An acoustically-motivated spatial prior for under-determined reverberant source separation

N. Q. K. Duong, E. Vincent, and R. Gribonval

METISS project team,

INRIA, Centre de Rennes - Bretagne Atlantique, France

May. 2011.

Content

- Problem introduction
- General Gaussian modeling framework
- Acoustically-motivated spatial prior
- MAP parameter estimation
- Experimental result and conclusion

Under-determined source separation

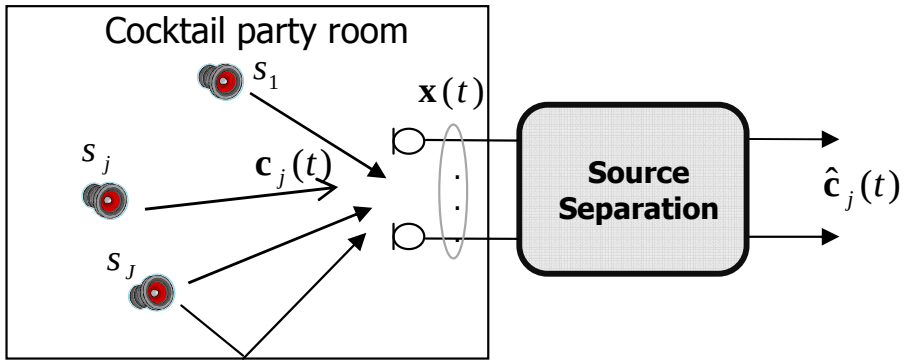
- Use I-channel mixture signal $\mathbf{x}(t) = [x_1(t), \ldots, x_I(t)]^T$ to separate J sources $s_j(t)$, where $I < J$

- Convolutive mixing model:

  Denoting by $\mathbf{c}_j(t)$ the contribution of $s_j(t)$ to all microphones, $\mathbf{x}(t)$ includes the contribution of several sources:

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t)$$

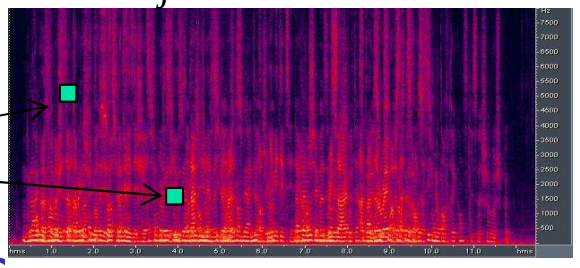
$$\mathbf{c}_j(t) = \sum_{\tau} \mathbf{h}_j(\tau) s_j(t-\tau)$$



where $\mathbf{h}_j(t) = [h_{1j}(t), \ldots, h_{Ij}(t)]^T$ the vector of mixing filters from source j to microphone array.

Baseline approaches

$$\mathbf{x}(t) = \sum_j \overbrace{\mathbf{h}_j * s_j(t)}^{\mathbf{c}_j(t)} \quad \xrightarrow{\text{STFT with narrowband approximation}} \quad \mathbf{x}(n,f) \approx \sum_j \overbrace{\mathbf{h}_j(f)\, s_j(n,f)}^{\mathbf{c}_j(n,f)}$$



Sparsity assumption: only FEW sources are active at each *time-frequency point*

The popular DUET algorithm includes 2 steps

(1) Estimate the mixing parameters

(2) Recover source coefficient by *binary masking* (only ONE source is considered to be active at each time-frequency point)

→ These techniques remain limited in the realistic reverberant environments since the narrowband approximation does not hold

Considered framework

Models the STFT coefficients of the source images as zero-mean Gaussian random variables, i.e.

$$\mathbf{c}_j(n,f) \sim N_c\left(\mathbf{0}, \boldsymbol{\Sigma}_j(n,f)\right)$$

$$\boldsymbol{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(f)$$

$I \times I$ *spatial covariance matrices* encoding spatial position and spatial spread of sources

Scalar *source variances* encoding spectro-temporal power of sources

Spatial covariance parameterizations:

- Rank-1 matrices (resulting from the narrowband assumption)

$$\mathbf{R}_j(f) = \mathbf{h}_j(f)\mathbf{h}_j^H(f)$$

- Full-rank matrices: the coefficients of $\mathbf{R}_j(f)$ are not deterministically related.

Source separation architecture

1. Time-frequency (T-F) transform

our focus ⟹ 2. Estimation of the model parameters $\boldsymbol{\theta} = \left\{ v_j(n,f), \mathbf{R}_j(f) \right\}_{j,n,f}$

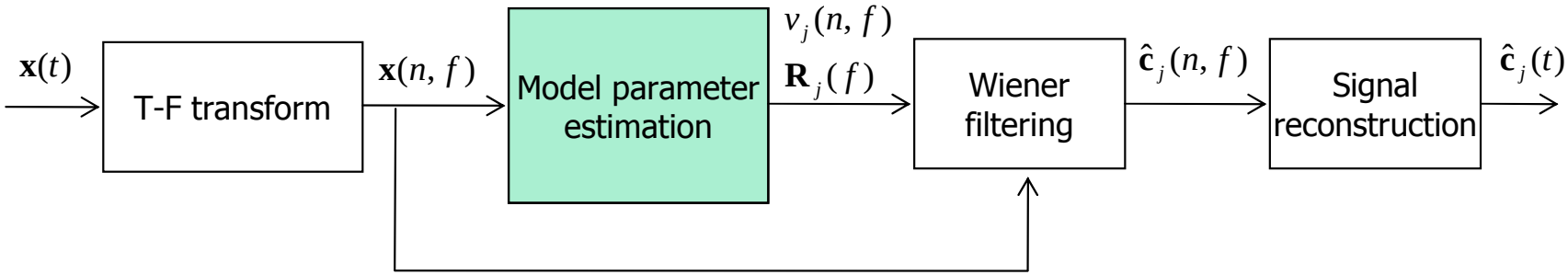
Likelihood $p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{n,f} \frac{1}{|\pi \boldsymbol{\Sigma}_{\mathbf{x}}(n,f)|} e^{-\mathbf{x}^H(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f)}$

where $\boldsymbol{\Sigma}_{\mathbf{x}}(n,f) = \sum_{j=1}^{J} v_j(n,f)\mathbf{R}_j(f)$

3. Source separation by multichannel Wiener filtering

$$\hat{\mathbf{c}}_j(n,f) = \left(v_j(n,f)\mathbf{R}_j(f) \right) \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f)\mathbf{x}(n,f)$$

4. Time-domain signal reconstruction



Acoustically-motivated spatial prior

- Motivation

  ✓ *Spatial prior: exploit knowledge about the source positions and the room characteristics to enhance the source separation performance*

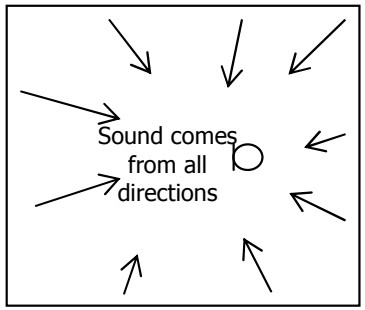
Known geometric setting

- Knowledge from the theory of statistical room acoustics

$$\mathrm{E}\left\{\mathbf{R}_j(f)\right\} = \underbrace{\mathbf{h}_j^{\mathrm{ane}}(f)\left(\mathbf{h}_j^{\mathrm{ane}}\right)^H(f)}_{\text{covariance of the direct part}} + \underbrace{\sigma_{\mathrm{rev}}^2\boldsymbol{\Omega}(f)}_{\text{covariance of the reverberant part}}$$

Underlying assumptions ⇐

The direct part and the reverberant part are uncorrelated and the reverberant part is diffuse



where $\mathbf{h}_j^{\mathrm{ane}}(f)$, $\boldsymbol{\Omega}(f)$, and σ_{rev}^2 can be computed directly given the geometric setting.

Acoustically-motivated spatial prior

- Inverse Wishart prior over the spatial covariance matrices

$$p\left(\mathbf{R}_j(f)\right) = IW\left(\mathbf{R}_j(f) \mid \mathbf{\Psi}_j(f), m\right)$$

where

$$IW\left(\mathbf{R} \mid \mathbf{\Psi}, m\right) = \frac{\left|\mathbf{\Psi}\right|^{m} \left|\mathbf{R}\right|^{-(m+I)} e^{-\mathrm{trace}\left(\mathbf{\Psi}\mathbf{R}^{-1}\right)}}{\pi^{I(I-1)/2} \prod_{i=1}^{I} \Gamma(m-i+1)}$$

$\leftarrow$ conjugate prior to the considered likelihood

The mean of $\mathbf{R}_j(f)$ is given by

$$\mathrm{E}\left\{\mathbf{R}_j(f)\right\} = \frac{\mathbf{\Psi}_j(f)}{m-I} = \mathbf{h}_j^{\mathrm{ane}}(f)\left(\mathbf{h}_j^{\mathrm{ane}}\right)^{H}(f) + \sigma_{\mathrm{rev}}^2 \mathbf{\Omega}(f)$$

the variance controlled by m learned from training data in the maximum likelihood (ML) sense.

MAP parameter estimation by EM

- Estimation of $\boldsymbol{\theta} = \left\{ v_j(n,f), \mathbf{R}_j(f) \right\}_{j,n,f}$

- E-step: estimation of the empirical covariance of each source

$$\mathbf{W}_j(n,f) = \boldsymbol{\Sigma}_j(n,f)\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(n,f), \text{ where } \boldsymbol{\Sigma}_j(n,f) = v_j(n,f)\mathbf{R}_j(n,f)$$

$$\hat{\boldsymbol{\Sigma}}_j(n,f) = \mathbf{W}_j(n,f)\hat{\boldsymbol{\Sigma}}_{\mathbf{x}}(n,f)\mathbf{W}_j^H(n,f) + \left(\mathbf{I} - \mathbf{W}_j(n,f)\right)\boldsymbol{\Sigma}_j(n,f)$$

- M-step: MAP parameter update for each source

$$v_j(n,f) = \frac{1}{I}\mathrm{tr}\left(\mathbf{R}_j^{-1}(f)\hat{\boldsymbol{\Sigma}}_j(n,f)\right)$$

$$\mathbf{R}_j(f) = \frac{1}{\gamma(m+I)+N}\left(\gamma\boldsymbol{\Psi}_j(f) + \sum_{n=1}^{N}\frac{\hat{\boldsymbol{\Sigma}}_j(n,f)}{v_j(n,f)}\right)$$

Trade-off parameter determining the contribution of the prior; $\gamma = 0$ for ML parameter estimation

Experiment

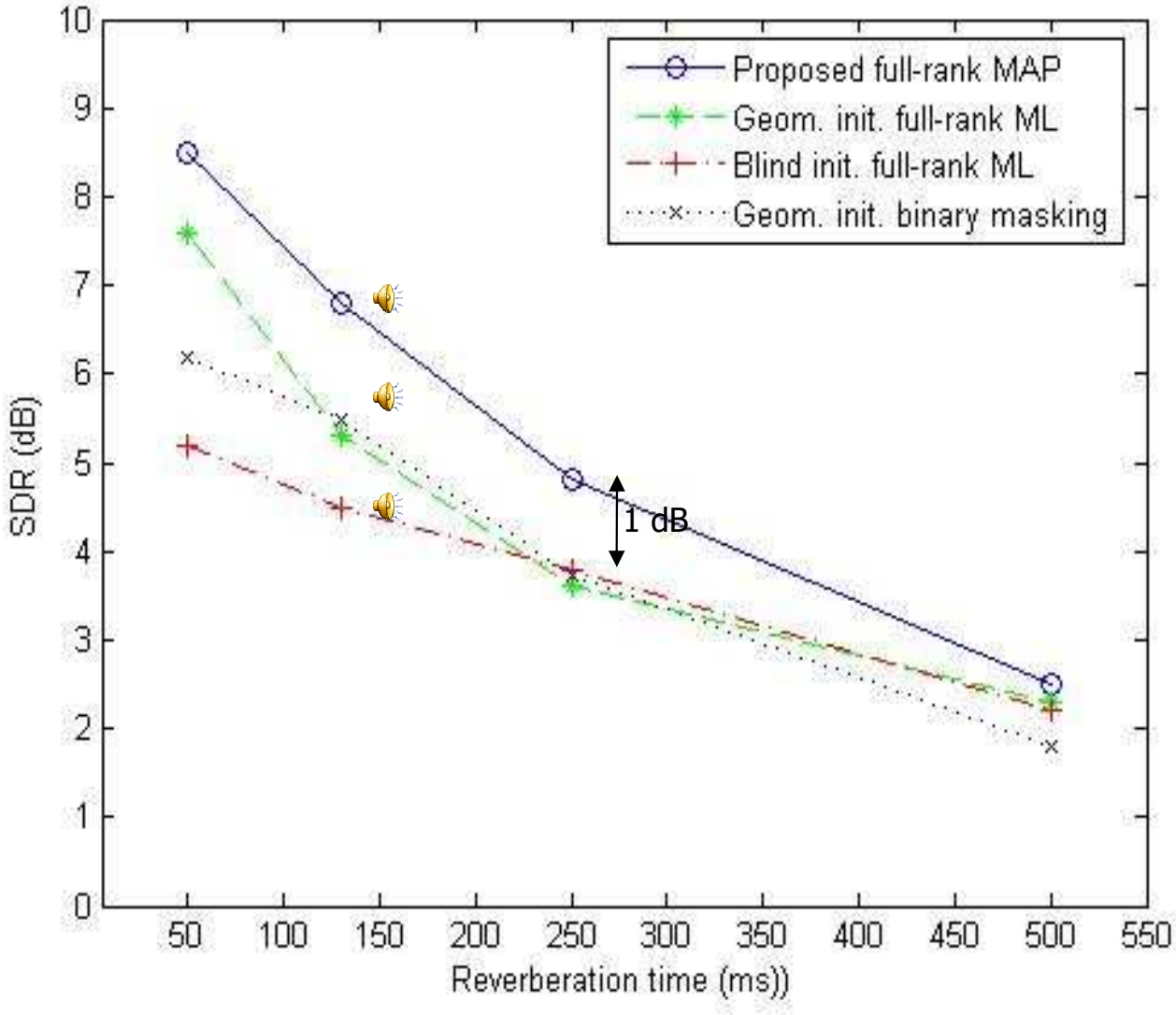
Purpose: compare the source separation performance of the proposed MAP based algorithm with

- ML based algorithm where $\mathbf{R}_j(f)$ are blindly initialized
- ML based algorithm where $\mathbf{R}_j(f)$ are initialized from geometric setting
- Baseline binary masking where the mixing vectors are fixed to the first eigenvector of $\mathbf{\Psi}_j(f)$

Parameter settings:

Speech length	20 s
Sampling rate	16 kHz
STFT window type	Sine
Window length	1024
Number of EM iterations	10
Trade-off parameter	$\gamma = 50$

Experimental result



Average separation performance over 3 stereo mixtures of 4 sources with DoA=[20°, 80°, 120°, 150°], mic spacing=5cm, and source-to-mic distance =50cm.

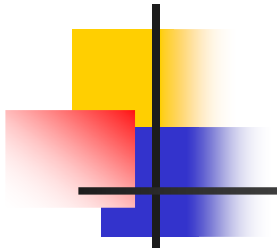
Conclusion & future work

Contributions:

- Propose an *acoustically-motivated spatial prior* from the theory of statistical room acoustics
- Derive a MAP based algorithm which offers a principled solution to the estimation of model parameters and to the permutation problem
- We showed that the proposed algorithm outperforms state-of-the-art approaches.

Future work

- Fully blind source separation by estimating all acoustical parameters (e.g. reverberation time, σ^2_{rev}, etc).



Thanks for your attention!

& Your comments…?

Learned value of m

T_{60}	50 ms	130 ms	250 ms	500 ms
m	2.1	2.8	4.2	6.4
σ^2_{rev}	0.011	0.057	0.131	0.287