Modeling Reverberant Mixtures for Multichannel Audio Source Separation

Simon Leglaive

Ph.D. supervisors: Roland Badeau and Gaël Richard

LTCI, Télécom ParisTech, Université Paris Saclay

Ph.D. defense, Télécom ParisTech, Paris

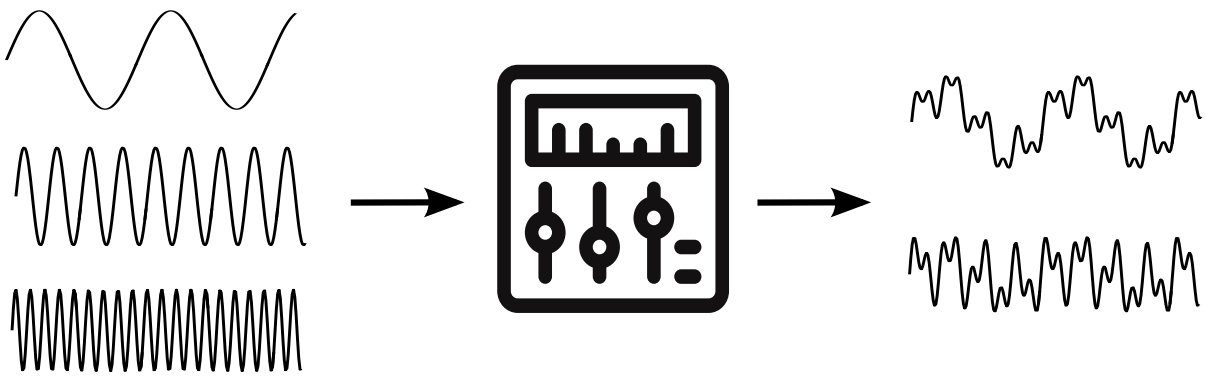
December 12, 2017

Outline

Outline

Source separation

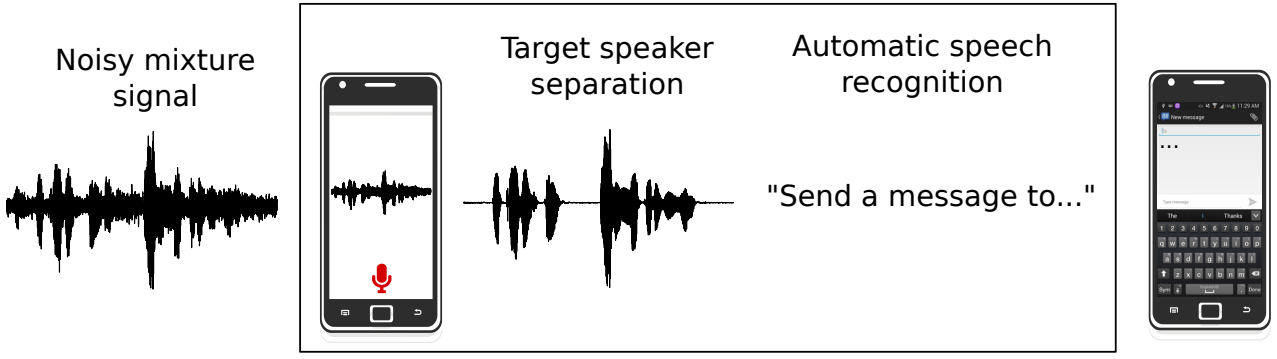
Objective: Recover source signals from one or several mixtures.



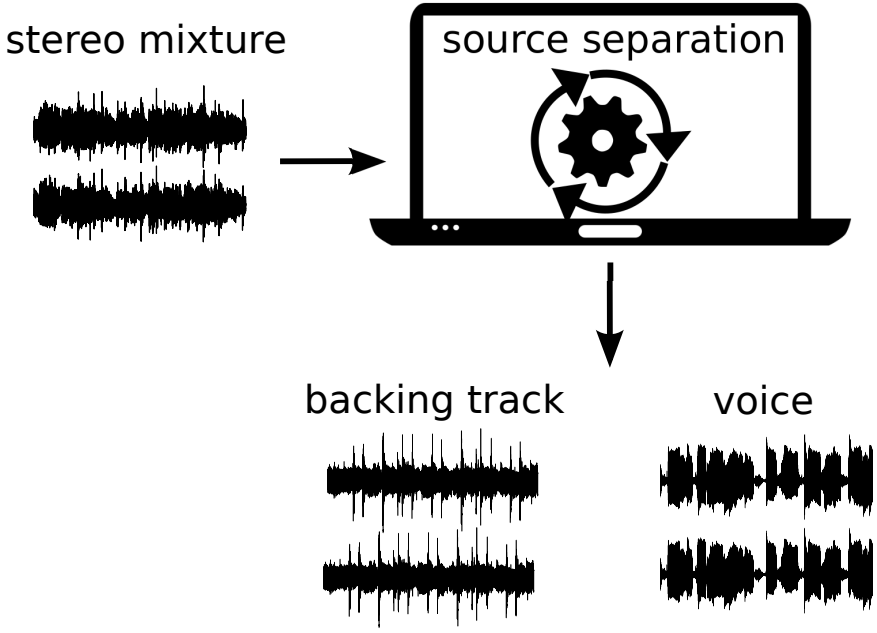
Many applications:

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

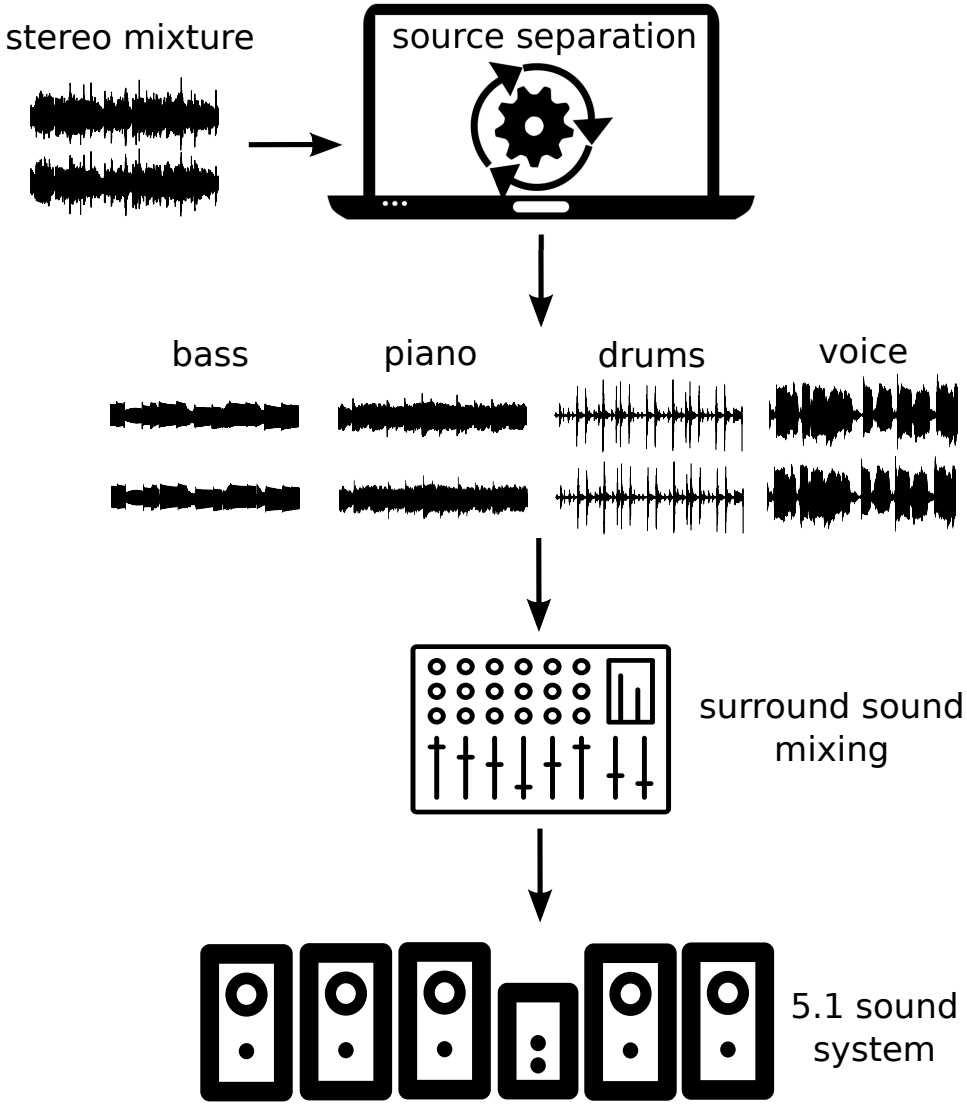
Audio source separation in everyday life



Audio source separation for karaoke



Audio source separation for music upmixing

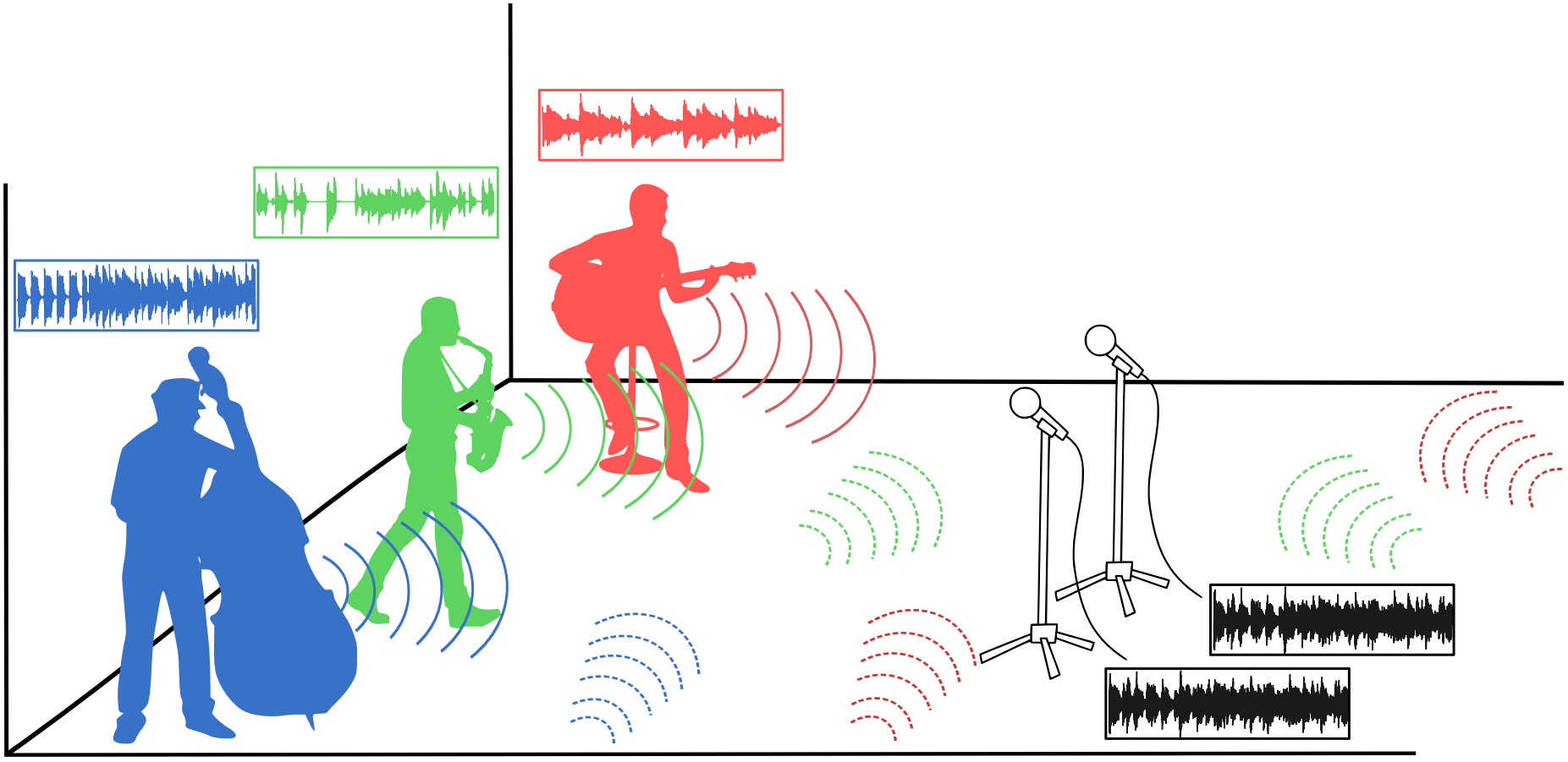


Use case of the ANR project

3DISON
EDISON 3D

Targeted scenario

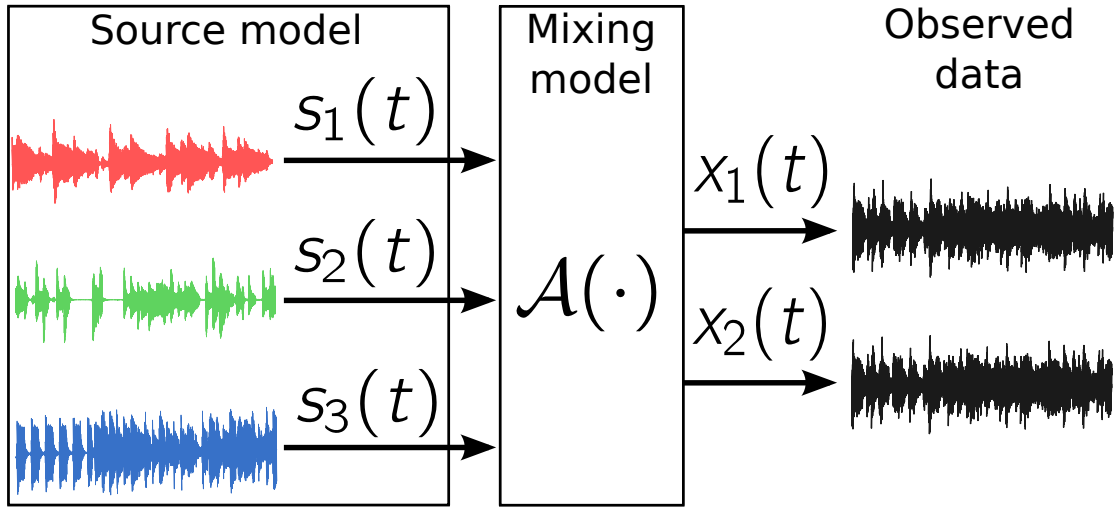
Under-determined and reverberant multichannel mixture.



Outline

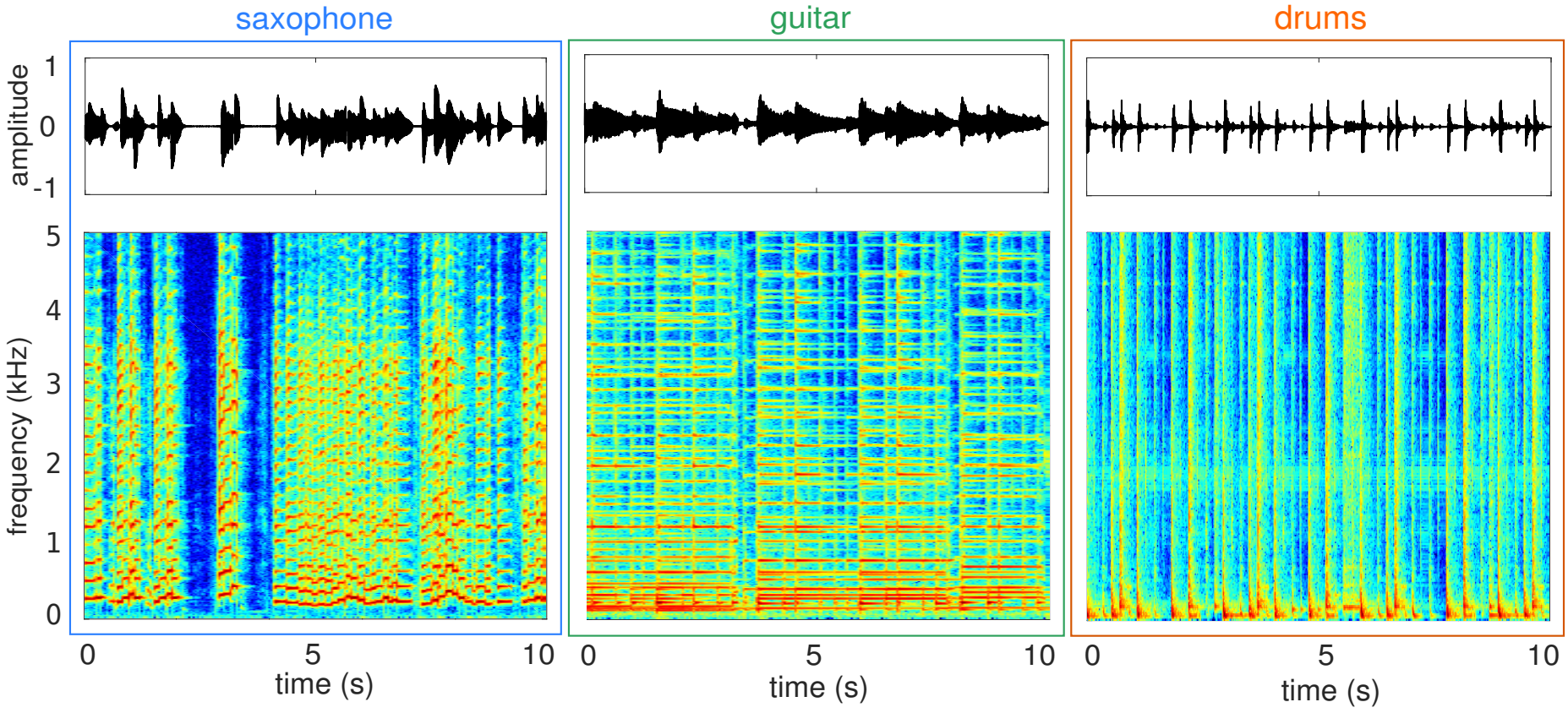
Model-based approach

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



Time-frequency source representation

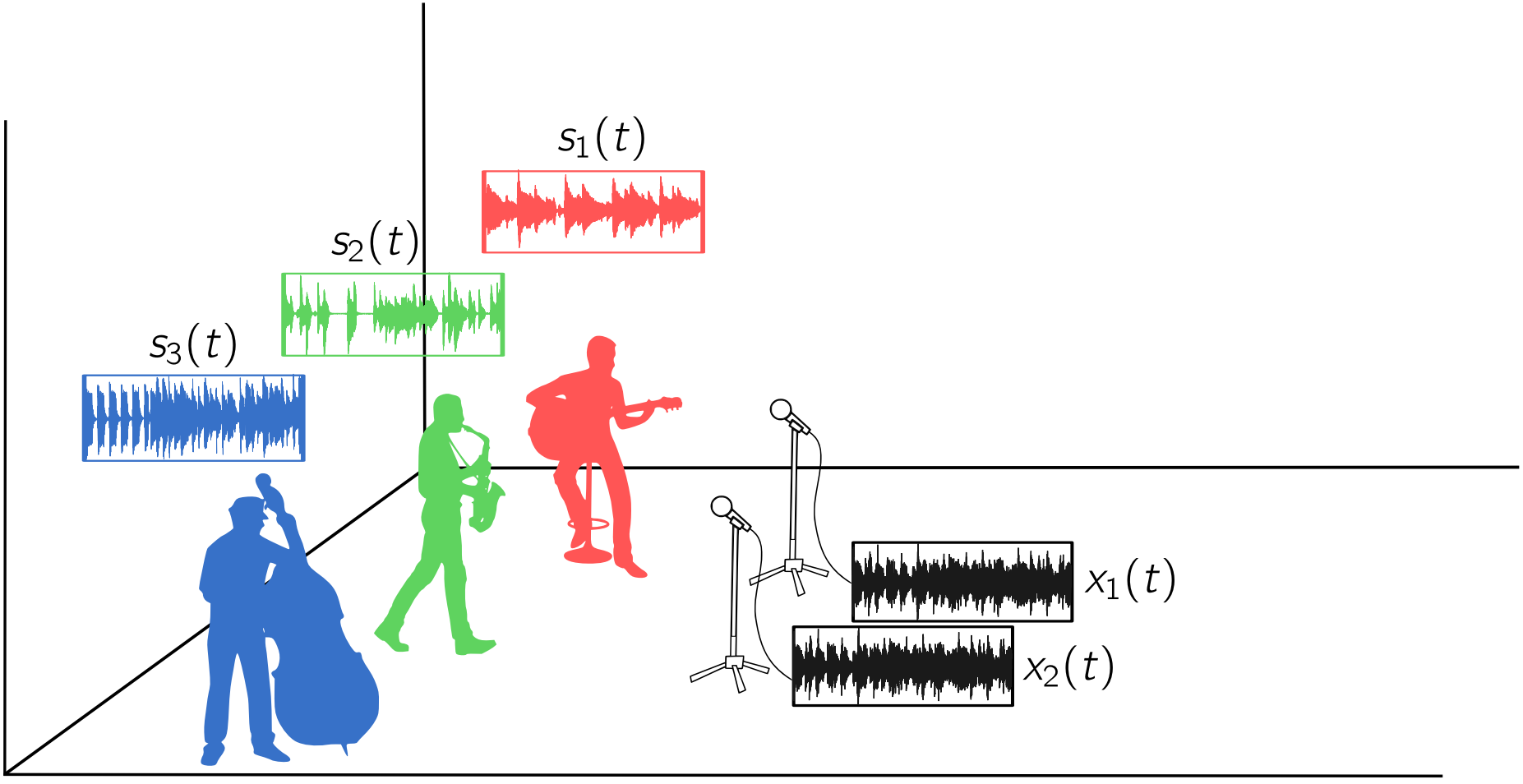
Time-frequency (TF) transforms provide meaningful representations.



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

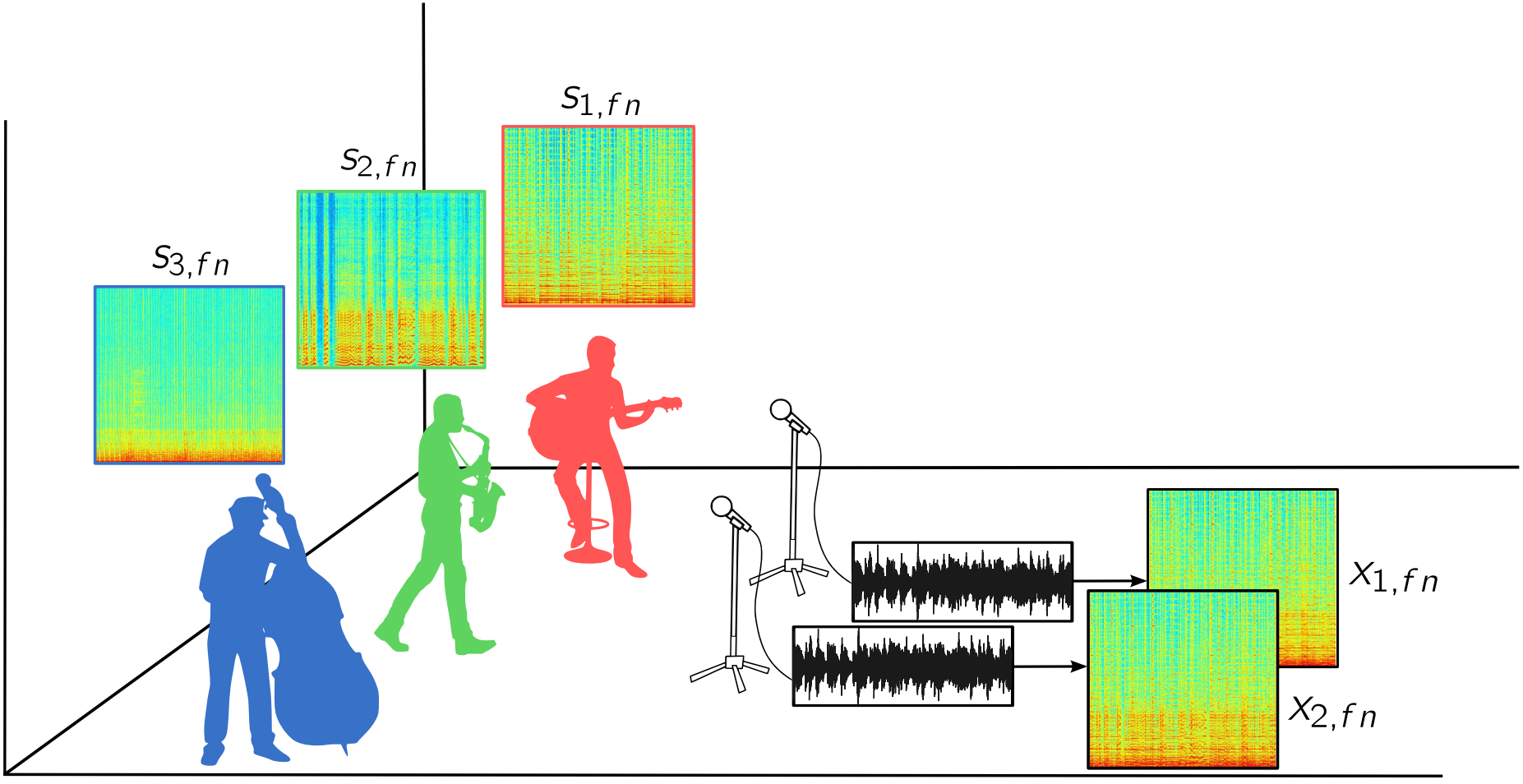
Reverberant mixtures (1)

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



Outline

Probabilistic model with latent variables (1)

- Observed random variables: $\mathbf{x}$ (e.g. mixture coefficients)
- Latent random variables: $\mathbf{z}$ (e.g. source coefficients)

Defining the probabilistic model

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

where $\boldsymbol{\theta}$ is the set of model parameters.

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

Probabilistic model with latent variables (2)

- Observed random variables: $\mathbf{x}$ (e.g. mixture coefficients)
- Latent random variables: $\mathbf{z}$ (e.g. source coefficients)

Defining the probabilistic model

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

where $\boldsymbol{\theta}$ is the set of model parameters.

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

Statistical inference

Posterior inference

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

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

with $\boldsymbol{\theta}^\star$ an estimate of the model parameters.

Parameters estimation

- Maximum likelihood (ML) - parameters are treated as deterministic variables:

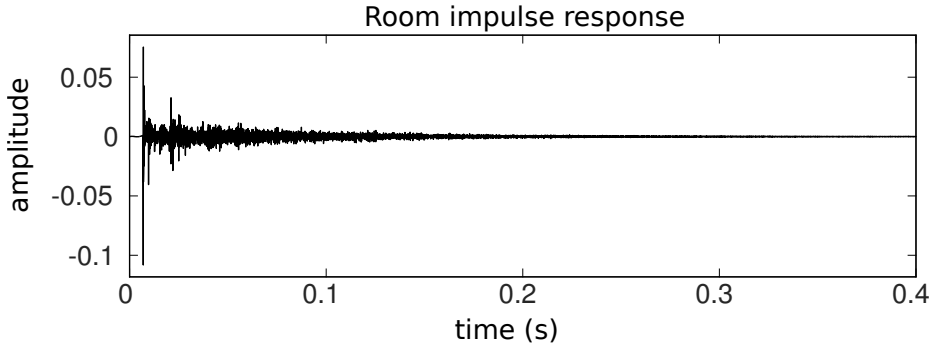
$$\boldsymbol{\theta}^\star_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta}).$$

- Maximum a posteriori (MAP) - parameters are treated as random variables:

$$\boldsymbol{\theta}^\star_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Research problem

- Mixing filters are usually treated as deterministic parameters only estimated from the observed data.
- We know that they correspond to room responses.



How can we guide the estimation of the mixing filters?

Two approaches:

1. STFT-domain convolutive mixture model;
2. Time-domain convolutive mixture model.

Outline

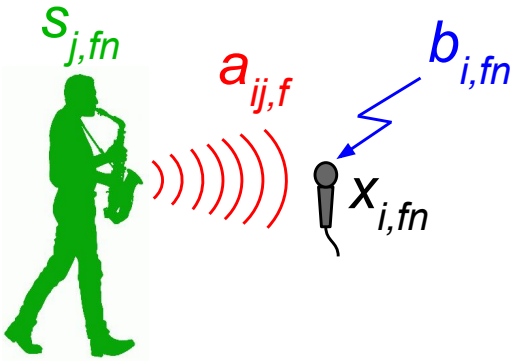
Outline

Mixture model

Convolutive noisy mixture of J sources on I channels

STFT domain:
$\forall (f, n) \in \{0, ..., F-1\} \times \{0, ..., N-1\}$

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



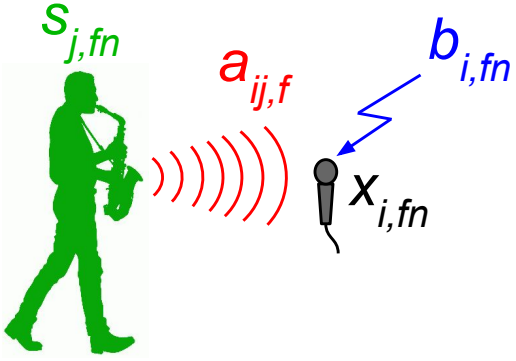
- Frequency response of the mixing filters: $a_{ij,f}$
- Source STFT coefficients: $s_{j,fn}$
- Additive Gaussian noise: $b_{i,fn} \sim \mathcal{N}_c(0, \sigma^2_{b,f})$

Mixture model

Convolutive noisy mixture of J sources on I channels

STFT domain:
$\forall (f, n) \in \{0, ..., F-1\} \times \{0, ..., N-1\}$

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



- Frequency response of the mixing filters: $a_{ij,f}$
- Source STFT coefficients: $s_{j,fn}$
- Additive Gaussian noise: $b_{i,fn} \sim \mathcal{N}_c(0, \sigma^2_{b,f})$

In matrix form

$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn}, \quad (1)$$

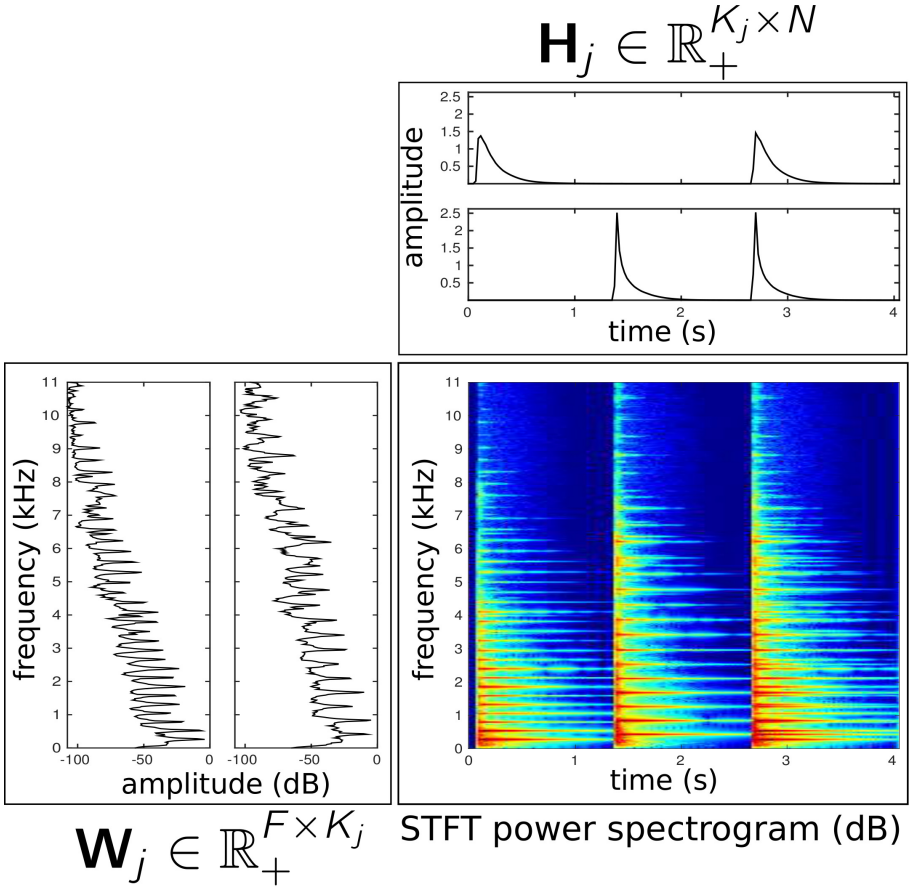
where $\mathbf{x}_{fn} = [x_{i,fn}]_i \in \mathbb{C}^I$, $\mathbf{s}_{fn} = [s_{j,fn}]_j \in \mathbb{C}^J$, $\mathbf{A}_f = [a_{ij,f}]_{ij} \in \mathbb{C}^{I \times J}$ and $\mathbf{b}_{fn} = [b_{i,fn}]_i \in \mathbb{C}^I$.

Source model

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

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

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



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

Statistical inference (1)

- Latent source random variables: $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$;
- Observed random variables: $\mathbf{x} = \{x_{i,fn}\}_{i,f,n}$;
- Model parameters: $\boldsymbol{\theta} = \left\{ \{\mathbf{W}_j, \mathbf{H}_j\}_j, \left\{\mathbf{A}_f, \sigma^2_{b,f}\right\}_f \right\}$.

Source posterior mean

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star_{\mathrm{ML}}}[\mathbf{s}]$$

Maximum likelihood estimation of the parameters

$$\boldsymbol{\theta}^\star_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x};\boldsymbol{\theta})$$

$\rightarrow$ Expectation-maximization (EM) algorithm [2].

[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.

Statistical inference (2)

- Latent source random variables: $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$;
- Observed random variables: $\mathbf{x} = \{x_{i,fn}\}_{i,f,n}$;
- Model parameters: $\boldsymbol{\theta} = \left\{ \{\mathbf{W}_j, \mathbf{H}_j\}_j, \left\{ \mathbf{A}_f, \sigma^2_{b,f} \right\}_f \right\}$.

Source posterior mean

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\theta}^\star_{\text{MAP}}}[\mathbf{s}]$$

Maximum a posteriori estimation of the parameters

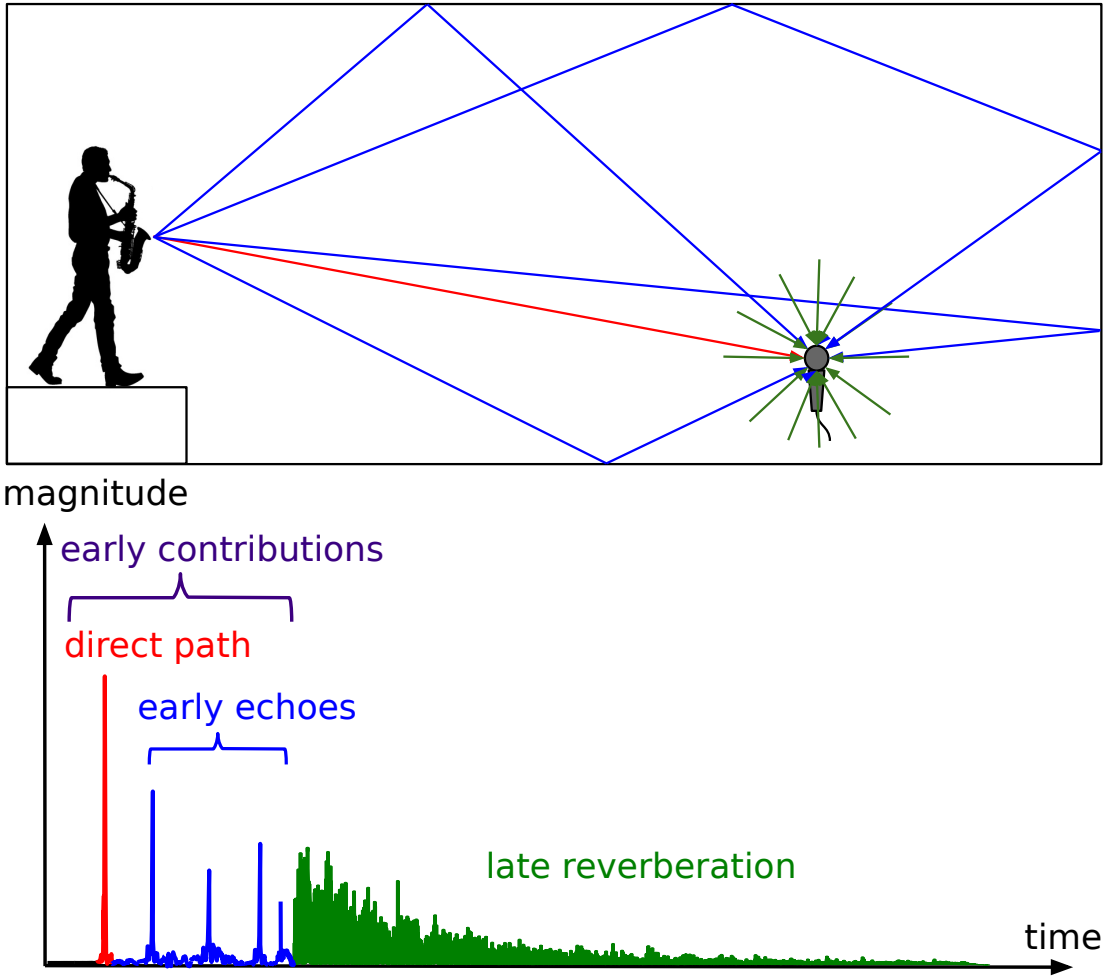
$$\boldsymbol{\theta}^\star_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{x}|\boldsymbol{\theta}) p\left(\{\mathbf{A}_f\}_f\right)$$

$\rightarrow$ Expectation-maximization (EM) algorithm.

Outline

Room impulse response

Mixing filters are room responses. They exhibit a simple specific structure in the time domain.



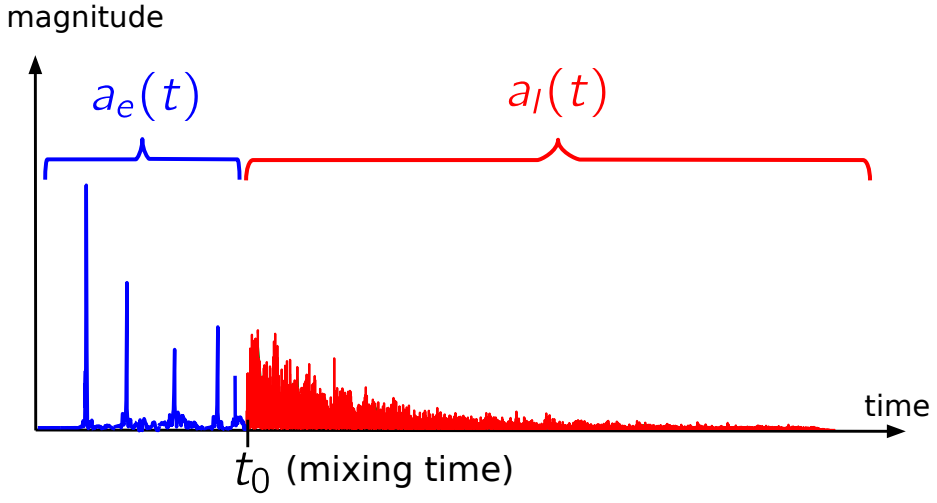
Room impulse and frequency responses

Room impulse and frequency responses

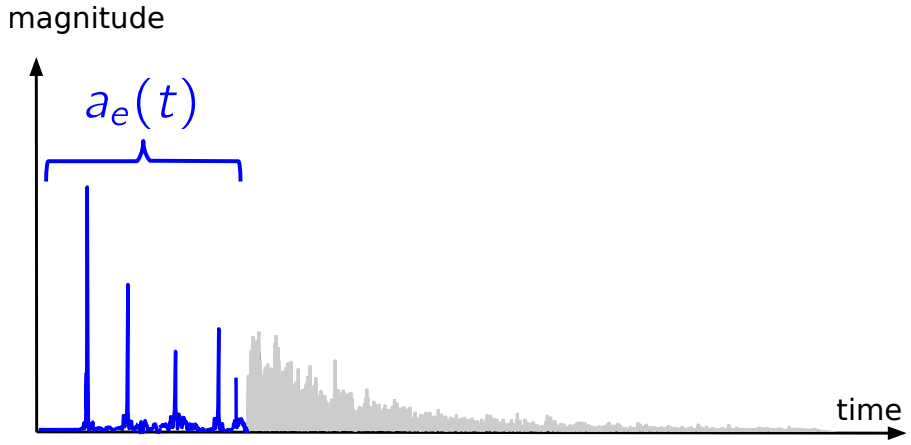
$\forall t, f \in \{0, ..., T-1\}$:

$$\underbrace{a(t) = a_e(t) + a_l(t)}_{\text{Room impulse response (RIR)}} \overset{\mathcal{F}_T}{\underset{\mathcal{F}_T^{-1}}{\rightleftarrows}} \underbrace{A(f) = A_e(f) + A_l(f)}_{\text{Room frequency response (RFR)}}$$

$\mathcal{F}_T$: Discrete Fourier transform (DFT) over T points.



Early contributions model (1)



Geometrical room acoustics

The k-th early contribution is characterized by an amplitude ρ_k and a delay τ_k:

$$A_e(f) = \sum_{k=0}^{R-1} \rho_k \delta_k^f \quad \text{with} \quad \delta_k = e^{-\imath 2\pi \tau_k / T}. \tag{2}$$

Early contributions model (2)

It follows (see, e.g., [3])

$$\{A_e(f)\}_{f=R,\ldots,T-1} \quad \text{satisfies} \quad \sum_{r=0}^{R} \varphi_r^e A_e(f-r) = 0, \tag{3}$$

where $\{\varphi_r^e\}_{r=0}^{R}$ and $\{\delta_k\}_{k=0}^{R-1}$ are the coefficients and roots of a same polynomial of order R.

Adding an error term $\rightarrow$ autoregressive model

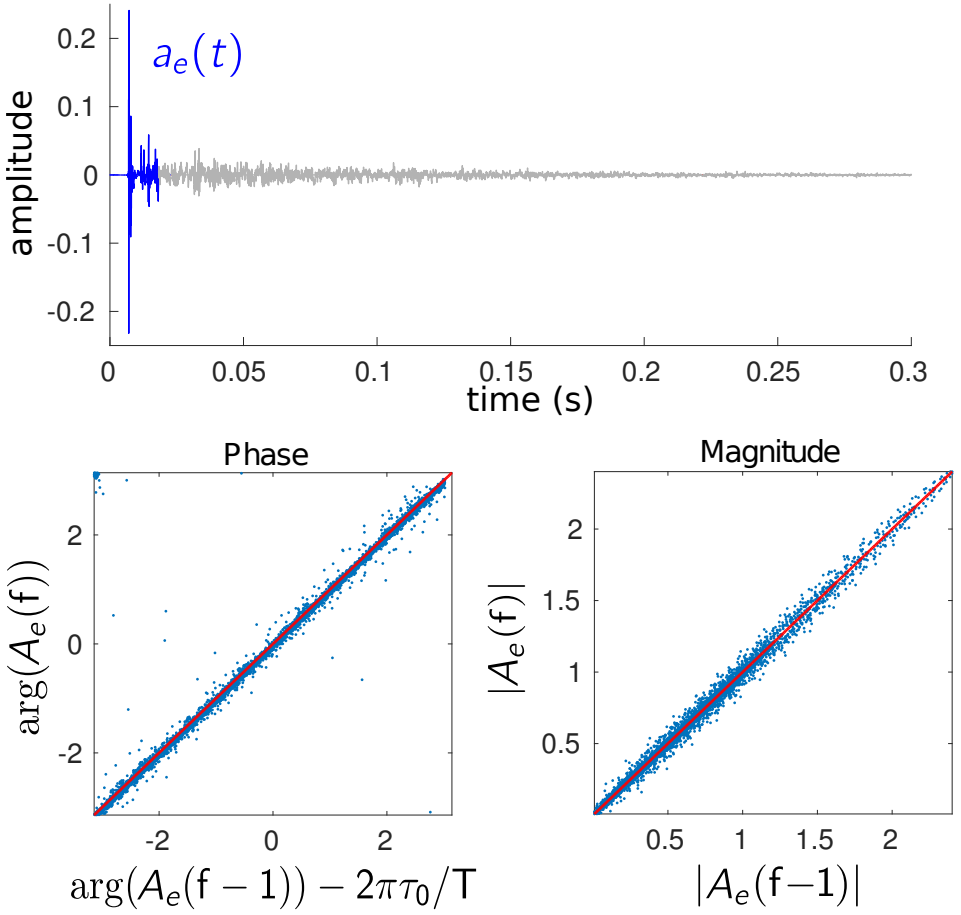
$$\sum_{r=0}^{R} \varphi_r^e A_e(f-r) = \kappa(f) \quad \text{with} \quad \kappa(f) \overset{i.i.d}{\sim} \mathcal{N}_c(0, \sigma_\kappa^2). \tag{4}$$

[3] R. Kumaresan, "On the zeros of the linear prediction-error filter for deterministic signals", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1983.

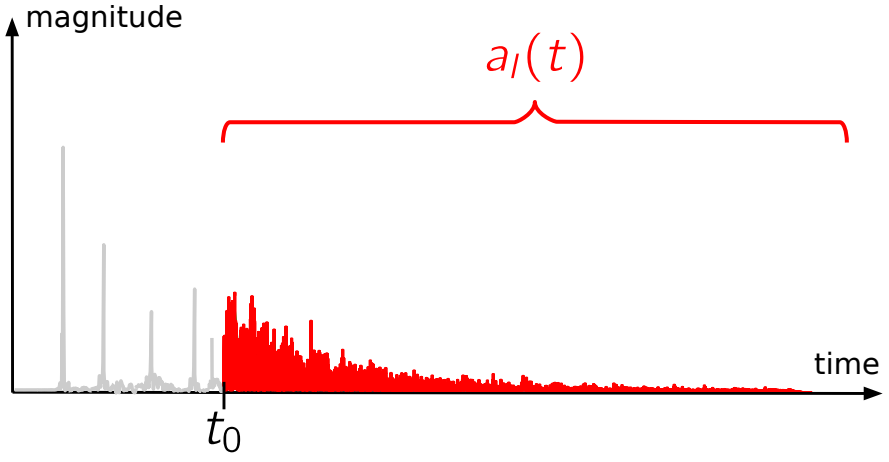
Autoregressive model of order 1

Assuming that the direct path dominates the early echoes:

$$|A_e(f)| \approx |A_e(f-1)| \quad \text{and} \quad \arg(A_e(f)) \approx \arg(A_e(f-1)) - 2\pi\frac{\tau_0}{T}.$$



Late reverberation model (time domain)



Statistical room acoustics

Exponentially decaying energy temporal profile (ETP):

$$\bar{a}_l(t) = \mathbb{E}\left[a_l^2(t)\right] \propto e^{-2t/\tau}\mathbb{1}_{t\geq t_0}(t).$$

- τ is related to the reverberation time T_{60} in seconds.
- $\mathbb{E}[\cdot]$: spatial averaging.

Late reverberation model (frequency domain)

Statistical room acoustics: $\{A_l(f)\}_f$ is a proper, centered and wide-sense stationary complex Gaussian random process.

Theoretical power spectral density (PSD)

We can show that the PSD is related to the ETP by:

$$\phi(t) = T\bar{a}_l(T-t).$$

Theoretical autocovariance function (ACVF)

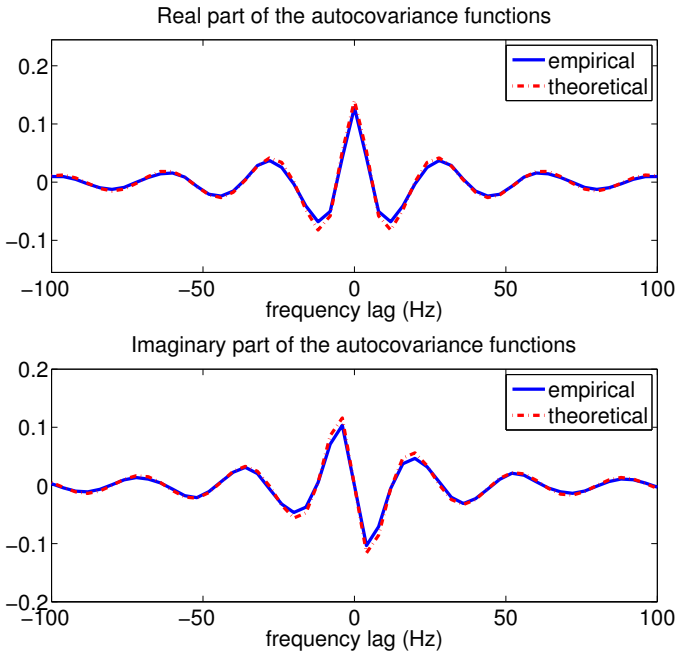
Applying the Wiener-Khinchin theorem we can obtain a theoretical expression of the ACVF:

$$\gamma(m) = \mathcal{F}_T^{-1}\{\phi(t)\}.$$

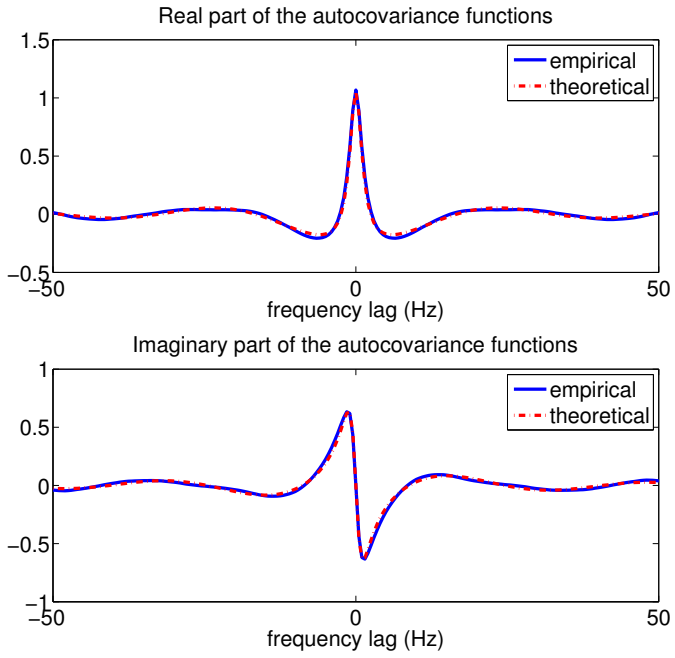
These quantities are theoretically defined according to some room parameters (reverberation time, dimensions).

Experimental validation

Empirical autocovariance functions computed from a Monte Carlo simulation on synthesized and real room responses.



(a) 196 simulated RIRs
$T_{60} = 0.25$ s
$10 \times 6.6 \times 3.3$ m



(b) 130 real RIRs
$T_{60} = 1.8$ s
$7.5 \times 9 \times 3.5$ m

Figure: Theoretical and empirical autocovariance functions

ARMA parametrization

ARMA representation of late reverberation in the frequency domain

We assume that $\{A_l(f)\}_f$ follows an ARMA(P, Q) model:

$$\Phi(L)A_l(f) = \Theta(L)\epsilon(f),$$

- $\Phi(L) = \sum_{p=0}^{P} \varphi_p^l L^p$ and $\Theta(L) = \sum_{q=0}^{Q} \vartheta_q L^q$ with $\varphi_0^l = \vartheta_0 = 1$;
- L is the lag operator, i.e. $LA_l(f) = A_l(f-1)$;
- $\epsilon(f) \overset{i.i.d}{\sim} \mathcal{N}_c(0, \sigma_\epsilon^2)$.

We can compute the ARMA parameters from the theoretical ACVF.

Experimental validation

Same room parameters as used before for simulated RIRs.

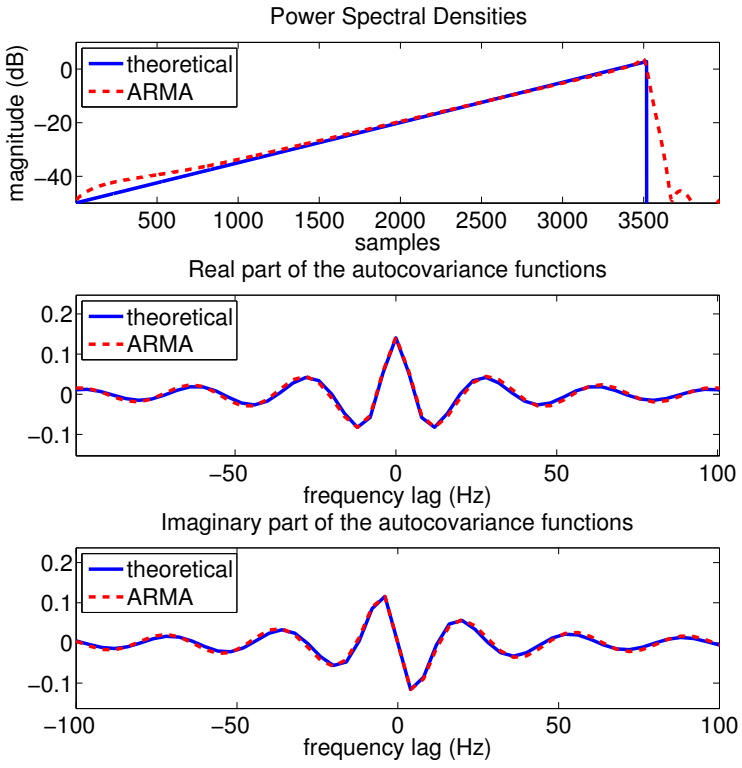


Figure: ARMA(7,2) parametrization

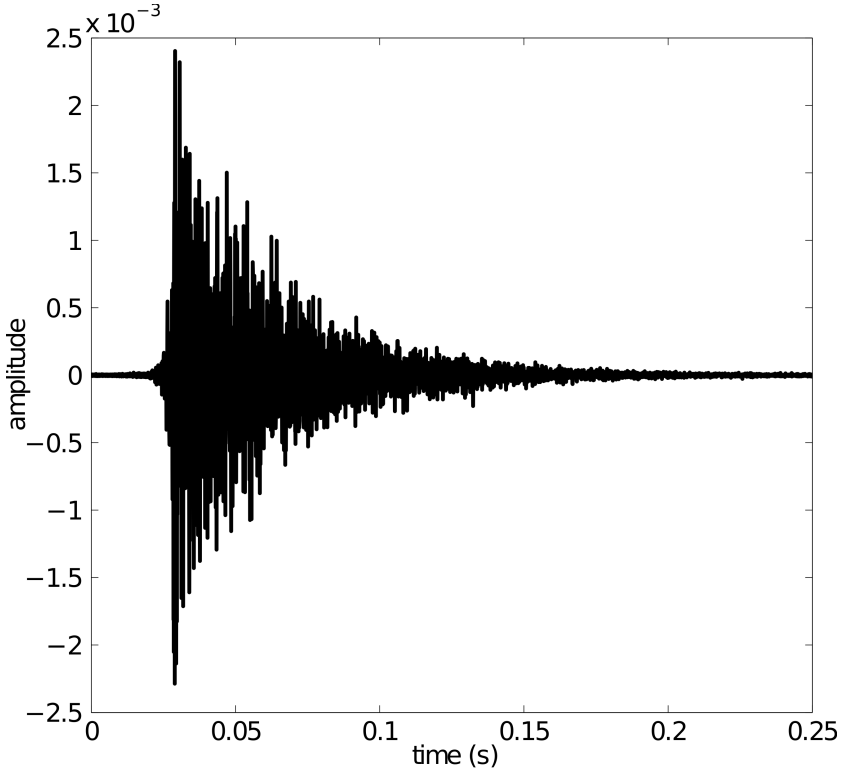


Figure: Synthesized late RIR

Outline

EM algorithm

- Mixing matrix: $\mathbf{A}_f = \underbrace{\mathbf{A}_{e,f}}_{\text{early reverb.}} + \underbrace{\mathbf{A}_{l,f}}_{\text{late reverb.}}$

E-step

$$Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) = \mathbb{E}_{\mathbf{s}|\mathbf{x},\boldsymbol{\theta}_{\text{old}}} \left[\ln p(\mathbf{x}, \mathbf{s}|\boldsymbol{\theta}) \right]$$

M-step

ML estimation:

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}})$$

MAP estimation:

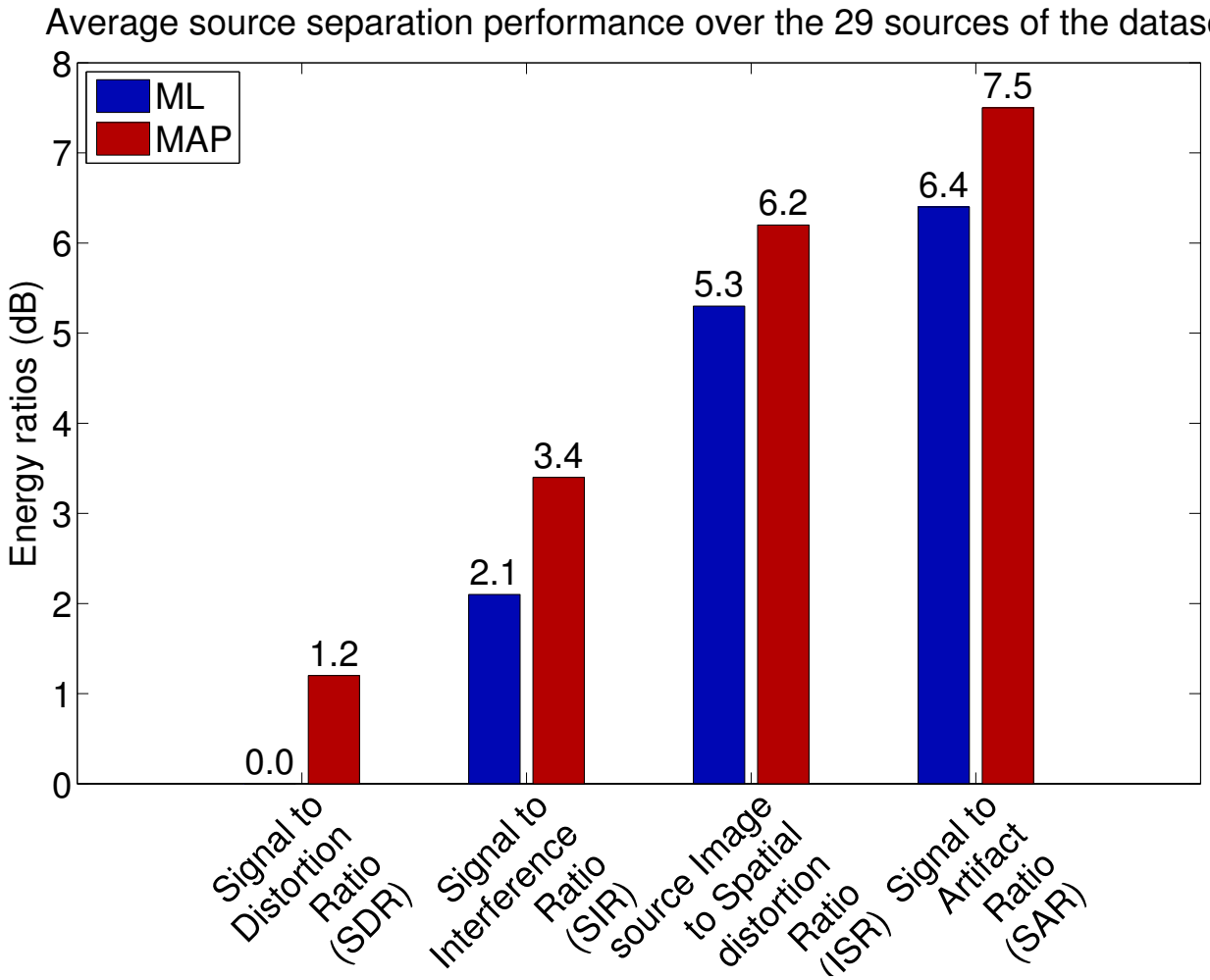
$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}_{\text{old}}) + \ln p(\{\mathbf{A}_{e,f}\}_f) + \ln p(\{\mathbf{A}_{l,f}\}_f)$$

$\Rightarrow$ ML and MAP estimations only differ in the mixing filters update at the M-step.

Experiments

- Dataset:
  - 8 stereo mixtures created with synthetic room impulse responses;
  - Reverberation time: 128 ms;
  - Number of sources per mixture: 3 to 5;
  - Musical sources: drums, piano, bass, guitar, voice;
  - Duration: 12 to 28 seconds.
- Source separation results: ML (w/o priors) vs. MAP (w/ priors).
- Both algorithms are run from the same initialization.

Source separation results

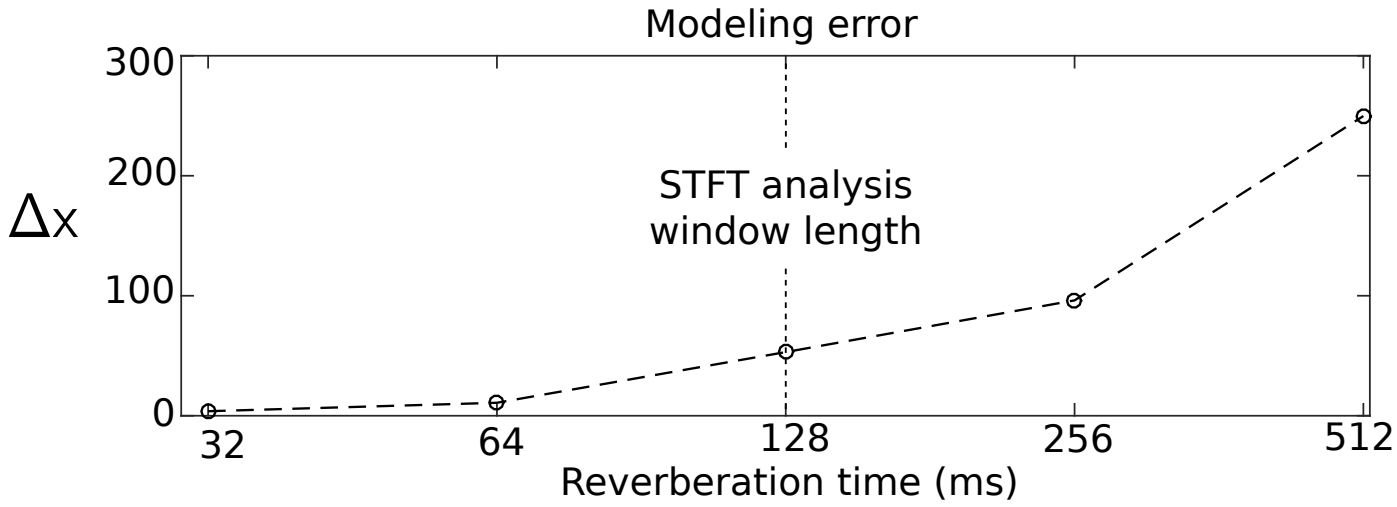


Remark: Some hyperparameters of the priors (noise variances) are optimized knowing the true source signals.

Outline

Main limitation

Error due to the STFT approximation of the convolutive mixing process.



Average relative squared error: $\Delta x = \frac{1}{IFN} \sum_{i,f,n} \frac{|\, x_{i,fn} - \hat{x}_{i,fn} \,|^2}{|\, x_{i,fn} \,|^2}$,

- $x_{i,fn} = \mathrm{STFT}\{x_i(t)\}$;
- $\hat{x}_{i,fn} = \sum_{j=1}^{J} a_{ij,f}\, s_{j,fn}$.

Overcoming this limitation

- More accurate time-frequency convolutive mixture representations:
  - 2-dimensional filtering [4];
  - Subband filtering (convolutive transfer function) [5, 6].
- **Time-domain convolutive mixture representation** [7]:
  - Exact representation of the convolutive mixing process;
  - Suitable for incorporating simple priors on the mixing filters.

[4] R. Badeau and M. Plumbley, "Multichannel high-resolution NMF for modeling convolutive mixtures of non-stationary signals in the time-frequency domain", *IEEE Transactions on Audio, Speech and Language Processing*, 2014.

[5] X. Li, L. Girin and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain Lasso optimization", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[6] X. Li, L. Girin and R. Horaud, "An EM algorithm for audio source separation based on the convolutive transfer function", *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

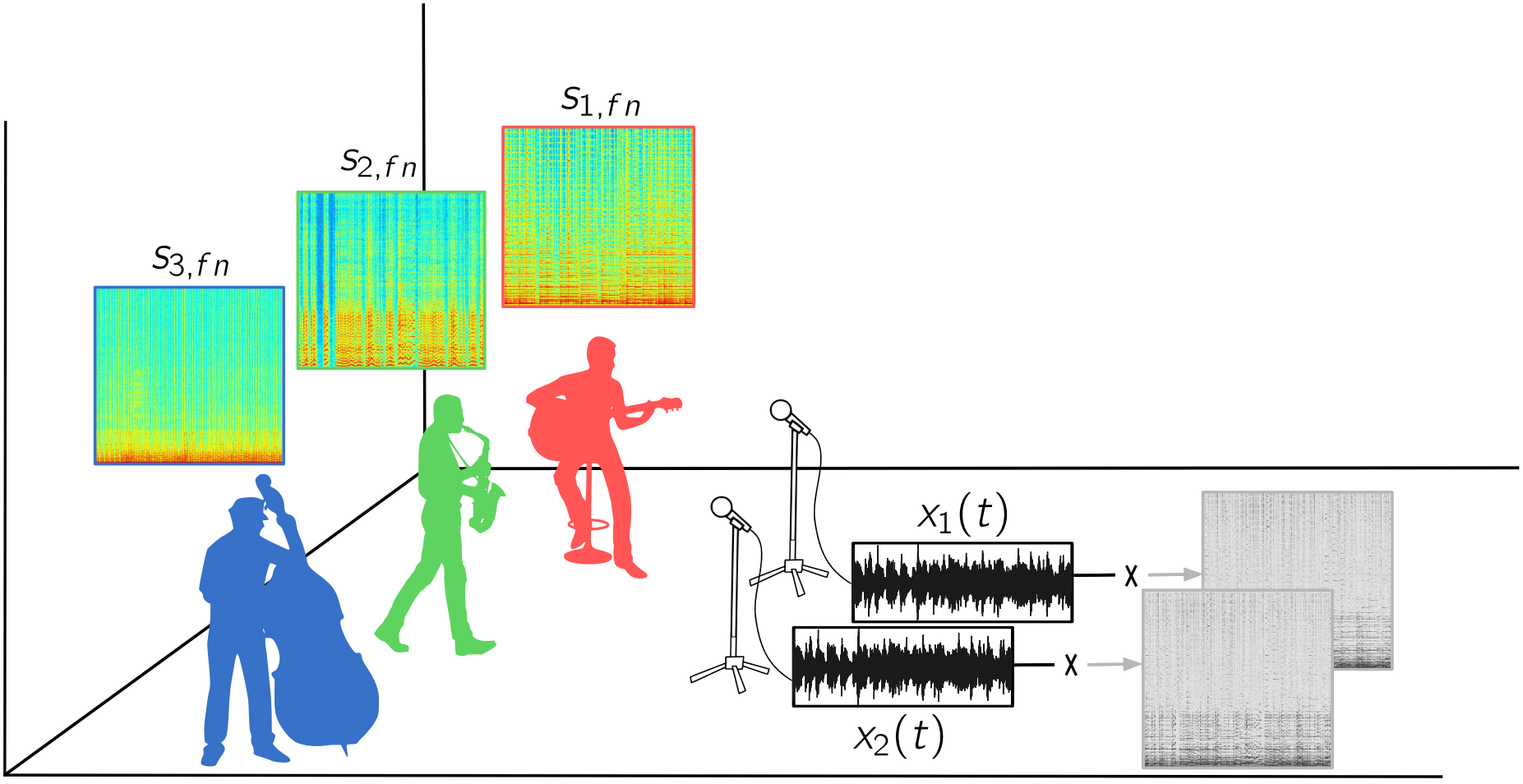
[7] M. Kowalski, E. Vincent and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation". *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

Outline

Outline

Proposed approach

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



Mixture model

Time-domain convolutive mixture model

$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t) + b_i(t), \tag{5}$$

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

Time-frequency source representation

$$s_j(t) = \sum_{(f,n) \in \mathcal{B}_j} s_{j,fn} \psi_{j,fn}(t), \tag{6}$$

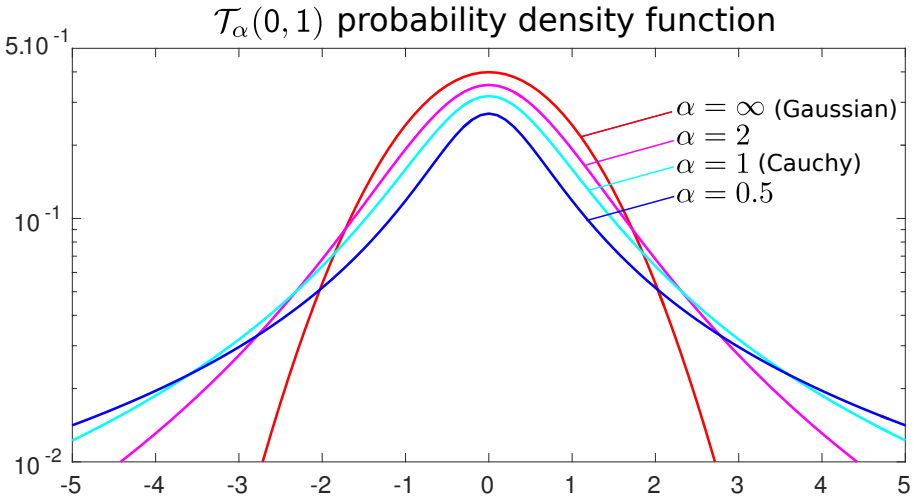
with $\psi_{j,fn}(t) \in \mathbb{R}$ a **source-dependent** modified discrete cosine transform (MDCT) atom and $\mathcal{B}_j = \{0, ..., F_j - 1\} \times \{0, ..., N_j - 1\}$.

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

Student's t distribution

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

- Shape: α;
- Location: μ;
- Scale: σ.



Scale mixture of Gaussians

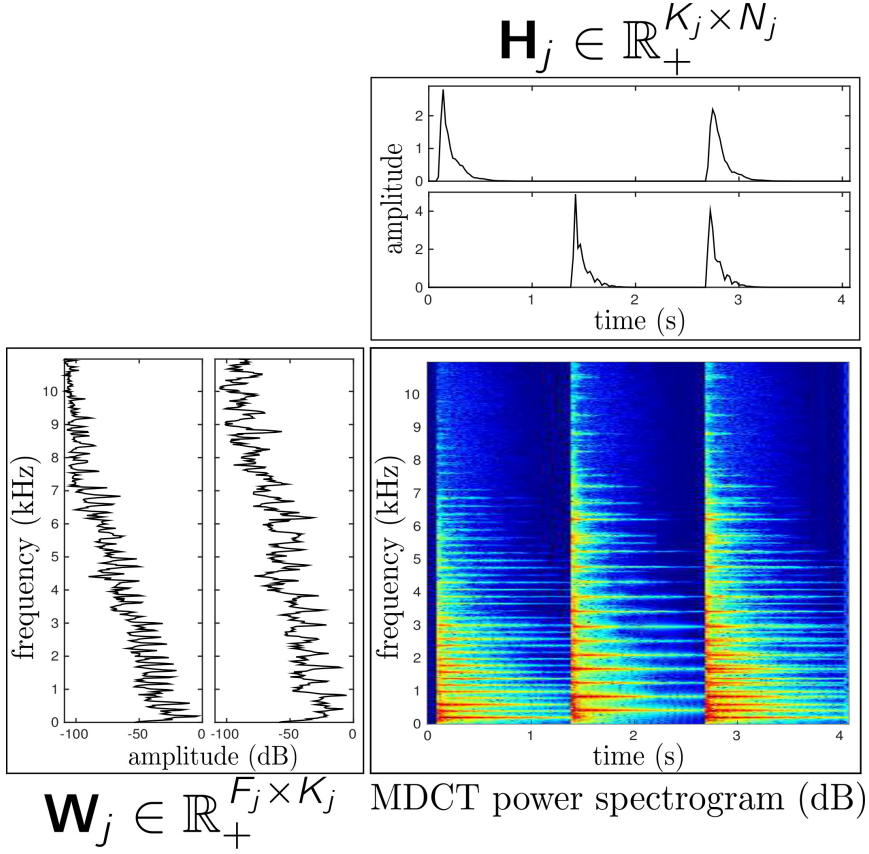
$$z \sim \mathcal{T}_\alpha(\mu, \sigma) \quad \Leftrightarrow \quad \begin{cases} z|v & \sim \mathcal{N}\left(\mu, v\sigma^2\right) \\ v & \sim \mathcal{IG}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right) \end{cases}$$

Source model

Student's t source model based on NMF [8]:

$$s_{j,fn} \sim \mathcal{T}_{\alpha_v}\left(0, \lambda_{j,fn}\right),$$

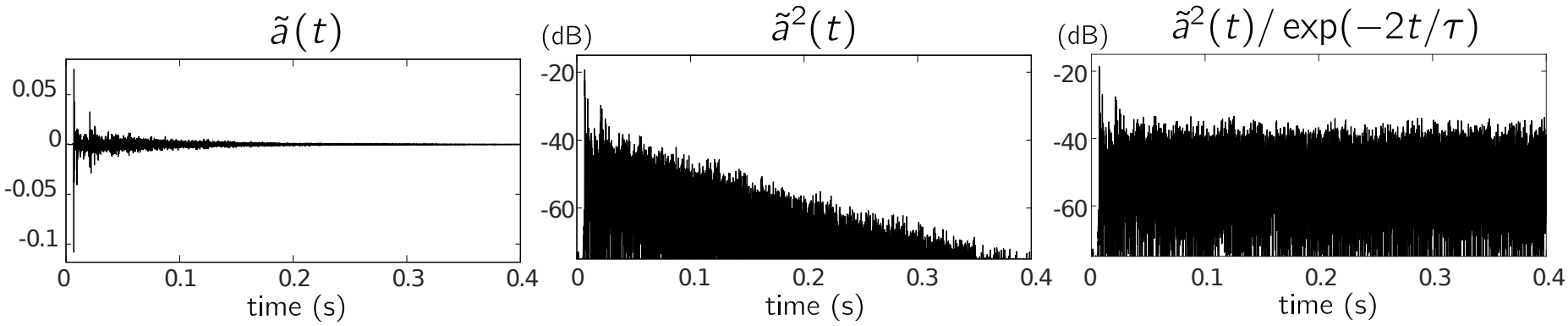
$$\text{with } \lambda^2_{j,fn} = (\mathbf{W}_j\mathbf{H}_j)_{f,n}.$$



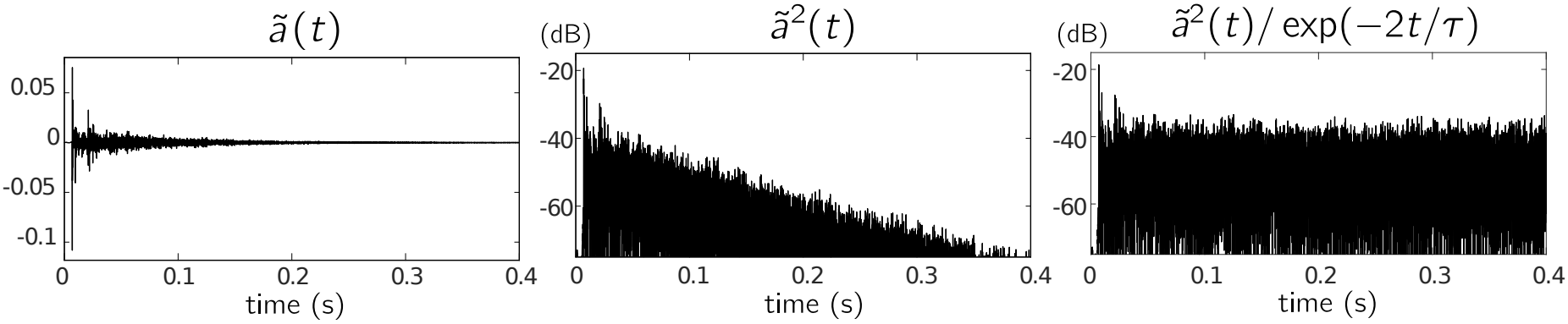
Remark: Source time-frequency coefficients are real-valued.

[8] 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 [9]

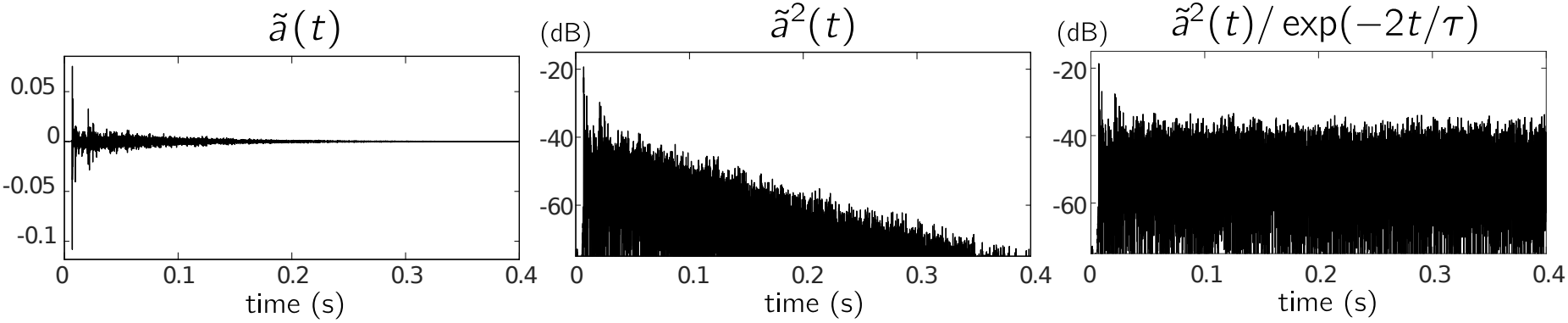
Independently for all $t \in \{0, ..., L_a - 1\}$:

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

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

[9] 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 [9]

Equivalently:

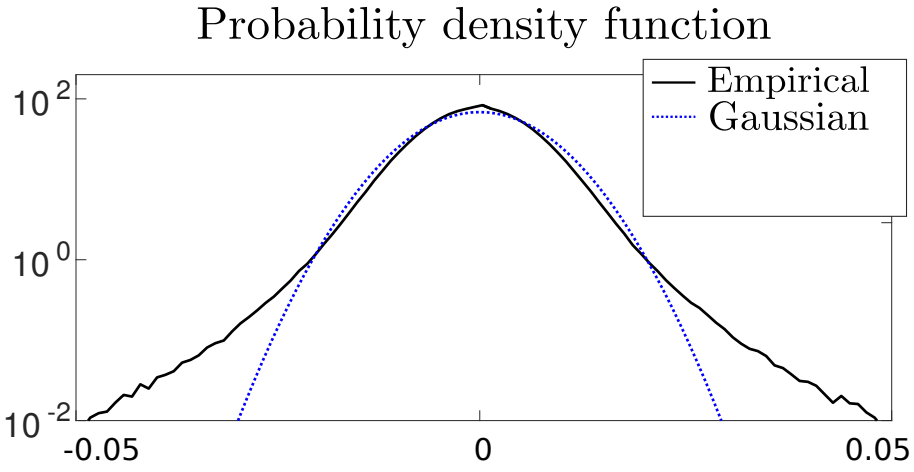
$$a(t)/\exp(-t/\tau) \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_r^2).$$

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

[9] 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 $a(t)/\exp(-t/\tau)$

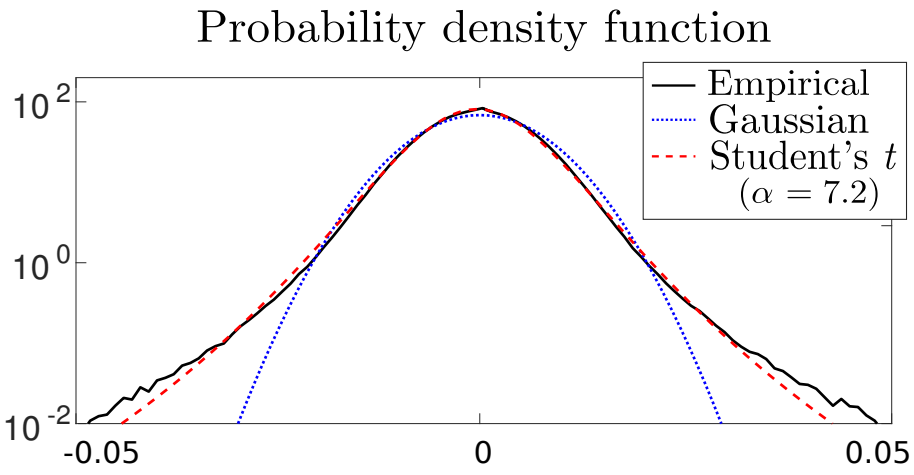


- 624 RIRs from the MIRD database [10];
- Reverberation time equals 610 ms.

[10] 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 $a(t)/\exp(-t/\tau)$

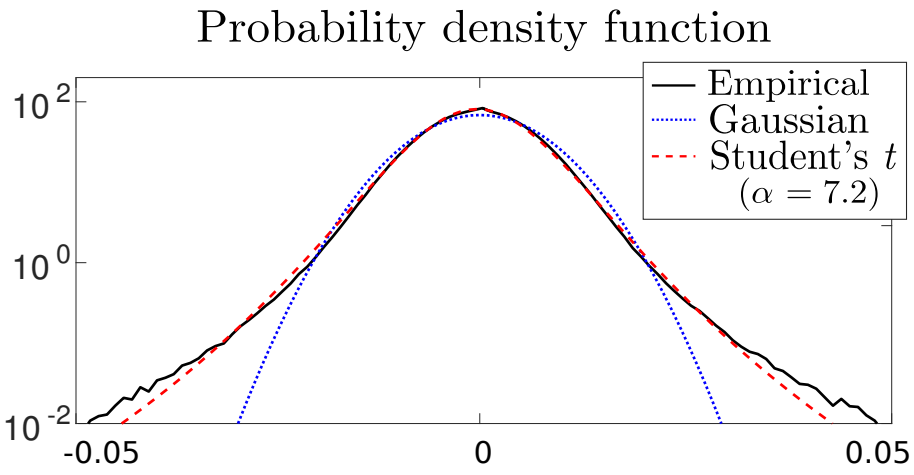


- 624 RIRs from the MIRD database [10];
- Reverberation time equals 610 ms.

[10] 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 $a(t)/\exp(-t/\tau)$



- 624 RIRs from the MIRD database [10];
- Reverberation time equals 610 ms.

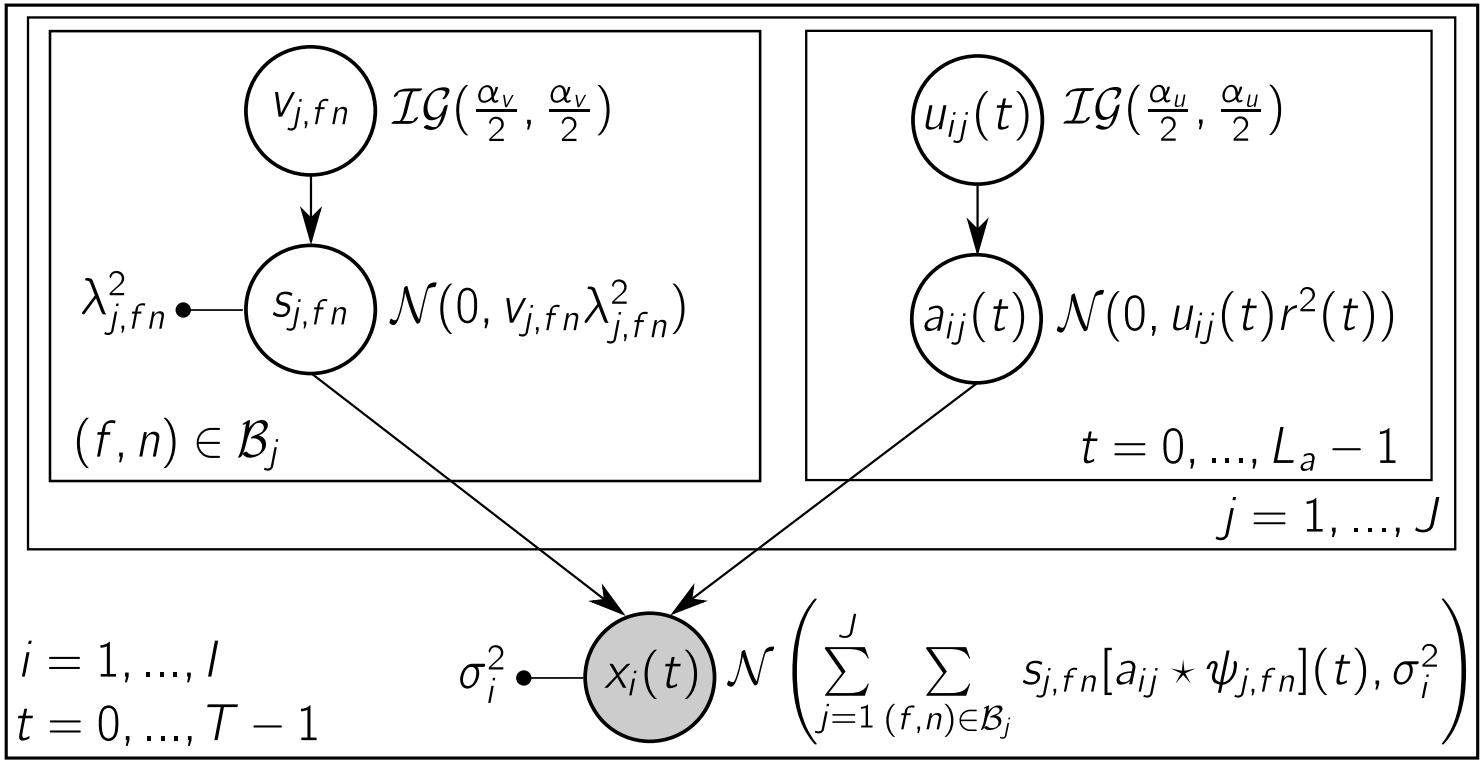
Student's t model with exponential decay

Independently for all i, j and t:

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

[10] 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).

Remark: We recall that $\lambda^2_{j,fn} = (\mathbf{W}_j\mathbf{H}_j)_{f,n}$.

Outline

Variational inference

- Exact posterior inference is analytically intractable.

Variational inference

- Exact posterior inference is analytically intractable.
- **Variational inference**: Find $q \in \mathcal{F}$ which approximates $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$.

Variational inference

- Exact posterior inference is analytically intractable.
- **Variational inference**: Find $q \in \mathcal{F}$ which approximates $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$.
- We take the Kullback-Leibler divergence as a measure of fit:

$$D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-(marginal) likelihood}} - \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{variational free energy}}, \quad (7)$$

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

Variational inference

- Exact posterior inference is analytically intractable.
- **Variational inference**: Find $q \in \mathcal{F}$ which approximates $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})$.
- We take the Kullback-Leibler divergence as a measure of fit:

$$D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-(marginal) likelihood}} - \underbrace{\mathcal{L}(q;\boldsymbol{\theta})}_{\text{variational free energy}}, \quad (7)$$

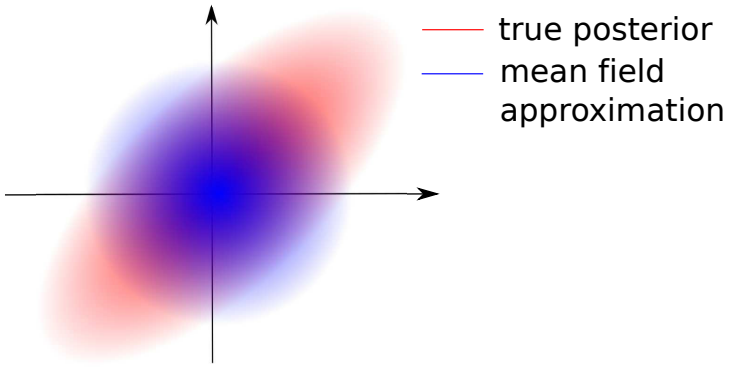
where $\mathcal{L}(q;\boldsymbol{\theta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{q(\mathbf{z})}\right)\right\rangle_q$ and $\langle f(\mathbf{z})\rangle_q = \int f(\mathbf{z})q(\mathbf{z})d\mathbf{z}$.

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

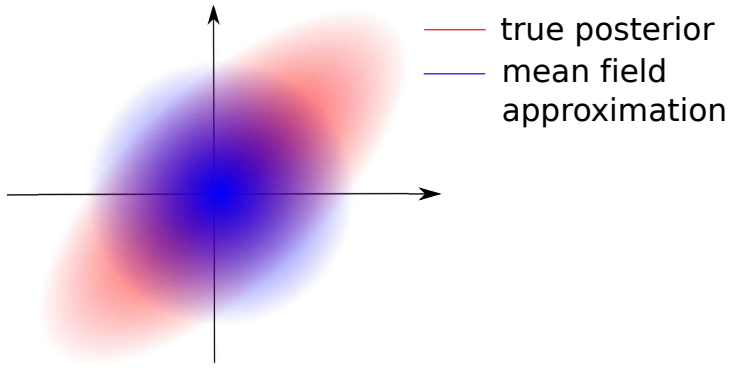
$$q(\mathbf{z}) = \prod_{z_k \in \mathbf{z}} q_k(z_k).$$



Mean-field approximation

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

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



Under the mean-field 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

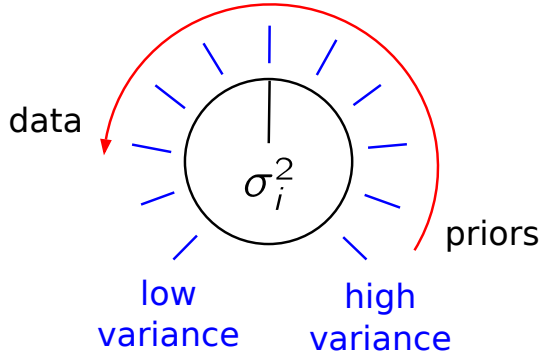
$$\min_{\mathbf{W}_j, \mathbf{H}_j \geq 0} \sum_{(f,n) \in \mathcal{B}_j} d_{IS}\left(\left\langle v_{j,fn}^{-1} \right\rangle_{q^\star} \left\langle s_{j,fn}^2 \right\rangle_{q^\star}, (\mathbf{W}_j \mathbf{H}_j)_{f,n} \right),$$

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.



Outline

Experimental setup

- Dataset:
  - 8 stereo mixtures created with RIRs from the MIRD database [10];
  - 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.

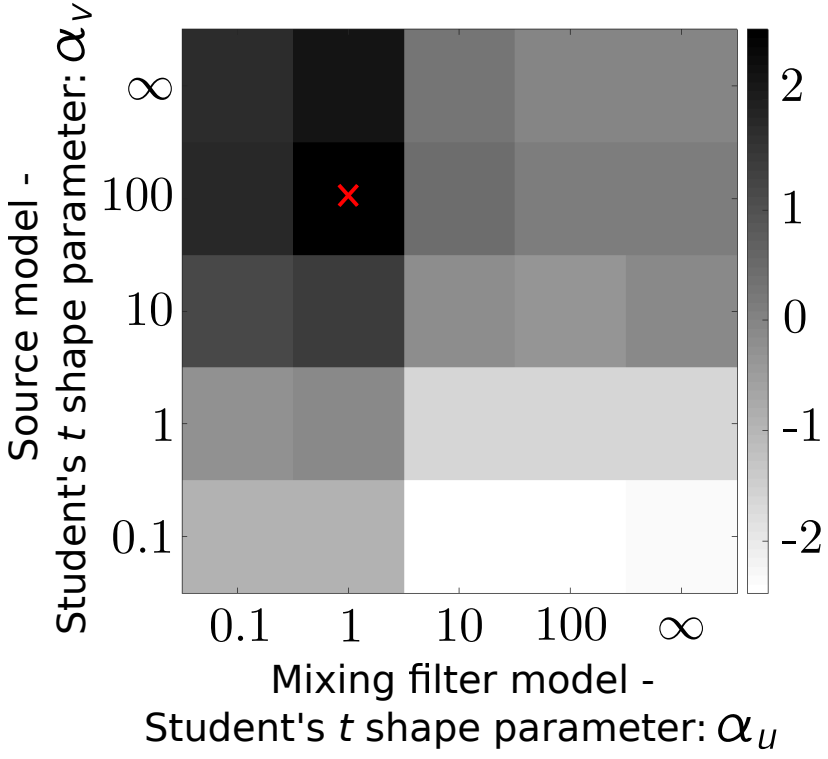
Model hyperparameters

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 SDR in dB using the mixtures with a reverberation time of 360 ms

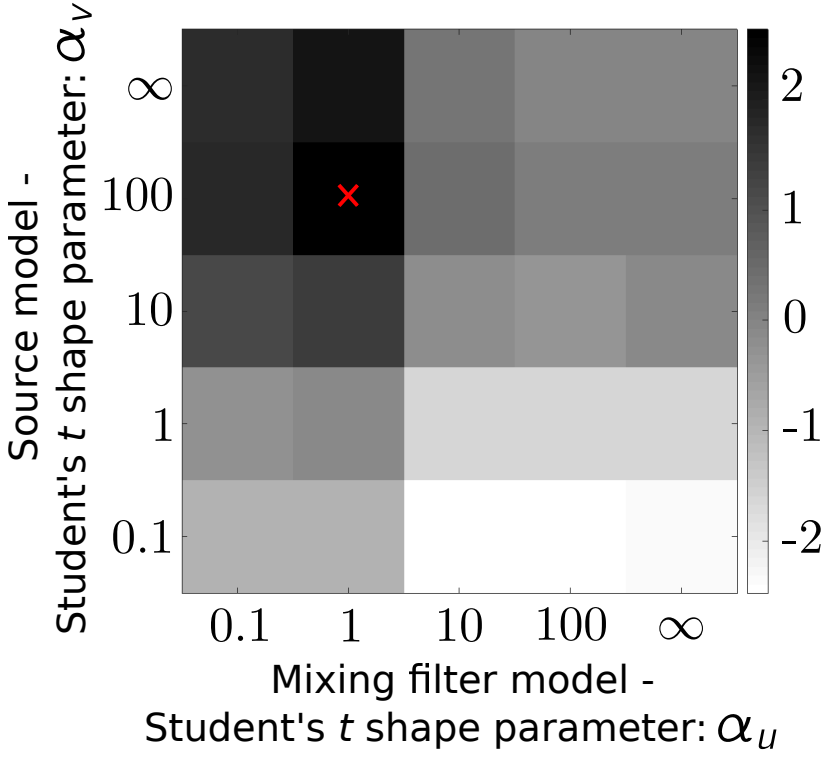


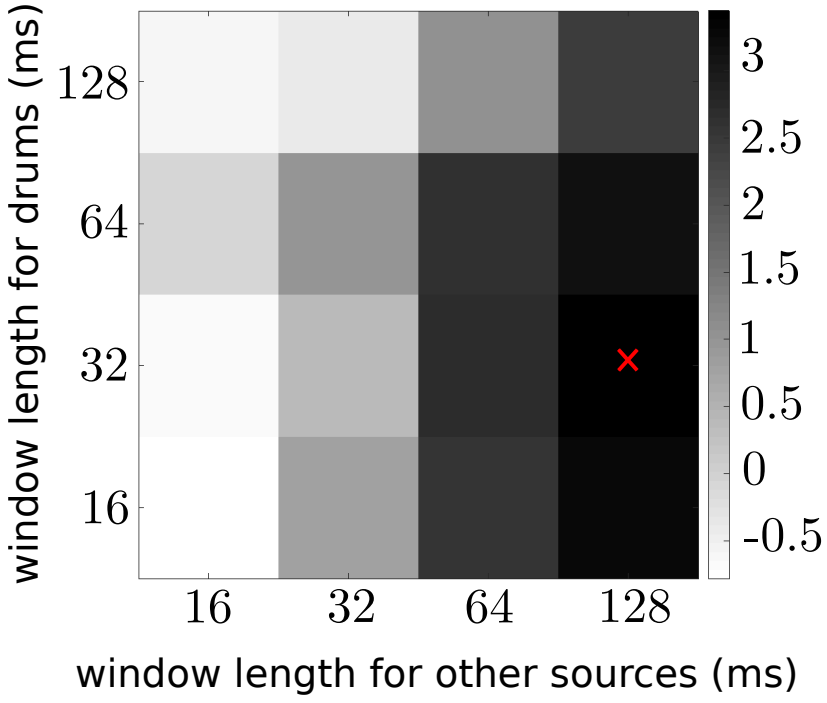
Remark: MDCT window length is fixed to 64 ms.

Model hyperparameters

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

Average SDR in dB 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 $(100, 1)$.

Baseline methods

First method: "Gaussian - SCM rank 1-2 " [11]

- **Source model**: STFT - Gaussian - NMF.
- **Convolutive mixture model**:
  - Approximate STFT-domain convolutive mixture representation.
  - Spatial covariance matrix (SCM) model:
    - Rank 1: punctual source;
    - Rank 2: diffuse source.

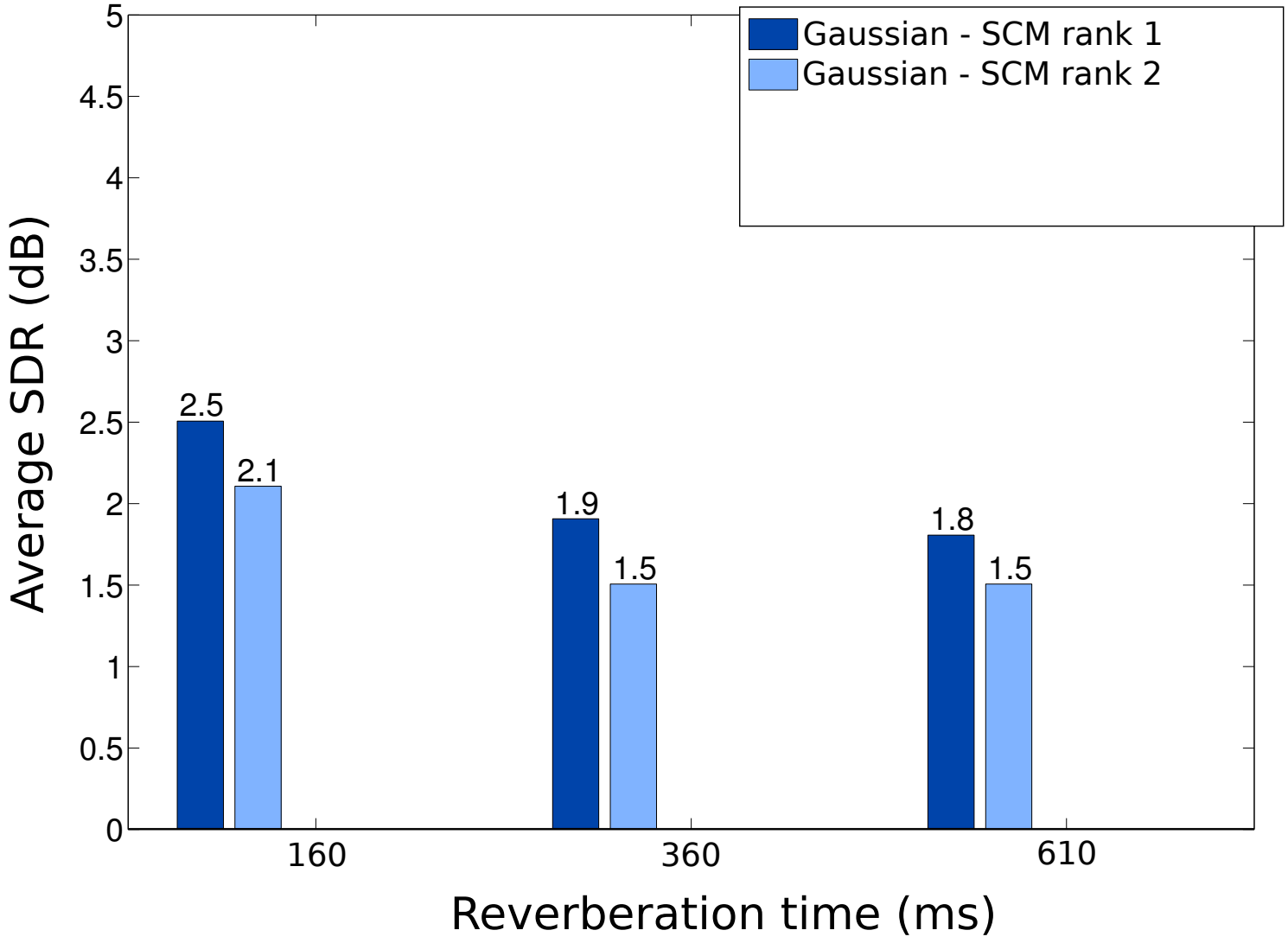
Second method: "Gaussian - unconst. TD filters" [12]

- **Source model**: MDCT - Gaussian - NMF.
- **Convolutive mixture model**:
  - Exact time-domain (TD) convolutive mixture representation.
  - Unconstrained mixing filters (w/o prior).

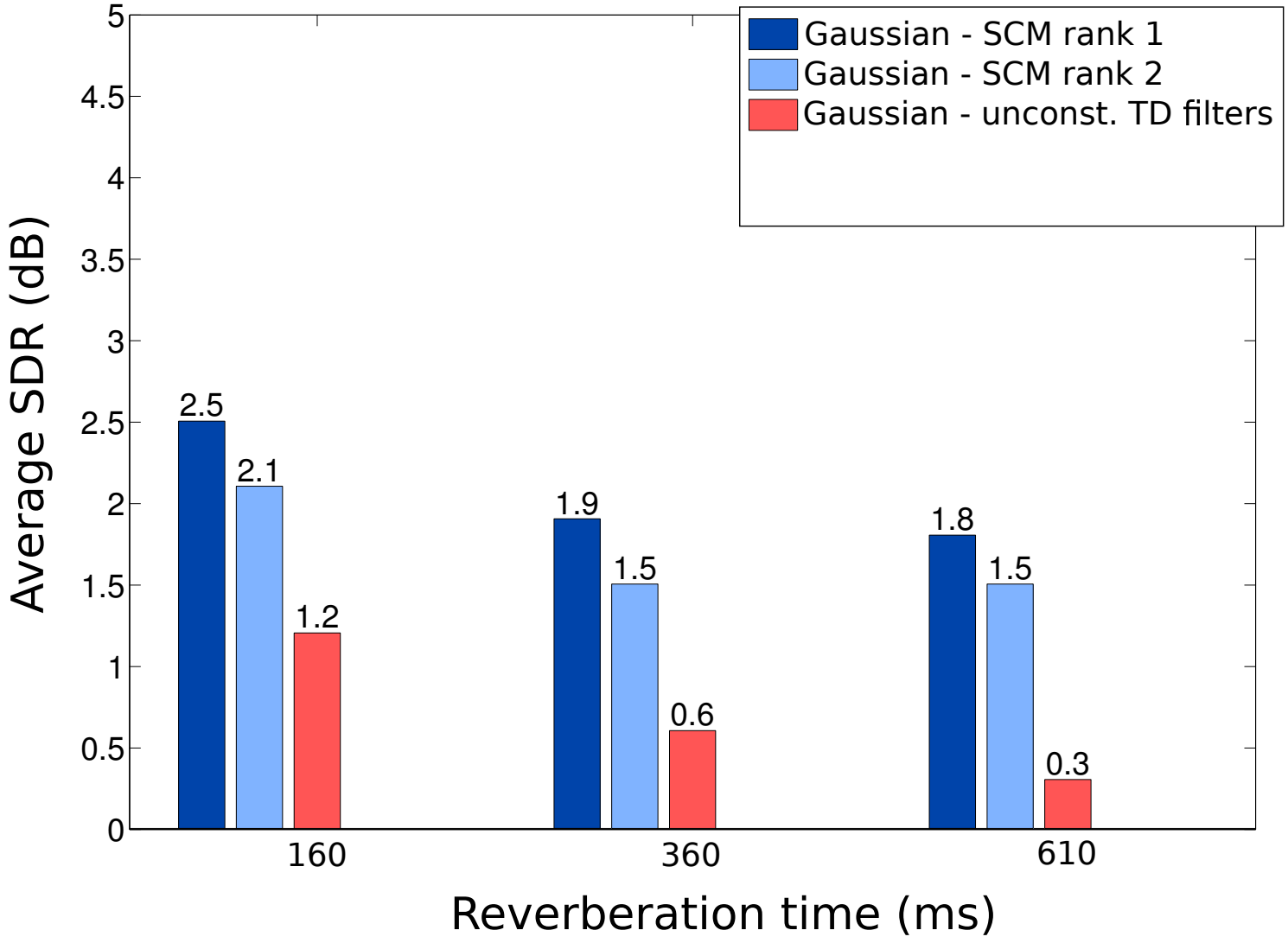
[11] A. Ozerov, E. Vincent and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation", *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[12] S. L., R. Badeau and G. Richard, "Multichannel audio source separation: Variational inference of time-frequency sources from time-domain observations", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

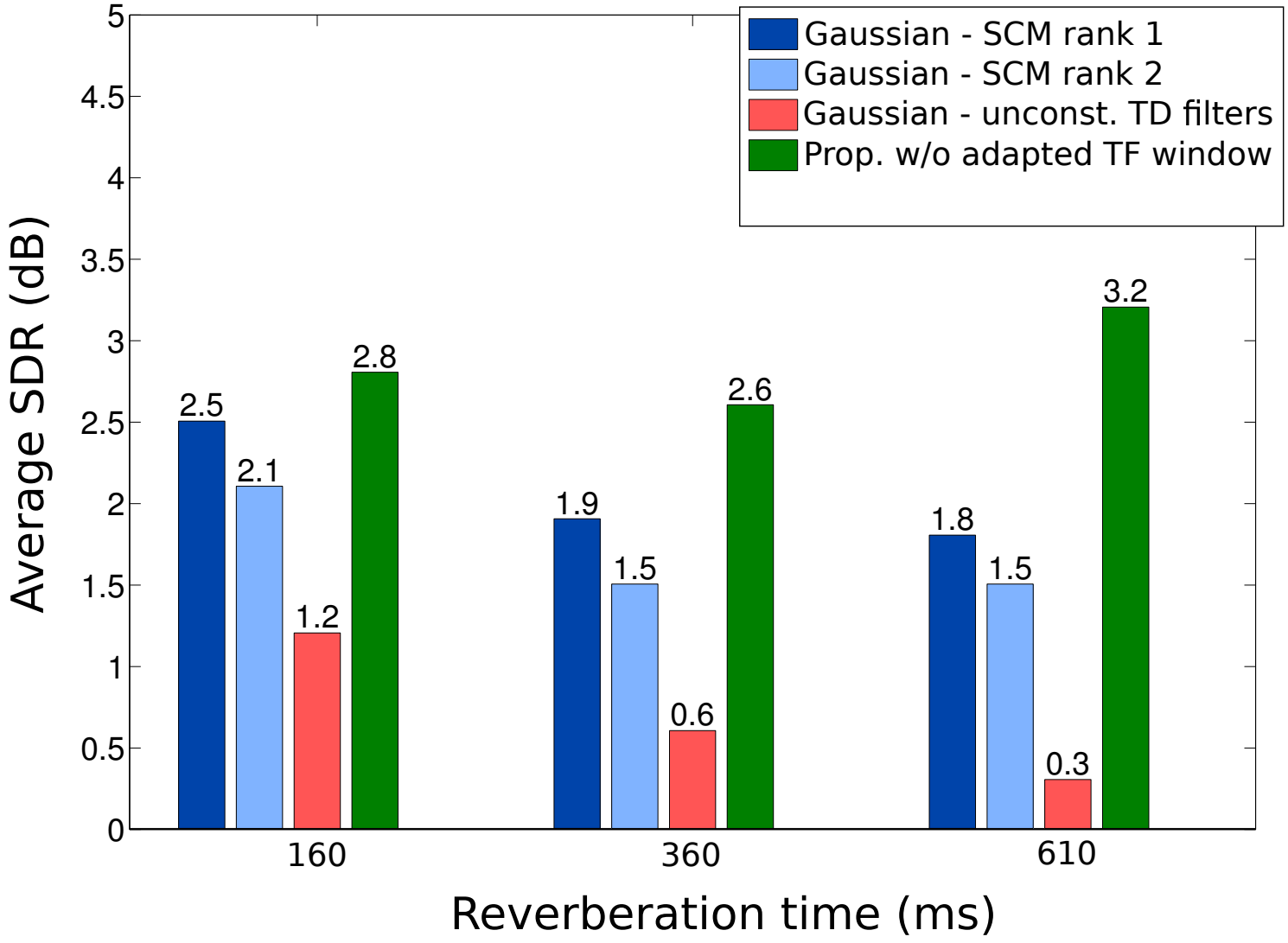
Source separation results



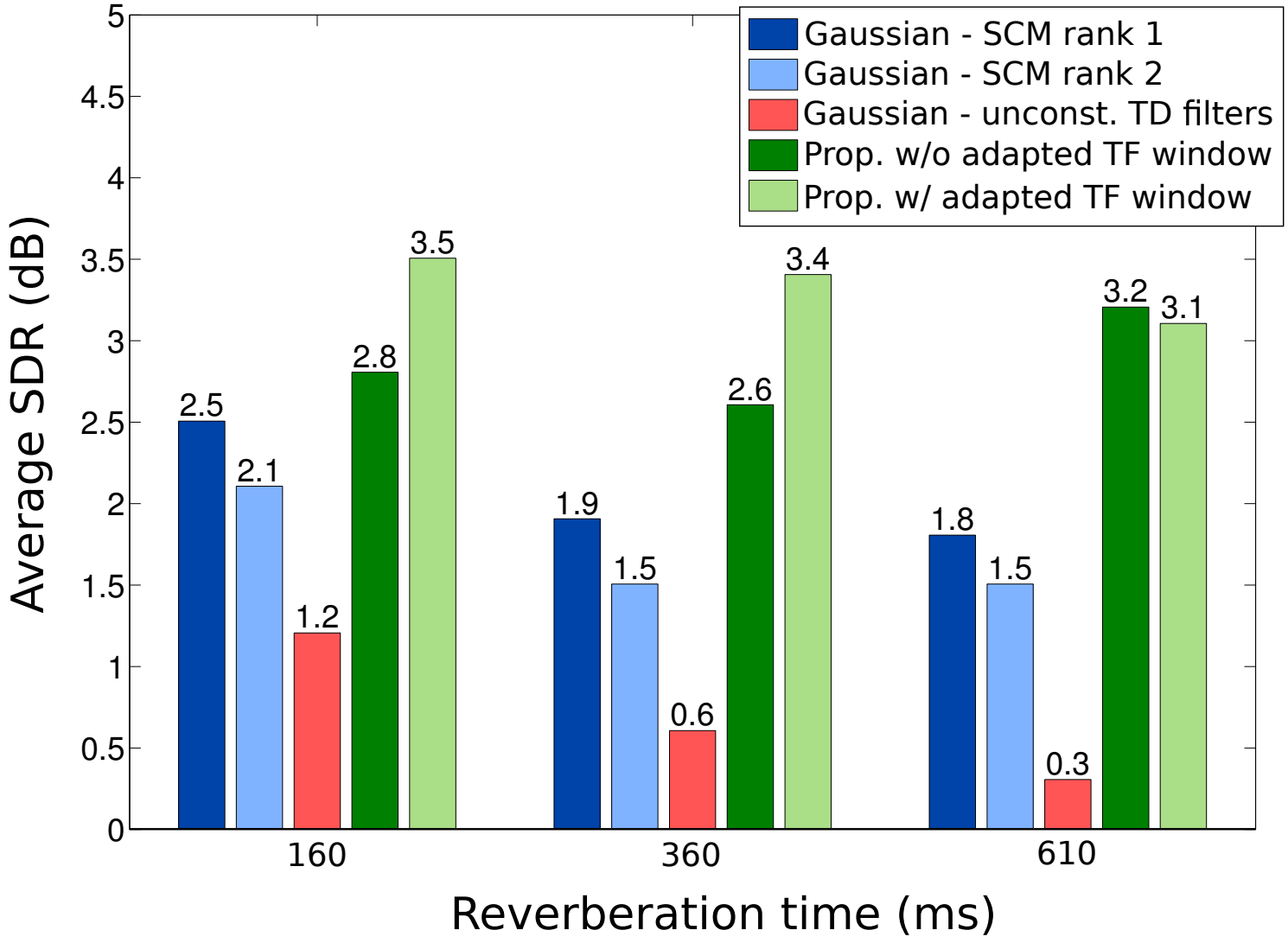
Source separation results



Source separation results

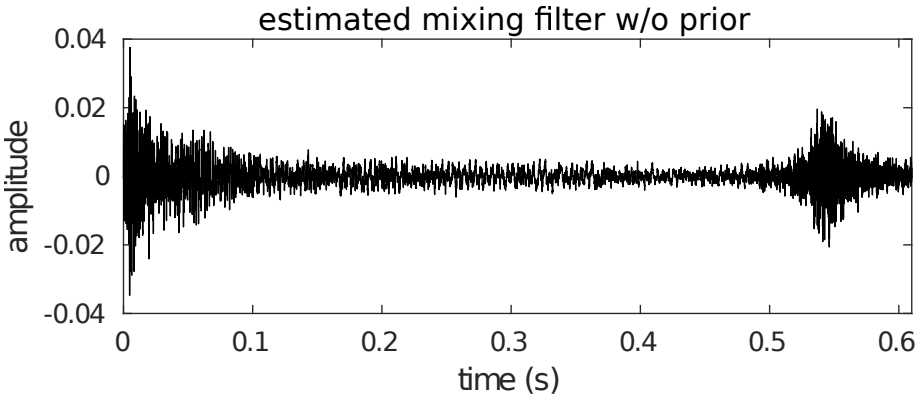


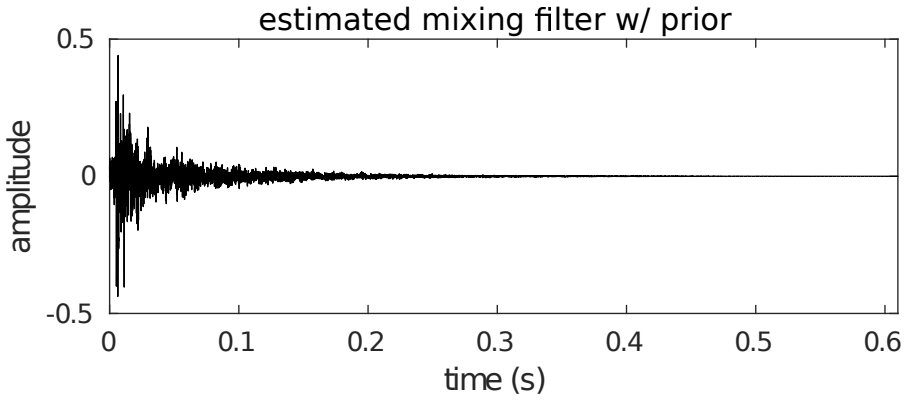
Source separation results



Mixing filters: Influence of the prior

- Convolutive mixture equation: $x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t)$.
- Multiple solutions can explain the same observed data.





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**.

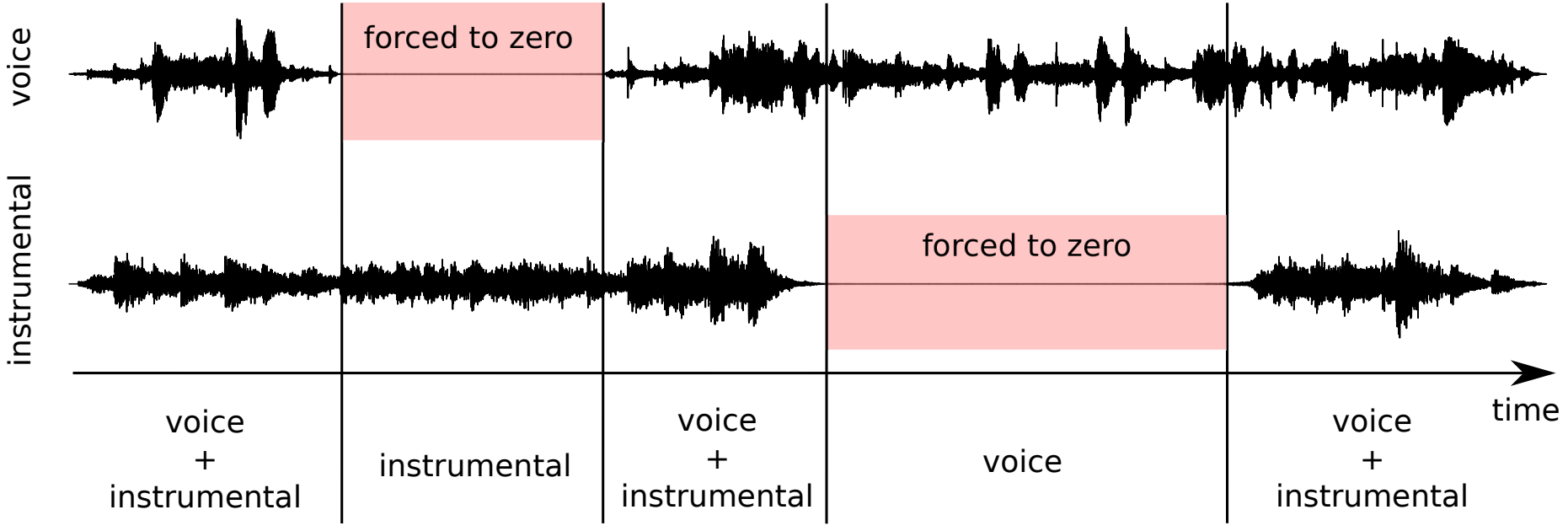
Stereo mixture:

source image	drums	guitar	bass
original			
Gaussian - SCM rank-1			
Gaussian - unconstrained 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).

Outline

Conclusion

Probabilistic modeling of reverberation for audio source separation

---

Guiding the estimation of the mixing filters can help improving source separation results.

- Convolutive mixture model in the STFT domain:
  - Time-domain dynamics of the mixing filters $\rightarrow$ frequency-domain correlations;
  - Limited to weakly reverberant mixtures.

- Convolutive mixture model in the time domain:
  - Appropriate for highly reverberant mixtures;
  - Simple priors on the mixing filters, in the time domain;
  - Multi-resolution source modeling.

Perspectives (1)

Supervised source models:

- Learn NMF dictionaries on an external dataset;
- Neural network model for the source scale parameters [13];
- Variational autoencoder for learning the source prior distributions [14].

Speech enhancement and recognition:

- Robustness to interfering noise and reverberant conditions is still an open issue for automatic speech recognition (ASR);
- In ASR, features are computed from the enhanced signal $\rightarrow$ uncertainty propagation from the speech posterior distribution [15].

[13] A. Nugraha, A. Liutkus and E. Vincent, "Multichannel music separation with deep neural networks", *European Signal Processing Conference (EUSIPCO)*, 2016.

[14] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization", *arXiv:1710.11439 [cs.SD]*, 2017.

[15] K. Adiloğlu and E. Vincent, "Variational Bayesian inference for source separation and robust feature extraction", *IEEE Transactions on Audio, Speech, and Language Processing*, 2016.

Perspectives (2)

Deconvolution:

- Can we recover $s(t)$ from $x(t) = [a \star s](t)$?
- Dictionary learning problem [16]:

$$x(t) = \sum_{f,n} s_{fn} \, [a \star \psi_{fn}](t).$$

Source localization:

- The latent variables $u_{ij}(t)$ are supposed to "encode" the early contributions;
- Can we use features computed from their posterior distribution to perform source localization?
- Could we extend the hierarchical model so that they depend on the source location?

[16] D. Barchiesi and M. Plumbley, "Dictionary learning of convolved signals", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

Publications

International peer-reviewed journals:

- S. L., R. Badeau, G. Richard, "Student's t source and mixing models for multichannel audio source separation", *IEEE Transactions on Audio, Speech and Language Processing*, 2017 (**submitted**).
- S. L., R. Badeau, G. Richard, "Multichannel audio source separation with probabilistic reverberation priors", *IEEE Transactions on Audio, Speech and Language Processing*, 2016.

International peer-reviewed conferences and workshops:

- S. L., R. Badeau, G. Richard, "Separating time-frequency sources from time-domain convolutive mixtures using non-negative matrix factorization", *IEEE WASPAA*, 2017.
- S. L., R. Badeau, G. Richard, "Semi-blind Student's t source separation for multichannel audio convolutive mixtures", *EUSIPCO*, 2017.
- S. L., U. Şimşekli, A. Liutkus, R. Badeau, G. Richard, "Alpha-stable multichannel audio source separation", *IEEE ICASSP*, 2017.
- S. L., R. Badeau, G. Richard, "Multichannel audio source separation: Variational inference of time-frequency sources from time-domain observations", *IEEE ICASSP*, 2017.
- S. L., R. Badeau, G. Richard, "Autoregressive moving average modeling of late reverberation in the frequency domain", *EUSIPCO*, 2016.
- S. L., R. Badeau, G. Richard, "Multichannel audio source separation with probabilistic reverberation modeling", *IEEE WASPAA*, 2015.

Thank you

Audio examples and Matlab code available at:

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