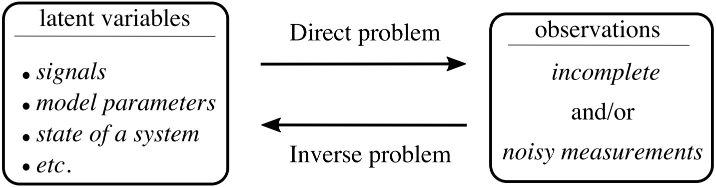
Audio Signal Modeling for Source Separation

From Hand-designed to Learned Probabilistic Priors

Simon Leglaive

Inria Grenoble Rhône-Alpes
Perception Team

Introduction

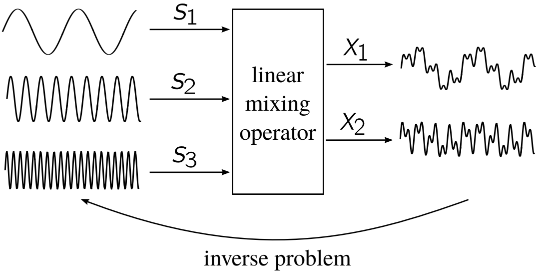


Many applications:

- ▷ Biomedical signal processing and imaging;
- ▷ Astrophysics;
- ▷ Underwater acoustics;
- ▷ Audio signal processing;
- ▷ etc.

Audio source separation

Objective: Recover source signals from one or multiple mixtures.



Usually an ill-posed inverse problem (in Hadamard sense).

Bayesian vs. deep learning approach

Bayesian approach:

- Need for prior knowledge (physically inspired, signal model, etc.).
- Solving the inverse problem: Posterior computation.
- Flexible.

[1] `https://sisec18.unmix.app`

Bayesian vs. deep learning approach

Bayesian approach:

- Need for prior knowledge (physically inspired, signal model, etc.).
- Solving the inverse problem: Posterior computation.
- Flexible.

Discriminative deep learning approach:

- Need for training data.
- Solving the inverse problem: Mapping $(x_1, x_2) \xrightarrow{\text{DNN}} (s_1, s_2, s_3)$.
- State-of-the-art[1].
- Not flexible (e.g. retrain if microphone added or SNR changed).

[1] `https://sisec18.unmix.app`

Bayesian vs. deep learning approach

Bayesian approach:

- Need for prior knowledge (physically inspired, signal model, etc.).
- Solving the inverse problem: Posterior computation.
- Flexible.

Discriminative deep learning approach:

- Need for training data.
- Solving the inverse problem: Mapping $(x_1, x_2) \xrightarrow{\text{DNN}} (s_1, s_2, s_3)$.
- State-of-the-art[1].
- Not flexible (e.g. retrain if microphone added or SNR changed).

Best of both worlds: Deep-learning-based generative models as priors.

[1] `https://sisec18.unmix.app`

Outline

1. **Hand-designed priors** for multichannel and reverberant audio source separation

   Joint work with Roland Badeau and Gaël Richard (Leglaive et al. 2018b)

2. **Deep-learning-based priors** for single-channel speech enhancement

   Joint work with Laurent Girin and Radu Horaud (Leglaive et al. 2018a)

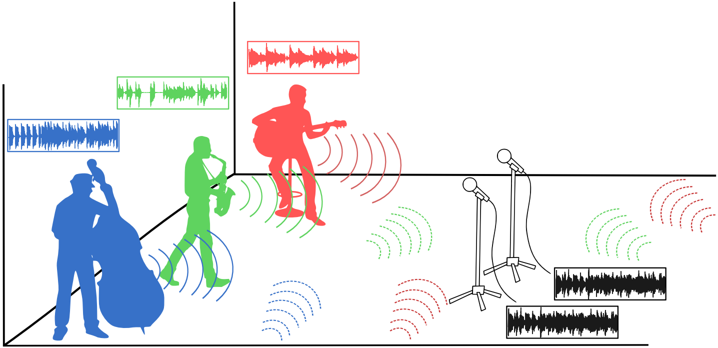
Hand-designed priors for multichannel and reverberant audio source separation

Hand-designed priors for multichannel and reverberant audio source separation

Introduction

Targeted scenario

Under-determined and reverberant multichannel mixture.

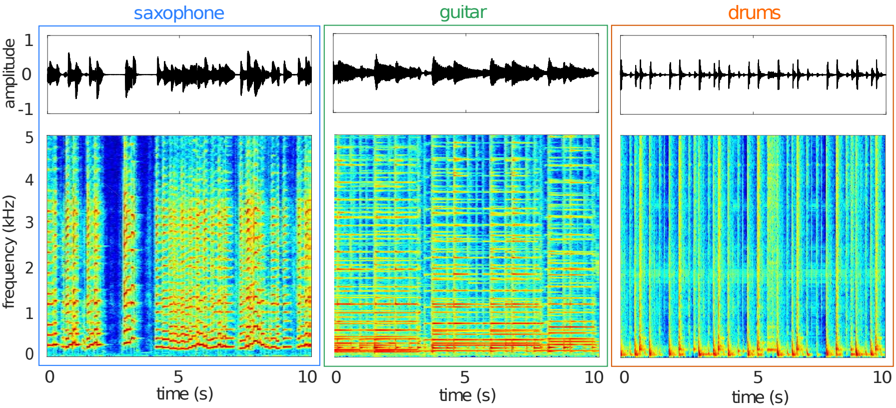


Two-step modeling approach:

- ▷ Source modeling;
- ▷ Mixing process modeling.

Time-frequency source representation

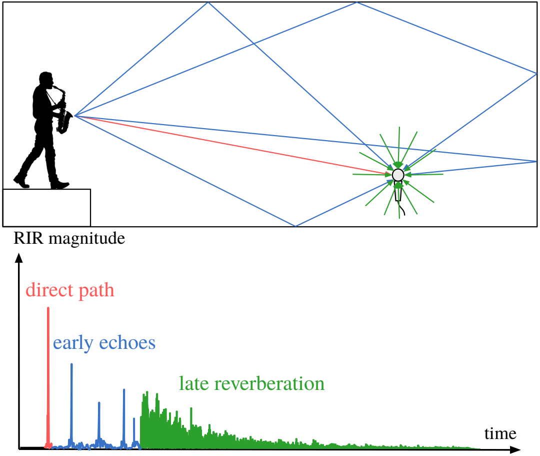
A time-frequency (TF) transform provides a meaningful representation.



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

Room impulse response (RIR)

recorded signal = source signal $\star$ room impulse response



Finite impulse response whose length equals the reverberation time.

Multichannel reverberant mixtures

Convolutive mixture in the time domain

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

for all $i \in \{1, ..., I\}$, $t \in \{0, ..., T-1\}$.

Ph.D. research problem: *Taking priors over the mixing filters into account.*

Multichannel reverberant mixtures

Convolutive mixture in the time domain

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

for all $i \in \{1, ..., I\}$, $t \in \{0, ..., T-1\}$.

Convolutive mixture in the STFT domain

$$x_{i,fn} \approx \sum_{j=1}^{J} a_{ij,f}\, s_{j,fn}, \tag{2}$$

for all $(f, n) \in \mathbb{B} = \{0, ..., F-1\} \times \{0, ..., N-1\}$.

Ph.D. research problem: *Taking priors over the mixing filters into account.*

Mixed time domain / time-frequency domain modeling

Time-domain convolutive mixture model

$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t). \tag{3}$$

Related to (Kowalski et al. 2010; Févotte and Kowalski 2014)

Mixed time domain / time-frequency domain modeling

Time-domain convolutive mixture model

$$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t). \tag{3}$$

Time-frequency synthesis source representation

$$s_j(t) = \sum_{(f,n) \in \mathbb{B}_j} s_{j,fn} \psi_{j,fn}(t). \tag{4}$$

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

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

Related to (Kowalski et al. 2010; Févotte and Kowalski 2014)

Probabilistic modeling with latent variables

- **Latent** time-frequency source coefficients: $\mathbf{s} = \{s_{j,fn} \in \mathbb{R}\}_{j,f,n}$
- **Latent** time-domain mixing filters: $\mathbf{a} = \{a_{ij}(t) \in \mathbb{R}\}_{i,j,t}$
- **Observed** time-domain mixture coefficients: $\mathbf{x} = \{x_i(t) \in \mathbb{R}\}_{i,t}$

Defining the probabilistic model

$$p\left(\mathbf{x}, \mathbf{s}, \mathbf{a}; \boldsymbol{\theta}\right) = p(\mathbf{s}; \boldsymbol{\theta}_s) \times p(\mathbf{a}; \boldsymbol{\theta}_a) \times p(\mathbf{x}|\mathbf{s}, \mathbf{a}; \boldsymbol{\theta}_m)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_s, \boldsymbol{\theta}_a, \boldsymbol{\theta}_m\}$ is a set of deterministic model parameters.

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

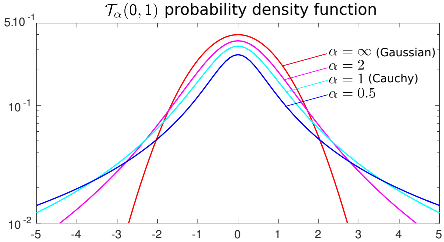
Hand-designed priors for multichannel and reverberant audio source separation

Model

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

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

Student's t NMF source model (Yoshii et al. 2016)

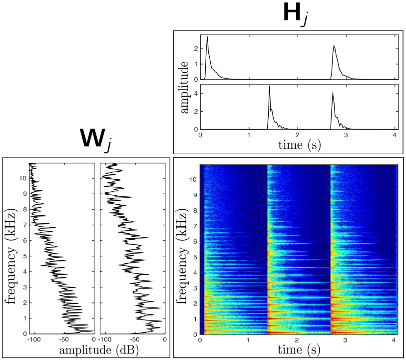
Student's t source model with non-negative matrix factorization (NMF).

Independently for all j, f, n:

$$s_{j,fn} \sim \mathcal{T}_{\alpha_v}\left(0, (\mathbf{W}_j\mathbf{H}_j)_{f,n}^{\frac{1}{2}}\right), \quad (5)$$

where

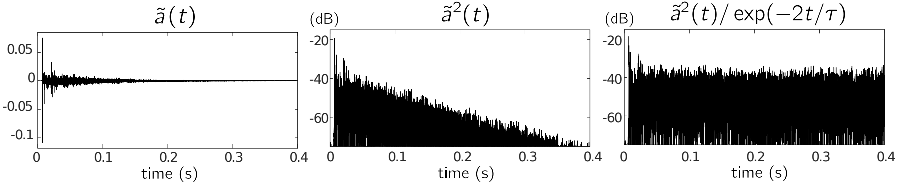
- $\mathbf{W}_j \in \mathbb{R}_+^{F_j \times K_j}$;
- $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N_j}$;
- K_j is the factorization rank.



MDCT power spectrogram (dB)

Related to (Benaroya et al. 2003; Févotte et al. 2009), among many other works.

Gaussian RIR model (Polack 1988)

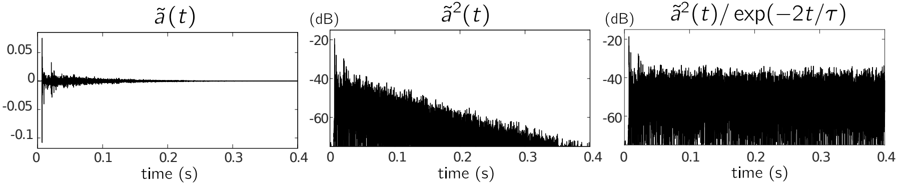


Gaussian model with exponential decay

Independently for all microphones i, sources j and time instants t:

$$a_{ij}(t) \sim \mathcal{N}\left(0, r^2(t)\right), \qquad r^2(t) = \sigma_r^2 \exp(-2t/\tau), \tag{6}$$

where τ is defined according to the reverberation time.



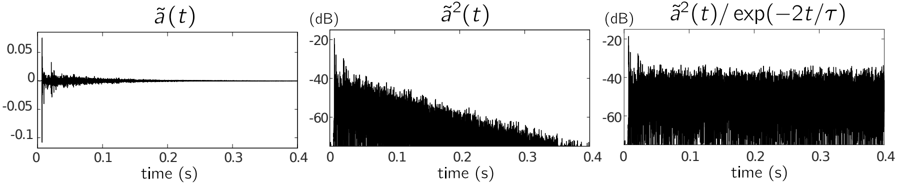
Student's t model with exponential decay

Independently for all microphones i, sources j and time instants t:

$$a_{ij}(t) \sim \mathcal{T}_{\alpha_u}(0, r(t)), \qquad r(t) = \sigma_r \exp(-t/\tau). \tag{7}$$

<u>Remark</u>: Generalization of the previous Gaussian model.

Student's t RIR model (2)



Student's t model with exponential decay

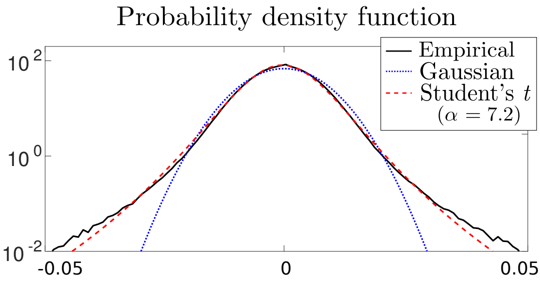
Equivalently:

$$a_{ij}(t)/\exp(-t/\tau) \overset{i.i.d}{\sim} \mathcal{T}_{\alpha_u}(0, \sigma_r). \tag{8}$$

Remark: Generalization of the previous Gaussian model.

Experimental validation

- 624 RIRs from the MIRD database (Hadad et al. 2014);
- Reverberation time equals 610 ms.
- Empirical distribution of the normalized RIR coefficients.



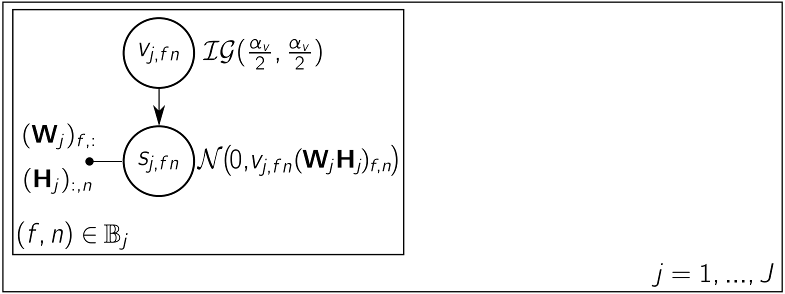
Conditional mixture distribution

Independently for all microphones i and time instants t:

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

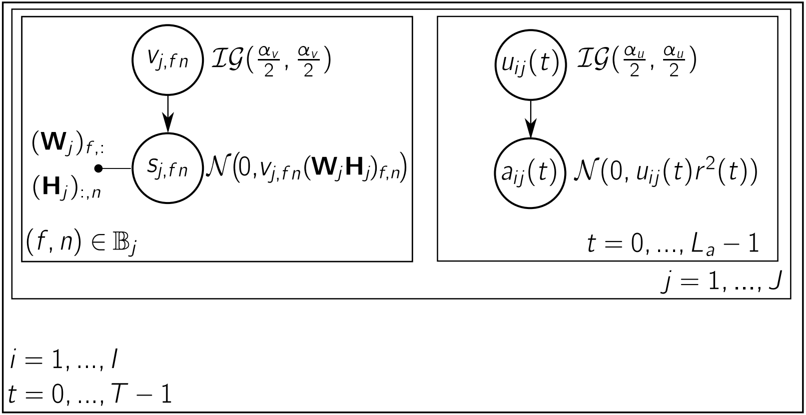
where we recall that $s_j(t) = \sum_{(f,n) \in \mathbb{B}_j} s_{j,fn} \psi_{j,fn}(t)$.

Bayesian network

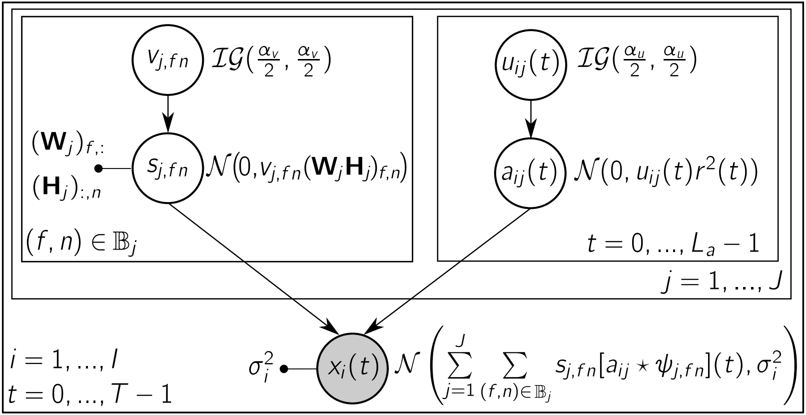


- $\mathbf{z}$: set of all latent variables (empty circles);
- $\mathbf{x}$: set of observations (shaded circles);
- $\boldsymbol{\theta}$: set of model parameters to be estimated (dots).

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).



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

Hand-designed priors for multichannel and reverberant audio source separation

Inference

Variational inference with the mean field approximation

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

Variational inference with the mean field approximation

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

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

$$D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-marginal likelihood}} - \underbrace{\mathcal{L}\left(q(\mathbf{z});\boldsymbol{\theta}\right)}_{\text{variational free energy}} \geq 0, \tag{9}$$

where $\mathcal{L}\left(q(\mathbf{z});\boldsymbol{\theta}\right) = \mathbb{E}_{q(\mathbf{z})}\left[\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \ln q(\mathbf{z})\right]$.

Variational inference with the mean field approximation

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

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

$$D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-marginal likelihood}} - \underbrace{\mathcal{L}\left(q(\mathbf{z});\boldsymbol{\theta}\right)}_{\text{variational free energy}} \geq 0, \tag{9}$$

  where $\mathcal{L}\left(q(\mathbf{z});\boldsymbol{\theta}\right) = \mathbb{E}_{q(\mathbf{z})}\left[\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \ln q(\mathbf{z})\right]$.

- Variational expectation-maximization algorithm:
  - **E-step**: $q^\star(\mathbf{z}) = \arg\max_{q(\mathbf{z})\in\mathcal{F}} \mathcal{L}(q(\mathbf{z});\boldsymbol{\theta}^\star)$
  - **M-step**: $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star(\mathbf{z});\boldsymbol{\theta})$

Variational inference with the mean field approximation

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

$$D_{\mathrm{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})) = \underbrace{\ln p(\mathbf{x};\boldsymbol{\theta})}_{\text{log-marginal likelihood}} - \underbrace{\mathcal{L}(q(\mathbf{z});\boldsymbol{\theta})}_{\text{variational free energy}} \geq 0, \tag{9}$$

where $\mathcal{L}(q(\mathbf{z});\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z})}[\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) - \ln q(\mathbf{z})]$.

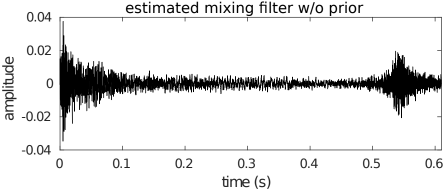
- Variational expectation-maximization algorithm:
  - **E-step**: $q^\star(\mathbf{z}) = \arg\max_{q(\mathbf{z})\in\mathcal{F}} \mathcal{L}(q(\mathbf{z});\boldsymbol{\theta}^\star)$
  - **M-step**: $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q^\star(\mathbf{z});\boldsymbol{\theta})$
- $\mathcal{F}$ is the set of pdfs that can be factorized as $q(\mathbf{z}) = \prod_{z_k\in\mathbf{z}} q_k(z_k)$.

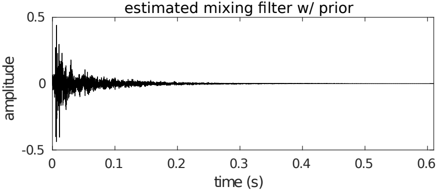
Hand-designed priors for multichannel and reverberant audio source separation

Qualitative experimental results

Necessity of the prior for the mixing filters

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



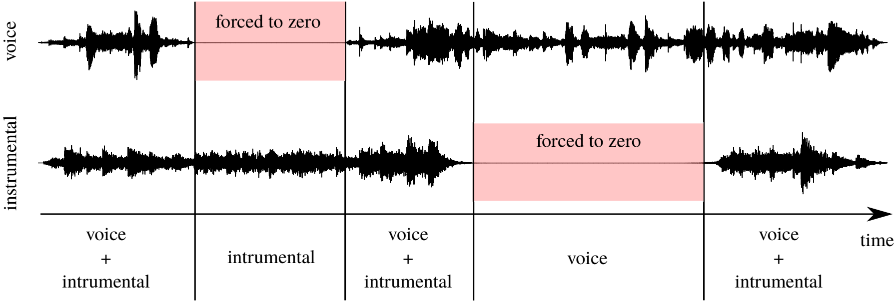


original filter	w/o prior	w/ prior

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

Blind audio source separation example

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

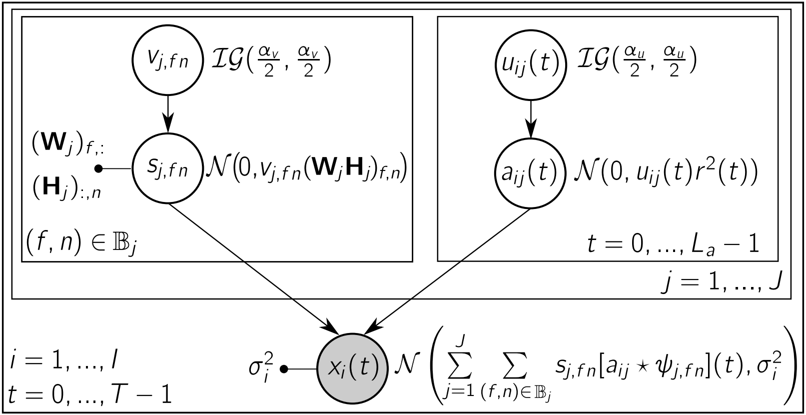


Song: "C'est magnifique" by Ella Fitzgerald (Nice Jazz Festival 1972 - Recording: ORTF).

Hand-designed priors for multichannel and reverberant audio source separation

Conclusion

Postdoc research problem



How can we include neural networks in such probabilistic models?

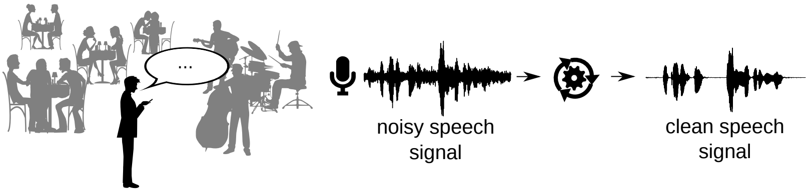
Deep-learning-based priors for single-channel speech enhancement

Deep-learning-based priors for single-channel speech enhancement

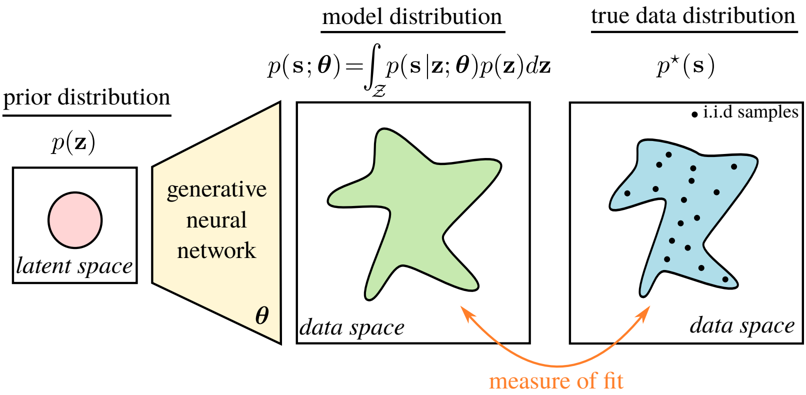
Introduction

Objective

- ▷ Learn a generative speech model directly from the data.
- ▷ Speaker-independent model.
- ▷ Deep learning approach.
- ▷ Application: semi-supervised speech enhancement.



Deep-learning-based generative models



Examples:

- Variational autoencoders (Kingma and Welling 2014);
- Generative adversarial networks (Goodfellow et al. 2014);
- etc.

Deep-learning-based priors for single-channel speech enhancement

Speech model

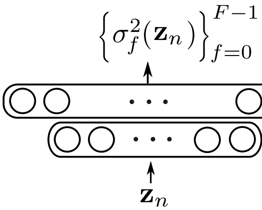
Deep-learning-based generative speech model

In the STFT domain, independently for all $(f, n) \in \mathbb{B}$, we define:

$$\mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L) \tag{10}$$

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c(0, \sigma_f^2(\mathbf{z}_n)), \tag{11}$$

where $s_{fn} \in \mathbb{C}$ and $\mathbf{z}_n \in \mathbb{R}^L$ is a low-dimensional latent random vector.



Generative network

We denote by $\boldsymbol{\theta}_s$ the weights and the biases.

Related to (Kingma and Welling 2014; Bando et al. 2018)

Relation to supervised Itakura-Saito NMF (Févotte et al. 2009)

NMF-based variance parametrization: $s_{fn} \sim \mathcal{N}_c(0, (\mathbf{WH})_{f,n} = \mathbf{w}_f^\top \mathbf{h}_n)$

- **Training time**: Learn $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ from clean signals.
- **Test time**: Estimate $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ from the noisy observations.
- The variance is a linear function of $\mathbf{h}_n \in \mathbb{R}_+^{K}$ (low-dimensional).
- Interpretable / linear / constrained number of trainable parameters.

Relation to supervised Itakura-Saito NMF (Févotte et al. 2009)

NMF-based variance parametrization: $s_{fn} \sim \mathcal{N}_c(0, (\mathbf{WH})_{f,n} = \mathbf{w}_f^\top \mathbf{h}_n)$

- **Training time**: Learn $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ from clean signals.
- **Test time**: Estimate $\mathbf{H} \in \mathbb{R}_+^{K \times N}$ from the noisy observations.
- The variance is a linear function of $\mathbf{h}_n \in \mathbb{R}_+^K$ (low-dimensional).
- Interpretable / linear / constrained number of trainable parameters.

Deep-learning-based variance parametrization: $s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c(0, \sigma_f^2(\mathbf{z}_n))$.

- **Training time**: Learn the neural network parameters $\boldsymbol{\theta}_s$.
- **Test time**: Estimate the posterior distribution of $\mathbf{z} = \{\mathbf{z}_n\}_{n=0}^{N-1}$
- The variance is a non-linear function of $\mathbf{z}_n \in \mathbb{R}^L$ (low-dimensional).
- Interpretable / non-linear / free number of trainable parameters.

Learning the model parameters with variational autoencoders

- **Training dataset** of STFT speech time frames: $\mathbf{s} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$.

- **Problem**:
  - Learn the parameters $\boldsymbol{\theta}_s$ of the generative model.
  - Intractable likelihood $p(\mathbf{s}; \boldsymbol{\theta}_s)$.

- **Solution**: Variational autoencoders (Kingma and Welling 2014).

Learning the model parameters with variational autoencoders

- **Training dataset** of STFT speech time frames: $\mathbf{s} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$.
- **Problem**:
  - Learn the parameters $\boldsymbol{\theta}_s$ of the generative model.
  - Intractable likelihood $p(\mathbf{s}; \boldsymbol{\theta}_s)$.
- **Solution**: Variational autoencoders (Kingma and Welling 2014).
- Variational free energy $\mathcal{L}(\phi, \boldsymbol{\theta}_s) \leq \ln p(\mathbf{s}; \boldsymbol{\theta}_s)$:

$$\mathcal{L}(\phi, \boldsymbol{\theta}_s) = \mathbb{E}_{q(\mathbf{z}|\mathbf{s};\phi)}\left[\ln p\left(\mathbf{s}|\mathbf{z}; \boldsymbol{\theta}_s\right)\right] - D_{\mathrm{KL}}\left(q(\mathbf{z}|\mathbf{s}; \phi) \parallel p(\mathbf{z})\right), \quad (12)$$

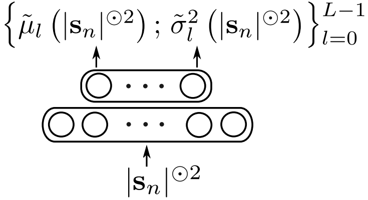
where $\mathbf{z} = \{\mathbf{z}_n\}_{n=0}^{N-1}$ and $q(\mathbf{z}|\mathbf{s}; \phi)$ is an approximation of $p(\mathbf{z}|\mathbf{s}; \boldsymbol{\theta}_s)$.

Variational distribution

$q(\mathbf{z} \mid \mathbf{s}; \phi)$ is defined independently for all time frames $n \in \{0, ..., N-1\}$ and all latent dimensions $l \in \{0, ..., L-1\}$ by:

$$(\mathbf{z}_n)_l \mid \mathbf{s}_n \sim \mathcal{N}\left(\tilde{\mu}_l\left(|\mathbf{s}_n|^{\odot 2}\right), \tilde{\sigma}_l^2\left(|\mathbf{s}_n|^{\odot 2}\right)\right), \tag{13}$$

where $\odot$ denotes element-wise exponentiation;



Recognition network

ϕ denotes the weights and the biases.

Variational free energy

$$\mathcal{L}\left(\boldsymbol{\theta}_s, \phi\right) \stackrel{c}{=} -\sum_{f=0}^{F-1} \sum_{n=0}^{N_{tr}-1} \mathbb{E}_{q(\mathbf{z}_n|\mathbf{s}_n;\phi)} \left[d_{IS}\left(\left|s_{fn}\right|^2 ; \sigma_f^2(\mathbf{z}_n) \right) \right]$$
$$+\frac{1}{2} \sum_{l=1}^{L} \sum_{n=0}^{N_{tr}-1} \left[\ln \tilde{\sigma}_l^2\left(|\mathbf{s}_n|^{\odot 2} \right) - \tilde{\mu}_l\left(|\mathbf{s}_n|^{\odot 2} \right)^2 - \tilde{\sigma}_l^2\left(|\mathbf{s}_n|^{\odot 2} \right) \right], \tag{14}$$

where $d_{IS}(x; y) = x/y - \ln(x/y) - 1$ is the Itakura-Saito (IS) divergence.

- Intractable expectation approximated by a sample average ("reparametrization trick" (Kingma and Welling 2014)).
- Differentiable with respect to both $\boldsymbol{\theta}_s$ and ϕ (backpropagation).
- Optimized using gradient-ascent-based algorithm.

Deep-learning-based priors for single-channel speech enhancement

Speech enhancement

Speech enhancement problem

▷ **Mixture model**: For all $(f, n) \in \mathbb{B}$,

$$x_{fn} = \sqrt{g_n} s_{fn} + b_{fn}, \tag{15}$$

where $g_n \in \mathbb{R}_+$ is a gain parameter.

Speech enhancement problem

▷ **Mixture model**: For all $(f, n) \in \mathbb{B}$,

$$x_{fn} = \sqrt{g_n} s_{fn} + b_{fn}, \tag{15}$$

where $g_n \in \mathbb{R}_+$ is a gain parameter.

▷ **Supervised**[2] **speech model**: Independently for all $(f, n) \in \mathbb{B}$,

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c(0, \sigma_f^2(\mathbf{z}_n)), \qquad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L). \tag{16}$$

---

[2] "supervised": parameters are learned beforehand (Smaragdis et al. 2007).

Speech enhancement problem

▷ **Mixture model**: For all $(f, n) \in \mathbb{B}$,

$$x_{fn} = \sqrt{g_n} s_{fn} + b_{fn}, \tag{15}$$

where $g_n \in \mathbb{R}_+$ is a gain parameter.

▷ **Supervised**[2] **speech model**: Independently for all $(f, n) \in \mathbb{B}$,

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c(0, \sigma_f^2(\mathbf{z}_n)), \qquad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L). \tag{16}$$

▷ **Unsupervised noise model**: Independently for all $(f, n) \in \mathbb{B}$,

$$b_{fn} \sim \mathcal{N}_c\left(0, (\mathbf{W}_b \mathbf{H}_b)_{f,n}\right), \tag{17}$$

where $\mathbf{W}_b \in \mathbb{R}_+^{F \times K_b}$ and $\mathbf{H}_b \in \mathbb{R}_+^{K_b \times N}$.

[2] "supervised": parameters are learned beforehand (Smaragdis et al. 2007).

Parameters estimation

- Unsupervised model parameters:

$$\boldsymbol{\theta}_u = \left\{ \mathbf{W}_b \in \mathbb{R}_+^{F \times K_b},\, \mathbf{H}_b \in \mathbb{R}_+^{K_b \times N},\, \mathbf{g} = \left[g_0, ..., g_{N-1} \right]^\top \in \mathbb{R}_+^N \right\}$$

- Observed data: $\mathbf{x} = \{x_{fn} \in \mathbb{C}\}_{(f,n) \in \mathbb{B}}$

<u>Direct maximum likelihood estimation is intractable</u>

Parameters estimation

- Unsupervised model parameters:

$$\boldsymbol{\theta}_u = \left\{ \mathbf{W}_b \in \mathbb{R}_+^{F \times K_b},\, \mathbf{H}_b \in \mathbb{R}_+^{K_b \times N},\, \mathbf{g} = [g_0, ..., g_{N-1}]^\top \in \mathbb{R}_+^N \right\}$$

- Observed data: $\mathbf{x} = \{x_{fn} \in \mathbb{C}\}_{(f,n) \in \mathbb{B}}$

<u>Direct maximum likelihood estimation is intractable</u>

- Latent data: $\mathbf{z} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$

- Expectation-maximization (EM) algorithm.

Monte Carlo EM algorithm

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) \quad = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_s,\boldsymbol{\theta}_u^\star)} \left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u)\right]$$

Monte Carlo EM algorithm

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$
\begin{aligned}
Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_s,\boldsymbol{\theta}_u^\star)}\left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u)\right] \\
&\approx \frac{1}{R}\sum_{r=1}^{R} \ln p\left(\mathbf{x}, \mathbf{z}^{(r)}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u\right),
\end{aligned} \tag{18}
$$

where the samples $\left\{\mathbf{z}^{(r)}\right\}_{r=1,\dots,R}$ are asymptotically drawn from $p(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u^\star)$ using a Markov chain Monte Carlo method.

Monte Carlo EM algorithm

- **E-Step.** From the current value of the parameters $\boldsymbol{\theta}_u^\star$, compute:

$$\begin{aligned} Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) &= \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_s,\boldsymbol{\theta}_u^\star)} \left[\ln p(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u)\right] \\ &\approx \frac{1}{R} \sum_{r=1}^{R} \ln p\left(\mathbf{x}, \mathbf{z}^{(r)}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u\right), \end{aligned} \tag{18}$$

where the samples $\left\{\mathbf{z}^{(r)}\right\}_{r=1,\dots,R}$ are asymptotically drawn from $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_s,\boldsymbol{\theta}_u^\star)$ using a Markov chain Monte Carlo method.

- **M-Step.**

$$\boldsymbol{\theta}_u^\star \leftarrow \underset{\boldsymbol{\theta}_u}{\arg\max} \; Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star), \tag{19}$$

with $\boldsymbol{\theta}_u = \{\mathbf{H}_b \in \mathbb{R}_+^{K_b \times N}, \mathbf{W}_b \in \mathbb{R}_+^{F \times K_b}, \mathbf{g} \in \mathbb{R}_+^N\}$.

Speech estimation

Let $\tilde{s}_{fn} = \sqrt{g_n} s_{fn}$ be the scaled speech STFT coefficients.

Posterior mean estimation (Wiener-like filtering)

$$\begin{aligned} \hat{\tilde{s}}_{fn} &= \mathbb{E}_{p(\tilde{s}_{fn} | x_{fn}; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u)}[\tilde{s}_{fn}] \\ &= \mathbb{E}_{p(\mathbf{z}_n | \mathbf{x}_n; \boldsymbol{\theta}_s, \boldsymbol{\theta}_u)} \left[\frac{g_n \sigma_f^2(\mathbf{z}_n)}{g_n \sigma_f^2(\mathbf{z}_n) + (\mathbf{W}_b \mathbf{H}_b)_{f,n}} \right] x_{fn}. \end{aligned} \tag{20}$$

Intractable expectation $\rightarrow$ Markov chain Monte Carlo.

Deep-learning-based priors for single-channel speech enhancement

Experiments

Dataset

- **Clean speech signals**: TIMIT database 🔊.
- **Noise signals**: DEMAND database (domestic environment, nature, office, indoor public spaces, street and transportation) 🔊.
- **Training**:
  - training set of TIMIT database;
  - $\sim$ 4 hours of speech;
  - 462 speakers.
- **Testing**:
  - 168 noisy mixtures at 0 dB signal-to-noise ratio;
  - Different speakers and sentences than in the training set.

Semi-supervised NMF baseline

Independently for all $(f, n) \in \mathbb{B}$:

$$s_{fn} \sim \mathcal{N}_c(0, (\mathbf{W}_s \mathbf{H}_s)_{f,n}) \qquad \text{and} \qquad b_{fn} \sim \mathcal{N}_c(0, (\mathbf{W}_b \mathbf{H}_b)_{f,n}),$$

with $\mathbf{W}_s \in \mathbb{R}_+^{F \times K_s}$, $\mathbf{H}_s \in \mathbb{R}_+^{K_s \times N}$, $\mathbf{W}_b \in \mathbb{R}_+^{F \times K_b}$ and $\mathbf{H}_b \in \mathbb{R}_+^{K_b \times N}$.

▷ **Training**: Learn $\mathbf{W}_s$ from a dataset of clean speech signals.

▷ **Test**: Estimate $\mathbf{H}_s, \mathbf{W}_b, \mathbf{H}_b$ from the noisy mixture signal.

▷ **Speech reconstruction**:

$$\hat{s}_{fn} = \frac{(\mathbf{W}_s \mathbf{H}_s)_{f,n}}{(\mathbf{W}_s \mathbf{H}_s + \mathbf{W}_b \mathbf{H}_b)_{f,n}} x_{fn}$$

Related to (Smaragdis et al. 2007; Févotte et al. 2009)

Fully-supervised deep-learning reference method (Xu et al. 2015)

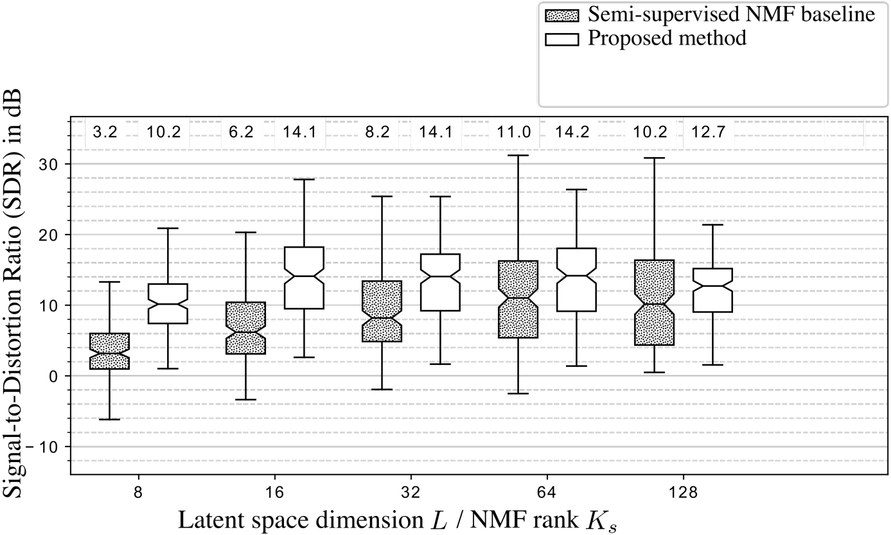
- A deep neural network is trained to map noisy speech log-power spectrograms to clean speech log-power spectrograms.

- From (Xu et al. 2015):

> *"to improve the generalization capability we include more than 100 different noise types in designing the training set"*

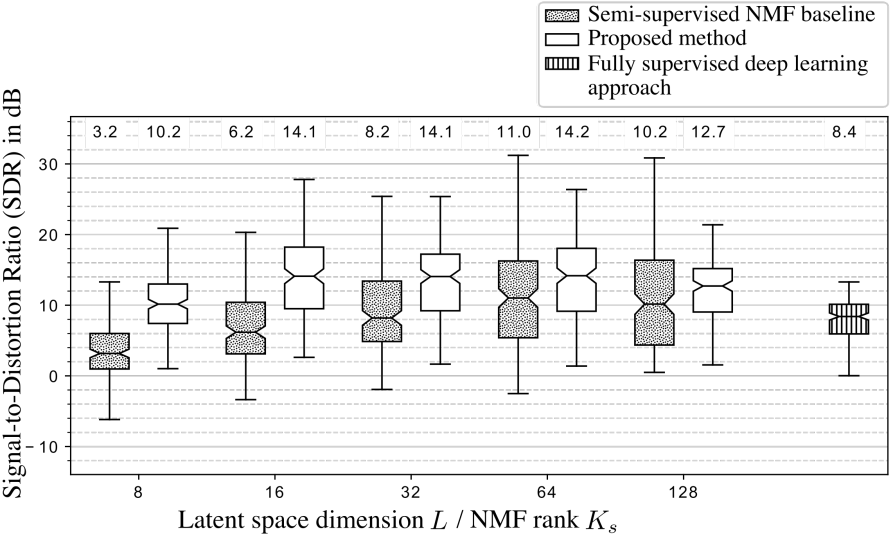
- We used a different noise database (with overlapping noise types) for testing.

Median value indicated above each boxplot.



Experimental results

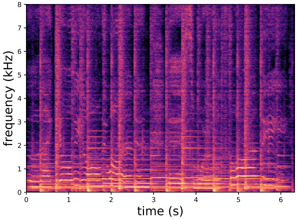
Median value indicated above each boxplot.



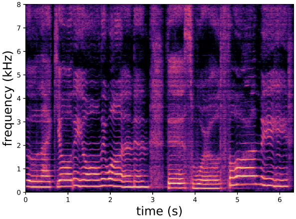
Musical audio example

All models were trained on speech (not singing voice).

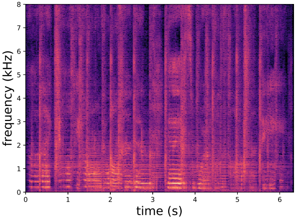
mixture



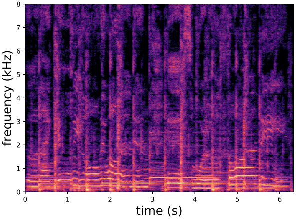
original voice



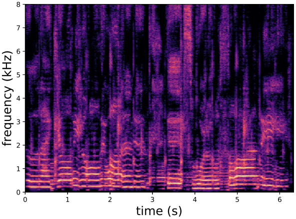
fully-supervised DNN



semi-supervised NMF



proposed



Conclusion

Conclusion

Deep-learning-based generative models can be used as priors for solving ill-posed inverse problems.

A flexible approach:

- Semi-supervision: mixing supervised and unsupervised models.
- Easy to adapt to other problems.

For example, multichannel extension $\mathbf{s}_{fn} \in \mathbb{C}^I$ (submitted to ICASSP 2019):

$$
\begin{aligned}
\mathbf{z}_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\
\mathbf{s}_{fn} \mid \mathbf{z}_n &\sim \mathcal{N}_c\Big(\mathbf{0}, \underbrace{\sigma_f^2(\mathbf{z}_n)}_{\substack{\text{supervised} \\ \text{spectro-temporal} \\ \text{model}}} \times \underbrace{\mathbf{R}_{\mathbf{s},f}}_{\substack{\text{unsupervised} \\ \text{spatial} \\ \text{model}}}\Big).
\end{aligned}
\tag{21}
$$

Spatial covariance matrix model from (Duong et al. 2010).

Thank you

Audio examples and code available online:

`https://sleglaive.github.io`

References

Bando, Y., M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara (2018). "Statistical Speech Enhancement Based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*.

Benaroya, L., L. Mcdonagh, F. Bimbot, and R. Gribonval (2003). "Non negative sparse representation for Wiener based source separation with a single sensor". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*.

Duong, N. Q. K., E. Vincent, and R. Gribonval (2010). "Under-determined reverberant audio source separation using a full-rank spatial covariance model". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 18.7.

Févotte, C. and M. Kowalski (2014). "Low-rank time-frequency synthesis". In: *Proc. Adv. Neural Information Process. Syst. (NIPS)*.

Févotte, C., N. Bertin, and J.-L. Durrieu (2009). "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis". In: *Neural computation* 21.3.

Goodfellow, I. et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*.

Hadad, E., F. Heese, P. Vary, and S. Gannot (2014). "Multichannel audio database in various acoustic environments". In: *Proc. IEEE Int. Workshop Acoustic Signal Enhancement (IWAENC)*.

Kingma, D. P. and M. Welling (2014). "Auto-encoding variational Bayes". In: *Proc. Int. Conf. Learning Representations (ICLR)*.

Kowalski, M., E. Vincent, and R. Gribonval (2010). "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 18.7.

Leglaive, S., L. Girin, and R. Horaud (2018a). "A variance modeling framework based on variational autoencoders for speech enhancement". Proc. IEEE Int. Workshop Machine Learning Signal Process. (MLSP).

Leglaive, S., R. Badeau, and G. Richard (2018b). "Student's t Source and Mixing Models for Multichannel Audio Source Separation". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26.6.

Polack, J.-D. (1988). "La transmission de l'énergie sonore dans les salles". PhD thesis. Université du Maine.

Smaragdis, P., B. Raj, and M. Shashanka (2007). "Supervised and semi-supervised separation of sounds from single-channel mixtures". In: *Proc. Int. Conf. Indep. Component Analysis and Signal Separation*.

Xu, Y., J. Du, L.-R. Dai, and C.-H. Lee (2015). "A regression approach to speech enhancement based on deep neural networks". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 23.1.

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