



HAL
open science

Unsupervised Speech Enhancement using Dynamical Variational Autoencoders

Xiaoyu Bie, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin

► **To cite this version:**

Xiaoyu Bie, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin. Unsupervised Speech Enhancement using Dynamical Variational Autoencoders. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2022, 30, pp.2993 - 3007. 10.1109/TASLP.2022.3207349 . hal-03295630

HAL Id: hal-03295630

<https://inria.hal.science/hal-03295630>

Submitted on 16 Dec 2022

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.

Unsupervised Speech Enhancement using Dynamical Variational Autoencoders

Xiaoyu Bie,¹ Simon Leglaive,² *Member, IEEE*, Xavier Alameda-Pineda,¹ *Senior Member, IEEE*, Laurent Girin³

Abstract—Dynamical variational autoencoders (DVAEs) are a class of deep generative models with latent variables, dedicated to model time series of high-dimensional data. DVAEs can be considered as extensions of the variational autoencoder (VAE) that include temporal dependencies between successive observed and/or latent vectors. Previous work has shown the interest of using DVAEs over the VAE for speech spectrograms modeling. Independently, the VAE has been successfully applied to speech enhancement in noise, in an unsupervised noise-agnostic setup that requires neither noise samples nor noisy speech samples at training time, but only requires clean speech signals. In this paper, we extend these works to DVAE-based single-channel unsupervised speech enhancement, hence exploiting both speech signals unsupervised representation learning and dynamics modeling. We propose an unsupervised speech enhancement algorithm that combines a DVAE speech prior pre-trained on clean speech signals with a noise model based on nonnegative matrix factorization, and we derive a variational expectation-maximization (VEM) algorithm to perform speech enhancement. The algorithm is presented with the most general DVAE formulation and is then applied with three specific DVAE models to illustrate the versatility of the framework. Experimental results show that the proposed DVAE-based approach outperforms its VAE-based counterpart, as well as several supervised and unsupervised noise-dependent baselines, especially when the noise type is unseen during training.

Index Terms—Speech enhancement, dynamical variational autoencoders, nonnegative matrix factorization, variational inference.

I. INTRODUCTION

SPEECH enhancement is a classical and fundamental problem in speech processing [1], [2], which aims to recover the clean speech signal from a noisy recording. Classical signal-processing-based solutions include spectral subtraction [3] and Wiener filtering [4] (which use noise and clean speech power spectral density estimates obtained from the noisy signal), and the short-time spectral amplitude estimator [5].

Recently, the advances in deep learning (DL) have opened new possibilities to tackle this task. The most widely studied approach to DL-based speech enhancement is that of a regression problem in the time-frequency (TF) domain where a deep neural network (DNN) is trained to map an input noisy speech signal into an output clean speech signal or into a denoising

TF mask that is applied on the noisy signal (see a review in [6]). We can refer to this general approach as a noisy-to-clean mapping (N2C). Recent works have considered N2C directly in the time domain instead of the TF domain [7] or leveraging generative adversarial networks (GANs) [8]–[10].

In the N2C approach, model training is typically done in a *supervised* manner using a *parallel* noisy-clean dataset, i.e., with noisy and clean versions of the same speech signal. Such parallel dataset must be prepared beforehand, the noisy version being obtained by summing the clean speech signal with noise. With a large amount of training data, DNNs can efficiently learn the N2C denoising mapping [6], [11]. However, supervised methods, whether they work in the TF domain or in the time domain, tend to have difficulties for generalizing to noise types and acoustic conditions that were not seen during training. And it is difficult, not to say impossible, to generate a dataset that includes all possible types and levels of noise (e.g., urban noise vs. office noise) and all possible acoustic conditions (e.g., different recording equipments, varying mouth-to-microphone distance and orientation, different reverberation characteristics, etc.).

Recent works in DL-based speech enhancement have tried to relax the constraints regarding the degree of supervision to ease the design of datasets and/or improve the generalization capability of the models. In the present speech enhancement context, relaxing the degree of supervision means that we go from methods using carefully aligned parallel noisy-clean data to methods using *non-parallel* noisy-clean data (i.e., the noisy examples are not the noisy version of the clean examples), which are easier to prepare, or noisy-only data, which are both easy to record and prepare, or clean-only data, which are not so easy to record, but which can lead to good generalization capabilities, as seen below. For example, the GANs employed in [12] and [13] use non-parallel noisy and clean speech examples. In the present context, such methods that do not require a parallel noisy-clean dataset, are referred to as *unsupervised*. They can be divided into two groups.

Unsupervised noise-dependent methods use noise examples or noisy speech examples only (they do not use clean speech signals). For example, a noisy-to-noisy (N2N) mapping approach, originally proposed for image denoising in [14], was applied to speech enhancement in [15], [16]. In this approach, the DNN input is still a noisy signal but the output clean signal is replaced with another noisy version of the same clean signal. This is supported by theoretical considerations: If the noises in the noisy input and output are zero-mean and uncorrelated, and an infinite number of examples is provided to the DNN, the latter will learn to output an average denoised version

¹ Inria Grenoble Rhône-Alpes, Univ. Grenoble-Alpes, France

² CentraleSupélec, IETR (UMR CNRS 6164), France

³ Univ. Grenoble Alpes, Grenoble-INP, CNRS, GIPSA-lab, France

This research was supported by ANR-3IA MIAI (ANR-19-P3IA-0003), ANR-JCJC ML3RI (ANR-19-CE33-0008-01), H2020 SPRING (funded by EC under GA #871245). This work was performed using HPC resources from the “Mésocentre” computing center of CentraleSupélec and Ecole Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France.

of the signal. The motivation for adopting this approach for speech enhancement is that clean audio signals are difficult and expensive to record (in studio condition) compared to noisy speech signals. However, the required assumptions are not met for audio signals. Not only the different channels of multichannel recordings do not contain the exact same clean speech signal, but they also contain correlated noise [17]–[20]. Moreover, in their experiments, the authors of [15], [16] have to rely on simulated clean speech plus noise signals, which questions the interest of the N2N approach compared to the conventional N2C mapping. However, the N2N approach inspired the more realistic noisy-target training method (NyTT) [21], in which noisy speech and extra noise is used. In the training step, the NyTT input is noisy speech further corrupted by an additional noise and the network is trained to recover the noisy speech at the output. Then in the test step, the network is supposed to recover the clean speech signal from a noisy speech input. Although NyTT lacks theoretical support, it was shown to obtain good results in practice [21]. In a different spirit, another approach not using clean speech signals is the MetricGAN-U method proposed by Fu *et al.* [22], an unsupervised version of their previous model MetricGAN [9], [10]. MetricGAN-U relies on the non-intrusive speech quality metric DNSMOS [23], which does not require using the clean speech signal, in contrast to the intrusive PESQ metric [24] used in MetricGAN. One problem with the unsupervised noise-dependent methods in general is that they learn the noise characteristics and acoustic conditions, and thus may generalize poorly to unseen noise and acoustic conditions, just like supervised N2C methods.

Alternatively, *unsupervised noise-agnostic* methods are based on a (deep) model of clean speech signals and do not learn the noise characteristics during training. Instead, the latter are estimated at test time on each speech sequence to denoise, hence conceptually letting the speech enhancement method the potential to adapt to any kind of noise. This setting was originally referred to as *semi-supervised* in the audio source separation literature [25]–[27], because it exploits a dataset of isolated signals for one of the sources in the mixture. This dataset is thus labeled with the class of the sound source, e.g., clean speech for speech enhancement. In the present paper, we choose to call this setting *unsupervised* because in the machine learning literature, semi-supervised refers to methods that are trained from both labeled and unlabeled datasets (e.g., [28]). In this context, a semi-supervised speech enhancement method would be trained from both a labeled dataset of noisy and clean speech signal pairs, and an unlabeled dataset containing only noisy or clean speech. While very interesting, this setting is not considered in this paper.

Examples of unsupervised noise-agnostic methods include that of Bando *et al.* [29], who proposed to use a variational autoencoder (VAE) [30], [31] to learn a prior distribution of the clean speech signals. At test time, the noise signal is modeled with Bayesian nonnegative matrix factorization (NMF) [32] whose parameters, as well as the VAE latent variables, are estimated with a Markov chain Monte Carlo (MCMC) algorithm, including a sampling of the NMF parameters and the VAE latent variables. The same general approach

was considered and extended in [33] within an expectation-maximization optimization framework, in [34] with an alpha-stable noise model, in [35] with efficient inference and learning algorithms, and in [36]–[38] for a multi-channel configuration. More recently, a guided VAE was proposed in [39], where the clean speech prior is defined conditionally on a voice activity detection or an ideal binary mask. This guiding information is provided by a supervised classifier, separately trained on noisy speech signals. Other supervised extensions of the speech enhancement framework combining a VAE clean speech model and an NMF noise model include [40] and [41].

Most of the above VAE-based unsupervised noise-agnostic speech enhancement methods focused on exploiting different distributions and algorithms. Very few works dealt with the inherent limitation of the VAE to handle sequential data correlated in time, as is the case of speech data. To the best of our knowledge, only two papers proposed generative approaches to speech enhancement based on VAE variants that can learn temporal dependencies: A recurrent VAE (RVAE) based on recurrent neural networks (RNNs) was proposed in [42] and stochastic temporal convolutional networks (TCNs) [43], [44] were used in [45], allowing the latent variables to have both hierarchical and temporal dependencies. Yet, a series of works have focused on developing extensions of the original VAE for time series (completely independently of the speech enhancement problem). The deep Kalman filter (DKF) [46], [47] is a DNN-based state-space model that combines a VAE with a non-linear first-order Markov model on the latent vectors. The variational RNN (VRNN) [48] and the stochastic RNN (SRNN) [49] are other temporal extensions of the VAE, with more complex temporal dependencies between the observed and latent data sequences, implemented with DNNs and RNNs. Actually, RVAE, DKF, VRNN, SRNN, and several other models [50]–[53] can all be seen as particular instances of a general class of models called dynamical variational autoencoders (DVAEs), which have been recently reviewed in [54]. In the present paper, we propose an unsupervised noise-agnostic speech enhancement algorithm based on the modeling of the clean speech signal with a DVAE and the use of the variational inference methodology [55]–[57]. We present this algorithm in the general context of the DVAE class of models, and then we apply it on three specific DVAE models in our experiments: RVAE (extending our preliminary work in [42]), DKF, and SRNN. To our knowledge, the present paper is the first in-depth study on the use of DVAEs, as a general class of models, for unsupervised speech enhancement.

The rest of the paper is organized as follows. Section II presents the technical background of DVAE models and their application to speech signals modeling. Section III presents the proposed DVAE-based speech enhancement algorithm. Section IV presents a series of experiments conducted with the three example DVAE models and their comparison with several state-of-the-art supervised and unsupervised speech enhancement methods. This includes cross-dataset experiments that investigate the generalization capabilities of the methods to unseen types of noise. Section V concludes the paper.

II. DVAE AND SPEECH MODELING

In this section, we first review the standard VAE and its extensions to temporal models, which are referred to as DVAEs. Then, we briefly introduce speech modeling using DVAE models. In the end of this section, we describe the practical implementation of three typical DVAE models.

A. VAEs and DVAEs

In a VAE [30], [31], an observed variable \mathbf{s} of high dimension F is assumed to be generated from an unobserved, or latent, random variable \mathbf{z} of low dimension $L \ll F$. Let $p_{\theta}(\mathbf{s}, \mathbf{z}) = p_{\theta_s}(\mathbf{s}|\mathbf{z})p_{\theta_z}(\mathbf{z})$ be the parametric generative model of their joint distribution, where $\theta = \theta_s \cup \theta_z$ denotes the set of parameters. In general, the latent vector \mathbf{z} is assumed to be generated from a very simple prior distribution, typically the multivariate standard Gaussian distribution $p_{\theta_z}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ (in that case, $\theta_z = \emptyset$). The parameters of $p_{\theta_s}(\mathbf{s}|\mathbf{z})$ are provided by a complex nonlinear function of \mathbf{z} , implemented with a deep neural network (DNN) (and θ_s is the set of parameters of this DNN).

Given a dataset $\mathbf{S} = \{\mathbf{s}_n\}_{n=1}^N$ of N i.i.d. samples of \mathbf{s} , a probabilistic model is traditionally optimized by maximizing the log-marginal likelihood (also called evidence), $\log p_{\theta}(\mathbf{S}) = \sum_{n=1}^N \log p_{\theta}(\mathbf{s}_n)$, over the parameter set θ . In the VAE case, the complexity of $p_{\theta_s}(\mathbf{s}|\mathbf{z})$ makes the marginalization over the latent variable, and thus the computation of $\log p_{\theta}(\mathbf{S})$, intractable, and the same for the posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{s})$. Therefore, instead of directly maximizing $\log p_{\theta}(\mathbf{S})$, an inference model $q_{\phi}(\mathbf{z}|\mathbf{s}) \approx p_{\theta}(\mathbf{z}|\mathbf{s})$ is introduced, which is also defined by a DNN (of parameters ϕ). Then, the following evidence lower bound (ELBO) is computed:

$$\begin{aligned} \mathcal{L}(\theta, \phi; \mathbf{S}) &= \sum_{n=1}^N \mathbb{E}_{q_{\phi}(\mathbf{z}_n|\mathbf{s}_n)} [\log p_{\theta}(\mathbf{s}_n, \mathbf{z}_n) - \log q_{\phi}(\mathbf{z}_n|\mathbf{s}_n)] \\ &= \log p_{\theta}(\mathbf{S}) - \sum_{n=1}^N D_{\text{KL}}(q_{\phi}(\mathbf{z}_n|\mathbf{s}_n) || p_{\theta}(\mathbf{z}_n|\mathbf{s}_n)), \end{aligned} \quad (1)$$

where $D_{\text{KL}}(\cdot)$ denotes the Kullback-Leibler (KL) divergence, which is always non-negative. The generative model $p_{\theta_s}(\mathbf{s}|\mathbf{z})$ and the inference model $q_{\phi}(\mathbf{z}|\mathbf{s})$ are jointly trained by maximizing the ELBO with respect to θ_s and ϕ , using stochastic gradient descent combined with sampling [30], [31].

While the vanilla VAE assumes statistical independence among observation vectors, DVAEs can be seen as an extension of the VAE for modeling sequential data correlated in time [54]. A DVAE keeps the global encoder-decoder philosophy of the VAE, but considers a sequence of (high-dimensional) observed random vectors $\mathbf{s}_{1:T} = \{\mathbf{s}_t\}_{t=1}^T$ and a corresponding sequence of (low-dimensional) latent vectors $\mathbf{z}_{1:T} = \{\mathbf{z}_t\}_{t=1}^T$. A DVAE is thus defined by the joint probability density function (pdf) of observed and latent sequences $p_{\theta}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T})$,

which can be factorized using the chain rule:¹

$$\begin{aligned} p_{\theta}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T}) &= \prod_{t=1}^T p_{\theta}(\mathbf{s}_t, \mathbf{z}_t | \mathbf{s}_{1:t-1}, \mathbf{z}_{1:t-1}) \\ &= \prod_{t=1}^T p_{\theta_s}(\mathbf{s}_t | \mathbf{s}_{1:t-1}, \mathbf{z}_{1:t}) p_{\theta_z}(\mathbf{z}_t | \mathbf{s}_{1:t-1}, \mathbf{z}_{1:t-1}). \end{aligned} \quad (2)$$

$$(3)$$

As for the VAE, the exact posterior distribution $p_{\theta}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ is not analytically tractable. Consequently, an approximate posterior distribution $q_{\phi}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ is introduced and it can be factorized using the chain rule as:

$$q_{\phi}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T}) = \prod_{t=1}^T q_{\phi}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{s}_{1:T}). \quad (4)$$

Chaining the inference and generation, the training of DVAEs is done by maximizing the ELBO on a set of training vector sequences, the ELBO being here defined by (for a single observed and latent data sequence):

$$\mathcal{L}(\theta, \phi; \mathbf{s}_{1:T}) = \mathbb{E}_{q_{\phi}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})} [\ln p_{\theta}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T}) - \ln q_{\phi}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})]. \quad (5)$$

When writing a joint distribution as a product of conditional distributions using the chain rule, a specific ordering of the variables has to be chosen. Among different possibilities, we chose a *causal* ordering to write the factorization in (2) and (3): The generation of \mathbf{s}_t and \mathbf{z}_t uses their past values $\mathbf{s}_{1:t-1}$ and $\mathbf{z}_{1:t-1}$ (plus \mathbf{z}_t for generating \mathbf{x}_t). In the DVAE literature, almost all models are causal [54]. Each of them can be seen as a special case of the general expression (3) where the dependencies in $p_{\theta}(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t})$ and $p_{\theta}(\mathbf{z}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t-1})$ are simplified, which may also affect the choice of the inference model $q_{\phi}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})$ in (4). In addition, a given DVAE model can have different implementations with various types of DNNs, see [54] for an extensive discussion on this topic.

B. Speech modeling using DVAEs

The VAE and the DVAE class of models have been used to model different kinds of data. In this subsection, we discuss the use of DVAEs for modeling speech signals in the short-term Fourier transform (STFT) domain. Fig. 1 illustrates this process. Let $\mathbf{s}_{1:T} = \{\mathbf{s}_t \in \mathbb{C}^F\}_{t=1}^T$ denote an $F \times T$ sequence of complex-valued STFT frames, where t is the time-frame index. Each vector $\mathbf{s}_t = \{s_{ft} \in \mathbb{C}\}_{f=1}^F$ represents the speech short-term spectrum at time index t , and f is the frequency bin. As indicated above, $\mathbf{s}_{1:T}$ is associated with an $L \times T$ sequence of latent variables $\mathbf{z}_{1:T} = \{\mathbf{z}_t \in \mathbb{R}^L\}_{t=1}^T$, with $L \ll F$.

In speech and audio processing, the Fourier coefficients in $\mathbf{s}_t \in \mathbb{C}^F$ are usually assumed to be independent and distributed according to a complex Gaussian circularly symmetric distribution [58] (denoted below by \mathcal{N}_c), whose variance vary over time and frequency [5], [59] (the circularly symmetric assumption means that the phase follows a uniform distribution

¹Here and in all the following, we take the convention that $\mathbf{s}_{1:0} = \mathbf{z}_{1:0} = \emptyset$. For $t = 1$ the first term of the product in (2) and (3) is thus $p_{\theta}(\mathbf{s}_1, \mathbf{z}_1)$ and $p_{\theta}(\mathbf{s}_1|\mathbf{z}_1)p_{\theta}(\mathbf{z}_1)$, respectively.

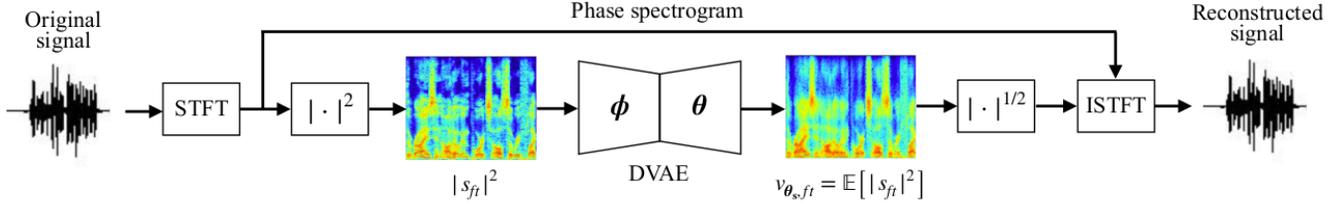


Fig. 1. Speech analysis-resynthesis with a DVAE model in the STFT domain. The speech power spectrogram is used as input to the DVAE and the output is the variance of a complex Gaussian model. The audio waveform is reconstructed by inverse STFT using the phase of the original signal.

in $[0, 2\pi)$). Thus, for all time frames $t \in \{1, \dots, T\}$, the DVAE generative model of speech signal is defined as follows:

$$p_{\theta_s}(s_t | s_{1:t-1}, \mathbf{z}_{1:t}) = \mathcal{N}(s_t; \mathbf{0}, \Sigma_{\theta_s, t}), \quad (6)$$

$$p_{\theta_z}(z_t | s_{1:t-1}, \mathbf{z}_{1:t-1}) = \mathcal{N}(z_t; \boldsymbol{\mu}_{\theta_z, t}, \Sigma_{\theta_z, t}), \quad (7)$$

where the diagonal covariance matrix $\Sigma_{\theta_s, t} = \text{diag}\{\mathbf{v}_{\theta_s, t}\}$ is provided by a DNN that takes as input the conditioning variables in (6), namely $(s_{1:t-1}, \mathbf{z}_{1:t})$. Similarly, $\boldsymbol{\mu}_{\theta_z, t}$ and $\Sigma_{\theta_z, t} = \text{diag}\{\mathbf{v}_{\theta_z, t}\}$ are provided by a DNN that takes as input the conditioning variables in (7), namely $(s_{1:t-1}, \mathbf{z}_{1:t-1})$.² We denote by $\boldsymbol{\theta} = \boldsymbol{\theta}_s \cup \boldsymbol{\theta}_z$ the parameters of the DNNs involved in (6) and (7).

As for the inference model, it is given by:

$$q_{\phi}(z_t | z_{1:t-1}, s_{1:T}) = \mathcal{N}(z_t; \boldsymbol{\mu}_{\phi, t}, \Sigma_{\phi, t}), \quad (8)$$

where $\boldsymbol{\mu}_{\phi, t}$ and $\Sigma_{\phi, t} = \text{diag}\{\mathbf{v}_{\phi, t}\}$ are provided by a DNN taking $(z_{1:t-1}, s_{1:T})$ as input and whose parameters are denoted by ϕ .

Even if the complex-valued vector sequence $s_{1:t-1}$ or $s_{1:T}$ is used as a conditioning variable in (6)-(8), in practice we use the modulus-squared values of these variables at the encoder and decoder input. In other words, the DVAE encoder and decoder distribution parameters are computed using sequences of vectors with entries equal to $|s_{ft}|^2$, as illustrated in Fig. 1. Note that the modulus-squared of data is homogeneous with the decoder output (the variance vector $\mathbf{v}_{\theta_s, t}$).

Given the generative model (6), (7) and the inference model (8), we can develop the ELBO in (5) as follows (for one data sequence):

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}, \phi; s_{1:T}) \stackrel{c}{=} & - \sum_{f,t=1}^{F,T} \mathbb{E}_{q_{\phi}(z_{1:t} | s_{1:T})} [d_{\text{IS}}(|s_{ft}|^2, v_{\theta_s, ft})] \\ & + \frac{1}{2} \sum_{l,t=1}^{L,T} \left[\ln \frac{v_{\phi, lt}}{v_{\theta_z, lt}} - \frac{v_{\phi, lt} + (\boldsymbol{\mu}_{\phi, lt} - \boldsymbol{\mu}_{\theta_z, lt})^2}{v_{\theta_z, lt}} \right], \quad (9) \end{aligned}$$

where $\stackrel{c}{=}$ denotes equality up to an additive constant w.r.t. $\boldsymbol{\theta}$ and ϕ , $d_{\text{IS}}(q, p) = q/p - \ln(q/p) - 1$ is the Itakura-Saito (IS) divergence [32], $v_{\theta_s, ft} \in \mathbb{R}_+$ is the f -th entry of $\mathbf{v}_{\theta_s, t}$, and $\{\boldsymbol{\mu}_{\phi, lt} \in \mathbb{R}, \boldsymbol{\mu}_{\theta_z, lt} \in \mathbb{R}, v_{\phi, lt} \in \mathbb{R}_+, v_{\theta_z, lt} \in \mathbb{R}_+\}$ are the l -th entry of $\{\boldsymbol{\mu}_{\phi, t}, \boldsymbol{\mu}_{\theta_z, t}, \mathbf{v}_{\phi, t}, \mathbf{v}_{\theta_z, t}\}$, respectively.

²It is important to note that, in DVAEs, a parameter of a distribution is always a function of the variables that come after the conditioning bar. In the rest of the paper, we will generally omit to rewrite these variables in the right-hand-side of the probabilistic modeling equations, for concision, but we may punctually make these dependencies explicit when it eases the understanding.

C. Three representative DVAE models

In this subsection, we briefly present three DVAE models, namely DKF, RVAE, and SRNN, which are representative of the versatility of the DVAE class (see the review in [54]) and which we used in practice in our speech enhancement method, with results reported in Section IV.

1) *DKF*: As briefly stated in the introduction, DKF [46], [47] is the simplest DVAE model, following the structure of a basic state-space model with a first-order Markov model on the latent vectors, i.e., \mathbf{z}_t is generated from \mathbf{z}_{t-1} , and an instantaneous observation model, i.e., s_t is generated from \mathbf{z}_t :

$$p_{\theta}(s_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^T p_{\theta_s}(s_t | z_t) p_{\theta_z}(z_t | z_{t-1}). \quad (10)$$

The inference model is defined by:

$$q_{\phi}(z_{1:T} | s_{1:T}) = \prod_{t=1}^T q_{\phi}(z_t | z_{t-1}, s_{t:T}). \quad (11)$$

2) *RVAE*: As opposed to DKF, RVAE [42] does not consider any dynamical model for the latent vectors, which are assumed i.i.d., with $p(z_t) = \mathcal{N}(z_t; \mathbf{0}, \mathbf{I})$. This prior therefore lacks parameters and does not involve a neural network, as in standard VAEs. However, the observation model in RVAE is more complex than in DKF, as s_t is generated from the current and previous latent variables $\mathbf{z}_{1:t}$. Formally, the RVAE generative and inference models are defined by:

$$p_{\theta}(s_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^T p_{\theta_s}(s_t | \mathbf{z}_{1:t}) p(z_t), \quad (12)$$

$$q_{\phi}(z_{1:T} | s_{1:T}) = \prod_{t=1}^T q_{\phi}(z_t | z_{1:t-1}, s_{t:T}). \quad (13)$$

RVAE is the only model that was presented in the literature with both a causal and non-causal form [42], [54]. The non-causal form is obtained by simply replacing $\mathbf{z}_{1:t}$ with $\mathbf{z}_{1:T}$ in (12) and by replacing $s_{t:T}$ with $s_{1:T}$ in (13).

3) *SRNN*: In SRNN [49], the latent vector \mathbf{z}_t is generated not only from \mathbf{z}_{t-1} (as in DKF), but also from $s_{1:t-1}$. And s_t is generated not only from \mathbf{z}_t (as in DKF) but also from

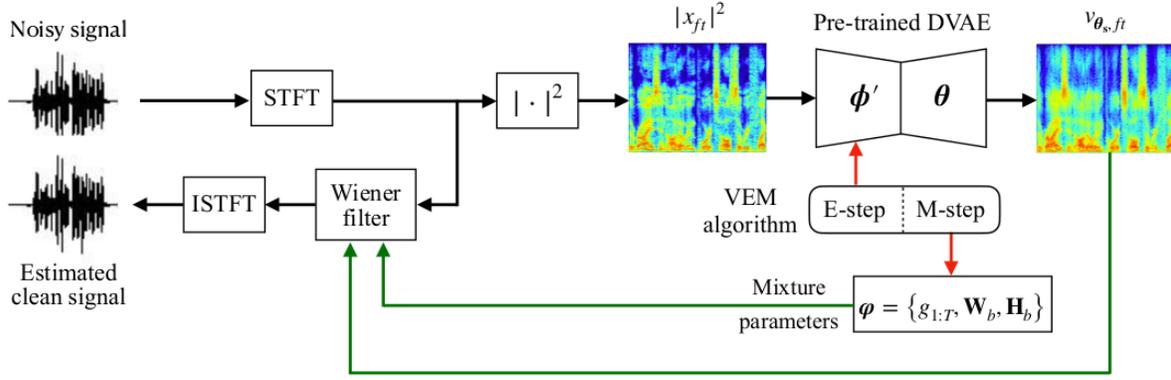


Fig. 2. Overview of the proposed speech enhancement method. The pre-trained DVAE is used within a VEM algorithm for speech enhancement. The DVAE encoder is fine-tuned during the E-step and the mixture parameters are estimated in the M-step (red arrows), see Section III-C and Algorithm 1. The clean speech signal is estimated by filtering the noisy signal with a Wiener filter combining the DVAE output parameters and the estimated mixture parameters (green arrows).

$\mathbf{s}_{1:t-1}$, so that SRNN actually corresponds to an autoregressive model. The generative and inference models are defined by:

$$p_{\theta}(\mathbf{s}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^T p_{\theta_s}(s_t | \mathbf{z}_t, \mathbf{s}_{1:t-1}) p_{\theta_z}(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{s}_{1:t-1}), \quad (14)$$

$$q_{\phi}(\mathbf{z}_{1:T} | \mathbf{s}_{1:T}) = \prod_{t=1}^T q_{\phi}(\mathbf{z}_t | \mathbf{z}_{t-1}, \mathbf{s}_{1:T}). \quad (15)$$

III. DVAE FOR SPEECH ENHANCEMENT

This section presents the proposed unsupervised noise-agnostic DVAE-based speech enhancement algorithm, where the clean speech signal is modeled with a DVAE and the noise is modeled with nonnegative matrix factorization (NMF) [32]. It is an extended version of the algorithm proposed in [42] for the RVAE model, with a more general formulation applicable to any other DVAE model. The proposed method is illustrated in Fig. 2. We assume that the DVAE-based clean speech generative and inference models defined in (6)–(7) and (8), respectively, have been learned, i.e., the DNN parameters θ and ϕ have been estimated from a dataset of clean speech signals during an independent training stage (see Section II-B). The objective of speech enhancement is to use this pre-trained DVAE model to estimate the clean speech signal when only the noisy mixture is observed. This is done with a variational expectation-maximization (VEM) algorithm [55]–[57]. We recall that this method is unsupervised, since no pair of clean and noisy speech examples are used. Moreover it is noise-agnostic since it does not make any assumption on the noise type, except that it can be modeled with an NMF model, and the noise model NMF parameters are estimated independently for each noisy sequence to process.

In the rest of this section, we first introduce the noise and mixture models, then we develop the general strategy to estimate the clean speech signal modeled by a DVAE when only the mixture signal is available, and finally we present the VEM algorithm used to estimate the remaining unknown

model parameters. Throughout this section, $\mathbf{s}_{1:T} = \{\mathbf{s}_t \in \mathbb{C}^F\}_{t=1}^T$, $\mathbf{b}_{1:T} = \{\mathbf{b}_t \in \mathbb{C}^F\}_{t=1}^T$, and $\mathbf{x}_{1:T} = \{\mathbf{x}_t \in \mathbb{C}^F\}_{t=1}^T$ respectively denote the STFT of the clean speech signal, the noise signal, and the noisy speech signal.

A. Noise and mixture models

As in [33], [42], we consider a Gaussian noise model with NMF parameterization of the variance [32]. Independently for all time frames $t \in \{1, \dots, T\}$, we define:

$$p(\mathbf{b}_t) = \mathcal{N}_c(\mathbf{b}_t; \mathbf{0}, \Sigma_{\mathbf{b},t}), \quad (16)$$

where $\Sigma_{\mathbf{b},t} = \text{diag}\{(\mathbf{W}_b \mathbf{H}_b)_{:,t}\}$ with $\mathbf{W}_b \in \mathbb{R}_+^{F \times K}$ and $\mathbf{H}_b \in \mathbb{R}_+^{K \times T}$. The rank of the factorization K is usually chosen such that $K(F+T) \ll FT$.

We consider that the noisy speech is a mixture of the noise defined in (16) and the clean speech defined in (6) and (7):

$$\mathbf{x}_t = \sqrt{g_t} \mathbf{s}_t + \mathbf{b}_t, \quad (17)$$

where $g_t \in \mathbb{R}_+$ is a frame-dependent frequency-independent gain parameter scaling the speech signal level at each time frame. This parameter enables to take into account the potentially different loudness between the clean speech training examples used to learn the DVAE model and the speech signal in the test noisy sequence we have to denoise [33].

From (6), (16) and (17), and by assuming the independence of the speech and noise signals, we have for all $t \in \{1, \dots, T\}$:

$$p_{\theta_x}(\mathbf{x}_t | \mathbf{s}_{1:t-1}, \mathbf{z}_{1:t}) = \mathcal{N}_c(\mathbf{x}_t; \mathbf{0}, \Sigma_{\theta_x,t}), \quad (18)$$

where $\Sigma_{\theta_x,t} = \text{diag}\{g_t v_{\theta_s,t} + (\mathbf{W}_b \mathbf{H}_b)_{:,t}\}$ and θ_x is the union of the speech generative model parameters θ_s and the mixture model parameters $\varphi = \{g = [g_1, \dots, g_T]^T, \mathbf{W}_b, \mathbf{H}_b\}$. As already mentioned in Section II-B, $v_{\theta_s,t}$ is actually a function of $(\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t})$. Note that it is clear from (16) and (17) that given the clean speech frame \mathbf{s}_t , the noisy speech frame \mathbf{x}_t is characterized by:

$$p_{\varphi}(\mathbf{x}_t | \mathbf{s}_t) = \mathcal{N}_c(\mathbf{x}_t; \sqrt{g_t} \mathbf{s}_t, \Sigma_{\mathbf{b},t}). \quad (19)$$

B. Speech reconstruction

Now we consider the problem of reconstructing the clean speech signal from the observed mixture signal, which consists in computing the following posterior mean vector:

$$\hat{\mathbf{s}}_t = \mathbb{E}_{p_{\theta}(\mathbf{s}_t|\mathbf{x}_{1:T})}[\mathbf{s}_t]. \quad (20)$$

However, we cannot write the posterior $p_{\theta}(\mathbf{s}_t|\mathbf{x}_{1:T})$ analytically, which makes the above expectation intractable. However, leveraging the speech model defined previously, we can approximate it by introducing random variables that are then marginalized.

1) *Introducing the past and current latent variables:* We start from marginalizing with respect to $\mathbf{z}_{1:t}$:

$$\begin{aligned} p_{\theta}(\mathbf{s}_t|\mathbf{x}_{1:T}) &= \int p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:T}) p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T}) d\mathbf{z}_{1:t} \\ &= \mathbb{E}_{p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})} [p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:T})]. \end{aligned} \quad (21)$$

Using (21) to rewrite (20), the estimate of the clean speech signal at time t is given by:

$$\hat{\mathbf{s}}_t = \mathbb{E}_{p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})} [\mathbb{E}_{p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:T})}[\mathbf{s}_t]]. \quad (22)$$

Let us now focus on the inner expectation, taken with respect to $p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:T})$. We will come back later on the outer expectation taken with respect to $p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})$. Using Bayes rule, we have:

$$p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:T}) = \frac{p_{\theta}(\mathbf{x}_{1:T}|\mathbf{s}_t, \mathbf{z}_{1:t}) p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}) p_{\theta}(\mathbf{z}_{1:t})}{p_{\theta}(\mathbf{z}_{1:t}, \mathbf{x}_{1:T})} \quad (23)$$

$$\propto p_{\theta}(\mathbf{x}_{1:T}|\mathbf{s}_t, \mathbf{z}_{1:t}) p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}) \quad (24)$$

$$\approx p_{\theta}(\mathbf{x}_t|\mathbf{s}_t) p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}). \quad (25)$$

The exact computation of $p_{\theta}(\mathbf{x}_{1:T}|\mathbf{s}_t, \mathbf{z}_{1:t})$ requires the marginalisation of $p_{\theta}(\mathbf{x}_{1:T}, \mathbf{s}_{1:t-1, t+1:T}, \mathbf{z}_{t+1:T}|\mathbf{s}_t, \mathbf{z}_{1:t})$ w.r.t. the undesired variables. This would require not only marginalising from future latent codes, but also from past and future clean speech, which is clearly not feasible. Instead, we approximate (24) with (25) by considering only the signal mixture model $p_{\theta}(\mathbf{x}_t|\mathbf{s}_t)$, as defined in (19).

2) *Introducing the past speech vectors:* Then, it comes to estimating $p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t})$ in (25). To do so, we introduce and then marginalize the past speech vectors $\mathbf{s}_{1:t-1}$:

$$\begin{aligned} p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}) &= \int p_{\theta}(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t}) p_{\theta}(\mathbf{s}_{1:t-1}|\mathbf{z}_{1:t}) d\mathbf{s}_{1:t-1} \\ &= \int p_{\theta}(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t}) \left[\prod_{\tau=1}^{t-1} p_{\theta}(\mathbf{s}_{\tau}|\mathbf{s}_{1:\tau-1}, \mathbf{z}_{1:\tau}) \right] d\mathbf{s}_{1:t-1} \\ &= \mathbb{E}_{\prod_{\tau=1}^{t-1} p_{\theta}(\mathbf{s}_{\tau}|\mathbf{s}_{1:\tau-1}, \mathbf{z}_{1:\tau})} [p_{\theta}(\mathbf{s}_t|\mathbf{s}_{1:t-1}, \mathbf{z}_{1:t})], \end{aligned} \quad (26)$$

where in the second line we used the fact that \mathbf{s}_{τ} is conditionally independent of $\mathbf{z}_{\tau+1:t}$.

When computing (26), we are facing two problems: First, the expectation is intractable; and second, in a speech enhancement framework, we do not have access to the past ground-truth clean speech vectors $\mathbf{s}_{1:t-1}$ (as opposed to the DVAE training procedure which is done using sequences of

clean speech signals). Therefore, we approximate $p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t})$ as follows:

$$\begin{aligned} p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}) &\approx p_{\theta}(\mathbf{s}_t|\tilde{\mathbf{s}}_{1:t-1}, \mathbf{z}_{1:t}) \\ &= \mathcal{N}_c(\mathbf{s}_t; \mathbf{0}, \Sigma_{\theta_s, t}(\tilde{\mathbf{s}}_{1:t-1}, \mathbf{z}_{1:t})), \end{aligned} \quad (27)$$

where $\Sigma_{\theta_s, t}(\tilde{\mathbf{s}}_{1:t-1}, \mathbf{z}_{1:t}) = \text{diag}\{\mathbf{v}_{\theta_s, t}(\tilde{\mathbf{s}}_{1:t-1}, \mathbf{z}_{1:t})\}$ and $\tilde{\mathbf{s}}_t$ is computed recursively as $\tilde{\mathbf{s}}_t = \mathbf{v}_{\theta_s, t}(\tilde{\mathbf{s}}_{1:t-1}, \mathbf{z}_{1:t})$.³ In practice, the decoder output at time frame $t-1$ is re-injected at the decoder input at the next time frame t . This part of the process is necessary only for SRNN, and more generally for any autoregressive DVAE. For non-autoregressive DVAE models, such as RVAE and DKF, $\Sigma_{\theta_s, t}$ is only computed from the sequence of latent vectors.

3) *Computing the conditional posterior:* Substituting (19) and (27) into (25), we have:

$$\begin{aligned} p_{\theta}(\mathbf{s}_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:T}) &\approx \mathcal{N}_c(\mathbf{x}_t; \sqrt{g_t} \mathbf{s}_t, \Sigma_{\mathbf{b}, t}) \mathcal{N}_c(\mathbf{s}_t; \mathbf{0}, \Sigma_{\theta_s, t}) \\ &= \mathcal{N}_c(\mathbf{s}_t; \mathbf{m}_{s, t}, \Sigma_{s, t}), \end{aligned} \quad (28)$$

where

$$\mathbf{m}_{s, t} = \sqrt{g_t} \Sigma_{\theta_s, t} (g_t \Sigma_{\theta_s, t} + \Sigma_{\mathbf{b}, t})^{-1} \mathbf{x}_t, \quad (29)$$

$$\Sigma_{s, t} = \Sigma_{\theta_s, t} \Sigma_{\mathbf{b}, t} (g_t \Sigma_{\theta_s, t} + \Sigma_{\mathbf{b}, t})^{-1}. \quad (30)$$

Finally, from (22), (28) and (29), the estimate of the clean speech signal is given by:

$$\hat{\mathbf{s}}_t \approx \mathbb{E}_{p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})} \left[\sqrt{g_t} \Sigma_{\theta_s, t} (g_t \Sigma_{\theta_s, t} + \Sigma_{\mathbf{b}, t})^{-1} \right] \mathbf{x}_t, \quad (31)$$

where we recall that $\Sigma_{\theta_s, t}$ is actually a function of $(\tilde{\mathbf{s}}_{1:t-1}, \mathbf{z}_{1:t})$. This speech signal estimate can be seen as the output of a ‘‘probabilistic’’ Wiener filter, i.e., a Wiener filter averaged over all possible realizations of the latent variables according to their posterior distribution $p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})$.

The expectation in (31) is intractable, but similarly as before we can approximate it by

$$\hat{\mathbf{s}}_t \approx \sqrt{g_t} \Sigma_{\theta_s, t} (g_t \Sigma_{\theta_s, t} + \Sigma_{\mathbf{b}, t})^{-1} \mathbf{x}_t, \quad (32)$$

where $\Sigma_{\theta_s, t} = \Sigma_{\theta_s, t}(\tilde{\mathbf{s}}_{1:t-1}, \tilde{\mathbf{z}}_{1:t})$ and $\tilde{\mathbf{z}}_{1:t}$ is sampled from $p_{\theta}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})$. This posterior distribution is also intractable. We thus propose to use instead a variational approximation $q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ whose parameters ϕ' need to be jointly estimated together with the noisy mixture model parameters φ . In the next section, we propose a VEM algorithm to do that. This generalizes the algorithm developed for RVAE in [42] to the whole class of DVAE models.

C. VEM algorithm for model parameters estimation

Now that we have an expression for the clean speech signal estimate, what remains to be estimated is the set of mixture model parameters φ (the NMF noise model parameters and the gains) and the parameters ϕ' of the variational distribution $q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$. Using a VEM algorithm, we will maximize

³Here we explicitly write the dependency of the covariance matrix on $\tilde{\mathbf{s}}_{1:t-1}$ and $\mathbf{z}_{1:t}$, to make the use of the DVAE model clear. In the following we will omit it again for concision of presentation.

the following ELBO defined for the noisy speech observations $\mathbf{x}_{1:T}$:

$$\mathcal{L}(\phi', \varphi; \mathbf{x}_{1:T}) = \mathbb{E}_{q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})} [\ln p_{\varphi}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) - \ln q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})]. \quad (33)$$

It can be shown that this corresponds to (i) maximizing with respect to φ a lower bound of the intractable log-marginal likelihood $\ln p_{\varphi}(\mathbf{x}_{1:T})$, and (ii) minimizing with respect to ϕ' the KL divergence between the variational distribution $q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ and the intractable exact posterior $p_{\theta}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ [55]. The proposed VEM algorithm thus consists in iterating between the following variational E and M steps.

1) *Variational E-step*: We consider a variational distribution of the same form as the DVAE inference model:

$$q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^T q_{\phi'}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T}), \quad (34)$$

where $q_{\phi'}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T})$ is defined as in (8), except that \mathbf{s}_t is replaced by \mathbf{x}_t . The noisy speech frames can be considered as out-of-sample data for the DVAE model trained on clean speech signals [60]. Therefore, similarly as in [42], we can fine-tune the pre-trained DVAE inference network on the noisy speech test signal, by maximizing the ELBO in (33) w.r.t. ϕ' . This objective function can be developed by marginalizing and sampling, similarly to what was done in Section III-B. This leads to the following expression:

$$\begin{aligned} \mathcal{L}(\phi', \varphi^*; \mathbf{x}_{1:T}) \stackrel{c}{=} & - \sum_{f=1}^F \sum_{t=1}^T \mathbb{E}_{q_{\phi'}} \left[\ln v_{\varphi^*, ft} + \frac{|x_{ft}|^2}{v_{\varphi^*, ft}} \right] + \\ & \frac{1}{2} \sum_{l=1}^L \sum_{t=1}^T \left[\ln v_{\phi', lt} - \ln v_{\theta_{\mathbf{z}}, lt} - \frac{v_{\phi', lt} + (\mu_{\phi', lt} - \mu_{\theta_{\mathbf{z}}, lt})^2}{v_{\theta_{\mathbf{z}}, lt}} \right], \end{aligned} \quad (35)$$

where x_{ft} denotes the f -th entry of \mathbf{x}_t , φ^* denotes the current estimate of the mixture model parameters, and $v_{\varphi^*, ft}$ is the f -th diagonal entry of $\Sigma_{\theta_{\mathbf{x}}, t}$, whose expectation is intractable and is approximated with:

$$v_{\varphi^*, ft} = g_t v_{\theta_{\mathbf{s}}, ft}(\tilde{\mathbf{s}}_{1:t-1}, \tilde{\mathbf{z}}_{1:t}) + (\mathbf{W}_b \mathbf{H}_b)_{ft}. \quad (36)$$

We remind that $\tilde{\mathbf{s}}_{1:t-1}$ is computed recursively from the output of the decoder network, as explained after equation (27), and $\tilde{\mathbf{z}}_{1:t}$ is recursively sampled from $q_{\phi'}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T}) = \prod_{\tau=1}^t q_{\phi'}(\mathbf{z}_{\tau}|\mathbf{z}_{1:\tau-1}, \mathbf{x}_{1:T})$, as defined in (34). During the variational E-step, the parameters ϕ' are updated with a gradient ascent technique, and we denote by ϕ'^* the resulting parameters that will be fixed in the M-step.

We recall that the recursive computation of $\tilde{\mathbf{s}}_t$ is required only for SRNN (actually for the DVAE models with an autoregressive form). DKF and RVAE, as non-autoregressive models, do not require estimating these quantities. They only require the sampling of the latent variables $\tilde{\mathbf{z}}_{1:t}$.

2) *M-step*: The M-step consists in maximizing $\mathcal{L}(\phi'^*, \varphi)$ w.r.t φ under a non-negativity constraint. Replacing the intractable expectation in (35) with a Monte Carlo estimate (using one single sample), the M-step can be recast as minimizing the following criterion [33]:

$$\mathcal{C}(\varphi) = \sum_{f=1}^F \sum_{t=1}^T d_{\text{IS}}(|x_{ft}|^2, v_{\varphi, ft}), \quad (37)$$

where $v_{\varphi, ft}$ is defined in (36). This optimization problem can be tackled using a majorize-minimize approach [61], which leads to the multiplicative update rules derived in [33] using the methodology proposed in [62]:

$$\mathbf{H}_b \leftarrow \mathbf{H}_b \odot \left[\frac{\mathbf{W}_b^{\top} (|\mathbf{X}|^{\odot 2} \odot (\mathbf{V}_{\mathbf{x}})^{\odot -2})}{\mathbf{W}_b^{\top} (\mathbf{V}_{\mathbf{x}})^{\odot -1}} \right]^{\odot 1/2}, \quad (38)$$

$$\mathbf{W}_b \leftarrow \mathbf{W}_b \odot \left[\frac{(|\mathbf{X}|^{\odot 2} \odot (\mathbf{V}_{\mathbf{x}})^{\odot -2}) \mathbf{H}_b^{\top}}{(\mathbf{V}_{\mathbf{x}})^{\odot -1} \mathbf{H}_b^{\top}} \right]^{\odot 1/2}, \quad (39)$$

$$\mathbf{g}^{\top} \leftarrow \mathbf{g}^{\top} \odot \left[\frac{\mathbf{1}^{\top} [|\mathbf{X}|^{\odot 2} \odot (\mathbf{V}_{\mathbf{s}} \odot (\mathbf{V}_{\mathbf{x}})^{\odot -2})]}{\mathbf{1}^{\top} [(\mathbf{V}_{\mathbf{s}} \odot (\mathbf{V}_{\mathbf{x}})^{\odot -1})]} \right]^{\odot 1/2}, \quad (40)$$

where \odot denotes element-wise multiplication and exponentiation, and matrix division is also element-wise, $\mathbf{V}_{\mathbf{s}}, \mathbf{V}_{\mathbf{x}} \in \mathbb{R}_+^{F \times T}$ are the matrices of entries $v_{\theta_{\mathbf{s}}, ft}$ and $v_{\varphi, ft}$ respectively, $\mathbf{X} \in \mathbb{C}^{F \times T}$ is the matrix of entries x_{ft} and $\mathbf{1}$ is an all-ones column vector of dimension F . Note that non-negativity is ensured provided that these parameters are initialized with non-negative values.

D. Summary

In summary, the clean speech signal estimation consists in approximating the posterior $p_{\theta}(\mathbf{s}_t|\mathbf{x}_{1:T})$ and taking the mean of the resulting approximate distribution (i.e., the Wiener filter output). The estimation of the involved parameters is made with the VEM algorithm, which consists in iteratively fine-tuning the inference network of the pre-trained DVAE (E-step) and updating the mixture model parameters φ (M-step). The complete proposed speech enhancement method is summarized in Algorithm 1. For non-causal DVAEs, we can simply replace $\mathbf{z}_{1:t}$ with $\mathbf{z}_{1:T}$ when generating \mathbf{s}_t .

IV. EXPERIMENTS

A. Datasets

We use the WSJ0-QUT dataset and the VoiceBank-DEMAND (VB-DMD) dataset, described below. Each dataset has a “clean” version used to pre-train the DVAE models and a “noisy” version used to test the proposed speech enhancement algorithm and the reference methods. The clean and noisy versions are actually used together to compute the speech enhancement objective performance measures (see Section IV-B) and for the training of the supervised reference methods. When using only the clean version, we refer to it as WSJ0 or VB.

Algorithm 1 DVAE-based unsupervised speech enhancement

Inputs:

- ▷ Pre-trained DVAE model: $p_{\theta}(\mathbf{z}_{1:T}, \mathbf{s}_{1:T})$ and $q_{\phi}(\mathbf{z}_{1:T}|\mathbf{s}_{1:T})$
- ▷ Noisy speech STFT $\mathbf{x}_{1:T}$

Initialization:

- ▷ Initialize NMF noise parameters \mathbf{H}_b and \mathbf{W}_b with random nonnegative values
- ▷ Initialize gain parameters $\mathbf{g} = \mathbf{1}$
- ▷ Initialize $q_{\phi'}(\mathbf{z}_{1:t}|\mathbf{x}_{1:T})$ with pre-trained inference network $q_{\phi}(\mathbf{z}_{1:t}|\mathbf{s}_{1:T})$

while stopping criterion not reached **do**

E-step:

- ▷ Fine-tune $q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ by maximizing (35) w.r.t. ϕ'
- ▷ Sample $\tilde{\mathbf{z}}_{1:T}$ from $q_{\phi'}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$
- ▷ Compute $\Sigma_{\theta_{s,t}}$ for $t = 1$ to T using the DVAE decoder

M-step:

- ▷ Update \mathbf{H}_b , \mathbf{W}_b and \mathbf{g} using (38)-(40)

end while

Output:

- ▷ Compute the clean speech signal estimate $\hat{\mathbf{s}}_t$ for $t = 1$ to T using (32)
-

1) *WSJ0-QUT*: We used the Wall Street Journal dataset (WSJ0) [63], which is composed of 16kHz clean speech signals (read Wall Street Journal news). WSJ0-QUT is the noisy version already presented and used in [42]. It was obtained by mixing clean speech signals from WSJ0 with various types of noise signals from the QUT-NOISE dataset [64], with three signal-to-noise ratio (SNR) values: -5 , 0 , and 5 dB. The full description of the dataset, including training/test splits and noise types, can be found in [42]. Note that we mixed the speech and noise signals using the ITU-R BS.1770-4 protocol [65]. An SNR computed with this protocol is 2.5 dB lower (in average) than that computed with sums of squared signal samples.

2) *VB-DMD*: We also used the publicly available VB-DMD dataset [66]. This dataset contains a training set with 11,572 utterances performed by 28 speakers and a test set with 824 utterances performed by 2 speakers, different from the training set. The noisy train set consists of mixture signals mixed at four different SNRs, namely 15, 10, 5, and 0 dB, whereas the noisy test speech signals are corrupted with 17.5, 12.5, 7.5, and 2.5 dB SNR and different noise types. The full description of the dataset, including training/test splits and noise types, can be found in [66]. Following [22], we also selected two speakers (p226 and p287) from the clean training set as the validation set for the training of the DVAEs.

3) *Data preprocessing*: In all our experiments, the STFT was computed with a 64-ms sine window (1,024 samples) and a 75%-overlap (256 samples hop length), resulting in a sequence of 513-dimensional discrete Fourier coefficients (for positive frequencies). The DVAEs were trained with STFT power spectrograms of clean speech signals extracted from either WSJ0 or VB, and obtained with the following

preprocessing. We first removed the silence at the beginning and ending of the files, using an energy-based voice activity detection threshold of -30 dB. The waveform signal was then normalized by its maximum absolute value. We set $T = 50$, meaning that speech segments of 0.8s were used to train the DVAE models. In summary, each training data sequence is a 513×50 STFT power spectrogram. For WSJ0, this data preprocessing resulted in a set of $N_{\text{tr}} = 93,393$ training sequences (representing about 20.8 hours of speech signal) and $N_{\text{val}} = 7,775$ validation sequences (about 1.7 hours). For VB, we obtained $N_{\text{tr}} = 29,389$ training sequences (6.5 hours) and $N_{\text{val}} = 2,152$ validation sequences (0.5 hour). For the evaluation of the speech enhancement methods, we used the STFT spectrogram of each complete noisy test sequence (with normalization), which can be of variable length, most often larger than 2s. For WSJ0-QUT, the total duration of the test dataset is 1.5 hours, and for VB-DMD it is 0.6 hour.

B. Evaluation metrics

We used three metrics to evaluate the quality of the estimated speech signals: The scale-invariant signal-to-distortion ratio (SI-SDR) in dB [67], the perceptual evaluation of speech quality (PESQ) score [24], and the extended short-time objective intelligibility (ESTOI) score (in $[0, 1]$) [68]. The PESQ measure is declined in three different variants, depending on different protocols:⁴ the narrow-band PESQ MOS value (PESQ MOS, in $[-0.5, 4.5]$), the narrow-band PESQ LQ0 value (PESQ NB, in $[1, 5]$), and the wide-band PESQ LQ0 value (PESQ WB, in $[1, 5]$). We report all of them in the following experiments. For all measures, a higher value indicates a better result.

C. Models implementation

Here we present the implementation of the three example DVAEs that we used in practice in the proposed DVAE-based speech enhancement algorithm, namely DKF, RVAE, and SRNN. As indicated in [54], we can have various implementations for each DVAE model, thus we only present the model configurations that showed the best performance in our experiments (for the latent space dimension selected below).

1) *Dimension of the latent space*: In the present experiments, we set $L = 16$. We recall that the data dimension is $F = 513$. We also recall that \mathbf{z}_t is a real-valued vector that is modeled by a Gaussian distribution, so the DNNs modeling \mathbf{z}_t has to output two L -dimensional vectors, the mean and variance vectors, for both inference and generation (except for RVAE since \mathbf{z}_t is assumed i.i.d. with a standard Gaussian distribution and no DNN is used for its generation). In contrast, \mathbf{s}_t is a complex-valued vector modeled by a circular complex Gaussian distribution, which only leaves one F -dimensional variance vector to be provided by the decoder DNN. To guarantee the positivity of the entries of this output variance vector, we used log-parameterization (the output is the log-variance in \mathbb{R} , which is then converted to variance by taking

⁴See the explanation for ‘pesq’ at <https://github.com/ludlows/python-pesq> and for ‘pypesq’ and <https://github.com/vBaiCai/python-pesq>.

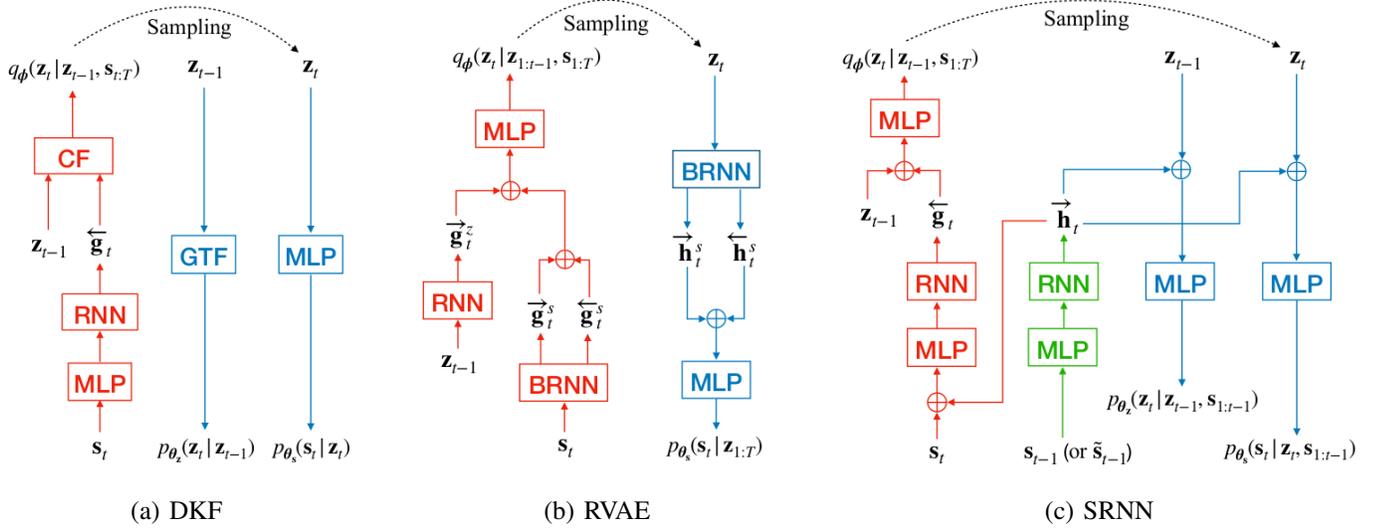


Fig. 3. Model implementation for the (a) deep Kalman filter (DKF) [46], [47], (b) recurrent variational autoencoder (RVAE) [42], and (c) stochastic recurrent neural network (SRNN) [49]. Each model consists of an inference (encoder) network (in red) and a generation (decoder) network (in blue). SRNN has a shared module between encoder and decoder (in green). CF: combiner function, GTF: gated transition function (see [46] for details), RNN: recurrent neural network, BRNN: bidirectional RNN, MLP: multi-layer perceptron, \oplus : vector concatenation. All RNNs are implemented with LSTM networks.

the exponential). The last layer predicting the mean and log-variance parameters is always a linear layer, with a dimension corresponding to that of z_t (16) or s_t (513). We omit this in the following description for simplicity.

2) *DKF*: Fig. 3(a) summarizes our implementation of DKF. The layers providing the parameters of the inference and generative models of z_t are respectively implemented with the specific *combiner function* and *gated transition function* described in [47]. For the inference model, we used a backward long short-term memory (LSTM) [69] layer with an internal state of dimension 128 to accumulate the information from $s_{t:T}$ in (11). Before being fed into the recurrent layer, each vector s_t passes through a multi-layer perceptron (MLP) with one hidden layer of dimension 256 and a tanh activation. The variance parameters of the generative model of s_t are provided by an MLP with 4 hidden layers of dimension 32, 64, 128 and 256, with a tanh activation function.

3) *RVAE*: We implemented the non-causal version of RVAE as schematized in Fig. 3(b). The inference model includes a bidirectional LSTM (BLSTM) layer with an internal state of dimension 128 to process the complete sequence $s_{1:T}$ and an LSTM layer to process the sampled past latent vectors sequence $z_{1:t-1}$. The output of these two layers are then concatenated and mapped into the parameters of the inference model over z_t by an MLP. The generative part of the model includes a BLSTM layer with an internal state of dimension 128, which takes the sampled $z_{1:T}$ as input. The output of this BLSTM layer is finally mapped to the parameters of the generative model over s_t by a single linear layer.

4) *SRNN*: SRNN is quite different from the two previous models. As shown in Fig. 3(c), the inference and generative models share a recurrent internal state vector \vec{h}_t (module in green) that is encoding the information from the past observed vectors $s_{1:t-1}$. This shared module is composed of an MLP with one layer of dimension 256 followed by a forward LSTM.

The dimension of \vec{h}_t is 128. For inference, the concatenation of \vec{h}_t and s_t is fed into a one-layer MLP of dimension 256 followed by a backward LSTM that provides the vector \overleftarrow{g}_t (of dimension 128). This vector is then concatenated with the sample of z_{t-1} and fed into an MLP with two hidden layers of dimension 64 and 32. For the generative part, we concatenate the shared state \vec{h}_t with the sampled latent vector at the previous or current time frame. An MLP with two hidden layers of dimension 64 and 32 is used for the generation of z_t , and an MLP with one hidden layer of dimension 128 is used for the generation of s_t . All MLP hidden layers use the tanh activation function.

D. DVAEs pre-training

For the pre-training of the three DVAE models on clean speech signals, we used the Adam optimizer [70] with a learning rate of $1e-3$ and $\beta_1 = 0.9$, $\beta_2 = 0.99$. On both datasets, we trained each model with a batch size of 128 during 300 epochs and kept the model snapshot with lowest validation loss. We applied a linear KL annealing for the first 20 epochs to warm-up the latent space [71].

As an autoregressive model, SRNN deserves a particular treatment during pre-training. Indeed, in the conventional training of autoregressive models, the ground-truth past clean speech vectors $s_{1:t-1}$ are used to generate the current one s_t , a strategy sometimes referred to as “teacher forcing” in the literature [72]. We have seen in Section III-B2 that it is not possible to do that in the proposed speech enhancement algorithm, where $s_{1:t-1}$ is replaced by its proxy $\tilde{s}_{1:t-1}$ (recursively computed from the decoder output). It is shown in [54] that directly using $\tilde{s}_{1:t-1}$ in the SRNN model trained with $s_{1:t-1}$ significantly decreases the quality of the reconstructed speech spectrogram, due to the mismatch between train and test conditions. To avoid such a mismatch (here between DVAE training and speech enhancement conditions), we trained SRNN using

TABLE I
RESULTS OF THE SPEECH ANALYSIS-RESYNTHESIS EXPERIMENT,
AVERAGED OVER THE TEST SUBSET OF WSJ0 AND VOICEBANK.

Models	Dataset	SI-SDR (dB)	PESQ MOS	PESQ WB	PESQ NB	ESTOI
VAE	WSJ0	8.0	3.33	2.95	3.31	0.89
DKF	WSJ0	9.0	3.55	3.39	3.61	0.91
RVAE	WSJ0	9.8	3.65	3.57	3.75	0.92
SRNN	WSJ0	8.2	3.48	3.24	3.52	0.90
VAE	VB	8.6	3.22	2.79	3.15	0.88
DKF	VB	9.4	3.35	2.96	3.34	0.90
RVAE	VB	9.6	3.41	3.00	3.42	0.90
SRNN	VB	9.1	3.39	2.99	3.39	0.89

$\tilde{s}_{1:t-1}$ (instead of $s_{1:t-1}$) to generate s_t . Such a training is difficult in practice and to make it efficient, we adopted a “scheduled sampling” approach [73], i.e., we progressively replace $s_{1:t-1}$ with $\tilde{s}_{1:t-1}$, with a proportion going from 0% to 100% along the training iterations. That is, at the beginning of the training, s_t is generated completely from $s_{1:t-1}$ and z_t . Then the probability to use $\tilde{s}_{1:t-1}$ increases during the training procedure. Finally, s_t is generated completely from $\tilde{s}_{1:t-1}$ and z_t . This takes 80 epochs after the KL annealing step.

Before we examine the speech enhancement performance, we can rapidly compare the speech modeling capacities of the three selected DVAE models (and the vanilla VAE) after their pre-training, by conducting a speech analysis-resynthesis experiment (i.e., chaining of the encoder and decoder) similar to [74]. The overall pipeline is shown in Fig. 1. For the VAE model, we used the baseline architecture already used in [42]. The results presented in Table I were obtained with the models being trained on the WSJ0 or VB train subsets and averaged over the corresponding test subsets. For SI-SDR scores, the noise is the modeling noise, i.e., the difference between original and reconstructed signal. We can see from Table I that all DVAE models outperform the VAE for all metrics, both on WSJ0 and VB, showing the benefits of introducing dynamics into VAE-based speech modeling.

E. DVAE-VEM algorithm settings

The rank of the NMF in the noise model (16) is set to $K = 8$. \mathbf{W}_b and \mathbf{H}_b were randomly initialized from a uniform distribution in $[0,1]$ and \mathbf{g} was initialized with an all-ones vector. In the E-step of the VEM algorithm, the encoder of the DVAE models is fine-tuned using the Adam optimizer [70] with a learning rate of $1e-3$. Fig. 4 illustrates the influence of the number of VEM iterations on the performance of the different DVAE models on the two test datasets. We observe that the performance of most models plateaus from 300 and 100 iterations for WSJ0-QUT and VB-DMD respectively. We fixed the number of iterations to these values as they represent a global optimal trade-off between performance and complexity and are neither beneficial nor disadvantageous for a particular DVAE model.

F. Baselines

Regarding baseline supervised methods, we used Open-Unmix (UMX) and MetricGAN+. UMX is an open-source

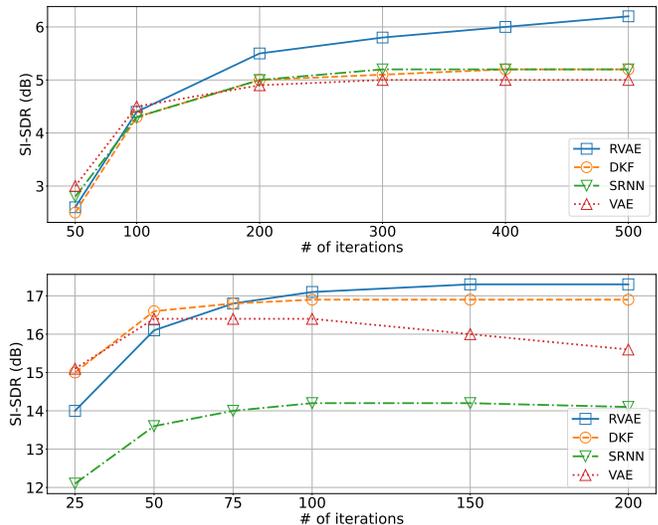


Fig. 4. Performance of the proposed DVAE-based speech enhancement method (and VAE baseline) as a function of the number of VEM iterations, for the WSJ0-QUT (top) and VB-DMD (bottom) test datasets. All DVAE models are pre-trained on the corresponding dataset of clean speech signals (WSJ0 and VB, respectively).

method based on a BLSTM network. It was originally proposed for music source separation [76] and was later adapted for speech enhancement [75]. MetricGAN+ [10] also adopts BLSTMs for mask-based prediction of the clean speech. In addition, it introduces a metric network that is trained to approximate the PESQ evaluation score, and then MetricGAN+ is trained to maximise this proxy of the PESQ score.

Regarding baseline unsupervised methods, we choose VAE-VEM [42], MetricGAN-U [22] and NyTT [21]. VAE-VEM, referred to as VAE-FFNN in [42], uses the same optimization methodology as our approach, except there is no temporal model. In practice, we tried with different model complexities for the VAE, and we report the one exhibiting the best performance. MetricGAN-U is the unsupervised version of MetricGAN+. Since the supervision in MetricGAN+ comes from the PESQ score computation using the paired enhanced and clean speech signal, MetricGAN-U adopts a non-intrusive speech quality metric instead, namely the DNSMOS measure [23], to bypass the paired supervision. Two versions of MetricGAN-U were proposed by the authors: the “full” one is trained entirely without supervision, while the “half” version monitors the PESQ measure to perform early-stopping. Since PESQ is an intrusive measure, this version of MetricGAN-U can be seen as weakly supervised (supervision is used only for validation and not for training *per se*) [22]. NyTT is based on a noisy speech target training strategy, where the network is trained to remove an additional noise added to the noisy speech. Since there is no need for noisy/clean speech pairs, NyTT can also be considered as an unsupervised speech enhancement method. It should be noted that both MetricGAN-U and NyTT are trained using noisy speech as input, so the resulting model is noise-dependent. This contrasts with the proposed method where only clean speech is used for the DVAE pre-training, resulting in a noise-agnostic speech enhancement method. Finally, we could not re-train MetricGAN-U or NyTT. Indeed,

TABLE II

SPEECH ENHANCEMENT RESULTS OBTAINED WITH MODELS TRAINED AND TESTED ON CORRESPONDING DATASETS. * INDICATES THE IMPLEMENTATION PROVIDED BY THE AUTHORS AND RE-TRAINED. THE ‘‘SUPERVISION’’ COLUMN INDICATES WHETHER THE TRAINING IS SUPERVISED (S), UNSUPERVISED NOISE-DEPENDENT (UD) OR NOISE-AGNOSTIC (UA).

Method	Supervision	Parameters	Train subset	Test subset	SI-SDR (dB)	PESQ MOS	PESQ WB	PESQ NB	ESTOI
Noisy mixture	-	-	-	WSJ0-QUT	-2.6	1.83	1.14	1.57	0.50
VAE-VEM [42]	UA	0.14M	WSJ0	WSJ0-QUT	5.0	2.13	1.45	1.86	0.58
Proposed DKF-VEM	UA	0.52M	WSJ0	WSJ0-QUT	5.1	2.23	1.46	1.95	0.62
Proposed RVAE-VEM	UA	1.06M	WSJ0	WSJ0-QUT	5.8	2.27	1.54	1.98	0.62
Proposed SRNN-VEM	UA	0.88M	WSJ0	WSJ0-QUT	5.2	2.23	1.48	1.95	0.63
UMX* [75]	S	1.55M	WSJ0-QUT	WSJ0-QUT	5.7	2.16	1.38	1.83	0.63
MetricGAN+* [10]	S	1.90M	WSJ0-QUT	WSJ0-QUT	3.6	2.83	2.18	2.61	0.60
Noisy mixture	-	-	-	VB-DMD	8.4	3.02	1.97	2.88	0.79
NyTT [21]	UD	-	VB-DMD + Extra noise	VB-DMD	17.7	-	2.30	-	-
NyTT [21]	UD	-	VB-DMD	VB-DMD	12.1	-	1.74	-	-
MetricGAN-U (full) [22]	UD	1.90M	VB-DMD	VB-DMD	6.5	3.13	2.13	3.03	0.74
MetricGAN-U (half) [22]	UD	1.90M	VB-DMD	VB-DMD	8.2	3.20	2.45	3.11	0.77
VAE-VEM [42]	UA	0.14M	VB	VB-DMD	16.4	3.18	2.37	3.10	0.80
Proposed DKF-VEM	UA	0.52M	VB	VB-DMD	16.9	3.22	2.42	3.14	0.81
Proposed RVAE-VEM	UA	1.06M	VB	VB-DMD	17.1	3.23	2.48	3.15	0.81
Proposed SRNN-VEM	UA	0.88M	VB	VB-DMD	14.2	3.20	2.32	3.12	0.80
UMX [75]	S	1.55M	VB-DMD	VB-DMD	14.0	3.18	2.35	3.08	0.83
MetricGAN+ [10]	S	1.90M	VB-DMD	VB-DMD	8.5	3.59	3.13	3.63	0.83

since MetricGAN-U uses the DNSMOS service, each training epoch needs 2 days for evaluation, which is impractical. Regarding NyTT, we cannot retrain it since there is no public release of the code.

G. Speech enhancement results

The speech enhancement scores obtained with the proposed method (for the three DVAE models) and with the baseline methods are reported in Tables II and III, along with the capacity of the models, reported in terms of number of parameters. Table II shows the results obtained when the test subset ‘‘corresponds’’ to the train subset, i.e., it originates from the same dataset (WSJ0 or WSJ0-QUT in the upper half of the table, and VB or VB-DMD in the lower half). Table III shows the results obtained with cross-dataset experiments conducted to evaluate the generalization capability of the different models. This means that WSJ0 or WSJ0-QUT is used for training and VB-DMD is used for testing, or alternatively, VB or VB-DMD is used for training and WSJ0-QUT is used for testing. Baseline models marked with * in Table II were retrained using the implementation provided by the authors. Other baseline results are obtained from the corresponding papers or from the pre-trained models if available. It can be seen that the number of parameters for the three DVAE models is lower than that of all the baselines but the VAE-VEM method. Among the DVAE models, RVAE is the one with the highest number of parameters (1.06M for RVAE, 0.88M for SRNN and 0.52M for DKF).

From the results in Table II, we first observe that the proposed DVAE-VEM algorithm outperforms the VAE-based counterpart for all three DVAE models, except SRNN-VEM on the VB dataset. This is consistent with the results of the analysis-resynthesis experiment and this shows the interest of modeling the speech signal dynamics within the proposed

speech enhancement method. Among the three tested DVAE models, RVAE performs the best for all evaluation metrics, except in terms of ESTOI on the WSJ0-QUT dataset, where SRNN obtains a slightly better score.

When comparing with the baseline unsupervised methods, hence only on the VB-DMD dataset, the proposed method achieves competitive results. Even if the highest SI-SDR performance is achieved by NyTT (17.7 dB), this is possibly due to the use of large amounts of noise, given that its performance drops significantly (12.1 dB) when only the noise from the DMD dataset is used. The proposed RVAE-VEM algorithm reaches very similar performance (17.1 dB) without training on any kind of noise. In terms of PESQ WB, the proposed approaches outperform NyTT independently of the amount of noise used at training time. Similar conclusions are drawn when comparing to MetricGAN-U, in which case the performance difference in terms of SI-SDR is very large (+8 to +10 dB). Remarkably, the proposed method achieves competitive performance with, and sometimes outperforms, MetricGAN-U (half) in terms of PESQ WB score, even though MetricGAN-U (half) uses the PESQ WB score on the validation set as a training stop criterion. In this regard, MetricGAN-U (half) is expected to exhibit higher PESQ WB values than MetricGAN-U (full).

When comparing with the supervised methods, we see that MetricGAN+ obtains PESQ scores that are significantly higher (e.g., 3.59 and 2.83 PESQ NB on VB-DMD and WSJ0-QUT respectively) than those of all the other methods (RVAE-VEM reaches a maximum of 3.23 and 2.27 on VB-DMD and WSJ0-QUT). This can be explained by the fact that PESQ is the criterion optimized during the MetricGAN+ model training. In contrast, when it is evaluated with other metrics, the results obtained by this method are considerably worse, especially in terms SI-SDR for which MetricGAN+ performs the worst. We found that this may be due to the fact that the energy in the

TABLE III

SPEECH ENHANCEMENT RESULTS WITH MODELS TRAINED AND TESTED ON DIFFERENT DATASETS. THE ‘‘SUPERVISION’’ COLUMN INDICATES WHETHER THE TRAINING IS SUPERVISED (S), UNSUPERVISED NOISE-DEPENDENT (UD) OR NOISE-AGNOSTIC (UA).

Method	Supervision	Parameters	Train subset	Test subset	SI-SDR (dB)	PESQ MOS	PESQ WB	PESQ NB	ESTOI
Noisy mixture	-	-	-	WSJ0-QUT	-2.6	1.83	1.14	1.57	0.50
MetricGAN-U (full) [22]	UD	1.90M	VB-DMD	WSJ0-QUT	-2.3	1.91	1.18	1.63	0.50
MetricGAN-U (half) [22]	UD	1.90M	VB-DMD	WSJ0-QUT	-1.6	2.01	1.25	1.71	0.49
VAE-VEM [42]	UA	0.14M	VB	WSJ0-QUT	3.8	1.89	1.31	1.68	0.54
Proposed DKF-VEM	UA	0.52M	VB	WSJ0-QUT	3.5	2.08	1.32	1.80	0.57
Proposed RVAE-VEM	UA	1.06M	VB	WSJ0-QUT	4.3	2.12	1.37	1.84	0.57
Proposed SRNN-VEM	UA	0.88M	VB	WSJ0-QUT	4.6	2.21	1.42	1.91	0.61
UMX [75]	S	1.55M	VB-DMD	WSJ0-QUT	4.1	2.06	1.34	1.76	0.61
MetricGAN+ [10]	S	1.90M	VB-DMD	WSJ0-QUT	1.8	2.31	1.61	2.02	0.56
Noisy mixture	-	-	-	VB-DMD	8.4	3.02	1.97	2.88	0.79
VAE-VEM [42]	UA	0.14M	WSJ0	VB-DMD	15.0	3.16	2.27	3.06	0.79
Proposed DKF-VEM	UA	0.52M	WSJ0	VB-DMD	16.8	3.17	2.34	3.08	0.81
Proposed RVAE-VEM	UA	1.06M	WSJ0	VB-DMD	17.3	3.21	2.41	3.13	0.81
Proposed SRNN-VEM	UA	0.88M	WSJ0	VB-DMD	16.8	3.17	2.34	3.08	0.81
UMX* [75]	S	1.55M	WSJ0-QUT	VB-DMD	10.4	3.10	2.21	2.98	0.78
MetricGAN+* [10]	S	1.90M	WSJ0-QUT	VB-DMD	3.9	3.41	2.51	3.39	0.73

speech signal estimated by MetricGAN+ is mostly concentrated in the low-frequency part of the spectrum, whereas the mid- and high-frequency parts are poorly recovered. Regarding UMX, we observe that its performance is systematically under the one of the proposed method, in terms of SI-SDR or any of the PESQ measures for both datasets.

To evaluate the generalization capability of the different models, we report in Table III their performance obtained when the training set and test set originate from different corpora. It is unsurprising that the performance of supervised methods significantly decreases in this setting. For example, MetricGAN+ obtains a PESQ WB value of 3.13 on the VB-DMD test dataset when trained on the VB-DMD train dataset, whereas it drops to 2.51 when trained on the WSJ0-QUT. In the same vein, UMX goes from 14.0 dB SI-SDR to 10.4 dB SI-SDR. Similar behavior is found for the other metrics and datasets. As for the unsupervised noise-dependent methods, we can also see a significant decrease of performance. For example, MetricGAN-U (trained on VB-DMD and provided by the authors) obtains a negative SI-SDR when tested on WSJ0-QUT. In contrast, the proposed DVAE-VEM algorithm is much less affected by the dataset mismatch (for all DKF, RVAE, and SRNN models). Even if, generally speaking, we observe a mild decrease in performance when the training and test sets do not match, the proposed method occasionally exhibits better performance when tested in a different dataset. For example, the RVAE-VEM algorithm tested on VB-DMD with RVAE trained on WSJ0 obtains 17.3 dB of SI-SDR, which is even better than when RVAE is trained on VB (17.1 dB). This is probably because the WSJ0 dataset is larger than the VB dataset. RVAE-VEM performs the best when trained on WSJ0 and tested on VB-DMD, while SRNN-VEM performs the best when trained on VB and tested on WSJ0-QUT. Overall, we can conclude that the proposed DVAE-based speech enhancement method exhibits competitive performance when compared to other supervised and unsupervised methods in the ‘‘corresponding dataset’’ setting, and superior performance in the cross-dataset setting.

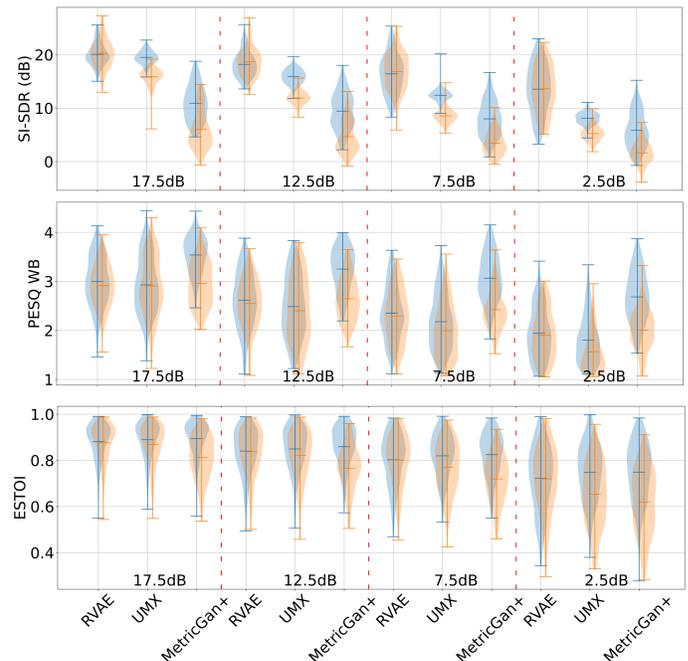


Fig. 5. Detailed performance of RVAE-VEM, UMX and MetricGAN+ on the VB-DMD dataset for different input SNRs. The blue and orange violin plots correspond to a training of the models on the VB-DMD and the WSJ0-QUT datasets, respectively.

Complementary to the above experimental analysis, we present in Fig. 5 violin plots showing the full distribution of the results obtained with RVAE-VEM, UMX and MetricGAN+, when evaluated on the VB-DMD dataset and trained either on the corresponding training set or on the WSJ0-QUT dataset. The results are presented separately for different test SNRs. As expected, we observe that the performance of all methods degrades as the SNR decreases. We can also see that the proposed method (RVAE) is much less affected by the mismatch between training and test sets, compared with

supervised methods (UMX and MetricGAN+) for which the distributions of the results clearly shift down.

At the light of the presented results, we extract the following concluding remarks. When there is no mismatch between the training and test datasets, the proposed method achieves state-of-the-art performance in unsupervised speech enhancement, and competitive results when compared with supervised speech enhancement methods (often outperforming them depending on the metric). As for cross-dataset experiments, where the training and test sets are coming from different datasets, we observe that the performance of the supervised methods is severely affected by the dataset mismatch, whereas the performance of the proposed unsupervised method is very robust to it. Overall, the results obtained in the various settings demonstrate the interest of the proposed DVAE-VEM methodology for speech enhancement. Audio examples and code are available at <https://team.inria.fr/robotlearn/unsupervised-speech-enhancement-using-dynamical-variational-auto-encoders>.

V. CONCLUSION

In this paper, we have proposed a general framework for unsupervised speech enhancement based on DVAEs. In our framework, the DVAEs are used to model the clean speech signal, while the noise is modeled via NMF. While DVAEs are pre-trained with a clean speech dataset, the noise parameters are estimated at test time, together with the clean speech, from the noisy speech sequence to process. To achieve that, we have derived a VEM algorithm for the most general formulation of a DVAE model, which can then be easily adapted to particular instances of DVAEs. We have illustrated this principle with DKF, RVAE and SRNN, and this can be extended to other DVAE models, e.g., STORN [51] or VRNN [48].

We have evaluated the speech enhancement performance obtained with those three example DVAEs. The proposed approach exhibits superior or competitive performance compared to supervised and unsupervised state-of-the-art methods when the training and test datasets are from the same corpora, and outperforms them on cross-dataset settings, i.e., when the training and test datasets are from different corpora. The RVAE model provided the best performance among the tested DVAEs. SRNN shows a great potential, provided that it is trained with scheduled sampling in order to reduce the gap between the training and speech enhancement conditions. If this gap could be further decreased, we believe that it could have even better performance than RVAE. This aspect should be further investigated, possibly including other autoregressive models in the DVAE family (e.g., VRNN [48]).

So far, the good performance of the proposed iterative DVAE-VEM algorithm comes at the cost of a high computational time. Indeed, processing one second of audio with 100 iterations of the algorithm takes approximately 14, 25 and 21 seconds for DKF, RVAE and SRNN, respectively, using a single core of an Intel Xeon Gold 6230 at 2.1GHz. Future work will include developing fast DVAE-based speech enhancement algorithms, for instance inspiring from [35]. Also, so far, the inference with DVAEs is non-causal, meaning that past, present, and future noisy speech observations are

required to enhance a given speech frame. Causal DVAE-based speech enhancement can also be investigated, but is out of the scope of this paper. Future work also includes using other powerful encoder-decoder networks, e.g., TCNs [44] and the Transformer [77], in the present unsupervised speech enhancement framework. The DVAE models may be further boosted with more expressive latent variables, e.g., introducing hierarchical multi-scale structure and normalizing flows [78]. We also plan to extend the proposed method to a multi-modal framework, using the speaker's lips motion and visual appearance, in the continuation of VAE-based audio-visual speech enhancement [79].

REFERENCES

- [1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006. I
- [2] P. C. Loizou, *Speech enhancement: Theory and practice*. CRC press, 2013. I
- [3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979. I
- [4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979. I
- [5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984. I, II-B
- [6] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018. I
- [7] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Asia-Pacific Signal Inform. Process. Assoc. Annual Conf. (APSIPA)*, Kuala Lumpur, Malaysia, 2017. I
- [8] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech Conf.*, Stockholm, Sweden, 2017. I
- [9] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, 2019. I
- [10] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech Conf.*, Brno, Czech Republik, 2021. I, IV-F, II, III
- [11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech Conf.*, Lyon, France, 2013. I
- [12] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1826–1838, 2020. I
- [13] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "CycleGAN-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism," in *Asia-Pacific Signal Inform. Process. Assoc. Annual Conf. (APSIPA)*, 2021. I
- [14] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018. I
- [15] N. Alamdari, A. Azarang, and N. Kehtarnavaz, "Improving deep speech denoising by noisy2noisy signal mapping," *Applied Acoustics*, vol. 172, p. 107631, 2021. I
- [16] M. M. Kashyap, A. Tambwekar, K. Manohara, and S. Natarajan, "Speech denoising without clean training data: a noise2noise approach," in *Proc. Interspeech Conf.*, Brno, Czech Republik, 2021. I
- [17] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech, Signal Process. Magazine*, vol. 5, no. 2, pp. 4–24, 1988. I
- [18] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer Science & Business Media, 2008. I

- [19] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, 2017. I
- [20] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. John Wiley & Sons, 2018. I
- [21] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, "Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech," in *Proc. Europ. Signal Process. Conf. (EUSIPCO)*, Dublin, Ireland (virtual conference), 2021. I, IV-F, II
- [22] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "MetricGAN-U: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2022. I, IV-A2, IV-F, IV-F, II, III
- [23] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, Canada, 2021. I, IV-F
- [24] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Salt Lake City, USA, 2001. I, IV-B
- [25] 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*, Charleston, USA, 2007. I
- [26] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, 2011. I
- [27] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013. I
- [28] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Montreal, Canada, 2014. I
- [29] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, Canada, 2018. I
- [30] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, Banff, Canada, 2014. I, II-A, II-A
- [31] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, 2014. I, II-A, II-A
- [32] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comp.*, vol. 21, no. 3, pp. 793–830, 2009. I, II-B, III, III-A
- [33] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Aalborg, Denmark, 2018. I, III-A, III-A, III-C2, III-C2
- [34] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brighton, UK, 2019. I
- [35] M. Pariente, A. Deleforge, and E. Vincent, "A statistically principled and computationally efficient approach to speech enhancement using variational autoencoders," in *Proc. Interspeech Conf.*, Graz, Austria, 2019. I, V
- [36] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multi-channel speech enhancement with a deep speech prior," in *Asia-Pacific Signal Inform. Process. Assoc. Annual Conf. (APSIPA)*, Honolulu, USA, 2018. I
- [37] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brighton, UK, 2019. I
- [38] M. Fontaine, A. A. Nugraha, R. Badeau, K. Yoshii, and A. Liutkus, "Cauchy multichannel speech enhancement with a deep speech prior," in *Proc. Europ. Signal Process. Conf. (EUSIPCO)*, A Coruna, Spain, 2019. I
- [39] G. Carbajal, J. Richter, and T. Gerkmann, "Guided variational autoencoder for speech enhancement with a supervised classifier," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, Canada, 2021. I
- [40] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *Proc. Interspeech Conf.*, Shanghai, China, 2020, pp. 2437–2441. I
- [41] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Toronto, Canada, 2021, pp. 676–680. I
- [42] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Barcelona, Spain, 2020. I, II-C2, II-C2, III, III-A, III-B3, III-C1, IV-A1, IV-A1, 3, IV-D, IV-F, IV-F, II, III
- [43] E. Aksan and O. Hilliges, "STCN: Stochastic temporal convolutional networks," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, New Orleans, USA, 2018. I
- [44] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Europ. Conf. Computer Vision (ECCV)*, Amsterdam, The Netherlands, 2016. I, V
- [45] J. Richter, G. Carbajal, and T. Gerkmann, "Speech enhancement with stochastic temporal convolutional networks," in *Proc. Interspeech Conf.*, Shanghai, China, 2020. I
- [46] R. G. Krishnan, U. Shalit, and D. Sontag, "Deep Kalman filters," *arXiv preprint arXiv:1511.05121*, 2015. I, II-C1, 3
- [47] R. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, San Francisco, USA, 2017. I, II-C1, 3, IV-C2
- [48] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Montreal, Canada, 2015. I, V
- [49] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther, "Sequential neural models with stochastic layers," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Barcelona, Spain, 2016. I, II-C3, 3
- [50] O. Fabius and J. R. van Amersfoort, "Variational recurrent autoencoders," *arXiv preprint arXiv:1412.6581*, 2014. I
- [51] J. Bayer and C. Osendorfer, "Learning stochastic recurrent networks," *arXiv preprint arXiv:1411.7610*, 2014. I, V
- [52] Y. Li and S. Mandt, "Disentangled sequential autoencoder," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Stockholm, Sweden, 2018. I
- [53] M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther, "A disentangled recognition and nonlinear dynamics model for unsupervised learning," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Long Beach, USA, 2017. I
- [54] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *Found. Trends Mach. Learn.*, vol. 15, no. 1-2, pp. 1–175, 2021. I, II-A, II-A, II-A, II-C, II-C2, IV-C, IV-D
- [55] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*. Springer, 1998, pp. 355–368. I, III, III-C
- [56] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1–2, p. 1–305, 2008. I, III
- [57] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin: Springer-Verlag, 2006. I, III
- [58] F. D. Neeser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993. II-B
- [59] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*. IGI global, 2011, pp. 162–185. II-B
- [60] P.-A. Mattei and J. Frellsen, "Refit your encoder when new data comes by," in *NeurIPS Workshop on Bayesian Deep Learning*, Montreal, Canada, 2018. III-C1
- [61] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004. III-C2
- [62] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β -divergence," *Neural Comp.*, vol. 23, no. 9, pp. 2421–2456, 2011. III-C2
- [63] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B. <https://catalog.ldc.upenn.edu/Ldc93s6b>," *Philadelphia: Linguistic Data Consortium*, 1993. IV-A1

- [64] D. Dean, A. Kanagasundaram, H. Ghaemmaghami, M. H. Rahman, and S. Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *Proc. Interspeech Conf.*, Dresden, Germany, 2015. IV-A1
- [65] ITU-R, "Recommendation BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level," *BS Series*, 2011. IV-A1
- [66] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Speech Synth. Workshop*, Sunnyvale, CA, 2016. IV-A2, IV-A2
- [67] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR: Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brighton, UK, 2019. IV-B
- [68] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011. IV-B
- [69] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comp.*, vol. 9, no. 8, pp. 1735–1780, 1997. IV-C2
- [70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Repres. (ICLR)*, San Diego, USA, 2015. IV-D, IV-E
- [71] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Barcelona, Spain, 2016. IV-D
- [72] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Comp.*, vol. 1, no. 2, pp. 270–280, 1989. IV-D
- [73] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Montreal, Canada, 2015. IV-D
- [74] X. Bie, L. Girin, S. Leglaive, T. Hueber, and X. Alameda-Pineda, "A benchmark of dynamical variational autoencoders applied to speech spectrogram modeling," in *Proc. Interspeech Conf.*, Brno, Czech Republic, 2021. IV-D
- [75] S. Uhlich and Y. Mitsufuji, "Open-Unmix for speech enhancement (UMX SE)," May 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3786908> IV-F, II, III
- [76] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix: A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019. IV-F
- [77] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Long Beach, USA, 2017. V
- [78] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Advances Neural Inform. Process. Systems (NeurIPS)*, Vancouver, Canada, 2020. V
- [79] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1788–1800, 2020. V



Simon Leglaive is an Assistant Professor (tenured) at CentraleSupélec and a researcher in the AIMAC team of the IETR laboratory, a CNRS joint research unit in Rennes, France. He received the Engineering degree from Télécom Paris (Paris, France) and the M.Sc. degree in acoustics, signal processing and computer science applied to music (ATIAM) from Sorbonne University (Paris, France) in 2014. He obtained the Ph.D. degree from Télécom Paris in the field of audio signal processing in 2017. He was then a post-doctoral researcher at Inria Grenoble Rhône-Alpes (Grenoble, France), in the Perception team. His research focuses on signal processing and machine learning for audio and speech applications. He is mainly interested in weakly-supervised approaches for problems that consist in estimating latent signals from noisy and/or incomplete observations (e.g., source separation, speech enhancement).



Xavier Alameda-Pineda is a (tenured) Research Scientist at Inria, and the Leader of the RobotLearn Team. He obtained the M.Sc. (equivalent) in Mathematics in 2008, in Telecommunications in 2009 from BarcelonaTech and in Computer Science in 2010 from Université Grenoble-Alpes (UGA). He then worked towards his Ph.D. in Mathematics and Computer Science, and obtained it 2013, from UGA. After a two-year post-doc period at the Multimodal Human Understanding Group, at University of Trento, he was appointed with his current position. Xavier is an active member of SIGMM, a senior member of IEEE and a member of ELLIS. He is the Coordinator of the H2020 Project SPRING: Socially Pertinent Robots in Gerontological Healthcare and is co-leading the "Audio-visual machine perception and interaction for companion robots" chair of the Multidisciplinary Institute of Artificial Intelligence. Xavier's research interests are at the cross-roads of machine learning, computer vision and audio processing for scene and behavior analysis and human-robot interaction.



Laurent Girin received the M.Sc. (1994) and Ph.D. (1997) degrees in signal processing from Institut National Polytechnique de Grenoble (INPG), France. In 1999 he joined Ecole Nationale Supérieure d'Electronique et de Radioélectrique de Grenoble, as an Associate Professor. Currently he is a Full Professor at Grenoble Institute of Technology (Grenoble-INP) in the Physics, Electronics, and Materials (PHELM) department, where he lectures signal processing theory and applications to audio. His research activity is carried out at GIPSA-Lab (Grenoble Laboratory of Image, Speech, Signal, and Automation). It deals with speech and audio processing (analysis, modeling, coding, transformation, synthesis, localization, enhancement and separation), with a special interest in multimodal speech processing (e.g., audiovisual processing or articulatory-acoustic modeling). Prof. Girin is currently the Head of the Speech and Cognition Pole at GIPSA-lab (one of the four scientific departments of GIPSA-lab). He is also a regular collaborator of the RobotLearn team at Inria Grenoble.



Xiaoyu BIE received the B.Sc. degree in optical and electronic information from Huazhong University of Science and Technology, Wuhan, China, in 2016, the Engineering degree in applied optics from Institut d'Optique/University of Paris-Saclay, Gif-sur-Yvette, France, in 2018 and the M.Sc. degree in signal and image processing from Centrale-Supélec/University of Paris-Saclay, Gif-sur-Yvette, France, in 2018. He is currently working toward the Ph.D. degree with INRIA and Univ. Grenoble Alpes. His research focuses on deep sequential generative

model, speech analysis and human understanding.