Notes on the use of variational autoencoders for speech and audio spectrogram modeling

Laurent Girin, Fanny Roche, Thomas Hueber, Simon Leglaive

NOTES ON THE USE OF VARIATIONAL AUTOENCODERS FOR SPEECH AND AUDIO SPECTROGRAM MODELING

Laurent Girin, Thomas Hueber
Univ. Grenoble Alpes, CNRS,
Grenoble INP, GIPSA-lab
Grenoble, France
laurent.girin@grenoble-inp.fr
thomas.hueber@grenoble-inp.fr

Fanny Roche *
Arturia
Meylan, France
fanny.roche@arturia.com

Simon Leglaive
Inria Grenoble Rhône-Alpes
Grenoble, France
simon.leglaive@inria.fr

ABSTRACT

Variational autoencoders (VAEs) are powerful (deep) generative artificial neural networks. They have been recently used in several papers for speech and audio processing, in particular for the modeling of speech/audio spectrograms. In these papers, very poor theoretical support is given to justify the chosen data representation and decoder likelihood function or the corresponding cost function used for training the VAE. Yet, a nice theoretical statistical framework exists and has been extensively presented and discussed in papers dealing with nonnegative matrix factorization (NMF) of audio spectrograms and its application to audio source separation. In the present paper, we show how this statistical framework applies to VAE-based speech/audio spectrogram modeling. This provides the latter insights on the choice and interpretability of data representation and model parameterization.

1. INTRODUCTION

Autoencoders (AEs) are a specific type of deep neural networks (DNNs) that can learn from data a non-linear projection of the signal space into a low-dimensional latent space (encoding step), followed by inverse non-linear transformation of the latent coefficients into the original signal space (decoding step) [1]. AEs have been essentially used as an unsupervised technique for data dimension reduction. More recently, variational autoencoders (VAEs) were proposed as a probabilistic/generative extension of AEs [2]: Instead of deterministically mapping the input vector $\mathbf{x}$ into a unique vector of latent coefficients $\mathbf{z}$, as done in AEs, the VAE *encoder network* maps $\mathbf{x}$ into the parameters of a conditional distribution $q_\phi(\mathbf{z}|\mathbf{x})$ of $\mathbf{z}$. Similarly, the *decoder network* maps a vector of latent coefficient $\mathbf{z}$ into the parameters of a conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$ of $\mathbf{x}$. A VAE decoder is thus intrinsically a (non-linear and deep) generative model of $\mathbf{x}$, conditioned on the latent variable $\mathbf{z}$ (which is itself conditioned on the input when decoding follows encoding). VAEs thus combine the modeling power of DNNs with the flexibility of generative models.

VAEs have recently received a strong interest for speech and audio processing, more specifically for modeling, transformation and synthesis of speech signals [3, 4, 5, 6], for music sound synthesis [7, 8], and for single-channel [9, 10, 11, 12] and multi-channel [13, 14, 15] speech enhancement and separation. In all those papers, VAEs are used to process a sequence of vectors encoding the short-time Fourier transform (STFT) spectrogram extracted from speech or music signals. For synthesis/transformation applications, the output audio signal is reconstructed using the decoded magnitude spectrogram, after possible modification of the latent coefficients, and either the phase of the original signal or some reconstructed phase more coherent with the decoded magnitude spectrogram. For speech enhancement application, the decoder of the VAE is used as a supervised generative model of the speech signal in the STFT domain, which is exploited in a probabilistic enhancement/separation method.

A keypoint is that in most of these papers, very few justification is given about the precise choice of the encoder and decoder conditional distributions, or the corresponding cost function used for VAE training. These distributions are generally chosen as Gaussian for convenience, but the choice for their parameters is not clearly justified. The same about the related issue of data representation: It is chosen a bit arbitrarily, without clear theoretical support, possibly more considering DNN training issues rather than fundamental signal processing ones.

Yet, this theoretical framework exists. In fact, it has been extensively presented and discussed in the seminal papers [16] and [17]. Those papers describe the statistical framework underlying the decomposition of audio magnitude/power spectrograms using Nonnegative Matrix Factorization (NMF) [18]. These developments have then been extensively used for audio source separation, see e.g. among many others [19, 20, 21, 22, 23, 24, 25]. In the present paper, we show how this theoretical statistical framework applies to the VAE model. Based on [16, 17], we describe the three main cases encountered in practice, with three modeling cost functions corresponding to three signal statistical models. We show how this provides interesting insights on the choice and interpretability of data representation and loss function for speech/audio spectrogram modeling with VAEs.

The remainder of this paper is organized as follows. Section 2 presents the VAE framework. In Section 3, we discuss the way VAEs are currently used to model speech/audio signals in the literature, and raise a set of related questions. In Section 4 we present the nonnegative representation and underlying signal statistical models as a general framework, of which NMF is a particular case, and we show how this framework also applies to VAE-based spectrogram modeling. Section 5 illustrates this discussion with some experiments on speech/audio analysis-synthesis with VAEs. Section 6 draws a series of conclusions and perspectives.

* This work is supported by a CIFRE PhD Grant funded by ANRT

2. VARIATIONAL AUTOENCODERS

As mentioned in the introduction, a VAE can be seen as a probabilistic autoencoder. In the original formulation of the seminal paper [2], a VAE delivers a parametric model of data distribution:

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

where $\mathbf{x} \in \mathbb{R}^F$ is a vector of observed data, $\mathbf{z} \in \mathbb{R}^L$ is a corresponding vector of latent data, with $L \ll F$, and θ denotes the set of distribution parameters. The likelihood function $p_\theta(\mathbf{x}|\mathbf{z})$ plays the role of a probabilistic decoder which models how the generation of observed data $\mathbf{x}$ is conditioned on the latent data $\mathbf{z}$. The prior distribution $p_\theta(\mathbf{z})$ is used to structure (or regularize) the latent space. Typically a standard Gaussian distribution is used: $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}_L)$, where $\mathbf{I}_L$ is the identity matrix of size L. This encourages the latent coefficients to be orthogonal and with similar range. Note that this prior actually lacks parameters. The likelihood $p_\theta(\mathbf{x}|\mathbf{z})$ is usually defined as Gaussian:

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\sigma}^2_\theta(\mathbf{z})), \tag{2}$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ denotes the probability density function (pdf) of the multivariate Gaussian distribution which is defined in the Appendix, and $\boldsymbol{\mu}_\theta(\mathbf{z}) \in \mathbb{R}^F$ and $\boldsymbol{\sigma}^2_\theta(\mathbf{z}) \in \mathbb{R}^F_+$ are the outputs of the decoder network. The parameter set θ is composed of the weights of this decoder network. Note that the entries of $\mathbf{x}$ are assumed independent as common in VAEs, so the vector $\boldsymbol{\sigma}^2_\theta(\mathbf{z})$ contains the diagonal coefficients of a diagonal covariance matrix.

The exact posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$ corresponding to the above model is intractable. It is approximated with a tractable parametric model $q_\phi(\mathbf{z}|\mathbf{x})$ that plays the role of the corresponding probabilistic encoder. This model generally has a form similar to the decoder:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \tilde{\boldsymbol{\mu}}_\phi(\mathbf{x}), \tilde{\boldsymbol{\sigma}}^2_\phi(\mathbf{x})), \tag{3}$$

where $\tilde{\boldsymbol{\mu}}_\phi(\mathbf{x}) \in \mathbb{R}^L$ and $\tilde{\boldsymbol{\sigma}}^2_\phi(\mathbf{x}) \in \mathbb{R}^L_+$ are the outputs of the encoder network. The parameter set ϕ is composed of the weights of this encoder network. As before, $\tilde{\boldsymbol{\sigma}}^2_\phi(\mathbf{x})$ is a vector containing the diagonal entries of a diagonal covariance matrix.

Training of the VAE model, i.e. estimation of θ and ϕ, is made by optimizing a lower-bound of the marginal log-likelihood $\log p_\theta(\mathbf{x})$ computed from a large training dataset of vectors $\mathbf{x}$. It is shown in [2] that the marginal log-likelihood for an individual vector $\mathbf{x}$ writes:

$$\log p_\theta(\mathbf{x}) = d_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}(\phi, \theta, \mathbf{x}), \tag{4}$$

where $d_{\text{KL}} \geq 0$ denotes the Kullback-Leibler (KL) divergence and $\mathcal{L}(\phi, \theta, \mathbf{x})$ is the variational lower bound (VLB) given by:

$$\mathcal{L}(\phi, \theta, \mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{reconstruction accuracy}} \underbrace{- d_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}))}_{\text{regularization}}. \tag{5}$$

We can see that the VLB is the sum of two terms. The first term represents the average reconstruction accuracy. The second term acts as a regularizer encouraging the approximate posterior $q_\phi(\mathbf{z}|\mathbf{x})$ to be close to the prior $p_\theta(\mathbf{z})$. Since the expectation taken with respect to $q_\phi(\mathbf{z}|\mathbf{x})$ in the reconstruction accuracy term is analytically intractable, it is approximated using a Monte Carlo estimate with R samples $\mathbf{z}^{(r)}$ independently and identically drawn from $q_\phi(\mathbf{z}|\mathbf{x})$:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \approx \frac{1}{R}\sum_{r=1}^{R} \log p_\theta(\mathbf{x}|\mathbf{z}^{(r)}). \tag{6}$$

In practice a training dataset $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N_{tr}}$ is used for the training of the VAE. Under the hypothesis of independent and identically distributed (i.i.d.) training vectors, the VAE training is done by maximizing the total VLB, which is the sum of individual VLBs over the training vectors. If we consider only one Monte Carlo sample per training vector (which is common practice provided that the batch size is sufficiently large [2]), or if we consider several Monte Carlo samples as additional training data, we can write the total VLB as:

$$\begin{aligned}\mathcal{L}(\phi, \theta, \mathbf{X}) = &\sum_{n=1}^{N_{tr}} \log p_\theta(\mathbf{x}_n|\mathbf{z}_n) \\ &- \sum_{n=1}^{N_{tr}} d_{\text{KL}}(q_\phi(\mathbf{z}_n|\mathbf{x}_n)|p_\theta(\mathbf{z}_n)).\end{aligned} \tag{7}$$

For the present case of Gaussian likelihood (2) and Gaussian encoding distribution (3), the VLB in (7) becomes:

$$\begin{aligned}\mathcal{L}(\phi, \theta, \mathbf{X}) = &-\sum_{n=1}^{N_{tr}}\sum_{f=0}^{F-1}\left(\log \sigma^2_{\theta,f}(\mathbf{z}_n) + \frac{(x_{fn} - \mu_{\theta,f}(\mathbf{z}_n))^2}{2\sigma^2_{\theta,f}(\mathbf{z}_n)}\right) \\ &+ \frac{1}{2}\sum_{n=1}^{N_{tr}}\sum_{l=1}^{L}\left(\log \tilde{\sigma}^2_{\phi,l}(\mathbf{x}_n) - \tilde{\mu}_{\phi,l}(\mathbf{x}_n)^2 - \tilde{\sigma}^2_{\phi,l}(\mathbf{x}_n)\right)\end{aligned} \tag{8}$$

where the subscript f or l denotes the f-th or l-th entry of a vector. Maximization of the total VLB is done by using the usual back-propagation technique and gradient-based optimization, which are not detailed in this paper. For more technical details that are not relevant here, the reader is referred to [2].

3. VAES FOR SPECTROGRAM MODELING: FACTS AND QUESTIONS

In this section, we analyze how VAEs are generally used for speech and audio spectrogram modeling in the recent literature. Although some of the points discussed below may seem trivial, they rise a series of fundamental questions that are poorly discussed in these papers and that we will address in the following.

3.1. Audio signal representation in the STFT domain

As shortly stated in the introduction, the processing is generally carried out in the STFT domain. Let $\mathbf{S} = [s_{fn}]_{f=0,n=1}^{F-1,N} \in \mathbb{C}^{F\times N}$ denote the STFT of a speech/audio signal, where f is the frequency bin index and n is the time frame index. Let $\mathbf{X} = [x_{fn}]_{f=0,n=1}^{F-1,N} \in \mathbb{R}_+^{F\times N}$ denote the corresponding real-valued and nonnegative *magnitude* or *power* spectrogram, i.e. $\mathbf{X} = |\mathbf{S}|$ or $\mathbf{X} = |\mathbf{S}|^2$, where $|.|$ and $.^2$ are to be understood as entry-wise operators. Note that we use the same notation as in the previous section on purpose, since the VAE modeling will precisely be applied on speech/audio spectrograms. Note also that $\mathbf{X} = |\mathbf{S}|^2$ is a sampled power spectrogram, aka a periodogram, i.e. an estimate of the power spectral density (PSD) $\mathbb{E}[|\mathbf{S}|^2]$ built from a single observation of the data in each time-frequency bin (and the same for the magnitude spectrogram).

3.2. Data representation, pre-processing and normalization

A VAE considers vectors as input and output. Hence an STFT spectrogram is processed as a sequence of successive spectral vectors $\mathbf{x}_n = [x_{fn}]_{f=0}^{F-1} \in \mathbb{R}_+^F$, each vector representing an STFT frame. Note that all x_{fn} are assumed independent across frequency bins and time frames, which is not to be confused with possible time-frequency structuration of the distribution parameters. An important practical question in VAEs is the choice of the audio STFT data representation. We did not observe any consensus in the literature.

For synthesis and transformation applications, e.g. [6], the observed/generated vector at time frame n generally corresponds to the short-term magnitude or power spectrum. There may be two explanations for that: (i) the original VAE formulation of [2] (i.e. the Gaussian models in (2) and (3)) considers real-valued and not complex-valued vectors, but in that case what about the non-negativity? and (ii) the magnitude or power spectrogram is the primary information used in the synthesis/transformation applications considered in the referenced papers (the phase spectrogram being processed separately).

For speech enhancement applications, the VAE speech model is generally plugged in a more general statistical framework including a noise model and a speech + noise mixture model, e.g. [9, 10]. In this framework, the original (real-valued) formulation of the VAE has been extended to model the complex-valued STFT vector $\mathbf{s}_n = [s_{fn}]_{f=0}^{F-1} \in \mathbb{C}^F$. This has been done by replacing the Gaussian distribution over real-valued vectors in (2) with the circularly symmetric complex Gaussian distribution that is widely used in speech enhancement and source separation probabilisitic methods [26, 27]. This important point is poorly commented in the referenced papers. Moreover, although $\mathbf{s}_n$ is here modeled by the VAE decoder, $\mathbf{x}_n$ as a short-term magnitude or power spectrum is still considered at the input of the encoder during VAE training.[1] The possible consequences (or absence of consequences) of this input/output mismatch are not discussed either. Note that here also, all s_{fn} are assumed independent across frequency bins and time frames, as is usually done in the speech enhancement and source separation literature.

It is important to note that in practice, the encoder input vector can contain magnitudes or squared magnitudes as discussed above, but also log-magnitudes as in [4], or actually any vector encoding a magnitude spectrum, possibly pre-processed and normalized in different manners. Normalization is a typical example of DNN-driven process, it has no theoretical justification from the signal processing point-of-view but it is known as helping a DNN training in general. So it is applied very frequently, and actually on purpose in VAEs. Also, the encoder input vector can be of different nature than the VAE decoder output vector, which is composed of probability distribution parameters; not to be confused with the output of the VAE as a generative model. Some of the output parameters may be homogeneous to the input data, e.g. mean vectors, and some others may not be, e.g. variance parameters. Moreover, data normalization can also be applied to output data, and the normalization/denormalization can be conducted in different manners at the input and at the output. Then, does data representation, pre-processing and normalization have any consequence on the theoretical foundations of the model?

[1] For speech enhancement applications, the encoder is only used for VAE training. During the speech signal inference process, only the decoder is used.

3.3. Statistical modeling and implications for VAE training

The choice of the reconstruction term of the loss function for the VAE training is often poorly discussed in papers dealing with VAE-based spectrogram modeling. A typical yet poorly justified approach could be: Let us choose a data representation that is appropriate for the considered application, for example a magnitude spectrum vector $\mathbf{x}_n$, and let us apply some normalization that is appropriate for DNNs. Then systematic application of the Gaussian model (2) is the easy way, leading to the weighted squared error form in the reconstruction term of (8). If we further set the variance parameters $\sigma^2_{\theta,f}(\mathbf{z}_n)$ to an arbitrarily fixed value σ^2 (i.e. we consider only the mean parameters $\mu_{\theta,f}(\mathbf{z}_n)$ as the free VAE outputs), then (8) becomes (up to an additive constant factor):

$$\mathcal{L}(\phi,\theta,\mathbf{X}) = -\frac{1}{\sigma^2}\sum_{n=1}^{N_{tr}}\sum_{f=0}^{F-1}\frac{1}{2}\big(x_{fn}-\mu_{\theta,f}(\mathbf{z}_n)\big)^2 + \frac{1}{2}\sum_{n=1}^{N_{tr}}\sum_{l=1}^{L}\Big(\log\tilde{\sigma}^2_{\phi,l}(\mathbf{x}_n) - \tilde{\mu}_{\phi,l}(\mathbf{x}_n)^2 - \tilde{\sigma}^2_{\phi,l}(\mathbf{x}_n)\Big) \quad (9)$$

This means that using the basic mean squared error (MSE) as the reconstruction term of the VAE loss function amounts to maximize the likelihood function under the present "fixed-variance free-mean" Gaussian model, hence providing some nice theoretical interpretation of the process. Yet this interpretation is poorly discussed in the papers. Does this approach have limitations? Does it make sense to model normalized magnitude vectors with a Gaussian distribution? Do other strategies exist? And what is the link with the problem of data representation?

As briefly mentioned in the introduction, a consistent theoretical framework exists that enables one to justify and interpret the choice of data representation, likelihood function and reconstruction term of the loss function, and how those points are related. This is what we present in the next section.

4. LINKING NMF AND VAE

In this section, we build on the existing statistical framework related to nonnegative representations, in particular Nonnegative Matrix Factorization (NMF), and its application to the modeling of speech/audio spectrograms. Most of the technical material presented here is extracted from [16] and [17]. We first shortly present the principle of NMF decomposition, then we go to the major point of this section which is to show that the underlying statistical framework directly applies to the VAE model, and can thus be used to give a solid theoretical interpretation of VAE-based modeling of speech/audio spectrograms. We finally report the three major NMF-based generative models considered in [16] and [17] and give their VAE counterparts.

4.1. The NMF model

NMF consists in modeling a matrix $\mathbf{V} = [v_{fn}]_{f,n} \in \mathbb{R}_+^{F\times N}$ of nonnegative entries as the product of two nonnegative matrices $\mathbf{W} = [w_{fk}]_{f,k} \in \mathbb{R}_+^{F\times K}$ and $\mathbf{H} = [h_{kn}]_{k,n} \in \mathbb{R}_+^{K\times N}$. In other words we have $\mathbf{V} \approx \widehat{\mathbf{V}} = \mathbf{WH}$, or equivalently $\widehat{v}_{fn} = (\mathbf{WH})_{fn} = \sum_{k=1}^{K} w_{fk}h_{kn}$. A low-rank approximation of $\mathbf{V}$, represented with a reduced number of parameters, is obtained by setting K such that $K(F+N) \ll FN$. In the speech/audio processing literature, $\widehat{\mathbf{V}}$ is typically used to model the signal ("true" or

"theoretical") PSD $\mathbb{E}[|\mathbf{S}|^2]$ based on the observed power spectrogram $\mathbf{X} = |\mathbf{S}|^2$ (or the same for the "true" magnitude spectrogram based on the observed magnitude spectrogram $\mathbf{X} = |\mathbf{S}|$). The interest of this approach is thus to provide a model of the signal PSD in each time-frequency bin with a very reasonable number of parameters (if K is chosen properly).

Calculating $\widehat{\mathbf{V}}$ from a given observed nonnegative matrix $\mathbf{X}$ is done by minimizing over $\mathbf{W}$ and $\mathbf{H}$ the following error under a non-negativity constraint:

$$D(\mathbf{X}|\widehat{\mathbf{V}}) = \sum_{n=1}^{N} \sum_{f=0}^{F-1} d(x_{fn}|\hat{v}_{fn}), \tag{10}$$

where $d(\cdot|\cdot)$ is a scalar divergence. The three most popular cost functions are the squared Euclidian distance $d_{\text{EUC}}(x|y) = 0.5(x-y)^2$, the generalized Kullback-Leibler (KL) divergence $d_{\text{KL}}(x|y) = x\log(x/y) - x + y$, and the Itakura-Saito (IS) divergence $d_{\text{IS}}(x|y) = x/y - \log(x/y) - 1$. For each of them, a set of algorithms have been proposed to solve the above minimization problem. Their presentation is out of the scope of this paper, where we focus on the link with the VAE and the underlying statistical models. For the same reason, we do not deal with the interpretation of NMF as a model of composite signals [16, 17], which is of primary importance in the source separation literature.

4.2. Linking NMF- and VAE-based spectrogram modeling

Now the major point of the present paper is the following: *The minimization of the global cost function* (10)*, the choice of the scalar cost function in* (10)*, the choice of data representation, and the interpretation in terms of underlying statistical model are problems that are all common to NMF and VAE.* In other words, *a common framework exists where* $\widehat{\mathbf{V}}$ *may as well be an NMF model* $\widehat{\mathbf{V}} = \mathbf{WH}$ *or the concatenation of successive (nonnegative) output vectors of a VAE*, e.g. $\widehat{\mathbf{V}} = [\boldsymbol{\sigma}_\theta^2(\mathbf{z}_1), \boldsymbol{\sigma}_\theta^2(\mathbf{z}_2), \cdots, \boldsymbol{\sigma}_\theta^2(\mathbf{z}_N)]$, which is the case for VAE-based spectrogram modeling. Indeed, as will be detailed below, for both NMF and VAE models, (10) is nothing but a reformulation of the negative log-likelihood function of the underlying generative model. More specifically, if $\widehat{\mathbf{V}}$ is the output of a VAE, the reconstruction accuracy in (7) and the cost function (10) are identical up to a constant multiplicative positive factor α, sign, and a constant additive factor. In short, (7) can be rewritten as:

$$\begin{aligned}\mathcal{L}(\phi, \theta, \mathbf{X}) = -\alpha &\sum_{n=1}^{N_{tr}} \sum_{f=0}^{F-1} d(x_{fn}|\hat{v}_{fn}) \\ &- \sum_{n=1}^{N_{tr}} d_{\text{KL}}(q_\phi(\mathbf{z}_n|\mathbf{x}_n)|p_\theta(\mathbf{z}_n)).\end{aligned} \tag{11}$$

In the VAE model framework, minimization of (10) *thus amounts to optimal estimation of the VAE parameters in the maximum-likelihood (ML) sense*. Let us temper a bit: (10) only concerns the VAE decoder, and the complete VAE is actually optimized by maximizing (7) (or (11)), i.e. the combination of (10) with the VLB regularization term. This latter is important to differentiate a VAE from a deterministic AE. Let us note that in the VAE framework, ML estimation of $\widehat{\mathbf{V}}$ is to be understood as a shortcut for ML estimation of θ, the decoder parameters, which requires the joint estimation of the encoder parameters ϕ during the VAE training. Finally, let us also note that α plays the role of balancing factor between reconstruction and regularization, and quite interestingly, it is very similar to the β factor of the β-VAE model proposed in [28] in an ad-hoc manner, for the same aim (though β is applied to the regularization term instead of the reconstruction term).

Although all these points may sound trivial to readers familiar with the statistical interpretation of NMF spectrogram modeling, to our knowledge they have never been pointed out in the literature on VAE-based speech/audio processing. One reasonable explanation for this may be that NMF studies often start with the cost function formulated as (10), and the interpretation in terms of underlying generative model comes in second (when it comes), whereas VAE studies start with a generative model then go to the cost function formulated as (7).

4.3. Practical cases

We now apply the above considerations to the three major cases considered in [16] and [17], which correspond to different divergences $d(\cdot|\cdot)$ in (10) and (11).

Euclidian distance case In the NMF context, it has been shown in [16, 17] that choosing and minimizing the squared Euclidian distance between $\mathbf{X}$ and $\widehat{\mathbf{V}} = \mathbf{WH}$ corresponds to ML estimation of $\mathbf{W}$ and $\mathbf{H}$ under the assumption of the Gaussian model

$$x_{fn} \sim \mathcal{N}(x_{fn}; \widehat{v}_{fn}, \sigma^2), \tag{12}$$

with $\widehat{v}_{fn} = (\mathbf{WH})_{fn} = \sum_{k=1}^{K} w_{fk}h_{kn}$. Similarly, in the VAE case, choosing and minimizing the squared Euclidian distance between x_{fn} and $\widehat{v}_{fn}$ in (11), with $\widehat{v}_{fn} = \mu_{\theta,f}(\mathbf{z}_n)$, corresponds to ML estimation of $\widehat{v}_{fn}$ under the assumption of the Gaussian model (2), with a fixed variance $\sigma^2_{\theta,f}(\mathbf{z}_n) = \sigma^2, \forall(f,n)$. Actually this is what we have already done at the end of Section 3, and formalized in (9). In both NMF and VAE cases, we have the following underlying model:

$$x_{fn} = \widehat{v}_{fn} + e_{fn}, \tag{13}$$

where e_{fn} is an i.i.d. additive white Gaussian noise, i.e. $e_{fn} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Moreover, identifying (11) and (9) leads to $\alpha = 1/\sigma^2$, hence σ^2 plays the role of balancing factor between reconstruction and regularization.

A Gaussian model is often favored because of its generality and its nice features in mathematical derivations. For instance, it has been used for VAE-based speech spectrogram modeling in [4, 6, 11, 29]. However, although this approach could work quite well in many settings, it suffers from what is referred to as an interpretation ambiguity in [16]: Although x_{fn} represents a magnitude or power spectrum, $\mathcal{N}(x_{fn}; \mu_{\theta,f}(\mathbf{z}_n), \sigma^2)$ may produce negative data (even if we somehow enforce $\mu_{\theta,f}(\mathbf{z}_n) \geq 0$). This problem may be partly fixed by appropriate data normalization (e.g. min-max rescaling within $[-1, 1]$) and/or with log-scaling. However, it is subject to discussion if the distribution of log-magnitude spectra of real-world speech and audio signals has a Gaussian shape or not.

Itakura-Saito divergence case Alternately, it was shown and largely discussed in [17] that using the IS divergence in (10) corresponds to maximizing the log-likelihood function under the assumption of a Gamma distribution for x_{fn}. More precisely, the statistical model is:

$$x_{fn} \sim \mathcal{G}(x_{fn}; \alpha, \alpha/\widehat{v}_{fn}), \tag{14}$$

where $\mathcal{G}(.;a,b)$ is the Gamma distribution with shape parameter $a>0$ and rate parameter $b>0$, and whose pdf is defined in the Appendix. In the NMF framework we have $\widehat{v}_{fn}=(\mathbf{WH})_{fn}=\sum_{k=1}^{K}w_{fk}h_{kn}$, but this result is still valid in the VAE framework where we now have $\widehat{v}_{fn}=\sigma^2_{\theta,f}(\mathbf{z}_n)$. In both NMF and VAE cases, we have the following underlying model:

$$x_{fn}=\widehat{v}_{fn}e_{fn}, \tag{15}$$

where e_{fn} is an i.i.d. multiplicative Gamma noise, i.e. $e_{fn}\overset{i.i.d.}{\sim}\mathcal{G}(e_{fn};\alpha,\alpha)$.

Importantly, it was also shown in [17] that if x_{fn} corresponds to a linear-scale squared magnitude, minimizing the IS divergence corresponds to ML estimation of $\widehat{v}_{fn}$ under a circularly symmetric complex Gaussian model for the STFT coefficients $s_{fn}\in\mathbb{C}$ corresponding to $x_{fn}=|s_{fn}|^2\in\mathbb{R}_+$, with a variance $\mathbb{E}[|s_{fn}|^2]$ equal to $\widehat{v}_{fn}$. In short, $s_{fn}\sim\mathcal{N}_c(s_{fn};0,\widehat{v}_{fn})$, where the pdf of the complex Gaussian distribution $\mathcal{N}_c$ is defined in the Appendix. This interpretation is quite important since this model and associated ML fitting procedure have been used extensively in speech enhancement and speech/audio source separation, in combination with NMF, e.g. [19, 21, 23], or not, e.g. [26, 20, 30]. Indeed, in such applications, we are interested in inferring the complex-valued source STFT coefficients s_{fn} from corrupted observations. Again, this result is valid for both NMF and VAE frameworks: In IS-based NMF, we have $\mathbb{E}[|s_{fn}|^2]=\mathbb{E}[x_{fn}]=\widehat{v}_{fn}=\sum_{k=1}^{K}w_{fk}h_{kn}$. In IS-based VAE, we have $\mathbb{E}[|s_{fn}|^2]=\mathbb{E}[x_{fn}]=\widehat{v}_{fn}=\sigma^2_{\theta,f}(\mathbf{z}_n)$ and the mean parameters $\mu_{\theta,f}(\mathbf{z}_n)$ are simply disregarded since (2) is implicitly replaced with the above Gamma model of x_{fn}. Note that IS-VAE was shown to outperform IS-NMF for speech enhancement in [10].

Generalized Kullback-Leibler divergence case Finally, minimizing the KL divergence between x_{fn} and $\widehat{v}_{fn}$ corresponds to ML estimation of $\widehat{v}_{fn}$ under the assumption of a Poisson distribution for x_{fn}:

$$x_{fn}\sim\mathcal{P}(x_{fn};\widehat{v}_{fn}), \tag{16}$$

where $\mathcal{P}(.;\lambda)$ is the Poisson distribution with scale parameter $\lambda>0$ and whose pdf is defined in the Appendix. Note that there is here no equivalent model in terms of additive or multiplicative noise. In theory, the Poisson distribution is defined for nonnegative integer-valued random variables, but this issue can be fixed by considering high-resolution fixed-point quantization of the spectrograms. As above, this result is valid for both NMF and VAE models. Here, $\widehat{v}_{fn}$ plays the role of a scale parameter, hence in principle the output of a KL-based VAE is a vector of scale parameters $\widehat{v}_{fn}=\sigma_{\theta,f}(\mathbf{z}_n)$ for $f=0,...,F-1$. Although, as stated above, arbitrary normalization and corresponding denormalization can be applied. Historically, KL-based NMF has been applied on (linear-scale) magnitude spectra instead of power spectra, see the seminal papers [31, 32], but in fact there is no underlying model on the complex-valued STFT coefficients s_{fn} to support this principle. In other words, in most papers on KL-based NMF, $\widehat{v}_{fn}$ is a scale parameter over magnitude spectra, because x_{fn} is a magnitude spectra, but it could as well be a scale parameter over a different representation. Of course, the same remark applies to a KL-based VAE.

In summary, in the speech/audio spectrogram NMF modeling framework, we had:

- EUC-NMF: $p_\theta(\mathbf{X}|\mathbf{Z})=\prod_{f,n}\mathcal{N}(x_{fn};(\mathbf{WH})_{fn},\sigma^2)$;
- IS-NMF: $p_\theta(\mathbf{X}|\mathbf{Z})=\prod_{f,n}\mathcal{G}(x_{fn};\alpha,\alpha/(\mathbf{WH})_{fn})$ and $p_\theta(\mathbf{S}|\mathbf{Z})=\prod_{f,n}\mathcal{N}_c(s_{fn};0,(\mathbf{WH})_{fn})$ with $x_{fn}=|s_{fn}|^2$;
- KL-NMF: $p_\theta(\mathbf{X}|\mathbf{Z})=\prod_{f,n}\mathcal{P}(x_{fn};(\mathbf{WH})_{fn})$.

In the VAE framework we have:

- EUC-VAE: $p_\theta(\mathbf{X}|\mathbf{Z})=\prod_{f,n}\mathcal{N}(x_{fn};\mu_{\theta,f}(\mathbf{z}_n),\sigma^2)$;
- IS-VAE: $p_\theta(\mathbf{X}|\mathbf{Z})=\prod_{f,n}\mathcal{G}(x_{fn};\alpha,\alpha/\sigma^2_{\theta,f}(\mathbf{z}_n))$ and $p_\theta(\mathbf{S}|\mathbf{Z})=\prod_{f,n}\mathcal{N}_c(s_{fn};0,\sigma^2_{\theta,f}(\mathbf{z}_n))$ with $x_{fn}=|s_{fn}|^2$;
- KL-VAE: $p_\theta(\mathbf{X}|\mathbf{Z})=\prod_{f,n}\mathcal{P}(x_{fn};\sigma_{\theta,f}(\mathbf{z}_n))$.

4.4. A practical note on the implementation of the VAE loss function

The above considerations have a practical consequence in the coding of the loss function when implementing a VAE with a deep learning library. Indeed, in practice, as stated above, input/output data are often pre-processed (e.g. log-scaled) and/or normalized to facilitate the VAE training. For the statistical interpretation considered in this paper to hold, the reconstruction term of the VAE loss function, as implemented in a deep learning toolkit, must have the form of the log-likelihood function $\log p_\theta(\mathbf{x}|\mathbf{z})$, and the data used in this loss function must be consistent with the model, i.e. if they have been previously normalized, then *they must be denormalized*. Using the normalized data would break the consistency of the underlying statistical model.

Let us give an example, by considering the Gamma model in (14) for the squared STFT magnitudes $x_{fn}=|s_{fn}|^2$. This model *implies that we have to* use the IS divergence in the reconstruction term of the loss function in (11). At training time, the VAE is fed with pre-processed/normalized data $x_{fn}^{\text{norm}}=g(x_{fn})$ and it provides pre-processed/normalized scale parameters $\widehat{v}_{fn}^{\text{norm}}=\tilde{g}(\widehat{v}_{fn})$. Note that the pre-processing/normalization of data and parameters may be different, as denoted by the different $g(\cdot)$ and $\tilde{g}(\cdot)$ functions. Then the implementation of the reconstruction term of the loss function based on the IS divergence and "applied to" x_{fn}^{norm} and $\widehat{v}_{fn}^{\text{norm}}$ should be of the form:

$$\frac{g^{-1}(x_{fn}^{\text{norm}})}{\tilde{g}^{-1}(\widehat{v}_{fn}^{\text{norm}})}-\log\frac{g^{-1}(x_{fn}^{\text{norm}})}{\tilde{g}^{-1}(\widehat{v}_{fn}^{\text{norm}})}-1=d_{\text{IS}}(x_{fn}|\widehat{v}_{fn}). \tag{17}$$

The denormalized outputs $\widehat{v}_{fn}=\tilde{g}^{-1}(\widehat{v}_{fn}^{\text{norm}})$ are then "automatically" homogeneous to scale parameters. In contrast, using directly the normalized values in the above reconstruction term (i.e. calculating $d_{\text{IS}}(x_{fn}^{\text{norm}}|\widehat{v}_{fn}^{\text{norm}})$) or using another distance (e.g. the MSE) on either the normalized or denormalized data would not be consistent with the Gamma model considered in this example.

5. EXPERIMENTS

In this section, we briefly present the results of experiments that were conducted to illustrate our discussion. We processed VAE-based analysis-synthesis of sound spectrograms for the three cases described in Section 4. Waveform resynthesis was done by combining the output magnitude spectrogram with the phase spectrogram of the original signal. We applied this on speech signals (TIMIT dataset [33], 10 utterances $\times$ 462 speakers in the training set, for a total of about 4h, and 10 different utterances $\times$ 168 different speakers in the test set, for a total of about 1.5h) and music

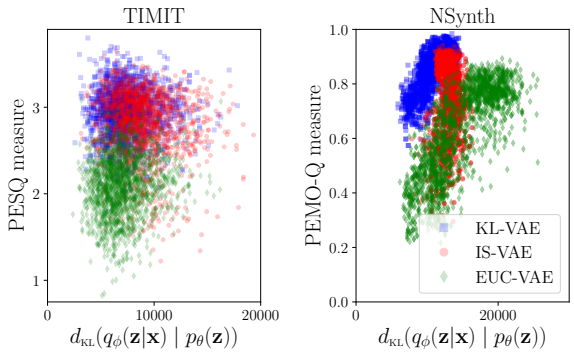

Figure 1: Audio quality as a function of the regularization term of (5).

signals (a subset of the large NSynth dataset [34], 88 notes with 4 different velocities from 17 instruments for the training set and 3 instruments for the test set all from the acoustic keyboards family, for a total of 9h of signals) at a 16 kHz sampling rate. The STFT was computed using a 64-ms sine window ($F = 513$) and a 75% overlap.

The VAE decoder network contains three layers of size [64, 128, 513] and the encoder network is the symmetric. Both networks use $\tanh$ and identity activation functions for the hidden and output layers respectively. The output of the encoder and decoder networks are thus real-valued, and as proposed in the original paper on VAEs [2], we output the logarithm of variance/scale parameters for the IS-VAE and KL-VAE cases. At the input of the encoder, we provide either magnitude spectrograms (KL-VAE and EUC-VAE) or power spectrograms (IS-VAE).

The results are plotted in Fig. 1. In order to measure the quality of the reconstructed signal independently of the nature of the cost function, PESQ scores [35] (for speech) and PEMO-Q scores [36] (for music) were calculated on the resynthesized signals in the test set. These scores are plotted in Fig. 1 as a function of the regularization term of (5). Each point represents either a utterance (left) or a music note (right) from the dataset. We set $\alpha = 0.1$ in (7) for the IS-VAE, and $\alpha = 1$ for both EUC-VAE and KL-VAE. This was to ensure (i) to keep a sufficiently small regularization term in the loss function so that VAEs are not turning into a deterministic autoencoders, and (ii) to obtain the same range of regularization term values for the 3 cost functions, so that the performance can be fairly compared in terms of reconstruction quality. We can see in Fig. 1 that for music signals (PEMO-Q scores) KL-VAE globally performs the best, followed by IS-VAE (with an overlapping zone of equal performance). For speech signals (PESQ scores), KL-VAE and IS-VAE are providing similar results. EUC-VAE generally provides lower scores.

6. CONCLUSION

We can now draw the following conclusions:

- The three presented cost functions usable for NMF or VAE modeling all correspond to an underlying statistical model of processed spectrogram $\mathbf{X} = [\mathbf{x}_n]_{n=1}^N$. For all three cases, training the VAE with data $\mathbf{X}$ corresponds to ML estimation of VAE decoder parameters under the corresponding statistical model of $\mathbf{X}$.
- Among these three cases, only one (IS-case) has an underlying statistical model of the speech/audio signal STFT coefficient s_{fn} (circularly symmetric complex Gaussian), which has proven to be of great interest for speech enhancement and source separation applications.
- The reconstruction accuracy and regularization of the VAE can be weighted using the α factor in (11). For EUC-VAE and IS-VAE this factor is naturally emerging as a parameter of the underlying statistical model, which provides a nice alternative (or interpretation) to the ad-hoc definition of the similar β factor introduced in [28]. This is not the case for KL-VAE, where $\alpha = 1$. For the interpretation of IS-VAE in terms of complex Gaussian model on s_{fn} to hold, we must also have $\alpha = 1$.
- In our experiments, KL-VAE and IS-VAE perform better than EUC-VAE according to perceptually-motivated objective measures.
- Although we necessarily presented this extension in the context of nonnegative representations, VAEs are not limited to nonnegative data. They can be applied to any real-valued data. This is what is done when processing log-scale spectrograms such as in [4]. The IS and KL divergences and associated Gamma and Poisson models are limited to nonnegative data, but the Euclidian distance and associated Gaussian model are not.
- In practice, input/output data are often pre-processed and/or normalized. If the pre-processed/normalized data are used in the VAE practical implementation, then the loss function should include denormalization and inverse pre-processing.
- All the points considered in this paper are valid for recurrent VAEs [37], which are likely to become popular in speech/audio processing as well. Also, generalization of NMF to more general divergences and corresponding statistical interpretation exist, e.g. [38, 39]. It is likely to be relevant for VAEs.

We spent time and effort to understand the correct form that a VAE loss function should have in a deep learning library to be consistent with a sounded signal statistical model. We believe that sharing the content of this paper (and code if the paper is accepted) with the speech/audio processing community can help colleagues to take VAEs into hand faster and in a principled manner. Also, we believe that the bridge we built in this paper can benefit to both the speech enhancement / source separation community and the musical sound processing community.

A. PROBABILITY DISTRIBUTIONS

A.1. Gaussian distributions

Let $\mathcal{N}(x; \mu, \sigma^2)$ denote the Gaussian distribution for a random variable $x \in \mathbb{R}$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+$. Its probability density function (pdf) is defined by:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \qquad (18)$$

Note that for simplicity we use the same notation to denote a probability distribution and its pdf.

Let $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ denote the multivariate Gaussian distribution for a real-valued random vector $\mathbf{x} \in \mathbb{R}^F$ of mean vector $\boldsymbol{\mu} \in \mathbb{R}^F$,

and with statistically independent entries such that $\boldsymbol{\sigma}^2 \in \mathbb{R}_+^F$ is the vector of variances (covariance terms are zero and thus omitted in the parametrization for simplicity). Its pdf is therefore equal to the product of univariate Gaussian pdfs:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \prod_{f=0}^{F-1} \mathcal{N}(x_f; \mu_f, \sigma_f^2), \tag{19}$$

where v_f denotes the f-th entry of a vector $\mathbf{v}$.

Let $\mathcal{N}_c(x; \mu, \sigma^2)$ denote the proper complex Gaussian distribution for a random variable $x \in \mathbb{C}$ with mean $\mu \in \mathbb{C}$ and variance $\sigma^2 \in \mathbb{R}_+$. Its pdf is defined by:

$$\mathcal{N}_c(x; \mu, \sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right). \tag{20}$$

This distribution is circularly symmetric (i.e. invariant to a phase shift for x) if $\mu = 0$

A.2. Gamma distribution

Let $\mathcal{G}(x; a, b)$ denote the Gamma distribution for a random variable $x \in \mathbb{R}_+$ with shape and rate parameters $a > 0$ and $b > 0$ respectively. Its pdf is defined by:

$$\mathcal{G}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \tag{21}$$

where $\Gamma(\cdot)$ is the Gamma function.

A.3. Poisson distribution

Let $\mathcal{P}(x; \lambda)$ denote the Poisson distribution for a random variable $x \in \mathbb{N}$ with rate parameter $\lambda > 0$. Its pdf is defined by:

$$\mathcal{P}(x; \lambda) = \exp(-\lambda)\frac{\lambda^x}{x!}. \tag{22}$$

7. REFERENCES

[1] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[2] D.P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Int. Conf. Learning Representations (ICLR)*, 2014.

[3] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders.," in *Conf. of the Int. Speech Comm. Association (Interspeech)*, San Francisco, CA, 2016.

[4] C.C. Hsu, H.T. Hwang, Y.C. Wu, Y. Tsao, and H.M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–6.

[5] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Conf. of the Int. Speech Comm. Association (Interspeech)*, Stockholm, Sweden, 2017.

[6] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Conf. of the Int. Speech Comm. Association (Interspeech)*, Hyderabad, India, 2018.

[7] F. Roche, T. Hueber, S. Limier, and L. Girin, "Autoencoders for music sound modeling: A comparison of linear, shallow, deep, recurrent and variational models," in *Sound and Music Computing Conference (SMC)*, Málaga, Spain, 2019.

[8] P. Esling, A. Chemla-Romeu-Santos, and A. Bitton, "Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, Paris, France, 2018.

[9] 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 Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Calgary, Canada, 2018.

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

[11] L. Pandey, A. Kumar, and V. Namboodiri, "Monaural audio source separation using variational autoencoders," in *Conf. of the Int. Speech Comm. Association (Interspeech)*, Hyderabad, India, 2018.

[12] S. Leglaive, U. Şimşekli, 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 Process. (ICASSP)*, 2019.

[13] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2018, pp. 1233–1239.

[14] L. Li, H. Kameoka, and S. Makino, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, Brighton, UK, 2019.

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

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

[17] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[18] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[19] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[20] N.Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[21] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Professionally-produced music separation guided by covers," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, Porto, Portugal, 2012.

[22] P. Smaragdis, C. Fevotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, 2014.

[23] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of moving sound sources," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NJ, USA, 2015.

[24] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2453–2465, 2016.

[25] S. Leglaive, R. Badeau, and G. Richard, "Student's t Source and Mixing Models for Multichannel Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1150–1164, 2018.

[26] 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.

[27] 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.

[28] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-vae: learning basic visual concepts with a constrained variational framework.," in *Int. Conf. on Learning Representations (ICLR)*, 2017.

[29] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection," in *Conf. of the Int. Speech Comm. Association (Interspeech)*, Hyderabad, India, 2018.

[30] X. Li, L. Girin, and R. Horaud, "An EM algorithm for audio source separation based on the convolutive transfer function," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NJ, USA, 2017.

[31] P. Smaragdis and J.C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NJ, 2003.

[32] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[33] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, and V. Zue, "TIMIT acoustic phonetic continuous speech corpus," in *Linguistic data consortium*, 1993.

[34] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," *arXiv preprint arXiv:1704.01279*, 2017.

[35] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *IEEE Int. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*, 2001, pp. 749–752.

[36] R. Huber and B. Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

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

[38] S. Sra and I.S. Dhillon, "Generalized nonnegative matrix approximations with bregman divergences," in *Advances in Neural Information Processing Systems*, 2006, pp. 283–290.

[39] A. Cichocki, S. Cruces, and S. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, 2011.