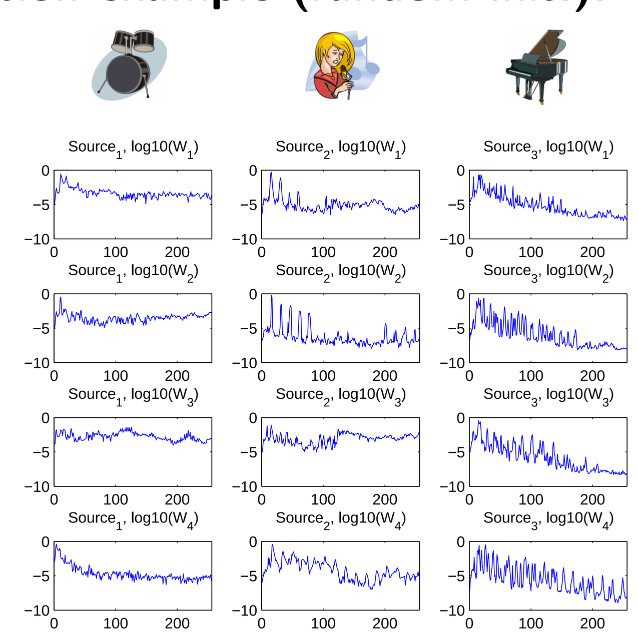
- Related to Nonnegative Tensor Factorization (NTF).

Convergence and Decomposition

- Convergence vs. source separation performance (with perturbed oracle initializations).



- Multichannel NMF decomposition example (random init.).



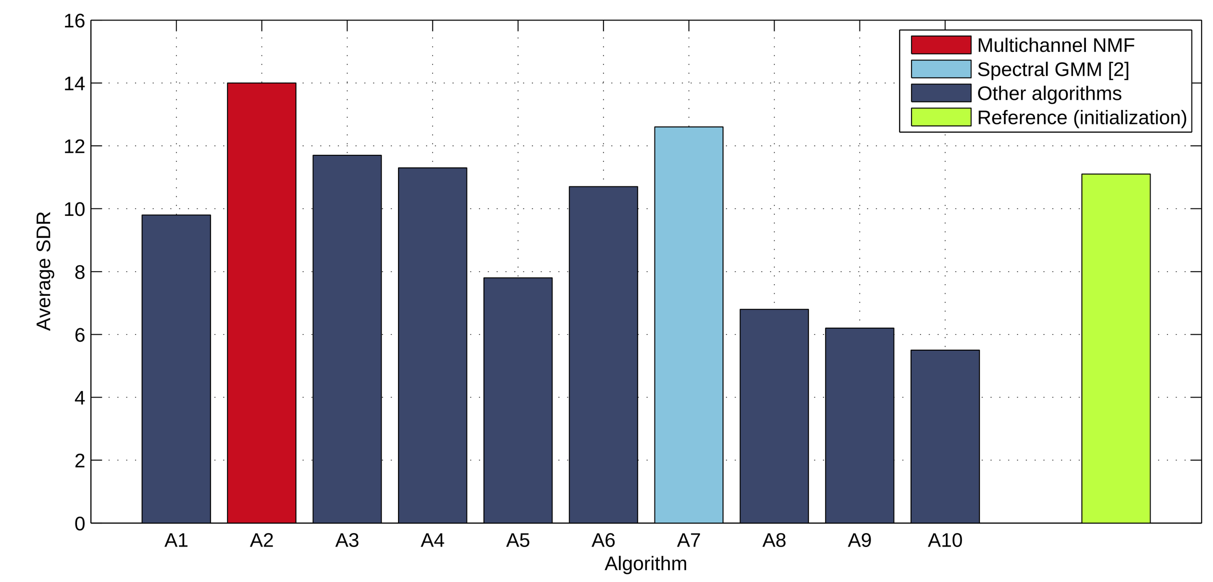
Source Separation Results with EM

- Results for different mixture types (initialization = baseline).

Mixture type	Music sources (3)		Speech sources (4)	
	Baseline	Proposed	Baseline	Proposed
Linear instantaneous	11.6	15.2	7.6	9.3
Synthetic convolutive	-0.8	-0.6	3.5	4.5
Live-recorded (convolutive)	3.3	4.2	3.6	4.3

Table 1: Separation in terms of average Signal to Distortion Ratio (SDR) (dB).

- Signal Separation Evaluation Campaign (SiSEC 2008) / ICA 2009.
  - "Under-determined speech and music (instantaneous) mixtures":



Conclusions

- Strong points of the proposed approach:
  - use of both spectral and spatial diversities for source separation,
  - joint and blind estimation of source and mixing models,
  - covers both underdetermined and (over)determined noisy cases,
  - source model frees us from convolutive BSS permutation ambiguity,
  - computational load growing linearly with number of components.
- Weak point: sensitive to the parameters initialization.

References

[1] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," Neural Computation, vol. 21, pp. 793–830, 2009.

[2] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in ICA'09, 2009.