A DEEP GENERATIVE MODEL OF SPEECH COMPLEX SPECTROGRAMS

Aditya Arie Nugraha[⋆] *Kouhei Sekiguchi*[†⋆] *Kazuyoshi Yoshii*[†⋆]

[⋆] RIKEN Center for Advanced Intelligence Project (AIP), Japan
[†] Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

This paper proposes an approach to the joint modeling of the short-time Fourier transform magnitude and phase spectrograms with a deep generative model. We assume that the magnitude follows a Gaussian distribution and the phase follows a von Mises distribution. To improve the consistency of the phase values in the time-frequency domain, we also apply the von Mises distribution to the phase derivatives, i.e., the group delay and the instantaneous frequency. Based on these assumptions, we explore and compare several combinations of loss functions for training our models. Built upon the variational autoencoder framework, our model consists of three convolutional neural networks acting as an encoder, a magnitude decoder, and a phase decoder. In addition to the latent variables, we propose to also condition the phase estimation on the estimated magnitude. Evaluated for a time-domain speech reconstruction task, our models could generate speech with a high perceptual quality and a high intelligibility.

Index Terms— deep generative model, magnitude, phase, group delay, instantaneous frequency

1. INTRODUCTION

Speech signal processing methods typically work in the time-frequency (TF) domain, and the most widely used TF representation is the short-time Fourier transform (STFT) [1–4]. The complex-valued STFT coefficients are typically decomposed into the real-valued magnitude and phase spectrograms. Most signal processing methods focus on magnitude modification or estimation. However, an increasing number of works have shown that phase, including its derivatives, is useful to improve the performance of various applications [5, 6]. In this paper, we are interested in the problem of joint magnitude and phase estimation in the context of speech enhancement.

There exists works on phase recovery given the magnitude, including the consistency-based approach [7, 8], the sinusoidal signal model based approaches [9, 10], and the deep neural network (DNN) based approaches [11, 12]. Takamichi et al. [11] optimize the estimations of the phase and the group delay assuming a von Mises distribution for each of them. The method outperforms the Griffin-Lim algorithm [7] given the true magnitude. Takahashi et al. [12] discretize the phase and view the phase estimation as a classification problem. The method performs well for source separation tasks given the true magnitude and the imperfect magnitude estimate. Both approaches train the DNNs in a supervised manner.

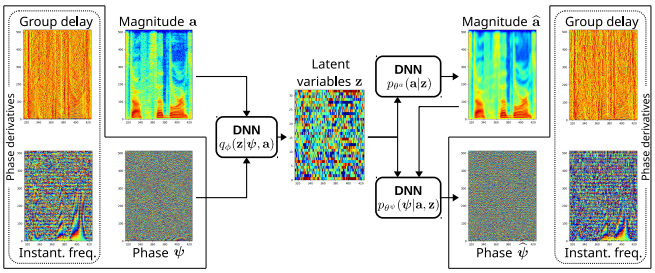


Fig. 1. Overview of the proposed model.

In contrast, a generative model assumes that some observation is generated by some latent variables, and the model learns those variables in an unsupervised manner. Several DNN-based generative models have been proposed recently, including the variational autoencoder (VAE) [13], the generative adversarial network [14], the flow-based model [15], and the autoregressive model, such as the WaveNet [16]. Among these models, the VAE is arguably the most popular. It has been used for various purposes, including speech separation [17, 18] and speech enhancement [19–21]. In these probabilistic approaches, the VAEs act as priors that allow an efficient estimation of the source power spectrograms. We could then employ a phase recovery method to estimate the source phase spectrograms. However, this cascading approach, i.e., a magnitude estimation followed by a phase recovery, is considered to be suboptimal. Therefore, we aim for a prior of the complex spectrogram. The WaveNet is possibly used to provide a time-domain prior. Nonetheless, we opt to work in the TF domain so that we can build upon the various speech enhancement methods that work in this domain [3].

Motivated by the success of VAEs in modeling the power spectrogram [17–21] and that of DNNs in estimating the phase given the magnitude [11, 12], we aim for a joint magnitude and phase deep generative model. Following the VAE framework, we define an encoder for estimating the latent variables given the magnitude and the phase spectrograms. We then define a magnitude decoder for reconstructing the magnitude given the latent variables, and a phase decoder for reconstructing the phase given the latent variables and the reconstructed magnitude. Figure 1 gives an overview of our model. We assume a Gaussian distribution for the magnitude, and a von Mises distribution for the phase and its derivatives, i.e., the group

delay (GD) and the instantaneous frequency (IF), each. The GD is the derivative along the frequency axis, and the IF is that along the time axis. Thus, there is an interdependence between the phase, the GD, and the IF that has to be satisfied. We explore different loss functions for training our models. The experimental results show that our models could reconstruct time-domain speech. The results also suggest that as long as the GD *and* the IF estimates are good, the phase estimates are not critical for obtaining a reasonable reconstruction.

The rest of this paper is organized as follows. Section 2 introduces the proposed approach. Section 3 presents the evaluation. Finally, Section 4 concludes this paper.

2. PROPOSED METHOD

Let $a_{f,n} \in \mathbb{R}_{\geq 0}$ and $\psi_{f,n} \in [-\pi, \pi)$ be the magnitude and the phase, respectively, of a complex-valued STFT spectrum $s_{f,n} = a_{f,n} e^{i\psi_{f,n}} \in \mathbb{C}$, where $f \in [1, F]$ and $n \in [1, N]$ are the frequency bin and the time frame indexes. For a generative modeling purpose, let us now assume that for a time frame n, the magnitude $\mathbf{a}_n = [a_{1,n}, \ldots, a_{F,n}]^\top$ and the phase $\boldsymbol{\psi}_n = [\psi_{1,n}, \ldots, \psi_{F,n}]^\top$ depend on latent variables $\mathbf{z}_n \in \mathbb{R}^D$ with $D < F$. Knowing that there is a generation process $p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n|\mathbf{z}_n)$ with some parameters θ, we want to maximize the joint probability $p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n)$.

2.1. Model formulation

We propose to represent the joint probability between the phase $\boldsymbol{\psi}_n$, the magnitude $\mathbf{a}_n$, and the latent variables $\mathbf{z}_n$ as

$$p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n) = p_{\theta^\psi}(\boldsymbol{\psi}_n|\mathbf{a}_n, \mathbf{z}_n) p_{\theta^a}(\mathbf{a}_n|\mathbf{z}_n) p_\theta(\mathbf{z}_n). \quad (1)$$

Note that $\boldsymbol{\psi}_n$ is conditioned on $\mathbf{a}_n$ and $\mathbf{z}_n$, while $\mathbf{a}_n$ is conditioned on $\mathbf{z}_n$ only. Following the VAE framework [13], to approximate the posterior $p_\theta(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)$, we introduce a variational inference process $q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n) \sim \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_n^q, (\boldsymbol{\sigma}_n^q)^2\mathbf{I})$ with parameters ϕ, where $\mathbf{I}$ is a D-dimensional identity matrix. We also assume a simple prior $p_\theta(\mathbf{z}_n) \sim \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{I})$.

We then obtain a VAE with an *encoder* $q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)$, and a decoder consisting of a *magnitude decoder* $p_{\theta^a}(\mathbf{a}_n|\mathbf{z}_n)$ and a *phase decoder* $p_{\theta^\psi}(\boldsymbol{\psi}_n|\mathbf{a}_n, \mathbf{z}_n)$. Thus, there are three DNNs to be trained. The combination of the encoder and the magnitude decoder is similar to the VAEs in [19,20]. Moreover, the phase decoder resembles the DNNs in [11, 12], that are trained in a supervised manner to estimate the phase given the magnitude spectrogram. In our work, the above three DNNs are jointly trained in an unsupervised manner. Using the encoder, we could estimate the latent variables $\mathbf{z}_n$ given some observations. Most importantly, we could obtain the complex-valued STFTs, reconstructed using the magnitude and the phase estimated by the decoder, given $\mathbf{z}_n$ sampled from the simple prior $p_\theta(\mathbf{z}_n)$.

2.2. Parameter estimation

The parameters could be jointly optimized by minimizing the negative log-likelihood (NLL) function:

$$\begin{aligned}
-\ln p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n) &= -\ln \int_{\mathbf{z}_n} p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n)\, \mathrm{d}\mathbf{z}_n \\
&\leq -\mathbb{E}_{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)} \left[\ln \frac{p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n)}{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)} \right] \\
&= \mathrm{KL}[q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)||p_\theta(\mathbf{z}_n)] \\
&\quad - \mathbb{E}_{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)} [\ln p_{\theta^a}(\mathbf{a}_n|\mathbf{z}_n)] \\
&\quad - \mathbb{E}_{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)} [\ln p_{\theta^\psi}(\boldsymbol{\psi}_n|\mathbf{a}_n, \mathbf{z}_n)], \quad (2)
\end{aligned}$$

where $\ln(\cdot)$ returns the natural logarithm, $\mathbb{E}[\cdot]$ returns the expectation, and $\mathrm{KL}[\triangle||\square]$ is the Kullback-Leibler divergence from $\square$ to $\triangle$ [22]. The first term is a regularization term $\mathcal{L}^{\text{reg}}$, the second term is a magnitude reconstruction loss $\mathcal{L}^{\text{mag}}$, and the third term is a phase reconstruction loss $\mathcal{L}^{\text{pha}}$.

The regularization term [13] is expressed as

$$\mathcal{L}^{\text{reg}} = \frac{1}{2N} \sum_{d,n} \left((\mu_{d,n}^q)^2 + (\sigma_{d,n}^q)^2 - \ln(\sigma_{d,n}^q)^2 - 1 \right), \quad (3)$$

where d is the latent variable dimension index.

The magnitude $a_{f,n}$ is assumed to follow a Gaussian distribution with mean $\mu_{f,n}^{\text{mag}} \in \mathbb{R}_{\geq 0}$ and variance $(\sigma_{f,n}^{\text{mag}})^2 \in \mathbb{R}_{\geq 0}$:

$$a_{f,n} \sim \mathcal{N}\left(a_{f,n} \middle| \mu_{f,n}^{\text{mag}}, \left(\sigma_{f,n}^{\text{mag}}\right)^2 \right). \quad (4)$$

The magnitude reconstruction loss is the NLL function:

$$\mathcal{L}^{\text{mag}} = \frac{1}{2N} \sum_{f,n} \left(\ln 2\pi \left(\widehat{\sigma}_{f,n}^{\text{mag}}\right)^2 + \frac{\left(a_{f,n} - \widehat{a}_{f,n}\right)^2}{\left(\widehat{\sigma}_{f,n}^{\text{mag}}\right)^2} \right), \quad (5)$$

where the estimate $\widehat{a}_{f,n}$ equals to the estimated mean $\widehat{\mu}_{f,n}^{\text{mag}}$. Additionally, we introduce a regularization term:

$$\mathcal{L}^{\text{var}} = \frac{1}{N} \sum_{f,n} \left(\widehat{\sigma}_{f,n}^{\text{mag}}\right)^2, \quad (6)$$

which enforces small variances for the distribution so that obtaining small estimation errors is more emphasized. Empirically, we observed that this term is crucial when we consider more loss components, i.e., the phase-related ones.

The phase $\psi_{f,n}$ is assumed to follow a von Mises distribution with mean $\mu_{f,n}^{\text{pha}} \in [-\pi, \pi)$ and concentration $\kappa_{f,n}^{\text{pha}} \in \mathbb{R}_{\geq 0}$:

$$\psi_{f,n} \sim \mathcal{VM}\left(\psi_{f,n} \middle| \mu_{f,n}^{\text{pha}}, \kappa_{f,n}^{\text{pha}} \right). \quad (7)$$

Several works have applied the same assumption for the phase [10,11,23]. The phase reconstruction loss is the NLL function:

$$\mathcal{L}^{\text{pha}} = \frac{1}{N} \sum_{f,n} \left(\ln 2\pi I_0\left(\widehat{\kappa}_{f,n}^{\text{pha}}\right) - \widehat{\kappa}_{f,n}^{\text{pha}} \cos\left(\psi_{f,n} - \widehat{\psi}_{f,n}\right) \right), \quad (8)$$

where the estimate $\widehat{\psi}_{f,n}$ is the estimated mean $\widehat{\mu}_{f,n}^{\text{pha}}$ and $I_0(\cdot)$ is the modified Bessel function of the first kind with order 0.

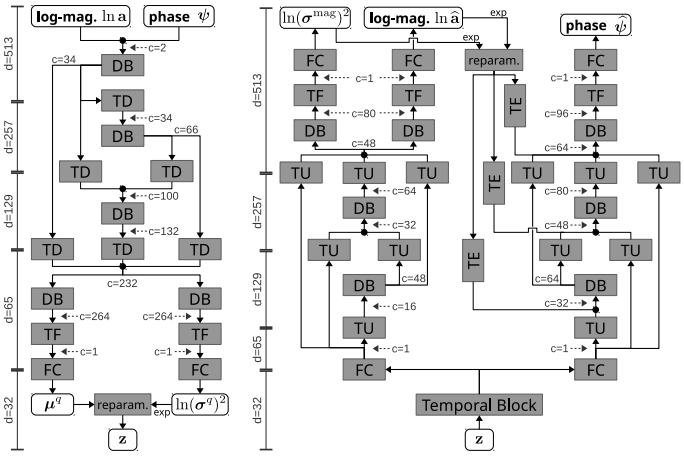
Furthermore, we consider the phase derivatives, i.e., the group delay [1] and the instantaneous frequency [24]. The group delay (GD) $\psi_{f,n}^{\text{grd}} \in [-\pi, \pi)$ is the phase derivative along the frequency axis:

$$\psi_{f,n}^{\text{grd}} = \text{wrap}(-\psi_{f+1,n} + \psi_{f,n}), \quad (9)$$

and the instantaneous frequency (IF) $\psi_{f,n}^{\text{ifr}} \in [-\pi, \pi)$ is the phase derivative along the time axis:

$$\psi_{f,n}^{\text{ifr}} = \text{wrap}(\psi_{f,n+1} - \psi_{f,n}), \quad (10)$$

where $\text{wrap}(\cdot)$ returns value in $[-\pi, \pi)$. Both derivatives capture the phase dynamics in the different axes.



(a) Encoder (b) Decoder

Fig. 2. Block diagram of the model architecture. The feature map dimension is (c, d, N), where c is the number of channels, d is the vector length for frame n, and N is the number of frames. Black circle concatenates the input channels. The reparameterization trick [13] is used during the training.

We also apply the von Mises distribution on the GD $\psi^{\text{grd}}_{f,n}$, with parameters $\mu^{\text{grd}}_{f,n}$ and $\kappa^{\text{grd}}_{f,n}$, and the IF $\psi^{\text{ifr}}_{f,n}$, with parameters $\mu^{\text{ifr}}_{f,n}$ and $\kappa^{\text{ifr}}_{f,n}$. We define $\mathcal{L}^{\text{grd}}$ for the GD by substituting $\psi_{f,n}$, $\widehat{\psi}_{f,n}$ and $\widehat{\kappa}^{\text{pha}}_{f,n}$ in (8) with $\psi^{\text{grd}}_{f,n}$, $\widehat{\psi}^{\text{grd}}_{f,n}$ and $\widehat{\kappa}^{\text{grd}}_{f,n}$, respectively. Similarly, we define $\mathcal{L}^{\text{ifr}}$ for the IF with $\psi^{\text{ifr}}_{f,n}$, $\widehat{\psi}^{\text{ifr}}_{f,n}$, and $\widehat{\kappa}^{\text{ifr}}_{f,n}$. Note that we do not directly estimate the GD and the IF. They are derived from the estimated phase. Thus, $\mathcal{L}^{\text{grd}}$ and $\mathcal{L}^{\text{ifr}}$ can be seen as constraints, or priors, during the training.

In this paper, we do not estimate any concentration parameter and opt to set $\widehat{\kappa}^{\text{pha}}_{f,n} = \widehat{\kappa}^{\text{grd}}_{f,n} = \widehat{\kappa}^{\text{ifr}}_{f,n} = \widehat{a}_{f,n} + 1$. This setting makes the estimation errors on $\widehat{\psi}^{\text{pha}}_{f,n}$, $\widehat{\psi}^{\text{grd}}_{f,n}$, and $\widehat{\psi}^{\text{ifr}}_{f,n}$ more important when the estimated magnitude $\widehat{a}_{f,n}$ is high, and vice versa. As a comparison, another work [11] sets $\widehat{\kappa}^{\text{pha}}_{f,n} = 1$.

2.3. DNN design and training

Figure 2 illustrates the model used in this paper. The number of frequency bins is $F = 513$ and the latent variable dimension is $D = 32$. The total number of parameters is about 1.7 million.

Our model implementation resembles the fully convolutional DenseNets [25], that combines the DenseNets [26] and the U-Net [27]. However, our model does not have skip connections between the encoder and the decoders. We employ the gated design [28] for all convolutional layers (CLs) and the weight normalization [29], instead of the batch normalization.

We follow the terminology in [25, 26]. A Dense Block (DB) consists of 4 two-dimensional CLs with a 3×3 kernel and a channel growth rate of 8. The output channel number is the input channel number plus 4×8. A Transition Down (TD) consists of a 1×1 CL followed by 1×1 average pooling with an adjustable stride for reducing the vector length while keeping the channel number. A Transition Expand (TE) is similar to a TD, but it returns 16 channels. Conversely, a Transition Up (TU) consists of a 3×3 transposed CL with an adjustable stride for expanding the vector length. It always returns 16 channels. A Transition Final (TF) consists of a 1×1 CL to reduce the channel number to 1. Additionally, we use a Temporal Block [30] consisting of 4 one-dimensional dilated CLs applied along the time frame axis. The kernel size is 3 with dilations of 1, 2, 4, and 8 for the different layers. This block is used to capture the temporal dynamic of the latent variables. A fully-connected (FC) layer uses a leaky ReLU activation function, except when it is used as the output layer. The phase decoder outputs are in $[-\pi, \pi)$.

The model training is done in two stages. In general, the first stage aims to model the magnitude, while the second one aims to jointly model the magnitude and the phase. In the first stage, the encoder and the magnitude decoder are randomly initialized and then trained with a loss $\mathcal{L}^{(\text{M})} = \mathcal{L}^{\text{reg}} + \mathcal{L}^{\text{mag}} + \mathcal{L}^{\text{var}}$. In the second stage, all encoder and decoders are trained with a loss $\mathcal{L}^{(\text{J})} = \mathcal{L}^{(\text{M})} + \mathcal{L}^{(\text{P})}$ given the pre-trained encoder, the pre-trained magnitude decoder, and the randomly initialized phase decoder. The loss $\mathcal{L}^{(\text{P})}$ may consist of $\mathcal{L}^{\text{pha}}$, $\mathcal{L}^{\text{grd}}$, or $\mathcal{L}^{\text{ifr}}$. Thus, we end up with several models (see Tables 1 and 2).

3. EVALUATION

To evaluate the proposed approach, we consider a speech signal reconstruction task, where the latent variables are estimated given a clean utterance and then used to recreate that utterance. The reconstructed speech quality is assessed in terms of the Mean Opinion Score (MOS) [31], obtained by mapping the Perceptual Evaluation of Speech Quality score [32]. The MOS ranging from 1 to 5 represents the quality ranging from bad to excellent. Additionally, the intelligibility is assessed using the Short-Time Objective Intelligibility (STOI) score [33].

3.1. Experimental settings

We use the speech utterances from the CHiME-4 dataset [34], which are taken from the 5k vocabulary subset of the Wall Street Journal corpus [35]. All data are sampled at 16 kHz. We only consider the clean speech from the channel 5 of the simulated utterances. The training, the development, and the test sets contain 7138, 1640, and 1320 utterances, respectively. Our models are trained on the training set, validated on the development set, and evaluated on the test set.

The STFT coefficients are extracted using a Hann window with a length of 512 and a 75% overlap. We then apply a 1024-point discrete Fourier transform on the windowed signals resulting in $F = 513$. The rather high zero-padding factor reveals useful features by oversampling the spectrum [1]. In our case, it exposes more evident patterns in the IF spectrogram.

The models are trained by backpropagation [36] with the Adam update rule whose parameters are fixed to $\alpha = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-6}$ [37]. The update is done for every minibatch of 4096 frames, composed of 256-frame segments from 16 randomly selected utterances. The phase of each segment is shifted with a random value sampled from $\mathcal{N}(0, 1)$ to increase its variation. The gradient is normalized with threshold $= 1$ [38]. The training is stopped after 20 consecutive epochs failed to obtain better validation error [39]. The latest model yielding the lowest error is kept.

Table 1. Average log-likelihood on the test set for the different training loss functions.

Model	Loss function	$\widehat{\mathbf{a}}_n$	$\widehat{\psi}_n$	$\widehat{\psi}_n^{\text{grd}}$	$\widehat{\psi}_n^{\text{ifr}}$
(M)	$\mathcal{L}^{\text{reg}} + \mathcal{L}^{\text{mag}} + \mathcal{L}^{\text{var}}$	**1400**	-1204	-1204	-1204
(J1)	(M) $+ \mathcal{L}^{\text{pha}}$	1366	**-964**	-712	-954
(J2)	(M) $+ \mathcal{L}^{\text{grd}}$	**1435**	-1201	**-607**	-1201
(J3)	(M) $+ \mathcal{L}^{\text{ifr}}$	**1401**	-1198	-1198	**-800**
(J4)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{pha}} + \frac{1}{2}\mathcal{L}^{\text{grd}}$	**1420**	-1053	-635	-1054
(J5)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{pha}} + \frac{1}{2}\mathcal{L}^{\text{ifr}}$	**1399**	-1191	-1194	-826
(J6)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{grd}} + \frac{1}{2}\mathcal{L}^{\text{ifr}}$	**1409**	-1198	-671	-894
(J7)	(M) $+ \frac{1}{3}\mathcal{L}^{\text{pha}} + \frac{1}{3}\mathcal{L}^{\text{grd}} + \frac{1}{3}\mathcal{L}^{\text{ifr}}$	**1403**	-1196	-690	-908

Table 2. Average objective perceptual performance on the test set for the different training loss functions. The Griffin-Lim algorithm (GLA) is also considered for post-processing.

Model	Loss function	Without GLA		With GLA	
		MOS	STOI	MOS	STOI
(M)	$\mathcal{L}^{\text{reg}} + \mathcal{L}^{\text{mag}} + \mathcal{L}^{\text{var}}$	1.96	0.690	3.97	**0.792**
(J1)	(M) $+ \mathcal{L}^{\text{pha}}$	3.34	0.770	3.83	**0.787**
(J2)	(M) $+ \mathcal{L}^{\text{grd}}$	2.18	0.734	4.00	**0.795**
(J3)	(M) $+ \mathcal{L}^{\text{ifr}}$	2.51	0.702	3.86	**0.789**
(J4)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{pha}} + \frac{1}{2}\mathcal{L}^{\text{grd}}$	**3.71**	**0.786**	**4.04**	**0.792**
(J5)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{pha}} + \frac{1}{2}\mathcal{L}^{\text{ifr}}$	2.39	0.690	3.89	**0.790**
(J6)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{grd}} + \frac{1}{2}\mathcal{L}^{\text{ifr}}$	3.54	0.777	3.90	**0.789**
(J7)	(M) $+ \frac{1}{3}\mathcal{L}^{\text{pha}} + \frac{1}{3}\mathcal{L}^{\text{grd}} + \frac{1}{3}\mathcal{L}^{\text{ifr}}$	3.13	0.766	3.86	**0.789**

3.2. Experimental results

Tables 1 and 2 show the experimental results for the different loss functions on the test set. Table 1 shows the average log-likelihood (LL). It is obtained by computing $-\mathcal{L}^{\text{mag}}$, $-\mathcal{L}^{\text{pha}}$, $-\mathcal{L}^{\text{grd}}$, and $-\mathcal{L}^{\text{ifr}}$ for each utterance and then averaging the results. These LL values reflect the estimation accuracy. Table 2 shows the objective perceptual performance. The magnitude and the phase are always estimated using our models, except for the model (M) where the phase is sampled randomly from a uniform distribution. This model (M) is obtained from the first training stage only and acts as the baseline. In addition, we consider the Griffin-Lim algorithm (GLA) [7] as post-processing. It is done for 100 iterations. Boldface numbers show the best performance for each column, taking into account the 95% confidence interval. A higher value is better for all metrics.

3.3. Discussion

Table 1 shows that the good magnitude reconstruction achieved by the model (M) could be preserved by the other models in most cases. Thus, we could focus on observing the estimation of the phase and its derivatives. The model (J1) shows that a good phase estimation naturally provides fair estimates of the derivatives. Conversely, the models (J4), (J5), and (J7) have better estimates of either or both of GD and IF, but worse phase estimate, than the model (J1). It suggests that the optimization of the phase derivatives strongly drives the overall optimization. Thus, a more elaborate weighting scheme might be useful.

Let us now observe Tables 1 and 2 together. The model (J4) provides the best performance. It suggests that minimizing $\mathcal{L}^{\text{grd}}$ is useful. The models (J1), (J4), (J6), and (J7) provide fair performance and all of them have a good GD estimation.

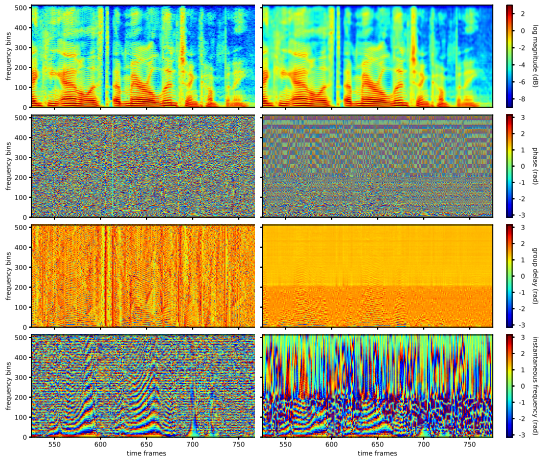


Fig. 3. The left and the right columns show spectrogram examples of a true speech and its reconstruction using the model (J4), respectively. From top to bottom, we show the log-magnitude, the phase, the group delay, and the instantaneous frequency spectrograms. The utterance is `F05_440C020I_PED` from the set `et05_ped_simu`.

However, the model (J2) shows that a good GD alone is not enough. Interestingly, the models (J6) and (J7) provide reasonable performance although the phase estimation is poor. It might suggest that estimating the absolute phase value is not critical, and capturing the phase interdependence on the frequency *and* the time axes is sufficient. Additionally, the GLA iterations effectively improve both the quality and the intelligibility. The performance of our models without the GLA is still below that of the GLA with random initial phase. However, our method does not need any iteration.

Figure 3 shows spectrogram examples of a speech segment and its reconstruction. The log-magnitude spectrograms show that the model reconstructs the harmonic structures well, although those for above the frequency bin 200 tend to be unclear. The phase and the phase derivative spectrograms clearly show that there are still opportunities for further improvement. The estimated spectrograms resemble the true ones only for the lower frequency bands. This might be an impact of associating the von Mises concentration parameters to the estimated magnitude. Therefore, the estimation of those parameters should be explored further. Audio samples are available online[1].

4. CONCLUSION

We proposed a deep generative model for jointly modeling the magnitude and the phase of STFT. We took into account the phase derivatives, i.e., the group delay and the instantaneous frequency. We found that good phase derivative estimates are sufficient to provide a fair speech quality. However, we also found that the phase derivative optimization strongly drives the overall optimization and thus, a more elaborate weighting scheme might be required. Additionally, future work includes incorporating the estimation of the von Mises concentration parameters and utilizing the proposed models for downstream tasks, e.g., speech enhancement and audio source separation.

[1]Demo webpage: `https://aanugraha.gitlab.io/demo/icassp19`

5. REFERENCES

[1] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 2nd edition, 2009.

[2] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication: Challenges and Perspectives*, Springer, 2010.

[3] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*, Wiley, 2018.

[4] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 235–238, 1977.

[5] T. Gerkmann, M. Krawczyk-Becker, and J. le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, pp. 55–66, 2015.

[6] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, July 2016.

[7] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[8] J. le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for stft spectrograms and their application to phase reconstruction," in *Proc. SAPA*, Brisbane, Australia, 2008.

[9] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 6, pp. 1095–1105, 2018.

[10] P. Magron and T. Virtanen, "On modeling the STFT phase of audio signals with the von Mises distribution," in *Proc. IWAENC*, Tokyo, Japan, 2018.

[11] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, and H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," in *Proc. IWAENC*, Tokyo, Japan, 2018.

[12] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "PhaseNet: Discretized phase modeling with deep neural networks for audio source separation," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 2713–2717.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, Banff, Canada, 2014.

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

[15] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, Lille, France, 2015, pp. 1530–1538.

[16] A. van den 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 e-prints*, 2016.

[17] H. Kameoka, L. Li, S. Inoue, and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," *arXiv e-prints*, 2018.

[18] S. Seki, H. Kameoka, L. Li, T. Toda, and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," *arXiv e-prints*, 2018.

[19] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE ICASSP*, Calgary, Canada, 2018, pp. 716–720.

[20] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. IEEE MLSP*, Aalborg, Denmark, 2018.

[21] K. Sekiguchi, Y. Bando, K. Yoshii, and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," in *Proc. APSIPA*, Honolulu, USA, 2018.

[22] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.

[23] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, 2014.

[24] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal – Part 1: Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 520–538, 1992.

[25] S. Jegou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proc. IEEE CVPR Workshops*, Honolulu, USA, 2017.

[26] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, Honolulu, USA, 2017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Munich, Germany, 2015.

[28] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, Sydney, Australia, 2017, pp. 933–941.

[29] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 901–909.

[30] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv e-prints*, 2018.

[31] ITU-T, "P.862.1 : Mapping function for transforming P.862 raw result scores to MOS-LQO," 2003.

[32] ITU-T, "P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[35] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," DVD, 2007, Philadelphia: Linguistic Data Consortium.

[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, USA, 2015.

[38] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, Atlanta, USA, June 2013, pp. 1310–1318.

[39] L. Prechelt, "Early stopping – but when?," in *Neural Networks: Tricks of the Trade*, G. Montavon, G. B. Orr, and K.-R. Müller, Eds., pp. 53–67. Springer, 2nd edition, 2012.