Speech enhancement with variational autoencoders and alpha-stable distributions
Simon Leglaive, Umut Şimşekli, Antoine Liutkus, Laurent Girin, Radu Horaud

To cite this version:
Simon Leglaive, Umut Şimşekli, Antoine Liutkus, Laurent Girin, Radu Horaud. Speech enhancement with variational autoencoders and alpha-stable distributions. ICASSP 2019 - 44th IEEE International Conference on Acoustics, Speech and Signal Processing, May 2019, Brighton, United Kingdom. pp.541-545, 10.1109/ICASSP.2019.8682546. hal-02005106

HAL Id: hal-02005106
https://inria.hal.science/hal-02005106
Submitted on 8 Feb 2019

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
SPEECH ENHANCEMENT WITH VARIATIONAL AUTOENCODERS AND ALPHA-STABLE DISTRIBUTIONS

Simon Leglaive\textsuperscript{1}, Umut Şimşekli\textsuperscript{2}, Antoine Liutkus\textsuperscript{3}, Laurent Girin\textsuperscript{1,4}, Radu Horaud\textsuperscript{1}

\textsuperscript{1}Inria Grenoble Rhône-Alpes, France, \textsuperscript{2}LTCI, Télécom ParisTech, Université Paris-Saclay, France
\textsuperscript{3}Inria and LIRMM, France, \textsuperscript{4}Univ. Grenoble Alpes, Grenoble INP, GIPSA-lab, France

\section*{ABSTRACT}

This paper focuses on single-channel semi-supervised speech enhancement. We learn a speaker-independent deep generative speech model using the framework of variational autoencoders. The noise model remains unsupervised because we do not assume prior knowledge of the noisy recording environment. In this context, our contribution is to propose a noise model based on alpha-stable distributions, instead of the more conventional Gaussian non-negative matrix factorization approach found in previous studies. We develop a Monte Carlo expectation-maximization algorithm for estimating the model parameters at test time. Experimental results show the superiority of the proposed approach both in terms of perceptual quality and intelligibility of the enhanced speech signal.

\textbf{Index Terms}— Speech enhancement, variational autoencoders, alpha-stable distribution, Monte Carlo expectation-maximization.

\section{1. INTRODUCTION}

Speech enhancement is one of the central problems in audio signal processing [1]. The goal is to recover a clean speech signal after observing a noisy mixture. In this work, we address single-channel speech enhancement, which can be seen as an under-determined source separation problem, where the sources are of different nature.

One popular statistical approach for source separation combines a local Gaussian model of the time-frequency signal coefficients with a variance model [2]. In this framework, non-negative matrix factorization (NMF) techniques have been used to model the time-frequency-dependent signal variance [3, 4]. Recently, discriminative approaches based on deep neural networks (DNNs) have also been investigated for speech enhancement, with the aim of estimating either clean spectrograms or time-frequency masks, given noisy spectrograms [5, 6, 7]. As a representative example, a DNN is used in [6] to map noisy speech log-power spectrograms into clean speech log-power spectrograms.

Even more recently, generative models based on deep learning, and in particular variational autoencoders (VAEs) [8], have been used for single-channel [9, 10] and multi-channel speech enhancement [11, 12]. These generative model-based approaches provide important advantages and justify the interest of semi-supervised methods for speech enhancement. Indeed, as shown in [9, 10], fully-supervised methods such as [6] may have issues for generalizing to unseen noise types. The method proposed in [10] was shown to outperform both a semi-supervised NMF baseline and the fully-supervised deep learning approach [6].

In most cases, probabilistic models for source separation or speech enhancement rely on a Gaussianity assumption, which turns out to be restrictive for audio signals [13]. As a result, heavy-tailed distributions have started receiving increasing attention in the audio processing community [14, 15, 16]. In particular, $\alpha$-stable distributions (cf. Section 3) are becoming popular heavy-tailed models for audio modeling due to their nice theoretical properties [17, 13, 14, 18, 19, 20, 21].

In this work, we investigate the combination of a deep learning-based generative speech model with a heavy-tailed $\alpha$-stable noise model. The rationale for introducing a noise model based on heavy-tailed distributions as opposed to a structured NMF approach as in [10] is to avoid relying on restricting assumptions regarding stationarity or temporal redundancy of the noisy environment, that may be violated in practice, leading to errors in the estimates. In addition, we let the noise model remain unsupervised in order to avoid the aforementioned generalization issues regarding the noisy recording environment. We develop a Monte Carlo expectation-maximization algorithm [22] for performing maximum likelihood estimation at test time. Experiments performed under challenging conditions show that the proposed approach outperforms the competing approaches in terms of both perceptual quality and intelligibility.

\section{2. SPEECH MODEL}

We work in the short-term Fourier transform (STFT) domain where $\mathbb{B} = \{0, \ldots, F-1\} \times \{0, \ldots, N-1\}$ denotes the set of time-frequency bins. For $(f, n) \in \mathbb{B}$, $f$ denotes the frequency index and $n$ the time-frame index. We use $s_{f,n}, b_{f,n}, x_{f,n} \in \mathbb{C}$ to denote the complex STFT coefficients of the speech, noise, and mixture signals, respectively.

As in [9, 10], independently for all $(f, n) \in \mathbb{B}$, we consider the following generative speech model involving a latent random vector $h_n \in \mathbb{R}^L$, with $L \ll F$:

\begin{align}
\begin{array}{c}
h_n \sim \mathcal{N}(0, I); \\
s_{f,n} \mid h_n \sim \mathcal{N}_c(0, \sigma_{s,f}^2(h_n)).
\end{array}
\end{align}

where $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ denotes the multivariate Gaussian distribution for a real-valued random vector, $I$ is the identity matrix of appropriate size, and $\mathcal{N}_c(x; \mu, \sigma^2)$ denotes the univariate complex proper Gaussian distribution. As represented in Fig. 1a, $\{\sigma_{s,f}^2 : \mathbb{R}^L \mapsto \mathbb{R}_+\}_{f=0}^{F-1}$ is a set of non-linear functions corresponding to a neural network which takes as input $h_n \in \mathbb{R}^L$. This variance term can be understood as a model for the short-term power spectral density of speech [23]. We denote by $\Theta$, the parameters of this generative neural network.

An important contribution of VAEs [8] is to provide an efficient way of learning the parameters of such a generative model. Let $s = \{s_{n} \in \mathbb{C}^F\}_{n=0}^{N-1}$ be a training dataset of clean-speech STFT time frames and $h = \{h_n \in \mathbb{R}^L\}_{n=0}^{N-1}$ the set of associated latent random
vectors. Taking ideas from variational inference, VAEs estimate the parameters $\theta$, by maximizing a lower bound of the log-likelihood $\ln p_{\theta}(s \mid h_{\theta})$, defined by:

$$L(\theta_s; \psi) = E_{q(h \mid s; \psi)}[\ln p(h \mid s; \psi) \mid p(h)] - \text{D}_{KL}(q(h \mid s; \psi) \mid p(h)).$$

where $q(h \mid s; \psi)$ is a variational approximation of the posterior $p(h \mid s, \theta)$, and $\text{D}_{KL}$ is the Kullback-Leibler divergence. Independently for all dimensions $l \in \{0, \ldots, L-1\}$ and all the time frames $n \in \{0, \ldots, N-1\}$, $q(h \mid s; \psi)$ is defined by:

$$h_{l,n} \mid s_{n} \sim \mathcal{N}(\tilde{\mu}_l(s_{n}^{\odot 2}), \tilde{\sigma}^2_l(s_{n}^{\odot 2})),

$$

where $h_{l,n} = (h_n)_l$ and $(\cdot)\odot$ denotes element-wise exponentiation. As represented in Figure 1b, $\tilde{\mu}_l: \mathbb{R}^{F \times L_0} \mapsto \mathbb{R}^{F \times L_1}$ and $\tilde{\sigma}^2_l: \mathbb{R}^{F \times L_0} \mapsto \mathbb{R}^{F \times L_1}$ are non-linear functions corresponding to a neural network which takes as input the speech power spectrum at a given time frame. $\psi$ denotes the parameters of this recognition network, which also have to be estimated by maximizing the variational lower bound defined in (3). Using (1), (2) and (4) we can develop this objective function as follows:

$$L(\theta_s; \psi) \equiv \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} E_{q(h_n \mid s_n; \psi)}[d_{\text{KL}}(|s_{fn}|^2; \tilde{\sigma}^2_l(h_n))] - \frac{1}{2} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \left[ \ln \tilde{\sigma}^2_l(s_{n}^{\odot 2}) - \tilde{\mu}_l(s_{n}^{\odot 2})^2 - \tilde{\sigma}^2_l(s_{n}^{\odot 2}) \right],

$$

where $d_{\text{KL}}(x; y) = x \ln(x/y) - y$ and $y$ is the Itakura-Saito divergence. Finally, using the so-called “reparametrization trick” [8] to approximate the intractable expectation in (5), we obtain an objective function which is differentiable with respect to both $\theta$ and $\psi$ and can be optimized using gradient-based algorithms. It is important to note that the only reason why the recognition network is introduced is to learn the parameters of the generative network.

### 3. NOISE AND MIXTURE MODELS

In the previous section we have seen how to learn the parameters of the generative model (1)-(2). This model can then be used as a speech signal probabilistic prior for a variety of applications. In this paper we are interested in single-channel speech enhancement.

We do not assume prior knowledge about the recording environment, so that the noise model remains unsupervised. Independently for all $(f,n) \in \mathbb{R}$, the STFT coefficients of noise are modeled as complex circularly symmetric $\alpha$-stable random variables [24]:

$$b_{f,n} \sim S_{\alpha}(\sigma_{b,f}),$$

where $\alpha \in [0, 2]$ is the characteristic exponent and $\sigma_{b,f} \in \mathbb{R}_+$ is the scale parameter. As proposed in [20] for multichannel speech enhancement, this scale parameter is only frequency-dependent, it does not depend on time. For algorithmic purposes, the noise model (6) can be conveniently rewritten in an equivalent scale mixture of Gaussians form [25], by making use of the product property of the symmetric $\alpha$-stable distribution [24]:

$$\phi_{f,n} \sim \frac{\alpha}{2}S(2\cos(\pi \alpha/4)^{1/\alpha});

$$

$$b_{f,n} \mid \phi_{f,n} \sim \mathcal{N}(0, \phi_{f,n} \sigma_{b,f}^2),$$

where $\phi_{f,n} \in \mathbb{R}_+$ is called the impulsive variable. It locally modulates the variance of the conditional distribution of $b_{f,n}$ given in (8). $\frac{\alpha}{2}S$ denotes a positive stable distribution of characteristic exponent $\alpha/2$. It corresponds to a right-skewed heavy-tailed distribution defined for non-negative random variables [26]. These impulse variables can be understood as carrying uncertainty about the stationary noise assumption made in the marginal model (6), where the scale parameter does not depend on the time-frame index.

The observed mixture signal is modeled as follows for all $(f,n) \in \mathbb{B}$:

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

where $g_n \in \mathbb{R}_+$ represents a frame-dependent but frequency-independent gain. The importance of this parameter was experimentally shown in [10]. We further consider the conditional independence of the speech and noise STFT coefficients so that:

$$x_{f,n} \mid h_n, \phi_{f,n} \sim \mathcal{N}(0, g_n \sigma_{s,f}^2(h_n) + \phi_{f,n} \sigma_{b,f}^2).$$

### 4. INFERENCE

Let $\theta_\ast \equiv \{g = \{g_n \in \mathbb{R}_+\}_{n=0}^{N-1}, \sigma_{b,f}^2 = \{\sigma_{b,f}^2 \in \mathbb{R}_+\}_{f=0}^{F-1}\}$ be the set of model parameters to be estimated. For maximum likelihood estimation, in this section we develop a Monte-Carlo expectation maximization (MCEM) algorithm [22], which iteratively applies the so-called E- and M-steps until convergence, which we detail below. Remember that the speech generative model parameters $\theta_s$ have been learned during a training phase (see Section 2). We denote by $x = (x_{f,n})_{f=0}^{F-1}$ the set of observed data while $z = (h_n, \phi_{f,n})_{n=0}^{N-1}$ is the set of latent variables. We will also use $z_n = (x_{f,n})_{f=0}^{F-1}$ and $z_n = (h_n, \phi_{f,n})$ to respectively denote the set of observed and latent variables at a given time frame $n$.

**Monte Carlo E-Step.** Let $\theta_\ast^\star$ be the current (or the initial) value of the model parameters. At the E-step of a standard expectation-maximization algorithm, we would compute the following conditional expectation of the complete-data log-likelihood $Q(\theta_\ast; \theta_\ast^\star) = E_{q(x,z; \theta_\ast^\star)}[\ln p(x, z; \theta_s, \theta_\ast)]$. However, this expectation cannot be here computed analytically. We therefore approximate $Q(\theta_\ast; \theta_\ast^\star)$ using an empirical average:

$$\tilde{Q}(\theta_\ast; \theta_\ast^\star) \equiv \frac{1}{R} \sum_{r=1}^{R} \sum_{(f,n)\in\mathbb{B}} \left[ \ln g_n \sigma_{s,f}^2(h_n^{(r)}) + \phi_{f,n} \sigma_{b,f}^2 \right] + |x_{f,n}|^2 \left[ g_n \sigma_{s,f}^2(h_n^{(r)}) + \phi_{f,n} \sigma_{b,f}^2 \right]^{-1},$$

where $\tilde{Q}$ denotes equality up to an additive constant, and $x_n^{(r)} = (h_n^{(r)}), \phi_{f,n}^{(r)} = (\phi_{f,n}^{(r)})_{f=0}^{F-1}$, $r \in \{1, \ldots, R\}$, is a sample drawn from the posterior $p(s_n \mid x_n, \theta_s, \theta_\ast^\star)$ using a Markov Chain Monte Carlo (MCMC) method. This approach forms the basis of the MCEM algorithm [22]. Note that unlike the standard EM algorithm, it does...
not ensure an improvement in the likelihood at each iteration. Nevertheless, some convergence results in terms of stationary points of the likelihood can be obtained under suitable conditions [27].

In this work we use a (block) Gibbs sampling algorithm [28]. From an initialization $x_0^{(0)}$, it consists in iteratively sampling from the so-called full conditionals. More precisely, at the $m$-th iteration of the algorithm and independently for all $n \in \{0, \ldots, N-1\}$, we first sample $h_n^{(m)} \sim \text{p}(h_n | x_n, \phi_n^{(m-1)}; \theta, \theta^*_n)$. Then, we sample $\phi_n^{(m)} \sim \text{p}(\phi_n | x_n, h_n^{(m)}; \theta, \theta^*_n)$ independently for all $n \in \{0, \ldots, N-1\}$. Those two full conditionals are unfortunately analytically intractable, but we can use one iteration of the Metropolis-Hastings algorithm to sample from them. This approach corresponds to the Metropolis-within-Gibbs sampling algorithm [28, p. 393]. One iteration of this method is detailed in Algorithm 1. The proposal distributions for $h_n$ and $\phi_n$ are respectively given in lines 2 and 6. The two acceptance probabilities required in lines 3 and 7 are computed as follows:

$$\alpha_n^{(h)} = \min \left( 1, \frac{p(h_n | x_n, \phi_n^{(m)}; \theta, \theta^*_n)}{p(h_n^{(m-1)} | x_n, \phi_n^{(m-1)}; \theta, \theta^*_n)} \right);$$

$$\alpha_n^{(\phi)} = \min \left( 1, \frac{p(x_n | h_n^{(m)}; \phi_n; \theta, \theta^*_n)}{p(x_n | h_n^{(m-1)}; \phi_n^{(m-1)}; \theta, \theta^*_n)} \right).$$

The two distributions involved in the computation of those acceptance probabilities are defined in (1) and (10). Finally, we only keep the last $R$ samples for computing $Q(\theta_n; \theta^*_n)$, i.e. we discard the samples drawn during a so-called burn-in period.

**M-step.** At the M-step we want to minimize $-\bar{Q}(\theta_n; \theta^*_n)$ with respect to $\theta_n$. Let $C(\theta_n)$ denote the cost function associated with this non-convex optimization problem. Similar to [29], we adopt a majorization-minimization approach. Let us introduce the two following sets of auxiliary variables $c = \{c_{f,n}^{(r)} \in \mathbb{R}_+\}_{r,f,n}$ and $\lambda = \{\lambda_{s,f,n}^{(r)} \in \mathbb{R}_+\}_{s,r,f,n}$. We can show using standard concave/convex inequalities (see e.g. [29]) that $C(\theta_n) \leq \bar{G}(\theta_n, c, \lambda)$, where:

$$\bar{G}(\theta_n, c, \lambda) = \frac{1}{R} \sum_{r=1}^{R} \sum_{(f,n)} \ln c_{f,n}^{(r)} + \frac{1}{c_{f,n}^{(r)}} \left( g_{s,n} \sigma_{s,n}^2(h_n^{(r)}) + \phi_{s,n}^2 \sigma_{b,f}^2 - c_{f,n}^{(r)} \right) + |x_{f,n}|^2 \left( \frac{\lambda_{1,f,n}}{g_{s,n} \sigma_{s,f}^2(h_n^{(r)})} + \frac{\lambda_{2,f,n}}{\phi_{s,n} \sigma_{b,f}^2} \right).$$

Moreover, this upper bound is tight, i.e. $C(\theta_n) = \bar{G}(\theta_n, c, \lambda)$, for:

$$c_{f,n}^{(r)} = g_{s,n} \sigma_{s,f}^2(h_n^{(r)}) + \phi_{s,n}^2 \sigma_{b,f}^2;$$

$$\lambda_{1,f,n} = g_{s,n} \sigma_{s,f}^2 \left( h_n^{(r)} \right) \left( g_{s,n}^2 \sigma_{s,f}^2 + \phi_{s,n}^2 \sigma_{b,f}^2 \right);$$

$$\lambda_{2,f,n} = \phi_{s,n}^2 \sigma_{b,f}^2 \left( g_{s,n} \sigma_{s,f}^2 \left( h_n^{(r)} \right) + \phi_{s,n} \sigma_{b,f}^2 \right).$$

Minimizing $ar{G}(\theta_n, c, \lambda)$ with respect to the model parameters $\theta_n$ is a convex optimization problem. By zeroing the partial derivatives of $\bar{G}$ with respect to each scalar in $\theta_n$ we obtain update rules that depend on the auxiliary variables. We then replace these auxiliary variables with the formulas given in (15)-(17), which makes the upper bound $\bar{G}$ tight. This procedure ensures that the cost $C(\theta_n)$ decreases [30]. The resulting final updates are given as follows:

$$\sigma_{b,f}^2 \leftarrow \sigma_{b,f}^2 + \frac{\sum_{r=1}^{R} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} |x_{f,n}|^2 \frac{\lambda_{2,f,n}}{g_{s,n}^2 \sigma_{s,f}^2 \left( h_n^{(r)} \right) + \phi_{s,n}^2 \sigma_{b,f}^2} }{\sum_{r=1}^{R} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \lambda_{2,f,n}};$$

$$\theta_n \leftarrow \theta_n + \frac{\sum_{r=1}^{R} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \frac{|x_{f,n}|^2 \lambda_{2,f,n}}{g_{s,n}^2 \sigma_{s,f}^2 \left( h_n^{(r)} \right) + \phi_{s,n}^2 \sigma_{b,f}^2} \phi_{f,n}^{(r)}}{\sum_{r=1}^{R} \sum_{f=0}^{F-1} \sum_{n=0}^{N-1} \lambda_{2,f,n}}.$$
The parameter setting: The STFT is computed using a 64-ms sine window (i.e. $F = 513$) with 75%-overlap. Based on [10], the latent dimension in the speech generative model (1)-(2) is fixed to $L = 64$. In this reference method, the NMF rank of the noise model is fixed to $K = 10$. The NMF parameters are randomly initialized. For both the proposed and the reference method, the gain $g_n$ is initialized to one for all time frames $n$. For the proposed method, the noise scale parameter $\sigma_{o,f}$ is also initialized to one for all frequency bins. We run 200 iterations of the MCEM algorithm. At each Monte-Carlo E-Step, we run 40 iterations of the Metropolis-within-Gibbs algorithm and we discard the first 30 samples as the burn-in period. The parameter $c^2$ in line 2 of Algorithm 1 is set to 0.01.

Neural network: The structure of the generative and recognition networks is the same as in [10] and is represented in Fig. 1. Hidden layers use hyperbolic tangent ($\tanh(\cdot)$) activation functions and output layers use identity activation functions ($I_d(\cdot)$). The output of these last layers is therefore real-valued, which is the reason why we consider logarithm of variances. For learning the parameters $\Theta$ and $\Phi$, we use the Adam optimizer [33] with a step size of $10^{-3}$, exponential decay rates for the first and second moment estimates of 0.9 and 0.999 respectively, and an epsilon of $10^{-7}$ for preventing division by zero. 20% of the TIMIT training set is kept as a validation set, and early stopping with a patience of 10 epochs is used. Weights are initialized using the uniform initializer described in [34].

In this work, we proposed a speech enhancement method exploiting a speech model based on VAEs and a noise model based on alpha-stable distributions. At the expense of more interferences, the proposed $\alpha$-stable noise model reduces the amount of artifacts in the estimated speech signal, compared to the use of a Gaussian NMF-based noise model as in [10]. Overall, it is shown that the proposed approach improves the intelligibility and perceptual quality of the enhanced speech signal. Future works include extending the proposed approach to a multi-microphone setting using multivariate $\alpha$-stable distributions [18, 20].

6. Conclusion

In this work, we proposed a speech enhancement method exploiting a speech model based on VAEs and a noise model based on alpha-stable distributions. At the expense of more interferences, the proposed $\alpha$-stable noise model reduces the amount of artifacts in the estimated speech signal, compared to the use of a Gaussian NMF-based noise model as in [10]. Overall, it is shown that the proposed approach improves the intelligibility and perceptual quality of the enhanced speech signal. Future works include extending the proposed approach to a multi-microphone setting using multivariate $\alpha$-stable distributions [18, 20].

1Actually $\alpha = 1.999$ because for $\alpha = 2$ the positive $\alpha$-stable distribution in (6) is degenerate.

2https://team.inria.fr/perception/icassp2019-ansvae/


7. REFERENCES


