Separating Time-Frequency Sources from Time-Domain Convolutive Mixtures Using Non-negative Matrix Factorization

Simon Leglaive, Roland Badeau, Gaël Richard

LTCI, Télécom ParisTech, Université Paris Saclay

IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)

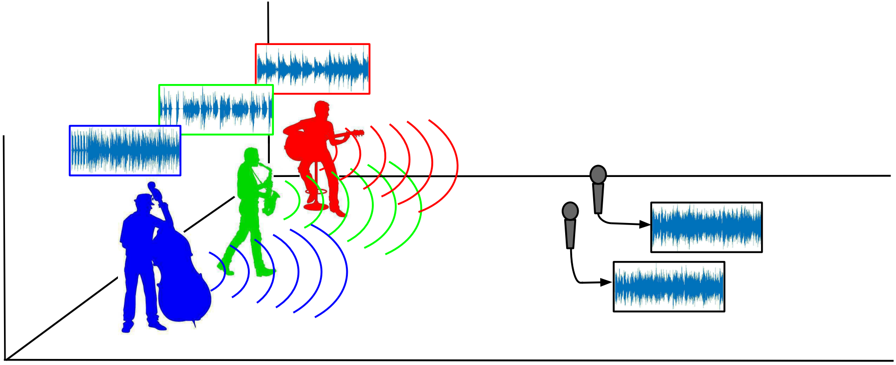
New Paltz, NY

October 17, 2017

Multichannel audio source separation

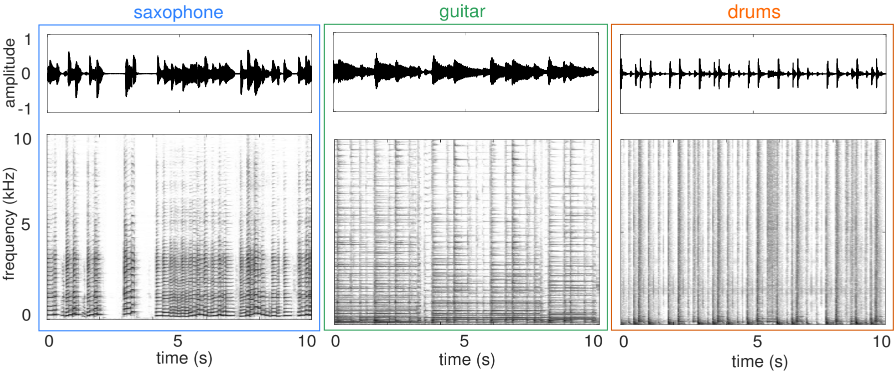
Objective: Recover source signals from the observation of several mixtures.

Context: Under-determined.



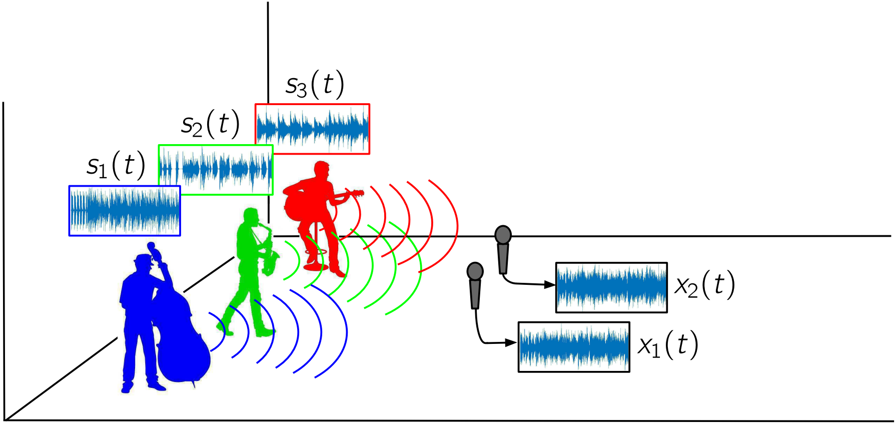
Time-frequency source representation

Time-frequency (TF) transforms provide meaningful representations.



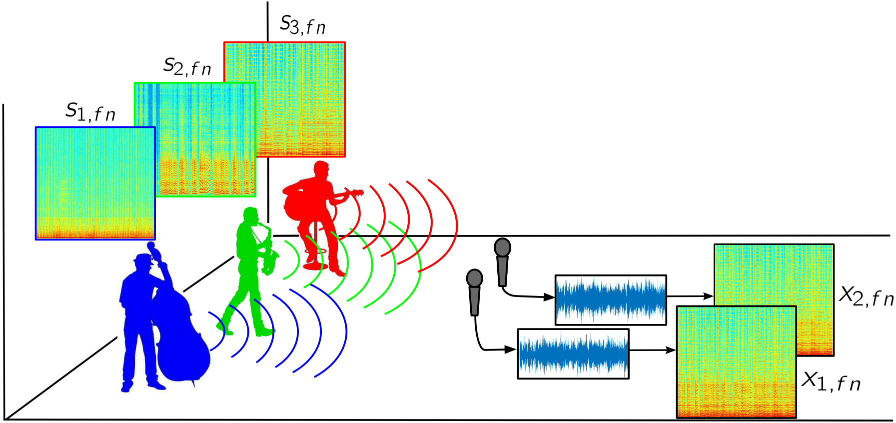
Reverberant mixtures (1)

Convolutive mixing process in the time domain: $x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$



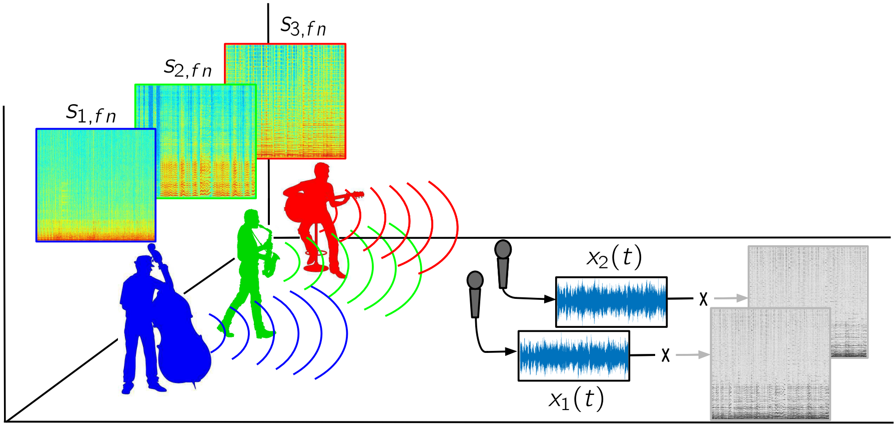
Reverberant mixtures (2)

Convolutive mixing process in the STFT domain: $x_{i,fn} \approx \sum_{j=1}^{J} a_{ij,f}\, s_{j,fn}$



Proposed approach

- Time-domain mixture representation: $x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$
- Time-frequency source representation: $s_j(t) = \mathcal{T}^{-1}(\{s_{j,fn}\}_{f,n})$



Choosing the time-frequency transform

- Modified Discrete Cosine Transform (MDCT)
  - Real valued;
  - Critically sampled;
  - No shift-invariance (phase information is contained in the amplitude of the MDCT coefficients).
- Odd-Frequency Short-Time Fourier Transform (OFSTFT)
  - Complex valued;
  - Redundant;
  - Shift-invariance.
- General TF synthesis equation:

$$s_j(t) = \frac{2}{\phi}\Re\left(\sum_{f=0}^{F-1}\sum_{n=0}^{N-1} s_{j,fn}\psi_{fn}(t)\right), \quad \text{with} \quad \phi = \begin{cases} 2 & \text{if MDCT } (s_{j,fn} \in \mathbb{R}) \\ 1 & \text{if OFSTFT } (s_{j,fn} \in \mathbb{C}) \end{cases}.$$

Outline

Probabilistic modeling with latent variables

- Latent TF source random variables: $\mathbf{s} = \{s_{j,fn} \in \mathbb{R} \text{ or } \mathbb{C}\}_{j,f,n}$
- Observed time-domain 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.

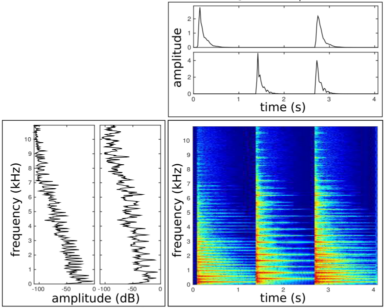
- What prior knowledge do we have on the latent source variables?
- How are the data generated from the latent unobserved variables?

Prior distribution of the latent variables

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

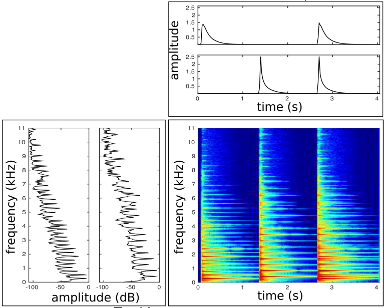
$$s_{j,fn} \sim \mathcal{N}(0, [\mathbf{W}_j\mathbf{H}_j]_{fn})$$

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



$\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ MDCT power spectrogram (dB)

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



$\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ OFSTFT power spectrogram (dB)

---

[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 $s_j(t) = \frac{2}{\phi} \Re \left(\sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \psi_{fn}(t) \right)$.

Conditional distribution

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

Outline

Inference

Posterior distribution

We are interested in the posterior distribution of the latent variables:

$$p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star) \qquad \text{with} \qquad \boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta})$$

- Model parameters: $\boldsymbol{\theta} = \{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i\}$
- Semi-blind setting: the mixing filters are assumed to be known.

The posterior distribution is Gaussian but with a high-dimensional full covariance matrix $\rightarrow$ **variational inference** to reduce the computational cost.

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:

$$KL\Big(q(\mathbf{s})\,||\,p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})\Big) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{Log-likelihood}} \quad - \quad \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{Variational Free Energy}} \quad , \tag{1}$$

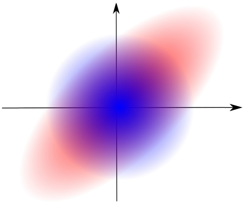
where $\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})}{q(\mathbf{s})}\right)\right\rangle_q$ and $\langle f(\mathbf{z})\rangle_q = \int f(\mathbf{z})q(\mathbf{z})d\mathbf{z}$.

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

Mean-field approximation

$\mathcal{F}$ is the set of pdfs 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:

$$q^{\star}_{jfn}(s_{j,fn}) = \begin{cases} N_{\mathbb{R}}(\hat{s}^{r}_{j,fn}, \gamma^{r}_{j,fn}) & \text{if MDCT} \\ N_{\mathbb{C}}(\rho_{j,fn}, \hat{s}^{r}_{j,fn}, \hat{s}^{\imath}_{j,fn}, \gamma^{r}_{j,fn}, \gamma^{\imath}_{j,fn}) & \text{if OFSTFT} \end{cases}. \quad (2)$$

In the OFSTFT case, $\Re(s_{j,fn})$ and $\Im(s_{j,fn})$ are correlated a posteriori.

Outline

Experiments

- Dataset:
  - 8 stereo mixtures created with measured room impulse responses from the RWCP database [2].
  - Reverberation time: 470 ms.
  - Number of sources per mixture: 3 to 5.
  - Mixture length: 12 to 28 seconds.

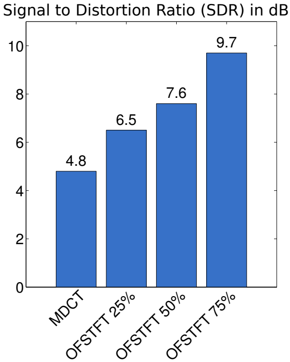
- Semi-blind setting: Mixing filters are known while all other parameters are blindly estimated.

[2] S. Nakamura et al. "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition". *Proc. of LREC*, 2000.

MDCT *vs.* OFSTFT

Comparison of the source separation results using:

- the MDCT;
- the OFSTFT with several overlap ratios: 25%, 50% and 75%.

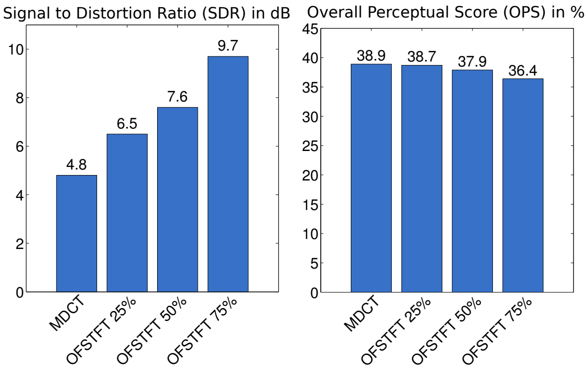


Audio examples available online (website address in the paper)

MDCT *vs.* OFSTFT

Comparison of the source separation results using:

- the MDCT;
- the OFSTFT with several overlap ratios: 25%, 50% and 75%.



Audio examples available online (website address in the paper)

Baseline methods

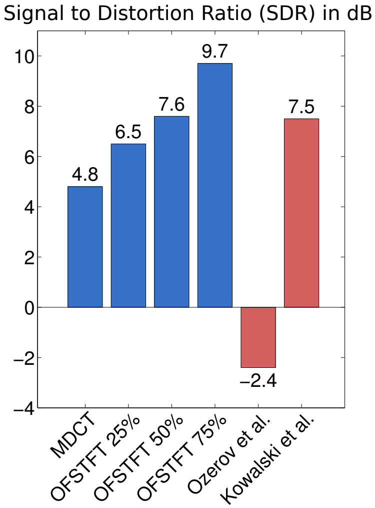
	Time-frequency source model (STFT domain)	Convolutive mixture representation
Ozerov et al. [3]	Gaussian NMF-based	approximate (STFT)
Kowalski et al. [4]	sparse (ℓ_1 norm)	exact (time)

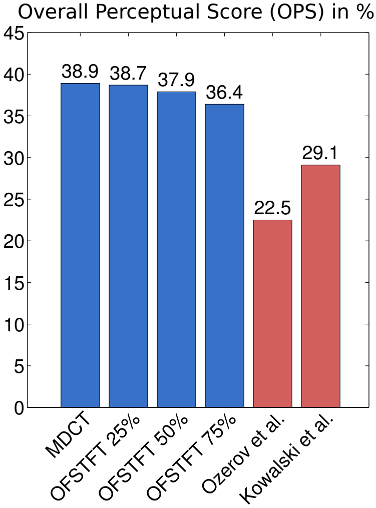
Length of the TF analysis/synthesis window: 128 ms.

[3] A. Ozerov, C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation", *IEEE Trans. Audio, Speech, Language Process.*, 2010.

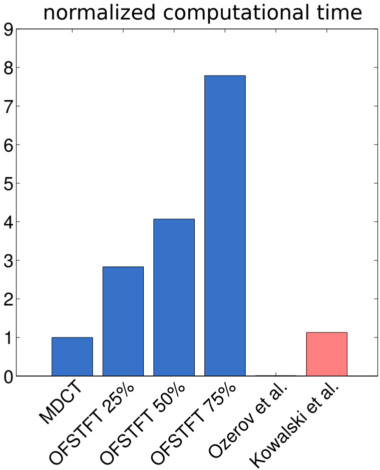
[4] M. Kowalski, E. Vincent, R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation", *IEEE Trans. Audio, Speech, Language Process.*, 2010.

Source separation results





Computational time



Separation of a 12 second-long mixture of 3 sources at 16 kHz:
$\sim$ 2 hours with the MDCT-based method.

Audio example

Stereo mix:

	Guitar 1	Guitar 2	Voice	Drums	Bass
Original source (stereo)					
Ozerov et al.					
Kowalski et al.					
Proposed (MDCT)					

Excerpt from "Ana" by Vieux Farka Toure.

Outline

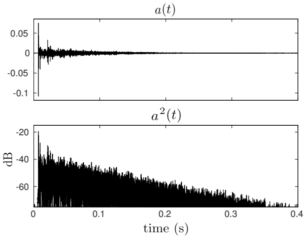
Conclusion

Conclusion:

- Working with the MDCT is computationally cheaper than with the OFSTFT and leads to similar perceived results.

Further work:

- Source-specific time-frequency resolution;
- Probabilistic priors on the mixing filters in the time domain.



Journal preprint:

- S. L., R. Badeau, G. Richard, "Student's t source and mixing models for multichannel audio source separation", submitted, 2017;
- Available online: `https://hal.archives-ouvertes.fr/hal-01584755`;
- Poster at the SANE Workshop.

Thank you

More audio examples and Matlab code available at:

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

Modified discrete cosine transform

- MDCT synthesis equation:

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

  - $\psi_{fn}(t) = \sqrt{\frac{2}{F}} w(t - nH) \cos\left(\frac{2\pi}{L_w}\left(t - nH + \frac{1}{2} + \frac{L_w}{4}\right)\left(f + \frac{1}{2}\right)\right);$
  - $w(t)$: synthesis window of length L_w;
  - $F = H = L_w/2$.

Short-time Fourier transform

- STFT synthesis equation:

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

  - $\psi_{fn}(t) = \sqrt{\frac{1}{L_w}} w(t - nH) \exp\left(\imath \frac{2\pi}{L_w} f(t - nH)\right);$
  - $F = L_w.$

- Hermitian symmetry: deterministic relation between TF coefficients.
- Using the Hermitian symmetry property:

$$s_j(t) = \sum_{n=0}^{N-1} \left[\underbrace{s_{j,0n}\, \psi_{0n}(t)}_{\text{Zero frequency}} + \underbrace{s_{j,\frac{F}{2}n}\, \psi_{\frac{F}{2}n}(t)}_{\text{Nyquist frequency}} + 2\Re \left(\sum_{f=1}^{F/2-1} s_{j,fn} \psi_{fn}(t) \right) \right].$$

Odd-frequency short-time Fourier transform

- OFSTFT synthesis equation:

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

  - $\psi_{fn}(t) = \sqrt{\frac{1}{L_w}}w(t-nH)\exp\left(\imath\frac{2\pi}{L_w}(f+\frac{1}{2})(t-nH)\right)$;
  - $F = L_w/2$;
  - H: hop size.

- **All TF coefficients are complex valued.**