A Recurrent Variational Autoencoder for Speech Enhancement

Simon LeGlaive\textsuperscript{1}, Xavier Alameda-Pineda\textsuperscript{2}, Laurent Girin\textsuperscript{3}, Radu Horaud\textsuperscript{2}

\textsuperscript{1}CentraleSupélec, IETR, France \hspace{1cm} \textsuperscript{2}Inria Grenoble Rhône-Alpes, France
\hspace{1cm} \textsuperscript{3}Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab, France

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

\[ x_{fn} = s_{fn} + b_{fn}, \]  

- \( s_{fn} \in \mathbb{C} \) 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.
Independently for all \((f, n) \in \mathcal{B}\):

\[
s_{fn} \sim \mathcal{N}_c(0, v_{s,fn}) \quad \perp \quad b_{fn} \sim \mathcal{N}_c(0, v_{b,fn}).
\] (2)

Consequently, we also have:

\[
x_{fn} \sim \mathcal{N}_c(0, v_{s,fn} + v_{b,fn}).
\] (3)
Non-stationary Gaussian source model (Pham and Garat 1997; Cardoso 2001)

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

\[
s_{fn} \sim \mathcal{N}_c(0, \nu_{s,fn}) \quad \perp \quad b_{fn} \sim \mathcal{N}_c(0, \nu_{b,fn}).
\]  

Consequently, we also have:

\[
x_{fn} \sim \mathcal{N}_c\left(0, \nu_{s,fn} + \nu_{b,fn}\right).
\]

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

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

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

$$s_n|z \sim N(c(0, \text{diag}\{v_{sn}(z)\}), \text{with } z_n \text{i.i.d } \sim N(0, I))$$

and where $v_{sn}(z)$ is provided by a decoder/generative neural network.

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

$\mathbf{s} = \mathbf{s}_{0:N-1} = \{s_n \in \mathbb{C}^F\}_{n=0}^{N-1}$ is a sequence of $N$ STFT speech time frames.

$\mathbf{z} = \mathbf{z}_{0:N-1} = \{z_n \in \mathbb{R}^L\}_{n=0}^{N-1}$ is a corresponding sequence of $N$ latent random vectors.

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

$$s_n \mid z \sim \mathcal{N}_c(0, \text{diag}\{v_{s,n}(z)\}), \quad \text{with} \quad z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I),$$

(4)

and where $v_{s,n}(z)$ is provided by a decoder/generative neural network.
Deep generative speech model

\( s = s_{0:N-1} = \{s_n \in \mathbb{C}^F\}_{n=0}^{N-1} \) is a sequence of \( N \) STFT speech time frames.

\( z = z_{0:N-1} = \{z_n \in \mathbb{R}^L\}_{n=0}^{N-1} \) is a corresponding sequence of \( N \) latent random vectors.

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

\[
\begin{align*}
    s_n \mid z & \sim \mathcal{N}_c(0, \text{diag } \{v_{s,n}(z)\}), \quad \text{with} \quad z_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I),
\end{align*}
\]

(4)

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

Multiple choices can be made to define this neural network, leading to different probabilistic graphical models.
The speech STFT time frames are not only conditionally independent, but also marginally independent: 

$$p(s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(s_{n} | z_{n}; \theta_{\text{dec}}) p(z_{n}).$$
Recurrent neural network (RNN)

Variance model

\[ \mathbf{v}_{s,n}(\mathbf{z}) = \varphi_{\text{dec},n}^{\text{RNN}}(\mathbf{z}_{0:n}; \theta_{\text{dec}}) \]

Probabilistic graphical model

\[
\ln \mathbf{v}_{s,n} \quad \rightarrow \quad \mathbf{h}_n \quad \rightarrow \quad \mathbf{h}_{n-1} \quad \mathbf{z}_n
\]

\[
\begin{align*}
\rightarrow \quad \text{Dense (linear)} & \\
\rightarrow \quad \text{RNN} & \\
\end{align*}
\]

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

The speech STFT time frames are **not marginally independent** anymore.
Bidirectional recurrent neural network (BRNN)

**Variance model**

\[ \nu_{s,n}(z) = \varphi_{\text{dec},n}^{\text{BRNN}}(z; \theta_{\text{dec}}) \]

**Probabilistic graphical model**

\[ p(s, z; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(s_n|z; \theta_{\text{dec}}) p(z_n). \]

The speech STFT time frames are not marginally independent anymore.
Learning the model parameters

**Training dataset**

\[ \{ s^{(i)} \in \mathbb{C}^{F \times N} \}_{i=1}^{l} : \text{i.i.d sequences of } N \text{ STFT speech time frames.} \]

**Maximum marginal likelihood**

\[ \max_{\theta_{\text{dec}}} \frac{1}{l} \sum_{i=1}^{l} \ln p \left( s^{(i)} ; \theta_{\text{dec}} \right) \]
Learning the model parameters

Training dataset

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

Maximum marginal likelihood

\[
\max_{\theta_{\text{dec}}} \ln p(s; \theta_{\text{dec}})
\]
Learning the model parameters

Training dataset

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

Maximum marginal likelihood

\[
\max_{\theta_{\text{dec}}} \ln p(s; \theta_{\text{dec}})
\]

Intractability issue

\[
p(s; \theta_{\text{dec}}) = \int p(s, z; \theta_{\text{dec}}) dz
\]

Solution

Variational inference (Jordan et al. 1999) + neural networks = variational autoencoder (VAE) (Kingma and Welling 2014)
Learning the model parameters

**Training dataset**

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

**Maximum marginal likelihood**

\[
\max_{\theta_{\text{dec}}} \ln p( s; \theta_{\text{dec}} )
\]

**Intractability issue**

\[
p(s; \theta_{\text{dec}}) = \int p(s, z; \theta_{\text{dec}}) dz
\]

**Solution**

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

= variational autoencoder (VAE) (Kingma and Welling 2014)
Variational lower bound

\[ \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathbb{E}_{q(z|s; \theta_{\text{enc}})} \left[ \ln p(s|z; \theta_{\text{dec}}) \right] - D_{KL} \left( q(z|s; \theta_{\text{enc}}) \parallel p(z) \right). \]  

(5)

The variational lower bound consists of three components: reconstruction accuracy, denoted by the expectation term inside the logarithm, and two regularization terms: the Kullback-Leibler divergence, denoted by the KL divergence term, and the hyperparameter \( \theta_{\text{dec}} \).

Problem #1

\[ \max_{\theta_{\text{dec}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \]

where

\[ \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \leq \ln p(s; \theta_{\text{dec}}). \]

Problem #2

\[ \max_{\theta_{\text{enc}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \iff \min_{\theta_{\text{enc}}} D_{KL} \left( q(z|s; \theta_{\text{enc}}) \parallel p(z) \right) \]

\( q(z|s; \theta_{\text{enc}}) \) is an approximation of the intractable posterior \( p(z|s; \theta_{\text{dec}}) \), and it is defined by an encoder/recognition network (Kingma and Welling 2014).
Variational lower bound

\[ \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathbb{E}_{q(z|s; \theta_{\text{enc}})} \left[ \ln p(s|z; \theta_{\text{dec}}) \right] - D_{\text{KL}} \left( q(z|s; \theta_{\text{enc}}) \parallel p(z) \right). \]  

(5)

reconstruction accuracy

regularization

Problem #1

\[ \max_{\theta_{\text{dec}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \]

where \( \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \leq \ln p(s; \theta_{\text{dec}}). \)
Variational lower bound

\[ \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathbb{E}_{q(z|s; \theta_{\text{enc}})} \left[ \ln p(s|z; \theta_{\text{dec}}) \right] - D_{\text{KL}}(q(z|s; \theta_{\text{enc}}) \mid\mid p(z)). \quad (5) \]

reconstruction accuracy

regularization

Problem #1

\[
\max_{\theta_{\text{dec}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}})
\]

where \( \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \leq \ln p(s; \theta_{\text{dec}}). \)

Problem #2

\[
\max_{\theta_{\text{enc}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \iff \min_{\theta_{\text{enc}}} D_{\text{KL}}(q(z|s; \theta_{\text{enc}}) \mid\mid p(z|s; \theta_{\text{dec}}))
\]
Variational lower bound

\[
\mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) = \mathbb{E}_{q(z|s; \theta_{\text{enc}})} \left[ \ln p(s|z; \theta_{\text{dec}}) \right] - D_{\text{KL}} \left( q(z|s; \theta_{\text{enc}}) \parallel p(z) \right). 
\]

(5)

**Problem #1**

\[
\max_{\theta_{\text{dec}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}})
\]

where \( \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \leq \ln p(s; \theta_{\text{dec}}). \)

**Problem #2**

\[
\max_{\theta_{\text{enc}}} \mathcal{L}_s(\theta_{\text{enc}}, \theta_{\text{dec}}) \iff \\
\min_{\theta_{\text{enc}}} D_{\text{KL}} \left( q(z|s; \theta_{\text{enc}}) \parallel p(z|s; \theta_{\text{dec}}) \right)
\]

\( q(z|s; \theta_{\text{enc}}) \) is an approximation of the intractable posterior \( p(z|s; \theta_{\text{dec}}) \), and it is defined by an **encoder/recognition network** (Kingma and Welling 2014).
Looking at posterior dependencies

BRNN model

\[
p(z|s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(z_n|z_0:n-1, s_0:N-1; \theta_{\text{dec}})
\]
Looking at posterior dependencies

**BRNN model**

\[
p(z|s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(z_n|z_{0:n-1}, s_0:N-1; \theta_{\text{dec}})
\]

**RNN model**

\[
p(z|s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(z_n|z_{0:n-1}, s_n:N-1; \theta_{\text{dec}})
\]
Looking at posterior dependencies

BRNN model

\[ p(z|s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(z_n|z_{0:n-1}, s_{0:N-1}; \theta_{\text{dec}}) \]

RNN model

\[ p(z|s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(z_n|z_{0:n-1}, s_{n:N-1}; \theta_{\text{dec}}) \]

FFNN model

\[ p(z|s; \theta_{\text{dec}}) = \prod_{n=0}^{N-1} p(z_n|s_n; \theta_{\text{dec}}) \]
Inference model and encoder network

BRNN model

\[ q(z|s; \theta_{\text{enc}}) = \prod_{n=0}^{N-1} q(z_n|z_{0:n-1}, s_{0:N-1}; \theta_{\text{enc}}) \]

RNN model

\[ q(z|s; \theta_{\text{enc}}) = \prod_{n=0}^{N-1} q(z_n|z_{0:n-1}, s_n:N-1; \theta_{\text{enc}}) \]

FFNN model

\[ q(z|s; \theta_{\text{enc}}) = \prod_{n=0}^{N-1} q(z_n|s_n; \theta_{\text{enc}}) \]
Inference model and encoder network

**BRNN model**

\[
q(z|s; \theta_{\text{enc}}) = \prod_{n=0}^{N-1} \mathcal{N}(z_n; \mu_{z,n}, \text{diag} \{v_{z,n}\})
\]

\[
\{\mu_{z,n}, v_{z,n}\} = \varphi_{\text{enc},n}^{\text{BRNN}}(z_0:n-1, s_0:N-1; \theta_{\text{enc}})
\]

**RNN model**

\[
q(z|s; \theta_{\text{enc}}) = \prod_{n=0}^{N-1} \mathcal{N}(z_n; \mu_{z,n}, \text{diag} \{v_{z,n}\})
\]

\[
\{\mu_{z,n}, v_{z,n}\} = \varphi_{\text{enc},n}^{\text{RNN}}(z_0:n-1, s_n:N-1; \theta_{\text{enc}})
\]

**FFNN model**

\[
q(z|s; \theta_{\text{enc}}) = \prod_{n=0}^{N-1} \mathcal{N}(z_n; \mu_{z,n}, \text{diag} \{v_{z,n}\})
\]

\[
\{\mu_{z,n}, v_{z,n}\} = \varphi_{\text{enc}}^{\text{FFNN}}(s_n; \theta_{\text{enc}})
\]
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

\[ s_{fn} \mid z \sim \mathcal{N}_c (0, v_{s,fn}(z)), \quad z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I), \]  

where \( v_{s,fn} \) 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

\[ s_{fn} \mid z \sim \mathcal{N}_c \left(0, \nu_{s,fn}(z) \right), \quad z_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I), \quad (6) \]

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

NMF-based noise model

\[ b_{fn} \sim \mathcal{N}_c \left(0, (W_b H_b)_{f,n} \right), \quad (7) \]

where \( W_b \in \mathbb{R}_{+}^{F \times K_b} \) and \( H_b \in \mathbb{R}_{+}^{K_b \times N} \).
Models for semi-supervised speech enhancement

Pre-trained deep generative speech model

\[
s_{fn} \mid z \sim \mathcal{N}_c \left(0, \nu_{s,fn}(z) \right), \quad z_n \overset{i.i.d}{\sim} \mathcal{N}(0, I),
\]

(6)

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

NMF-based noise model

\[
b_{fn} \sim \mathcal{N}_c \left(0, (W_bH_b)_{f,n} \right),
\]

(7)

where \( W_b \in \mathbb{R}^{F \times K_b} \) and \( H_b \in \mathbb{R}^{K_b \times N} \).

Likelihood

\[
\chi_{fn} \mid z \sim \mathcal{N}_c \left(0, g_n \nu_{s,fn}(z) + (W_bH_b)_{f,n} \right),
\]

(8)

where \( g_n \in \mathbb{R}_+ \) is a gain parameter (Leglaive et al. 2018).
Speech estimation with Wiener-like filtering

Wiener-like filtering

\[ \hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x_{fn};\phi)} \{s_{fn} \} = \mathbb{E}_{p(z|x;\phi)} \left[ \frac{\sqrt{g_n v_{s,fn}(z)}}{g_n v_{s,fn}(z) + (W_b H_b)_{f,n}} \right] x_{fn}. \]
Speech estimation with Wiener-like filtering

Wiener-like filtering

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

Two problems:

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

\[\phi = \{g_0, ..., g_{N-1}, W_b, H_b\},\]

but the marginal likelihood \(p(x;\phi)\) is intractable.

2. We need to find an approximation to the intractable posterior \(p(z|x;\phi)\).
Proposed variational EM algorithm

Variational lower bound at test time

\[ \mathcal{L}_x(\theta_{enc}, \phi) = \mathbb{E}_{q(z|x; \theta_{enc})} [\ln p(x|z; \phi)] - D_{KL} (q(z|x; \theta_{enc}) \parallel p(z)), \quad (10) \]

where \( q(z|x; \theta_{enc}) \) 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 \( \theta_{enc} \) (E-Step) and \( \phi \) (M-Step).
Temporal dynamic

Wiener-like filtering

\[
\hat{s}_{fn} = \mathbb{E}_{p(z|x;\phi)} \left[ \frac{\sqrt{g_n v_{s,fn}(z)}}{g_n v_{s,fn}(z) + (W_b H_b)_{f,n}} \right] x_{fn}
\]

\[
\approx \mathbb{E}_{q(z|x;\theta_{enc})} \left[ \frac{\sqrt{g_n v_{s,fn}(z)}}{g_n v_{s,fn}(z) + (W_b H_b)_{f,n}} \right] x_{fn}.
\] (11)

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

\[
q(z|x;\theta_{enc}) = \prod_{n=0}^{N-1} q(z_n|z_{0:n-1}, x;\theta_{enc}).
\] (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).
**Monte Carlo E-Step**

For the FFNN generative model only, a Markov chain Monte Carlo method was used to sample from the intractable posterior $p(z|x; \phi)$ (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):

$$z^* = \arg \max_z \{ p(z|x; \phi) \propto p(x|z; \phi)p(z) \},$$

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: {”caf´e”, ”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

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

- SI-SDR in dB:
  - MCEM: -2.6 ± 0.5
  - PEEM: 5.4 ± 0.3
  - VEM: 5.3 ± 0.2

- PESQ in [-0.5, 4.5]:
  - MCEM: 2.2 ± 0.05
  - PEEM: 2.1 ± 0.05
  - VEM: 2.0 ± 0.05

- ESTOI in [0, 1]:
  - MCEM: 0.8 ± 0.01
  - PEEM: 0.8 ± 0.01
  - VEM: 0.8 ± 0.01

- Oracle Wiener:
  - SI-SDR: 12.1 ± 0.3
  - PESQ: 3.13 ± 0.02
  - ESTOI: 0.88 ± 0.01
Results

- 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.
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.
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:


