A Recurrent Variational Autoencoder for Speech Enhancement

Simon Leglaive[1], Xavier Alameda-Pineda[2], Laurent Girin[3], Radu Horaud[2]

[1]CentraleSuplec, IETR, France [2]Inria Grenoble Rhône-Alpes, France
[3]Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab, France

2020 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)

Introduction

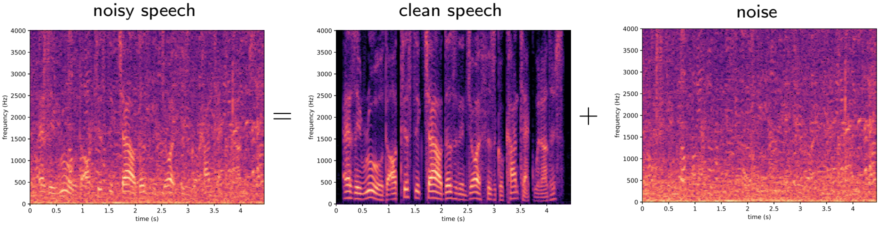
Semi-supervised speech enhancement

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

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

We also want the method to be speaker independent.

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 source 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 source 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:

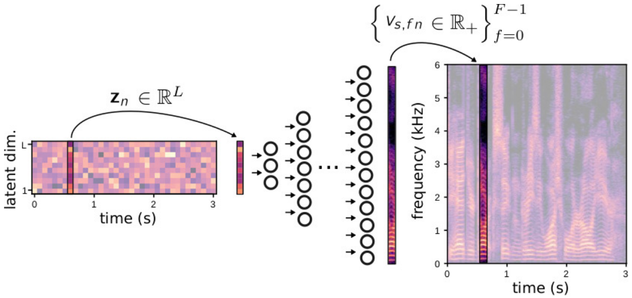
$$x_{fn} \sim \mathcal{N}_c\left(0, v_{s,fn} + v_{b,fn}\right). \tag{3}$$

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

- structured sparsity penalties
  (Févotte et al. 2006; Kowalski and Torrésani 2009)
- non-negative matrix factorization (NMF)
  (Benaroya et al. 2003; Févotte et al. 2009; Ozerov et al. 2012)
- deep generative neural networks
  (Bando et al. 2018)

Deep generative speech model for speech enhancement

It was recently proposed to model the speech variance by a generative neural network (variational autoencoder) (Bando et al. 2018).



- ▷ single-microphone semi-supervised speech enhancement (Bando et al. 2018; Leglaive et al. 2018; Leglaive et al. 2019b; Pariente et al. 2019).
- ▷ multi-microphone semi-supervised speech enhancement (Sekiguchi et al. 2018; Leglaive et al. 2019a; Fontaine et al. 2019; Sekiguchi et al. 2019).

Previous works only considered a feed-forward and fully-connected generative neural network, thus neglecting speech temporal dynamic.

Previous works only considered a feed-forward and fully-connected generative neural network, thus neglecting speech temporal dynamic.

In this work,

- we propose a recurrent VAE speech model trained on clean speech signals;
- at test time, it is combined with an NMF noise model;
- we derive a variational expectation-maximization algorithm where the pre-trained encoder of the VAE is fine-tuned from the noisy mixture signal;
- experiments show that the temporal dynamic induced over the estimated speech signal improves the speech enhancement performance.

Deep generative speech model

Deep generative speech model

- $\mathbf{s} = \mathbf{s}_{0:N-1} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$ is a sequence of N STFT speech time frames.
- $\mathbf{z} = \mathbf{z}_{0:N-1} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$ is a corresponding sequence of N latent random vectors.

Deep generative speech model

- $\mathbf{s} = \mathbf{s}_{0:N-1} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$ is a sequence of N STFT speech time frames.
- $\mathbf{z} = \mathbf{z}_{0:N-1} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$ is a corresponding sequence of N latent random vectors.

Deep generative speech model

Independently for all time frames, in its most general form, we have:

$$\mathbf{s}_n \mid \mathbf{z} \sim \mathcal{N}_c\left(\mathbf{0}, \text{diag}\left\{\mathbf{v}_{\mathbf{s},n}(\mathbf{z})\right\}\right), \quad \text{with} \quad \mathbf{z}_n \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{4}$$

and where $\mathbf{v}_{\mathbf{s},n}(\mathbf{z})$ is provided by a decoder/generative neural network.

Deep generative speech model

- $\mathbf{s} = \mathbf{s}_{0:N-1} = \{\mathbf{s}_n \in \mathbb{C}^F\}_{n=0}^{N-1}$ is a sequence of N STFT speech time frames.
- $\mathbf{z} = \mathbf{z}_{0:N-1} = \{\mathbf{z}_n \in \mathbb{R}^L\}_{n=0}^{N-1}$ is a corresponding sequence of N latent random vectors.

Deep generative speech model

Independently for all time frames, in its most general form, we have:

$$\mathbf{s}_n \mid \mathbf{z} \sim \mathcal{N}_c\left(\mathbf{0}, \text{diag}\left\{\mathbf{v}_{\mathbf{s},n}(\mathbf{z})\right\}\right), \quad \text{with} \quad \mathbf{z}_n \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right), \tag{4}$$

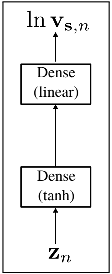
and where $\mathbf{v}_{\mathbf{s},n}(\mathbf{z})$ is provided by a decoder/generative neural network.

Multiple choices can be made to define this neural network, leading to different probabilistic graphical models.

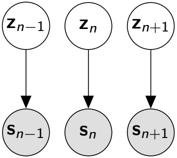
Feed-forward fully-connected neural network (FFNN)

Variance model

$$\mathbf{v}_{\mathbf{s},n}(\mathbf{z}) = \varphi_{\text{dec}}^{\text{FFNN}}(\mathbf{z}_n; \boldsymbol{\theta}_{\text{dec}})$$



Probabilistic graphical model



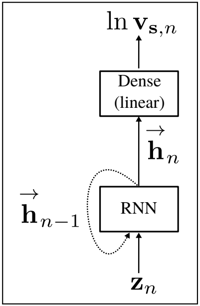
$$p(\mathbf{s}, \mathbf{z}; \boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{s}_n | \mathbf{z}_n; \boldsymbol{\theta}_{\text{dec}}) p(\mathbf{z}_n).$$

The speech STFT time frames are not only conditionally independent, but also marginally independent: $p(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{s}_n; \boldsymbol{\theta}_{\text{dec}})$.

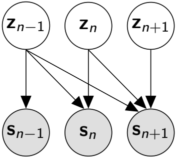
Recurrent neural network (RNN)

Variance model

$$\mathbf{v}_{\mathbf{s},n}(\mathbf{z}) = \varphi^{\mathrm{RNN}}_{\mathrm{dec},n}(\mathbf{z}_{0:n}; \boldsymbol{\theta}_{\mathrm{dec}})$$



Probabilistic graphical model



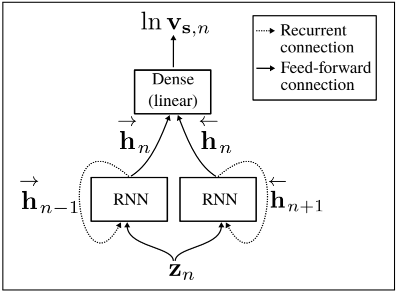
$$p(\mathbf{s}, \mathbf{z}; \boldsymbol{\theta}_{\mathrm{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{s}_n | \mathbf{z}_{0:n}; \boldsymbol{\theta}_{\mathrm{dec}}) p(\mathbf{z}_n),$$

The speech STFT time frames are not marginally independent anymore.

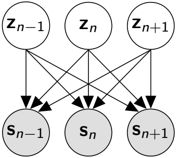
Bidirectional recurrent neural network (BRNN)

Variance model

$$\mathbf{v}_{\mathbf{s},n}(\mathbf{z}) = \varphi^{\text{BRNN}}_{\text{dec},n}(\mathbf{z}; \boldsymbol{\theta}_{\text{dec}})$$



Probabilistic graphical model



$$p(\mathbf{s}, \mathbf{z}; \boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{s}_n | \mathbf{z}; \boldsymbol{\theta}_{\text{dec}}) p(\mathbf{z}_n).$$

The speech STFT time frames are not marginally independent anymore.

Learning the model parameters

Training dataset

$\left\{\mathbf{s}^{(i)} \in \mathbb{C}^{F \times N}\right\}_{i=1}^{I}$: i.i.d sequences of N STFT speech time frames.

Maximum marginal likelihood

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \frac{1}{I} \sum_{i=1}^{I} \ln p\left(\mathbf{s}^{(i)}; \boldsymbol{\theta}_{\text{dec}}\right)$$

Learning the model parameters

Training dataset

$\{\mathbf{s}^{(i)} \in \mathbb{C}^{F \times N}\}_{i=1}^{I}$: i.i.d sequences of N STFT speech time frames.

Maximum marginal likelihood

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \ \ln p\left(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}\right)$$

Learning the model parameters

Training dataset

$\{\mathbf{s}^{(i)} \in \mathbb{C}^{F \times N}\}_{i=1}^{I}$: i.i.d sequences of N STFT speech time frames.

Maximum marginal likelihood

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \ln p\left(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}\right)$$

Intractability issue

$$p(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}) = \int p(\mathbf{s}, \mathbf{z}; \boldsymbol{\theta}_{\text{dec}}) d\mathbf{z}$$

Learning the model parameters

Training dataset

$\{\mathbf{s}^{(i)} \in \mathbb{C}^{F\times N}\}_{i=1}^{I}$: i.i.d sequences of N STFT speech time frames.

Maximum marginal likelihood

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \ \ln p\left(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}\right)$$

Intractability issue

$$p(\mathbf{s}; \boldsymbol{\theta}_{\text{dec}}) = \int p(\mathbf{s}, \mathbf{z}; \boldsymbol{\theta}_{\text{dec}}) d\mathbf{z}$$

Solution

Variational inference (Jordan et al. 1999) + neural networks

= variational autoencoder (VAE) (Kingma and Welling 2014)

Variational lower bound

Variational lower bound

$$\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}})}\left[\ln p\left(\mathbf{s}|\mathbf{z};\boldsymbol{\theta}_{\text{dec}}\right)\right]}_{\text{reconstruction accuracy}} - \underbrace{D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z})\right)}_{\text{regularization}}. \quad (5)$$

Variational lower bound

$$\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}})} \left[\ln p\left(\mathbf{s}|\mathbf{z};\boldsymbol{\theta}_{\text{dec}}\right)\right]}_{\text{reconstruction accuracy}} - \underbrace{D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z})\right)}_{\text{regularization}}. \quad (5)$$

Problem #1

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}})$$

where $\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) \leq \ln p(\mathbf{s};\boldsymbol{\theta}_{\text{dec}})$.

Variational lower bound

$$\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}})}\left[\ln p\left(\mathbf{s}|\mathbf{z};\boldsymbol{\theta}_{\text{dec}}\right)\right]}_{\text{reconstruction accuracy}} - \underbrace{D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z})\right)}_{\text{regularization}}. \quad (5)$$

Problem #1

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}})$$

where $\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) \leq \ln p(\mathbf{s};\boldsymbol{\theta}_{\text{dec}})$.

Problem #2

$$\max_{\boldsymbol{\theta}_{\text{enc}}} \mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}})$$

$$\Leftrightarrow$$

$$\min_{\boldsymbol{\theta}_{\text{enc}}} D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}})\right)$$

Variational lower bound

Variational lower bound

$$\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}})}\left[\ln p\left(\mathbf{s}|\mathbf{z};\boldsymbol{\theta}_{\text{dec}}\right)\right]}_{\text{reconstruction accuracy}} - \underbrace{D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z})\right)}_{\text{regularization}}. \quad (5)$$

Problem #1

$$\max_{\boldsymbol{\theta}_{\text{dec}}} \mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}})$$

where $\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}}) \leq \ln p(\mathbf{s};\boldsymbol{\theta}_{\text{dec}})$.

Problem #2

$$\max_{\boldsymbol{\theta}_{\text{enc}}} \mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}_{\text{enc}}, \boldsymbol{\theta}_{\text{dec}})$$

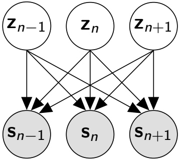
$$\Leftrightarrow$$

$$\min_{\boldsymbol{\theta}_{\text{enc}}} D_{\text{KL}}\left(q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}})\right)$$

$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}})$ is an approximation of the intractable posterior $p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}})$, and it is defined by an encoder/recognition network (Kingma and Welling 2014).

Looking at posterior dependencies

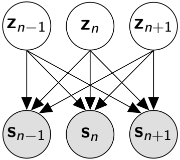
BRNN model



$$p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{z}_n|\mathbf{z}_{0:n-1}, \mathbf{s}_{0:N-1}; \boldsymbol{\theta}_{\text{dec}})$$

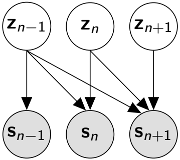
Looking at posterior dependencies

BRNN model



$$p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{z}_n|\mathbf{z}_{0:n-1}, \mathbf{s}_{0:N-1}; \boldsymbol{\theta}_{\text{dec}})$$

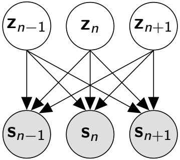
RNN model



$$p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{z}_n|\mathbf{z}_{0:n-1}, \mathbf{s}_{n:N-1}; \boldsymbol{\theta}_{\text{dec}})$$

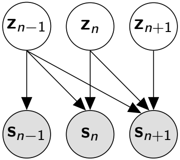
Looking at posterior dependencies

BRNN model



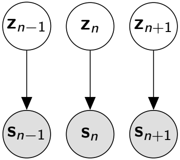
$$p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{z}_n|\mathbf{z}_{0:n-1}, \mathbf{s}_{0:N-1}; \boldsymbol{\theta}_{\text{dec}})$$

RNN model



$$p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{z}_n|\mathbf{z}_{0:n-1}, \mathbf{s}_{n:N-1}; \boldsymbol{\theta}_{\text{dec}})$$

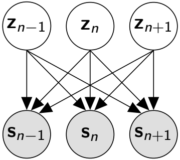
FFNN model



$$p(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{dec}}) = \prod_{n=0}^{N-1} p(\mathbf{z}_n|\mathbf{s}_n; \boldsymbol{\theta}_{\text{dec}})$$

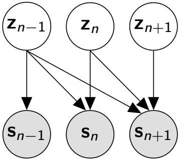
Inference model and encoder network

BRNN model



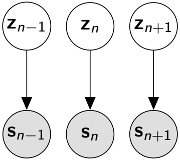
$$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} q(\mathbf{z}_n|\mathbf{z}_{0:n-1},\mathbf{s}_{0:N-1};\boldsymbol{\theta}_{\text{enc}})$$

RNN model



$$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} q(\mathbf{z}_n|\mathbf{z}_{0:n-1},\mathbf{s}_{n:N-1};\boldsymbol{\theta}_{\text{enc}})$$

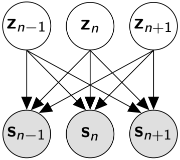
FFNN model



$$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} q(\mathbf{z}_n|\mathbf{s}_n;\boldsymbol{\theta}_{\text{enc}})$$

Inference model and encoder network

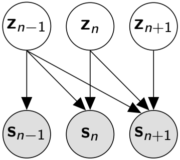
BRNN model



$$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} \mathcal{N}\Big(\mathbf{z}_n; \boldsymbol{\mu}_{\mathbf{z},n}, \text{diag}\left\{\mathbf{v}_{\mathbf{z},n}\right\}\Big)$$

$$\{\boldsymbol{\mu}_{\mathbf{z},n}, \mathbf{v}_{\mathbf{z},n}\} = \varphi_{\text{enc},n}^{\text{BRNN}}(\mathbf{z}_{0:n-1}, \mathbf{s}_{0:N-1}; \boldsymbol{\theta}_{\text{enc}})$$

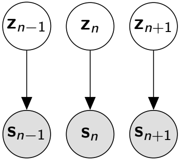
RNN model



$$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} \mathcal{N}\Big(\mathbf{z}_n; \boldsymbol{\mu}_{\mathbf{z},n}, \text{diag}\left\{\mathbf{v}_{\mathbf{z},n}\right\}\Big)$$

$$\{\boldsymbol{\mu}_{\mathbf{z},n}, \mathbf{v}_{\mathbf{z},n}\} = \varphi_{\text{enc},n}^{\text{RNN}}(\mathbf{z}_{0:n-1}, \mathbf{s}_{n:N-1}; \boldsymbol{\theta}_{\text{enc}})$$

FFNN model



$$q(\mathbf{z}|\mathbf{s};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} \mathcal{N}\Big(\mathbf{z}_n; \boldsymbol{\mu}_{\mathbf{z},n}, \text{diag}\left\{\mathbf{v}_{\mathbf{z},n}\right\}\Big)$$

$$\{\boldsymbol{\mu}_{\mathbf{z},n}, \mathbf{v}_{\mathbf{z},n}\} = \varphi_{\text{enc}}^{\text{FFNN}}(\mathbf{s}_n; \boldsymbol{\theta}_{\text{enc}})$$

Training

With this inference model defined, the variational lower bound is completely specified and it can be optimized using gradient-ascent based algorithms.

We used around 25 hours of clean speech data, from the Wall Street Journal (WSJ0) dataset.

Semi-supervised speech enhancement

Models for semi-supervised speech enhancement

Pre-trained deep generative speech model

$$s_{fn} \mid \mathbf{z} \sim \mathcal{N}_c\left(0, v_{\mathbf{s},fn}(\mathbf{z})\right), \qquad \mathbf{z}_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

where $v_{\mathbf{s},fn}$ is the decoder neural network (FFNN, RNN or BRNN) whose parameters were learned during the training phase.

Models for semi-supervised speech enhancement

Pre-trained deep generative speech model

$$s_{fn} \mid \mathbf{z} \sim \mathcal{N}_c\left(0, v_{\mathbf{s},fn}(\mathbf{z})\right), \qquad \mathbf{z}_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

where $v_{\mathbf{s},fn}$ is the decoder neural network (FFNN, RNN or BRNN) whose parameters were learned during the training phase.

NMF-based noise model

$$b_{fn} \sim \mathcal{N}_c\left(0, (\mathbf{W}_b\mathbf{H}_b)_{f,n}\right), \tag{7}$$

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

Models for semi-supervised speech enhancement

Pre-trained deep generative speech model

$$s_{fn} \mid \mathbf{z} \sim \mathcal{N}_c\left(0, v_{\mathbf{s},fn}(\mathbf{z})\right), \qquad \mathbf{z}_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6}$$

where $v_{\mathbf{s},fn}$ is the decoder neural network (FFNN, RNN or BRNN) whose parameters were learned during the training phase.

NMF-based noise model

$$b_{fn} \sim \mathcal{N}_c\left(0, (\mathbf{W}_b\mathbf{H}_b)_{f,n}\right), \tag{7}$$

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

Likelihood

$$x_{fn} \mid \mathbf{z} \sim \mathcal{N}_c\left(0, g_n v_{\mathbf{s},fn}(\mathbf{z}) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}\right), \tag{8}$$

where $g_n \in \mathbb{R}_+$ is a gain parameter (Leglaive et al. 2018).

Speech estimation with Wiener-like filtering

Wiener-like filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn};\phi)}[s_{fn}] = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\phi)}\left[\frac{\sqrt{g_n} v_{\mathbf{s},fn}(\mathbf{z})}{g_n v_{\mathbf{s},fn}(\mathbf{z}) + (\mathbf{W}_b \mathbf{H}_b)_{f,n}}\right] x_{fn}. \tag{9}$$

Speech estimation with Wiener-like filtering

Wiener-like filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn};\phi)}[s_{fn}] = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\phi)}\left[\frac{\sqrt{g_n}v_{\mathbf{s},fn}(\mathbf{z})}{g_n v_{\mathbf{s},fn}(\mathbf{z}) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}}\right] x_{fn}. \quad (9)$$

Two problems:

1. We need to estimate the remaining unknown model parameters:

$$\phi = \{g_0, ..., g_{N-1}, \mathbf{W}_b, \mathbf{H}_b\},$$

   but the marginal likelihood $p(\mathbf{x};\phi)$ is intractable.

2. We need to find an approximation to the intractable posterior $p(\mathbf{z}|\mathbf{x};\phi)$.

Proposed variational EM algorithm

Variational lower bound at test time

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}_{\text{enc}}, \phi) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{\text{enc}})}\left[\ln p(\mathbf{x}|\mathbf{z};\phi)\right] - D_{\text{KL}}\big(q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{\text{enc}}) \parallel p(\mathbf{z})\big), \quad (10)$$

where $q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{\text{enc}})$ corresponds to the pre-trained inference model from the training phase, but now the encoder takes noisy speech as input.

Alternate maximization with respect to $\boldsymbol{\theta}_{\text{enc}}$ (E-Step) and ϕ (M-Step).

Temporal dynamic

Wiener-like filtering

$$\hat{s}_{fn} = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\phi)}\left[\frac{\sqrt{g_n}v_{\mathbf{s},fn}(\mathbf{z})}{g_n v_{\mathbf{s},fn}(\mathbf{z}) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}}\right]x_{fn}$$
$$\approx \mathbb{E}_{q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{\text{enc}})}\left[\frac{\sqrt{g_n}v_{\mathbf{s},fn}(\mathbf{z})}{g_n v_{\mathbf{s},fn}(\mathbf{z}) + (\mathbf{W}_b\mathbf{H}_b)_{f,n}}\right]x_{fn}. \tag{11}$$

The expectation is intractable, so it is approximated by an empirical average using samples drawn from:

$$q(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{\text{enc}}) = \prod_{n=0}^{N-1} q(\mathbf{z}_n|\mathbf{z}_{0:n-1},\mathbf{x};\boldsymbol{\theta}_{\text{enc}}). \tag{12}$$

For the RNN and BRNN generative models, this sampling is done recursively. There is a temporal dynamic that is propagated from the latent vectors to the estimated speech signal, through the expectation in (11).

Alternative E-steps

Monte Carlo E-Step

For the FFNN generative model only, a Markov chain Monte Carlo method was used to sample from the intractable posterior $p(\mathbf{z}|\mathbf{x};\phi)$ (Bando et al. 2018; Leglaive et al. 2018).

"Point-estimate" E-Step

In (Kameoka et al. 2019), it was proposed to only rely on a "point estimate" of the latent variables, based on the maximum a posteriori (MAP):

$$\mathbf{z}^\star = \arg\max_{\mathbf{z}} \big\{ p(\mathbf{z}|\mathbf{x};\phi) \propto p(\mathbf{x}|\mathbf{z};\phi)p(\mathbf{z}) \big\},$$

which can be obtained with gradient-based optimization techniques.

Experiments

Experimental setting

Dataset:

- About 1.5 hours of noisy speech @ 16 kHz using the WSJ0 (unseen speakers) and QUT-NOISE datasets.
- Noise types: {"café", "home", "street", "car"}.
- Signal-to-noise ratios (SNRs): {-5, 0, 5} dB.

Experimental setting

Dataset:

- About 1.5 hours of noisy speech @ 16 kHz using the WSJ0 (unseen speakers) and QUT-NOISE datasets.
- Noise types: {"café", "home", "street", "car"}.
- Signal-to-noise ratios (SNRs): {-5, 0, 5} dB.

Performance measures (higher is better):

- scale-invariant signal-to-distortion ratio (SI-SDR) in dB
- perceptual evaluation of speech quality (PESQ) (between -0.5 and 4.5)
- extended short-time objective intelligibility (ESTOI) (between 0 and 1)

Experimental setting

Dataset:

- About 1.5 hours of noisy speech @ 16 kHz using the WSJ0 (unseen speakers) and QUT-NOISE datasets.
- Noise types: {"café", "home", "street", "car"}.
- Signal-to-noise ratios (SNRs): {-5, 0, 5} dB.

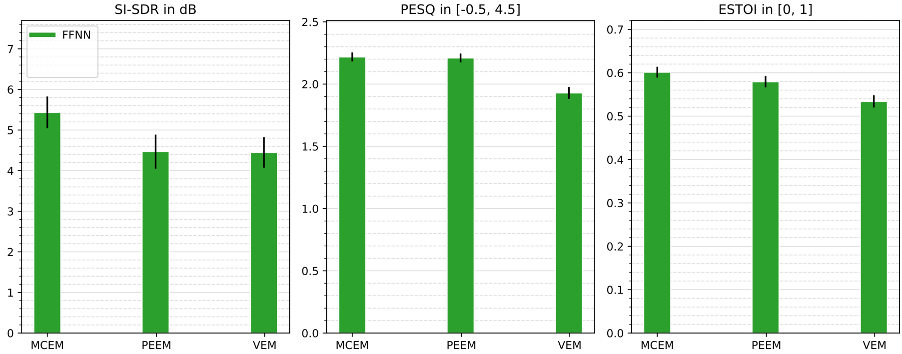
Performance measures (higher is better):

- scale-invariant signal-to-distortion ratio (SI-SDR) in dB
- perceptual evaluation of speech quality (PESQ) (between -0.5 and 4.5)
- extended short-time objective intelligibility (ESTOI) (between 0 and 1)

Methods:

- Monte Carlo EM `MCEM - FFNN`
- "Point-estimate" EM `PEEM - {FFNN, RNN, BRNN}`
- Proposed variational EM `VEM - {FFNN, RNN, BRNN}`

Results



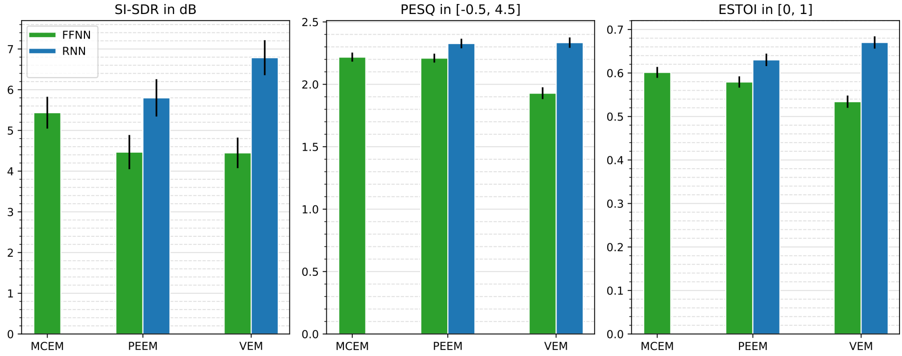
input: -2.6 ± 0.5
oracle Wiener: 12.1 ± 0.3

input: 1.82 ± 0.03
oracle Wiener: 3.13 ± 0.02

input: 0.49 ± 0.01
oracle Wiener: 0.88 ± 0.01

▷ For the FFNN generative model, the MCEM algorithm gives the best results.

Results



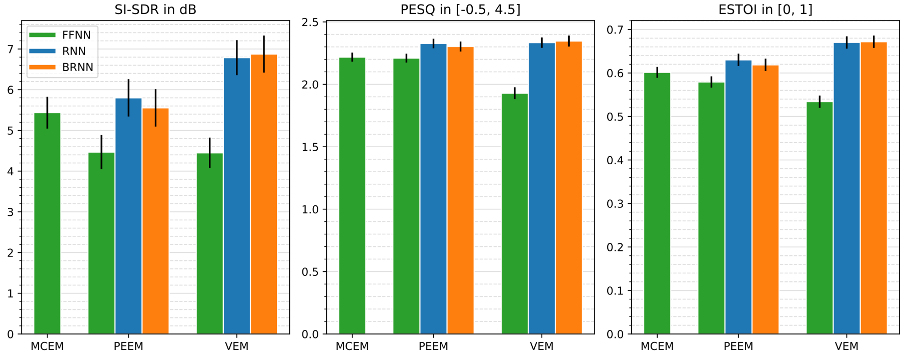
input: -2.6 ± 0.5
oracle Wiener: 12.1 ± 0.3

input: 1.82 ± 0.03
oracle Wiener: 3.13 ± 0.02

input: 0.49 ± 0.01
oracle Wiener: 0.88 ± 0.01

- For the FFNN generative model, the MCEM algorithm gives the best results.
- The RNN model outperforms the FFNN model.
- The VEM algorithm outperforms the PEEM algorithm.

Results



input: -2.6 ± 0.5	input: 1.82 ± 0.03	input: 0.49 ± 0.01
oracle Wiener: 12.1 ± 0.3	oracle Wiener: 3.13 ± 0.02	oracle Wiener: 0.88 ± 0.01

- For the FFNN generative model, the MCEM algorithm gives the best results.
- The RNN model outperforms the FFNN model.
- The VEM algorithm outperforms the PEEM algorithm.
- The BRNN model does not perform significantly better than the RNN model.

Conclusion

- We combined a recurrent VAE with an NMF noise model for semi-supervised speech enhancement.
- The inference model (encoder network) should be carefully designed in order to preserve posterior temporal dependencies between the latent variables.
- The temporal dynamic induced over the estimated speech signal is beneficial in terms of speech enhancement results.

Audio examples and code:

`https://sleglaive.github.io/demo-icassp2020.html`

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., 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.

Fontaine, M., A. A. Nugraha, R. Badeau, K. Yoshii, and A. Liutkus (2019). "Cauchy Multichannel Speech Enhancement with a Deep Speech Prior". In: *Proc. European Signal Processing Conference (EUSIPCO)*.

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.

Kameoka, H., L. Li, S. Inoue, and S. Makino (2019). "Supervised Determined Source Separation with Multichannel Variational Autoencoder". In: *Neural Computation* 31.9.

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.

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).

— (2019a). "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization". In: *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*.

Leglaive, S., U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud (2019b). "Speech enhancement with variational autoencoders and alpha-stable distributions". 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.

Pariente, M., A. Deleforge, and E. Vincent (2019). "A Statistically Principled and Computationally Efficient Approach to Speech Enhancement using Variational Autoencoders". In: *Proc. Interspeech*.

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.

Sekiguchi, K., Y. Bando, K. Yoshii, and T. Kawahara (2018). "Bayesian Multichannel Speech Enhancement with a Deep Speech Prior". In: *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Sekiguchi, K., Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara (2019). "Semi-Supervised Multichannel Speech Enhancement With a Deep Speech Prior". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.12.

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.