A Variational EM Algorithm for the Separation of Moving Sound Sources

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

Source Separation from Convolutive Mixtures

- Problem: J source signals are filtered and summed at I microphones $\rightarrow$ We want to recover the source signals!
- 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:
  - moving sources
  - moving microphones
  - 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 the STFT domain, the mixture is approximated by:

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

- mixture $[I \times 1]$ *observed*
- mixing matrix $[I \times J]$ *unknown!*
- source $[J \times 1]$ *unknown!*
- sensor noise $[I \times 1]$ *unknown!*

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

Proposed Dynamic Mixture Formulation (I)

- We start from the probabilistic framework of Local Composite Gaussian Model of sources, plugged in the (static) convolutive mixture model [Ozerov & Févotte 2010]: adapted to underdetermined mixtures ($I < J$), EM-based estimation, the entries of $\mathbf{A}_f$ are parameters.
- Our approach: Dynamic mixing filters: $\mathbf{A}_f$ replaced with $\mathbf{A}_{f1}, \dots, \mathbf{A}_{f\ell}, \dots, \mathbf{A}_{fL}$. The mixing becomes:

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

$\mathbf{A}_{f\ell}$ is modeled as a random latent variable.
→ Provides compact parametrization.
→ Flexibility on the source-microphone path model.
→ Estimate is a distribution instead of a single value.

Proposed Dynamic Mixture Formulation (II)

- $\mathbf{A}_{f1}, \ldots, \mathbf{A}_{f\ell}, \ldots, \mathbf{A}_{fL}$ are modeled as complex-Gaussian with first-order temporal model:

$$\mathbf{A}_{f1} \sim \mathcal{N}_c\left(\text{vec}(\mathbf{A}_{f1}); \boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a\right) \text{ (1}^{\text{st}}\text{ frame prior)}$$

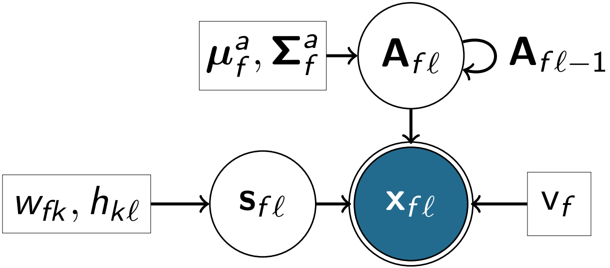
$$\mathbf{A}_{f\ell}|\mathbf{A}_{f\ell-1} \sim \mathcal{N}_c\left(\text{vec}(\mathbf{A}_{f\ell}); \text{vec}(\mathbf{A}_{f\ell-1}), \boldsymbol{\Sigma}_f^a\right) \text{ (evolution).}$$

- $\text{vec}(\mathbf{A}_{f\ell})$: vectorization for computational simplicity.
- $\boldsymbol{\Sigma}_f^a \in \mathbb{C}^{IJ \times IJ}$ encodes temporal correlation between successive filters.
- Limited number of parameters to be estimated, IJ is small!

The NMF Source Model

- Same as in [Ozerov & Févotte 2010]:
  - Each source is a sum of elementary components:
    $$s_{j,f\ell} = \sum_{k \in \mathcal{K}_j} c_{k,f\ell}$$
  - Component vector is assumed complex-Gaussian:
    $$p(\mathbf{c}_{f\ell}) = \mathcal{N}_c\Big(\mathbf{c}_{f\ell}; \mathbf{0}, \mathrm{diag}_K\left(w_{fk} h_{k\ell}\right)\Big)$$
  - Hence, source vector is complex-Gaussian:
    $$p(\mathbf{s}_{f\ell}) = \mathcal{N}_c\left(\mathbf{s}_{f\ell}; \mathbf{0}, \mathrm{diag}_J\left(\sum_{k \in \mathcal{K}_j} w_{fk} h_{k\ell}\right)\right).$$
- Benefits:
  - Reduces the number of source parameters to be estimated.
  - Provides very simple update rules for both $w_{fk}, h_{k\ell}$.
  - Avoids permutation of sources between frequencies.

Associated Graphical Model



Inference & EM Algorithm

- Probabilistic inference of:

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

- We have $p(\mathcal{A})$ and $p(\mathcal{S})$
- Observation density: $p(\mathcal{X}|\mathcal{A},\mathcal{S}) = \prod_{f,\ell}^{F,L} \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_{f\ell}\mathbf{s}_{f\ell}, \mathrm{v}_f \mathbf{I}_I)$.

- Standard EM would alternate between:
  - Inference of $p(\mathcal{A},\mathcal{S}|\mathcal{X})$.
  - Estimation of $\theta = \left\{\mathrm{v}_f, w_{fk}, h_{k\ell}, \boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a\right\}_{f,\ell,k}$.

- 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 $p(\mathcal{A}|\mathcal{X})$

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

  - Filters E-step: Estimate $p(\mathcal{A}|\mathcal{X})$ given $p(\mathcal{S}|\mathcal{X})$

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

- M-step: parameter estimation via maximization of the complete-data expected log-likelihood.

Expectation Steps

- $p(\mathcal{X}, \mathcal{A}, \mathcal{S}) = p(\mathcal{X}|\mathcal{A}, \mathcal{S})p(\mathcal{A})p(\mathcal{S})$
- Sources E-step: $p(\mathcal{S}|\mathcal{X}) \propto p(\mathcal{S}) \exp\left(\mathbb{E}_{p(\mathcal{A}|\mathcal{X})}\left[\log p(\mathcal{X}|\mathcal{A}, \mathcal{S})\right]\right)$
  This expression yields:
  $$p(\mathbf{s}_{f\ell}|\mathcal{X}) = \mathcal{N}_c(\mathbf{s}_{f\ell}; \hat{\mathbf{s}}_{f\ell}, \mathbf{\Sigma}^{\eta s}_{f\ell}),$$
  with $\hat{\mathbf{s}}_{f\ell}, \mathbf{\Sigma}^{\eta s}_{f\ell}$ having closed-form expressions involving mixing filters posterior moments and observations (Wiener filtering).
- Filters E-step: $p(\mathcal{A}|\mathcal{X}) \propto p(\mathcal{A}) \exp\left(\mathbb{E}_{p(\mathcal{S}|\mathcal{X})}\left[\log p(\mathcal{X}|\mathcal{A}, \mathcal{S})\right]\right)$
  This expression yields:
  $$p(\mathbf{A}_{f1:L}|\mathcal{X}) \propto p(\mathbf{A}_{f1:L}) \prod_{\ell=1}^{L} \mathcal{N}_c(\boldsymbol{\mu}^{\iota a}_{f\ell}; vec(\mathbf{A}_{f\ell}), \mathbf{\Sigma}^{\iota a}_{f\ell}),$$
  with $\boldsymbol{\mu}^{\iota a}_{f\ell}, \mathbf{\Sigma}^{\iota a}_{f\ell}$ having closed-form expressions involving sources posterior moments and observations. This is an LDS, solved with a Kalman smoother:
  $$p(\mathbf{A}_{f\ell}|\mathcal{X}) = \mathcal{N}_c\left(\mathrm{vec}(\mathbf{A}_{f\ell}); \mathrm{vec}(\hat{\mathbf{A}}_{f\ell}), \mathbf{\Sigma}^{\eta a}_{f\ell}\right).$$

Maximization Step

- The parameter set θ estimated by maximizing the complete data expected log-likelihood:

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

- Closed-form updates for: $\{\boldsymbol{\Sigma}_f^a, \boldsymbol{\mu}_f^a, \mathrm{v}_f\}_f$.
- Closed-from alternating updates for the source NMF parameters: $\{w_{fk}, h_{k\ell}\}_{f,\ell,k}$.
- 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 [Hummersone 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 power spectra of true source corrupted by the other sources, with SNR of: 20dB, 10dB, 0dB.
- Performance evaluation using SDR [Vincent et al. 2007].

Quantitative Results

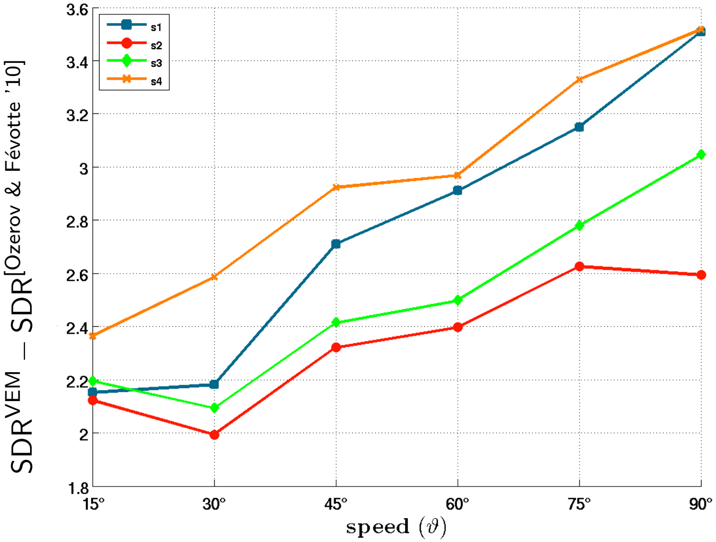
Average SDR (dB) scores (10 sets of speakers):

	Proposed				[Ozerov & Févotte 2010]			
SNR	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

Input SDR (dB)

s_1	s_2	s_3	s_4
-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 !