Modeling Reverberant Mixtures for Multichannel Audio-Source Separation

Simon Leglaive

Ph.D. supervisors: Roland Badeau and Gaël Richard

LTCI, Télécom ParisTech, Université Paris Saclay

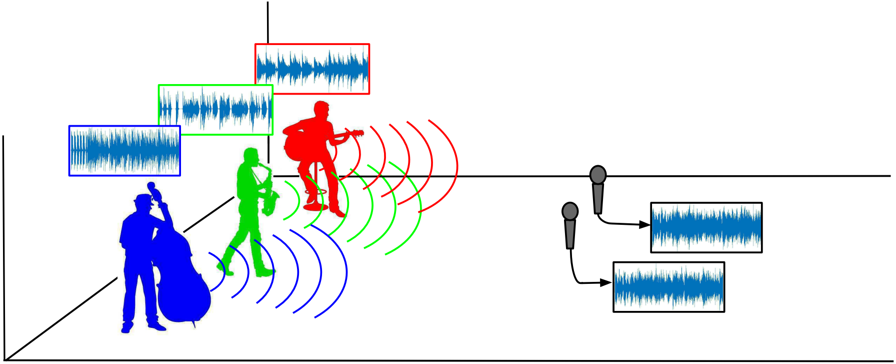
Seminar at PERCEPTION group, INRIA Grenoble Rhône-Alpes,
Montbonnot Saint-Martin

March 22, 2017

Multichannel audio source separation

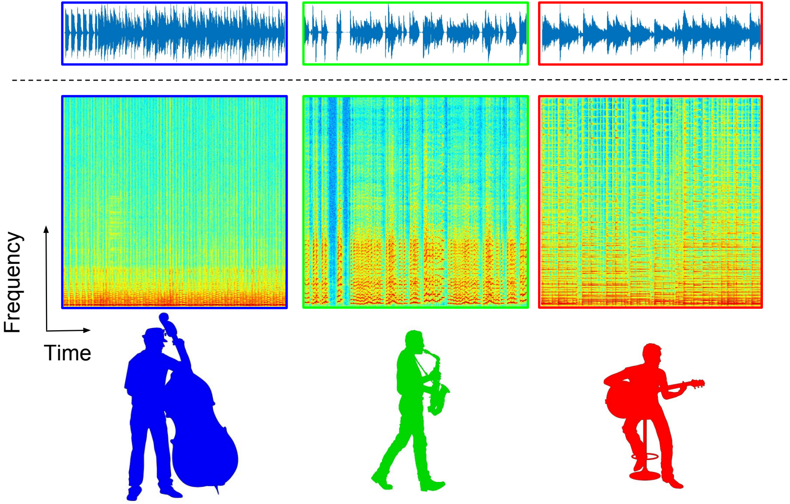
Objective: Recover source signals from the observation of several mixtures.

Context: Under-determined and reverberant.



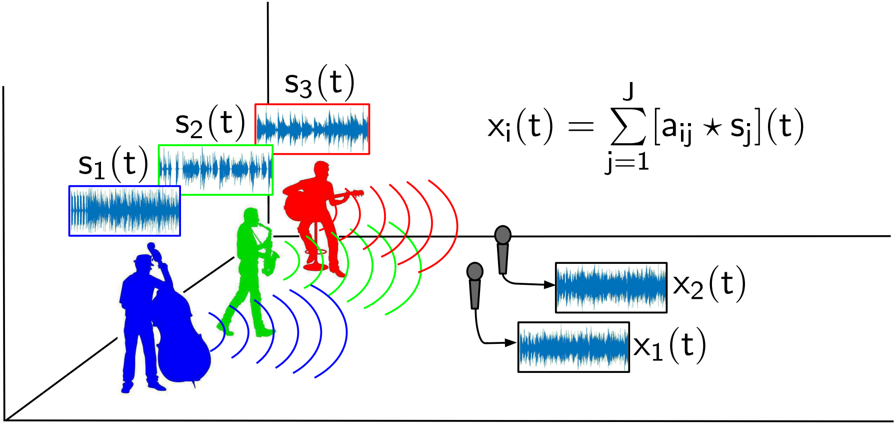
Time-frequency source representation

Time-frequency (TF) transforms provide meaningful representations.



Modeling reverberant mixtures (1)

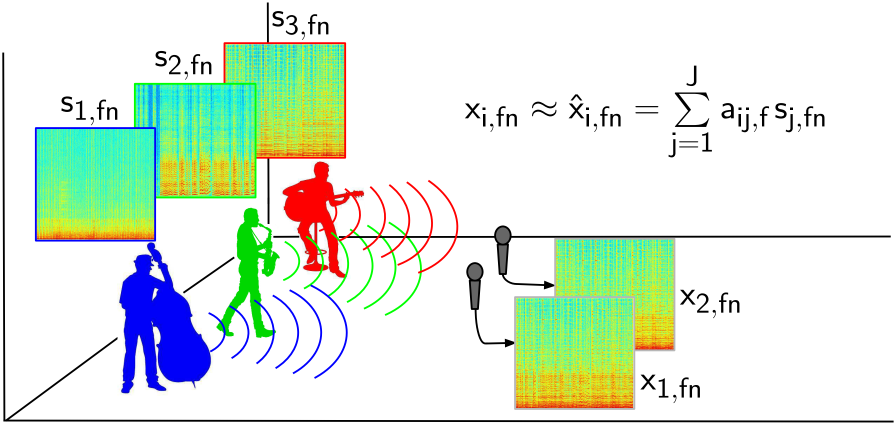
Convolutive model in the time domain:



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

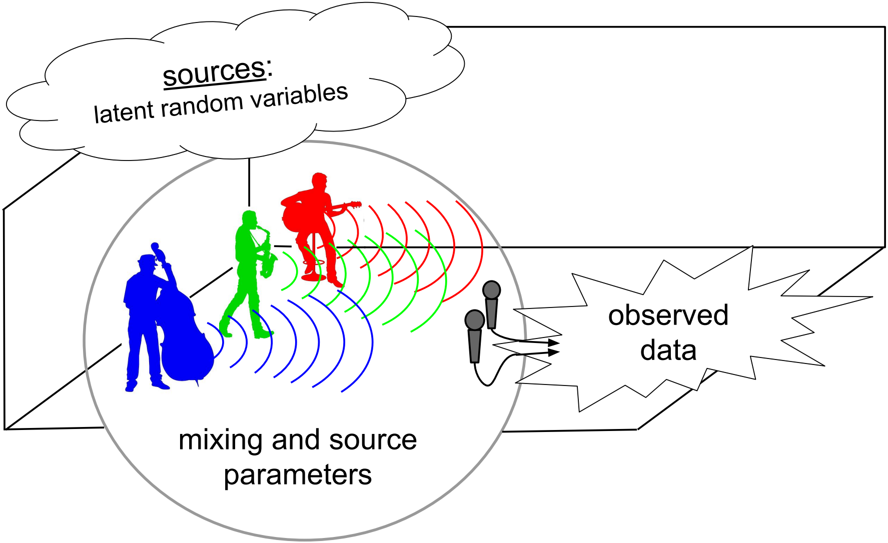
Modeling reverberant mixtures (2)

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



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

Probabilistic framework



Outline

Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Outline

Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

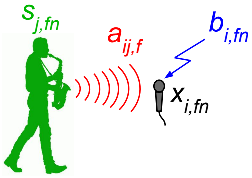
Conclusion

Mixture model

Convolutive noisy mixture of J sources on I channels

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

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



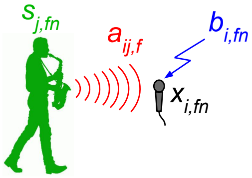
- Frequency response of the mixing filters: $a_{ij,f}$;
- Source STFT coefficients: $s_{j,fn}$;
- Additive Gaussian noise: $b_{i,fn} \sim \mathcal{N}_c(0, \sigma^2_{b,f})$.

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

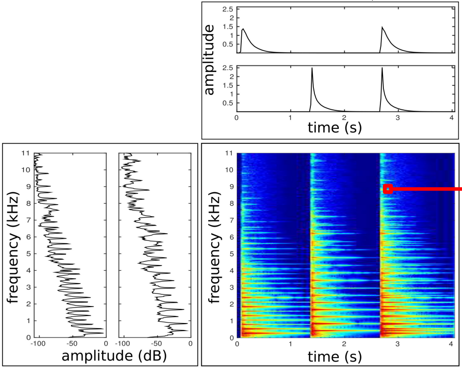
$$\mathbf{x}_{fn} = \mathbf{A}_f \mathbf{s}_{fn} + \mathbf{b}_{fn}, \tag{1}$$

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 [Févotte et al., 2009].

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



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

STFT power spectrogram (dB)

temporal activations

spectral templates

$$\boxed{\mathsf{s}_{\mathsf{j,fn}} \sim \mathcal{N}_{\mathsf{c}}(0, [\mathbf{W}_{\mathsf{j}}\mathbf{H}_{\mathsf{j}}]_{\mathsf{fn}})}$$

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

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

Statistical inference - ML

- 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: $\boldsymbol{\eta} = \left\{ \{\mathbf{W}_j, \mathbf{H}_j\}_j, \left\{\mathbf{A}_f, \sigma^2_{b,f}\right\}_f \right\}$.

Source estimation according to the posterior mean

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\eta}^\star}[\mathbf{s}]$$

Maximum Likelihood parameters estimation

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

$\rightarrow$ Expectation-Maximization (EM) algorithm [Ozerov and Févotte, 2010].

[Ozerov and Févotte, 2010] "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation". *IEEE Transactions on Audio, Speech and Language Processing*.

Statistical inference - MAP

- 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: $\boldsymbol{\eta} = \left\{ \{\mathbf{W}_j, \mathbf{H}_j\}_j, \left\{\mathbf{A}_f, \sigma^2_{b,f}\right\}_f \right\}$.

Source estimation according to the posterior mean

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\eta}^\star}[\mathbf{s}]$$

Maximum A Posteriori parameters estimation

$$\boldsymbol{\eta}^\star = \arg\max_{\boldsymbol{\eta}} p(\mathbf{x}|\boldsymbol{\eta}) p\left(\{a_{ij,f}\}_{i,j,f}\right)$$

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

Outline

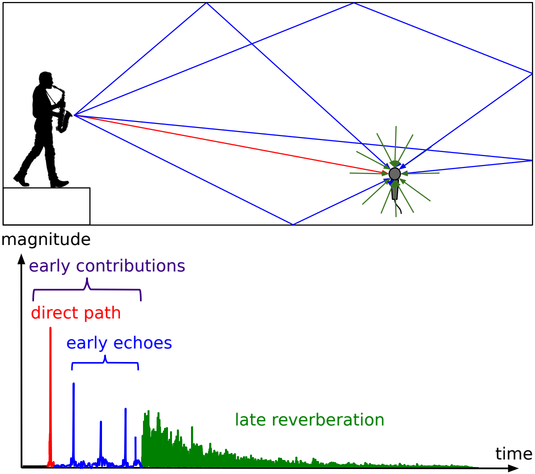
Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Room impulse response

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



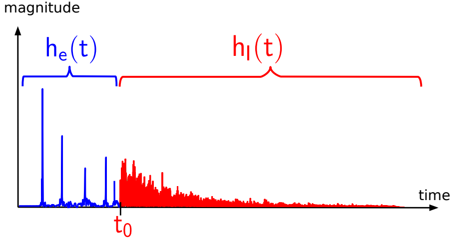
Room impulse and frequency responses

Room impulse and frequency responses

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

$$\underbrace{h(t) = h_e(t) + h_l(t)}_{\text{Room impulse response (RIR)}} \underset{\mathcal{F}_T^{-1}}{\overset{\mathcal{F}_T}{\rightleftarrows}} \underbrace{H(f) = H_e(f) + H_l(f)}_{\text{Room frequency response (RFR)}}$$

$\mathcal{F}_T$: Discrete Fourier Transform (DFT)

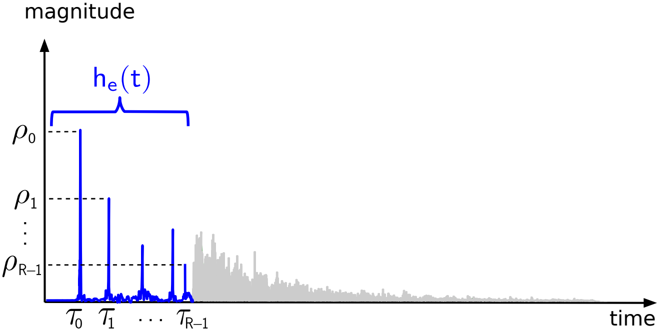


Mixing time:

$t_0 = \lfloor C_0 \sqrt{V} f_s \rfloor$ samples

- $C_0 = 2 \times 10^{-3}$
- V: volume of the room
- f_s: sampling rate

Early contributions model (1)



k-th early contribution: amplitude ρ_k and delay τ_k

$$H_e(f) = \sum_{k=0}^{R-1} \rho_k \delta_k^f \quad \text{with} \quad \delta_k = e^{-j2\pi\tau_k/T}. \tag{2}$$

Early contributions model (2)

It follows (see, e.g., [Kumaresan, 1983])

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

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

Adding an error term $\rightarrow$ autoregressive model

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

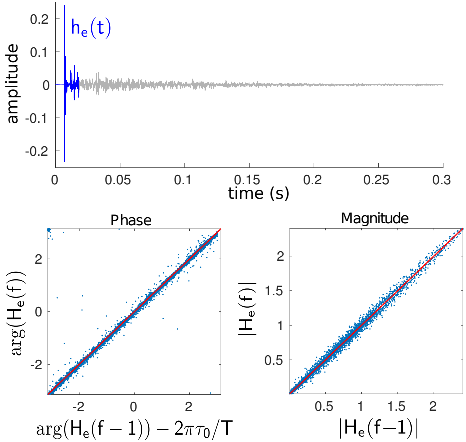
[Kumaresan, 1983] "On the zeros of the linear prediction-error filter for deterministic signals". *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

For more details: [Leglaive et al., 2015] "Multichannel audio source separation with probabilistic reverberation modeling". *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

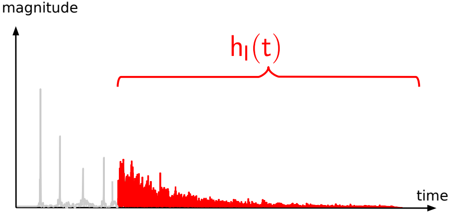
Autoregressive model of order 1

Assuming that the direct path dominates the early echoes:

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



Late reverberation model (time domain)



Power Temporal Profil (PTP) $\rightarrow$ exponential decay

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

- $\tau = \frac{T_{60} f_s}{3\ln(10)}$ samples, with T_{60} the reverberation time in seconds;
- $\mathbb{E}[\cdot]$: spatial averaging.

Late reverberation model (frequency domain)

Statistical room acoustics: $\{H_l(f)\}_f$ is a proper centered and WSS complex Gaussian random process.

Late reverberation model (frequency domain)

Statistical room acoustics: $\{H_l(f)\}_f$ is a proper centered and WSS complex Gaussian random process.

Theoretical Power Spectral Density (PSD)

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

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

Late reverberation model (frequency domain)

Statistical room acoustics: $\{H_l(f)\}_f$ is a proper centered and WSS complex Gaussian random process.

Theoretical Power Spectral Density (PSD)

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

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

Theoretical Autocovariance function (ACVF)

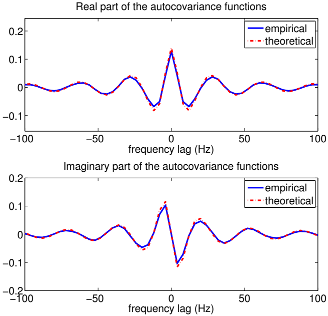
Applying the Wiener-Khinchin theorem we can obtain a theoretical expression of the ACVF:

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

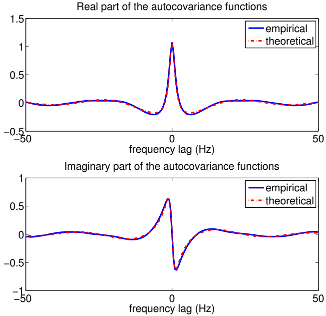
These quantities are theoretically defined according to some room parameters (reverberation time, dimensions).

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 $\{H_l(f)\}_f$ follows an ARMA(P, Q) model:

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

- $\Phi(L) = \sum_{p=0}^{P} \varphi_p^l L^p$ and $\Theta(L) = \sum_{q=0}^{Q} \theta_q L^q$ with $\varphi_0^l = \theta_0 = 1$;
- L is the lag operator, i.e. $LH_l(f) = H_l(f-1)$;
- $\epsilon(f) \sim \mathcal{N}_c(0, \sigma_\epsilon^2)$.

We can compute the ARMA parameters from the theoretical ACVF.

Experimental validation

Same room parameters as used before for simulated RIRs

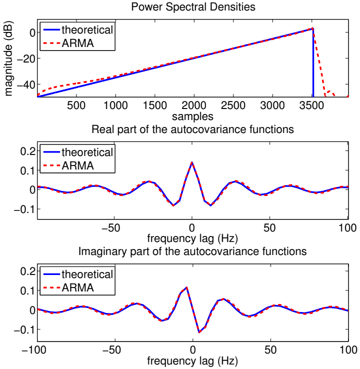


Figure: ARMA(7,2) parametrization

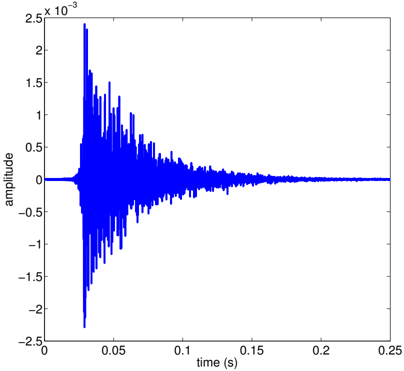


Figure: Synthesized late RIR

For more details: [Leglaive et al., 2016] "Autoregressive moving average modeling of late reverberation in the frequency domain". *European Signal Processing Conference (EUSIPCO)*.

Outline

Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

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

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

M-step

ML estimation:

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

MAP estimation:

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

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

Early reverberation prior

Early reverberation prior

We consider an AR(1) model for the early part of the mixing filters:

$$\ln p(\{\mathbf{A}_{e,f}\}_f) \stackrel{c}{=} -\frac{1}{\sigma_\kappa^2} \sum_{f=1}^{F-1} \left|\left| \mathbf{A}_{e,f} - \mathbf{\Delta} \circ \mathbf{A}_{e,f-1} \right|\right|_F^2, \tag{5}$$

where $\mathbf{\Delta} = [\delta_{ij}]_{ij} \in \mathbb{C}^{I \times J}$, $|| \cdot ||_F^2$ is the Frobenius norm and $\circ$ is the element-wise matrix product.

Hyperparameters

- AR coefficients $\{\delta_{ij}\}_{ij}$: Estimated within the M-step.
- Noise variance σ_κ^2: Expresses how confident we are about the prior (assumed to be fixed).

Late reverberation prior

Late reverberation prior

We consider an ARMA(7,2) model for the late part of the mixing filters:

$$\ln p(\{\mathbf{A}_f^l\}_f) \stackrel{c}{=} -\sum_{f=0}^{F-1} \mathrm{Trace}\left[\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{A}_{l,f}\right)^H \mathbf{\Sigma}_{\epsilon,f}^{-1}\left(\frac{\Phi(L)}{\Theta(L)}\mathbf{A}_{l,f}\right)\right], \qquad (6)$$

where $\mathbf{\Sigma}_{\epsilon,f} = \sigma_\epsilon^2 \mathbf{I}_I$.

Hyperparameters

- ARMA coefficients: **Learned and fixed** from the theoretical ACVF, knowing some room parameters.
- Noise variance σ_ϵ^2: Expresses how confident we are about the prior (assumed to be fixed).

Outline

Convolutive mixture model in the STFT domain

- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain

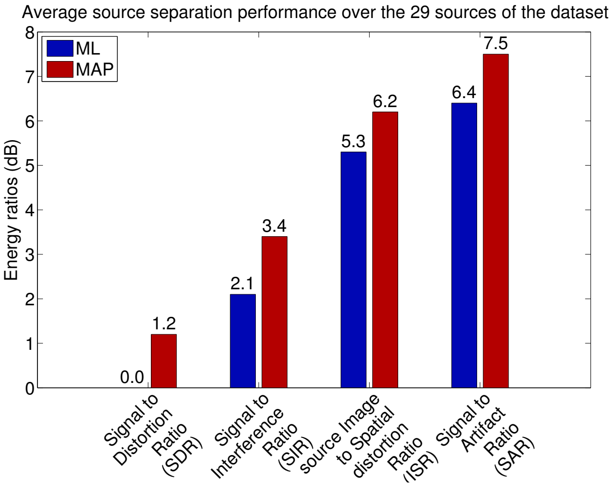
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Experiments

- Dataset composed of 8 stereo mixtures:
  - Created using synthetic room impulse responses;
  - Reverberation time: 128 ms;
  - Duration: 12 to 28 seconds;
  - Number of musical sources: 3 to 5.
- Source separation results: ML (w/o priors) vs. MAP (w/ priors).
- Both algorithms are run from the same blind initialization.

Source separation results



For more details: [Leglaive et al., 2016] "Multichannel audio source separation with probabilistic reverberation priors". *IEEE Transactions on Audio, Speech and Language Processing*.

Audio example

- Separation of 4 sources from a stereo mixture.
- Both algorithms are run from the same blind initialization.

Stereo mixture:

source	drums		guitar 1		guitar 2		voice	
original								
estimated	ML	MAP	ML	MAP	ML	MAP	ML	MAP

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

Outline

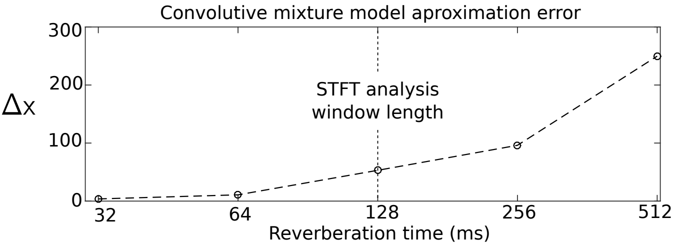
Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Main limitation

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



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

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

Overcoming this limitation

- More accurate time-frequency convolutive mixture models:
  - 2D filtering [Badeau and Plumbley, 2014];
  - Subband filtering (convolutive transfer function) [Li et al., 2017];
- **Time-domain convolutive mixture model** [Kowalski et al., 2010].

[Badeau and Plumbley, 2014] "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*.

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

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

Outline

Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Outline

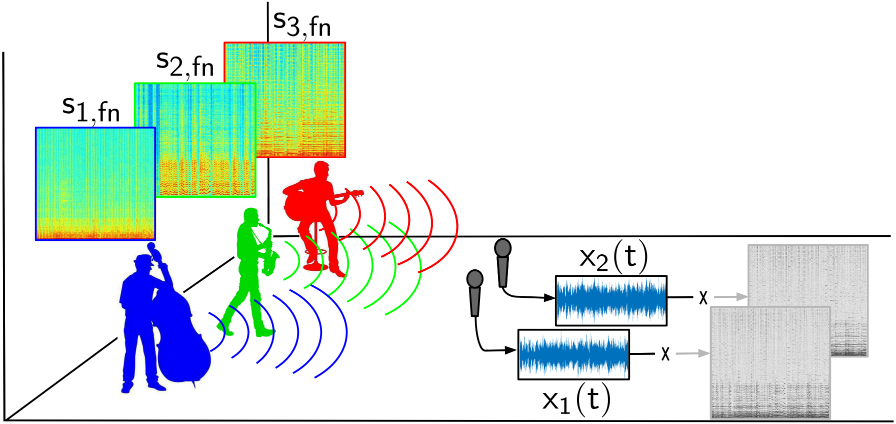
Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Proposed approach

TF source model and time-domain convolutive mixture model.



Mixture model

Time-domain convolutive mixture model

$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t) + b_i(t), \tag{7}$$

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

Time-frequency source representation

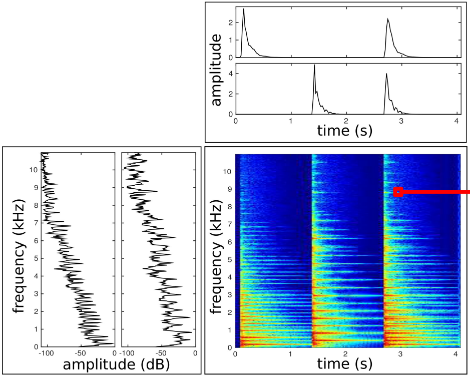
$$s_j(t) = \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} s_{j,fn} \psi_{fn}(t), \tag{8}$$

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

Source model

Gaussian source model based on Non-negative Matrix Factorization.

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



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

MDCT power spectrogram (dB)

temporal activations

spectral templates

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

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

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

Outline

Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Statistical inference

- Latent **time-frequency** random variables: $\mathbf{s} = \{s_{j,fn}\}_{j,f,n}$;
- Observed **time-domain** random variables: $\mathbf{x} = \{x_i(t)\}_{i,t}$;
- Model parameters: $\boldsymbol{\eta} = \big\{\{\mathbf{W}_j, \mathbf{H}_j\}_j, \{a_{ij}(t)\}_{i,j,t}, \{\sigma_i^2\}_i\big\}$.

Source and parameter estimation

- Source estimation according to the posterior mean:

$$\hat{\mathbf{s}} = \mathbb{E}_{\mathbf{s}|\mathbf{x};\boldsymbol{\eta}^\star}[\mathbf{s}].$$

- Maximum likelihood estimation of the parameters:

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

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

Variational inference

- We want to find $q \in \mathcal{F}$ which approximates $p(\mathbf{s}|\mathbf{x};\boldsymbol{\eta})$.

- Taking the KL divergence as a measure of fit, we can show that:

$$KL(q||p(\mathbf{s}|\mathbf{x};\boldsymbol{\eta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\eta})}_{\text{Log-likelihood}} \quad - \quad \underbrace{\mathcal{L}(q;\boldsymbol{\eta})}_{\text{Variational Free Energy}} \quad , \tag{9}$$

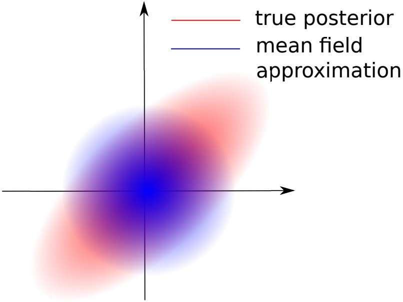
where $\mathcal{L}(q;\boldsymbol{\eta}) = \left\langle \ln\left(\frac{p(\mathbf{x},\mathbf{s};\boldsymbol{\eta})}{q(\mathbf{s})}\right)\right\rangle_q$ and $\langle f(\mathbf{s})\rangle_q = \int f(\mathbf{s})q(\mathbf{s})d\mathbf{s}$.

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

Mean-field approximation

$\mathcal{F}$: set of pdfs over $\mathbf{s}$ 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:

$$q^{\star}_{jfn}(s_{j,fn}) = \arg\max_{q_{jfn}} \mathcal{L}(q; \boldsymbol{\eta}_{\text{old}}) = N(s_{j,fn}; m_{j,fn}, \gamma_{j,fn}).$$

source estimate

M-Step

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

NMF parameters

Compute an NMF with the Itakura-Saito divergence on:

$$\left\langle s_{j,fn}^2 \right\rangle_{q^\star} = m_{j,fn}^2 + \gamma_{j,fn},$$

$\rightarrow$ standard multiplicative update rules.

Mixing filters

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

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

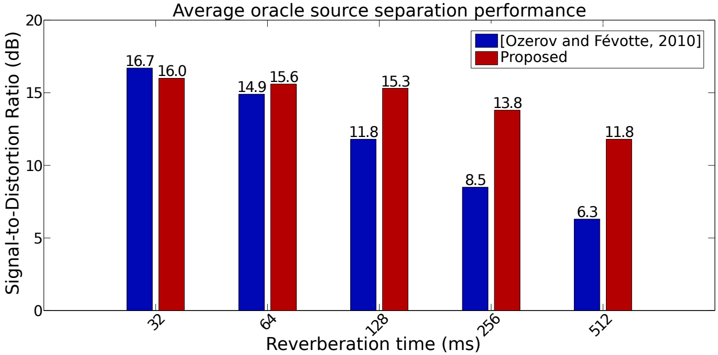
Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Oracle experiment

- **Dataset**: Same as before with different reverberation times;
- **Oracle initialization** of the parameters;
- STFT and MDCT analysis/synthesis window length: 128 ms.



For more details: [Leglaive et al., 2017] "Multichannel audio source separation: variational inference of time-frequency sources from time-domain observations". *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*.

Semi-blind experiment

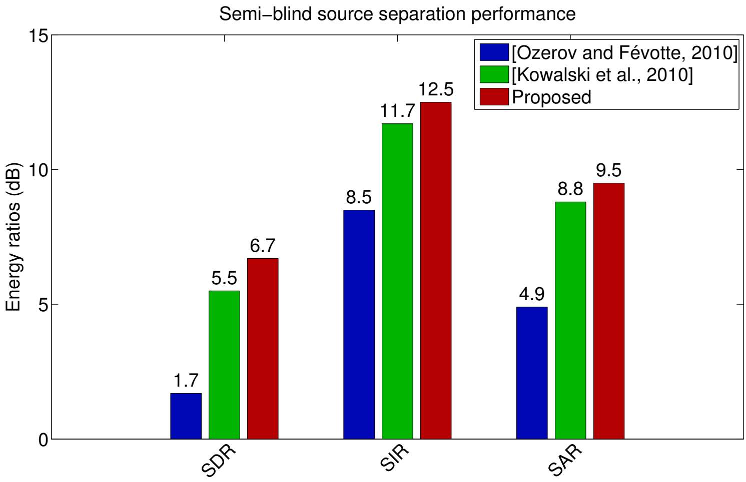
- Mixing filters are known and fixed;
- All other parameters are blindly estimated;
- Compared methods:

		Convolutive mixture model	
		Exact (time)	Approximate (TF)
Source model (TF)	Sparse (ℓ_1)	[Kowalski et al., 2010]	-
	Gaussian NMF-based	Proposed	[Ozerov and Févotte, 2010]

- Dataset: Same as before with a reverberation time of 256 ms.

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

Semi-blind experiment results



Semi-blind audio example

Stereo mixture:

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

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

Outline

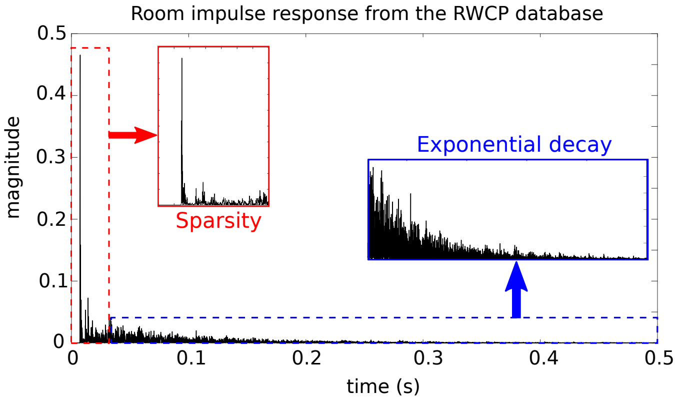
Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Ongoing work

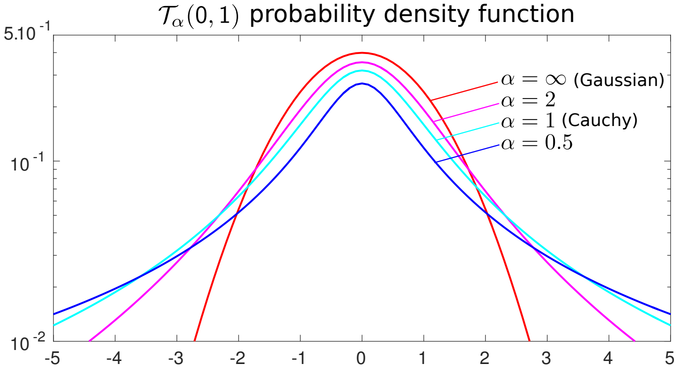
Probabilistic priors on the mixing filters **in the time domain**.



Student's t distribution

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

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



Outline

Convolutive mixture model in the STFT domain
- Baseline source separation framework
- Room frequency response modeling
- Source separation with reverberation priors
- Experiments
- Limitations

Convolutive mixture model in the time domain
- Model
- Inference
- Experiments
- Ongoing work

Conclusion

Conclusion

Convolutive mixture model in the STFT domain:

- Time-domain dynamics of the mixing filter $\rightarrow$ frequency-domain correlations:
  - Early reverberation: AR model;
  - Late reverberation: ARMA model.
- This approach is limited to low reverberation conditions.

Convolutive mixture model in the time domain:

- Accurate for long reverberation times;
- Good separation quality in a semi-blind setting;
- Suitable for incorporating simple priors on the mixing filters, in the time domain.

Thank you

More audio examples and Matlab code available at:

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