Semi-Supervised Speech Enhancement with Variational Autoencoders

Simon Leglaive

CentraleSupélec, IETR

Joint work with Xavier Alameda-Pineda, Laurent Girin, Radu Horaud, Mostafa Sadeghi

Seminar at École Normale Supérieure (ENS), Paris

December 3rd, 2019

CentraleSupélec

IETR

Institut d'électronique et de télécommunications de Rennes

Outline

Bayesian and Deep Learning

Inverse problems

latent variables of interest

- ▹ signals;
- ▹ model parameters;
- ▹ state of a system;
- ▹ etc.

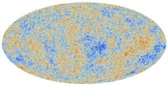
direct problem →

← inverse problem

observations

incomplete and/or noisy measurements

Astrophysics

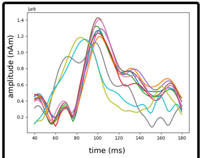


Credit: ESA - Planck Collaboration

Medical imaging

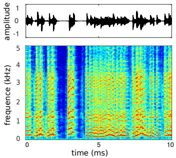
Credit: Wikipedia

Biomedical signal processing



Credit: MNE Python

Audio signal processing



Ill-posed inverse problem: requires external information.

Bayesian methodology vs deep learning

Bayesian methodology

External information:	Prior	$p(\text{latent})$
	Likelihood	$p(\text{obs.} \mid \text{latent})$
Problem solving:	Posterior	$p(\text{latent} \mid \text{obs.})$
Advantages:	Flexible, explanatory	
Drawbacks:	Performance, strong hypotheses	

Bayesian methodology vs deep learning

Bayesian methodology

External information:	Prior	$p(\text{latent})$
	Likelihood	$p(\text{obs.} \mid \text{latent})$
Problem solving:	Posterior	$p(\text{latent} \mid \text{obs.})$
Advantages:	Flexible, explanatory	
Drawbacks:	Performance, strong hypotheses	

Discriminative deep learning approach

External information:	Training data
Problem solving:	Observations $\xrightarrow[\text{network}]{\text{neural}}$ latent variables of interest
Advantages:	State-of-the-art, fast at test time
Drawbacks:	Poorly flexible once trained, poorly explanatory

Bayesian methodology vs deep learning

Bayesian methodology

External information:	Prior	$p(\text{latent})$
	Likelihood	$p(\text{obs.} \mid \text{latent})$
Problem solving:	Posterior	$p(\text{latent} \mid \text{obs.})$
Advantages:	Flexible, explanatory	
Drawbacks:	Performance, strong hypotheses	

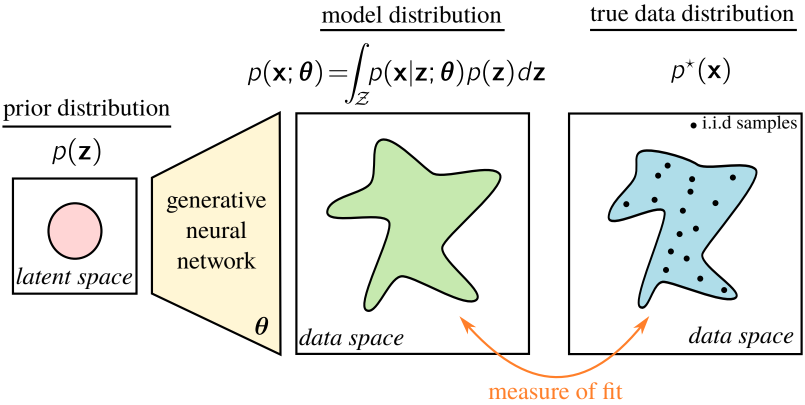
Discriminative deep learning approach

External information:	Training data
Problem solving:	Observations $\xrightarrow[\text{network}]{\text{neural}}$ latent variables of interest
Advantages:	State-of-the-art, fast at test time
Drawbacks:	Poorly flexible once trained, poorly explanatory

How to exploit the best of both worlds?

Learn the prior directly from data using deep generative models.

Deep-learning-based generative models (with latent variables)



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

Application: semi-supervised speech enhancement

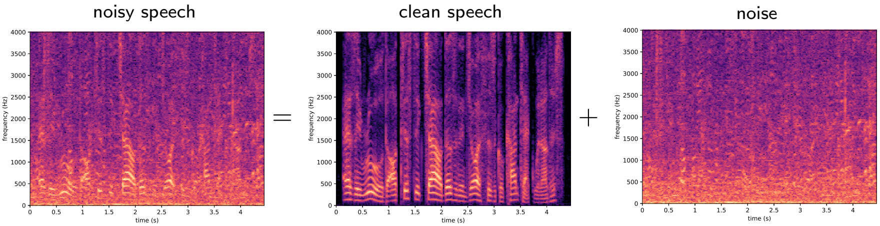
Semi-supervised approach (Smaragdis et al. 2007):

- ◇ Training from clean speech signals only.
- ◇ Free of generalization issues regarding the noisy recording environment.

We want the method to be speaker independent.

Speech Enhancement with Non-negative Matrix Factorization

Speech enhancement as a source separation problem



In the short-term Fourier transform (STFT) domain, we observe:

$$x_{fn} = s_{fn} + b_{fn}, \tag{1}$$

- ▷ $s_{fn} \in \mathbb{C}$ is the clean speech signal.
- ▷ $b_{fn} \in \mathbb{C}$ is the noise signal.
- ▷ $(f, n) \in \mathbb{B} = \{0, ..., F-1\} \times \{0, ..., N-1\}$.
- ▷ f is the frequency index and n the time-frame index.

Non-stationary Gaussian model

- **Non-stationary Gaussian model** (Pham and Garat 1997; Cardoso 2001)

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

$$s_{fn} \sim \mathcal{N}_c(0, v_{s,fn}) \qquad \perp \qquad b_{fn} \sim \mathcal{N}_c(0, v_{b,fn}). \tag{2}$$

  Consequently, we also have:

$$x_{fn} \sim \mathcal{N}_c\left(0, v_{s,fn} + v_{b,fn}\right). \tag{3}$$

Non-stationary Gaussian model

- **Non-stationary Gaussian model** (Pham and Garat 1997; Cardoso 2001)

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

  $$s_{fn} \sim \mathcal{N}_c(0, v_{s,fn}) \qquad \perp \qquad b_{fn} \sim \mathcal{N}_c(0, v_{b,fn}). \tag{2}$$

  Consequently, we also have:

  $$x_{fn} \sim \mathcal{N}_c\left(0, v_{s,fn} + v_{b,fn}\right). \tag{3}$$

- **Spectro-temporal variance modeling** (Vincent et al. 2010; Vincent et al. 2014)

  - persistence with structured sparsity penalties;

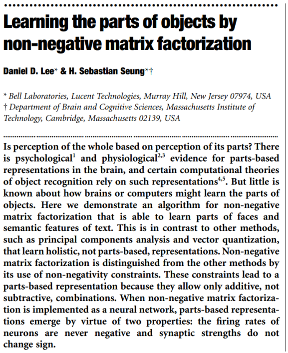
    (Févotte et al. 2006; Kowalski and Torrésani 2009)
  - redundancy with non-negative matrix factorization;

    (Benaroya et al. 2003; Févotte et al. 2009; Ozerov et al. 2012)
  - more complex structures with deep neural networks.

    (Bando et al. 2018; Leglaive et al. 2018)

Non-negative matrix factorization (NMF) (Lee and Seung 1999)

Low-rank matrix factorization technique with non-negativity constraints.

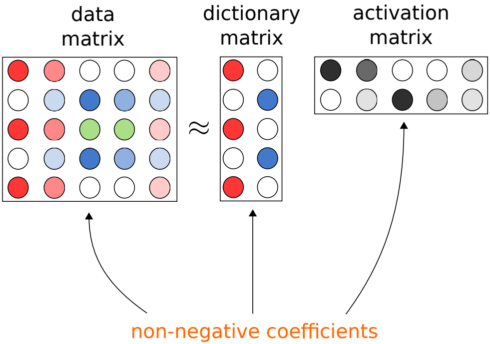


Learning the parts of objects by non-negative matrix factorization

Daniel D. Lee* & H. Sebastian Seung*†

* *Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA*
† *Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

Is perception of the whole based on perception of its parts? There is psychological[1] and physiological[2,3] evidence for parts-based representations in the brain, and certain computational theories of object recognition rely on such representations[4,5]. But little is known about how brains or computers might learn the parts of objects. Here we demonstrate an algorithm for non-negative matrix factorization that is able to learn parts of faces and semantic features of text. This is in contrast to other methods, such as principal components analysis and vector quantization, that learn holistic, not parts-based, representations. Non-negative matrix factorization is distinguished from the other methods by its use of non-negativity constraints. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. When non-negative matrix factorization is implemented as a neural network, parts-based representations emerge by virtue of two properties: the firing rates of neurons are never negative and synaptic strengths do not change sign.

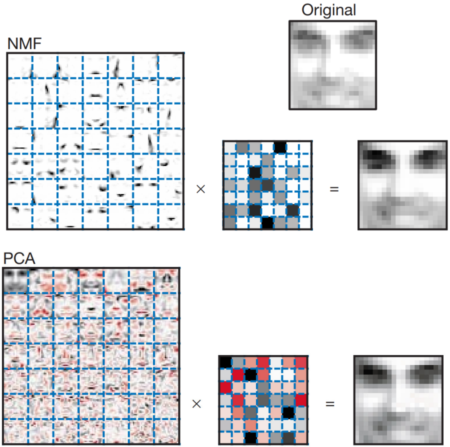


Principal component analysis (PCA) is also a low-rank matrix factorization, with different constraints.

NMF for face images (Lee and Seung 1999)

A face can be represented a linear combination of basis images.

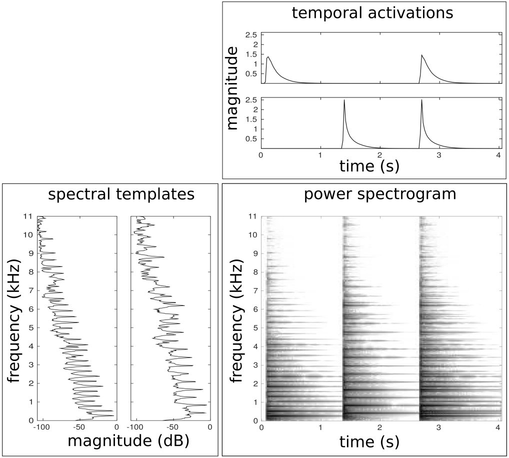
- with NMF: localized features representing intuitive notions of parts of faces.
- with PCA: eigenfaces.



Reproduced from (Lee and Seung, 1999)

NMF for audio spectrograms (Smaragdis and Brown 2003)

A spectrogram can be represented a linear combination of spectral templates.



NMF-based variance modeling for audio (Févotte et al. 2009)

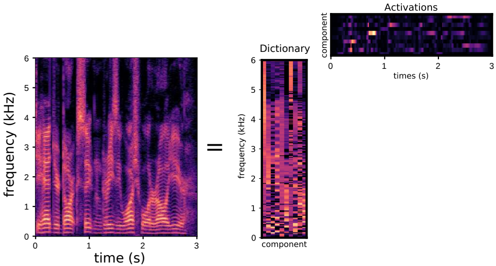
- **Non-stationary Gaussian model**:

$$s_{fn} \sim \mathcal{N}_c(0, v_{s,fn}) \qquad \text{and} \qquad b_{fn} \sim \mathcal{N}_c(0, v_{b,fn}). \tag{4}$$

- **Variance model** based on non-negative matrix factorization (NMF):

$$v_{j,fn} = (\mathbf{W}_j \mathbf{H}_j)_{f,n}, \qquad j \in \{s, b\}, \tag{5}$$

  - $\mathbf{W}_j \in \mathbb{R}_+^{F \times K_j}$ is a dictionary matrix of spectral templates;
  - $\mathbf{H}_j \in \mathbb{R}_+^{K_j \times N}$ is the activation matrix;
  - K_j is the rank of the factorization (usually $K_j(F+N) \ll FN$).



Semi-supervised NMF-based speech enhancement

(Smaragdis et al. 2007; Mysore and Smaragdis 2011)

Speech enhancement with Wiener filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn})}[s_{fn}] = \frac{(\mathbf{W}_s\mathbf{H}_s)_{f,n}}{(\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_b\mathbf{H}_b)_{f,n}} x_{fn}. \tag{6}$$

Semi-supervised NMF-based speech enhancement

(Smaragdis et al. 2007; Mysore and Smaragdis 2011)

Speech enhancement with Wiener filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn})}[s_{fn}] = \frac{(\mathbf{W}_s\mathbf{H}_s)_{f,n}}{(\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_b\mathbf{H}_b)_{f,n}} x_{fn}. \tag{6}$$

Training: learn $\mathbf{W}_s$ from a dataset of clean speech signals

$$\min_{\mathbf{W}_s \geq 0} \sum_{(f,n)\in\mathbb{B}} d_{\mathrm{IS}}\left(|s_{fn}|^2, (\mathbf{W}_s\mathbf{H}_s)_{f,n}\right), \tag{7}$$

- $d_{\mathrm{IS}}(\cdot,\cdot)$ is the Itakura-Saito (IS) divergence.
- equivalent to maximizing the likelihood of $\mathbf{s} = \{s_{fn}\}_{(f,n)\in\mathbb{B}}$ (Févotte et al. 2009).
- majorize-minimize algorithm (Févotte and Idier 2011).

Semi-supervised NMF-based speech enhancement

(Smaragdis et al. 2007; Mysore and Smaragdis 2011)

Speech enhancement with Wiener filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn})}[s_{fn}] = \frac{(\mathbf{W}_s\mathbf{H}_s)_{f,n}}{(\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_b\mathbf{H}_b)_{f,n}} x_{fn}. \tag{6}$$

Training: learn $\mathbf{W}_s$ from a dataset of clean speech signals

$$\min_{\mathbf{W}_s \geq 0} \sum_{(f,n)\in\mathbb{B}} d_{\text{IS}}\left(|s_{fn}|^2, (\mathbf{W}_s\mathbf{H}_s)_{f,n}\right), \tag{7}$$

- $d_{\text{IS}}(\cdot,\cdot)$ is the Itakura-Saito (IS) divergence.
- equivalent to maximizing the likelihood of $\mathbf{s} = \{s_{fn}\}_{(f,n)\in\mathbb{B}}$ (Févotte et al. 2009).
- majorize-minimize algorithm (Févotte and Idier 2011).

Test: estimate $\mathbf{H}_s, \mathbf{W}_b, \mathbf{H}_b$ from the noisy mixture signal

$$\min_{\mathbf{H}_s,\mathbf{W}_b,\mathbf{H}_b \geq 0} \sum_{(f,n)\in\mathbb{B}} d_{\text{IS}}\left(|x_{fn}|^2, (\mathbf{W}_s\mathbf{H}_s + \mathbf{W}_b\mathbf{H}_b)_{f,n}\right). \tag{8}$$

Research problem: from NMF to neural networks

NMF-based supervised speech model

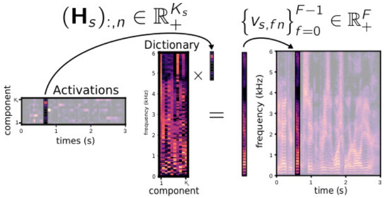
$$v_{s,fn} = (\mathbf{W}_s\mathbf{H}_s)_{f,n} = (\mathbf{W}_s)_{f,:}^{\top} \times (\mathbf{H}_s)_{:,n}$$

Negative aspects:

- ▷ linear function of $(\mathbf{H}_s)_{:,n} \in \mathbb{R}_+^{K_s}$.
- ▷ # trainable parameters $= F \times K_s$.

Positive aspect:

- ▷ Interpretability.



In this work, we explore the use of neural networks in order to overcome the limitations of this variance model.

Speech Enhancement with Variational Autoencoders

Speech Enhancement with Variational Autoencoders

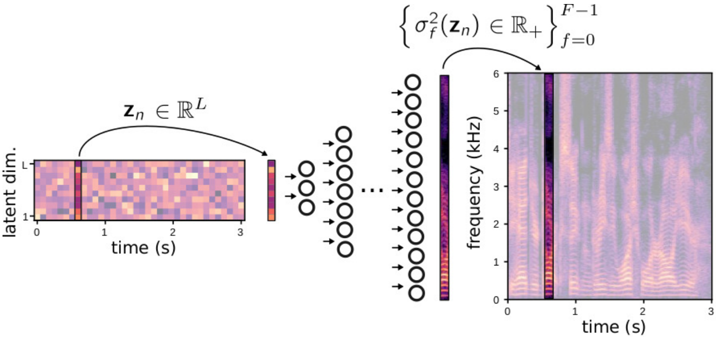
Deep Generative Speech Modeling

Deep generative speech model (Bando et al. 2018)

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

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(0, \sigma_f^2(\mathbf{z}_n)\right), \qquad \text{with } \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), \tag{9}$$

and $\sigma_f^2 : \mathbb{R}^L \mapsto \mathbb{R}_+$ corresponds to a neural network of parameters $\boldsymbol{\theta}_s$.



How to learn the parameters $\boldsymbol{\theta}_s$ of this generative neural network?

Learning the model parameters with variational autoencoders

- **Training dataset** of STFT speech time frames: $\mathbf{s} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$.
- **Associated latent variables**: $\mathbf{z} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$.
- **Difficulty**: Intractable marginal likelihood $p(\mathbf{s}; \boldsymbol{\theta}_s) = \displaystyle\int p(\mathbf{s}|\mathbf{z}; \boldsymbol{\theta}_s) p(\mathbf{z}) d\mathbf{z}$.
- **Solution**: Variational autoencoder (VAE) (Kingma and Welling 2014).

Learning the model parameters with variational autoencoders

- **Training dataset** of STFT speech time frames: $\mathbf{s} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$.
- **Associated latent variables**: $\mathbf{z} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$.
- **Difficulty**: Intractable marginal likelihood $p(\mathbf{s}; \boldsymbol{\theta}_s) = \displaystyle\int p(\mathbf{s}|\mathbf{z}; \boldsymbol{\theta}_s) p(\mathbf{z}) d\mathbf{z}$.
- **Solution**: Variational autoencoder (VAE) (Kingma and Welling 2014).

Maximize a lower bound of $\ln p(\mathbf{s}; \boldsymbol{\theta}_s)$, which can be recast as:

$$\min_{\boldsymbol{\theta}_s} \sum_{(f,n) \in \mathbb{B}} \mathbb{E}_{q(\mathbf{z}_n|\mathbf{s}_n;\phi)} \left[d_{IS}\left(|s_{fn}|^2 ; \sigma_f^2(\mathbf{z}_n) \right) \right], \tag{10}$$

where $q(\mathbf{z}_n|\mathbf{s}_n; \phi)$ is an approximation of the intractable posterior $p(\mathbf{z}_n|\mathbf{s}_n; \boldsymbol{\theta}_s)$ and is defined by an "encoder network" of parameters ϕ.

The dependency of $\sigma_f^2(\cdot)$ on $\boldsymbol{\theta}_s$ is not made explicit to avoid cluttered notations.

Variational inference (Jordan et al. 1999; Blei et al. 2017)

For any variational distribution $q(\mathbf{z}|\mathbf{s};\phi)$, we have:

$$\ln p(\mathbf{s};\boldsymbol{\theta}_s) = \mathcal{L}(\phi,\boldsymbol{\theta}_s) + D_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{s};\phi) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_s)\big), \tag{11}$$

where $D_{\mathrm{KL}}\big(q \parallel p\big) = \mathbb{E}_q[\ln q - \ln p] \geq 0$.

Variational free energy

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

Variational inference (Jordan et al. 1999; Blei et al. 2017)

For any variational distribution $q(\mathbf{z}|\mathbf{s};\phi)$, we have:

$$\ln p(\mathbf{s};\boldsymbol{\theta}_s) = \mathcal{L}(\phi,\boldsymbol{\theta}_s) + D_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{s};\phi) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_s)\big), \tag{11}$$

where $D_{\mathrm{KL}}\big(q \parallel p\big) = \mathbb{E}_q[\ln q - \ln p] \geq 0$.

Variational free energy

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

Problem #1

$$\max_{\boldsymbol{\theta}_s} \ \mathcal{L}(\phi,\boldsymbol{\theta}_s)$$

where $\mathcal{L}(\phi,\boldsymbol{\theta}_s) \leq \ln p(\mathbf{s};\boldsymbol{\theta}_s)$.

Variational inference (Jordan et al. 1999; Blei et al. 2017)

For any variational distribution $q(\mathbf{z}|\mathbf{s};\phi)$, we have:

$$\ln p(\mathbf{s};\boldsymbol{\theta}_s) = \mathcal{L}(\phi,\boldsymbol{\theta}_s) + D_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{s};\phi) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_s)\big), \tag{11}$$

where $D_{\mathrm{KL}}\big(q \parallel p\big) = \mathbb{E}_q[\ln q - \ln p] \geq 0$.

Variational free energy

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

Problem #1

$$\max_{\boldsymbol{\theta}_s} \; \mathcal{L}(\phi,\boldsymbol{\theta}_s)$$

where $\mathcal{L}(\phi,\boldsymbol{\theta}_s) \leq \ln p(\mathbf{s};\boldsymbol{\theta}_s)$.

Problem #2

$$\max_{\phi} \; \mathcal{L}(\phi,\boldsymbol{\theta}_s)$$

$$\Leftrightarrow$$

$$\min_{\phi} \; D_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{s};\phi) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_s)\big)$$

Variational inference (Jordan et al. 1999; Blei et al. 2017)

For any variational distribution $q(\mathbf{z}|\mathbf{s};\phi)$, we have:

$$\ln p(\mathbf{s};\boldsymbol{\theta}_s) = \mathcal{L}(\phi,\boldsymbol{\theta}_s) + D_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{s};\phi) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_s)\big), \tag{11}$$

where $D_{\mathrm{KL}}\big(q \parallel p\big) = \mathbb{E}_q[\ln q - \ln p] \geq 0$.

Variational free energy

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

Problem #1

$$\max_{\boldsymbol{\theta}_s} \ \mathcal{L}(\phi,\boldsymbol{\theta}_s)$$

where $\mathcal{L}(\phi,\boldsymbol{\theta}_s) \leq \ln p(\mathbf{s};\boldsymbol{\theta}_s)$.

Problem #2

$$\max_{\boldsymbol{\phi}} \ \mathcal{L}(\phi,\boldsymbol{\theta}_s)$$
$$\Leftrightarrow$$
$$\min_{\boldsymbol{\phi}} \ D_{\mathrm{KL}}\big(q(\mathbf{z}|\mathbf{s};\phi) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_s)\big)$$

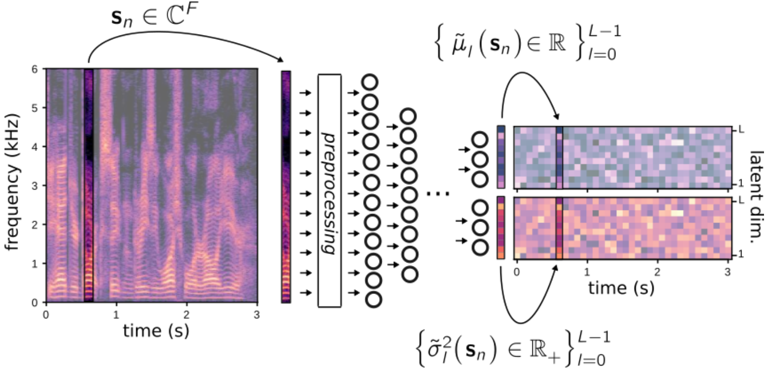
To define the objective function, we need to define $q(\mathbf{z}|\mathbf{s};\phi)$.

The encoder network

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

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

where $\tilde{\mu}_l : \mathbb{C}^F \mapsto \mathbb{R}$ and $\tilde{\sigma}_l^2 : \mathbb{C}^F \mapsto \mathbb{R}_+$ correspond to a neural network of parameters ϕ.



Variational free energy: full expression

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

Variational free energy: full expression

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

- ▷ Intractable expectation replaced by a sample average:

$$\mathbb{E}_{q(\mathbf{z}_n|\mathbf{s}_n;\phi)}\left[d_{\mathrm{IS}}\left(\left|s_{fn}\right|^2; \sigma_f^2(\mathbf{z}_n)\right)\right] \approx \frac{1}{R}\sum_{r=1}^{R}\left[d_{\mathrm{IS}}\left(\left|s_{fn}\right|^2; \sigma_f^2\left(\tilde{\mathbf{z}}_n^{(r)}\right)\right)\right], \quad (15)$$

where $\{\tilde{\mathbf{z}}_n^{(r)}\}_{r=1}^{R}$ are i.i.d. realizations drawn[1] from $q\left(\mathbf{z}_n|\mathbf{s}_n;\phi\right)$.

[1] using the so-called "reparametrization trick" (Kingma and Welling 2014).

Variational free energy: full expression

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

▷ Intractable expectation replaced by a sample average:

$$\mathbb{E}_{q(\mathbf{z}_n|\mathbf{s}_n;\phi)}\left[d_{\mathrm{IS}}\left(|s_{fn}|^2; \sigma_f^2(\mathbf{z}_n)\right)\right] \approx \frac{1}{R}\sum_{r=1}^{R}\left[d_{\mathrm{IS}}\left(|s_{fn}|^2; \sigma_f^2\left(\tilde{\mathbf{z}}_n^{(r)}\right)\right)\right], \tag{15}$$

where $\{\tilde{\mathbf{z}}_n^{(r)}\}_{r=1}^R$ are i.i.d. realizations drawn[1] from $q(\mathbf{z}_n|\mathbf{s}_n;\phi)$.

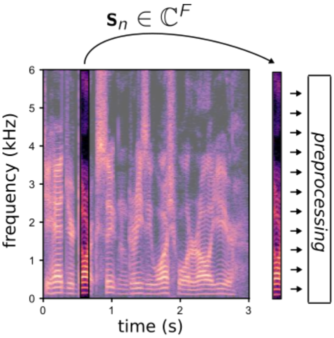
▷ In practice $R = 1$:

$$\mathbb{E}_{q(\mathbf{z}_n|\mathbf{s}_n;\phi)}\left[d_{\mathrm{IS}}\left(|s_{fn}|^2; \sigma_f^2(\mathbf{z}_n)\right)\right] \approx d_{\mathrm{IS}}\left(|s_{fn}|^2; \sigma_f^2(\tilde{\mathbf{z}}_n)\right). \tag{16}$$

[1] using the so-called "reparametrization trick" (Kingma and Welling 2014).

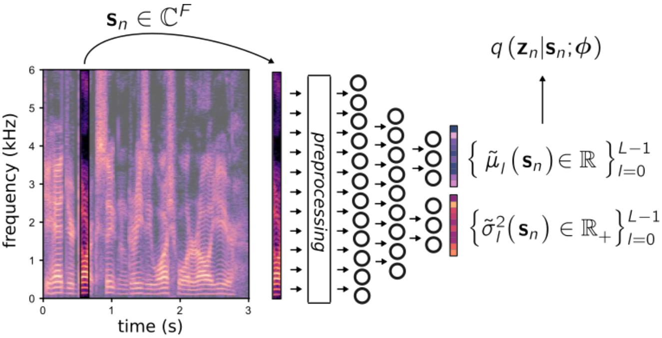
Training procedure step by step (0)

$$\mathcal{L}(\boldsymbol{\theta}_s, \phi) \stackrel{c}{=} -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[d_{\mathrm{IS}}\left(\left|s_{fn}\right|^2; \sigma_f^2(\tilde{\mathbf{z}}_n)\right)\right] + \frac{1}{2}\sum_{l=1}^{L}\sum_{n=0}^{N-1}\left[\ln\tilde{\sigma}_l^2(\mathbf{s}_n) - \tilde{\mu}_l^2(\mathbf{s}_n) - \tilde{\sigma}_l^2(\mathbf{s}_n)\right]. \tag{17}$$



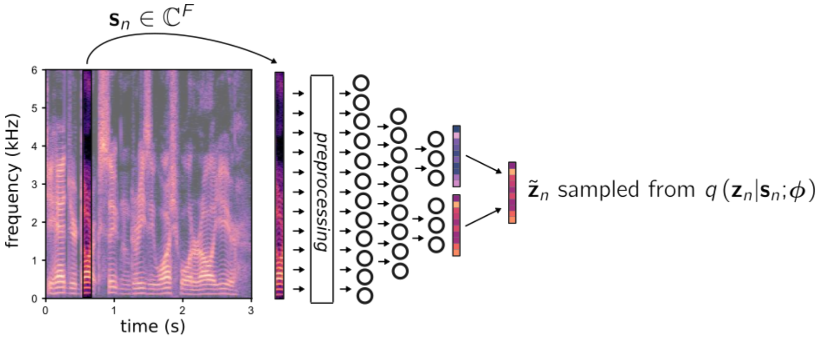
Training procedure step by step (1)

$$\mathcal{L}(\boldsymbol{\theta}_s, \phi) \stackrel{c}{=} -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[d_{\text{IS}}\left(|s_{fn}|^2; \sigma_f^2(\tilde{\mathbf{z}}_n)\right)\right]$$
$$+\frac{1}{2}\sum_{l=1}^{L}\sum_{n=0}^{N-1}\left[\ln\tilde{\sigma}_l^2(\mathbf{s}_n) - \tilde{\mu}_l^{\,2}(\mathbf{s}_n) - \tilde{\sigma}_l^2(\mathbf{s}_n)\right]. \quad (18)$$



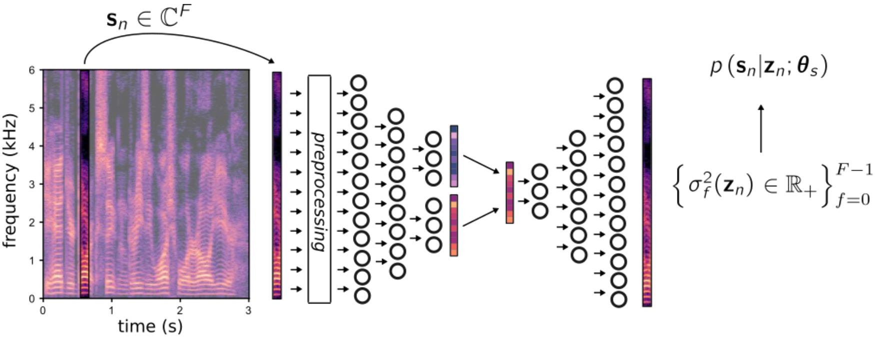
Training procedure step by step (2)

$$\mathcal{L}(\boldsymbol{\theta}_s, \phi) \stackrel{c}{=} -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[d_{\mathrm{IS}}\left(\left|s_{fn}\right|^2; \sigma_f^2(\tilde{\mathbf{z}}_n)\right)\right]$$
$$+\frac{1}{2}\sum_{l=1}^{L}\sum_{n=0}^{N-1}\left[\ln\tilde{\sigma}_l^2(\mathbf{s}_n) - \tilde{\mu}_l{}^2(\mathbf{s}_n) - \tilde{\sigma}_l^2(\mathbf{s}_n)\right]. \quad (19)$$



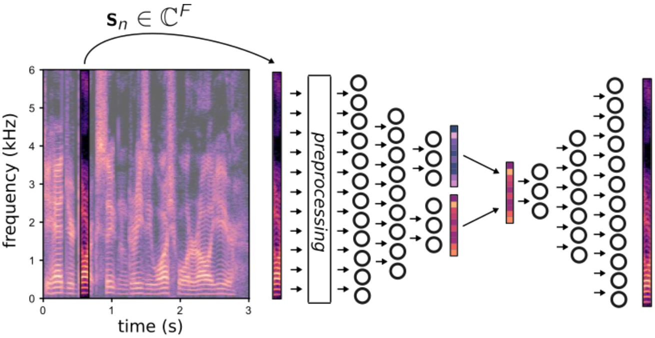
Training procedure step by step (3)

$$\mathcal{L}(\boldsymbol{\theta}_s, \phi) \stackrel{c}{=} -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[d_{\mathrm{IS}}\left(\left|s_{fn}\right|^2; \sigma_f^2(\tilde{\mathbf{z}}_n)\right)\right]$$
$$+\frac{1}{2}\sum_{l=1}^{L}\sum_{n=0}^{N-1}\left[\ln \tilde{\sigma}_l^2(\mathbf{s}_n) - \tilde{\mu}_l^{\,2}(\mathbf{s}_n) - \tilde{\sigma}_l^2(\mathbf{s}_n)\right]. \qquad (20)$$



Training procedure step by step (4)

$$\mathcal{L}(\boldsymbol{\theta}_s, \phi) \stackrel{c}{=} -\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\left[d_{\mathrm{IS}}\left(|s_{fn}|^2 ; \sigma_f^2(\tilde{\mathbf{z}}_n)\right)\right] + \frac{1}{2}\sum_{l=1}^{L}\sum_{n=0}^{N-1}\left[\ln \tilde{\sigma}_l^2(\mathbf{s}_n) - \tilde{\mu}_l^2(\mathbf{s}_n) - \tilde{\sigma}_l^2(\mathbf{s}_n)\right]. \quad (21)$$



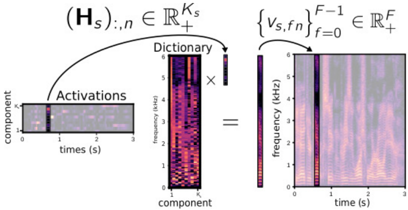
Iterative optimization with a gradient-ascent-based algorithm.

Summary

NMF-based model

$$v_{s,fn} = (\mathbf{W}_s)^\top_{f,:} \times (\mathbf{H}_s)_{:,n}$$

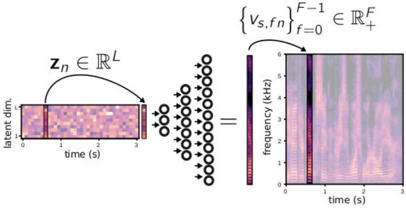
- linear function of $(\mathbf{H}_s)_{:,n} \in \mathbb{R}_+^{K_s}$.
- $\#$ trainable parameters $= F \times K_s$.
- IS divergence minimization.
- Interpretability.



VAE-based model

$$v_{s,fn} = \sigma_f^2(\mathbf{z}_n)$$

- non-linear function of $\mathbf{z}_n \in \mathbb{R}^L$.
- $\#$ trainable parameters is free.
- IS divergence minimization.
- Lack of (direct) interpretability.



Speech Enhancement with Variational Autoencoders

Speech Enhancement

Models for semi-supervised speech enhancement

Supervised speech model

$$s_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(0, \sigma_f^2(\mathbf{z}_n)\right), \qquad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{22}$$

where $\sigma_f^2(\cdot)$ corresponds to the decoder network of parameters $\boldsymbol{\theta}_s$.

Unsupervised noise model

$$b_{fn} \sim \mathcal{N}_c\left(0, (\mathbf{W}_b\mathbf{H}_b)_{f,n}\right), \tag{23}$$

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

Likelihood

$$x_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(0, \sigma_f^2(\mathbf{z}_n) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}\right). \tag{24}$$

S. L., L. Girin, R. Horaud "A variance modeling framework based on variational autoencoders for speech enhancement", *IEEE MLSP, 2018.*

Semi-supervised VAE-based speech enhancement

Speech enhancement with Wiener-like filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn};\boldsymbol{\theta})}[s_{fn}] = \mathbb{E}_{p(\mathbf{z}_n|\mathbf{x}_n;\boldsymbol{\theta})}\left[\frac{\sigma_f^2(\mathbf{z}_n)}{\sigma_f^2(\mathbf{z}_n) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}}\right] x_{fn}, \tag{25}$$

where the expectation is intractable: Markov chain Monte Carlo (MCMC).

S. L., L. Girin, R. Horaud "A variance modeling framework based on variational autoencoders for speech enhancement", *IEEE MLSP, 2018.*

Semi-supervised VAE-based speech enhancement

Speech enhancement with Wiener-like filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn};\boldsymbol{\theta})}[s_{fn}] = \mathbb{E}_{p(\mathbf{z}_n|\mathbf{x}_n;\boldsymbol{\theta})}\left[\frac{\sigma_f^2(\mathbf{z}_n)}{\sigma_f^2(\mathbf{z}_n) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}}\right] x_{fn}, \qquad (25)$$

where the expectation is intractable: Markov chain Monte Carlo (MCMC).

Training: learn $\sigma_f^2(\cdot)$ from a dataset of clean speech signals

Introduce an encoder network and maximize a lower bound of $p(\mathbf{s}; \boldsymbol{\theta}_s)$.

S. L., L. Girin, R. Horaud "A variance modeling framework based on variational autoencoders for speech enhancement", *IEEE MLSP, 2018.*

Semi-supervised VAE-based speech enhancement

Speech enhancement with Wiener-like filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn};\boldsymbol{\theta})}[s_{fn}] = \mathbb{E}_{p(\mathbf{z}_n|\mathbf{x}_n;\boldsymbol{\theta})}\left[\frac{\sigma_f^2(\mathbf{z}_n)}{\sigma_f^2(\mathbf{z}_n) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}}\right] x_{fn}, \qquad (25)$$

where the expectation is intractable: Markov chain Monte Carlo (MCMC).

Training: learn $\sigma_f^2(\cdot)$ from a dataset of clean speech signals

Introduce an encoder network and maximize a lower bound of $p(\mathbf{s};\boldsymbol{\theta}_s)$.

Test: estimate $\mathbf{W}_b, \mathbf{H}_b$ from the noisy mixture signal

We would like to maximize w.r.t $\mathbf{W}_b \in \mathbb{R}_+^{F\times K_b}, \mathbf{H}_b \in \mathbb{R}_+^{K_b\times N}$:

$$p(\mathbf{x};\boldsymbol{\theta}) = \int p(\mathbf{x}|\mathbf{z};\boldsymbol{\theta})p(\mathbf{z})d\mathbf{z}. \qquad (26)$$

We develop a Monte Carlo EM algorithm (see paper for further details).

S. L., L. Girin, R. Horaud "A variance modeling framework based on variational autoencoders for speech enhancement", *IEEE MLSP, 2018.*

Speech Enhancement with Variational Autoencoders

Experiments

Dataset

- **Clean speech signals**: TIMIT database (Garofolo et al. 1993).
- **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.
- **Test**:
  - 168 noisy mixtures at 0 dB signal-to-noise ratio;
  - Different speakers and sentences than in the training set.

Reference methods

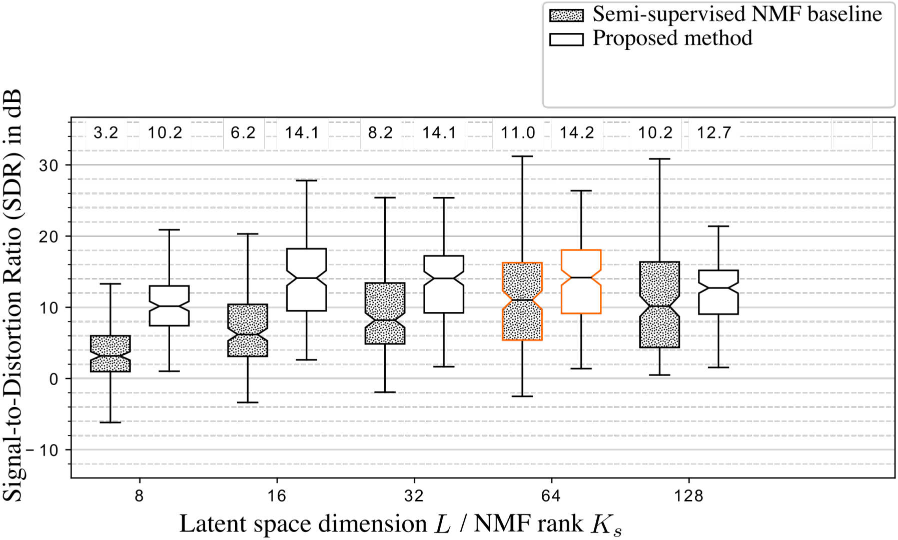
1. **Semi-supervised NMF baseline.**
2. **Fully-supervised deep-learning-based method** (Xu et al. 2015):
   - ▷ Deep neural network for mapping noisy speech log-power spectrograms to clean speech log-power spectrograms.
   - ▷ From (Xu et al. 2015):

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

   - ▷ Here, we use different noise datasets for training and testing (with overlapping noise types).

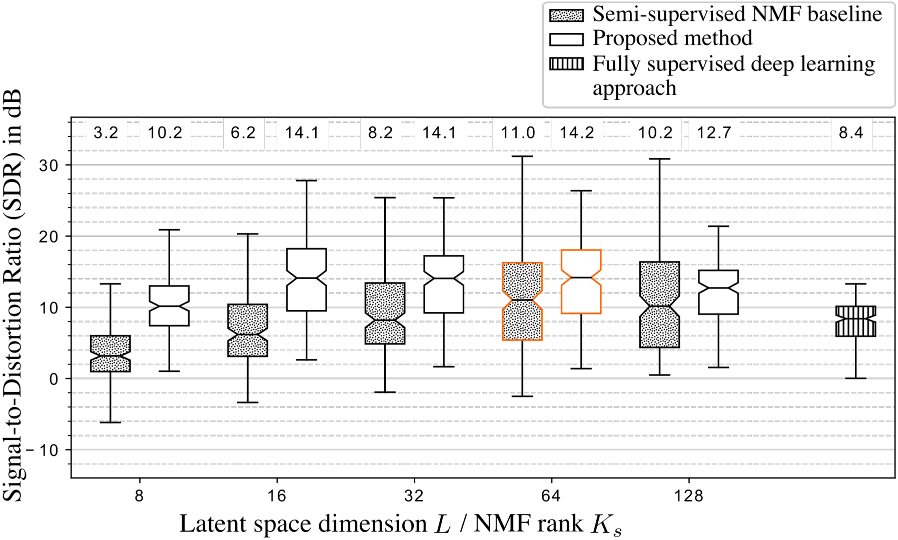
Experimental results

Median value indicated above each boxplot.



Experimental results

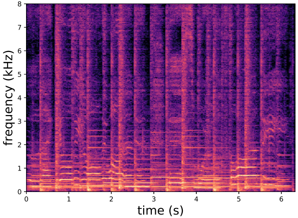
Median value indicated above each boxplot.



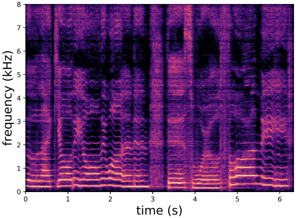
Singing voice separation in a monophonic mixture

All models were trained on speaking and not singing voice.

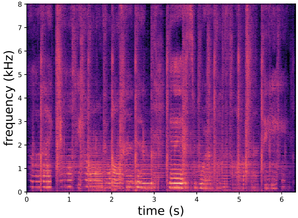
mixture



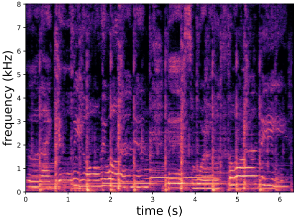
original voice



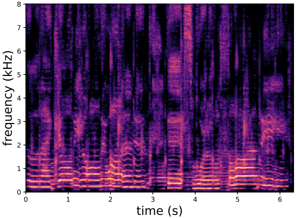
fully-supervised DNN



semi-supervised NMF



proposed

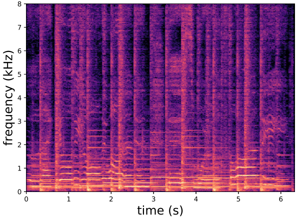


Song: "Sunrise" by Shannon Hurley, from the MGT Music Audio Signal Separation (MASS) dataset.

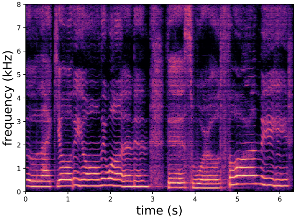
Singing voice separation in a monophonic mixture

All models were trained on speaking and not singing voice.

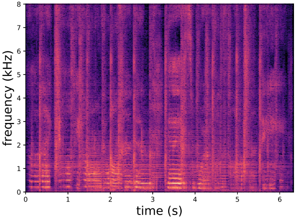
mixture



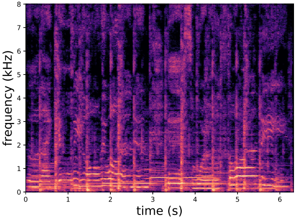
original voice



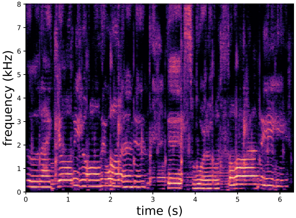
fully-supervised DNN



semi-supervised NMF



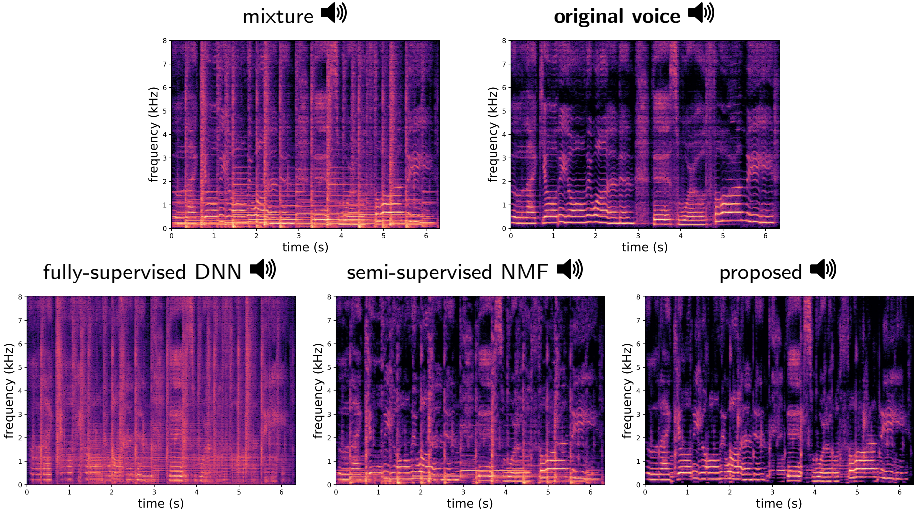
proposed



Song: “Sunrise” by Shannon Hurley, from the MGT Music Audio Signal Separation (MASS) dataset.

Singing voice separation in a monophonic mixture

All models were trained on speaking and not singing voice.

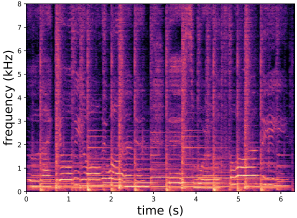


Song: "Sunrise" by Shannon Hurley, from the MGT Music Audio Signal Separation (MASS) dataset.

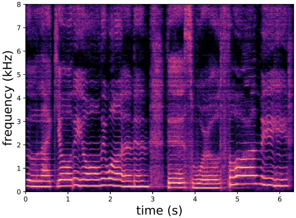
Singing voice separation in a monophonic mixture

All models were trained on speaking and not singing voice.

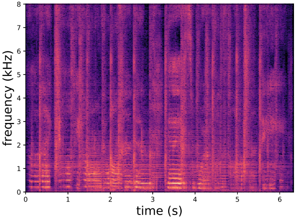
mixture



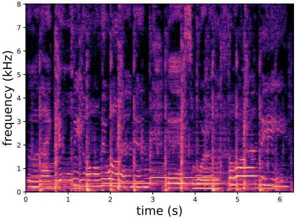
original voice



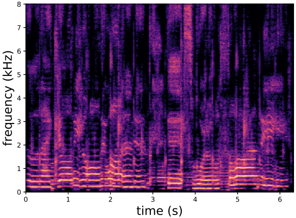
fully-supervised DNN



semi-supervised NMF



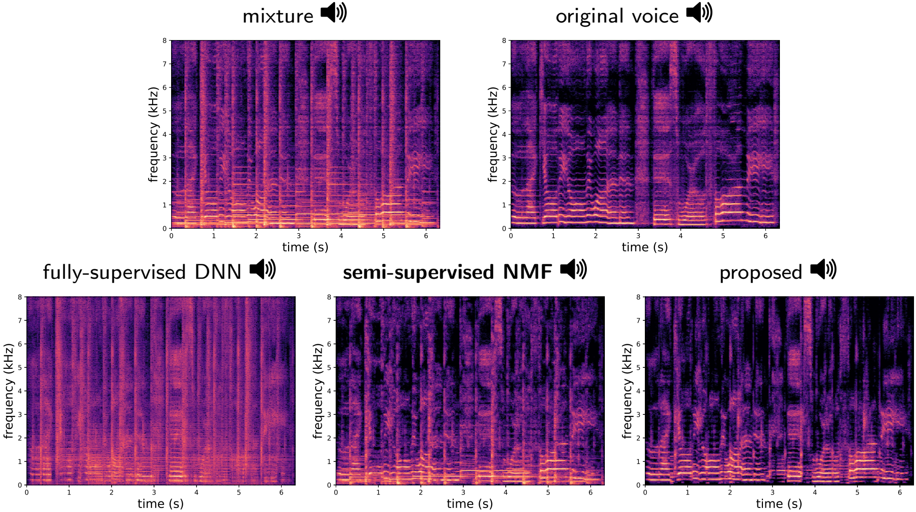
proposed



Song: "Sunrise" by Shannon Hurley, from the MGT Music Audio Signal Separation (MASS) dataset.

Singing voice separation in a monophonic mixture

All models were trained on speaking and not singing voice.

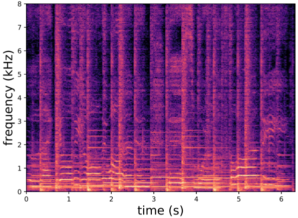


Song: "Sunrise" by Shannon Hurley, from the MGT Music Audio Signal Separation (MASS) dataset.

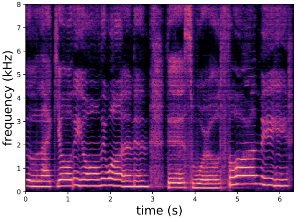
Singing voice separation in a monophonic mixture

All models were trained on speaking and not singing voice.

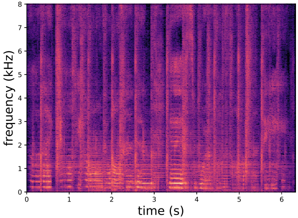
mixture



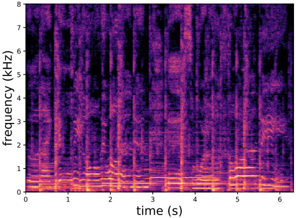
original voice



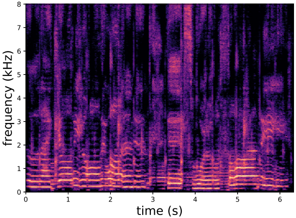
fully-supervised DNN



semi-supervised NMF



proposed



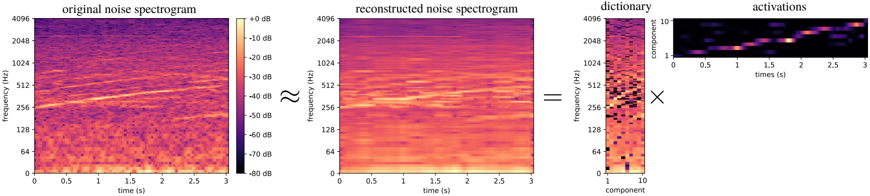
Song: "Sunrise" by Shannon Hurley, from the MGT Music Audio Signal Separation (MASS) dataset.

Extensions

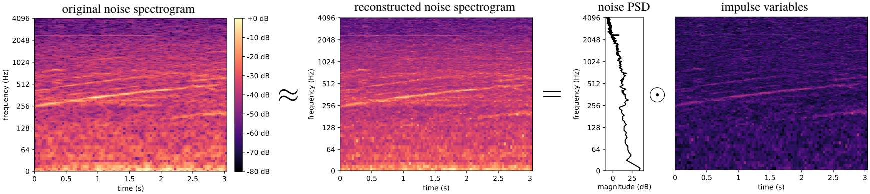
Alpha-stable noise model

Example noise signal recorded within an accelerating subway.

- ▷ Gaussian NMF-based noise model:



- ▷ Alpha-stable noise model:



S. L., U. Şimşekli, A. Liutkus, L. Girin, R. Horaud "Speech enhancement with variational autoencoders and alpha-stable distributions", *IEEE ICASSP, 2019.*

Multi-microphone recording setup

- ▷ A fully-supervised model would need to be retrained. We might even need to collect new data.

- ▷ Our semi-supervised approach can be easily adapted to this new configuration.

Multichannel speech model

Let $\mathbf{s}_{fn} \in \mathbb{C}^I$ be the multichannel speech signal, we have:

$$\mathbf{s}_{fn} \mid \mathbf{z}_n \sim \mathcal{N}_c\left(\mathbf{0}, \sigma_f^2(\mathbf{z}_n) \times \mathbf{R}_{\mathbf{s},f}\right), \qquad \mathbf{z}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{27}$$

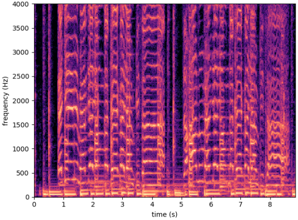
- ▷ $\sigma_f^2(\cdot)$ is learned during the training stage.
- ▷ $\mathbf{R}_{\mathbf{s},f}$ is the spatial covariance matrix and is estimated at test time.

S. L., L. Girin, R. Horaud "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization ", *IEEE ICASSP, 2019.*

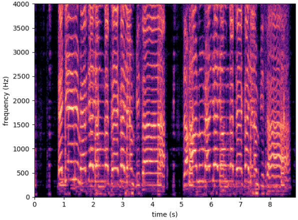
Singing voice separation in a stereo mixture

The VAE model was trained on speaking and not singing voice.

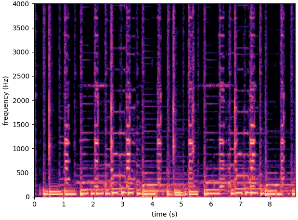
Mixture



Estimated voice



Estimated accompaniment

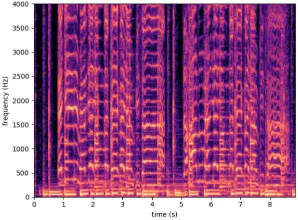


Song: “Ana” by Vieux Farka Toure, from the MGT Music Audio Signal Separation (MASS) dataset.

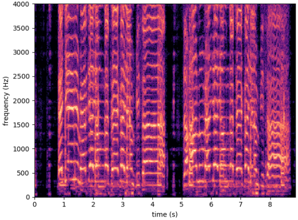
Singing voice separation in a stereo mixture

The VAE model was trained on speaking and not singing voice.

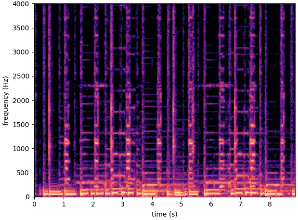
Mixture



Estimated voice



Estimated accompaniment

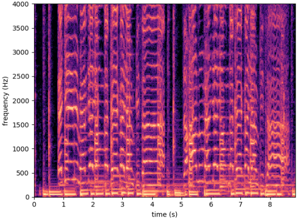


Song: "Ana" by Vieux Farka Toure, from the MGT Music Audio Signal Separation (MASS) dataset.

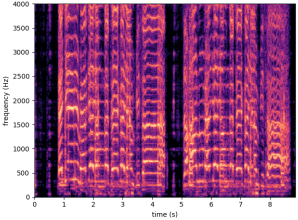
Singing voice separation in a stereo mixture

The VAE model was trained on speaking and not singing voice.

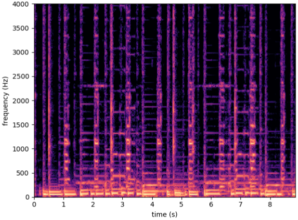
Mixture



Estimated voice



Estimated accompaniment

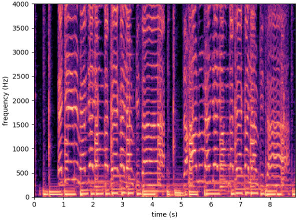


Song: "Ana" by Vieux Farka Toure, from the MGT Music Audio Signal Separation (MASS) dataset.

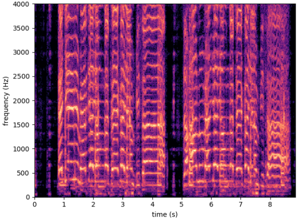
Singing voice separation in a stereo mixture

The VAE model was trained on speaking and not singing voice.

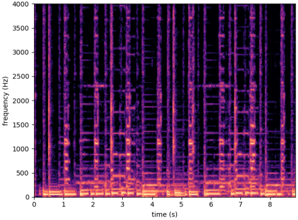
Mixture



Estimated voice



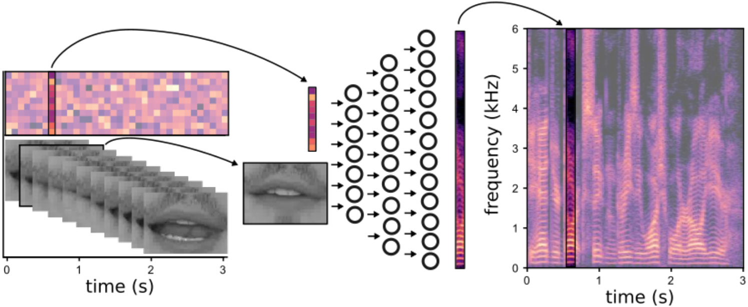
Estimated accompaniment



Song: "Ana" by Vieux Farka Toure, from the MGT Music Audio Signal Separation (MASS) dataset.

Audio-visual speech enhancement

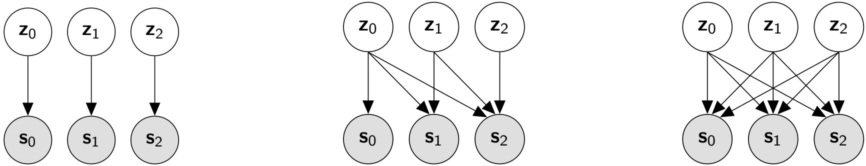
The speech generative process is conditioned on visual information of the lip region, which is invariant to the acoustic noise.



Mostafa Sadeghi, S. L., Xavier Alameda-Pineda, Laurent Girin, Radu Horaud "Audio-visual speech enhancement using conditional variational auto-encoder", *submitted to IEEE Transactions on Audio, Speech and Language Processing, 2019.*

Recurrent deep generative speech model

Generate a sequence of speech STFT time frames from a sequence of latent vectors.



(a) feed-forward NN **(b)** recurrent NN **(c)** bidirectional recurrent NN

Recurrent models induce a temporal dynamic over the reconstructed speech, with Wiener filtering.

S. L., X. Alameda-Pineda, L. Girin, R. Horaud "A recurrent variational autoencoder for speech enhancement ", *submitted to IEEE ICASSP, 2020.*

Conclusion

Conclusion

We combined the learning capabilities of neural networks with the flexibility of probabilistic models for speech enhancement.

- Variational autoencoders are more expressive than NMF models due to their non-linear nature and due to the freedom in the number of trainable parameters.
- Semi-supervised approaches are flexible and can easily adapt to different situations at test time, in terms of noise and number of microphones.

Future work

Some **challenges** that we would like to address:

- to account for phase information;
- to develop deep generative spatial models of multi-microphone signals;
- to encode multi-level and multi-time-scale properties of speech signals in the deep generative process;
- to develop more efficient statistical inference algorithms.

Future work

Some **challenges** that we would like to address:

- to account for phase information;
- to develop deep generative spatial models of multi-microphone signals;
- to encode multi-level and multi-time-scale properties of speech signals in the deep generative process;
- to develop more efficient statistical inference algorithms.

Thank you for your attention

Audio examples and code:
`https://sleglaive.github.io`

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).*

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). “Variational inference: A review for statisticians”. In: *Journal of the American Statistical Association* 112.518.

Cardoso, J.-F. (2001). “The three easy routes to independent component analysis; contrasts and geometry”. In: *Proc. ICA*. Vol. 2001.

Févotte, C. and J. Idier (2011). “Algorithms for nonnegative matrix factorization with the β-divergence”. In: *Neural computation* 23.9.

Févotte, C., L. Daudet, S. J. Godsill, and B. Torrésani (2006). “Sparse regression with structured priors: Application to audio denoising”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP).* IEEE.

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.

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue (1993). “TIMIT Acoustic Phonetic Continuous Speech Corpus”. In: *Linguistic data consortium.*

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

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). “An introduction to variational methods for graphical models”. In: *Machine learning* 37.2.

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

Kowalski, M. and B. Torrésani (2009). “Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients”. In: *Signal, image and video processing* 3.3.

Lee, D. D. and H. S. Seung (1999). “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755.

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

Mysore, G. J. and P. Smaragdis (2011). “A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics”. In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP).*

Ozerov, A., E. Vincent, and F. Bimbot (2012). “A general flexible framework for the handling of prior information in audio source separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 20.4.

Pham, D. T. and P. Garat (1997). “Blind separation of mixture of independent sources through a quasi-maximum likelihood approach”. In: *IEEE transactions on Signal Processing* 45.7.

Smaragdis, P. and J. C. Brown (2003). “Non-negative matrix factorization for polyphonic music transcription”. In: *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA).*

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.*

Vincent, E., M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies (2010). “Probabilistic modeling paradigms for audio source separation”. In: *Machine Audition: Principles, Algorithms and Systems*. Ed. by W. Wang. IGI Global.

Vincent, E., N. Bertin, R. Gribonval, and F. Bimbot (2014). “From blind to guided audio source separation: How models and side information can improve the separation of sound”. In: *IEEE Signal Processing Magazine* 31.3.

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.