Semi-Blind Student's t Source Separation for Multichannel Audio Convolutive Mixtures

Simon Leglaive, Roland Badeau, Gaël Richard

LTCI, Télécom ParisTech, Université Paris Saclay

European Signal Processing Conference

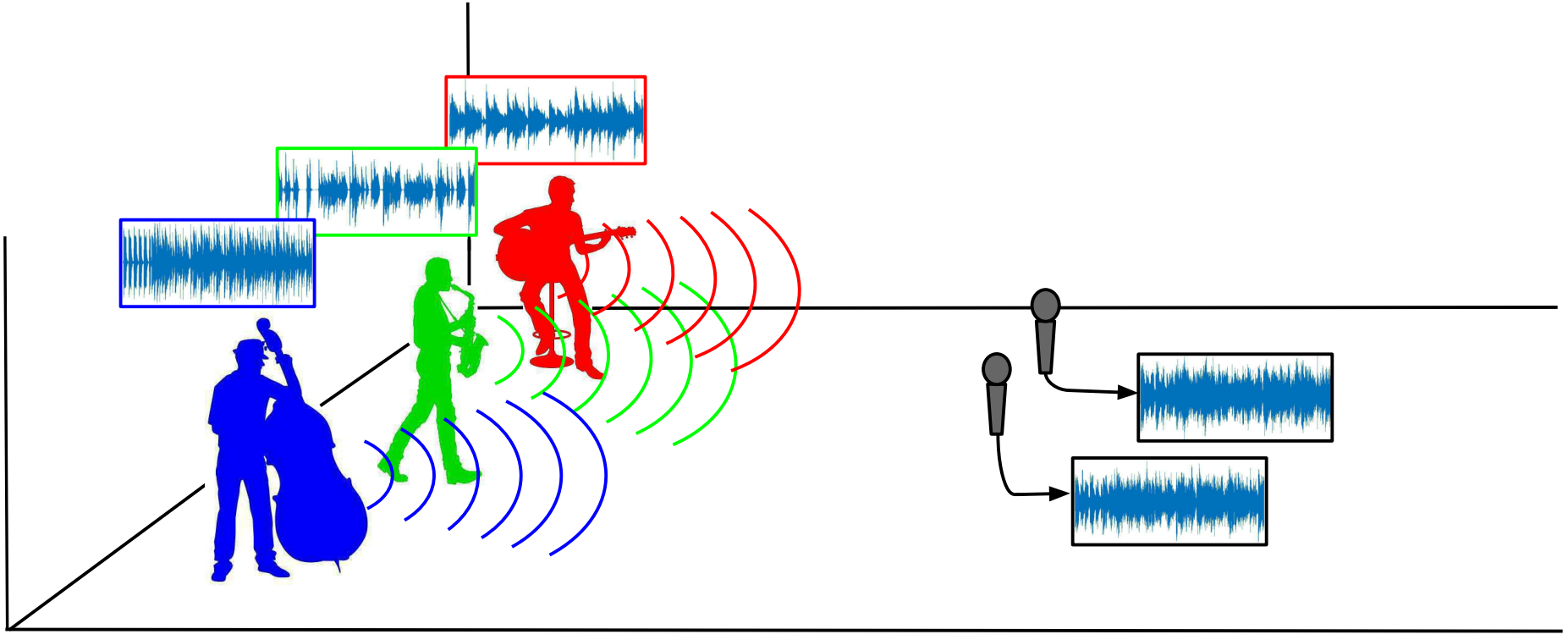
Kos island, Greece

August 31, 2017

Multichannel audio source separation

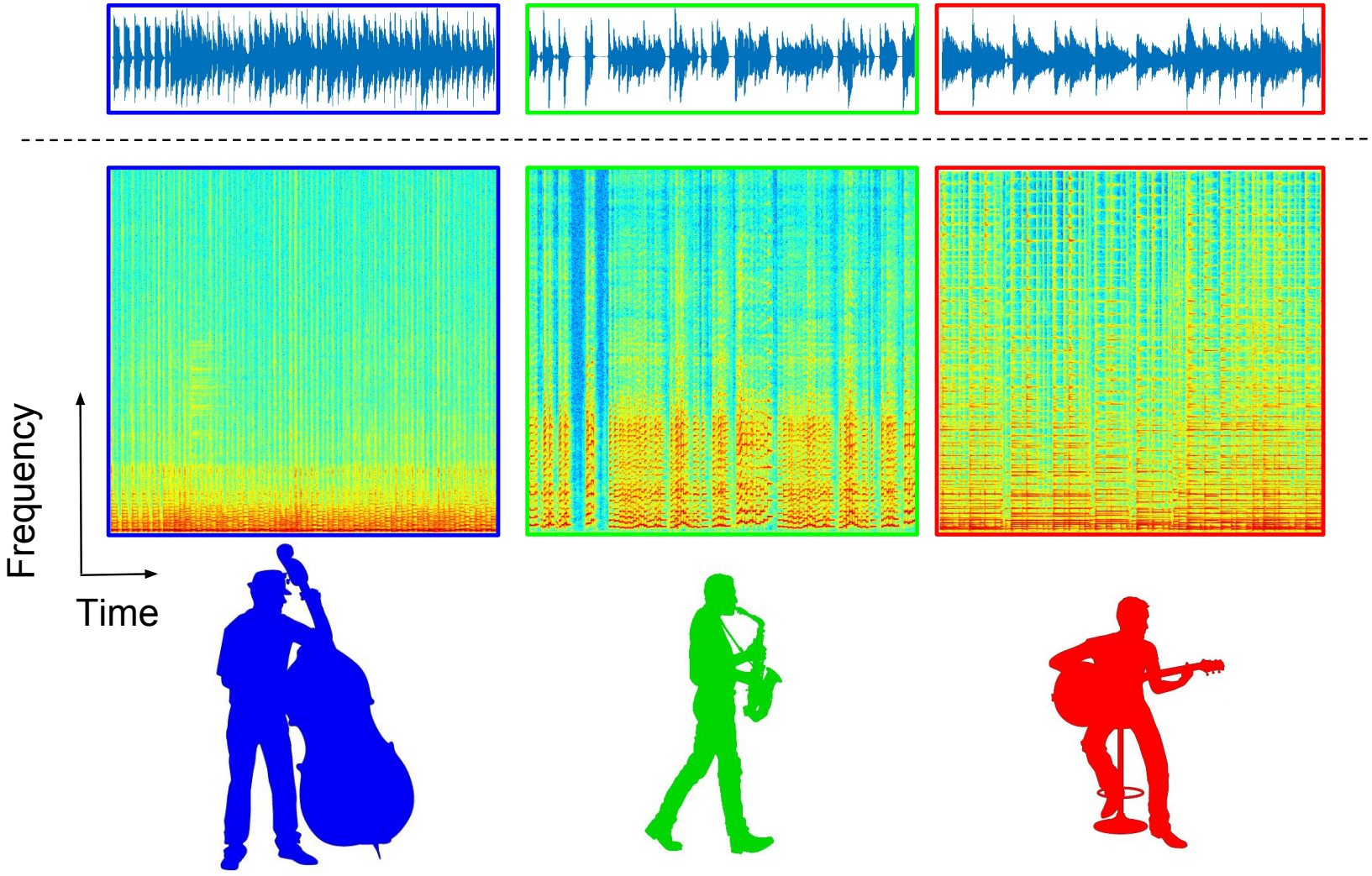
Objective: Recover source signals from the observation of several mixtures.

Context: Under-determined and reverberant.



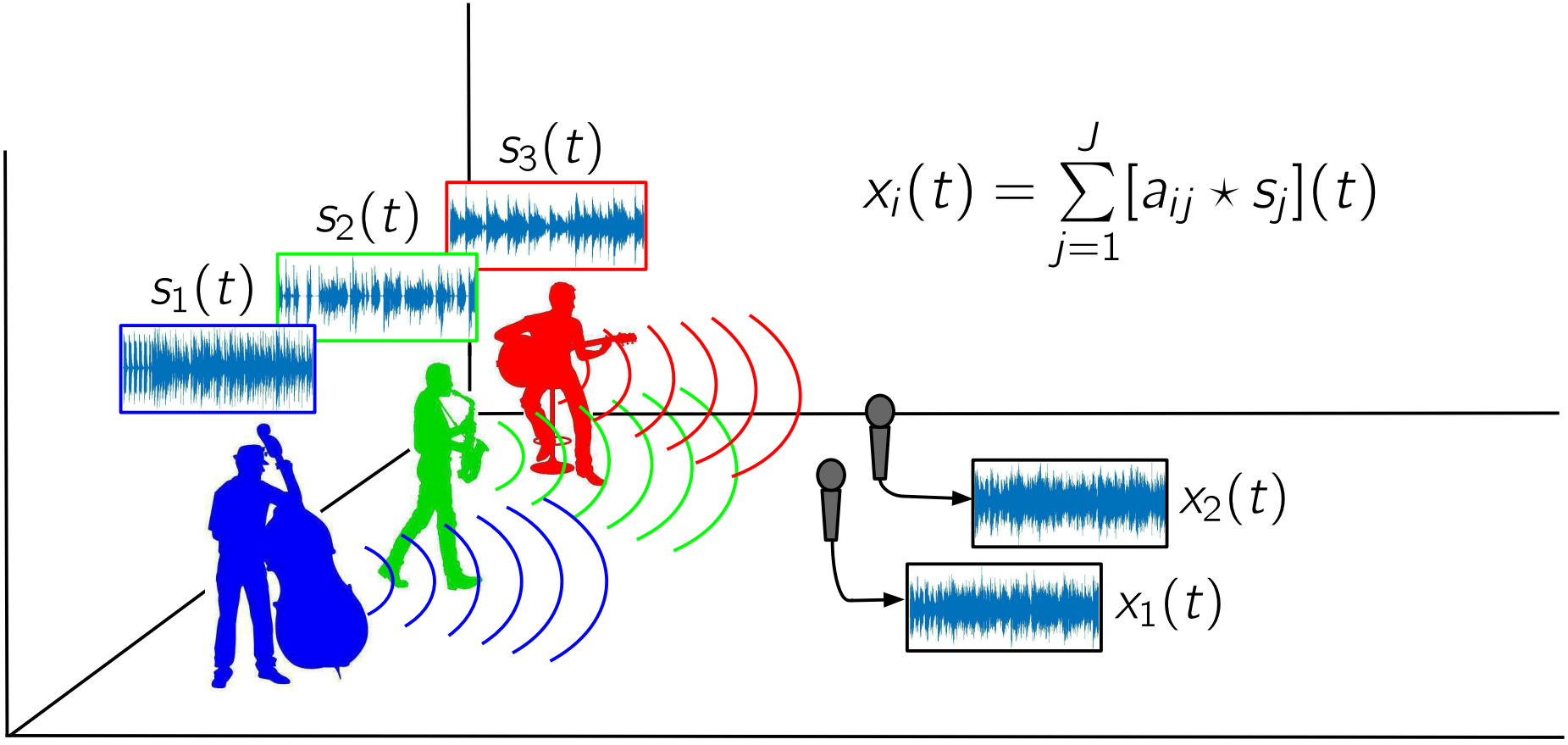
Time-frequency source representation

Time-frequency (TF) transforms provide meaningful representations.



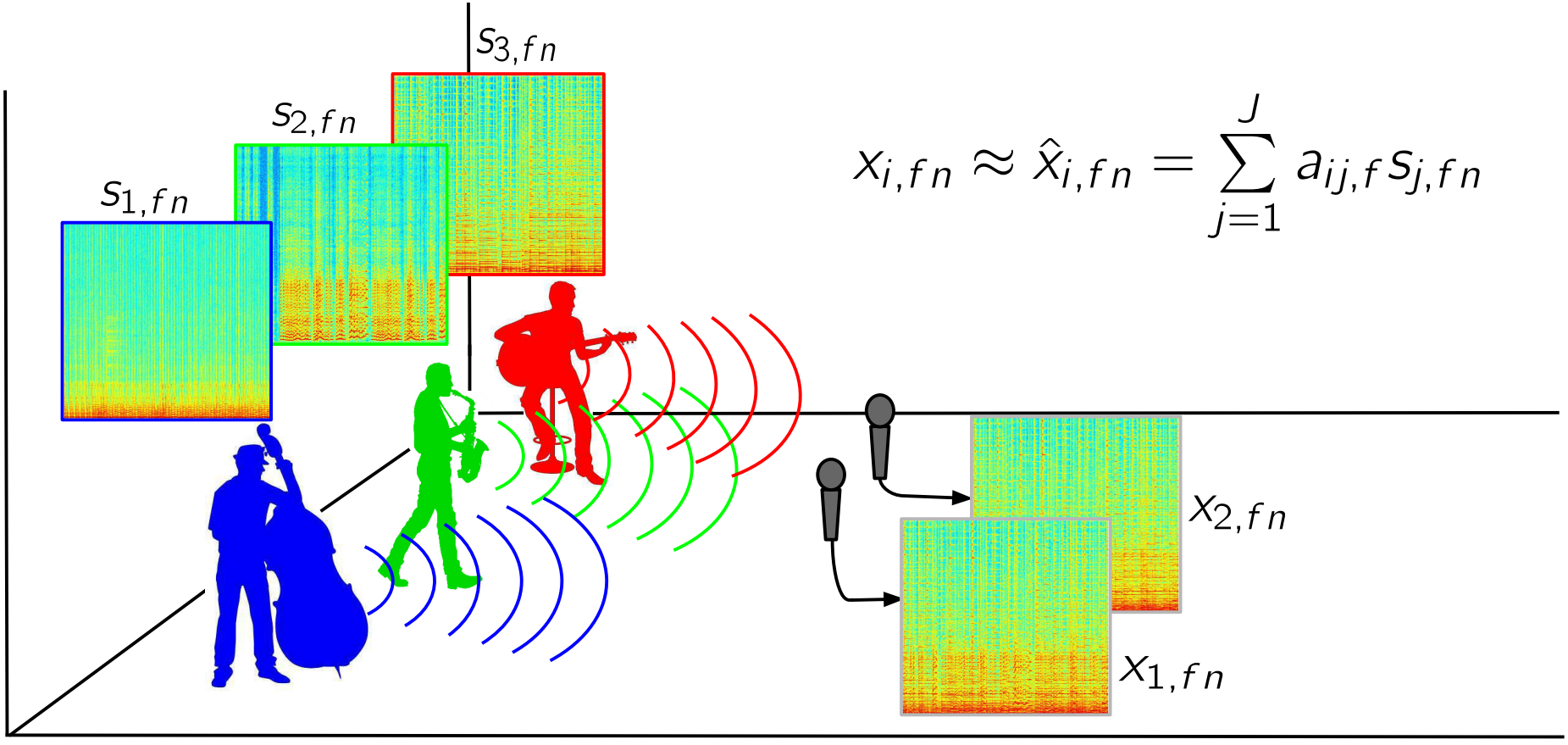
Reverberant mixtures (1)

Convolutive mixing process in the time domain:



Reverberant mixtures (2)

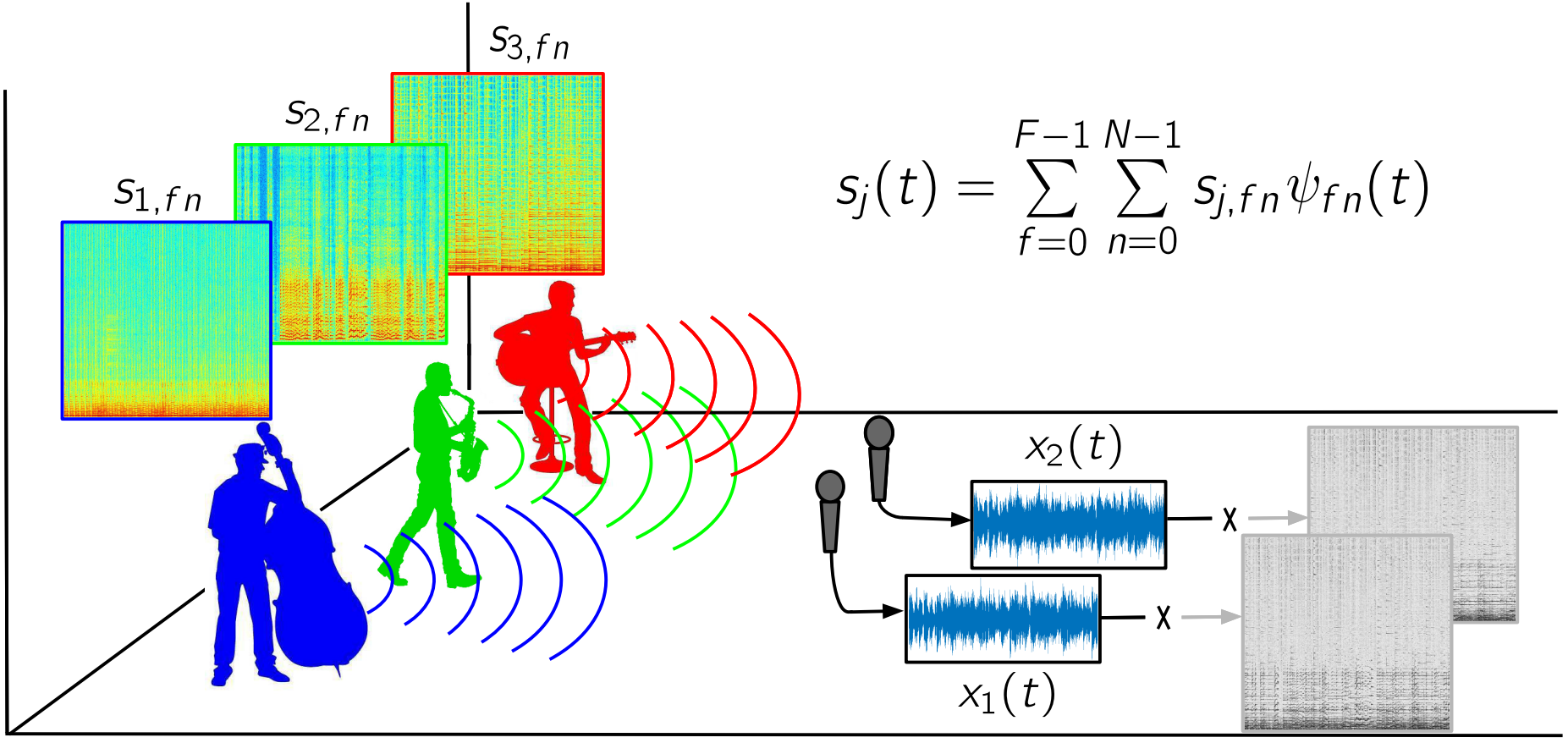
Convolutive mixing process in the Short-Term Fourier Transform (STFT) domain:



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

Proposed approach

TF source model and time-domain convolutive mixture representation.



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

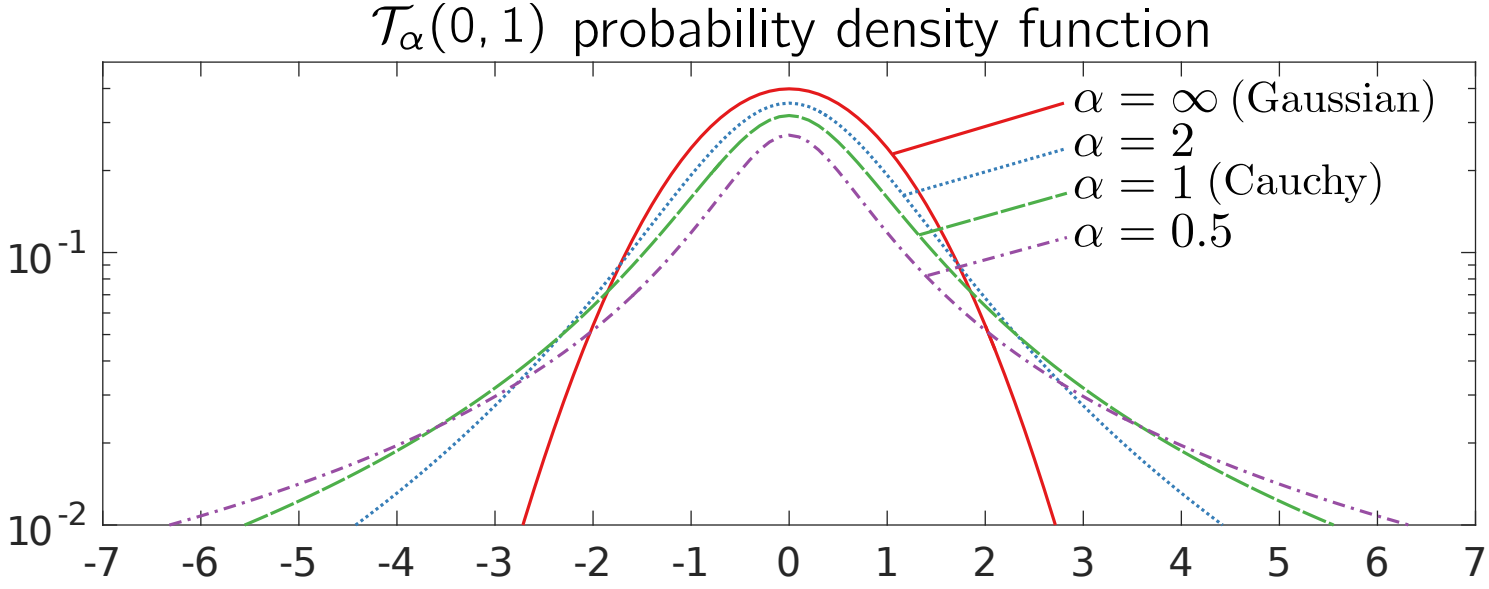
$\psi_{fn}(t)$ is a Modified Discrete Cosine Transform (MDCT) atom.

Outline

Student's t source model

Student's t model as conditionally Gaussian

$$s_{j,fn} \sim \mathcal{T}_\alpha(0, \lambda_{j,fn}) \quad \Leftrightarrow \quad \begin{cases} s_{j,fn}|v_{j,fn} & \sim \mathcal{N}(0, v_{j,fn}) \\ v_{j,fn} & \sim \mathcal{IG}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\lambda_{j,fn}^2\right) \end{cases}$$



Sparsity and NMF-based source models

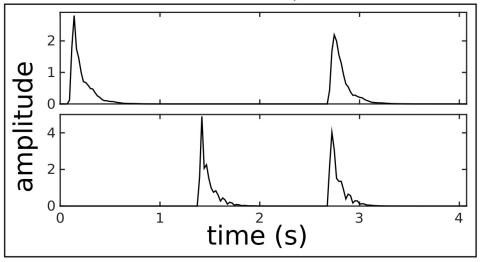
Student's t sparse source model [Févotte and Godsill, 2006]

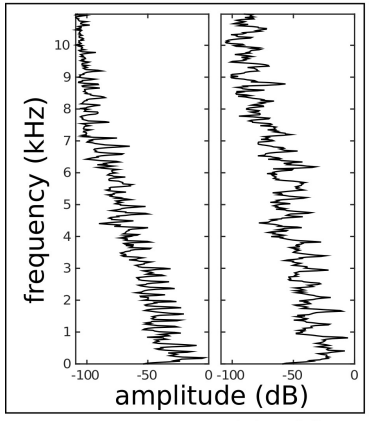
$$\lambda_{j,fn}^2 = \lambda_j^2$$

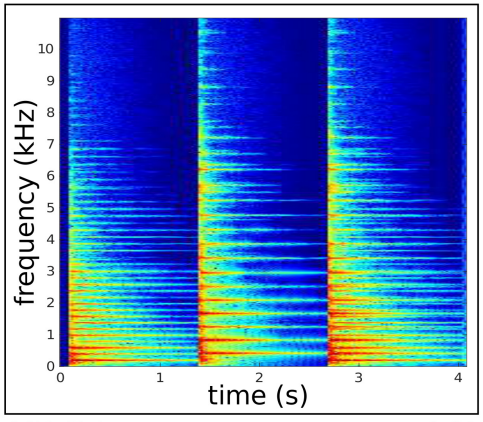
Student's t non-negative matrix factorization (NMF) [Yoshii et al., 2016]

$$\lambda_{j,fn}^2 = [\mathbf{W}_j \mathbf{H}_j]_{f,n}$$

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







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

MDCT power spectrogram (dB)

Convolutive mixture model

Gaussian modeling error

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

with $b_i(t) \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_i^2)$.

We recall that:

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

with $\psi_{fn}(t)$ an MDCT atom.

Outline

Inference

- TF latent variables: $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$ and $\mathbf{v} = \{v_{j,fn}\}_{j,f,n}$
- Time-domain observed variables: $\mathbf{x} = \{x_i(t)\}_{i,t}$
- Model parameters: $\boldsymbol{\theta} = \left\{\{\lambda^2_{j,fn}\}_{j,f,n}, \{a_{ij}(t)\}_{i,j,t}, \{\sigma^2_i\}_i\right\}$
- Semi-blind setting: the mixing filters are assumed to be known.

Exact posterior inference

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

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

Exact posterior inference is analytically intractable with the proposed model $\rightarrow$ **approximate inference**.

Variational inference

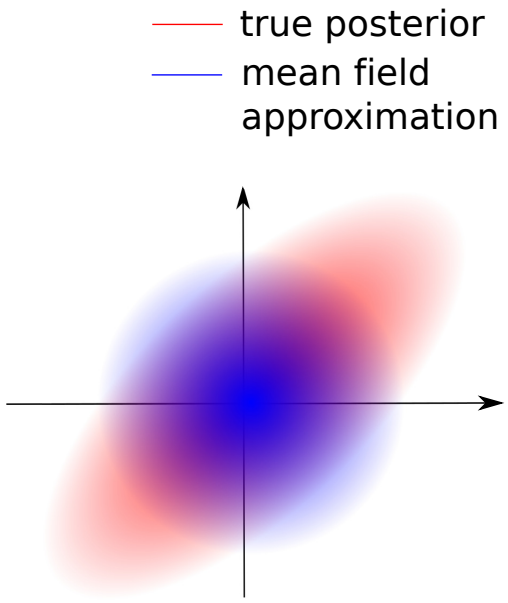
- We want to find $q \in \mathcal{F}$ which approximates $p(\mathbf{s}, \mathbf{v}|\mathbf{x}; \boldsymbol{\theta})$.
- Taking the KL divergence as a measure of fit, we can show that:

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

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

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

Mean-field approximation



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

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

Under the mean-field approximation we can show that:

$$q^{s^\star}_{jfn}(s_{j,fn}) = \arg\max_{q^s_{jfn}} \mathcal{L}(q; \boldsymbol{\theta}^\star) = N(s_{j,fn};\ \hat{s}_{j,fn}, \gamma_{j,fn})$$

$$q^{v^\star}_{jfn}(v_{j,fn}) = \arg\max_{q^v_{jfn}} \mathcal{L}(q; \boldsymbol{\theta}^\star) = IG(v_{j,fn};\ \delta, \beta_{j,fn})$$

M-Step

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

Source model parameters

- Sparse model: $\lambda_j^2 = \left(\frac{1}{FN} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \left\langle v_{j,fn}^{-1} \right\rangle_{q^\star} \right)^{-1}$
- NMF model: $\min_{\mathbf{W}_j, \mathbf{H}_j \geq 0} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} d_{IS} \left([\mathbf{W}_j \mathbf{H}_j]_{fn}, \left\langle v_{j,fn}^{-1} \right\rangle_{q^\star}^{-1} \right)$

  $\rightarrow$ Majorize-minimize approach to obtain multiplicative update rules.

Noise variance

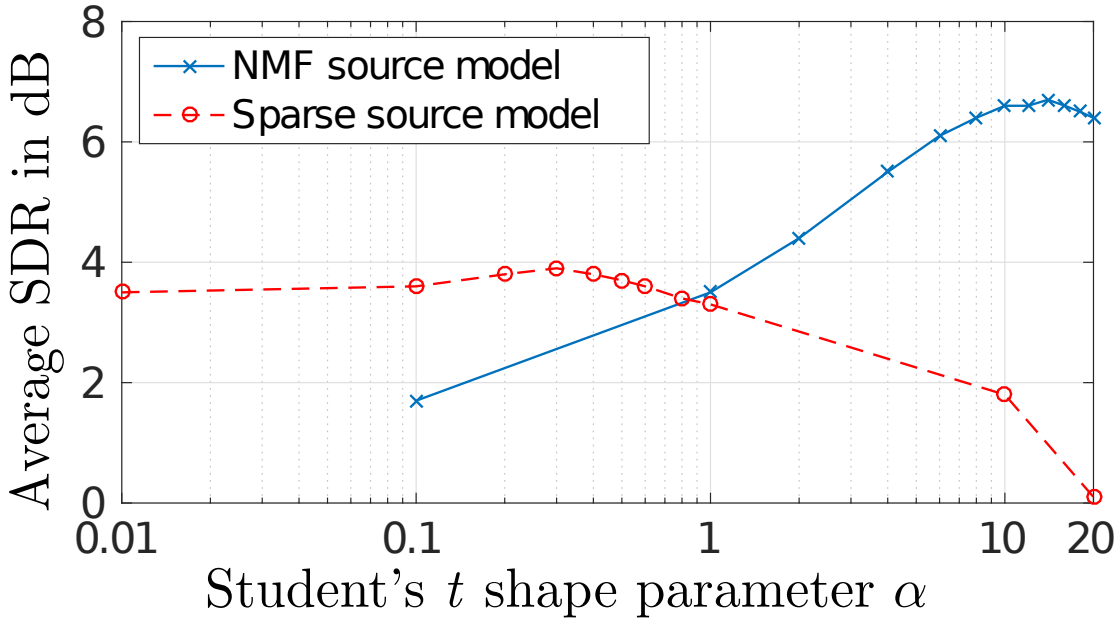
$$\sigma_i^2 = \frac{1}{T} \sum_{t=0}^{T-1} \left\langle \left(x_i(t) - \sum_{j=1}^{J} [a_{ij} \star s_j](t) \right)^2 \right\rangle_{q^\star}$$

Outline

Experiments

- Dataset:
  - 8 stereo mixtures created with synthetic room impulse responses
  - Reverberation time: 256 ms
  - Number of sources: 3 to 5
  - Mixture length: 12 to 28 seconds
- Semi-blind setting:
  - Mixing filters are known while all the other parameters are blindly estimated.
- Performance measures:
  - Signal-to-Distortion (SDR), Interferences (SIR) and Artifacts (SAR) Ratios, in decibels (dB).

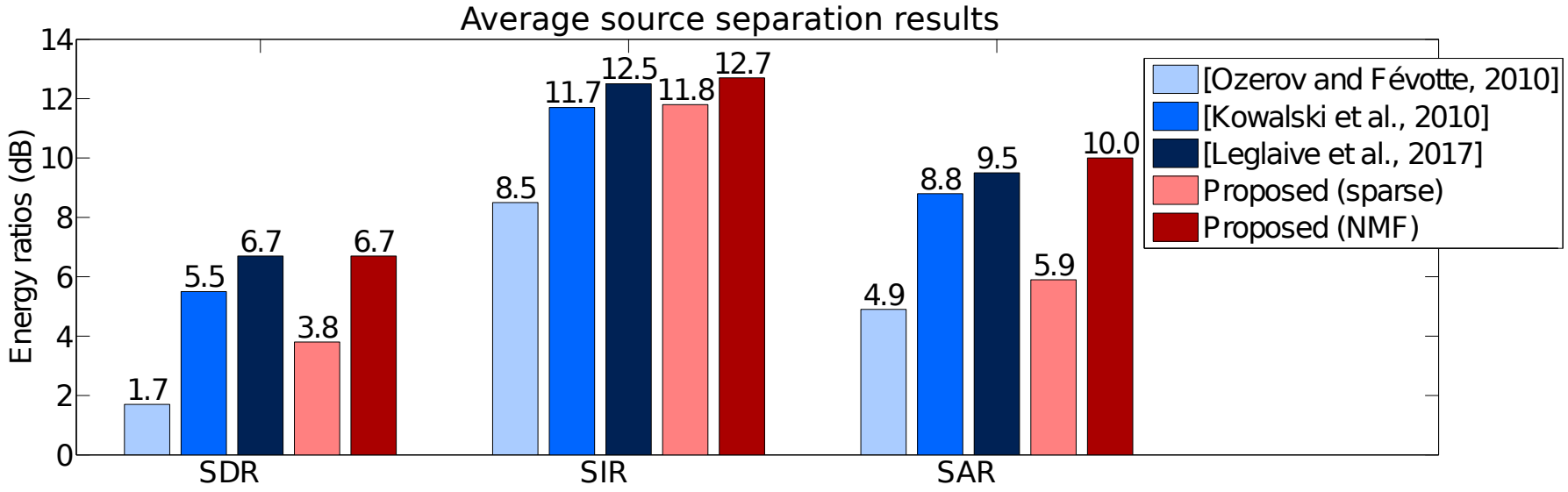
Sparse *versus* NMF-based source models



- The sparse model requires a smaller shape parameter compared with the NMF-based source model.
- The best results are obtained by modeling the spectro-temporal characteristics of the sources with NMF.

Source separation results

Reference	Source model (TF domain)	Convolutive mixture representation
[Kowalski et al., 2010]	Sparse (ℓ_1 norm)	Exact (time domain)
[Ozerov and Févotte, 2010]	Gaussian NMF-based	Approximate (STFT domain)
[Leglaive et al., 2017]	Gaussian NMF-based	Exact (time domain)



TF analysis/synthesis window length: 128 ms.

Audio example

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

Stereo mixture: 

	Guitar 1	Guitar 2	Voice	Drums	Bass
Original source					
[Ozerov and Févotte, 2010]					
[Kowalski et al., 2010]					
Proposed (NMF)					

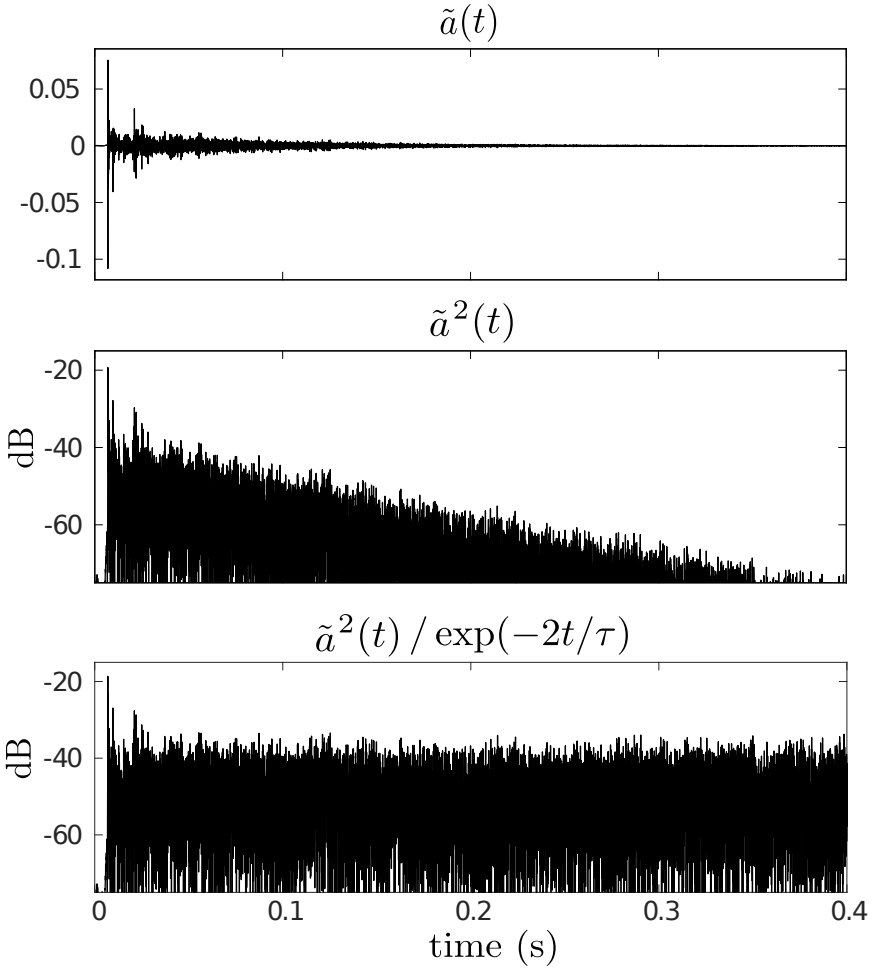
More audio examples available at:

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

Outline

Future work

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



Thank you

References

[Févotte and Godsill, 2006] C. Févotte, S.J. Godsill, "A Bayesian approach for blind separation of sparse sources", *IEEE Trans. Audio, Speech, Language Process.*, 2006.

[Yoshii et al., 2016] K. Yoshii, K. Itoyama, M. Goto, "Student's *t* nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation", in Proc. *IEEE Int. Conf. Acoust., Speech., Signal Process. (ICASSP)*, 2016.

[Kowalski et al., 2010] 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.

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

[Leglaive et al., 2017] S. Leglaive, R. Badeau, G. Richard, "Multichannel audio source separation: Variational inference of time-frequency sources from time-domain observations", in Proc. *IEEE Int. Conf. Acoust., Speech., Signal Process. (ICASSP)*, 2017.