Multichannel Audio Source Separation:
Variational Inference of Time-Frequency Sources from Time-Domain Observations

Simon Leglaive, Roland Badeau, Gaël Richard

LTCI, Télécom ParisTech, Université Paris Saclay

IEEE International Conference on Acoustics Speech and Signal Processing

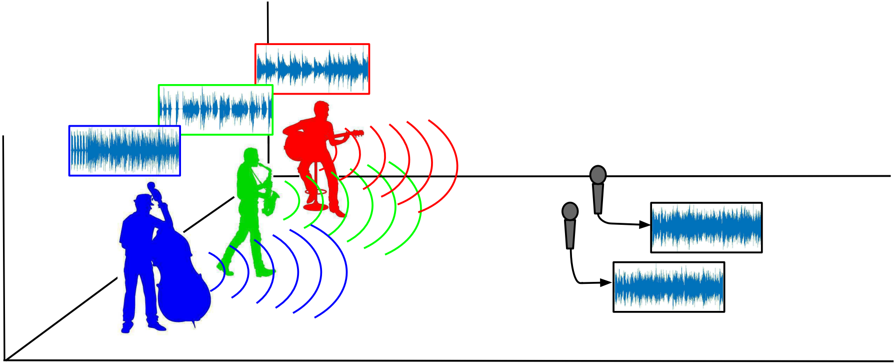
New-Orleans, LA, USA

March 6, 2017

Multichannel audio source separation

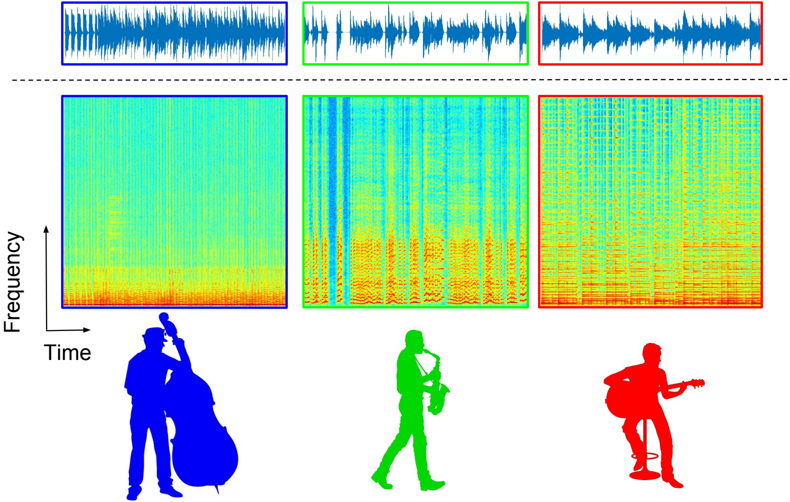
Objective: Recover source signals from the observation of several mixtures.

Context: Under-determined and reverberant.



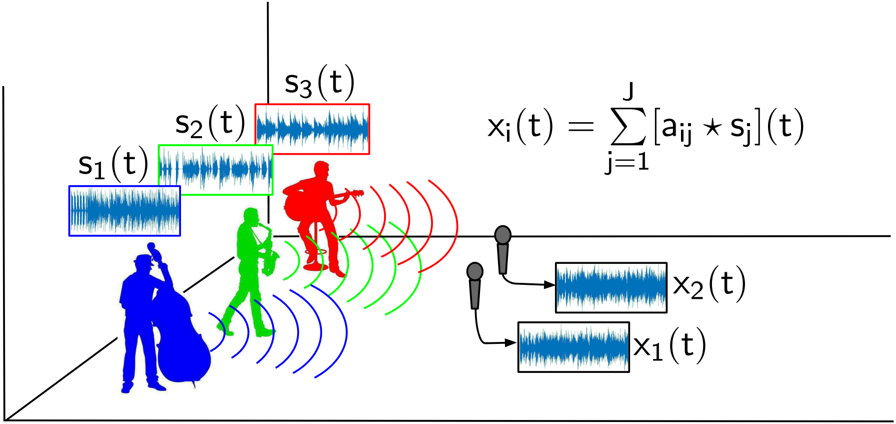
Time-frequency source representation

Time-frequency (TF) transforms provide meaningful representations.



Modeling reverberant mixtures (1)

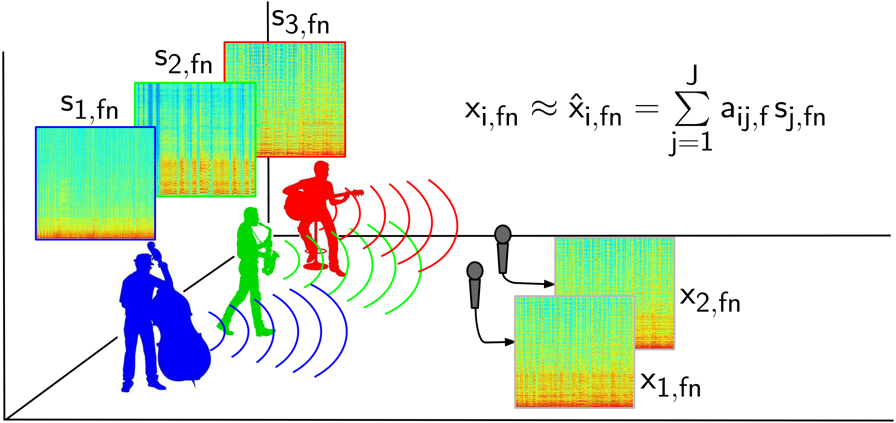
Convolutive model in the time domain:



$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$$

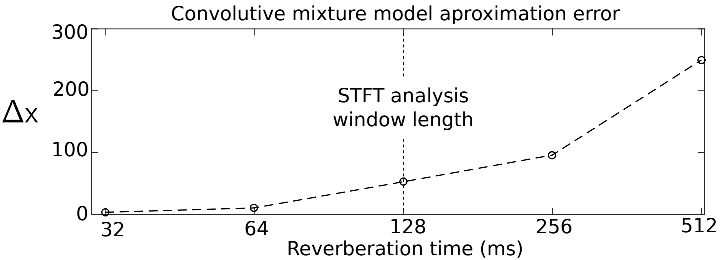
Modeling reverberant mixtures (2)

Convolutive model in the Short-Term Fourier Transform (STFT) domain:



$$x_{i,fn} \approx \hat{x}_{i,fn} = \sum_{j=1}^{J} a_{ij,f}\, s_{j,fn}$$

Error due to the STFT approximation

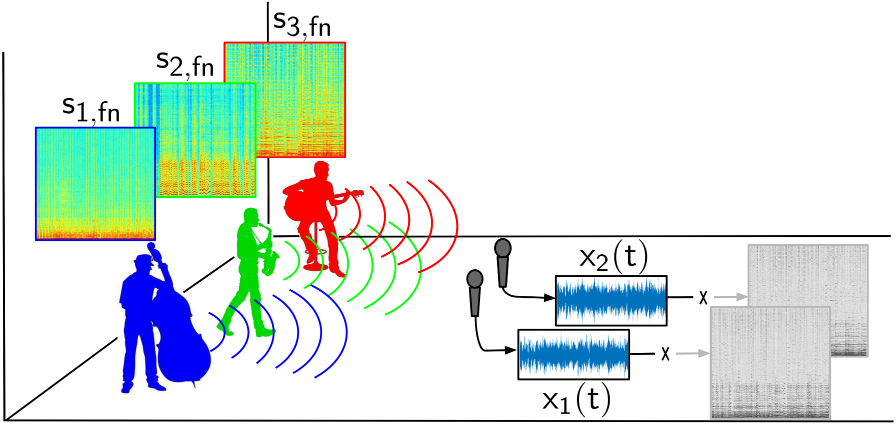


Average relative squared error:

$$\Delta x = \frac{1}{IFN} \sum_{i,f,n} \frac{|\ x_{i,fn} - \hat{x}_{i,fn}\ |^2}{|\ x_{i,fn}\ |^2} \text{ with } x_{i,fn} = \mathrm{STFT}[x_i(t)]$$

Proposed approach

TF source model and time-domain convolutive mixture model.

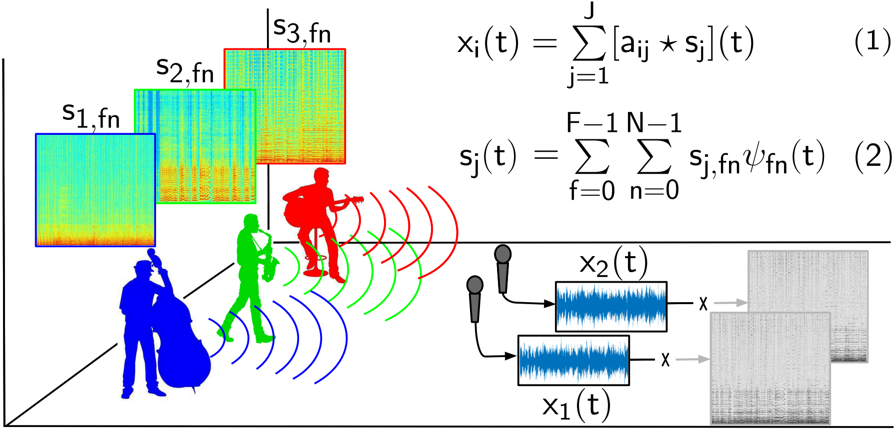


Proposed approach

TF source model and time-domain convolutive mixture model.

$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t) \qquad (1)$$

$$s_j(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \psi_{fn}(t) \qquad (2)$$



$\psi_{fn}(t)$ is a Modified Discrete Cosine Transform (MDCT) atom.

Outline

Probabilistic model

Inference

Experiments

Future work

Probabilistic modeling with latent variables

- Latent source random variables: $\mathbf{s} = \{s_{j,fn} \in \mathbb{R}\}_{j,f,n}$;
- Observed random variables: $\mathbf{x} = \{x_i(t) \in \mathbb{R}\}_{i,t}$.

Defining the probabilistic model

$$p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) = \underbrace{p(\mathbf{s}; \boldsymbol{\theta})}_{\text{prior distribution of } \mathbf{s}} \times \overbrace{p(\mathbf{x}|\mathbf{s}; \boldsymbol{\theta})}^{\text{conditional distribution of } \mathbf{x} \text{ given } \mathbf{s}}$$

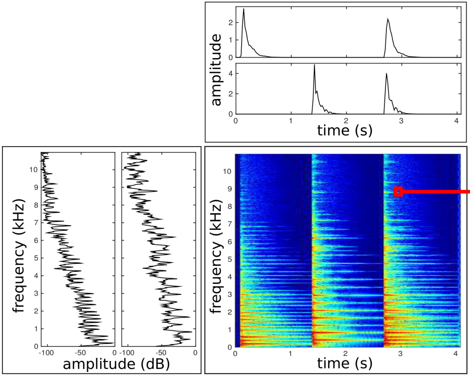
where $\boldsymbol{\theta}$ is a set of deterministic parameters.

- What prior knowledge do we have on the latent source variables?
- How are the data generated from the latent unobserved variables?

Prior distribution of the latent variables

Gaussian source model based on Non-negative Matrix Factorization [1].

$$\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$$



$$\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$$

MDCT power spectrogram (dB)

temporal activations

spectral templates

$$\boxed{s_{j,fn} \sim \mathcal{N}(0, [\mathbf{W}_j \mathbf{H}_j]_{fn})}$$

All the sources and TF points are further assumed to be independent.

[1] C. Févotte, N. Bertin, J.-L. Durrieu. "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis". *Neural computation*, 2009.

Conditional distribution of x given s

Gaussian modeling error

$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t) + b_i(t),$$

with $b_i(t) \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_i^2)$ and $s_j(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \psi_{fn}(t)$.

Conditional distribution

$$x_i(t) | \mathbf{s}; \boldsymbol{\theta} \sim \mathcal{N}\left(\sum_{j=1}^{J} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} [a_{ij} \star \psi_{fn}](t), \sigma_i^2 \right)$$

Outline

Statistical inference

We are interested in the posterior distribution $p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})$,

with $\boldsymbol{\theta} = \left\{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i\right\}$.

Source and parameter estimation

- Source estimation according to the posterior mean:

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star}[\mathbf{s}].$$

- Maximum likelihood estimation of the parameters:

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}).$$

The posterior distribution is Gaussian but with a high-dimensional full covariance matrix $\rightarrow$ **Variational inference**.

Variational inference

- We want to find $q \in \mathcal{F}$ which approximates $p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})$.
- Taking the KL divergence as a measure of fit, we can show that:

$$KL(q||p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{Log-likelihood}} - \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{Variational Free Energy}}, \tag{3}$$

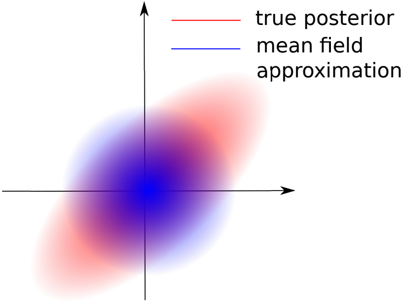
where $\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})}{q(\mathbf{s})}\right)\right\rangle_q$ and $\langle f(\mathbf{s})\rangle_q = \int f(\mathbf{s})q(\mathbf{s})d\mathbf{s}$.

- Variational Expectation-Maximization algorithm:
  - **E-step**: $q^\star = \underset{q\in\mathcal{F}}{\arg\min}\, KL(q||p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}_{\text{old}})) = \underset{q\in\mathcal{F}}{\arg\max}\, \mathcal{L}(q;\boldsymbol{\theta}_{\text{old}})$;
  - **M-step**: $\boldsymbol{\theta}_{\text{new}} = \underset{\boldsymbol{\theta}}{\arg\max}\, \mathcal{L}(q^\star;\boldsymbol{\theta})$.

Mean-field approximation

$\mathcal{F}$: set of pdfs over $\mathbf{s}$ that factorize as

$$q(\mathbf{s}) = \prod_{j=1}^{J} \prod_{f=0}^{F-1} \prod_{n=0}^{N-1} q_{jfn}(s_{j,fn}).$$



Under the mean-field approximation we can show that:

$$q^{\star}_{jfn}(s_{j,fn}) = \arg\max_{q_{jfn}} \mathcal{L}(q; \boldsymbol{\theta}_{\text{old}}) = N(s_{j,fn}; m_{j,fn}, \gamma_{j,fn}).$$

source estimate

M-Step

Maximize (or only increase) the variational free energy w.r.t the $\boldsymbol{\theta}$.

NMF parameters

Compute an NMF with the Itakura-Saito divergence on:

$$\left\langle s_{j,fn}^2 \right\rangle_{q^\star} = m_{j,fn}^2 + \gamma_{j,fn},$$

$\rightarrow$ standard multiplicative update rules.

Mixing filters

Solve a Toeplitz system of equations for $\mathbf{a}_{ij} = [a_{ij}(0), \ldots, a_{ij}(L_a - 1)]^T$.

Noise variance

$$\sigma_i^2 = \frac{1}{T} \sum_{t=0}^{T-1} \left\langle \left(x_i(t) - \sum_{j=1}^{J} [a_{ij} \star s_j](t) \right)^2 \right\rangle_{q^\star}.$$

Outline

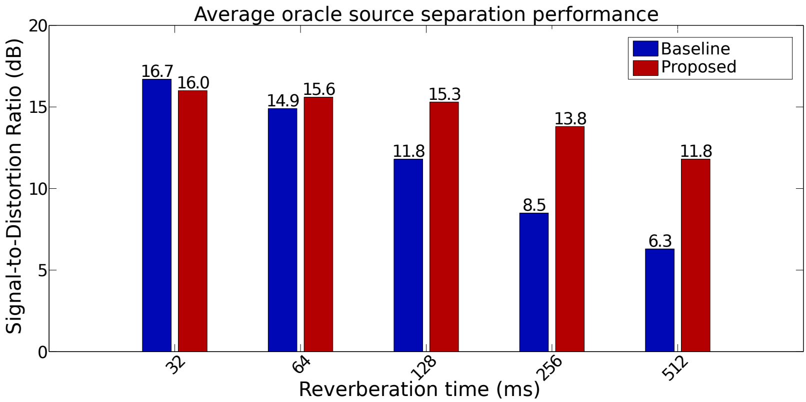
Experiments

- Dataset:
  - 5 reverberation times: 32, 64, 128, 256, 512 ms;
  - 5 $\times$ 8 stereo mixtures created with synthetic room impulse responses;
  - Number of sources: 3 to 5;
  - Mixture length: 12 to 28 seconds.
- Baseline approach [2]
  - Gaussian NMF-based source model with STFT approximation of the convolutive mixing process.
- Oracle initialization:
  - Parameters are initialized using the true source signals and mixing filters.
- Performance measure:
  - Signal-to-Distortion Ratio (SDR).

[2] A. Ozerov and C. Févotte. "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". *IEEE Transactions on Audio, Speech and Language Processing*, 2010.

Oracle source separation results

STFT and MDCT analysis/synthesis window length: 128 ms.



Other standard energy ratios (ISR, SIR, SAR) are in the paper.

Semi-blind audio example

Room impulse responses from the RWCP database: recorded in a real room (reverberation time of 470 ms).

Semi-blind setting: The mixing filters are known and fixed while all the other parameters are blindly estimated.

Stereo mixture:

	Original	Baseline	Proposed
Drums			
Guitar 1			
Guitar 2			
Voice			
Bass			

Musical excerpt from "Ana" by Vieux Farka Toure. MTG MASS database.

Outline

Probabilistic model

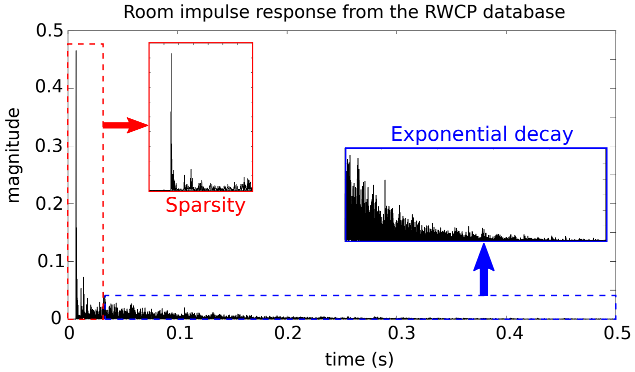
Inference

Experiments

Future work

Future work

- **Multi-resolution** time-frequency source modeling;
- Probabilistic priors on the mixing filters in the **time-domain**;



- **Blind** source separation method.

Thank you

More audio examples and Matlab code available at:

`https://perso.telecom-paristech.fr/leglaive/`