Speech Enhancement with Variational Autoencoders and Alpha-Stable Distributions

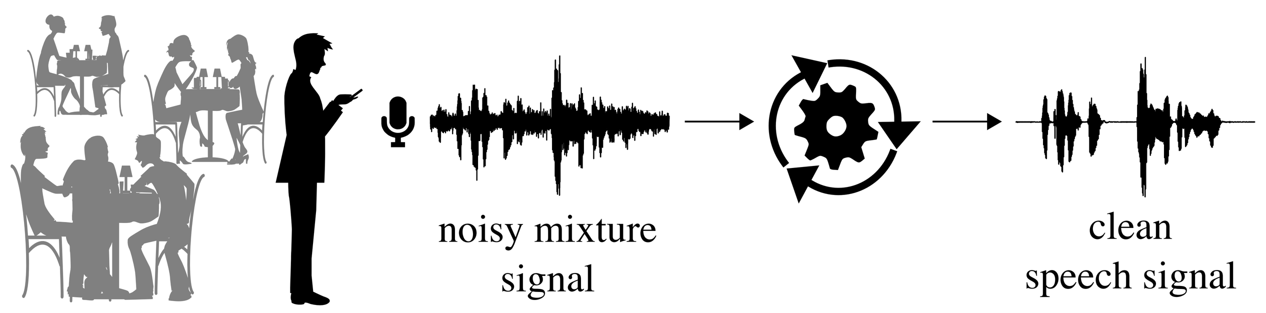
Simon Leglaive[1] Umut Şimşekli[2] Antoine Liutkus[1,3] Laurent Girin[1,4] Radu Horaud[1]

1: Inria 2: LTCI, Télécom ParisTech, Université Paris-Saclay 3: LIRMM 4: Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019

1. Introduction

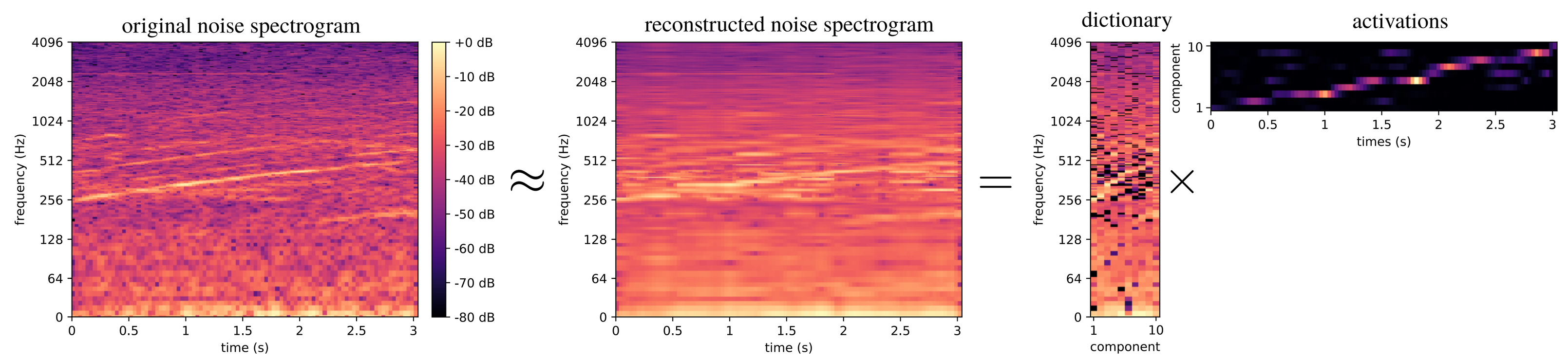
Speech enhancement aims to recover a clean speech signal from a noisy mixture signal.



Motivation

Gaussian noise modeling based on nonnegative matrix factorization (NMF) is common in semi-supervised speech enhancement, but it is limiting for certain types of noise.

Example noise signal recorded inside an accelerating subway:

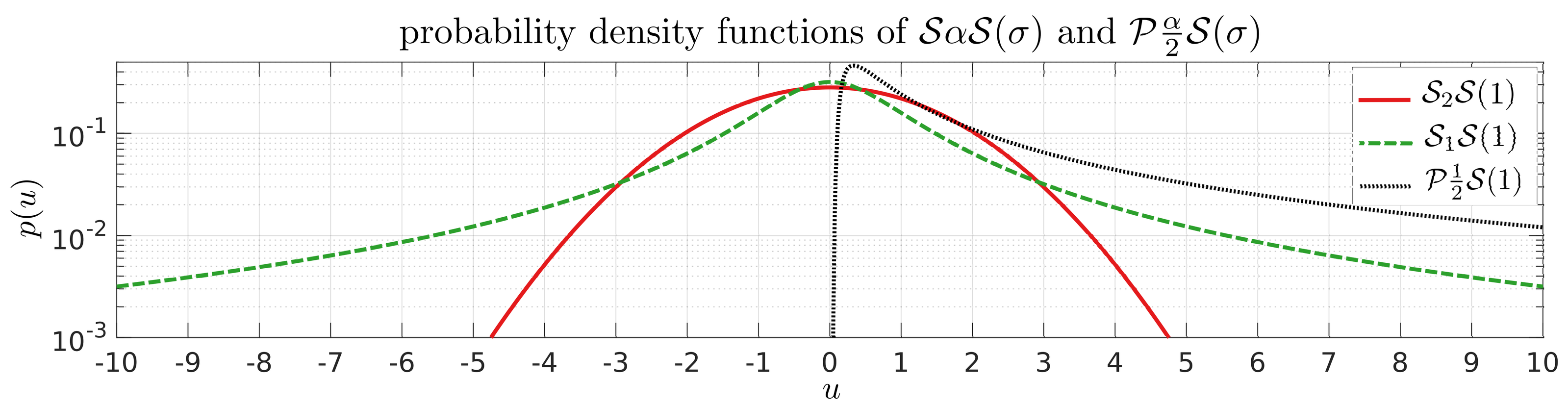


Contribution

- ▷ Flexible alpha-stable noise model combined with a deep generative speech model for semi-supervised speech enhancement.
- ▷ Monte Carlo expectation-maximization (MCEM) algorithm.
- ▷ Outperforms the counterpart approach [1] based on Gaussian noise modeling with NMF variance parametrization.

2. Symmetric and positive alpha-stable distributions

Alpha-stable distributions are *heavy-tailed* distributions.



$\alpha \in]0,2]$ is the characteristic exponent and $\sigma \in \mathbb{R}_+$ the scale parameter.

For $\alpha = 2$, we recover the Gaussian distribution: $\mathcal{S}_2\mathcal{S}(\sigma) = \mathcal{N}(0, 2\sigma^2)$.

References

[1] S. Leglaive *et al.*, "A variance modeling framework based on variational autoencoders for speech enhancement", *IEEE MLSP*, 2018.

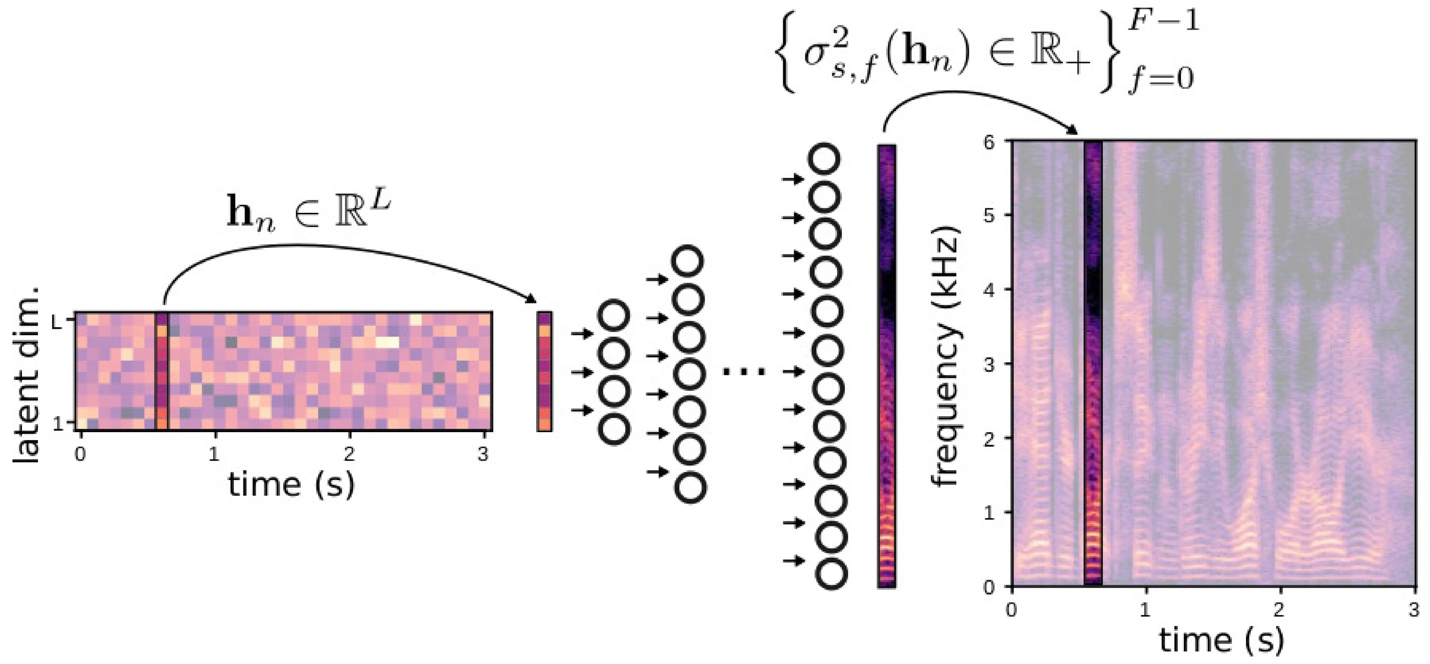
[2] Y. Bando *et al.*, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization", *IEEE ICASSP*, 2018.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes", *ICLR*, 2014.

3. Supervised deep generative speech model

In the short-term Fourier transform domain, independently for all $(f,n) \in \mathbb{B} = \{0,...,F-1\} \times \{0,...,N-1\}$ we have [1, 2]:

$$s_{fn} \mid \mathbf{h}_n \sim \mathcal{N}_c(0, \sigma^2_{s,f}(\mathbf{h}_n)), \qquad \text{where } \mathbf{h}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$



The parameters $\boldsymbol{\theta}_s$ of the neural network are learned from a dataset of clean speech signals. They are estimated by maximizing a lower bound of the log-likelihood, in the framework of variational autoencoders [3].

4. Unsupervised alpha-stable noise model

Marginal circularly symmetric alpha-stable noise model

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

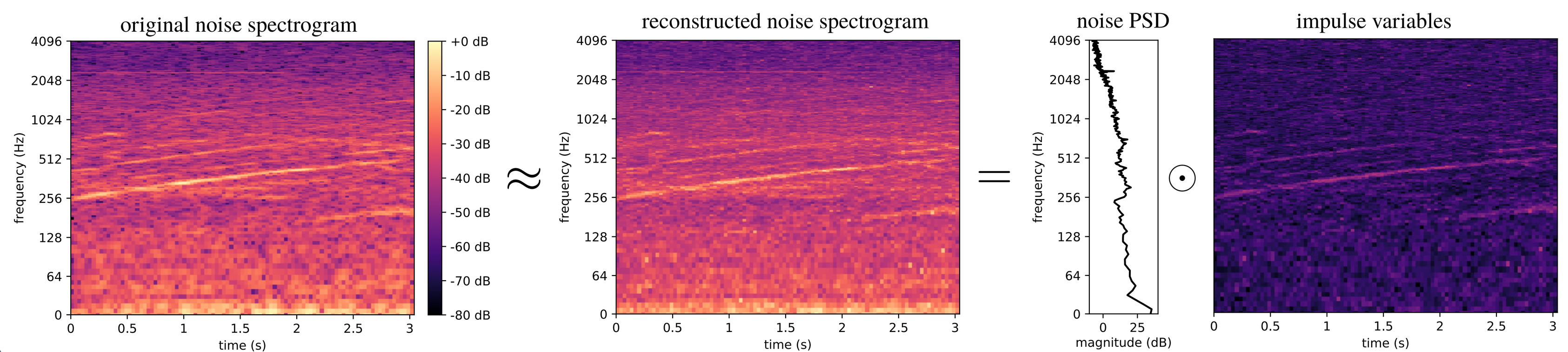
$$b_{fn} \sim \mathcal{S}\alpha\mathcal{S}_c(\sigma_{b,f}),$$

where $\sigma^2_{b,f}$ can be understood as the noise power spectral density (PSD).

Equivalent conditionally Gaussian noise model

$$b_{fn} \mid \phi_{fn} \sim \mathcal{N}_c(0, \phi_{fn}\sigma^2_{b,f}), \qquad \text{where } \phi_{fn} \sim \mathcal{P}\frac{\alpha}{2}\mathcal{S}\Big(2\cos(\pi\alpha/4)^{2/\alpha}\Big).$$

$\phi_{fn} \in \mathbb{R}_+$ is an *impulse variable carrying uncertainty* about the stationarity assumption of the marginal noise model.



5. Mixture model

The observed mixture signal is modeled as follows:

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

where $g_n \in \mathbb{R}_+$ represents a frame-dependent gain. We further consider the conditional independence of the speech and noise signals so that:

$$x_{fn} \mid \mathbf{h}_n, \phi_{fn} \sim \mathcal{N}_c\left(0, g_n\sigma^2_{s,f}(\mathbf{h}_n) + \phi_{fn}\sigma^2_{b,f}\right).$$

6. Inference

▷ Unsupervised model parameters to be estimated:

$$\boldsymbol{\theta}_u = \left\{\mathbf{g} = \{g_n \in \mathbb{R}_+\}_{n=0}^{N-1}, \boldsymbol{\sigma}_b^2 = \{\sigma^2_{b,f} \in \mathbb{R}_+\}_{f=0}^{F-1}\right\}$$

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

▷ Latent variables: $\mathbf{z} = \left\{\mathbf{h}_n, \{\phi_{fn}\}_{f=0}^{F-1}\right\}_{n=0}^{N-1}$

MCEM algorithm

From an initialization $\boldsymbol{\theta}_u^\star$ of the parameters, iterate:

- **E-Step:** $Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_s,\boldsymbol{\theta}_u^\star)}[\ln p(\mathbf{x},\mathbf{z};\boldsymbol{\theta}_s,\boldsymbol{\theta}_u)]$.
  Intractable expectation → Markov chain Monte Carlo method.
- **M-Step:** $\boldsymbol{\theta}_u^\star \leftarrow \arg\max_{\boldsymbol{\theta}_u} Q(\boldsymbol{\theta}_u; \boldsymbol{\theta}_u^\star)$.
  Minorize-maximize approach leading to multiplicative update rules.

Posterior mean speech estimate with Wiener-like filtering

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

$$\hat{\tilde{s}}_{fn} = \mathbb{E}_{p(\tilde{s}_{fn}|\mathbf{x};\boldsymbol{\theta}_u,\boldsymbol{\theta}_s)}[\tilde{s}_{fn}] = \mathbb{E}_{p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_u,\boldsymbol{\theta}_s)}\left[\frac{g_n\sigma^2_{s,f}(\mathbf{h}_n)}{g_n\sigma^2_{s,f}(\mathbf{h}_n) + \phi_{fn}\sigma^2_{b,f}}\right] x_{fn}.$$

7. Reference method [1]

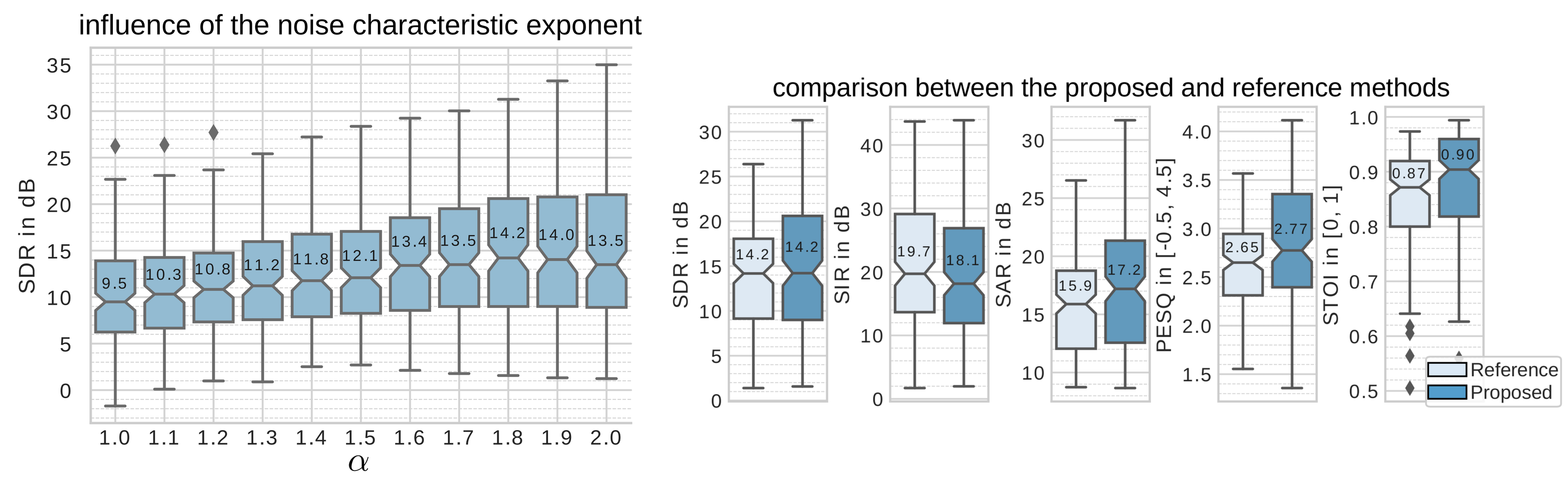
The speech model is the same, only the noise model differs. It is Gaussian with NMF-based variance parametrization:

$$b_{fn} \sim \mathcal{N}_c(0, (\mathbf{WH})_{f,n}),$$

where both $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ are estimated from the noisy mixture signal using an MCEM algorithm.

8. Experiments

- ▷ Training set (∼ 4 hours): 462 speakers × 10 sentences × 3 seconds.
- ▷ Test set: 168 noisy mixtures (∼ 3 seconds) at a 0 dB SNR.
- ▷ Noise types: Domestic or office environments, nature, indoor public spaces, street, transportation.



- ▷ (SDR, SIR, SAR) measure (global quality, interferences, artifacts).
- ▷ (PESQ, STOI) measure (perceptual quality, intelligibility).