INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Complex-Valued Variational Autoencoder: A Novel Deep Generative Model for Direct Representation of Complex Spectra

Toru Nakashika

The University of Electro-Communications, Japan
nakashika@uec.ac.jp

Abstract

In recent years, variational autoencoders (VAEs) have been attracting interest for many applications and generative tasks. Although the VAE is one of the most powerful deep generative models, it still has difficulty representing complex-valued data such as the complex spectra of speech. In speech synthesis, we usually use the VAE to encode Mel-cepstra, or raw amplitude spectra, from a speech signal into normally distributed latent features and then synthesize the speech from the reconstruction by using the Griffin-Lim algorithm or other vocoders. Such inputs are originally calculated from complex spectra but lack the phase information, which leads to degradation when recovering speech. In this work, we propose a novel generative model to directly encode the complex spectra by extending the conventional VAE. The proposed model, which we call the complex-valued VAE (CVAE), consists of two complex-valued neural networks (CVNNs) of an encoder and a decoder. In the CVAE, not only the inputs and the parameters of the encoder and decoder but also the latent features are defined as complex-valued to preserve the phase information throughout the network. The results of our speech encoding experiments demonstrated the effectiveness of the CVAE compared to the conventional VAE in both objective and subjective criteria.

Index Terms: complex neural networks, deep learning, variational autoencoder, speech synthesis, speech encoding

1. Introduction

Deep learning has been enormously successful in the fields of image processing, speech signal processing, and more [1]. Recently, generative models such as generative adversarial networks (GANs) [2, 3], variational autoencoders (VAEs) [4, 5], and restricted Boltzmann machines (RBMs) [6, 7] have been attracting attention because they are more interpretable and require less labelled data than discriminative models.

The VAE is especially easy to implement and train and has a powerful representation ability. The VAE consists of an encoder that encodes the input into latent variables and a decoder that reconstructs the input from the latent variables in a probabilistic manner. Both the encoder and decoder usually stack multiple layers to represent high-order abstraction, which results in deep neural networks (DNNs). The most popular type of VAE assumes Gaussian-distributed latent variables as the posterior given inputs and the standard normal distribution as the prior. The latent variables can also be modeled as other distributions such as categorical distribution [8, 9], vector quantization (VQ) [10, 11], Gaussian mixture models (GMMs) [12, 13], and the von-Mises-Fisher distribution [14].

Although many variations of the VAE have been proposed so far, to the best of our knowledge there is not yet a VAE with a complex-valued variable prior. For the other machine learning models, various extensions that represent complex-valued data have been proposed [15, 16, 17]. There are still many cases where we need to deal with complex-valued actual data such as medical images, radar images, wireless signals, and acoustic intensity. In the speech community, typically used acoustic features such as MFCC, Mel-cepstra, Mel-spectra, and amplitude spectra are all calculated from the complex spectra of speech. In other words, these features lack phase information and can no longer represent the original complex spectra. Especially in speech synthesis, we need to estimate the phase by using the Griffin-Lim algorithm [18] or recover the signal from the amplitude-based acoustic features by using vocoders such as a Mel-log spectrum approximation (MLSA) filter [19], WORLD [20], STRAIGHT [21], or WaveNet [22], which results in degraded reconstruction of speech.

In this paper, we propose an extension of the VAE, which we called complex-valued VAE (CVAE), that can directly encode complex-valued spectra and learn the distribution of complex-valued latent variables. The encoder and decoder consist of complex neural networks [15] and output complex normal distributions of the latent variable and the observation, respectively. In addition, the CVAE imposes the standard complex normal distribution with zero mean, unit covariance, and zero pseudo-covariance as a prior of the latent variables. The KL divergence between the prior and the posterior of the latent variables can still be derived into a quite simple form. We also propose a reparameterization trick in the CVAE training, similar to the conventional VAE. As this does not involve implicit gradients [23], the gradients of the decoder can be directly propagated back to the encoder during the training.

Some studies have reported the use of VAEs for representing a distribution of complex-valued output data [24, 25, 26, 27, 28, 29]. These methods assume a zero-mean complex normal distribution whose variance parameters are output by a decoder, whereas in this paper, we propose a complete complex-valued VAE consisting of complex-valued output, latent variables, and weights of the DNN encoder and decoder.

In Section 2, we briefly review the conventional VAE. In Section 3, we present our proposed model, CVAE, and its reparameterization trick. In Section 4, we report our experimental results. We conclude in Section 5 with a brief summary.

2. Preliminary: VAE

The variational autoencoder (VAE) [4] is a generative model that defines two paired distributions $q_\phi(\boldsymbol{l}|\boldsymbol{x})$ and $p_\theta(\boldsymbol{x}|\boldsymbol{l})$ of H-dimensional latent variables $\boldsymbol{l} \in \mathbb{R}^H$ and D-dimensional observation $\boldsymbol{x} \in \mathbb{R}^D$, where ϕ and θ are model parameters of an encoder and a decoder, respectively. $q_\phi(\boldsymbol{l}|\boldsymbol{x})$ is actually an approximation of the real posterior distribution $p(\boldsymbol{l}|\boldsymbol{x})$. The encoder and decoder are typically composed of neural networks (NNs), and their parameters $\{\theta, \phi\}$ are estimated using the auto-encoding variational Bayes (AEVB) algorithm. Given the ob-

http://dx.doi.org/10.21437/Interspeech.2020-1964

servation $\boldsymbol{x}$, the lower bound of the log-likelihood can be found by using the Jensen's inequality, as

$$\begin{aligned}\log p(\boldsymbol{x}) &= D_{KL}(q_\phi(\boldsymbol{l}|\boldsymbol{x})||p_\theta(\boldsymbol{l}|\boldsymbol{x})) + \mathcal{L}(\theta,\phi;\boldsymbol{x}) \\ &\geq \mathcal{L}(\theta,\phi;\boldsymbol{x}) \triangleq \mathbb{E}_{q_\phi(\boldsymbol{l}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\boldsymbol{l})}{q_\phi(\boldsymbol{l}|\boldsymbol{x})}\right], \quad (1)\end{aligned}$$

where $D_{KL}(q||p)$ denotes the Kullback-Leibler (KL) divergence between distributions q and p. The lower bound $\mathcal{L}(\theta,\phi;\boldsymbol{x})$ can be further rewritten as

$$\begin{aligned}\mathcal{L}(\theta,\phi;\boldsymbol{x}) = &\mathbb{E}_{q_\phi(\boldsymbol{l}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{l})\right] \\ &- D_{KL}(q_\phi(\boldsymbol{l}|\boldsymbol{x})||p(\boldsymbol{l})). \quad (2)\end{aligned}$$

The first term on the right side of Eq. (2) indicates the expectation of the conditional log-likelihood of the observation given the latent variables that are encoded from the observation, while the second term indicates the constraint that the posterior and prior distributions of the latent variables are close to each other. From the point of view of an auxiliary function, the optimum $\{\theta,\phi\}$ that maximize the lower bound $\mathcal{L}(\theta,\phi;\boldsymbol{x})$ is also the reasonable solution for the log-likelihood $\log p(\boldsymbol{x})$. Therefore, in the VAE training, each parameter is optimized so as to maximize $\mathcal{L}(\theta,\phi;\boldsymbol{x})$ using the gradient method. However, it is difficult to back-propagate the gradients with respect to ϕ due to the sampling process. This is typically resolved by utilizing a reparameterization trick in the Gaussian case, as discussed later.

2.1. The VAE with Gaussian latent variables

There has been much prior research on the VAE adopt the Gaussian distribution as the posterior and prior of latent variables. In this case, the encoder NN outputs a concatenated vector of the mean $\boldsymbol{\mu} \in \mathbb{R}^H$ and the variance $\boldsymbol{\sigma} \in \mathbb{R}+^H$. In the forward pass of the VAE, we obtain reconstructed data $\boldsymbol{x}'$ as the output of the decoder NN that inputs a sample of latent variables $\tilde{\boldsymbol{l}}$:

$$\tilde{\boldsymbol{l}} \sim q_\phi(\boldsymbol{l}|\boldsymbol{x}) \triangleq \mathcal{N}(\boldsymbol{l};\boldsymbol{\mu},\Delta(\boldsymbol{\sigma})), \quad (3)$$

where $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma}))$ denotes the multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, and $\Delta(\cdot)$ is the function that returns a diagonal matrix whose diagonal elements are the input. This kind of VAE also imposes a standard Gaussian prior on the latent variable distribution as $p(\boldsymbol{l}) \triangleq \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. Therefore, the second term on the right side of Eq. (2) can be calculated analytically as:

$$\begin{aligned}D_{KL}(q_\phi(\boldsymbol{l}|\boldsymbol{x})||p(\boldsymbol{l})) &= D_{KL}(\mathcal{N}(\boldsymbol{\mu},\Delta(\boldsymbol{\sigma}))||\mathcal{N}(\boldsymbol{0},\boldsymbol{I})) \\ &= \frac{1}{2}(\boldsymbol{\mu}^\top\boldsymbol{\mu} + \|\boldsymbol{\sigma} - 1 - \log\boldsymbol{\sigma}\|_1).\end{aligned}$$

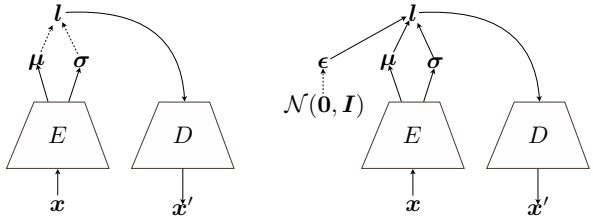
For continuous data such as Mel-cepstra, the conditional distribution of the data is also often modeled as Gaussian with a unit covariance, as

$$p_\theta(\boldsymbol{x}|\boldsymbol{l}) \triangleq \mathcal{N}(\boldsymbol{x};\boldsymbol{a},\boldsymbol{I}), \quad (4)$$

where $\boldsymbol{a} \in \mathbb{R}^D$ is the output of the decoder NN. In addition, the Monte Carlo method approximates an expectation $\mathbb{E}_{q_\phi(\boldsymbol{l}|\boldsymbol{x})}[f(\boldsymbol{l})]$ of L samples as

$$\mathbb{E}_{q_\phi(\boldsymbol{l}|\boldsymbol{x})}[f(\boldsymbol{l})] \approx \frac{1}{L}\sum_{i=1}^{L} f(\boldsymbol{l}^{(i)}). \quad (5)$$

Note that the approximation with only a sample $L = 1$ performs sufficiently as long as the minibatch size is large enough [4].



(a) w/o reparameterization (b) w/ reparameterization

Figure 1: *Comparison of forward pass of the vanilla VAE (a) without reparameterization and (b) with reparameterization. Solid and dotted lines indicate the forward pass and the sampling process where the gradients cannot be back-propagated, respectively.*

From the above, the first term on the right side of Eq. (2) can be made simpler:

$$\mathbb{E}_{q_\phi(\boldsymbol{l}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{l})\right] \approx -\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{a}\|_2^2 + K, \quad (6)$$

where K is a constant independent of model parameters.

2.2. Reparameterization trick

The model parameters are optimized to maximize the lower bound by using the gradient method. However, the gradients of the decoder cannot be back-propagated toward the encoder due to the sampling process of Eq. (3). To circumvent this, we utilize the reparameterization trick, as shown in Fig. 1. With a normal sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I})$, a sample of latent variables in Eq. (3) becomes equivalent to

$$\tilde{\boldsymbol{l}} = \boldsymbol{\mu} + \sqrt{\boldsymbol{\sigma}} \circ \boldsymbol{\epsilon}, \quad (7)$$

which is differential to the outputs of the encoder. Therefore, the gradients from the decoder can be back-propagated to the encoder.

3. Proposed model: CVAE

The VAE discussed above represents real-valued data due to the assumption of Gaussian-distributed observations. We can also properly represent binary data by assuming a Bernoulli distribution, which will change the loss function in Eq. (6) into a cross entropy. However, these models cannot represent complex-valued data correctly under the assumption of the distribution, although they can feed a concatenated vector of the real and imaginary parts of complex-valued data. We propose, therefore, a new generative model that directly represents complex-valued data through an encoder complex-valued neural network (CVNN) and a decoder CVNN, similarly to the VAE, as shown in Fig. 2. We call this model a complex-valued variational autoencoder (CVAE). Unlike the vanilla VAE, both the encoder and decoder of CVAE output the distribution of complex-valued variables.

Let $\boldsymbol{z} \in \mathbb{C}^D$ and $\boldsymbol{h} \in \mathbb{C}^H$ be complex-valued data and complex-valued latent variables, respectively. The same as the conventional VAE, the objective of the CVAE $\mathcal{L}(\theta,\phi;\boldsymbol{z})$ is the variational lower bound of the log-likelihood $\log p(\boldsymbol{z})$, as

$$\begin{aligned}\log p(\boldsymbol{z}) \geq &\mathcal{L}(\theta,\phi;\boldsymbol{z}) \quad (8) \\ = &\mathbb{E}_{q_\phi(\boldsymbol{h}|\boldsymbol{z})}\left[\log p_\theta(\boldsymbol{z}|\boldsymbol{h})\right] - D_{KL}(q_\phi(\boldsymbol{h}|\boldsymbol{z})||p(\boldsymbol{h})).\end{aligned}$$

Note that ϕ and θ, model parameters of the CVNN encoder and decoder, are all complex-valued here.

First, the CVAE defines the data conditional distribution as the multivariate complex normal distribution, as

$$p_\theta(\boldsymbol{z}|\boldsymbol{h}) \triangleq \mathcal{N}_c(\boldsymbol{z}; \boldsymbol{a}, \boldsymbol{\Gamma}, \mathbf{C}) \tag{9}$$

$$= \frac{1}{\pi^D \sqrt{\det(\boldsymbol{\Gamma})\det(\bar{\boldsymbol{\Gamma}} - \mathbf{C}^H \boldsymbol{\Gamma}^{-1} \mathbf{C})}} \tag{10}$$

$$\cdot \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{z}-\boldsymbol{a} \\ \bar{\boldsymbol{z}}-\bar{\boldsymbol{a}} \end{bmatrix}^H \begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{C} \\ \mathbf{C}^H & \boldsymbol{\Gamma}^H \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{z}-\boldsymbol{a} \\ \bar{\boldsymbol{z}}-\bar{\boldsymbol{a}} \end{bmatrix} \right\},$$

where $\boldsymbol{a} \in \mathbb{C}^D$, $\boldsymbol{\Gamma} \in \mathbb{C}^{D\times D}$, and $\mathbf{C} \in \mathbb{C}^{D\times D}$ denote the parameters of the complex normal distribution $\mathcal{N}_c(\cdot; \boldsymbol{a}, \boldsymbol{\Gamma}, \mathbf{C})$ of mean, covariance, pseudo-covariance, respectively. For simplicity, we use unit covariance and zero pseudo-covariance; i.e. $p_\theta(\boldsymbol{z}|\boldsymbol{h}) = \mathcal{N}_c(\boldsymbol{z}; \boldsymbol{a}, \boldsymbol{I}, \mathbf{O})$, and the decoder outputs only the mean $\boldsymbol{a}$ in this paper. This provides the following deformation in regards to the first term on the right side of Eq. (8):

$$\mathbb{E}_{q_\phi(\boldsymbol{h}|\boldsymbol{z})}\left[\log p_\theta(\boldsymbol{z}|\boldsymbol{h})\right] \approx -\|\boldsymbol{z} - \boldsymbol{a}\|_2^2 + K. \tag{11}$$

Second, the CVAE also defines the complex normal distribution on the latent variables to sample. As a simple but effective form, we assume the complex normal distribution with diagonal covariance and pseudo-covariance matrices, as

$$\tilde{\boldsymbol{h}} \sim q_\phi(\boldsymbol{h}|\boldsymbol{z}) \triangleq \mathcal{N}_c(\boldsymbol{h}; \boldsymbol{\mu}, \Delta(\boldsymbol{\sigma}), \Delta(\boldsymbol{\delta})), \tag{12}$$

where $\boldsymbol{\mu} \in \mathbb{C}^H$, $\boldsymbol{\sigma} \in \mathbb{R}+^H$, and $\boldsymbol{\delta} \in \mathbb{C}^H$ are the outputs of the encoder. As a prior of the latent variables, we assume the standard complex normal as $p(\boldsymbol{h}) \triangleq \mathcal{N}_c(\mathbf{0}, \boldsymbol{I}, \mathbf{O})$, which is one of the simplest and most representative distributions of complex random variables. As a result, the second term on the right side of Eq. (8) can still be computed in a simple and closed form, as

$$D_{KL}(q_\phi(\boldsymbol{h}|\boldsymbol{z})||p_\theta(\boldsymbol{h})) \tag{13}$$

$$= D_{KL}(\mathcal{N}_c(\boldsymbol{\mu}, \Delta(\boldsymbol{\sigma}), \Delta(\boldsymbol{\delta}))||\mathcal{N}_c(\mathbf{0}, \boldsymbol{I}, \mathbf{O})) \tag{14}$$

$$= \boldsymbol{\mu}^H \boldsymbol{\mu} + \|\boldsymbol{\sigma} - 1 - \frac{1}{2}\log(\boldsymbol{\sigma}^2 - |\boldsymbol{\delta}|^2)\|_1, \tag{15}$$

where $\cdot^2$ and $|\cdot|$ denote the element-wise square and absolute operations, respectively. The term of Eq. (14) indicates the constraint that makes the encoder outputs close to the simple standard complex normal while the pseudo-variance as well as the mean and the variance can change by different input $\boldsymbol{z}$.

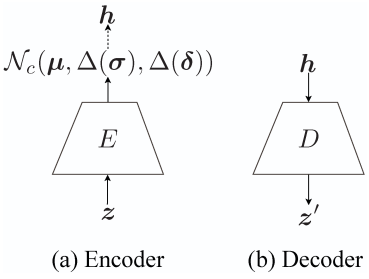
In this paper, we estimate the parameters of the CVAE $\{\phi, \theta\}$ by using the complex-valued gradient method so as to maximize Eq. (8). The simplest one is the complex-valued steepest ascent [30, 31], which iteratively updates each parameter with a complex-valued learning rate $\alpha \in \mathbb{C}, \Re(\alpha) > 0$, as

$$\theta^{(\text{new})} \leftarrow \theta^{(\text{old})} + \alpha \cdot 2\frac{\partial \mathcal{L}}{\partial \bar{\theta}}, \tag{16}$$

where the partial derivative in Eq. (16) is the Wirtinger derivative:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{1}{2}\left(\frac{\partial \mathcal{L}}{\partial \Re(\theta)} - i\frac{\partial \mathcal{L}}{\partial \Im(\theta)}\right). \tag{17}$$

The same is true of ϕ. In our experiments, we utilized the complex Adam [17] for more efficient learning.



(a) Encoder (b) Decoder

Figure 2: *The CVAE consists of (a) an encoder E that inputs complex-valued observation $\boldsymbol{z}$ and outputs the distribution of complex-valued latent variables, and (b) a decoder D that reconstructs the observation $\boldsymbol{z}'$ from the latent variables.*

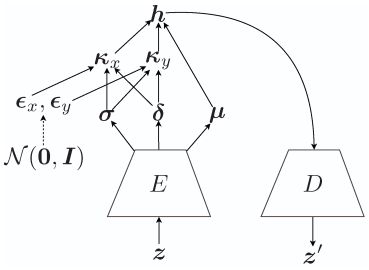


Figure 3: *Reparameterization trick in CVAE. The solid and dotted lines are forward pass and sampling process, respectively.*

3.1. Reparameterization trick in CVAE

As in the conventional VAE, the sampling process in Eq. (12) makes it impossible to back-propagate the gradients from the decoder side to the encoder. Therefore, we propose a reparameterization trick for the CVAE as shown in Fig. 3.

The complex-valued latent variables $\boldsymbol{h}$ can be decomposed into the real part $\boldsymbol{x} \in \mathbb{R}^H$ and the imaginary part $\boldsymbol{y} \in \mathbb{R}^H$ as $\boldsymbol{h} = \boldsymbol{x} + i\boldsymbol{y}$. Under the assumption of Eq. (12), the elements of $\boldsymbol{h}$ are independent of each other, and $\boldsymbol{x}$ and $\boldsymbol{y}$ follow the Gaussian distribution with the mean of $\Re(\boldsymbol{\mu})$ and $\Im(\boldsymbol{\mu})$ and the variance of $\boldsymbol{\sigma}_x \triangleq \frac{\boldsymbol{\sigma}+\Re(\boldsymbol{\delta})}{2}$ and $\boldsymbol{\sigma}_y \triangleq \frac{\boldsymbol{\sigma}-\Re(\boldsymbol{\delta})}{2}$, respectively. Because there are correlations $\boldsymbol{\rho} \triangleq \frac{\Im(\boldsymbol{\delta})}{\boldsymbol{\sigma}+\Re(\boldsymbol{\delta})}$ between $\boldsymbol{x}$ and $\boldsymbol{y}$, the latent variables follow the probability

$$\mathcal{N}\left(\Im(\boldsymbol{\mu}) + \boldsymbol{\rho} \circ \sqrt{\frac{\boldsymbol{\sigma}_y}{\boldsymbol{\sigma}_x}} \circ (\tilde{\boldsymbol{x}} - \Re(\boldsymbol{\mu})), (\mathbf{1} - \boldsymbol{\rho}^2) \circ \boldsymbol{\sigma}_y\right)$$

after we sample $\tilde{\boldsymbol{x}} = \Re(\boldsymbol{\mu}) + \boldsymbol{\sigma}_x \circ \boldsymbol{\epsilon}_x$ where $\boldsymbol{\epsilon}_x \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Therefore, we can sample $\tilde{\boldsymbol{y}}$ using another standard normal random variable $\boldsymbol{\epsilon}_y \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ as

$$\tilde{\boldsymbol{y}} = \Im(\boldsymbol{\mu}) + \boldsymbol{\rho} \circ \sqrt{\frac{\boldsymbol{\sigma}_y}{\boldsymbol{\sigma}_x}} \circ (\tilde{\boldsymbol{x}} - \Re(\boldsymbol{\mu})) + \sqrt{(\mathbf{1} - \boldsymbol{\rho}^2) \circ \boldsymbol{\sigma}_y} \circ \boldsymbol{\epsilon}_y,$$

where $\sqrt{\cdot}$ denotes the element-wise square. Summarizing the above, we can sample latent variables $\tilde{\boldsymbol{h}}$ as follows:

$$\tilde{\boldsymbol{h}} = \boldsymbol{\mu} + \boldsymbol{\kappa}_x \circ \boldsymbol{\epsilon}_x + \boldsymbol{\kappa}_y \circ \boldsymbol{\epsilon}_y \tag{18}$$

$$\boldsymbol{\kappa}_x \triangleq \frac{\boldsymbol{\sigma} + \boldsymbol{\delta}}{\sqrt{2\boldsymbol{\sigma} + 2\Re(\boldsymbol{\delta})}} \tag{19}$$

$$\boldsymbol{\kappa}_y \triangleq i\frac{\sqrt{\boldsymbol{\sigma}^2 - |\boldsymbol{\delta}|^2}}{\sqrt{2\boldsymbol{\sigma} + 2\Re(\boldsymbol{\delta})}}. \tag{20}$$

Table 1: *Experimental conditions of each method.*

	CVAE	**VAE(R+I)**	**VAE(GL)**
input features	complex spectra	real & imag of complex spectra	amplitude spectra
speech reconstruction	inverse STFT & OLA	inverse STFT & OLA	Griffin-Lim
no. of epochs	80	540	290
optimizer (learning rate)	CAdam (0.0001)	Adam (0.001)	Adam (0.001)
encoder architecture	255-100-[50,50,50]	510-200-[100,100]	255-100-[50,50]
decoder architecture	50-100-255	100-200-510	50-100-255

Table 2: *Experimental results of each method. Values after* $\pm$ *indicate the 95% confidential intervals.*

Method	PESQ	MOS
VAE(GL)	1.90	1.52 ± 0.08
VAE(R+I)	1.80	1.83 ± 0.10
CVAE	$\mathbf{2.44}$	$\mathbf{3.08} \pm 0.13$
CVAE(w/o $\boldsymbol{\delta}$)	2.39	–
Original	–	4.81 ± 0.04

As shown in Fig. 3, the gradients from the decoder can be back-propagated to the encoder side.

4. Experiments

4.1. Setup

To evaluate our proposed model, we conducted analysis-by-synthesis experiments using 50 sentence speech signals for training and another 53 for tests pronounced by a female announcer ("FTK") from set "B" of the ATR speech corpora [32]. The speech signals were downsampled from the original 20kHz to 16kHz and then processed into 255-dimensional complex spectra using the short-time Fourier transform (STFT) with a window length of 512 and a hop size of 64 as the input features. We used the remaining two-dimensional (first and last) real-valued spectra as originals in the generation stage. After we trained the CVAE, we evaluated it using the perceptual evaluation of speech quality (PESQ) and the 5-scale mean opinion score (MOS) of 11 participants, comparing it with two kinds of VAE: one that feeds the concatenated vector of real and imaginary parts of the complex spectra ("VAE(R+I)") and one that feeds the magnitude spectra and recovers signals using the Griffin-Lim ("VAE(GL)"). For each method, we stopped the training when the loss did not go down. CVAE and "VAE(R+I)" restored speech using the inverse STFT from the reconstructed complex spectra followed by the overlap-add (OLA) method. The number of Griffin-Lim iterations for "VAE(GL)" was 100. Table 1 summarizes the experimental setup. The notation "255-100-[50,50,50]," for example, indicates that the model has three fully connected layers having 255, 100, 50×3 units in order. Note that "VAE(R+I)" has twice as many units as the CVAE having two-degrees-of-freedom of real and imaginary units for a fair comparison.

4.2. Results and discussion

The CVAE significantly outperformed the two VAE methods in both objective and subjective criteria, as shown in Table 2. "VAE(R+I)" could not model the high frequencies very well, as shown in Fig. 4. In contrast, the CVAE generated superior

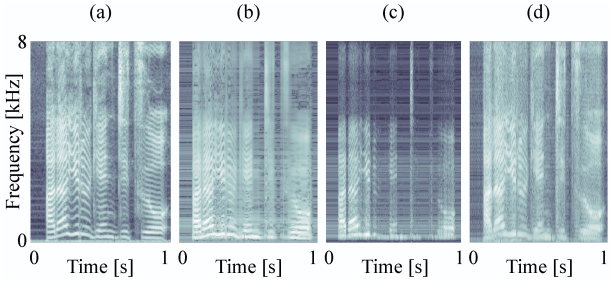


Figure 4: *(a) Original amplitude spectrum and the reconstructed spectra by (b) VAE(GL), (c) VAE(R+I), and (d) CVAE.*

complex spectra that had fine structures and formants. This is because the CVAE can capture the frequent complex spectral patterns due to its direct complex encoding system and the complex gradient method that keep the complex structures of the data. When we compare the results of the two conventional VAE methods, the performance of "VAE(GL)" was worse than that of "VAE(R+I)" in the MOS criterion, as the Griffin-Lim algorithm generates perceptually poor signals. For all methods, we feel that the performance could be improved by a deeper architecture, convolution layers, skip connections, and other techniques. This will be investigated in our future work.

As a reference, we also took a look at our method without using the pseudo-variance $\boldsymbol{\delta}$ as the output of the encoder (i.e., always $\boldsymbol{\delta} = \mathbf{0}$). The absence of $\boldsymbol{\delta}$ degraded the performance, as depicted in the "CVAE(w/o $\boldsymbol{\delta}$)" row in Table 2. This means that capturing the correlations between the real and imaginary parts of the latent variables is important in the CVAE.

5. Conclusion

In this paper, we proposed a new generative model, the CVAE, to directly represent complex spectra by extending the VAE. The CVAE is based on the assumption that the complex-valued latent variables follow the complex normal with diagonal covariance and pseudo-covariance matrices. We showed that the sampled complex-valued latent variable can be back-propagated by using our reparameterization trick. We demonstrated the effectiveness of CVAE through analysis-by-synthesis experiments. Our findings demonstrate that the CVAE has the potential to be just as a fundamental model as the VAE and can be applied to many tasks such as speech synthesis, voice conversion, source separation, and even image or other signal processing.

6. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 18K18069.

7. References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[3] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICML*, 2014.

[5] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[6] Y. Freund and D. Haussler, "Unsupervised learning of distributions of binary vectors using two layer networks," 1994.

[7] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," pp. 873–880, 2008.

[8] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-softmax," in *ICLR*, 2017.

[9] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *ICLR*, 2017.

[10] A. van den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *NeurIPS*, 2017, pp. 6306–6315.

[11] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion," *Interspeech*, pp. 724–728, 2019.

[12] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," in *ICLR*, 2017.

[13] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," in *IJCAI*, 2017, pp. 1965–1972.

[14] J. Xu and G. Durrett, "Spherical latent spaces for stable variational autoencoders," in *EMNLP*, 2018.

[15] I. Nemoto and T. Kono, "Complex neural networks," *Systems and computers in Japan*, vol. 23, no. 8, pp. 75–84, 1992.

[16] H. Kameoka, N. Ono, and K. Kashino, "Complex NMF: A new sparse representation for acoustic signals," *ICASSP*, pp. 3437–3440, 2009.

[17] T. Nakashika, S. Takaki, and J. Yamagishi, "Complex-valued restricted Boltzmann machine for speaker-dependent speech parameterization from complex spectra," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 244–254, 2018.

[18] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[19] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan*, vol. 66, no. 2, pp. 10–18, 1983.

[20] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[21] H. Kawahara, "STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[22] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[23] M. Figurnov, S. Mohamed, and A. Mnih, "Implicit reparameterization gradients," in *NeurIPS*, 2018, pp. 441–452.

[24] 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 *ICASSP*, 2018, pp. 716–720.

[25] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Supervised determined source separation with multichannel variational autoencoder," *Neural computation*, vol. 31, no. 9, pp. 1891–1914, 2019.

[26] S. Leglaive, L. Girin, and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," in *ICASSP*, 2019, pp. 101–105.

[27] L. Girin, F. Roche, T. Hueber, and S. Leglaive, "Notes on the use of variational autoencoders for speech and audio spectrogram modeling," in *DAFx*, 2019, pp. 1–8.

[28] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *APSIPA ASC*. IEEE, 2018, pp. 1233–1239.

[29] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *MLSP*, 2018, pp. 1–6.

[30] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proceedings F-Communications, Radar and Signal Processing*, vol. 130, no. 1, pp. 11–16, 1983.

[31] H. Zhang and D. P. Mandic, "Is a complex-valued stepsize advantageous in complex-valued gradient learning algorithms?" *IEEE transactions on neural networks and learning systems*, vol. 27, no. 12, pp. 2730–2735, 2015.

[32] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.