

# A Variational EM Algorithm for the Separation of Moving Sound Sources

Dionyssos Kounades-Bastian, Laurent Girin,  
Xavier Alameda-Pineda, Sharon Gannot, Radu Horaud



UNIVERSITY  
OF TRENTO



# Source Separation from Convolutional Mixtures

- Problem:  $J$  Source signals, mixed with filters and summed, are recorded at  $I$  microphones: Recover the original sources!
- Existing approaches mainly deal with **static** setups, e.g., [Ozerov & Févotte 2010], [Duong et al. 2010], [Ozerov et al. 2012].
- We want to address **dynamic** setups, for example:
  - moving sources, or
  - moving microphones, or
  - changes in the environment.
- Existing techniques consider either block-wise adaptation of static models, e.g., [Simon & Vincent 2012], or DOA-based discrete temporal models, e.g. [Higuchi et al. 2014].
- We propose a continuous temporal formulation based on linear dynamical systems (LDS)

## Formulation of Static Mixtures

- Separate a mixture of  $J$  sources with  $I$  microphones.
- In STFT domain the problem becomes:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}$$

mixture  $[I \times 1]$   
*observed*

mixing matrix  $[I \times J]$   
*unknown!*

source STFT  $[J \times 1]$   
*unknown!*

sensor noise  $[I \times 1]$   
*unknown!*

- $f = [1, F]$ : frequency bins,  $\ell = [1, L]$ : time frames.

## Proposed Dynamic Mixture Formulation (I)

- The mixture signal at a microphone:

$$x_{i,fl} = \dots + A_{ij,f} s_{j,fl} + \dots$$

- In [Ozerov & Févotte 2010] the entries ( $A_{ij,f}$ ) of  $\mathbf{A}_f$  are parameters
- Our approach:

$\mathbf{A}_f$  replaced with  $\mathbf{A}_{f1}, \dots, \mathbf{A}_{fl}, \dots, \mathbf{A}_{fL}$ .

The mixing becomes:

$$\mathbf{x}_{fl} = \mathbf{A}_{fl} \mathbf{s}_{fl} + \mathbf{b}_{fl}.$$

- The entries of  $A_{fl}$  are modeled as random latent variables.

## Proposed Dynamic Mixture Formulation (II)

- The mixing matrix  $\mathbf{A}_{f\ell}$  is a random variable:
  - Flexibility on the source-microphone path model.
  - Estimate is a distribution instead of a single value.
- The mixing matrix  $\mathbf{A}_{f\ell}$  is complex-Gaussian:
  - Provides compact parametrization.

## Proposed Dynamic Mixture Formulation (III)

- $\mathbf{A}_{f1}, \dots, \mathbf{A}_{f\ell}, \dots, \mathbf{A}_{fL}$  are complex-Gaussian r.v.'s with LDS:
  - $\mathbf{A}_{f1} \sim \mathcal{N}_c(\text{vec}(\mathbf{A}_{f1}); \boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a)$  (1<sup>st</sup> frame prior).
  - $\mathbf{A}_{f\ell} | \mathbf{A}_{f\ell-1} \sim \mathcal{N}_c(\text{vec}(\mathbf{A}_{f\ell}); \text{vec}(\mathbf{A}_{f\ell-1}), \boldsymbol{\Sigma}_f^a)$  (evolution).
- $\text{vec}(\mathbf{A}_{f\ell})$ : vectorization for computational simplicity.
- $\boldsymbol{\Sigma}_f^a \in \mathbb{C}^{IJ \times IJ}$  encodes temporal correlation between filters.
- Limited number of parameters to be estimated,  $IJ$  is small!

# The NMF Source Model

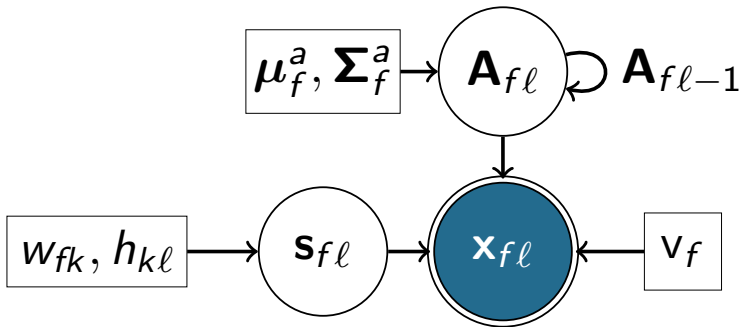
- Same as in [Ozerov & Févotte 2010]:

- Each source: sum of elementary components  $s_{j,fl} = \sum_{k=1}^{K_j} c_{k,fl}$
- Each component follows  $c_{k,fl} \sim \mathcal{N}_c(c_{k,fl}; 0, w_{fk} h_{kl})$ .

- Benefits:

- Reduces the number of parameters to be estimated!
- Provides very simple update rules for both  $w_{fk}$ ,  $h_{kl}$ .
- Avoids permutation of sources between frequencies!

## Associated Graphical Model





## Inference & EM Algorithm

- Probabilistic inference of:

$$\mathcal{A} = \{\mathbf{A}_{fl}\}_{f,\ell=1}^{F,L}, \mathcal{S} = \{\mathbf{s}_{fl}\}_{f,\ell=1}^{F,L} \text{ given } \mathcal{X} = \{\mathbf{x}_{fl}\}_{f,\ell=1}^{F,L}.$$

- Gaussian sensor noise:  $p(\mathcal{X}|\mathcal{A}, \mathcal{S}) = \mathcal{N}_c(\mathbf{x}_{fl}; \mathbf{A}_{fl}\mathbf{s}_{fl}, \mathbf{v}_f \mathbf{I}_l)$ .
- Standard EM alternates between:

- Inference of  $p(\mathcal{A}, \mathcal{S}|\mathcal{X})$ .

- Estimation of  $\theta = \left\{ \mathbf{v}_f, w_{fk}, h_{kl}, \boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a \right\}_{f,\ell,k=1}^{F,L,(\sum_{j=1}^J K_j)}$ .

- Inference of  $p(\mathcal{A}, \mathcal{S}|\mathcal{X})$  is intractable in our case.

# Variational EM

- Variational approximation:  $p(\mathcal{A}, \mathcal{S}|\mathcal{X}) \approx p(\mathcal{A}|\mathcal{X})p(\mathcal{S}|\mathcal{X})$ ,
- E-step split into two steps:
  - Sources E-step: Estimate  $p(\mathcal{S}|\mathcal{X})$  given  $p(\mathcal{A}|\mathcal{X})$
  - Filters E-step: Estimate  $p(\mathcal{A}|\mathcal{X})$  given  $p(\mathcal{S}|\mathcal{X})$ .
- M-step: parameter estimation via maximization of the complete-data expected log-likelihood.

## Expectation Steps

- Sources E-step:

$$p(\mathcal{S}|\mathcal{X}) \propto p(\mathcal{S}) \exp \left( \mathbb{E}_{p(\mathcal{A}|\mathcal{X})} [\log p(\mathcal{X}|\mathcal{A}, \mathcal{S})] \right)$$

This expression results:

$$p(\mathbf{s}_{fl}|\mathcal{X}) = \mathcal{N}_c(\mathbf{s}_{fl}; \hat{\mathbf{s}}_{fl}, \boldsymbol{\Sigma}_{fl}^{\eta^s}).$$

- Filters E-step:

$$p(\mathcal{A}|\mathcal{X}) \propto p(\mathcal{A}) \exp \left( \mathbb{E}_{p(\mathcal{S}|\mathcal{X})} [\log p(\mathcal{X}|\mathcal{A}, \mathcal{S})] \right).$$

This expression, solved with a [Kalman smoother](#), yields:

$$p(\mathbf{A}_{fl}|\mathcal{X}) = \mathcal{N}_c \left( \text{vec}(\mathbf{A}_{fl}); \text{vec}(\hat{\mathbf{A}}_{fl}), \boldsymbol{\Sigma}_{fl}^{\eta^a} \right).$$

## Maximization Step

- The parameter set  $\theta$  estimated by maximizing the **complete data expected log-likelihood**:

$$\mathbb{E}_{p(\mathcal{S}|\mathcal{X})p(\mathcal{A}|\mathcal{X})} [\log p(\mathcal{X}, \mathcal{A}, \mathcal{S})].$$

- Closed-form updates for:  $\{\boldsymbol{\Sigma}_f^a, \boldsymbol{\mu}_f^a, \nu_f\}_{f=1}^F$ .
- Closed-form **alternating** updates for the source-spectra parameters:  $\{w_{fk}, h_{kl}\}_{f,\ell,k=1}^{F,L,(\sum_{j=1}^J K_j)}$ .
- The detailed derivations are in <http://arxiv.org/abs/1510.04595>

# Experimental Setup

- Time-varying convolutive stereo mixtures containing 4 speech signals from TIMIT (length = 2s),
- Source motions simulated using BRIRs [Hummerson et al. 2013].
- Comparison with block-wise implementation of [Ozerov & Févotte 2010]
- Blind initialization of filter parameters ( $\mathbf{A}_{f\ell}$  entries set to 1).
- Initialization of NMF using true source spectra, corrupted by the other sources, with SNR of: 20dB, 10dB, 0dB.
- Performance evaluation using SDR (higher the better) [Vincent et al. 2007].

## Quantitative Results

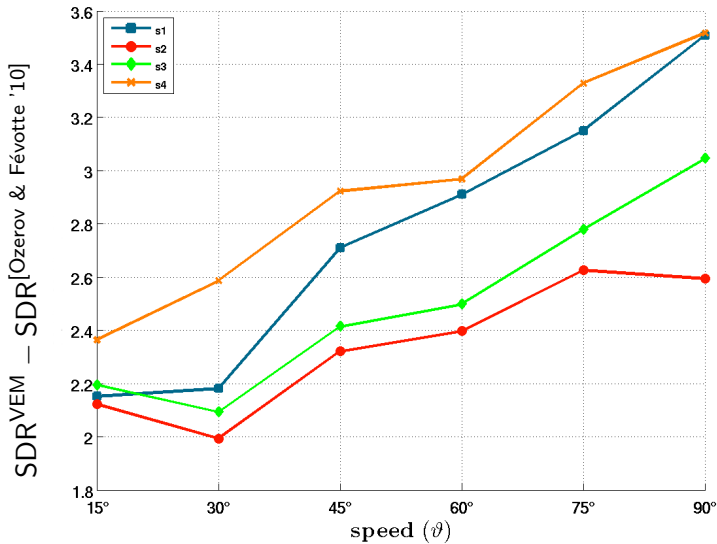
Average SDR (dB) scores (10 sets of speakers):

SNR	Proposed				[Ozerov & Févotte 2010]			
	$s_1$	$s_2$	$s_3$	$s_4$	$s_1$	$s_2$	$s_3$	$s_4$
20dB	7.0	6.6	7.6	9.2	3.8	3.9	4.9	5.8
10dB	6.1	6.0	6.9	8.2	3.7	3.9	4.6	5.4
0 dB	1.8	1.7	3.4	3.8	0.7	1.0	1.7	2.3

SDR measured at the input: The mix-signal is the estimate!

	$s_1$	$s_2$	$s_3$	$s_4$
SDR(dB)	-7.8	-7.6	-5.3	-4.1

## Effect of Circular Speed of Source



## Example of Separation Results

- $J = 4$  sources,  $I = 2$  microphones
- Sources move, forward and backward, along circular trajectories
- Sources #3 and #4 move twice faster than sources #1 and #2



## Conclusions and Future Work

- We addressed separation of moving acoustic sources;
- We proposed a generalization of the successful time-invariant convolutive model of [Ozerov & Févotte 2010];
- We devised a variational EM (VEM) inference procedure;
- Results obtained with 4 sources and 2 microphones (underdetermined mixtures) are quite encouraging;
- VEM is well known to be sensitive to initialization and less efficient than EM;
- We plan to thoroughly investigate initialization strategies and to improve the algorithm's speed of convergence;
- We also plan to combine diarization and separation.

Thank you !