Student's t Source and Mixing Models for Multichannel Audio Source Separation

Simon Leglaive

Inria Grenoble Rhône-Alpes, Perception Team

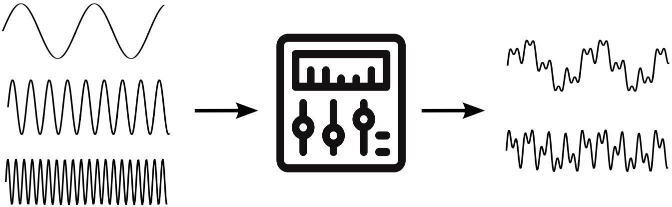
Bayes in Grenoble Seminar
May 15, 2018

This work was carried out during my Ph.D. at Télécom ParisTech, with my Ph.D. advisors Roland Badeau and Gaël Richard.

Introduction

Source separation

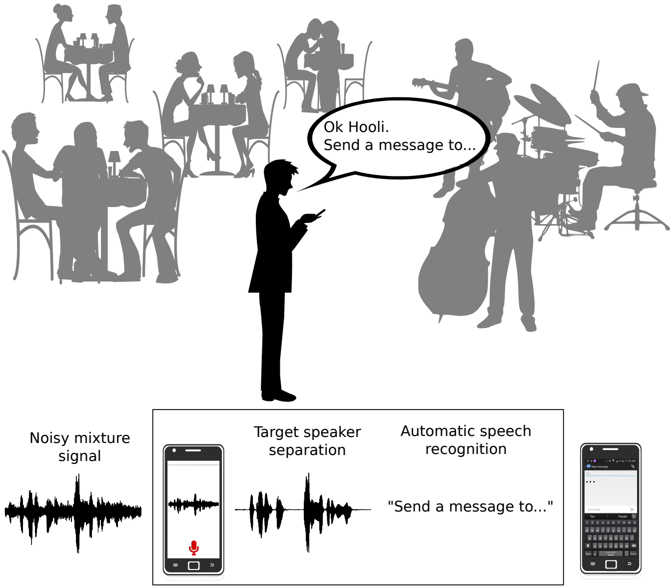
Objective: Recover source signals from one or several mixtures.



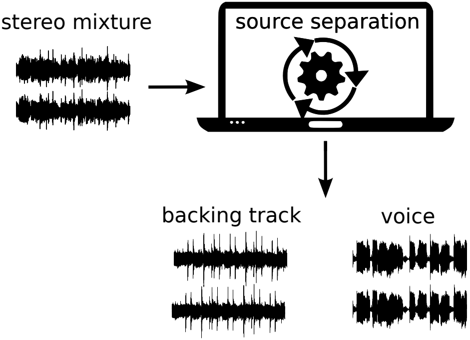
Many applications:

- Biomedical signal processing (ECG, EEG, MEG, MRI, etc.);
- Astrophysics;
- Underwater acoustics;
- **Audio signal processing**;
- etc.

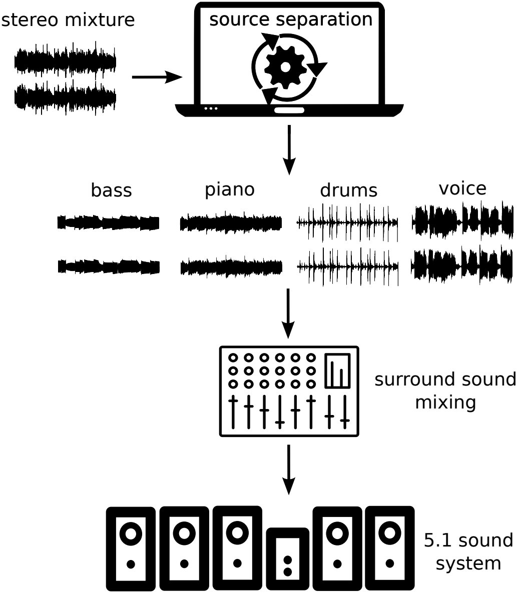
Audio source separation in everyday life



Audio source separation for karaoke



Audio source separation for music upmixing



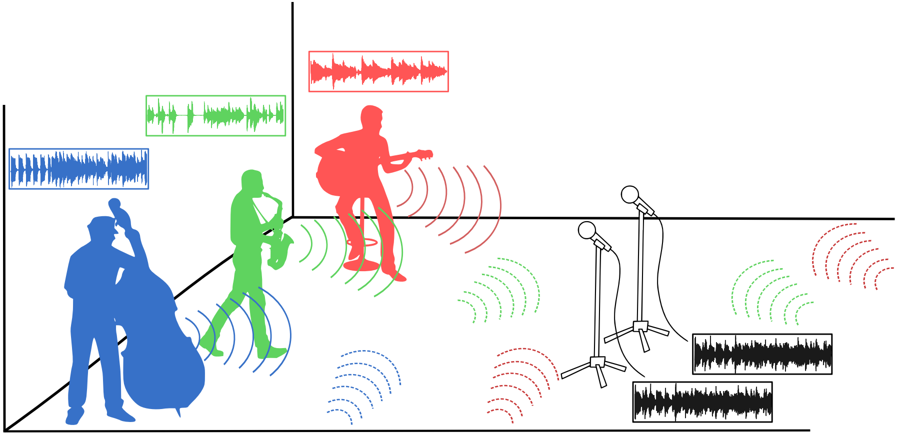
Use case of the ANR project

3DISON

EDISON 3D

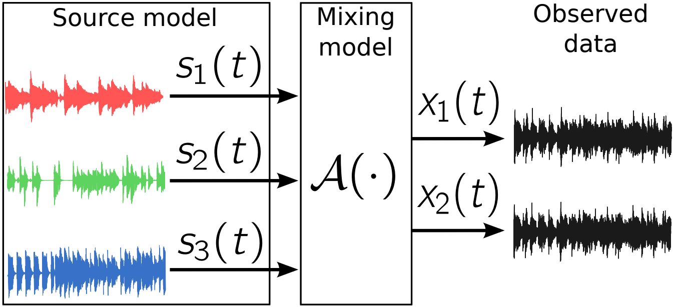
Targeted scenario

Under-determined and reverberant multichannel mixture.



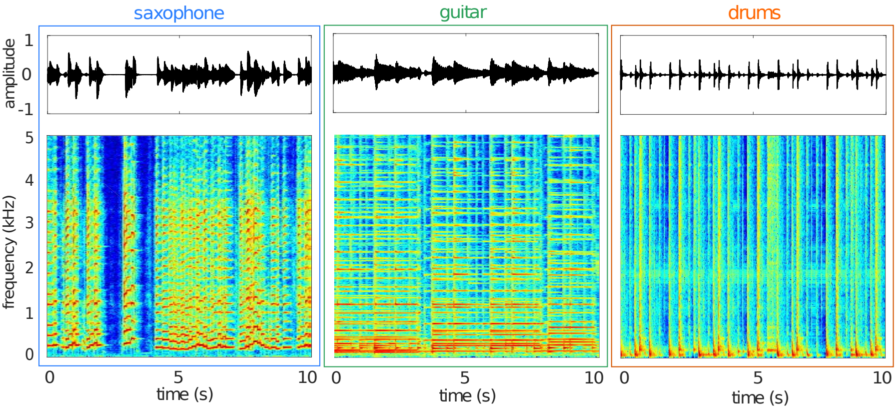
Model-based approach

A model aims to explain how are the observed data generated.



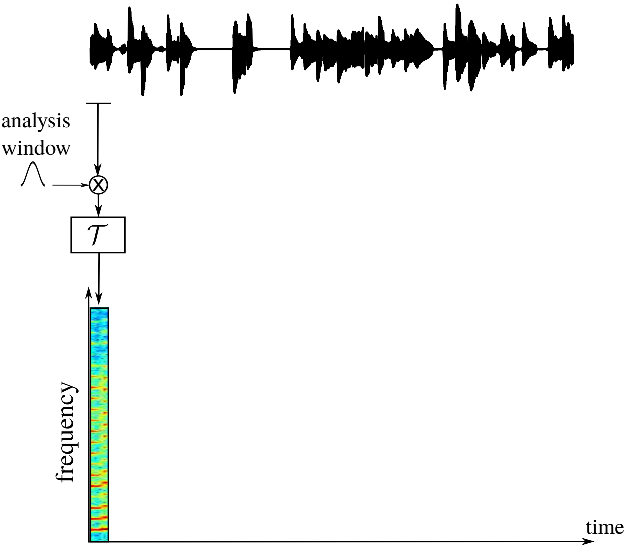
Time-frequency source representation

Time-frequency (TF) transforms provide meaningful representations.



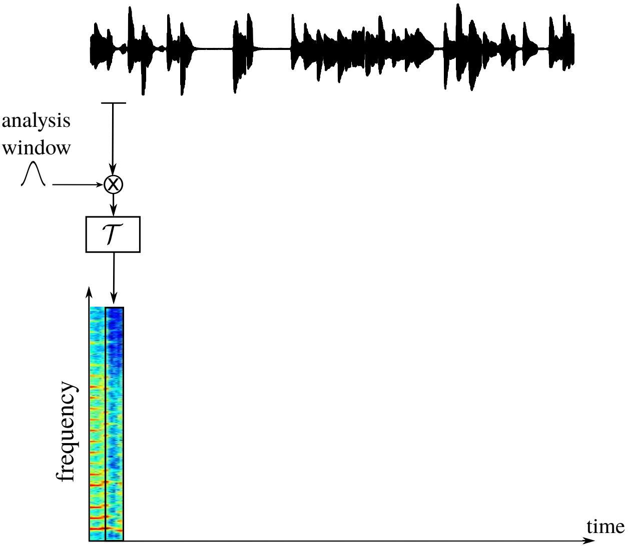
Spectrograms computed from the short-term Fourier transform (STFT).

Time-frequency transform



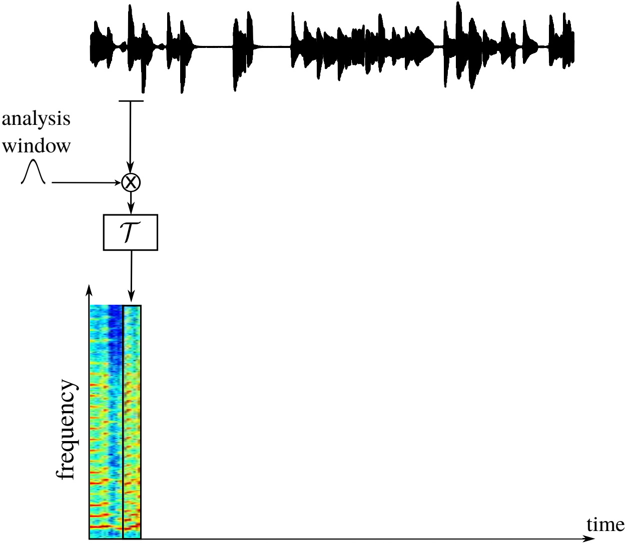
$\mathcal{T}$: discrete Fourier transform, discrete cosine transform, etc.

Time-frequency transform



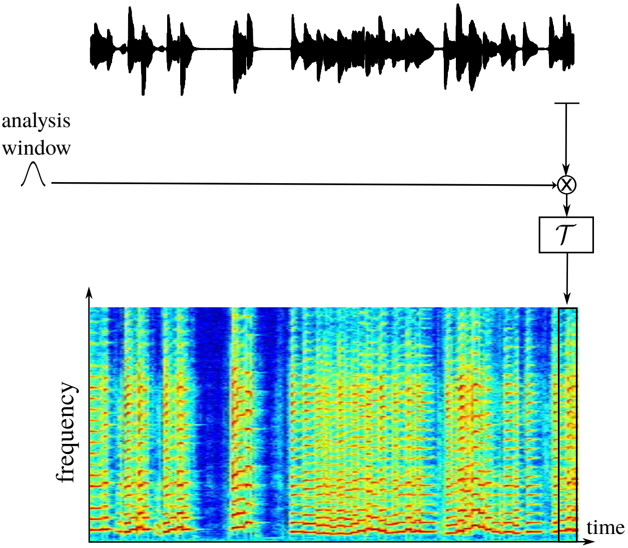
$\mathcal{T}$: discrete Fourier transform, discrete cosine transform, etc.

Time-frequency transform



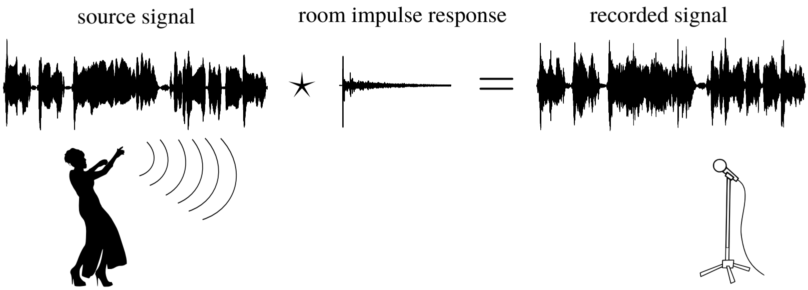
$\mathcal{T}$: discrete Fourier transform, discrete cosine transform, etc.

Time-frequency transform



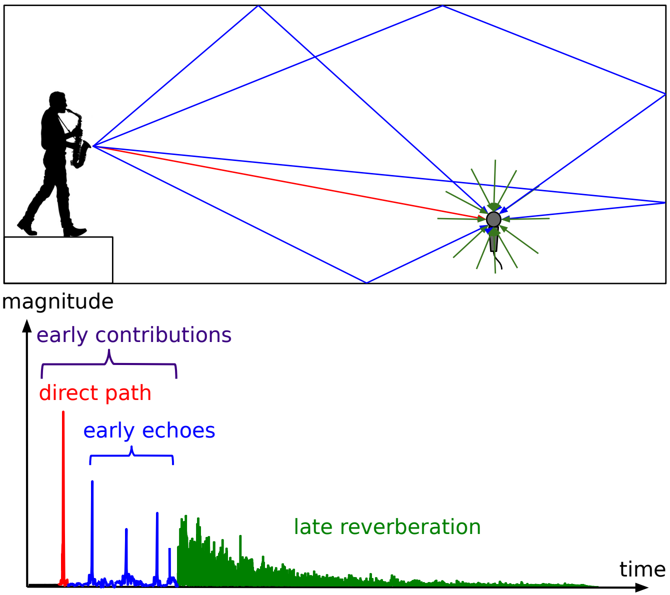
$\mathcal{T}$: discrete Fourier transform, discrete cosine transform, etc.

Room impulse response (RIR) (1)



- Characterizes the source-to-microphone acoustic path.
- Reverberation time:
  - between 0.1 and 0.8 s for domestic/office rooms
  - up to a few seconds for concert halls
  - 75 s for a Scottish oil storage tank (world record!)

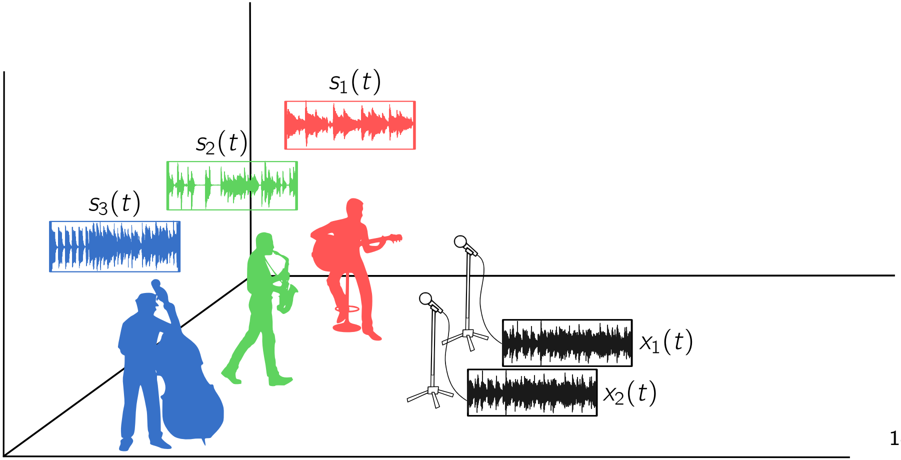
Room impulse response (RIR) (2)



Reverberant mixtures (1)

Convolutive mixing process in the time domain:

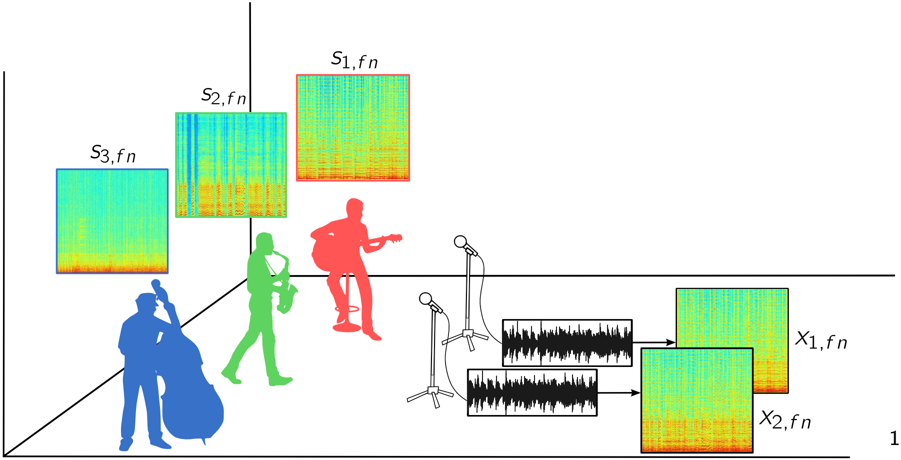
$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$$



Reverberant mixtures (2)

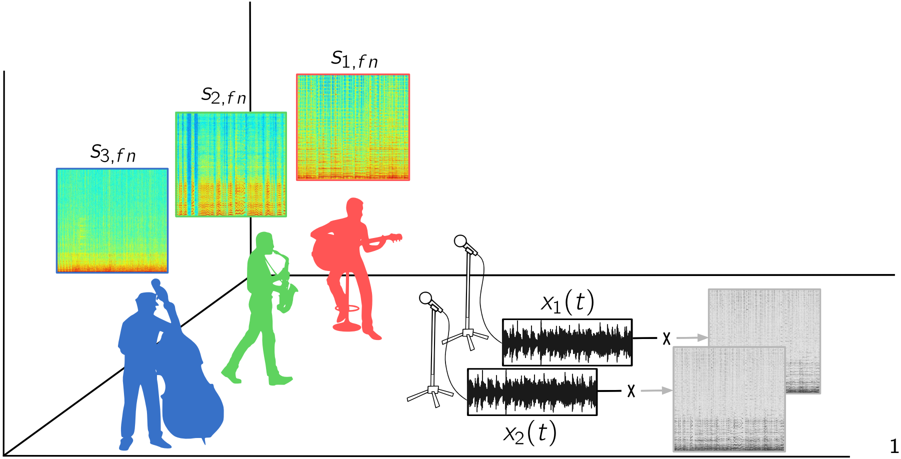
Convolutive mixing process in the STFT domain:

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



Proposed approach (1)

Time-domain mixture representation and time-frequency source representation.



Proposed approach (2)

Time-domain convolutive mixture model

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

Time-frequency source representation

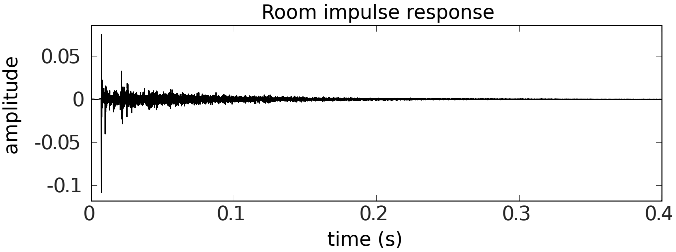
$$s_j(t) = \sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\psi_{j,fn}(t), \tag{2}$$

with $\psi_{j,fn}(t) \in \mathbb{R}$ a (source-dependent) modified discrete cosine transform (MDCT) atom and $\mathcal{B}_j = \{0, ..., F_j - 1\} \times \{0, ..., N_j - 1\}$.

Remark: Source time-frequency coefficients are real-valued.

Outline

1. **Deterministic** time-domain mixing filters
2. **Probabilistic** time-domain mixing filters



Deterministic time-domain mixing filters

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

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

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

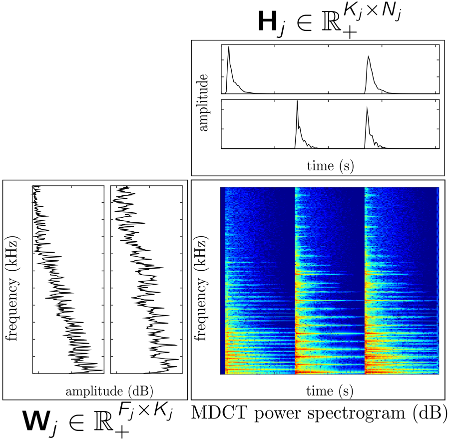
- What prior knowledge do we have on the latent source variables?
- How are the data generated from the latent unobserved variables?

Source model

Gaussian source model based on non-negative matrix factorization (NMF) [1].

Independently for all sources and TF points:

$$s_{j,fn} \sim \mathcal{N}\left(0, (\mathbf{W}_j \mathbf{H}_j)_{f,n}\right).$$



[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 mixture distribution

Independently for all microphones and time instants:

$$x_i(t) \mid \mathbf{s} \sim \sum_{j=1}^{J} [a_{ij} \star s_j](t) + \mathcal{N}\left(0, \sigma_i^2\right),$$

where we recall that $s_j(t) = \sum_{(f,n) \in \mathcal{B}_j} s_{j,fn} \psi_{j,fn}(t)$.

Statistical inference

Posterior inference

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

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

with $\boldsymbol{\theta}^\star$ an estimate of $\boldsymbol{\theta} = \big\{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i\big\}$.

Maximum likelihood parameters estimation

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

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

Variational inference

- Find $q(\mathbf{s}) \in \mathcal{F}$ which approximates $p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})$.

Variational inference

- Find $q(\mathbf{s}) \in \mathcal{F}$ which approximates $p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})$.

- We take the Kullback-Leibler divergence as a measure of fit:

$$D_{KL}(q(\mathbf{s})||p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-likelihood}} - \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{variational free energy}}, \tag{3}$$

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 inference

- Find $q(\mathbf{s}) \in \mathcal{F}$ which approximates $p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})$.
- We take the Kullback-Leibler divergence as a measure of fit:

$$D_{KL}(q(\mathbf{s})||p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-likelihood}} - \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{variational free energy}}, \tag{3}$$

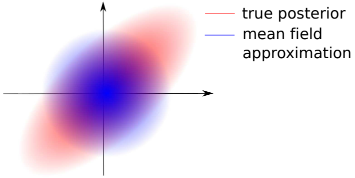
where $\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\theta})}{q(\mathbf{s})}\right)\right\rangle_q$ and $\langle f(\mathbf{s})\rangle_q = \int f(\mathbf{s})q(\mathbf{s})d\mathbf{s}$.

- Variational expectation-maximization algorithm:
  - **E-step**: $q^\star = \arg\min_{q\in\mathcal{F}} D_{KL}(q(\mathbf{s})||p(\mathbf{s}|\mathbf{x};\boldsymbol{\theta}_{\text{old}})) = \arg\max_{q\in\mathcal{F}} \mathcal{L}(q;\boldsymbol{\theta}_{\text{old}})$;
  - **M-step**: $\boldsymbol{\theta}_{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star;\boldsymbol{\theta})$.

Mean-field approximation

$\mathcal{F}$ is the set of probability density functions that factorize as:

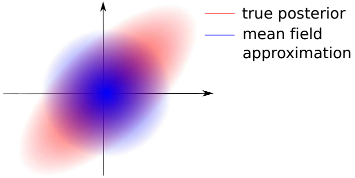
$$q(\mathbf{s}) = \prod_{j=1}^{J} \prod_{(f,n)\in\mathcal{B}_j} q_{jfn}(s_{j,fn}).$$



Mean-field approximation

$\mathcal{F}$ is the set of probability density functions that factorize as:

$$q(\mathbf{s}) = \prod_{j=1}^{J} \prod_{(f,n)\in\mathcal{B}_j} q_{jfn}(s_{j,fn}).$$



Under the mean-field approximation we can show that:

$$q^{\star}_{jfn}(s_{j,fn}) = N(\hat{s}_{j,fn}, \gamma_{j,fn}).$$

E-Step: update the variational parameters.

Source estimate: approximate posterior mean $\hat{s}_{j,fn}$.

M-Step

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

NMF parameters

$$\min_{\mathbf{W}_j,\mathbf{H}_j \geq 0} \sum_{(f,n)\in\mathcal{B}_j} d_{IS}\left(\left\langle s_{j,fn}^2 \right\rangle_{q^\star}, (\mathbf{W}_j\mathbf{H}_j)_{f,n}\right),$$

where $d_{IS}(\cdot,\cdot)$ is the Itakura-Saito divergence.

$\rightarrow$ standard multiplicative update rules [1].

M-Step

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

NMF parameters

$$\min_{\mathbf{W}_j, \mathbf{H}_j \geq 0} \sum_{(f,n) \in \mathcal{B}_j} d_{IS} \left(\left\langle s_{j,fn}^2 \right\rangle_{q^\star}, (\mathbf{W}_j \mathbf{H}_j)_{f,n} \right),$$

where $d_{IS}(\cdot, \cdot)$ is the Itakura-Saito divergence.

$\rightarrow$ standard multiplicative update rules [1].

Mixing filters

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

M-Step

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

NMF parameters

$$\min_{\mathbf{W}_j,\mathbf{H}_j \geq 0} \sum_{(f,n)\in\mathcal{B}_j} d_{IS}\left(\left\langle s_{j,fn}^2 \right\rangle_{q^\star}, (\mathbf{W}_j\mathbf{H}_j)_{f,n}\right),$$

where $d_{IS}(\cdot,\cdot)$ is the Itakura-Saito divergence.

$\rightarrow$ standard multiplicative update rules [1].

Mixing filters

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

Noise variance

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

Semi-oracle audio source separation example

Semi-oracle setting: mixing filters are known and fixed.

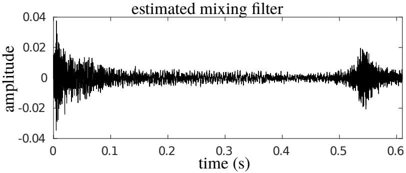
Musical excerpt from “Ana” by Vieux Farka Toure:

	Voice	Guitar 1	Guitar 2	Drums	Bass
Original source (stereo)					
Estimated source (stereo)					

Estimating the mixing filters - an ill posed problem

- Observations: $x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$.
- Estimating both the source signals and mixing filters is an ill-posed problem.
- The estimated mixing filter contains some part of the voice signal.

True: 🔊



Estimated: 🔊

drums separation example		
stereo mixture	original source	estimated source
🔊	🔊	🔊

Probabilistic time-domain mixing filters

Proposed approach

Time-domain convolutive mixture model

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

Time-frequency source representation

$$s_j(t) = \sum_{(f,n)\in\mathcal{B}_j} s_{j,fn}\psi_{j,fn}(t).$$

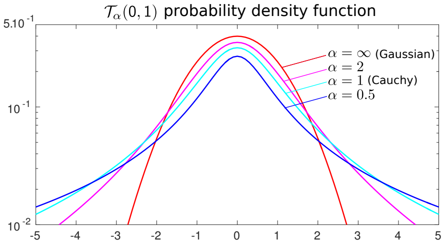
Latent random variables

- Time-frequency source coefficients $\{s_{j,fn}\}_{j,f,n}$;
- Time-domain mixing filters $\{a_{ij}(t)\}_{i,j,t}$.

Student's t distribution

Student's t distribution: $\mathcal{T}_\alpha(\mu, \sigma)$

- Shape: $\alpha > 0$;
- Location: $\mu \in \mathbb{R}$;
- Scale: $\sigma > 0$.



Scale mixture of Gaussians

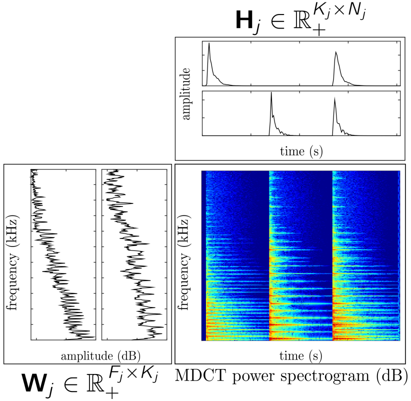
$$z \sim \mathcal{T}_\alpha(\mu, \sigma) \quad \Leftrightarrow \begin{cases} z|v & \sim \mathcal{N}\left(\mu, v\sigma^2\right) \\ v & \sim \mathcal{IG}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right) \end{cases}$$

Source model

Student's t source model based on NMF [2].

Independently for all sources and TF points:

$$s_{j,fn} \sim \mathcal{T}_{\alpha_v}\left(0, (\mathbf{W}_j \mathbf{H}_j)^{\frac{1}{2}}_{f,n}\right).$$

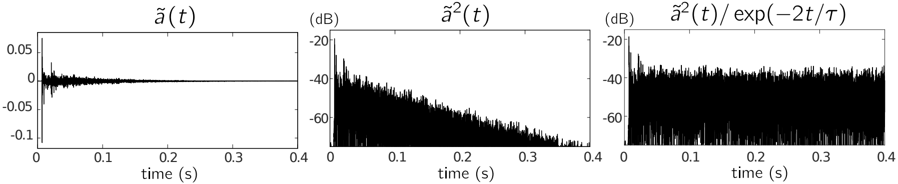


Remark: Generalization of the previous Gaussian model.

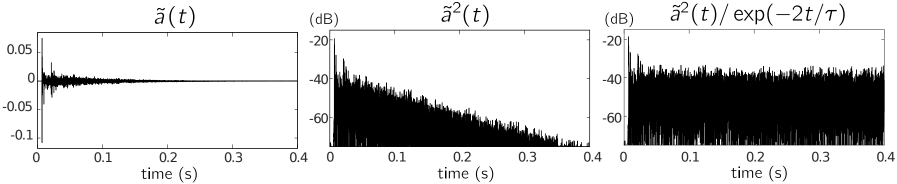
---

[2] K. Yoshii, K. Itoyama and M. Goto, "Student's t nonnegative matrix factorization and positive semidefinite tensor factorization for single-channel audio source separation", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

Gaussian RIR model (1)



Gaussian RIR model (1)



Gaussian model with exponential decay [3]

Independently for all microphones, sources and time instants:

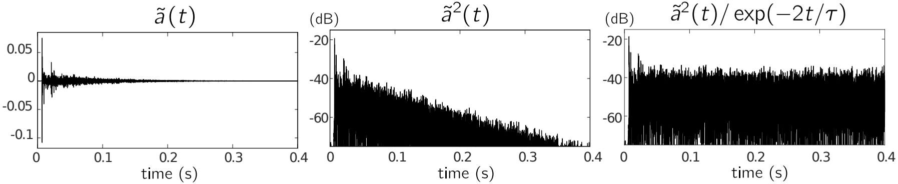
$$a_{ij}(t) \sim \exp(-t/\tau)\,\mathcal{N}(0, \sigma_r^2),$$

where τ is defined according to the reverberation time.

Theoretically valid only for late reverberation (diffuse sound field).

[3] J. D. Polack, "La transmission de l'énergie sonore dans les salles", Ph.D. dissertation, Université du Maine, 1988.

Gaussian RIR model (2)



Gaussian model with exponential decay [3]

Equivalently:

$$a_{ij}(t)/\exp(-t/\tau) \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_r^2),$$

where τ is defined according to the reverberation time.

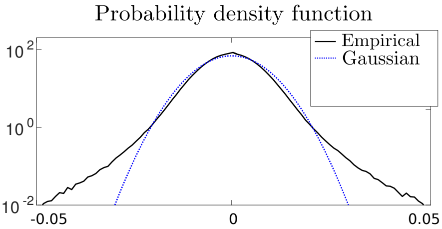
Theoretically valid only for late reverberation (diffuse sound field).

---

[3] J. D. Polack, "La transmission de l'énergie sonore dans les salles", Ph.D. dissertation, Université du Maine, 1988.

Student's t RIR model

Distribution of the normalized RIR coefficients

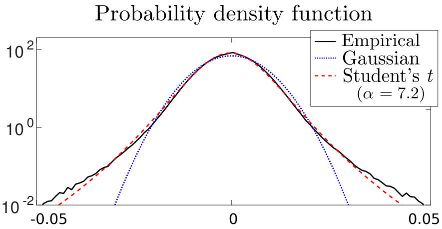


- 624 RIRs from the MIRD database [4];
- Reverberation time equals 610 ms.

[4] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments", *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

Student's t RIR model

Distribution of the normalized RIR coefficients

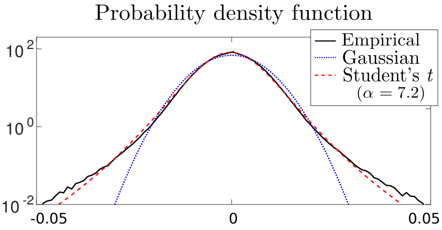


- 624 RIRs from the MIRD database [4];
- Reverberation time equals 610 ms.

[4] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments", *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

Student's t RIR model

Distribution of the normalized RIR coefficients



- 624 RIRs from the MIRD database [4];
- Reverberation time equals 610 ms.

Student's t model with exponential decay

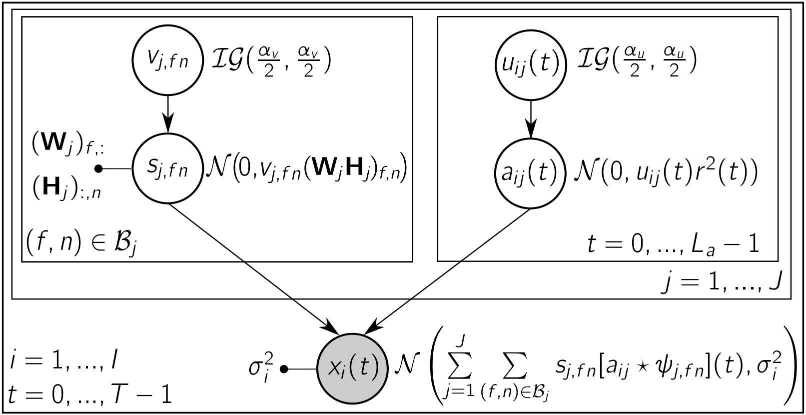
Independently for all microphones, sources and time instants:

$$a_{ij}(t) \sim \mathcal{T}_{\alpha_u}(0, r(t)), \qquad r^2(t) = \sigma_r^2 \exp(-2t/\tau).$$

---

[4] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments", *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

Bayesian network



- $\mathbf{z}$: set of all latent variables (empty circles);
- $\mathbf{x}$: set of observations (shaded circles);
- $\boldsymbol{\theta}$: set of model parameters to be estimated (dots).

Variational inference

- Exact posterior inference is analytically intractable.
- Variational inference with mean-field approximation:

$$q(\mathbf{z}) = \prod_{z_k \in \mathbf{z}} q_k(z_k).$$

- Under this approximation we can show that:

$$\begin{aligned}
q^{s}_{jfn}(s_{j,fn})^{\star} &= N(\hat{s}_{j,fn}, \gamma_{j,fn}); \\
q^{a}_{ijt}(a_{ij}(t))^{\star} &= N\left(\hat{a}_{ij}(t), \rho_{ij}(t)\right); \\
q^{v}_{jfn}(v_{j,fn})^{\star} &= IG(\nu_v, \beta_{j,fn}); \\
q^{u}_{ijt}(u_{ij}(t))^{\star} &= IG(\nu_u, d_{ij}(t)).
\end{aligned}$$

- **E-Step** $\rightarrow$ update all the variational parameters.

M-Step

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

NMF parameters

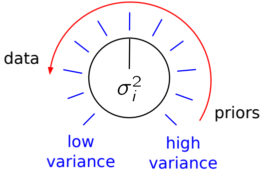
$$\min_{\mathbf{W}_j,\mathbf{H}_j \geq 0} \sum_{(f,n)\in\mathcal{B}_j} d_{IS}\left(\left\langle v_{j,fn}^{-1} \right\rangle_{q^\star} \left\langle s_{j,fn}^2 \right\rangle_{q^\star}, (\mathbf{W}_j\mathbf{H}_j)_{f,n} \right),$$

where $d_{IS}(\cdot,\cdot)$ is the Itakura-Saito divergence.

$\rightarrow$ multiplicative update rules [1].

Noise variance

σ_i^2 is manually decreased along the iterations.



Experimental setup

- Dataset:
  - 8 stereo mixtures created with RIRs from the MIRD database [4];
  - Reverberation times: 160, 360 and 610 ms;
  - Number of sources per mixture: 3 to 5;
  - Musical sources: drums, piano, bass, guitar, voice;
  - Duration: 12 to 28 seconds.
- Semi-blind scenario:
  - NMF dictionaries $\mathbf{W}_j$ are pre-trained using the true source signals;
  - Reverberation time is assumed to be known;
  - All other parameters are blindly estimated.

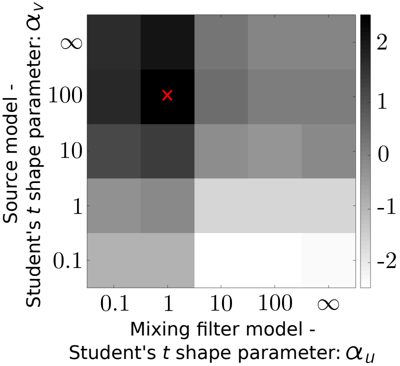
Model hyperparameters

How do we choose α_v, α_u and the MDCT window length?

Model hyperparameters

How do we choose α_v, α_u and the MDCT window length?

Average performance using the mixtures with a reverberation time of 360 ms

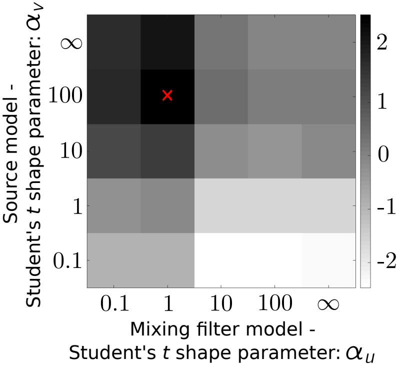


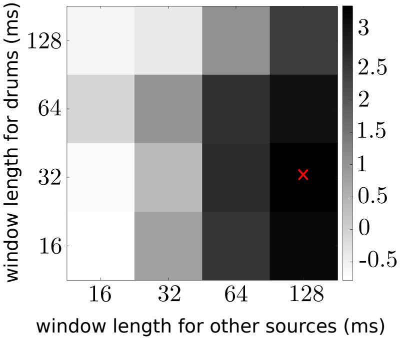
Remark: MDCT window length is fixed to 64 ms.

Model hyperparameters

How do we choose α_v, α_u and the MDCT window length?

Average performance using the mixtures with a reverberation time of 360 ms





Remark: MDCT window length is fixed to 64 ms.

Remark: (α_v, α_u) is fixed to $(100, 1)$.

Reference methods

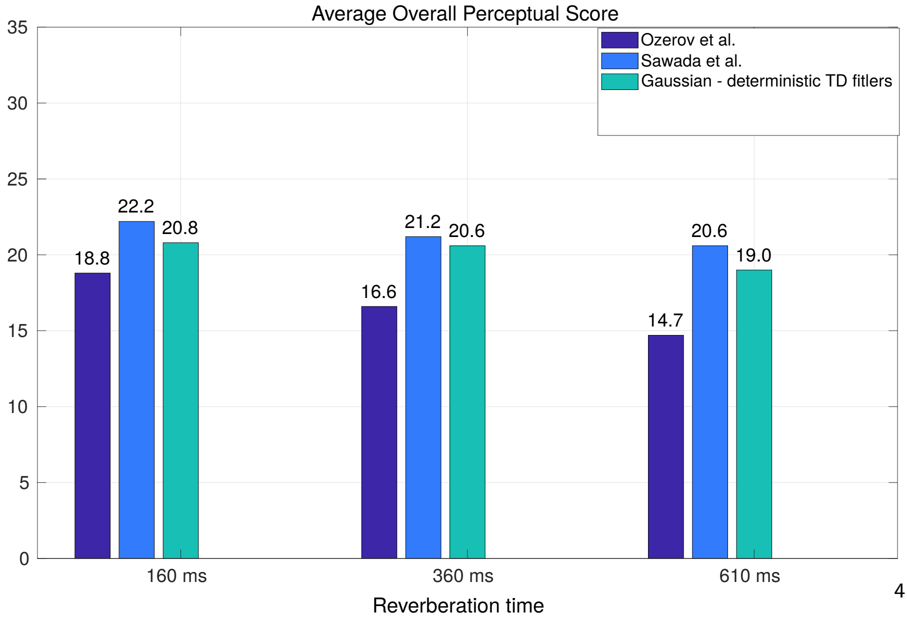
- Ozerov *et al.* [5] – Sawada *et al.* [6]:
  - **Similar models**:
    - Source model: STFT - Gaussian - NMF.
    - Convolutive mixture model: STFT - approximate.
  - **Different estimation algorithms**.
- Our previous method with deterministic and unconstrained time-domain mixing filters.

---

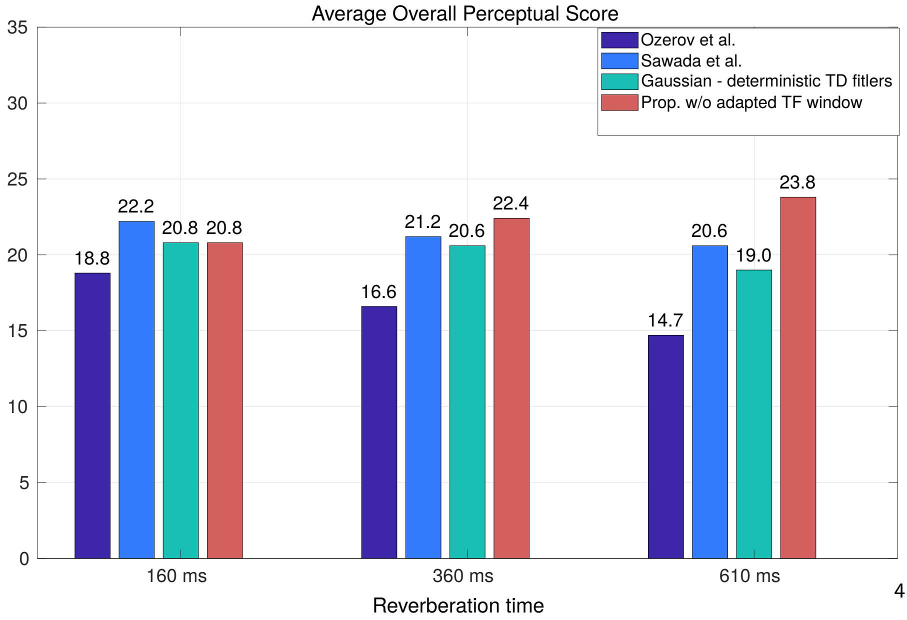
[5] A. Ozerov, E. Vincent and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation", *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[6] H. Sawada, H. Kameoka, S. Araki and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data". *IEEE Transactions on Audio, Speech, and Language Processing*, 2013.

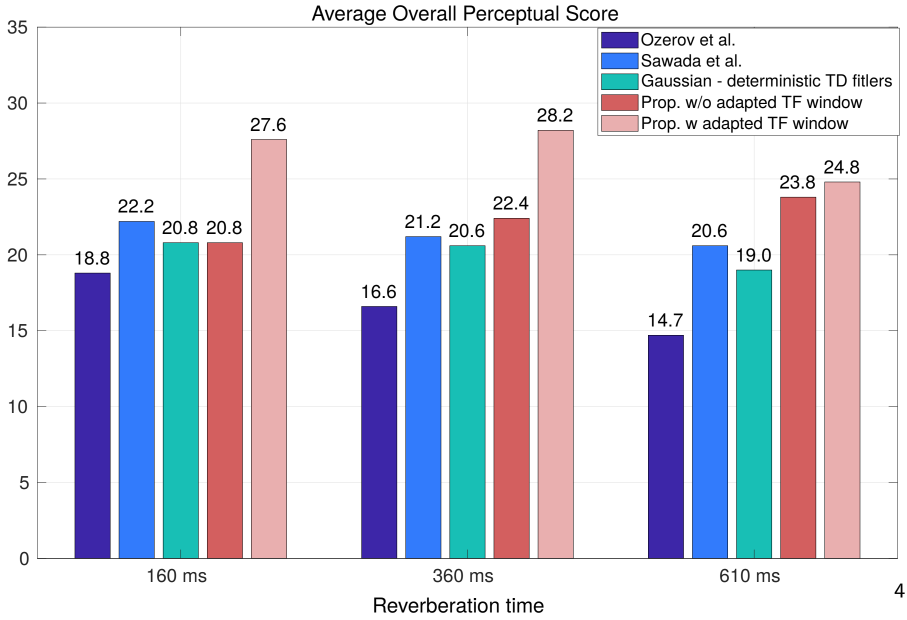
Source separation results



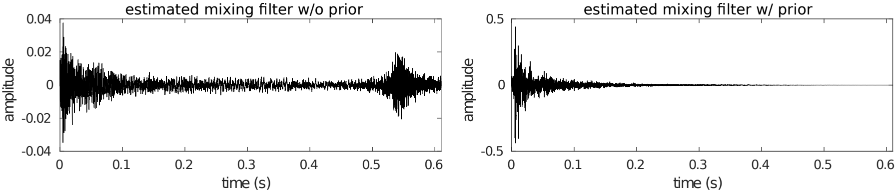
Source separation results



Source separation results



Mixing filters: Influence of the prior



drums separation example			
stereo mixture	original source	w/o prior	w/ prior
🔊	🔊	🔊	🔊

Audio example (1)

- Separation of 3 sources from a stereo mixture ($T_{60} = 610$ ms).
- All algorithms are run using **oracle NMF dictionaries**.

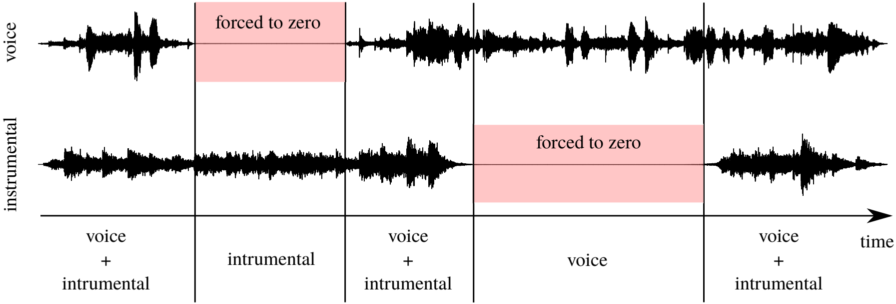
Stereo mixture: 🔊

source image	drums	guitar	bass
original	🔊	🔊	🔊
Ozerov et al.	🔊	🔊	🔊
Sawada et al.	🔊	🔊	🔊
Gaussian - deterministic TD filters	🔊	🔊	🔊
Prop. w/o adapted TF window	🔊	🔊	🔊

Song: "TV On" by Kismet. MTG MASS database.

Audio example (2)

- Stereo mixture provided by Radio France (Edison 3D ANR project).
- **Blind** separation of voice and instrumental.



Song: "C'est magnifique" by Ella Fitzgerald (Nice Jazz Festival 1972 - Recording: ORTF).

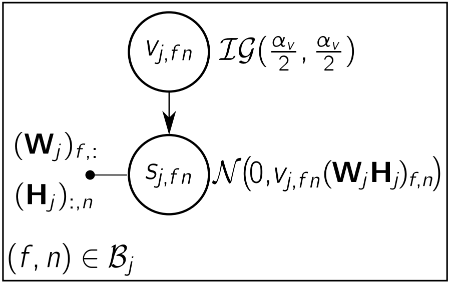
Conclusion

Conclusion

Multichannel audio source separation with time-domain convolutive mixture model:

- Appropriate for highly reverberant mixtures;
- Necessary to have priors on the mixing filters;
- Multi-resolution source modeling.

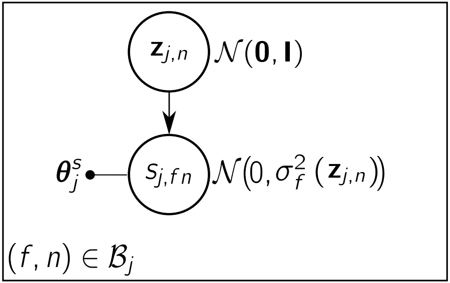
Perspectives: supervised source model



- Learn NMF dictionaries on an external dataset.
- What about neural networks?

Perspectives: supervised source model

Variational autoencoder as a generative source model.



$\sigma_f^2(\cdot)$: non-linear function parametrized by $\boldsymbol{\theta}_j^s \rightarrow$ neural network.

- Learning $\boldsymbol{\theta}_j^s$ is "easy" in the framework of variational autoencoders.
- The difficulty lies in the inference of $\{\mathbf{z}_{j,n}\}_n$ when the source signal is not directly observed $\rightarrow$ Markov chain Monte Carlo methods.

Thank you

Audio examples and Matlab code available online:

`https://sleglaive.github.io`