A Deep Generative Model of Speech Complex Spectrograms

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

[⋆] Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo 103-0027, Japan

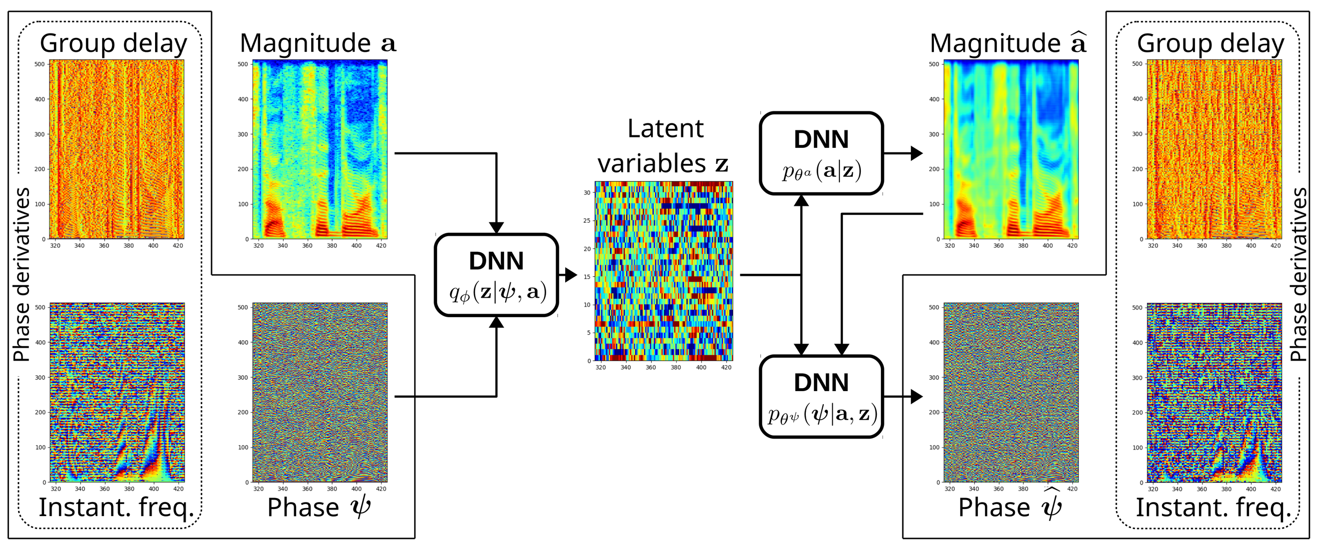
[†] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

MOTIVATION

- Probabilistic approaches to speech enhancement [1,2]: VAE as a prior of the speech *power* spectrograms
- How about a prior of *complex* spectrograms..? It might allow a better speech enhancement.
- Phase recovery approaches typically assume the magnitude is known, e.g., the Griffin-Lim algorithm and some DNN-based methods [3,4]
- **Let's develop a latent variable model for speech complex spectrogram generation!**

IDEA

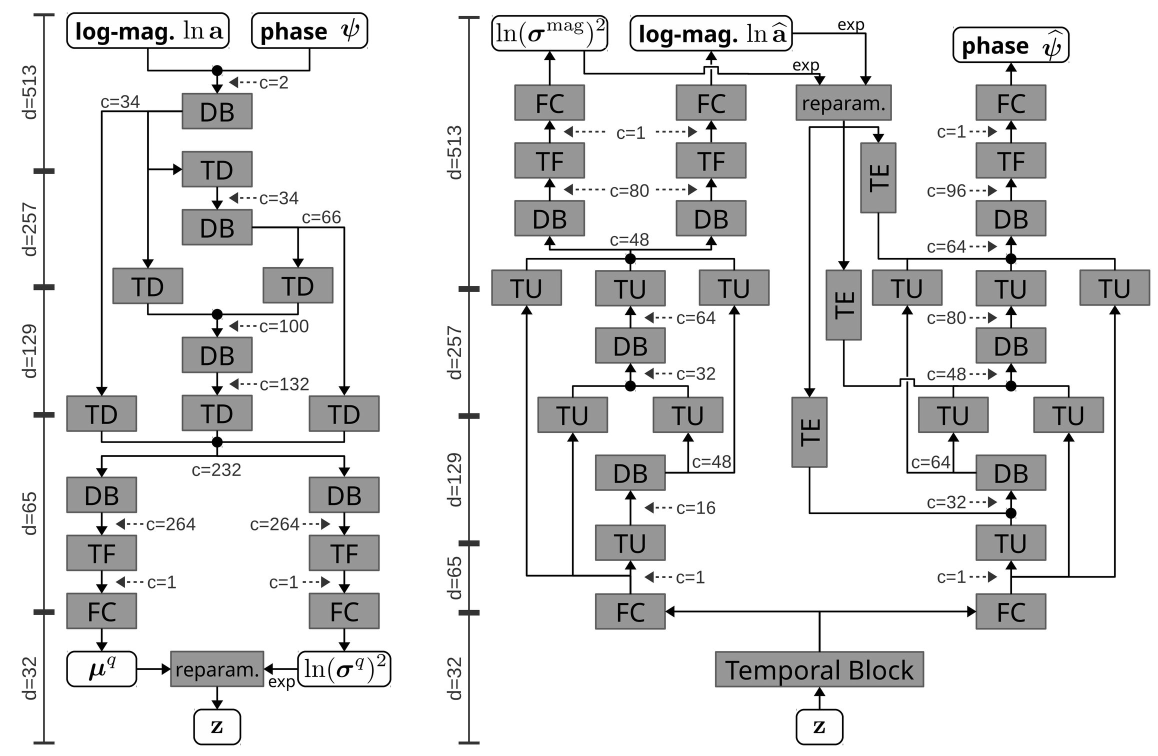
$$\underbrace{s_{f,n}}_{\text{complex STFT}} = \underbrace{a_{f,n}}_{\text{magnitude}} \exp(i \underbrace{\psi_{f,n}}_{\text{phase}})$$



- Phase derivatives:
  - **Group delay (GD)**: the derivative along the frequency axis
    $$\psi^{\text{grd}}_{f,n} = \text{wrap}(-\psi_{f+1,n} + \psi_{f,n})$$
  - **Instantaneous frequency (IF)**: the derivative along the time axis
    $$\psi^{\text{ifr}}_{f,n} = \text{wrap}(\psi_{f,n+1} - \psi_{f,n})$$
- **Let's exploit the interdependence between the phase, the GD, and the IF!**

PROPOSED METHOD

- Model formulation:
  $$p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n) = p_{\theta^\psi}(\boldsymbol{\psi}_n|\mathbf{a}_n, \mathbf{z}_n)\, p_{\theta^a}(\mathbf{a}_n|\mathbf{z}_n)\, p_\theta(\mathbf{z}_n) \quad (1)$$
- The model parameters are estimated by minimizing the negative log-likelihood (NLL):
  $$\begin{aligned} -\ln \int_{\mathbf{z}_n} p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n)\, d\mathbf{z}_n &= -\ln \int_{\mathbf{z}_n} \frac{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)}{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)} p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n)\, d\mathbf{z}_n \\ &\le -\mathbb{E}_{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n,\mathbf{a}_n)}\left[\ln \frac{p_\theta(\boldsymbol{\psi}_n, \mathbf{a}_n, \mathbf{z}_n)}{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)}\right] \\ &= \text{KL}[q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)||p_\theta(\mathbf{z}_n)] \\ &\quad - \mathbb{E}_{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n,\mathbf{a}_n)}[\ln p_{\theta^a}(\mathbf{a}_n|\mathbf{z}_n)] \\ &\quad - \mathbb{E}_{q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n,\mathbf{a}_n)}[\ln p_{\theta^\psi}(\boldsymbol{\psi}_n|\mathbf{a}_n, \mathbf{z}_n)] \\ &\triangleq \mathcal{L}^{\text{reg}} + \mathcal{L}^{\text{mag}} + \mathcal{L}^{\text{pha}} \end{aligned} \quad (2)$$
- Assuming a simple prior $p_\theta(\mathbf{z}_n) \sim \mathcal{N}(\mathbf{z}_n|\mathbf{0}, \mathbf{I})$, *the regularization term* $\mathcal{L}^{\text{reg}}$:
  $$\mathcal{L}^{\text{reg}} = \frac{1}{2N}\sum_{d,n}\left((\mu^q_{d,n})^2 + (\sigma^q_{d,n})^2 - \ln(\sigma^q_{d,n})^2 - 1\right) \quad (3)$$
- The magnitude follows a Gaussian distribution:
  $$a_{f,n} \sim \mathcal{N}\left(a_{f,n}\middle|\mu^{\text{mag}}_{f,n}, (\sigma^{\text{mag}}_{f,n})^2\right) \quad (4)$$
  The magnitude reconstruction loss $\mathcal{L}^{\text{mag}}$ is the NLL:
  $$\mathcal{L}^{\text{mag}} = \frac{1}{2N}\sum_{f,n}\left(\ln 2\pi(\widehat{\sigma}^{\text{mag}}_{f,n})^2 + \frac{(a_{f,n} - \widehat{a}_{f,n})^2}{(\widehat{\sigma}^{\text{mag}}_{f,n})^2}\right) \quad (5)$$
- The phase follows a von Mises distribution:
  $$\psi_{f,n} \sim \mathcal{VM}\left(\psi_{f,n}\middle|\mu^{\text{pha}}_{f,n}, \kappa^{\text{pha}}_{f,n}\right) \quad (6)$$
  The phase reconstruction loss $\mathcal{L}^{\text{pha}}$ is the NLL:
  $$\mathcal{L}^{\text{pha}} = \frac{1}{N}\sum_{f,n}\left(\ln 2\pi I_0(\widehat{\kappa}^{\text{pha}}_{f,n}) - \widehat{\kappa}^{\text{pha}}_{f,n}\cos(\psi_{f,n} - \widehat{\psi}_{f,n})\right) \quad (7)$$
- Additionally, each of the GD and the IF also follows a von Mises distribution.
  The GD and the IF reconstruction losses ($\mathcal{L}^{\text{grd}}$ and $\mathcal{L}^{\text{ifr}}$) are defined similarly to $\mathcal{L}^{\text{pha}}$.
- Concentration parameters: $\widehat{\kappa}^{\text{pha}}_{f,n} = \widehat{\kappa}^{\text{grd}}_{f,n} = \widehat{\kappa}^{\text{ifr}}_{f,n} = \widehat{a}_{f,n} + 1$



Encoder: $q_\phi(\mathbf{z}_n|\boldsymbol{\psi}_n, \mathbf{a}_n)$ **Decoders:** $p_{\theta^a}(\mathbf{a}_n|\mathbf{z}_n)$, $p_{\theta^\psi}(\boldsymbol{\psi}_n|\mathbf{a}_n, \mathbf{z}_n)$

- The model is based on the DenseNets design [5], mainly consisting of convolutional layers (see the paper for the details).
- The model training is done in two stages:
  - Stage 1 aims for a good magnitude estimation
  - Stage 2 aims for a good phase and magnitude estimation

EVALUATION

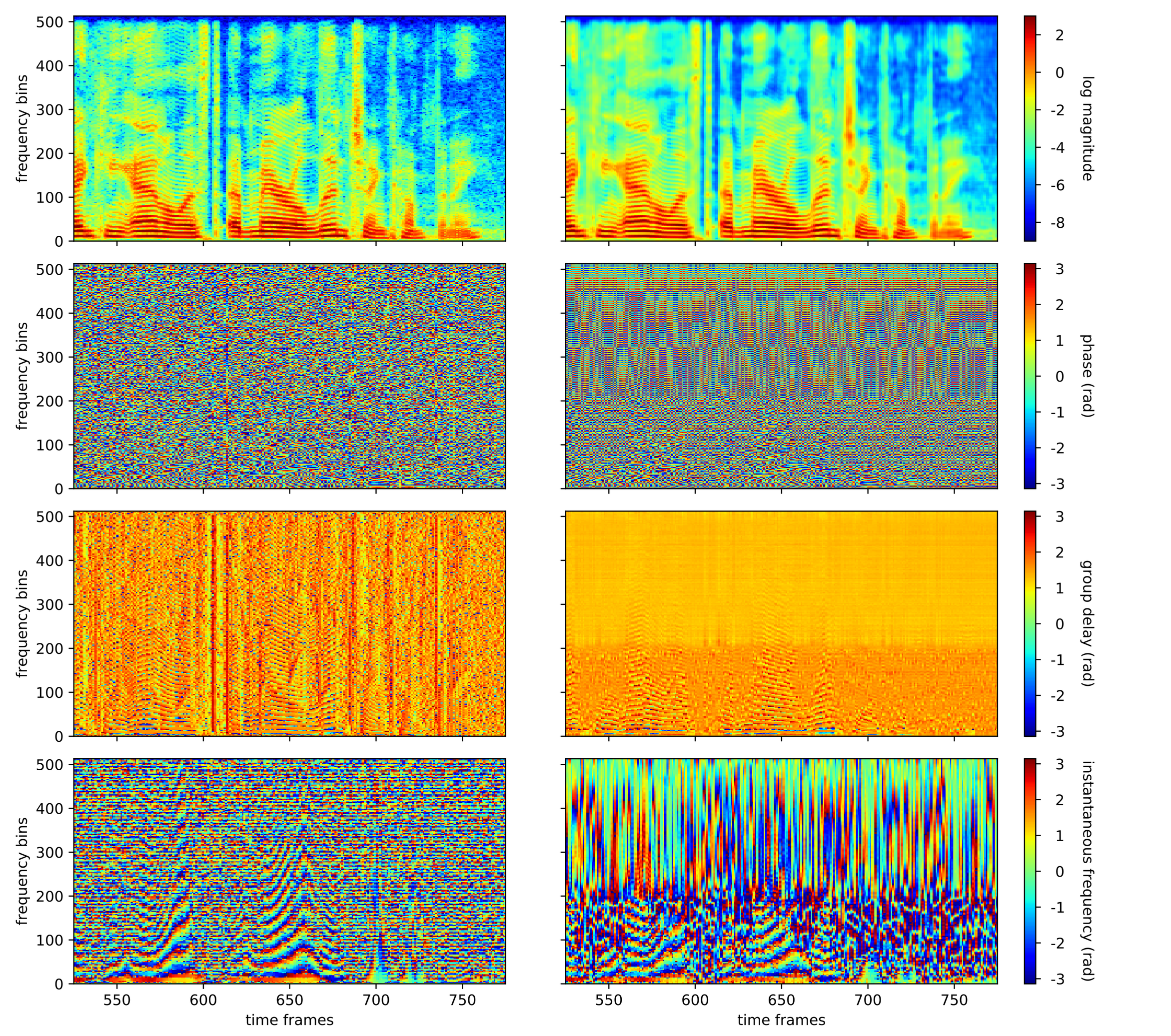
- Task: speech reconstruction
- Performance metrics:
  - Mean Opinion Score (MOS), mapped from Perceptual Evaluation of Speech Quality (PESQ) score
  - Short-Time Objective Intelligibility (STOI)
- Corpus: CHiME-4
  - all data are sampled at 16 kHz
  - only the clean speech of the channel 5 from the simulated datasets
  - subsets:
    - training set: 7138 utts. (± 15.0 hours)
    - dev. set: 1640 utts. (± 2.9 hours)
    - test set: 1320 utts. (± 2.3 hours)
- STFT analysis parameters:
  - 512-point Hann window (75% overlap)
  - 1024-point DFT

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

Model	Loss function	$\widehat{\mathbf{a}}_n$	$\widehat{\boldsymbol{\psi}}_n$	$\widehat{\boldsymbol{\psi}}^{\text{grd}}_n$	$\widehat{\boldsymbol{\psi}}^{\text{ifr}}_n$
(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

Average objective perceptual performance on the test set for the different training loss functions.

Model	Loss function	MOS	STOI
(M)	$\mathcal{L}^{\text{reg}} + \mathcal{L}^{\text{mag}} + \mathcal{L}^{\text{var}}$	1.96	0.690
(J1)	(M) $+ \mathcal{L}^{\text{pha}}$	3.34	0.770
(J2)	(M) $+ \mathcal{L}^{\text{grd}}$	2.18	0.734
(J3)	(M) $+ \mathcal{L}^{\text{ifr}}$	2.51	0.702
(J4)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{pha}} + \frac{1}{2}\mathcal{L}^{\text{grd}}$	**3.71**	**0.786**
(J5)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{pha}} + \frac{1}{2}\mathcal{L}^{\text{ifr}}$	2.39	0.690
(J6)	(M) $+ \frac{1}{2}\mathcal{L}^{\text{grd}} + \frac{1}{2}\mathcal{L}^{\text{ifr}}$	3.54	0.777
(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



(a) True speech (b) Reconstruction (J4)

Utt. ID: `F05_440C020I_PED` from the set `et05_ped_simu`

CONCLUSION

- The proposed method can reproduce time-domain speech with a high quality and a high intelligibility. Audio samples are available on the demo webpage: https://aanugraha.gitlab.io/demo/icassp19/.
- Good phase derivatives are sufficient to obtain a fair speech quality.
- The phase derivative optimization strongly drives the overall optimization and thus, a more elaborate weighting might be necessary.
- Future works include (1) estimating the von Mises concentration parameters, and (2) utilizing the model for speech enhancement.

References

[1] 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," IEEE ICASSP '18,

[2] S. Leglaive, L. Girin, & R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," IEEE MLSP '18.

[3] N. Takahashi, P. Agrawal, N. Goswami, & Y. Mitsufuji, "PhaseNet: Discretized phase modeling with deep neural networks for audio source separation," Interspeech '18.

[4] S. Takamichi, Y. Saito, N. Takamune, D. Kitamura, & H. Saruwatari, "Phase reconstruction from amplitude spectrograms based on von-Mises-distribution deep neural network," IWAENC '18.

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," CVPR '17.