A Weighted-Variance Variational Autoencoder Model for Speech Enhancement

Ali Golmakani[1], Mostafa Sadeghi[1], Romain Serizel[1], and Xavier Alameda-Pineda[2]

[1]Multispeech team, Inria Nancy - Grand Est, France, [2]RobotLearn team, Inria Grenoble Rhône-Alpes, France

Overview

- We address **unsupervised speech enhancement** (SE).
- A **weighted-variance** variational autoencoder (**VAE**) model is proposed as the speech generative model.
- A **Gamma prior** distribution is imposed on the weights, leading to a **Student's t-distribution** for time-frequency elements.
- **Efficient training and speech enhancement** algorithms are developed.
- Experimental results demonstrate the **effectiveness** and **robustness** of the proposed approach compared to the standard unweighted variance model.

Unsupervised speech enhancement

Separate the speech and noise signals without training on noise.

Short-time Fourier transform (STFT) domain: $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{b}$

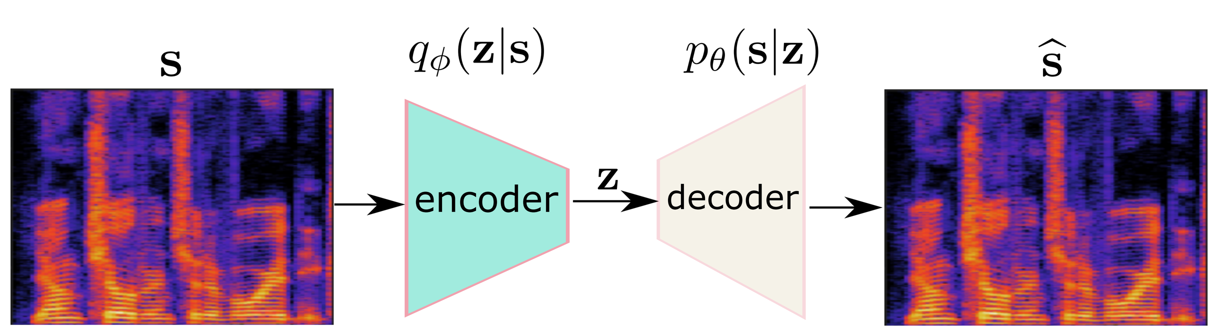
- $\boldsymbol{s} \to$ clean speech signal with prior $p_\theta(\boldsymbol{s})$
- $\boldsymbol{b} \to$ noise signal with prior $p_\psi(\boldsymbol{b})$

Training: Learn a parametric prior $p_\theta(\boldsymbol{s})$
Testing: Estimate $\boldsymbol{s}$ using $p_\psi(\boldsymbol{s}|\boldsymbol{x}) \propto p_\psi(\boldsymbol{x}|\boldsymbol{s}) \times p_\theta(\boldsymbol{s})$

Training: learning speech prior

VAE-based speech generative model for $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_T\}$ [1]:

$$p_\theta(\boldsymbol{s}) = \int p_\theta(\boldsymbol{s}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}, \quad p(\boldsymbol{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$



$$\begin{cases} p_\theta(\mathbf{s}_t|\mathbf{z}_t) = \mathcal{N}_c\Big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t))\Big) \\ p(\mathbf{z}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \end{cases}$$

▷ Learn encoder-decoder parameters over *clean* speech data:

$$\mathcal{L}(\Phi; \mathbf{s}) = \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{s})}\{\log p_\theta(\mathbf{s}|\mathbf{z})\} - \mathcal{D}_{\mathrm{KL}}(q_\psi(\mathbf{z}|\mathbf{s}) \| p(\mathbf{z}))$$

Testing: speech enhancement

Non-negative matrix factorization (**NMF**)-based noise model:

$$p_\psi(\boldsymbol{b}) \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\mathrm{vec}(\mathbf{WH}))), \quad \psi = \{\mathbf{W}, \mathbf{H}\}$$

Parameter inference: Expectation-maximization (**EM**)

- **E-step:** compute posterior $p_\psi(\mathbf{z}|\mathbf{x})$ Intractable to compute! $\to$ Sample from it.
- **M-step:** update parameters:

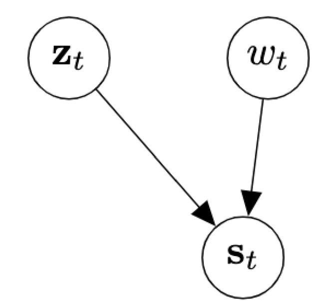
$$\max_\psi \ \mathbb{E}_{p_\psi(\mathbf{z}|\mathbf{x})}\{\log p_\psi(\mathbf{x}|\mathbf{z})\}$$

☞ *multiplicative update rules*

Proposed framework: StVAE model

- **Weighting the variance:** introduce scalar weights per variance components

$$\begin{cases} p_\theta(\mathbf{s}_t|\mathbf{z}_t, w_t) = \mathcal{N}_c\Big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t))/w_t\Big) \\ p(\mathbf{z}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ p(w_t) = \mathcal{G}(w_t; \alpha, \beta) \end{cases}$$



- **Gamma prior over the weights:**

$$\mathcal{G}(w; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{\alpha-1} \exp(-\beta w)$$

☞ An infinite *mixture* of Gaussian distributions: $p_\theta(\mathbf{s}_t|\mathbf{z}_t) = \int p_\theta(\mathbf{s}_t|\mathbf{z}_t, w_t)p(w_t)\mathrm{d}w_t$

Training phase of VAE:

- **Parameters to learn:** $\tilde{\Phi} = \{\theta, \psi, \alpha, \beta\}$
- **Posterior distribution:** $p_\theta(\mathbf{z}_t, w_t|\mathbf{s}_t) = p_\theta(w_t|\mathbf{s}_t, \mathbf{z}_t) \cdot p_\theta(\mathbf{z}_t|\mathbf{s}_t)$

The first term writes $p_\theta(w_t|\mathbf{s}_t, \mathbf{z}_t) \propto p_\theta(\mathbf{s}_t|\mathbf{z}_t, w_t) \cdot p(w_t) = \mathcal{G}(\alpha_t', \beta_t')$, where:

$$\begin{cases} \alpha_t' = \alpha + F \\ \beta_t' = \beta + \sum_f \frac{|s_{ft}|^2}{\sigma^2_{\theta,f}(\mathbf{z}_t)} \end{cases}$$

The intractable posterior $p_\theta(\mathbf{z}_t|\mathbf{s}_t)$ is approximated by a variational distribution: $p_\theta(\mathbf{z}_t|\mathbf{s}_t) \approx q_\psi(\mathbf{z}_t|\mathbf{s}_t)$.

$$p_\theta(\mathbf{z}, \boldsymbol{w}|\mathbf{s}) \approx q_\psi(\mathbf{z}, \boldsymbol{w}) = p_\theta(\boldsymbol{w}|\mathbf{s}, \mathbf{z})q_\psi(\mathbf{z}|\mathbf{s})$$

- **Training objective:**

$$\log p_\theta(\mathbf{s}) \geq \mathbb{E}_{q_\psi(\mathbf{z},\boldsymbol{w})}\left\{\log \frac{p_\theta(\mathbf{s}, \mathbf{z}, \boldsymbol{w})}{q_\psi(\mathbf{z}, \boldsymbol{w})}\right\} \triangleq \mathcal{L}(\tilde{\Phi}; \mathbf{s})$$

Speech enhancement phase:

E-step: Posterior $p_\psi(\mathbf{z}_t, w_t|\mathbf{x}_t)$ is intractable. Instead, we simply find the modes:

$$\mathbf{z}_t^*, w_t^* = \arg\max_{\mathbf{z}_t, w_t} \ \log p_\psi(\mathbf{z}_t, w_t|\mathbf{x}_t)$$

☞ **First-order optimization**

M-step: Update the NMF matrices using

$$\max_{\mathbf{W},\mathbf{H}} \ \sum_t \log p_\psi(\mathbf{x}_t|\mathbf{z}_t^*, w_t^*)$$

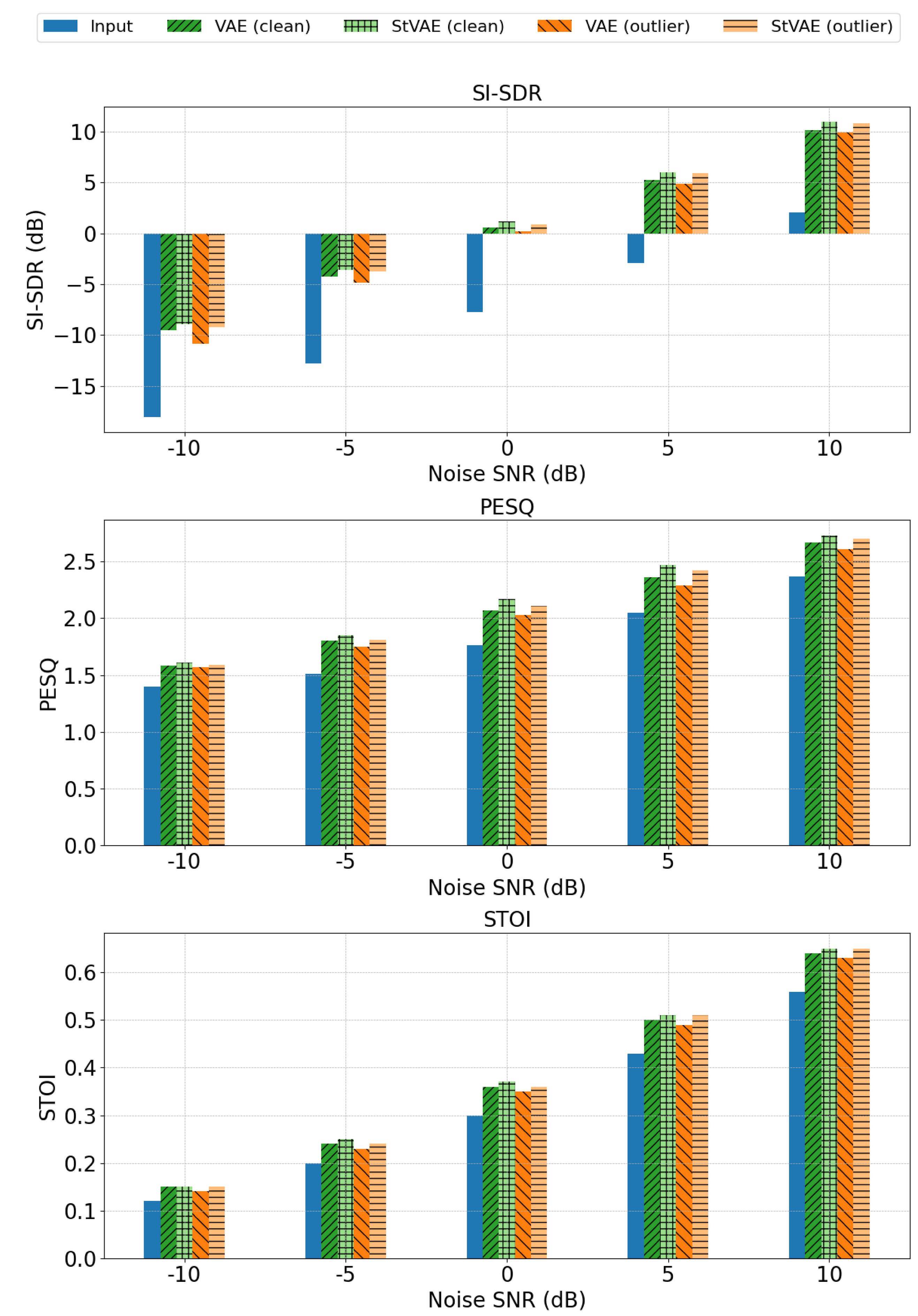
☞ **Multiplicative update rules**

Speech estimation: Posterior mean $\hat{\mathbf{s}}_t = \mathbb{E}_{p_{\psi^*}(\mathbf{s}_t|\mathbf{x}_t)}\{\mathbf{s}_t\}, \forall t$

$$\hat{\mathbf{s}}_t = \mathbb{E}_{p_{\psi^*}(\mathbf{z}_t^*, w_t^*|\mathbf{x}_t)}\{\mathbb{E}_{p_{\psi^*}(\mathbf{s}_t|\mathbf{x}_t, \mathbf{z}_t^*, w_t^*)}\{\mathbf{s}_t\}\} \approx \frac{(w_t^*)^{-1}\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t^*)}{(w_t^*)^{-1}\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t^*) + \mathbf{W}^*\boldsymbol{h}_t^*} \odot \mathbf{x}_t$$

Experiments

- **Datasets**: TCD-TIMIT (training & evaluation)
  - *Clean* setup: training on clean speech ($\sim$ 8 hrs)
  - *Outlier* setup: training on {clean, noise} data ($\sim$ 9.6 hrs)
- **Parameters**: STFT with 64 ms-long (1024 samples) sine window, 75% overlap ($F = 512$). $K = 100$ EM iterations, 10 iterations of posterior sampling. Latent dimension $L = 32$.
- **Baseline**: VAE [1] (single-layer, 128 nodes, encoder-decoder).



▷ StVAE surpasses VAE, especially at higher SNRs, highlighting its effective weighted variance Gaussian distribution.

▷ StVAE shows robust performance on noise-contaminated data.

▷ Training StVAE on noise-contaminated data outperforms the VAE trained on clean data, demonstrating StVAE's superior robustness.

Reference

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