# Séparation de sources audio en milieu réverbérant : Factorisation en matrices non-négatives et représentation temporelle du mélange convolutif

Simon Leglaive, Roland Badeau, Gaël Richard

LTCI, Télécom ParisTech, Université Paris Saclay

XXVI$^{\text{ème}}$ Colloque GRETSI

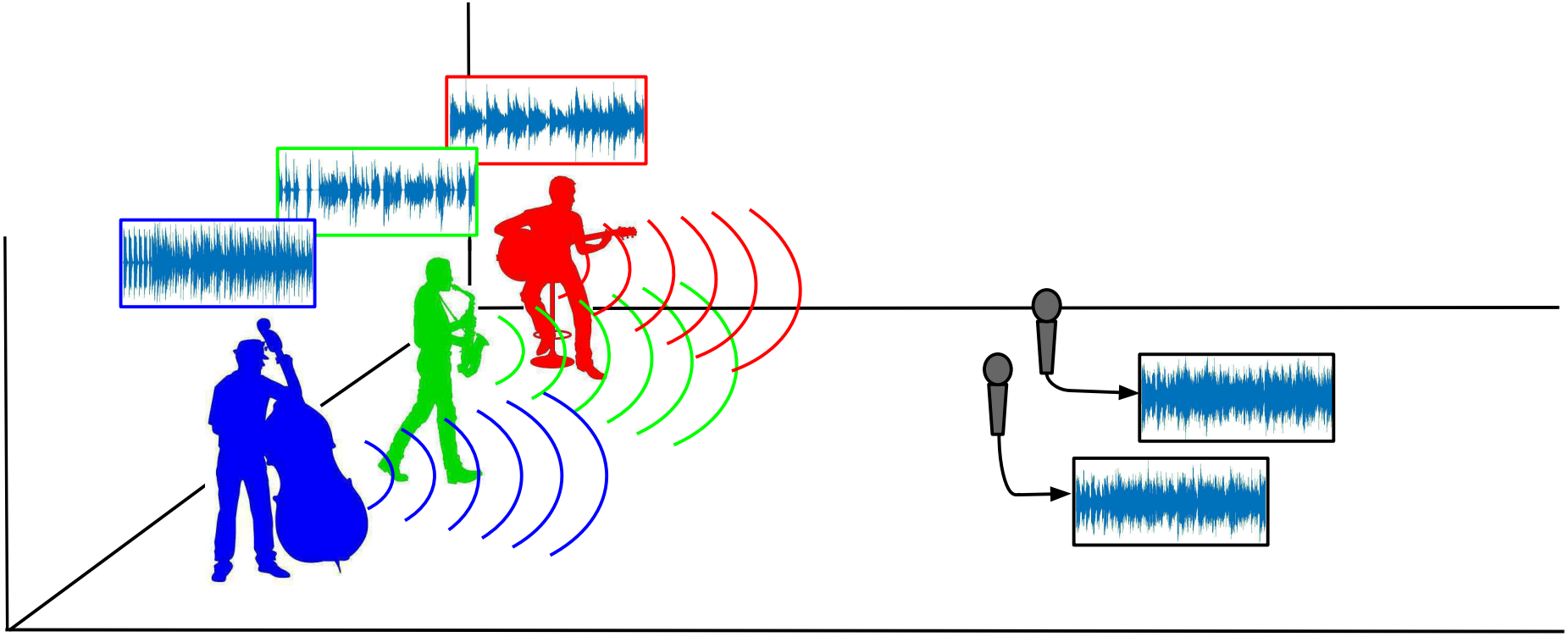
Juan-Les-Pins, France

8 septembre 2017

# Séparation multicanale de sources audio

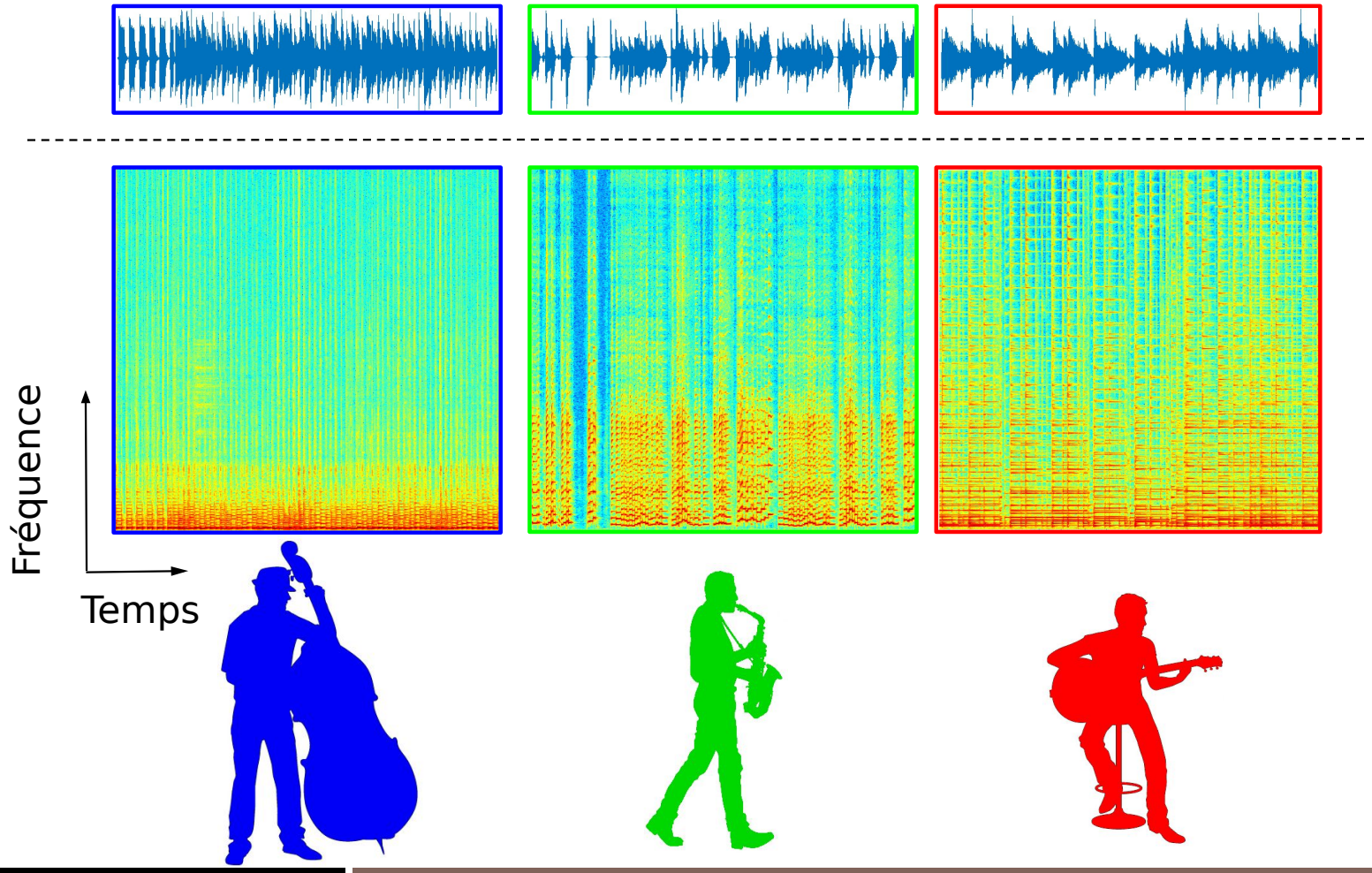
Objectif : Retrouver les signaux sources à partir de l'observation de leur mélange sur plusieurs canaux.

Contexte : Sous-déterminé et réverbérant.



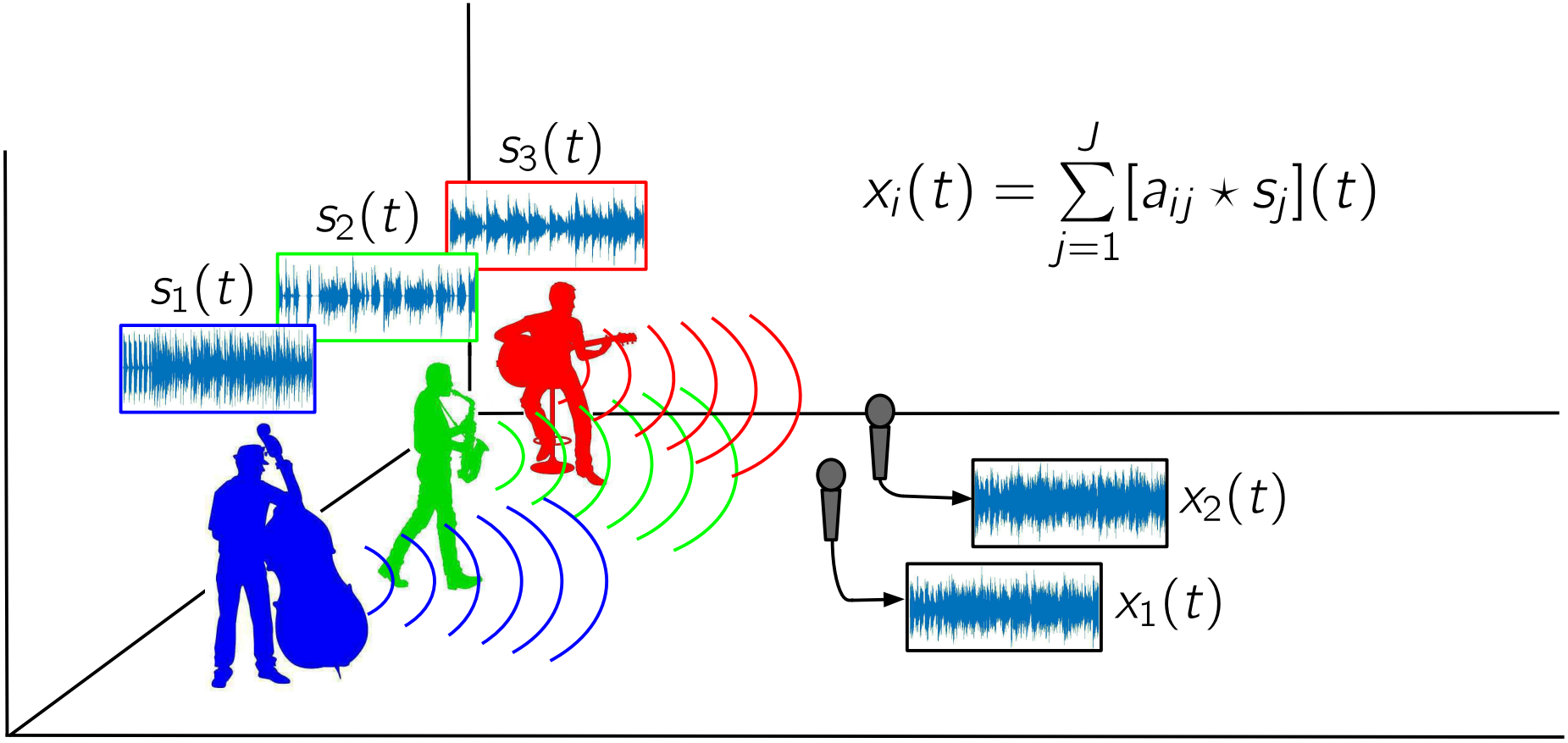
# Représentation temps-fréquence des sources

Une transformée temps-fréquence (TF) fournit une représentation qui permet de distinguer les caractéristiques spécifiques des sources.



# Mélange réverbérant (1)

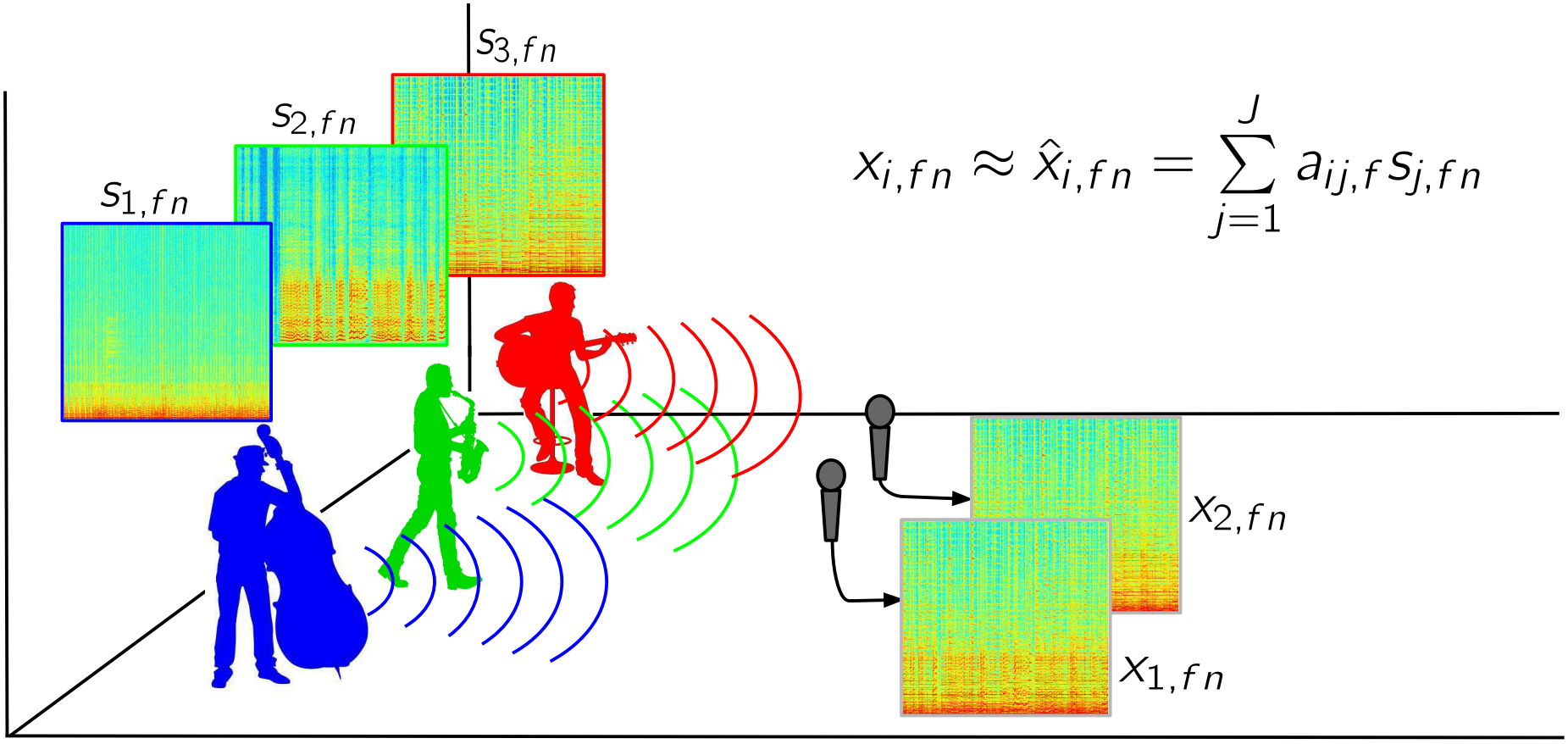
Mélange convolutif dans le domaine temporel :




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

# Mélange réverbérant (2)

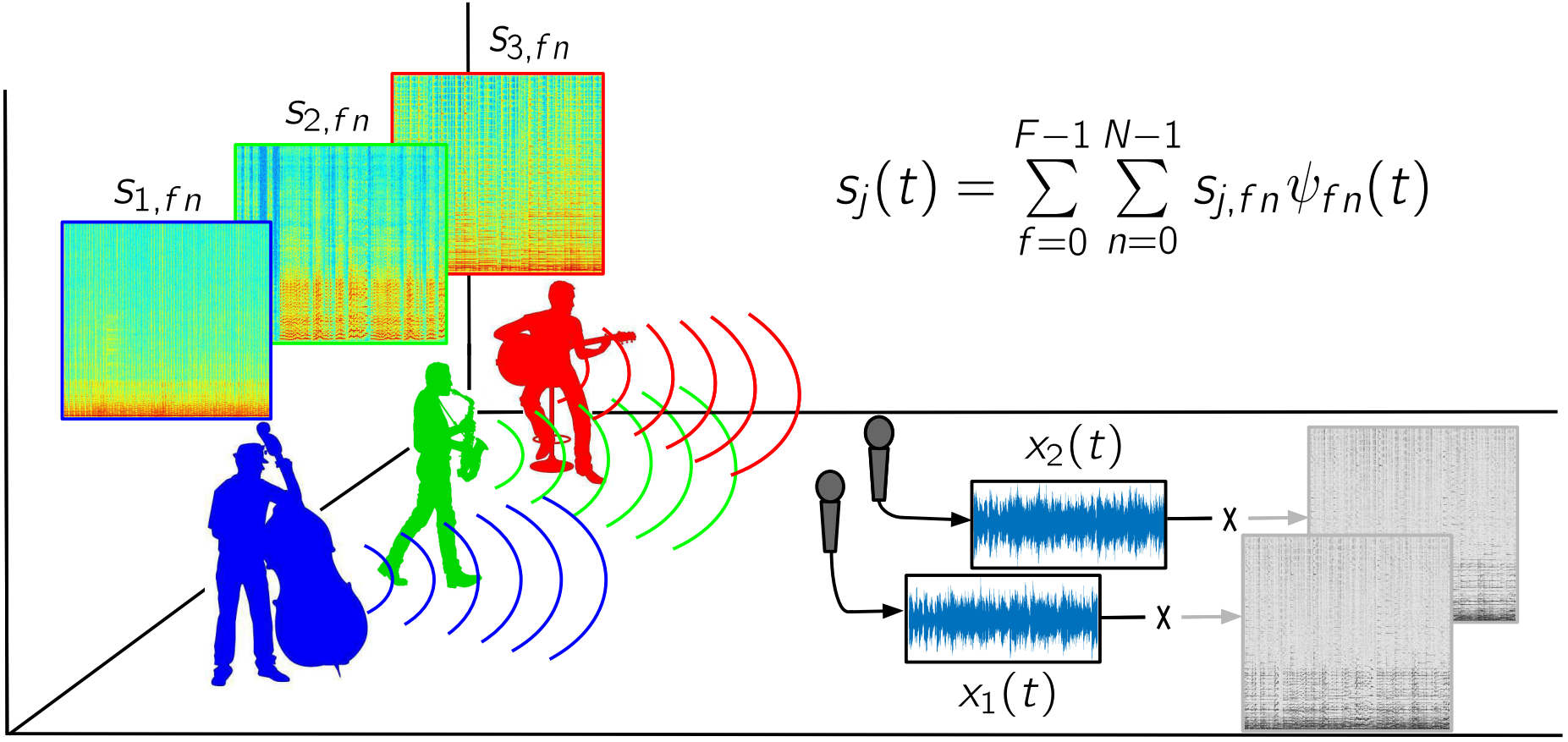
Mélange convolutif dans le domaine de la transformée de Fourier à court-terme (TFCT) :




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

# Approche proposée

Modèle TF de source et représentation temporelle du mélange convolutif.




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

$\psi_{fn}(t)$ est un atome de transformée en cosinus discrète modifiée (MDCT).

# Plan

Modèle probabiliste

Inférence

Expériences

Travaux futurs

# Modèle à variables latentes

- Variables latentes temps-fréquence : $\mathbf{s} = \{s_{j,fn} \in \mathbb{R}\}_{j,f,n}$;
- Observations temporelles : $\mathbf{x} = \{x_i(t) \in \mathbb{R}\}_{i,t}$.

## Définir un modèle probabiliste

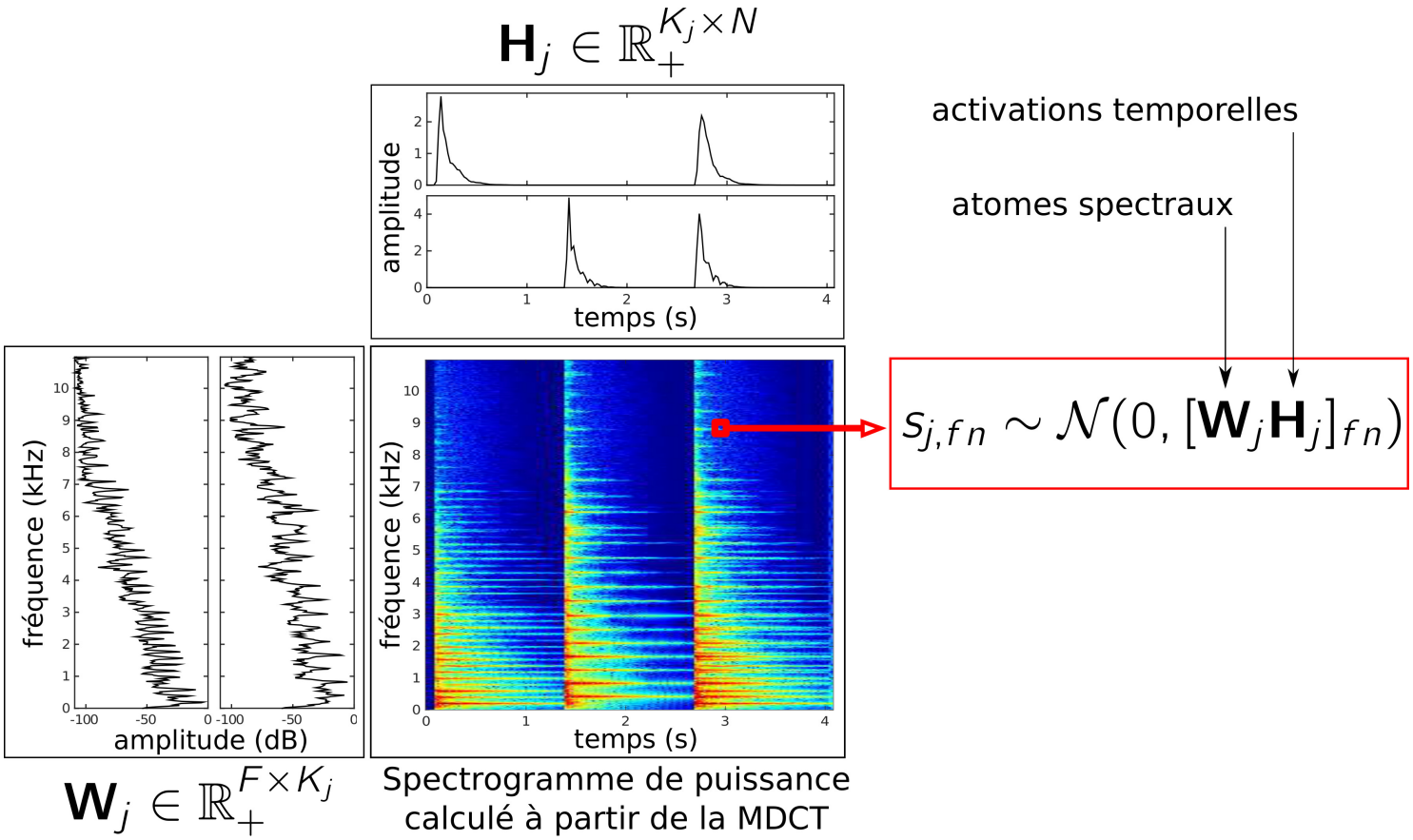
$$p(\mathbf{x}, \mathbf{s}; \boldsymbol{\theta}) = p(\mathbf{x}|\mathbf{s}; \boldsymbol{\theta})p(\mathbf{s}; \boldsymbol{\theta})$$

où $\boldsymbol{\theta}$ est un ensemble de paramètres déterministes.

- Quelle connaissance a priori avons-nous sur les variables latentes ?
- Comment les données observées sont-elles générées à partir des variables latentes ?

# Modèle de source

Modèle de source gaussien basé sur la factorisation en matrices non-négatives (NMF) [Févotte *et al.*, 2009].




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

# Modèle de mélange

## Erreur de modélisation gaussienne

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

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

## Distribution conditionnelle

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

# Plan

Modèle probabiliste

Inférence

Expériences

Travaux futurs

# Inférence

- Paramètres du modèle : $\boldsymbol{\theta} = \left\{ \{\mathbf{W}_j, \mathbf{H}_j\}_j, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i \right\}$.
- Application semi-aveugle : Les filtres de mélange sont supposés connus.

## Distribution a posteriori

On s'intéresse à la distribution a posteriori des variables latentes :

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

- Cette distribution est gaussienne mais paramétrée par une matrice de covariance pleine et de grande dimension $\rightarrow$ **inférence approchée**.

# Inférence variationnelle

- On cherche $q \in \mathcal{F}$ comme approximation de $p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})$.
- La divergence KL est utilisée comme mesure de similarité :

$$KL\Big(q(\mathbf{s})\,||\,p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})\Big) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{Log-vraisemblance}} - \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{Energie variationnelle libre}}, \tag{1}$$

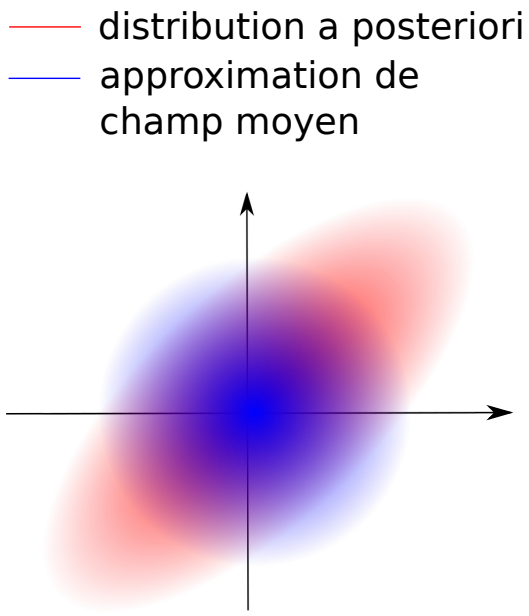
où $\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})}{q(\mathbf{s})}\right)\right\rangle_q$ et $\langle f(\mathbf{z})\rangle_q = \int f(\mathbf{z})q(\mathbf{z})d\mathbf{z}$.

- Algorithme espérance-maximisation variationnel :
  - **Etape E** : $q^\star = \arg\min_{q\in\mathcal{F}} KL\Big(q(\mathbf{s})\,||\,p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star)\Big) = \arg\max_{q\in\mathcal{F}} \mathcal{L}(q;\boldsymbol{\theta}^\star)$
  - **Etape M** : $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star;\boldsymbol{\theta})$

# Approximation de champ moyen

$\mathcal{F}$ est l'ensemble des densités de probabilité qui se factorisent sous la forme suivante :

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




Sous l'approximation de champ-moyen on peut montrer que :

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

# Etape M

Maximiser (ou augmenter) l'énergie variationnelle libre par rapport à $\boldsymbol{\theta}$.

## Paramètres NMF

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

$\rightarrow$ Règles de mise à jour multiplicatives.

## Variance du bruit

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

# Plan

Modèle probabiliste

Inférence

Expériences

Travaux futurs

# Expériences

- Base de données :
  - 8 mélanges stéréo créés à partir de réponses de salle simulées
  - Temps de réverbération : 256 ms
  - Nombre de sources : entre 3 et 5
  - Durée de chaque mélange : entre 12 et 28 secondes
- Scénario semi-aveugle :
  - Les filtres de mélange sont connus tandis que tous les autres paramètres sont estimés de façon aveugle.
- Mesures de performance :
  - Rapports signal sur distorsion (SDR), interférences (SIR) et artéfacts (SAR), exprimés en décibels (dB).

# Méthodes comparées

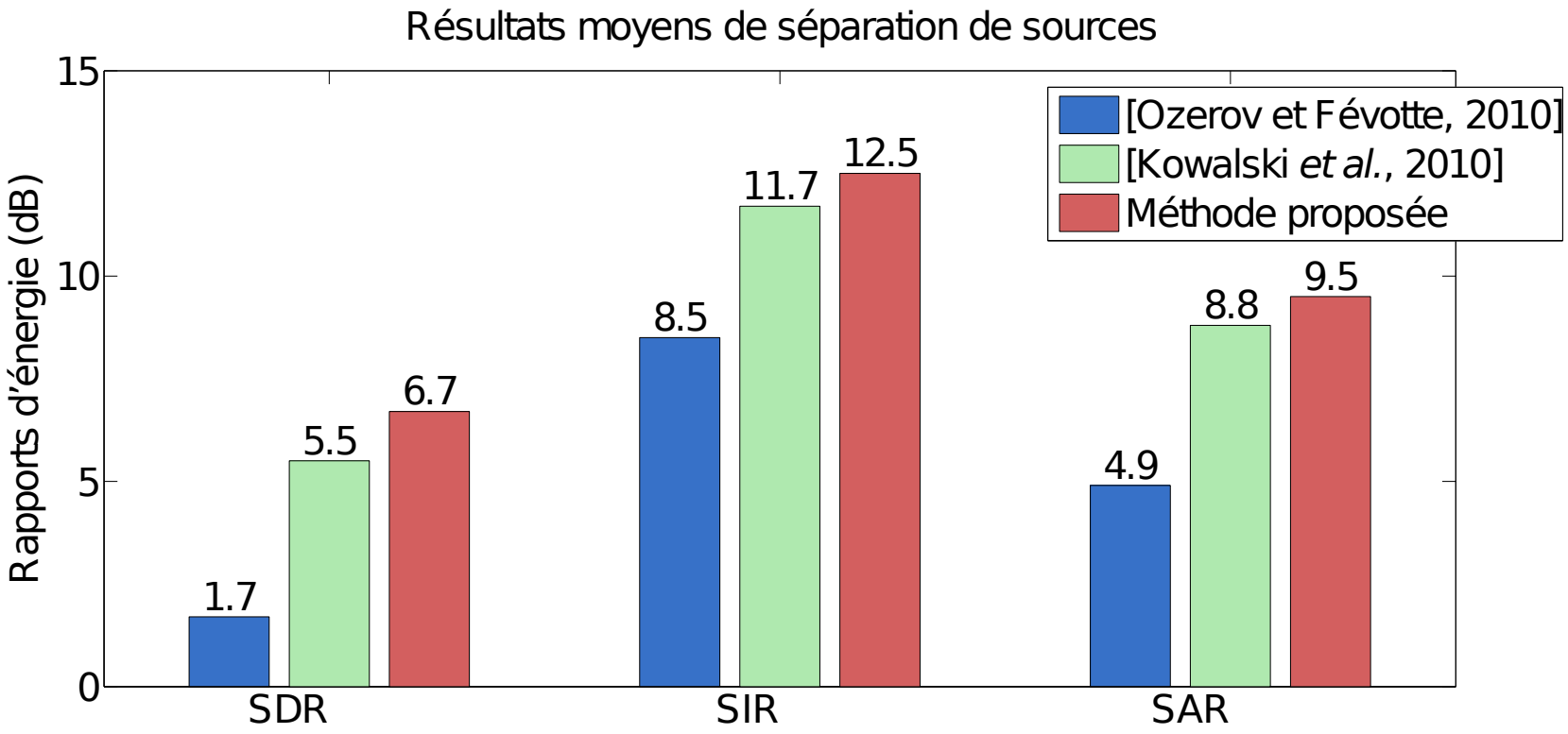
| | Modèle de source temps-fréquence | Représentation du mélange convolutif |
|---|---|---|
| [Kowalski *et al.*, 2010] | Parcimonieux (norme $\ell_1$) | Exacte (temps) |
| [Ozerov et Févotte, 2010] | Gaussien - NMF | Approchée (TFCT) |
| **Méthode proposée** | Gaussien - NMF | Exacte (temps) |

Longueur de la fenêtre d'analyse/synthèse TF : 128 ms.

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

# Résultats


Résultats moyens de séparation de sources
Rapports d'énergie (dB)
15
10
5
0
1.7
5.5
6.7
8.5
11.7
12.5
4.9
8.8
9.5
SDR
SIR
SAR
[Ozerov et Févotte, 2010]
[Kowalski et al., 2010]
Méthode proposée

# Exemple audio

Mélange stéréo:

| | Guitare 1 | Guitare 2 | Voix | Batterie | Basse |
|---|---|---|---|---|---|
| Source originale | | | | | |
| [Ozerov et Févotte, 2010] | | | | | |
| [Kowalski *et al.*, 2010] | | | | | |
| Méthode proposée | | | | | |

Extrait de «Ana» par Vieux Farka Toure. Base de données MTG MASS.

# Plan

Modèle probabiliste

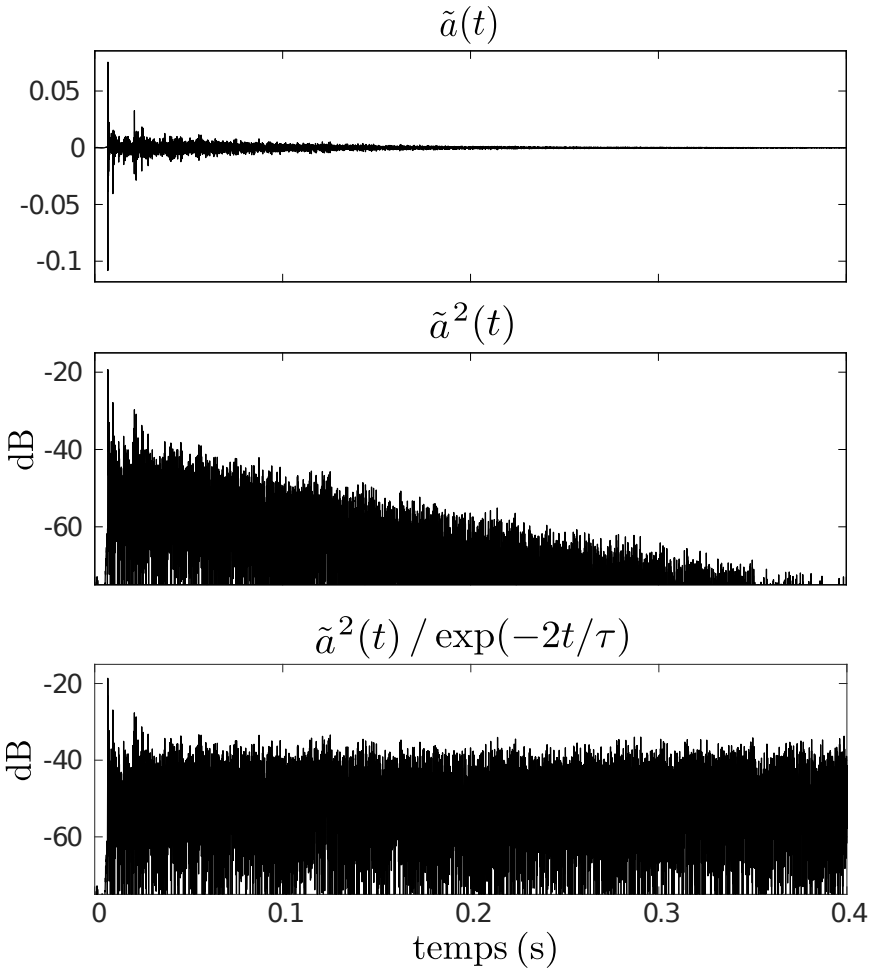
Inférence

Expériences

Travaux futurs

# Travaux futurs

- Résolution TF spécifique pour chaque source
- A priori sur les filtres de mélange dans le domaine temporel



# Merci

Code Matlab disponible à l'adresse suivante :

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