Fast and efficient speech enhancement with variational autoencoders

Mostafa Sadeghi, Romain Serizel

FAST AND EFFICIENT SPEECH ENHANCEMENT WITH VARIATIONAL AUTOENCODERS

Mostafa Sadeghi and Romain Serizel

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

ABSTRACT

Unsupervised speech enhancement based on variational autoencoders has shown promising performance compared with the commonly used supervised methods. This approach involves the use of a pre-trained deep speech prior along with a parametric noise model, where the noise parameters are learned from the noisy speech signal with an expectation-maximization (EM)-based method. The E-step involves an intractable latent posterior distribution. Existing algorithms to solve this step are either based on computationally heavy Monte Carlo Markov Chain sampling methods and variational inference, or inefficient optimization-based methods. In this paper, we propose a new approach based on Langevin dynamics that generates multiple sequences of samples and comes with a total variation-based regularization to incorporate temporal correlations of latent vectors. Our experiments demonstrate that the developed framework makes an effective compromise between computational efficiency and enhancement quality, and outperforms existing methods.

***Index Terms*—** Speech enhancement, deep generative model, variational autoencoder, Langevin dynamics.

1. INTRODUCTION

Speech enhancement based on deep generative models has attracted much attention recently [1–7]. In particular, unsupervised speech enhancement based on variational autoencoder (VAE) [8] consists of a pre-training phase during which a deep speech prior is learned using only clean speech data. At test time, the learned speech prior is used to infer a target clean speech from a given noisy audio recording, e.g., by computing the posterior mean. In this step, a parametric Gaussian noise model is considered whose parameters are learned from the observed noisy speech signal based on expectation-maximization (EM) [3, 4].

This framework stays in contrast with the supervised methods based on deep learning, which train a deep neural architecture to directly map a given noisy speech signal to a clean version or a time-frequency mask [9], without explicitly modeling the statistical characteristics of speech signals. On the other hand, unsupervised methods extend the traditional speech enhancement approaches based on linear statistical models, e.g., non-negative matrix factorization (NMF), by incorporating the expressive representation learning frameworks provided by deep neural architectures [10]. As such, they are more interpretable, and could potentially achieve better generalization performance than supervised methods, because noise characteristics are modeled and learned only at test time [3, 4].

A disadvantage of unsupervised methods is that they are inefficient at test time, because of the iterative EM process. The computational bottleneck comes from the expectation step, which involves an intractable posterior distribution over the latent vectors of the model. This is tackled by a Markov Chain Monte Carlo (MCMC) based sampling method in [4]. A variational EM (VEM) algorithm is proposed in [11] to approximate the posterior with a parametric Gaussian form. Nevertheless, these approaches are computationally expensive. An optimization-based method is also proposed in [11, 12] that approximates the intractable posterior using only its mode. This, however, leads to performance degradation as the involved posterior is usually multi-modal. An alternative approach is proposed in [13] which reuses the pre-trained model to approximate the intractable posterior, leading to a fast inference algorithm. However, its performance depends largely on the quality and generalization capability of the learned speech model.

In this paper, we develop a computationally efficient sampling framework based on Langevin dynamics [14] to approximate the intractable posterior in the EM phase. This consists of generating multiple sets of samples by using only the gradient information of the log posterior along with injecting stochastic noise to better explore high-density regions of the posterior. Moreover, to incorporate the temporal correlations between consecutive latent vectors, we propose a total variation (TV) based regularization that further improves the performance. The proposed framework makes an effective compromise between the MCMC and optimization-based methods in terms of output quality and computations during the EM phase. Our experiments on speech enhancement confirm that the proposed methodology demonstrates a promising performance and outperforms the previous approaches.

The rest of the paper is organized as follows. Section 2 reviews VAE-based speech prior learning and enhancement. The proposed speech enhancement framework is detailed in Section 3. Experimental results are presented in Section 4. Finally, Section 5 concludes the paper.

2. BACKGROUND

In this section, we first review the VAE-based speech generative modeling framework and then discuss different approaches to perform speech enhancement using the learned deep speech prior.

2.1. Speech prior learning

Speech signals are first transformed to the time-frequency domain by computing the short-time Fourier transform (STFT), yielding a sequence of time frames denoted $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_T\}$, where $\mathbf{s}_t = [s_{ft}]_{f=1}^F \in \mathbb{C}^F$. To model the generative process of these time frames, a latent vector, denoted $\mathbf{z}_t \in \mathbb{R}^L$, is attributed to each time frame $\mathbf{s}_t$, which encodes the generative information of $\mathbf{s}_t$ in a lower dimension, i.e., $L \ll F$. The way $\mathbf{s}_t$ is generated from $\mathbf{z}_t$ is modeled by a parametric Gaussian from, i.e., $p_\theta(\mathbf{s}_t|\mathbf{z}_t)$, where its mean and variance, as nonlinear functions of $\mathbf{z}_t$, are provided by a so-called decoder network. For the prior of the latent codes, i.e., $p(\mathbf{z}_t)$, a standard Gaussian distribution is usually assumed. Therefore, the joint distribution of the observed and latent variables can be written as $p_\theta(\mathbf{s}_t, \mathbf{z}_t) = p_\theta(\mathbf{s}_t|\mathbf{z}_t)p(\mathbf{z}_t)$.

Having a training set of time frames $\mathbf{s}$, the next step is to learn the parameters of the generative model, i.e., θ, via EM. This amounts to solving the following problem (M-step):

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{t=1}^{T} \mathbb{E}_{p_\theta(\mathbf{z}_t|\mathbf{s}_t)} \{\log p_\theta(\mathbf{s}_t, \mathbf{z}_t)\} \qquad (1)$$

which requires computation of the intractable posterior distributions $p_\theta(\mathbf{z}_t|\mathbf{s}_t)$ (E-step). The solution proposed in VAE is to approximate this distribution with a parametric Gaussian form, where, similarly to the decoder, the mean and variance are functions of $\mathbf{s}_t$ parameterized with a deep neural network (DNN), called the encoder, with parameters denoted ψ. That is, $q_\psi(\mathbf{z}_t|\mathbf{s}_t) \approx p_\theta(\mathbf{z}_t|\mathbf{s}_t)$. The encoder and decoder parameters are then jointly learned following a computationally efficient framework that involves optimizing a lower-bound on the intractable data log-likelihood $\log p_\theta(\mathbf{s})$ [8].

2.2. Speech enhancement

Let us denote the STFT representation of a given noisy speech signal as $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\tilde{T}}\}$, where $\mathbf{x}_t = \mathbf{s}_t + \mathbf{b}_t$, with $\mathbf{b}_t$ corresponding to noise. To complete the observation model, in addition to the pre-trained speech prior, i.e., $p_\theta(\mathbf{s}_t, \mathbf{z}_t)$, we need to specify a noise model. A common choice is to consider a circularly symmetric Gaussian form $p_\phi(\mathbf{b}_t) \sim \mathcal{N}_c(\mathbf{0}, \text{diag}([\mathbf{WH}]_t))$ whose variance is parameterized with an NMF model [3]. The two non-negative low-rank matrices, $\mathbf{W}, \mathbf{H}$, constitute the noise parameters ϕ. To learn ϕ from $\mathbf{x}$, one would need to solve a similar problem as (1), that is

$$\phi^* = \underset{\phi}{\operatorname{argmax}} \sum_{t=1}^{\tilde{T}} \mathbb{E}_{p_\phi(\mathbf{z}_t|\mathbf{x}_t)} \{\log p_\phi(\mathbf{x}_t, \mathbf{z}_t)\} \qquad (2)$$

where $p_\phi(\mathbf{x}_t, \mathbf{z}_t) = p_\phi(\mathbf{x}_t|\mathbf{z}_t)p(\mathbf{z}_t)$, and the likelihood $p_\phi(\mathbf{x}_t|\mathbf{z}_t)$ can easily be computed noting the independence of $\mathbf{s}_t$ and $\mathbf{b}_t$ [4]. Once learned, the speech signal is estimated as the posterior mean $\hat{\mathbf{s}} = \mathbb{E}_{p_{\phi^*}(\mathbf{s}|\mathbf{x})} \{\mathbf{s}\}$ [4]. Here, too, the posterior $p_\phi(\mathbf{z}_t|\mathbf{x}_t)$ in (2) is intractable to compute. However, noting the similarity between (2) and (1), an immediate solution would be to follow the VAE framework and learn an approximate distribution. One could also fine-tune $q_\psi(\mathbf{z}_t|\mathbf{s}_t)$ on $\mathbf{x}_t$, as with the variational EM (VEM) approach proposed in [11]. Nevertheless, the VEM method is not computationally efficient, especially when the encoder is very complex.

Another solution is to approximate the intractable expectation in (2) by sampling from $p_\phi(\mathbf{z}_t|\mathbf{x}_t) \propto p_\phi(\mathbf{x}_t|\mathbf{z}_t)p(\mathbf{z}_t)$ to form a Monte-Carlo (MC) average using the Metropolis-Hastings algorithm [15], resulting in the MCEM method proposed in [4]. The Metropolis-Hastings algorithm is an iterative MCMC-based method that enables sampling from distributions that are known up to a normalization factor, e.g., $p_\phi(\mathbf{z}_t|\mathbf{x}_t)$. However, it could take many iterations to obtain a sequence of samples that resemble those from the true distribution. As such, the resulting MCEM approach might be computationally expensive. A lightweight alternative is to rely on a single sample to approximate the expectation (2) by finding the mode of $p_\phi(\mathbf{z}_t|\mathbf{x}_t)$, e.g., using a gradient-based solver, leading to the point-estimate EM (PEEM) method [11]. Nevertheless, given that the posterior is most likely multi-modal, this approach might fail to achieve high-quality results.

3. PROPOSED FRAMEWORK

In this section, we present our proposed framework based on Langevin dynamics for approximating the expectation in (2) by efficiently sampling from the involved intractable distribution. To this end, we first present a direct application of Langevin dynamics, and then propose a more specialized version by taking into account the specific structure of our problem. The proposed framework combines the advantages of the MCEM and PEEM methods and outperforms them, as will be discussed in Section 4.

3.1. Langevin dynamics

With Langevin dynamics, we can generate a sequence of samples from the intractable posterior $p_\phi(\mathbf{z}_t|\mathbf{x}_t)$ using only its score function defined as follows:

$$f(\mathbf{z}_t) = \nabla_{\mathbf{z}_t} \log p_\phi(\mathbf{z}_t|\mathbf{x}_t) = \nabla_{\mathbf{z}_t} g(\mathbf{z}_t) \qquad (3)$$

where

$$g(\mathbf{z}_t) = \log p_\phi(\mathbf{x}_t|\mathbf{z}_t) + \log p(\mathbf{z}_t). \qquad (4)$$

Given an initial state $\mathbf{z}_t^{(0)}$, the next states (samples) are produced according to the following rule ($k \geq 1$):

$$\mathbf{z}_t^{(k)} = \mathbf{z}_t^{(k-1)} + \frac{\eta}{2} f(\mathbf{z}_t^{(k-1)}) + \sqrt{\eta}\boldsymbol{\zeta}, \tag{5}$$

where $\boldsymbol{\zeta} \sim \mathcal{N}(0, \mathbf{I})$, and $\eta > 0$ is a step size. When $k \to \infty$ and $\eta \to 0$, one would have $\mathbf{z}_t^{(k)} \sim p_\phi(\mathbf{z}_t|\mathbf{x}_t)$ under some regularity conditions [14]. Injecting noise $\boldsymbol{\zeta}$ into the gradient update prevents the final sample to collapse into the modes and helps better explore high-density regions of the posterior.

3.2. Extended Langevin dynamics

In contrast to the original formulation of Langevin dynamics, we propose a more general approach where multiple sequences of samples are obtained in parallel, instead of a single sequence. Starting from $\mathbf{z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_{\tilde{T}}\}$, we first draw m different states per each latent vector $\mathbf{z}_t$, denoted $\mathbf{z}_{t,1}, \cdots, \mathbf{z}_{t,m}$, in a random walk manner by sampling from the following proposal distribution:

$$\mathbf{z}_{t,i}|\mathbf{z}_t \sim \mathcal{N}(\mathbf{z}_t, \sigma^2\mathbf{I}), \quad \forall t, i \tag{6}$$

where $\sigma^2 > 0$ is a given variance parameter. The above sampling can be equivalently written as $\mathbf{z}_{t,i} = \mathbf{z}_t + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Compared with (5), this additional level of stochasticity further helps cover high-density regions.

Furthermore, a direct application of Langevin dynamics would ignore time dependencies (correlations) between the sequences of latent codes $\{\mathbf{z}_1, \cdots \mathbf{z}_{\tilde{T}}\}$. To solve this issue, we propose to use a TV-based regularization by extending the original optimization function as follows:

$$h_\lambda(\mathbf{z}) = \sum_{t=1}^{\tilde{T}} g(\mathbf{z}_t) + \lambda \sum_{t=2}^{\tilde{T}} \|\mathbf{z}_t - \mathbf{z}_{t-1}\|_1 \tag{7}$$

where $\|.\|$ computes the ℓ_1 norm of a vector, i.e., the sum of absolute values, and $\lambda \geq 0$ is a trade-off parameter balancing the impact of the TV regularization. This effectively imposes a proximity constraint on the samples of consecutive latent vectors to incorporate their correlations. So, we define

$$f_\lambda(\mathbf{z}) = \nabla_{\mathbf{z}} h_\lambda(\mathbf{z}), \tag{8}$$

which replaces $f(.)$ in (5).

The overall Langevin dynamics (LD) and the EM algorithm to solve (2) for speech enhancement (LDEM) are summarized in Alg. 1 and Alg. 2, respectively. Updating ϕ in line 9 of Alg. 2 is done using multiplicative update rules [4]. The original LD method [14] usually requires many iterations to converge, after which, the initial states corresponding to the so-called burn-in period are discarded. The remaining states could then be used to approximate desired intractable expectations as a weighted Monte-Carlo average. Nevertheless, for our problem, there is no need to run the LD method for many iterations, thanks to the fact that here, the LD iterations are performed inside another iterative process, i.e., LDEM. Therefore, it benefits from warm-starting.

Algorithm 1 LD

1: **Require:** $\bar{\mathbf{z}}^{(0)} = \left\{\mathbf{z}_{t,i}^{(0)}\right\}_{t,i}$, K (sampling steps), η (step-size).
2: **for** $k = 1, \cdots, K$ **do**
3: $\quad \boldsymbol{\zeta} = \left\{\boldsymbol{\zeta}_{t,i}\right\}_{t,i}$, with $\boldsymbol{\zeta}_{t,i} \sim \mathcal{N}(0, \mathbf{I})$
4: $\quad \bar{\mathbf{z}}^{(k)} = \bar{\mathbf{z}}^{(k-1)} + \frac{\eta}{2} f_\lambda(\bar{\mathbf{z}}^{(k-1)}) + \sqrt{\eta}\boldsymbol{\zeta}$,
5: **end for**
6: **Output:** $\bar{\mathbf{z}}^{(K)} = \left\{\mathbf{z}_{t,i}^{(K)}\right\}_{t,i}$

Algorithm 2 LDEM

1: **Require:** $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{\tilde{T}}$, f_λ, σ, K, η, m, J (EM iterations).
2: **Initialize:** $\mathbf{z}$, $\mathbf{W}$, $\mathbf{H}$.
3: **for** $j = 1, \cdots, J$ **do**
4: $\quad \mathbf{z}_{t,i} = \mathbf{z}_t + \sigma\boldsymbol{\epsilon}_{t,i}$, with $\boldsymbol{\epsilon}_{t,i} \sim \mathcal{N}(0, \mathbf{I})$, $\forall t, i$
5: $\quad \{\mathbf{z}_{t,i}\}_{t,i} \leftarrow \mathrm{LD}(\{\mathbf{z}_{t,i}\}_{t,i})$
6: $\quad \phi \leftarrow \mathrm{argmax}_\phi \ \sum_{t,i} \log p_\phi(\mathbf{x}_t|\mathbf{z}_{t,i})$
7: **end for**
8: **Output:** $\phi = \{\mathbf{W}, \mathbf{H}\}$

4. EXPERIMENTS

Baselines. In this section, we evaluate the performance of the proposed LDEM framework for speech enhancement, and compare it with the PEEM and MCEM methods [11]. We used the publicly available PyTorch implementations of the two latter methods[1], and implemented LDEM based on that. We also tested the VEM method of [11] using its available implementation, but, unfortunately, it did not work. The poor performance of VEM was already observed in [11]. Furthermore, there is no public code for the method of [13], and our own implementation did not work. Therefore, we excluded these two methods from the baselines.

Evaluation metrics. The speech enhancement performance is measured based on the standard metrics, including the short-term objective intelligibility (STOI) measure [16], ranging in $[0, 1]$, the perceptual evaluation of speech quality (PESQ) score [17], ranging in $[-0.5, 4.5]$, and the scale-invariant signal-to-distortion ratio (SI-SDR) [18] in dB. Moreover, as a rough measure of the computational complexity of different methods, we report the average runtime in seconds. Our experiments were performed on an Intel Xeon Gold 6130 CPU machine.

VAE architecture. The architecture of the VAE considered in our experiments follows the one proposed in [4], which consists of an encoder and decoder, each having a 128-node single fully-connected hidden layer with hyperbolic tangent activation functions. The dimension of the latent space is set

[1] https://gitlab-research.centralesupelec.fr/sleglaive/icassp-2020-se-rvae

Table 1: Average values of the SI-SDR, PESQ, and STOI metrics for the input (unprocessed) and enhanced speech signals.

Metric	SI-SDR (dB)					PESQ					STOI				
Noise SNR (dB)	-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
PEEM [11]	-9.66	-4.35	0.57	5.49	10.33	1.60	1.80	2.06	2.36	2.67	0.15	0.24	0.36	0.49	0.63
MCEM [4]	-7.67	-1.48	3.34	7.81	12.00	1.55	1.84	2.18	2.49	2.78	**0.17**	**0.27**	**0.40**	**0.54**	**0.66**
LDEM ($\lambda: 0, m: 1$)	**-7.20**	**-1.03**	**3.76**	**8.18**	**12.37**	1.54	1.85	2.18	2.50	2.78	0.16	0.25	0.38	0.52	0.65
LDEM ($\lambda: 0.5, m: 1$)	-7.17	-1.08	3.70	8.16	12.34	1.58	1.87	2.20	2.51	2.80	0.17	0.27	0.40	0.53	0.66
LDEM ($\lambda: 5, m: 1$)	-7.28	-1.41	3.42	7.93	12.13	1.70	1.96	2.25	2.56	2.83	0.17	0.27	0.40	0.53	0.66
LDEM ($\lambda: 5, m: 5$)	-7.10	-1.26	3.60	8.07	12.27	**1.73**	**2.01**	**2.30**	**2.59**	**2.85**	**0.17**	**0.27**	**0.40**	**0.54**	**0.67**

Table 2: Average runtimes (in seconds) of different speech enhancement methods per test sample ($\sim$ 5-second long).

Method	PEEM [11]	MCEM [4]	LDEM ($m: 1$)	LDEM ($m: 5$)
runtime	5	32	5.4	18

to $L = 32$.

Datasets. To train the VAE model, we used the speech data in the TCD-TIMIT corpus [19], which contains speech utterances from 56 English speakers (39 for training, 8 for validation, and 9 for testing) with an Irish accent. There are 98 different audio files per speaker, each with an approximate length of 5 seconds, and sampled at 16 kHz. This results in about 8 hours of data. The STFT of the speech data is computed with a 1024 samples-long (64 ms) sine window, 75% overlap, and without zero-padding, yielding STFT frames of length $F = 513$.

To evaluate the speech enhancement performance, we used some pre-computed noisy versions of the TCD-TIMIT data presented in [20]. This includes six different noise types, namely *Living Room (LR)*, *White*, *Café*, *Car*, *Babble*, and *Street*, with five noise levels: -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB. For each noise type and noise level, we randomly selected 5 audio files from each test speaker, which resulted in a total of 1350 test files.

Parameters setting. The VAE model was trained with a batch size of 128 using the Adam optimizer [21], with a learning rate of 0.0001. We used early stopping on the validation set with patience of 20 epochs (i.e., training stops if the validation loss does not improve after 20 epochs). The number of EM iterations is set to $J = 100$ for all the methods. The MCEM algorithm was run using the default setting considered in [11]. The number of iterations in the E-step has been set to $K = 10$ for both PEEM and LDEM. The Adam optimizer has been used for PEEM, as in [11], with a learning rate of 0.005. For LDEM, we set $\eta = 0.005$ and $\sigma^2 = 0.01$. We have tried different values for the regularization parameter, λ, and the number of samples, m.

Results. The speech enhancement results per noise level are reported in Table 1. The average runtimes of different methods per each test sample ($\sim$ 5-second long) are reported in Table 2. We can draw several conclusions by inspecting the results. First, it can clearly be seen that LDEM, even without TV regularization and with $m = 1$, significantly outperforms PEEM in all the metrics, e.g., 2.7 dB average performance gain in SI-SDR, with approximately the same runtime. Moreover, LDEM achieves approximately the same PESQ scores as MCEM, but outperforms it in terms of SI-SDR, while having a much lower runtime. However, the STOI scores of LDEM are lower than those of MCEM.

As we increase λ (TV regularization parameter), we see a clear performance improvement, such that for $\lambda = 5$, LDEM outperforms MCEM in terms of PESQ, while showing similar performance in terms of SI-SDR and STOI. Again, LDEM achieves this with a much lower runtime. Furthermore, increasing the number of samples, m, causes a performance boost, such that, with a lower runtime, LDEM demonstrates better enhancement metrics than MCEM.

5. CONCLUSIONS

In this paper, we addressed the EM step of speech enhancement based on VAEs, which involves approximating an intractable latent posterior distribution. While existing approximating methods suffer either from high computational complexity or low-quality output results, we developed an efficient sampling-based method that effectively compromises the complexity and quality. More precisely, the proposed framework builds on Langevin dynamics and extends it to have multiple sets of samples as well as a total variation regularization to take into account the temporal correlations of latent vectors. Our experimental results showed that the proposed method outperforms the previous sampling-based approaches.

6. ACKNOWLEDGEMENT

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).

7. REFERENCES

[1] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[2] Aditya Arie Nugraha, Kouhei Sekiguchi, and Kazuyoshi 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.

[3] Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[4] Simon Leglaive, Laurent Girin, and Radu Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2018.

[5] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu 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.

[6] Huajian Fang, Guillaume Carbajal, Stefan Wermter, and Timo Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[7] Guillaume Carbajal, Julius Richter, and Timo Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[8] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *Proc. International Conference on Learning Representations (ICLR)*, April 2014.

[9] DeLiang Wang and Jitong 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.

[10] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[12] Hirokazu Kameoka, Li Li, Shota Inoue, and Shoji Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[13] Manuel Pariente, Antoine Deleforge, and Emmanuel Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders.," in *INTERSPEECH*, 2019.

[14] Max Welling and Yee W Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 2011, pp. 681–688.

[15] Christian P Robert, George Casella, and George Casella, *Monte Carlo statistical methods*, vol. 2, Springer, 1999.

[16] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper 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.

[17] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2001.

[18] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "SDR–half-baked or well done?," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[19] Naomi Harte and Eoin Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[20] Ahmed Hussen Abdelaziz et al., "NTCD-TIMIT: A new database and baseline for noise-robust audio-visual speech recognition.," in *Interspeech*, 2017, pp. 3752–3756.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations ICLR*, May 2015.