



HAL
open science

A weighted-variance variational autoencoder model for speech enhancement

Ali Golmakani, Mostafa Sadeghi, Xavier Alameda-Pineda, Romain Serizel

► To cite this version:

Ali Golmakani, Mostafa Sadeghi, Xavier Alameda-Pineda, Romain Serizel. A weighted-variance variational autoencoder model for speech enhancement. ICASSP 2024 - International Conference on Acoustics Speech and Signal Processing, IEEE, Apr 2024, Seoul (Korea), South Korea. pp.1-5. hal-03833827v2

HAL Id: hal-03833827

<https://inria.hal.science/hal-03833827v2>

Submitted on 20 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

A WEIGHTED-VARIANCE VARIATIONAL AUTOENCODER MODEL FOR SPEECH ENHANCEMENT

Ali Golmakan¹, Mostafa Sadeghi¹, Xavier Alameda-Pineda², and Romain Serizel¹

¹Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

²Inria Grenoble & Univ. Grenoble Alpes, France

ABSTRACT

We address speech enhancement based on variational autoencoders, which involves learning a speech prior distribution in the time-frequency (TF) domain. A zero-mean complex-valued Gaussian distribution is usually assumed for the generative model, where the speech information is encoded in the variance as a function of a latent variable. In contrast to this commonly used approach, we propose a *weighted* variance generative model, where the contribution of each spectrogram time-frame in parameter learning is weighted. We impose a Gamma prior distribution on the weights, which would effectively lead to a Student's t-distribution instead of Gaussian for speech generative modeling. We develop efficient training and speech enhancement algorithms based on the proposed generative model. Our experimental results on spectrogram auto-encoding and speech enhancement demonstrate the effectiveness and robustness of the proposed approach compared to the standard unweighted variance model.

Index Terms: Speech enhancement, generative model, variational autoencoder, Student's t-distribution.

1. INTRODUCTION

Speech enhancement is a fundamental task in signal processing and machine learning, aiming to recover a clean speech signal from a noisy observation [1]. A classical approach to this problem involves statistical modeling of clean speech and noise signals, e.g., using non-negative matrix factorization (NMF), followed by an inference method such as maximum likelihood (ML) or Maximum a posteriori (MAP) estimation [2, 3]. However, with the advent of deep learning, there has been a significant shift towards supervised (discriminative) frameworks, which train a deep neural network (DNN) on a large collection of paired clean and noisy speech signals [4]. Nevertheless, these methods suffer from generalization issues, e.g., for unseen noise environments, as the train and test conditions might be significantly different.

Recently, there has been a growing interest in alternative approaches based on deep generative models, including variational autoencoders (VAEs) [5–9], generative adversarial networks (GANs) [10], and normalizing flows (NFs) [11], due to their *potential* generalization advantage. In particular, VAE-based speech enhancement involves learning a prior distribution of (time-frequency domain) clean speech data with a latent variable model. More precisely, the distribution of each speech time-frequency (TF) point is modeled as a circularly symmetric complex Gaussian distribution, where the

variance is a DNN-based parameterized function of a latent variable with a standard Gaussian prior. Given a noisy speech observation and the trained speech prior, a parametric statistical noise model is adaptively learned with an expectation-maximization (EM) approach followed by clean speech signal estimation. Therefore, noise characteristics are modeled at test time, giving them higher potential for performance generalization compared to supervised methods [5, 6].

In this paper, we propose to use a *weighted* variance circularly symmetric complex Gaussian distribution for VAE-based speech modeling, where the contribution of each spectrogram time-frame to parameter learning and inference is separately weighted. Assuming a Gamma prior for the weights and marginalizing them, the resulting model would become a Student's t-distribution. This brings more efficient, robust, and flexible modeling power than the standard unweighted Gaussian variance model. We develop computationally efficient training and speech enhancement methods based on the EM framework. Our experiments show that the proposed weighted variance VAE model outperforms the standard unweighted counterpart, both in terms of reconstruction quality and speech enhancement performance.

The rest of the paper is organized as follows. Section 2 reviews VAE-based speech generative modeling. The proposed speech generative and enhancement frameworks are detailed in Section 3. Section 4 discusses the related work. Experimental results are presented in Section 5. Finally, Section 6 concludes the paper.

2. VAE-BASED SPEECH MODELING

We denote the short-time Fourier transform (STFT) representation of clean speech signals as $\mathbf{s} = \{\mathbf{s}_1, \dots, \mathbf{s}_T\}$, consisting of complex-valued vectors $\mathbf{s}_t = [s_{ft}]_{f=1}^F \in \mathbb{C}^F$. The VAE framework associates a latent variable $\mathbf{z}_t \in \mathbb{R}^L$ to each time frame \mathbf{s}_t , where $L \ll F$. The joint distribution, that is $p(\mathbf{s}_t, \mathbf{z}_t) = p(\mathbf{s}_t|\mathbf{z}_t) \cdot p(\mathbf{z}_t)$, is modeled by some parametric Gaussian forms:

$$p_{\theta}(\mathbf{s}_t|\mathbf{z}_t) = \mathcal{N}_c\left(\mathbf{0}, \text{diag}(\sigma_{\theta}^2(\mathbf{z}_t))\right), \quad p(\mathbf{z}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1)$$

where $\mathcal{N}_c(\mathbf{0}, \Sigma)$ denotes a circularly symmetric complex Gaussian distribution, and \mathbf{I} is the identity matrix. Also, $\sigma_{\theta}(\cdot)$ (applied element-wise) is a non-linear function denoting the standard deviation, which is modeled by some DNN, called the *decoder*, with parameters θ . To learn θ , one needs to compute the posterior distribution $p_{\theta}(\mathbf{z}_t|\mathbf{s}_t)$, which is intractable. In the VAE framework, this term is approximated as follows:

$$q_{\psi}(\mathbf{z}_t|\mathbf{s}_t) = \mathcal{N}(\boldsymbol{\mu}_{\psi}(\mathbf{s}_t), \text{diag}(\sigma_{\psi}^2(\mathbf{s}_t))), \quad (2)$$

where $\boldsymbol{\mu}_{\psi}$ and σ_{ψ} are implemented using a DNN, called the *encoder*.

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER, and several Universities as well as other organizations (see <https://www.grid5000.fr>).

The set of parameters, $\Phi = \{\theta, \psi\}$, is learned by optimizing a lower-bound, denoted $\mathcal{L}(\Phi; \mathbf{s})$, on the intractable data log-likelihood $\log p_\theta(\mathbf{s})$. This is achieved by defining the evidence lower-bound (ELBO) as follows [12]:

$$\mathcal{L}(\Phi; \mathbf{s}) = \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{s})} \{\log p_\theta(\mathbf{s}|\mathbf{z})\} - \mathcal{D}_{\text{KL}}(q_\psi(\mathbf{z}|\mathbf{s})\|p(\mathbf{z})), \quad (3)$$

where $\mathcal{D}_{\text{KL}}(q\|p)$ stands for the Kullback–Leibler (KL) divergence between q and p . The first term in (3) measures the reconstruction quality of the model, and the second one is a regularization. Training proceeds by optimizing $\mathcal{L}(\Phi; \mathbf{s})$ over Φ using a gradient-based optimizer, along with the *reparametrization trick* [12].

3. PROPOSED FRAMEWORK

3.1. Generative model

As opposed to the commonly used unweighted variance model presented in (1), we propose a more flexible distribution, by introducing some weight parameters $w_t > 0$:

$$\begin{cases} p_\theta(\mathbf{s}_t|\mathbf{z}_t, w_t) = \mathcal{N}_c(\mathbf{0}, \text{diag}(\boldsymbol{\sigma}_\theta^2(\mathbf{z}_t))/w_t), \\ p(\mathbf{z}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ p(w_t) = \mathcal{G}(w_t; \alpha, \beta), \end{cases} \quad (4)$$

where, $\mathcal{G}(w; \alpha, \beta)$ is the Gamma distribution ($\alpha, \beta > 0$):

$$\mathcal{G}(w; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} w^{\alpha-1} \exp(-\beta w), \quad (5)$$

and $\Gamma(\cdot)$ denotes the gamma function. The mean and variance of this distribution are equal to α/β and α/β^2 , respectively. Note that $p_\theta(\mathbf{s}_t|\mathbf{z}_t)$ is essentially an infinite mixture of Gaussian distributions: $p_\theta(\mathbf{s}_t|\mathbf{z}_t) = \int p_\theta(\mathbf{s}_t|\mathbf{z}_t, w_t) p(w_t) dw_t$. This effectively takes the form of a Student’s t-distribution, which is well-known for its robustness and flexibility advantages over a standard Gaussian distribution [13].

3.2. Parameters inference

To learn the parameters of the proposed Student VAE (St-VAE) model, denoted $\tilde{\Phi} = \{\theta, \psi, \alpha, \beta\}$, we need to compute the posterior distribution of the latent variables \mathbf{z}_t, w_t :

$$p_\theta(\mathbf{z}_t, w_t|\mathbf{s}_t) = p_\theta(w_t|\mathbf{s}_t, \mathbf{z}_t) \cdot p_\theta(\mathbf{z}_t|\mathbf{s}_t). \quad (6)$$

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:

$$\begin{cases} \alpha'_t = \alpha + F \\ \beta'_t = \beta + \sum_f \frac{|s_{ft}|^2}{\sigma_{\theta,f}^2(\mathbf{z}_t)}. \end{cases} \quad (7)$$

The second posterior distribution, i.e., $p_\theta(\mathbf{z}_t|\mathbf{s}_t)$ cannot be computed in closed-form. We, therefore, resort to a variational approximation: $p_\theta(\mathbf{z}_t|\mathbf{s}_t) \approx q_\psi(\mathbf{z}_t|\mathbf{s}_t)$, with q_ψ defined similarly as in (2). Overall, we have:

$$p_\theta(\mathbf{z}, \mathbf{w}|\mathbf{s}) \approx q_\psi(\mathbf{z}, \mathbf{w}) = p_\theta(\mathbf{w}|\mathbf{s}, \mathbf{z}) q_\psi(\mathbf{z}|\mathbf{s}), \quad (8)$$

where, $\mathbf{w} = \{w_1, \dots, w_T\}$. We target a lower-bound on the data log-likelihood to learn $\tilde{\Phi}$:

$$\log p_\theta(\mathbf{s}) \geq \mathbb{E}_{q_\psi(\mathbf{z}, \mathbf{w})} \left\{ \log \frac{p_\theta(\mathbf{s}, \mathbf{z}, \mathbf{w})}{q_\psi(\mathbf{z}, \mathbf{w})} \right\} \triangleq \mathcal{L}(\tilde{\Phi}; \mathbf{s}), \quad (9)$$

which is simplified as

$$\begin{aligned} \mathcal{L}(\tilde{\Phi}; \mathbf{s}) &= \mathbb{E}_{q_\psi(\mathbf{z}, \mathbf{w})} \{\log p_\theta(\mathbf{s}|\mathbf{z}, \mathbf{w})\} - \\ &\mathcal{D}_{\text{KL}}(q_\psi(\mathbf{z}|\mathbf{s})\|p(\mathbf{z})) - \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{s})} \{\mathcal{D}_{\text{KL}}(p_\theta(\mathbf{w}|\mathbf{z}, \mathbf{s})\|p(\mathbf{w}))\}. \end{aligned} \quad (10)$$

The first and third terms can be further simplified. This will bring us to the following final form:¹

$$\begin{aligned} \mathcal{L}(\tilde{\Phi}; \mathbf{s}) &= \sum_{t=1}^T \mathbb{E}_{q_\psi(\mathbf{z}_t|\mathbf{s}_t)} \left\{ - \sum_{f=1}^F \log |\sigma_{\theta,f}^2(\mathbf{z}_t)| - \right. \\ &(\alpha + F) \log \left(\beta + \sum_{f=1}^F \frac{|s_{ft}|^2}{\sigma_{\theta,f}^2(\mathbf{z}_t)} \right) \left. \right\} + \sum_{\ell=0}^{F-1} \log(\alpha + \ell) + \\ &+ \alpha \log \beta - \mathcal{D}_{\text{KL}}(q_\psi(\mathbf{z}|\mathbf{s})\|p(\mathbf{z})). \end{aligned} \quad (11)$$

As in VAEs, we approximate the above expectation using a single sample $\mathbf{z}_t \sim q_\psi(\mathbf{z}_t|\mathbf{s}_t)$, followed by the reparametrization trick. The obtained objective function is then optimized over $\tilde{\Phi}$ using a stochastic gradient-based optimizer.

3.3. Speech Enhancement

The observed noisy speech STFT time frames are modeled as $\mathbf{x}_t = \mathbf{s}_t + \mathbf{b}_t$, $t = 1, \dots, \tilde{T}$, where \mathbf{b}_t corresponds to background noise. For \mathbf{s}_t , the pre-trained generative model in (4) is used. For \mathbf{b}_t , the following nonnegative matrix factorization (NMF) based model is considered:

$$\mathbf{b}_t \sim \mathcal{N}_c(\mathbf{0}, \text{diag}(\mathbf{W}\mathbf{h}_t)), \quad (12)$$

where, $\mathbf{W} \in \mathbb{R}_+^{F \times K}$, and \mathbf{h}_t is the t -th column of $\mathbf{H} \in \mathbb{R}_+^{K \times \tilde{T}}$.

3.3.1. Parameter estimation

To infer the model’s parameters, i.e., $\phi = \{\mathbf{W}, \mathbf{H}\}$, we follow an EM approach, where in the expectation (E) step, the intractable posterior distribution $p(\mathbf{z}_t, w_t|\mathbf{x}_t)$ needs to be computed. As an approximation, we find only the modes, i.e., the points that maximize this distribution [14, 15]:

$$\mathbf{z}_t^*, w_t^* = \underset{\mathbf{z}_t, w_t}{\text{argmax}} \log p_\phi(\mathbf{z}_t, w_t|\mathbf{x}_t), \quad (13)$$

or, equivalently,

$$\mathbf{z}_t^*, w_t^* = \underset{\mathbf{z}_t, w_t}{\text{argmax}} \log p_\phi(\mathbf{x}_t|\mathbf{z}_t, w_t) + \log p(\mathbf{z}_t) + \log p(w_t). \quad (14)$$

It is straightforward to show that:

$$p_\phi(\mathbf{x}_t|\mathbf{z}_t, w_t) = \mathcal{N}_c(\mathbf{0}, \text{diag}(w_t^{-1} \boldsymbol{\sigma}_\theta^2(\mathbf{z}_t) + \mathbf{W}\mathbf{h}_t)). \quad (15)$$

Problem (14) is then solved via a first-order optimizer, e.g., Adam [16]. In the maximization (M) step, the parameters are updated by solving the following problem:

$$\max_{\mathbf{W}, \mathbf{H}} \sum_t \mathbb{E}_{p_\phi(\mathbf{z}_t, w_t|\mathbf{x}_t)} \{\log p_\phi(\mathbf{x}_t, \mathbf{z}_t, w_t)\} \quad (16)$$

$$\equiv \max_{\mathbf{W}, \mathbf{H}} \sum_t \mathbb{E}_{p_\phi(\mathbf{z}_t, w_t|\mathbf{x}_t)} \{\log p_\phi(\mathbf{x}_t|\mathbf{z}_t, w_t)\}. \quad (17)$$

¹Due to the limited space, we provide the detailed derivations in Supplementary Material available online: <https://msaadeghii.github.io/files/stvae.pdf>.

We approximate the above expectation using \mathbf{z}_t^*, w_t^* as follows:

$$\max_{\mathbf{W}, \mathbf{H}} \sum_t \log p_\phi(\mathbf{x}_t | \mathbf{z}_t^*, w_t^*). \quad (18)$$

Substituting (15) and pursuing the approach proposed in [6], we obtain the following multiplicative update rules:

$$\mathbf{H} \leftarrow \mathbf{H} \odot \left(\frac{\mathbf{W}^\top (|\mathbf{X}|^{\odot 2} \odot \mathbf{V}^{\odot -2})}{\mathbf{W}^\top \mathbf{V}^{\odot -1}} \right)^{\odot 1/2}, \quad (19)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \left(\frac{(|\mathbf{X}|^{\odot 2} \odot \mathbf{V}^{\odot -2}) \mathbf{H}^\top}{\mathbf{V}^{\odot -1} \mathbf{H}^\top} \right)^{\odot 1/2}, \quad (20)$$

where \odot denotes element-wise operation, $\mathbf{V} \in \mathbb{R}_+^{F \times \tilde{T}}$ is a matrix with columns $\mathbf{v}_t = (w_t^*)^{-1} \sigma_\theta^2(\mathbf{z}_t^*) + \mathbf{W} \mathbf{h}_t$, and $\mathbf{X} \in \mathbb{C}^{F \times \tilde{T}}$ is a matrix with columns \mathbf{x}_t . The overall inference algorithm iterates between (14), (19), and (20).

3.3.2. Speech estimation

Having learned $\phi^* = \{\mathbf{W}^*, \mathbf{H}^*\}$, the speech signal is estimated as the posterior mean $\hat{\mathbf{s}}_t = \mathbb{E}_{p_{\phi^*}(\mathbf{s}_t | \mathbf{x}_t)} \{\mathbf{s}_t\}$, $\forall t$, which can be equivalently written as

$$\begin{aligned} \hat{\mathbf{s}}_t &= \mathbb{E}_{p_{\phi^*}(\mathbf{z}_t^*, w_t^* | \mathbf{x}_t)} \left\{ \mathbb{E}_{p_{\phi^*}(\mathbf{s}_t | \mathbf{x}_t, \mathbf{z}_t^*, w_t^*)} \{\mathbf{s}_t\} \right\} \\ &\approx \frac{(w_t^*)^{-1} \sigma_\theta^2(\mathbf{z}_t^*)}{(w_t^*)^{-1} \sigma_\theta^2(\mathbf{z}_t^*) + \mathbf{W}^* \mathbf{h}_t^*} \odot \mathbf{x}_t, \end{aligned} \quad (21)$$

with element-wise division.

4. RELATED WORK

The closest work to ours is [17], which presents a VAE for robust density estimation applications with a Gamma prior distribution on the variance of the Gaussian decoder. The parameters of this distribution are then modeled as functions of the latent codes, i.e., $\alpha(\mathbf{z})$ and $\beta(\mathbf{z})$, implemented by some DNNs. However, our approach is different, as we consider a variance model, $\sigma_\theta^2(\cdot)$, that is shared among all the data, and we instead consider separate scalar weights for each data point. We also do not model the Gamma parameters as functions of \mathbf{z} , because, it would highly complicate the optimization of \mathbf{z} in the speech enhancement phase, i.e., (14). Furthermore, in contrast to [17], we do not marginalize the weights and instead follow a variational approach, which is much more efficient.

5. EXPERIMENTS

5.1. Setup

In this section, we compare the performance of our proposed StVAE framework (4) against the standard VAE method based on (1) for both speech spectrogram auto-encoding and speech enhancement. The former consists of auto-encoding the *clean* speech spectrogram using the trained VAE model to measure how well the input spectrogram is reconstructed, as also considered in [8]. The reconstruction quality is measured based on the signal-to-noise ratio (SNR) in dB. Moreover, we plug the original phase into the reconstructed spectrogram and obtain the time-domain speech signal using inverse STFT. This is to evaluate the intelligibility and perceptual quality of

the reconstructed speech signal in terms of the short-term objective intelligibility (STOI) measure [18], ranging in $[0, 1]$, and the perceptual evaluation of speech quality (PESQ) score [19], ranging in $[-0.5, 4.5]$, respectively.

For speech enhancement performance evaluation, in addition to PESQ and STOI, we report the scale-invariant signal-to-distortion ratio (SI-SDR) [20] values. To have a fair comparison, the standard VAE-based speech enhancement (VAE-SE) considered as the baseline follows the same steps as those of StVAE-SE detailed in Section (3.3). It should be mentioned that we did not include the VAE model proposed in [17] as a baseline, because we could not get satisfactory results for the reasons mentioned in Section 4.

5.2. Datasets

For training the StVAE and VAE models, we used the speech data in the TCD-TIMIT corpus [21]. This dataset contains speech utterances from 56 English speakers (39 for training, 8 for validation, and 9 for testing) with an Irish accent, uttering 98 different sentences, each with an approximate length of 5 seconds, and sampled at 16 kHz (~ 8 hours of data). The STFT of the speech data was computed with a 64 ms-long (1024 samples) sine window, 75% overlap, without zero-padding, which results in $F = 513$.

To test the speech enhancement performance, we used some noisy versions of the TCD-TIMIT dataset [22], including six types of noise, namely *Living Room (LR)*, *White*, *Cafe*, *Car*, *Babble*, and *Street*. For each noise type, we considered five noise levels: -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB. From each test speaker, we randomly selected 5 utterances for each noise level and noise type, giving 1350 test samples.

Furthermore, to see how the two competing algorithms behave when the training data include some noise signals in addition to the clean speech data, we extended the training data by taking some noise data from the DEMAND dataset [23], including *STRAFFIC*, *DWASHING*, *SPSQUARE*, *NRIVER*, *TBUS*, *NPARK*, and *DKITCHEN*. The total amount of noise data is around 20% of the clean speech training data. As the speech generative model is supposed to be learned on only *clean* data, including noise signals in the training set aims to measure the robustness of the learned models. We emphasize that, here, noise signals are *not* added with clean speech signals to form mixtures. Instead, they are just intended to serve as some outlier training data.

5.3. Model architecture

The architectures of both StVAE and VAE follow the one proposed in [6], consisting of an encoder and decoder each having a single fully-connected hidden layer with 128 nodes and hyperbolic tangent activation functions. The dimension of the latent space was set as $L = 32$.

5.4. Parameters setting

Both VAE models are trained with stochastic gradient descent (batch size of 128) using Adam. The learning rate is equal to 0.0001. We used early stopping on the validation set with a patience of 20 epochs. The number of EM iterations for speech enhancement is set to 100. The learning rate for optimizing (14) is set to 0.005, with 10 iterations.

Although α and β in StVAE could be learned according to (11), we observed in our experiments that fixing these values during the

Table 1: Average values of the input and output SI-SDR, PESQ, and STOI metrics for speech enhancement. The results are presented separately for the VAE models trained on clean data and outlier-contaminated data.

| Metric | SI-SDR (dB) | | | | | PESQ | | | | | STOI | | | | |
|--|--------------|--------------|-------------|-------------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | -10 | -5 | 0 | 5 | 10 | -10 | -5 | 0 | 5 | 10 | -10 | -5 | 0 | 5 | 10 |
| Input (unprocessed) | -18.08 | -12.80 | -7.72 | -2.91 | 2.04 | 1.40 | 1.51 | 1.76 | 2.05 | 2.37 | 0.12 | 0.20 | 0.30 | 0.43 | 0.56 |
| Models trained on <i>clean data</i> | | | | | | | | | | | | | | | |
| VAE-SE | -9.56 | -4.25 | 0.57 | 5.23 | 10.13 | 1.58 | 1.80 | 2.07 | 2.36 | 2.67 | 0.15 | 0.24 | 0.36 | 0.50 | 0.64 |
| StVAE-SE | -8.92 | -3.56 | 1.16 | 5.97 | 10.97 | 1.61 | 1.85 | 2.17 | 2.47 | 2.73 | 0.15 | 0.25 | 0.37 | 0.51 | 0.65 |
| Models trained on <i>outlier-contaminated data</i> | | | | | | | | | | | | | | | |
| VAE-SE | -10.83 | -4.84 | 0.22 | 4.87 | 9.89 | 1.57 | 1.75 | 2.03 | 2.29 | 2.61 | 0.14 | 0.23 | 0.35 | 0.49 | 0.63 |
| StVAE-SE | -9.23 | -3.74 | 0.87 | 5.89 | 10.83 | 1.59 | 1.81 | 2.11 | 2.42 | 2.70 | 0.15 | 0.24 | 0.36 | 0.51 | 0.65 |

Table 2: Average values of the SNR, PESQ, and STOI metrics for speech spectrogram auto-encoding.

| Metric | SNR (dB) | PESQ | STOI |
|--|-------------|-------------|-------------|
| Models trained on <i>clean data</i> | | | |
| VAE | 6.94 | 3.29 | 0.85 |
| StVAE | 7.98 | 3.51 | 0.88 |
| Models trained on <i>outlier-contaminated data</i> | | | |
| VAE | 5.93 | 3.12 | 0.83 |
| StVAE | 7.15 | 3.32 | 0.86 |

whole training process leads to more stable and improved results. As such, we empirically set $\alpha = \beta = 100$, resulting in the mean and variance for the prior distribution of the weights equal to 1 and 0.01, respectively.

5.5. Speech enhancement results

The input (evaluated on unprocessed, noisy speech signals) and output (evaluated on the estimated speech signals) values of the speech enhancement metrics for different noise SNRs are reported in Table 1. Concerning the generative models learned on clean training data (without outlier data, i.e., noise signals), one can see that StVAE-SE outperforms VAE-SE in almost all the cases, demonstrating the efficiency of the proposed weighted variance Gaussian distribution compared to the standard, unweighted distribution. This is more noticeable for higher SNR levels. With respect to the outlier-contaminated data involving noise signals, we can also clearly see the advantage of StVAE. In addition, we note that StVAE-SE trained on outlier-contaminated data outperforms VAE-SE trained on clean data.

5.6. Spectrogram auto-encoding results

Table 2 summarizes the speech spectrogram auto-encoding results as a measure of reconstruction quality of the models. It can be seen that the proposed StVAE model performs considerably better than the standard VAE model, in terms of both reconstruction SNR and speech quality measures, PESQ and STOI. As also observed in the previous section, the results of StVAE when trained on outlier-contaminated data are better even than those of the VAE model trained on clean data. This confirms that the weighted vari-

ance model (Student’s t-distribution) is a better fit for speech generative modeling.

It is important to mention that the performance improvements discussed above are achieved with very little additional computational overhead when compared to standard VAE-based speech generative modeling and enhancement. This comparison includes the generative models presented in equations (1) and (4), as well as the new speech enhancement framework presented in equation (14), which involves optimizing over the additional scalar-valued weight parameters.

6. CONCLUSIONS

We presented a weighted variance Gaussian generative model for speech signals based on variational autoencoders. The proposed probabilistic generative model assumes a separate stochastic weight for each spectrogram time-frame with a Gamma prior distribution, providing a more flexible and effective modeling framework compared to the standard, unweighted Gaussian variance model. We also presented efficient parameter inference and speech enhancement methodologies. Our experimental results showed the superiority of the proposed model, both in terms of spectrogram auto-encoding reconstruction quality and speech enhancement results.

As future works, we plan to extend the proposed weighted variance generative model and speech enhancement frameworks to the dynamical VAE models [8], and to consider a Markovian dependency for the weights, which are not straightforward. This will allow for more efficient incorporation of the time-dynamics of the spectrogram time-frames, and consequently an improved performance. Furthermore, this will enable us to fairly compare the performance of the developed dynamical, weighted variance speech enhancement system with the supervised (discriminative), DNN-based approaches.

7. REFERENCES

- [1] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.
- [2] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis,” *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.
- [3] P. Smaragdis, B. Raj, and M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel

- mixtures,” in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 414–421.
- [4] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.
- [5] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, “Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization,” in *ICASSP*, 2018.
- [6] S. Leglaive, L. Girin, and R. Horaud, “A variance modeling framework based on variational autoencoders for speech enhancement,” in *IEEE MLSP*, September 2018.
- [7] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “Audio-visual speech enhancement using conditional variational auto-encoders,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.
- [8] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, “Unsupervised speech enhancement using dynamical variational autoencoders,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.
- [9] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, “Variational autoencoder for speech enhancement with a noise-aware encoder,” in *IEEE ICASSP*, 2021.
- [10] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” *arXiv preprint arXiv:1703.09452*, 2017.
- [11] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, “A flow-based deep latent variable model for speech spectrogram modeling and enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1104–1117, 2020.
- [12] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *Proc. International Conference on Learning Representations (ICLR)*, April 2014.
- [13] K. L. Lange, R. J. Little, and J. M. Taylor, “Robust statistical modeling using the t distribution,” *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.
- [14] H. Kameoka, L. Li, S. Inoue, and S. Makino, “Supervised determined source separation with multichannel variational autoencoder,” *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.
- [15] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “A recurrent variational autoencoder for speech enhancement,” in *IEEE ICASSP*, 2020.
- [16] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. International Conference on Learning Representations ICLR*, May 2015.
- [17] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, and S. Yagi, “Student-t variational autoencoder for robust density estimation,” in *IJCAI*, 2018, pp. 2696–2702.
- [18] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, February 2011.
- [19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *IEEE ICASSP*, May 2001.
- [20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR—half-baked or well done?” in *IEEE ICASSP*, 2019.
- [21] N. Harte and E. Gillen, “TCD-TIMIT: An audio-visual corpus of continuous speech,” *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.
- [22] A. H. Abdelaziz *et al.*, “NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition,” in *Interspeech*, 2017.
- [23] J. Thiemann, N. Ito, and E. Vincent, “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments,” in *Proc. Meetings Acoust.*, 2013, pp. 1–6.