A Benchmark of Dynamical Variational Autoencoders applied to Speech Spectrogram Modeling

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

A Benchmark of Dynamical Variational Autoencoders applied to Speech Spectrogram Modeling

Xiaoyu Bie[1]*, Laurent Girin*[2]*, Simon Leglaive*[3]*, Thomas Hueber*[2] *and Xavier Alameda-Pineda*[1]

[1] Inria, Univ. Grenoble Alpes, CNRS, LJK, 38000 Grenoble, France.
[2] Univ. Grenoble Alpes, CNRS, Grenoble-INP, GIPSA-lab, 38000 Grenoble, France.
[3] CentraleSupélec, IETR, 35576 Cesson-Sévigné, France.

Abstract

The Variational Autoencoder (VAE) is a powerful deep generative model that is now extensively used to represent high-dimensional complex data via a low-dimensional latent space learned in an unsupervised manner. In the original VAE model, input data vectors are processed independently. In recent years, a series of papers have presented different extensions of the VAE to process sequential data, that not only model the latent space, but also model the temporal dependencies within a sequence of data vectors and corresponding latent vectors, relying on recurrent neural networks. We recently performed a comprehensive review of those models and unified them into a general class called Dynamical Variational Autoencoders (DVAEs). In the present paper, we present the results of an experimental benchmark comparing six of those DVAE models on the speech analysis-resynthesis task, as an illustration of the high potential of DVAEs for speech modeling.

Index Terms: Speech signals modeling, dynamical variational autoencoders, speech spectrograms, speech analysis-resynthesis

1. Introduction

The Variational Autoencoder (VAE) introduced in [1, 2] is a powerful deep generative model that is now extensively used to represent high-dimensional data via a low-dimensional latent space learned in an unsupervised manner. It has been used for speech modeling in, e.g., [3, 4, 5, 6, 7, 8, 9, 10, 11].

The original VAE does not include temporal modeling. This means that every data vector from a dataset is processed independently of the other data vectors (and the corresponding latent vector is also processed independently of the other latent vectors). In recent years, a series of papers have presented different extensions of the VAE to process sequential data, that not only model the latent space, but also model the temporal dependencies within a sequence of data vectors and corresponding latent vectors, relying on recurrent neural networks (RNNs) [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. In practice, those different models vary in how they define the dependencies between the observed and latent variables, how they define and parameterize the corresponding generative distributions, and, importantly, how they define and parameterize the inference model, which is a key ingredient of the VAE methodology. They also differ on how they combine the variables with RNNs to model temporal dependencies, at both generation and inference. In contrast, and remarkably, the training phase is quite similar between models since it is consistently based on the VAE methodology: chaining the inference and generative models (the encoder and decoder) and maximizing a lower bound of the data likelihood over a training dataset.

In [24], we performed an extensive and comprehensive literature review of these models. We introduced a general class of models called Dynamical Variational Autoencoders (DVAEs) that encompasses and unifies the above-cited temporal VAE extensions. The objectives and contributions of the present paper are the following. First, it seems that the DVAE class of models is still relatively poorly known by the speech processing community, yet it has the potential to yield major advances in many speech processing applications such as speech signal synthesis and transformation. So far, it has been used only in a very few studies, see for example the pioneering works in speech coding [25] or speech denoising [23]. Therefore we want to disseminate the existence of the DVAE class of models to the speech processing community and foster research addressing speech processing applications with DVAEs. Then, as an illustration of their potential for speech modeling, in the present paper, we report the results of an experimental benchmark conducted on the speech (power spectrogram) analysis-resynthesis task. We selected six of the DVAE models that we detailed in [24], we reimplemented them and compared their performance on this task. The PyTorch code is made publicly available, and we believe it can be of interest for many researchers interested in joint unsupervised representation learning and dynamical modeling of speech signals.

2. Dynamical VAEs

2.1. The original VAE model

The seminal VAE model introduced in [1, 2] is defined by:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}), \tag{1}$$

where $p(\mathbf{z})$, the prior distribution of the latent variable $\mathbf{z}$, is a multivariate standard Gaussian distribution, $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is the (conditional) *likelihood function* of the observed variable $\mathbf{x}$, and the dimension L of $\mathbf{z}$ is (much) lower than the dimension F of $\mathbf{x}$. The parameters of $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ are provided by a deep neural network (DNN), called the decoder network, that takes $\mathbf{z}$ as input. $\boldsymbol{\theta}$ represents the parameters of this decoder network (e.g., the weights and biases of a multi-layer perceptron).

Because the relationship between $\mathbf{z}$ and $\mathbf{x}$ is highly nonlinear, the posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ is not analytically tractable. It is thus approximated with a parametric variational distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$, a.k.a. the inference model, whose parameters are provided by another DNN (called the encoder network, with weights $\boldsymbol{\phi}$ and input $\mathbf{x}$). A usual choice is to set $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ as a Gaussian distribution with diagonal covariance matrix. The parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$ are then jointly estimated by maximizing a lower bound of the data log-likelihood function, called the Variational Lower Bound (VLB) or Evidence Lower

Bound (ELBO), given by (for one single data vector):

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\ln p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - D_{\mathrm{KL}}\big[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big], \tag{2}$$

and evaluated on a training dataset (D_{KL} denotes the Kullback-Leibler divergence). Maximization of the VLB is done by combining stochastic gradient descent with sampling techniques. Such optimization process is now considered as routine within deep learning toolkits such as Keras and PyTorch.

2.2. From VAE to DVAEs

As already mentioned in the introduction, the Dynamical Variational Autoencoder (DVAE) is a class of deep generative models that generalizes the VAE to the modeling of sequential data [24]. The DVAE models that we consider here process an ordered time sequence of vector data $\mathbf{x}_{1:T} = \{\mathbf{x}_t\}_{t=1}^T$ and a corresponding ordered sequence of latent vectors $\mathbf{z}_{1:T} = \{\mathbf{z}_t\}_{t=1}^T$. A DVAE is thus defined by the following joint probability density function (pdf), which is a generalization of (1):

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}). \tag{3}$$

However, this form does not give much information about the generative process, and we prefer to use the chain rule to reformulate (3) as:

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}) p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}). \tag{4}$$

This particular reformulation (among many other possibilities) is a *causal* form: $\mathbf{z}_t$ is first generated from $\mathbf{z}_{1:t-1}$ and $\mathbf{x}_{1:t-1}$, and then $\mathbf{x}_t$ is generated from $\mathbf{x}_{1:t-1}$ and $\mathbf{z}_{1:t}$. In a general manner, the dependencies are implemented using RNNs : The parameters of the generative distributions of $\mathbf{z}_t$ and $\mathbf{x}_t$ are the outputs of RNNs that take $\mathbf{z}_{1:t-1}$ and $\mathbf{x}_{1:t-1}$ (and $\mathbf{z}_t$ for the generation of $\mathbf{x}_t$) as inputs.

The different DVAE models that we cited in the introduction are all special cases of the general expression (4) where the dependencies are possibly simplified. In the following of the paper, we consider the six following DVAE models: The Deep Kalman Filter model (DKF) [14, 18], the Stochastic Recurrent Neural Network (STORN) [12], the Variational Recurrent Neural Network (VRNN) [15, 20], another type of Stochastic Recurrent Neural Network (SRNN) [17], the Recurrent Variational Autoencoder (RVAE) [23] and the Disentangled Sequential Autoencoder (DSAE) [22]. The corresponding simplified forms of the generative distributions are given in Table 1. Note that VRNN is the "richer" possible DVAE model in terms of variable dependencies since all dependencies in (4) are kept, whereas in contrast, the original VAE can be seen as a DVAE where all temporal dependencies have been removed.

The inference and training methodology of a DVAE model follows the one of the VAE: Definition of an inference model $q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ (since the exact posterior distribution $p_{\boldsymbol{\theta}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$ is not analytically tractable), chaining of the encoder and decoder, and training by maximizing the VLB on training data. Similar to the generative model, it is convenient to reshape the inference model in the following general form, using the chain rule:

$$q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{x}_{1:T}). \tag{5}$$

Table 1: *Conditional independence assumptions for various models in the DVAE family. The * indicates that the inference model is compliant with the structure of the true posterior. For DSAE,* $\mathbf{v}$ *is an additional sequence-level variable (not detailed here, see [22, 24] for details).*

		$p_\theta(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$	$p_\theta(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$
VAE*	[1, 2]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{z}_t)$
RVAE*	[23]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{z}_{1:t})$
STORN	[12]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t)$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$
DKF*	[14, 18]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{z}_{t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{z}_t)$
DSAE	[22]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{z}_t, \mathbf{v})$
VRNN	[15, 20]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$
SRNN*	[17]	$p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$	$p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_t)$

We can see that (5) is causal regarding the past latent vectors $\mathbf{z}_{1:t-1}$ but not regarding the complete sequence of observed vectors $\mathbf{x}_{1:T}$. Similar to the DVAE generative model, the dependencies in (5) can be simplified (or not), depending on, e.g., if one wants the inference model to have the same structure as the exact posterior distribution [26], [27, Chapter 8], or if one wants to have a causal implementation to enable online inference. In the present paper, for each DVAE model we used the inference model defined in the corresponding original paper.

Still Similar to the VAE, the training of DVAE models is based on maximization of the VLB, here defined by (for one data sequence) [24]:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{x}_{1:T}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})}\big[\ln p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\big]. \tag{6}$$

The developed form of this VLB, obtained by reinjecting (4) and (5) into (6) and using some "cascade" trick, is given in [24]. Of course, depending on the specific (chosen) DVAE generative and inference models, this developed form can be simplified. In a general manner, expressing the VLB in a form that is differentiable w.r.t. $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ requires some sampling of $\mathbf{z}_{1:T}$, which is here done recursively. This sampling is alternated with calculation of the VLB gradient over a training dataset and parameter update. Again, see [24] for more detailed information.

3. Application to speech power spectrogram modeling

In the literature, the DVAE models have been applied to different kinds of data. Here we focus on the modeling of speech signals. This is done in the short-term Fourier transform (STFT) domain: The time-domain speech waveform is first transformed into a speech STFT spectrogram $\mathbf{s}_{1:T} = \{\mathbf{s}_t\}_{t=1}^T$, where each complex-valued vector $\mathbf{s}_t = \{s_{t,f}\}_{f=0}^{F-1}$ is the speech short-term spectrum at time index t, and f is the frequency bin. As is usual in speech/audio processing, $s_{t,f}$ is assumed to follow a circular-symmetric zero-mean complex-valued Gaussian distribution, see, e.g., [28, 29, 30]. Moreover, the STFT coefficients at different frequency bins are assumed to be (conditionally) independent, i.e. the covariance matrix of $\mathbf{s}_t$ is diagonal.

In practice, the data sequence $\mathbf{x}_{1:T}$ processed by a DVAE is the STFT *power* spectrogram, i.e. $x_{t,f} = |s_{t,f}|^2$ for all time-frequency bins. Given the above assumption on $s_{t,f}$, each power spectrogram coefficient $x_{t,f}$ follows a Gamma distribution with shape parameter 1 and scale parameter

$\sigma^2_{\mathbf{s},t,f}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$ [29, 31, 32].[1] In the most general DVAE context, the parameter vector $\boldsymbol{\sigma}^2_{\mathbf{s},t}(\cdot) = \{\sigma^2_{\mathbf{s},t,f}(\cdot)\}_{f=0}^{F-1}$ depends on $\mathbf{x}_{1:t-1}$ and $\mathbf{z}_{1:t}$, i.e. it is provided by RNNs taking $\mathbf{x}_{1:t-1}$ and $\mathbf{z}_{1:t}$ as inputs. Again, those dependencies can be simplified depending on the specific DVAE model, see Table 1. Note that by using the above probabilistic model, we assume that the phase of $s_{t,f}$ follows a uniform distribution in $[0, 2\pi[$. This is a very common assumption in speech/audio processing, since modeling the phase of STFT spectrograms is still a very challenging task, be it with classical statistical models or with deep-learning-based models (see [33] for an example in the VAE framework).

As for the latent vector generative distribution $p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$, it is set in its most general form as a (real-valued) Gaussian distribution, with mean vector $\boldsymbol{\mu}_{\mathbf{z},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$ and a diagonal covariance matrix with entries from vector $\boldsymbol{\sigma}^2_{\mathbf{z},t}(\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$. Those two vectors are provided by RNNs taking here $\mathbf{x}_{1:t-1}$ and $\mathbf{z}_{1:t-1}$ as inputs, and again, the dependencies can be simplified according to Table 1 depending on the specific DVAE model. Note that in all cases the entries of the latent vector $\mathbf{z}_t$ are assumed (conditionally) independent, which is in line with the principle of looking for a disentangled latent representation [24].

Finally, the encoder follows (5), where, in a general manner, $q_{\boldsymbol{\phi}}(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{s}_{1:T})$ is a (real-valued) Gaussian distribution with mean vector $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{z}_{1:t-1}, \mathbf{s}_{1:T})$ and a diagonal covariance matrix with entries from vector $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{z}_{1:t-1}, \mathbf{s}_{1:T})$. Those two vectors are provided by the encoder RNN. As stated in Section 2.2, the dependencies can be simplified, and in our experiments, for each specific DVAE model, we used the inference model described in the corresponding original paper.

4. Experimental benchmark

In the following, we present our experimental benchmark of the six DVAE models of Table 1 applied to speech power spectrogram analysis-resynthesis. In short, a speech power spectrogram $\mathbf{x}_{1:T}$, is encoded into, and then resynthesized from, a latent vector sequence $\mathbf{z}_{1:T}$.

4.1. Implementation of the DVAE models

Because of lack of room, it is not possible to detail here the implementation of each model. The detailed implementation equations of both the generative part and inference part of each DVAE model, involving the expression of RNN internal states, are detailed in [24]. Importantly, not only the different DVAE models differ in the way the dependencies in $p_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t})$ and $p_{\boldsymbol{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1})$ are simplified, but a given DVAE model can have different implementations. Indeed, the parameters of the generative distributions are provided by neural networks, and many different network implementations can be considered for the same dependency structure [24]. In a general manner, we have tried to find a good trade-off between respecting the architecture of the model as described in the original paper and ensuring a fair comparison between the different models for the speech analysis-resynthesis task. Similar modules across different DVAE models are thus implemented in the same manner, i.e. with the same number of layers, units per layers, and activation function (see [24] for details). Moreover, the following specifications are common to all DVAEs:

- The dimension of the observation vector $\mathbf{x}_t$ and output parameter $\boldsymbol{\sigma}^2_{\mathbf{s},t}(\cdot)$ is set to $F = 257$ (see next subsection);
- The dimension of the latent vector $\mathbf{z}_t$ is set to $L = 16$;
- The dimension of RNN hidden internal state vectors is set to 128; Unless specified in the original paper, all RNNs are instantiated as LSTM networks;

4.2. Dataset and pre/post-processing

We used the Wall Street Journal (WSJ0) dataset [34]. The *si_tr_s*, *si_dt_05* and *si_et_05* subsets were used for model training, validation, and test, respectively. The STFT was applied on 16-kHz signals with a 32-ms sine window (512 samples) and 50%-overlap to obtain sequences of 257-dimensional discrete spectra (for positive frequencies). We set $T = 150$ (i.e. 2.4-s speech sequences). In summary, each data sequence is a 150×257 STFT power spectrogram. This data preprocessing resulted in a set of $N_{\text{tr}} = 13,272$ training sequences (about 9h of speech signal) and $N_{\text{val}} = 2,143$ validation sequences (about 1.5h). For test, we used the STFT power spectrogram of each complete test sequence (with beginning and ending silence portions removed), which can be of variable length, most often larger than 2.4s (total of about 1.5h). For the evaluation, we used the reconstructed power spectrograms, as well as the reconstructed waveforms, obtained by combining the reconstructed magnitude spectrograms with the input phase spectrograms and applying inverse STFT with overlap-add.

4.3. Training and testing

All models were implemented in PyTorch [35]. Training was made with mini-batch stochastic gradient descent, using the Adam optimizer [36], with a learning rate of 0.0001 and a batch size of 32. We used early stopping on the validation set with a patience of 20 epochs. After training, we evaluated the average performance on the test set using the following three metrics: The root mean squared error (RMSE) between original and reconstructed waveforms,[2] Perceptual Evaluation of Speech Quality (PESQ) scores [37] and Short-Time Objective Intelligibility (STOI) scores [38]. The amplitude of each original speech waveform was normalized in $[-1, 1]$, so the RMSE (generally much lower than 1) directly represents a percentage of the maximum amplitude value. PESQ scores are in $[-0.5, 4.5]$ and STOI scores are in $[0, 1]$. For both, the higher the better.

4.4. Results and discussion

We first checked that the loss curves (i.e., VLB up to a constant term) obtained on the training data and the validation data show a successful convergence of the training for all the implemented DVAE models. Then we report in Table 2 the values of the three evaluation metrics averaged over the test dataset. From this table, and from the observation of reconstructed spectrograms (not shown here because of room limitation), we can draw the following comments:

- First, all tested DVAE models lead to correct signal reconstruction, with an RMSE that represents only a few percents (generally lower than 5%) of the maximum waveform amplitude. The quality of the reconstructed speech signals, as mea-

[1] $\sigma^2_{\mathbf{s},t,f}(\cdot)$ is also the variance of $s_{t,f}$, the mean of $x_{t,f}$, and the speech signal power spectral density.

[2] Because of the orthogonal properties of the Discrete Fourier Transform, and because the phase of the original spectrogram is combined with the reconstructed magnitude spectrogram, RMSE calculated between speech waveforms is equivalent to RMSE calculated between the corresponding magnitude spectrograms.

sured by PESQ, goes from fair to good. STOI scores, generally higher than 0.90 show their good intelligibility. Importantly, all DVAE models outperform the standard VAE model. This demonstrates the interest of including temporal modeling in the VAE framework for modeling speech signals.

- VRNN and SRNN are the two methods with highest performance, with a notable gain in performance over all other models, and SRNN is slightly better than VRNN. We recall that VRNN keeps all possible dependencies in the general DVAE formulation (4), and Table 1 shows that SRNN contains more dependencies with the past observed and latent variables than the other implemented DVAEs. We believe that this allows VRNN and SRNN to better capture the temporal patterns of speech spectrograms. As for SRNN performing slightly better than VRNN, this could be due to the fact that the inference model of SRNN respects the structure of the exact posterior distribution, which depends on all observations $\mathbf{x}_{1:T}$, whereas the inference model of VRNN (as proposed in the original paper) does not: it uses only the causal observations $\mathbf{x}_{1:t}$ [15, 17, 24].
- The performance scores of DKF and STORN are quite equivalent, but below those of SRNN and VRNN. This is likely due to the fact that the temporal dependencies in DKF and STORN are less rich than in VRNN and SRNN. DKF has the structure of a state-space model, where there is no explicit temporal dependency between $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$, but only between $\mathbf{z}_{t-1}$ and $\mathbf{z}_t$. In STORN $\mathbf{x}_t$ depends on $\mathbf{x}_{1,t-1}$ and $\mathbf{z}_{1:t}$, so one could think that this model is richer that DKF and should have better performance. However, we found out that DKF slightly outperforms STORN in terms of PESQ and STOI (but not in RMSE). We can hypothesize that, here also the difference in performance is (at least partly) due to the fact that the inference model of DKF does respect the structure of the exact posterior distribution whereas the inference model of STORN does not (for STORN, the inference of $\mathbf{z}_t$ depends only on $\mathbf{x}_{1:t}$ and not on $\mathbf{z}_{t-1}$ nor $\mathbf{x}_{t+1:T}$, see [12, 24]). Another possible explanation is that the prior distribution of $\mathbf{z}_t$ is i.i.d. in STORN, while it has temporal dependencies in DKF. In a general manner, we hypothesize that models with i.i.d. prior over time on $\mathbf{z}_t$ risk to underperform w.r.t. models that are defined via a temporal generative model of $\mathbf{z}_t$. We must keep in mind that the $\mathbf{z}_{1:T}$ sequence is assumed to encode high-level characteristics of the data $\mathbf{x}_{1:T}$ that generally evolve smoothly over time (at least for some of these characteristics). This is not ensured by the i.i.d. standard Gaussian prior distribution of $\mathbf{z}_t$ used in STORN.
- The performance of DSAE is quite disappointing, especially compared to the performance of DKF. Indeed, like DKF, DSAE also has the structure of a state-space model. Actually, DSAE can be seen as an improved version of DKF, with an additional sequence-level variable (not detailed here) and infinite-order temporal dependency of $\mathbf{z}_t$ (as opposed to first-order for DKF). Again, this poor performance could come from the structure of the inference model, which depends on $\mathbf{x}_{1:T}$, whereas the exact posterior distribution of $\mathbf{z}_t$ depends on $\mathbf{z}_{1:t-1}$ and $\mathbf{x}_{t:T}$ [24].
- In our experiments, RVAE exhibits the worst performance of all tested DVAE models. Here also, i.i.d. modeling of $\mathbf{z}_t$ may be suboptimal. In addition to that, there is no explicit modeling of the temporal dependencies on $\mathbf{x}_t$ (e.g., $\mathbf{x}_t$ does not depend on $\mathbf{x}_{t-1}$), hence leading to a model with weak "predictive power." However, we recall that for the present experiments, we set up the neural network architectures so that all compared DVAE models have a similar number of parameters. For RVAE, the architecture was made more complex than in the original paper [23], which drastically decreased the performance. With the original RVAE model in [23], the RMSE, PESQ and STOI performance are 0.0297, 3.47 and 0.95, respectively. This shows that two DVAE models with the same probabilistic dependencies but different neural network architectures can perform very differently.

Table 2: *Performance of the tested DVAE models in our speech analysis-resynthesis experiments (RMSE, PESQ and STOI scores averaged over the WSJ0 test subset).*

	VAE	DKF	STORN	VRNN	SRNN	RVAE	DSAE
RMSE ($\times 10^{-2}$)	5.10	3.44	3.38	2.67	**2.48**	4.99	4.69
PESQ	2.05	3.30	3.05	3.60	**3.64**	2.27	2.32
STOI	0.86	0.94	0.93	0.96	**0.97**	0.89	0.90

5. Conclusions

We presented a benchmark of several DVAE models that are of high interest for speech modeling. This benchmark shows that, in a practical application requiring the modeling of speech power/magnitude spectrograms, VRNN or SRNN seem particularly promising. Some of the other tested DVAE models show a slightly lower performance but they also show a reduced complexity that can be an interesting feature. Also, we can conjecture that having an inference model that respects the exact variable dependencies at inference time is very important for obtaining optimal performance. However this is not always possible, e.g. some applications may require a causal inference model for online processing.

We insist on the fact that the above "model ranking" is valid only for the presented experiments, which consist of pure analysis-resynthesis of speech power spectrograms. For other tasks such as speech signal generation (from new values of the latent vectors) or speech signal transformation (with modification of the latent vectors), we have no claim on how the implemented models would behave. So far, it is still quite difficult to know how much of the information about $\mathbf{x}_{1:T}$, and what kind of information, is encoded into $\mathbf{z}_{1:T}$, and in particular, we do not know about the "disentanglement power" of each model. And importantly, there is no clear methodology to address those issues, i.e., to evaluate the generated data and the representation power of the latent variable. This is part of our current works.

The code re-implementing the six tested DVAE models (+ the VAE) and used on the benchmark task is made available to the community. The open-source code and the best trained models can be downloaded at the following repository: `https://github.com/XiaoyuBIE1994/DVAE-speech`. We have taken care, in the code, to follow the unified presentation and notation used in our review paper [24], making it, hopefully, a useful resource to the speech processing community.

6. Acknowledgements

This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003) and by the European Commission (H2020 SPRING project under GA #871245).

7. References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *International Conference on Learning Representations (ICLR)*,

Banff, Canada, 2014.

[2] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning (ICML)*, Beijing, China, 2014.

[3] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," in *Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, CA, 2016.

[4] C. Hsu, H. Hwang, Y. Wu, Y. Tsao, and H. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Jeju, Korea, 2016.

[5] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[6] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018.

[7] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, Denmark, 2018.

[8] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India, 2018.

[9] L. Pandey, A. Kumar, and V. Namboodiri, "Monaural audio source separation using variational autoencoders," in *Conference of the International Speech Communication Association (INTERSPEECH)*, Hyderabad, India, 2018.

[10] S. Leglaive, U. Simsekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[11] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[12] J. Bayer and C. Osendorfer, "Learning stochastic recurrent networks," *arXiv preprint arXiv:1411.7610*, 2014.

[13] O. Fabius and J. R. van Amersfoort, "Variational recurrent auto-encoders," *arXiv preprint arXiv:1412.6581*, 2014.

[14] R. Krishnan, U. Shalit, and D. Sontag, "Deep Kalman filters," in *arXiv preprint arXiv:1511.05121*, 2015.

[15] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 2980–2988.

[16] S. Gu, Z. Ghahramani, and R. E. Turner, "Neural adaptive sequential Monte Carlo," in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 2629–2637.

[17] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther, "Sequential neural models with stochastic layers," in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016.

[18] R. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," in *AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017.

[19] M. Fraccaro, S. Kamronn, U. Paquet, and O. Winther, "A disentangled recognition and nonlinear dynamics model for unsupervised learning," in *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017, pp. 3601–3610.

[20] A. Goyal, A. Sordoni, M.-A. Côté, N. R. Ke, and Y. Bengio, "Z-forcing: Training stochastic recurrent networks," in *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017, pp. 6713–6723.

[21] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems*, Long Beach, CA, 2017, pp. 1878–1889.

[22] Y. Li and S. Mandt, "Disentangled sequential autoencoder," in *International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018, pp. 5670–5679.

[23] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[24] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *arXiv preprint arXiv:2008.12595*, 2020.

[25] Y. Yang, G. Sautière, J. J. Ryu, and T. S. Cohen, "Feedback recurrent autoencoder," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 3347–3351.

[26] D. Geiger, T. Verma, and J. Pearl, "Identifying independence in bayesian networks," *Networks*, vol. 20, no. 5, pp. 507–534, 1990.

[27] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[28] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[29] 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.

[30] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, no. 7, pp. 3155–3167, 2011.

[31] C. Févotte and A. Cemgil, "Nonnegative matrix factorizations as probabilistic inference in composite models," in *European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, 2009.

[32] L. Girin, F. Roche, T. Hueber, and S. Leglaive, "Notes on the use of variational autoencoders for speech and audio spectrogram modeling," in *Digital Audio Effects Conference (DAFx)*, Birmingham, UK, 2019.

[33] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A deep generative model of speech complex spectrograms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[34] 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.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019, pp. 8026–8037.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, Utah, 2001.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, 2010, pp. 4214–4217.