Exploring the Potential of VAE Decoders for Enhanced Speech Re-Synthesis

Omead Pooladzandi, XiLin Li, Yang Gao, Lalin Theverapperuma
opooladz@g.ucla.edu, {xilin, metag, lalin}@meta.com

2023 IEEE Statistical Signal Processing Workshop (SSP) | 978-1-6654-5245-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/SSP53291.2023.10207969

***Abstract*—In this paper, we study different Variational Autoencoders (VAEs) decoder distributions in the audio setting to see how to improve magnitude and phase reconstruction on speech resynthesis tasks. We first provide background on the existing decoder distributions, such as Complex Gaussian and Laplace, which are equivalent to a Gamma decoder under certain conditions. We then consider separately modeling speech's magnitude and phase information to see if we can improve the quality of either component, yielding an improvement in speech resynthesis. Extensive experiments show the Gamma decoder significantly improves magnitude reconstruction and that the von Mises decoder can weakly learn phase information. The novel Gamma decoder outperforms previous approaches, achieving a near-perfect PESQ of 4.4, representing a 42% improvement upon the state-of-the-art IS-VAE and an 86% decrease in the FAD metric. Our results demonstrate the effectiveness of the novel approach, improving the quality of speech resynthesis and compression capacity of VAEs.**

***Index Terms*—VAE, Speech Re-Synthesis**

I. Introduction

The goal of the VAE is to find an optimal trade-off between rate and distortion [1], such that the reconstructed data is as close as possible to the original data while still requiring a minimal amount of information to represent it.

In the audio setting, VAEs have been applied to tasks such as music generation, speech synthesis, speech source-filter representation, and audio denoising [2, 3, 4, 5, 6, 7]. One key aspect of VAE performance is the choice of the decoder loss function, which determines how well the model is able to capture the underlying distribution of the data. Different decoder distributions have been used in the context of VAEs for speech, including Gaussian decoders, complex-valued probability density functions, and a combination of magnitude and phase reconstruction [2, 3, 4, 5, 6, 7]. Yet, the Itakura-Saito (IS) VAE has emerged as the state-of-the-art (SOTA) decoder loss, thanks to its ability to accurately model the distribution of audio signals and generate high-quality samples.

This work considers the effect of different decoder distributions on audio-quality speech resynthesis, as well as compression capacity. We explore whether we can learn phase information in our decoder formulation and compare it to the IS-VAE. We believe that if we can find a better spectrum representation via a new decoder, it will improve current VAE-based audio processing methods that currently use IS-divergence.

To this end, we explore multiple novel decoder distributions that can more accurately capture the underlying distribution of speech signals. This results in improved reconstruction performance on speech resynthesis tasks. The role of the decoder distribution in the VAE architecture and its relationship to speech codec is discussed first. Our novel decoder distributions are then introduced, and results demonstrating their improved performance on speech resynthesis tasks compared to the IS-VAE are presented.

Our experimental findings suggest that: a) Phase information can be learned to a limited extent; b) There exists many distributions that can model speech leading to different properties; c) The Gamma distribution is the best-tested model for speech re-synthesis and compress-ability; d) Network architecture does not greatly affect performance. All models are evaluated using the Perceptual Evaluation of Speech Quality (PESQ) metric [8], which ranges from [-0.5, 4.5], and the Fŕechet Audio Distance (FAD) metric [9], whose range is unbounded, with a lower value indicating a greater similarity between the two sets of audio features. Unlike PESQ, FAD does not require a reference signal but instead takes the distance in the embedding space of an NN-based audio classifier between a distribution of natural audio and a generated dataset. The best-performing decoder explored in this paper is the Gamma decoder. This decoder achieves a near-perfect PESQ of 4.4, improves PESQ by 42% and reduces FAD by 86% compared to the IS-VAE.

II. Background

In recent years, the Itakura-Saito VAE (IS-VAE) has become the SOTA VAE used in the domain of speech modeling. Typically, the IS-VAE models the probability density function (pdf) of the magnitude spectrum of the Short-Time Fourier Transform (STFT) domain. Only considering the magnitude spectrum of the speech signal, rather than both the magnitude and phase, could potentially lead to a loss of quality of the synthesized speech. To investigate this hypothesis, Nakashika [6] proposed using a VAE to directly model the speech signal using a complex-valued Gaussian decoder. This approach may not be the most effective for capturing phase information because the magnitude and phase may follow different distributions. To this end, we propose modeling the magnitude and phase of speech separately using a positive distribution for the magnitude and a circular distribution for the phase. By inputting $x = [x_r, x_i]$, where $|x| = \sqrt{x_r^2 + x_i^2} = r$ represents the magnitude spectrum of the speech signal, and estimating both the magnitude, $|\hat{x}|$, and phase, $\angle x$, we aim to improve the performance of VAEs for speech re-synthesis tasks. This approach differs from the recently proposed method [6], which uses a singular complex-valued Gaussian distribution to model the magnitude and phase of the signal x.

Overall, our proposed method offers a new approach to modeling the magnitude and phase of speech separately, with the goal of improving the performance of VAEs in speech re-synthesis tasks.

III. Problem Setting

The VAE is a framework that learns a latent representation, z, of the data x, which captures the underlying structure of the data distribution. VAEs are trained to learn an encoder function, $e_\phi(z|x)$, which maps the input data x to the latent space z. The approximate posterior distribution over the latent variables is given by $e_\phi(z|x)$. The encoder estimates the parameters of a normal distribution in the latent space, such as the mean and variance, which define the location and scale of the distribution. The VAE also learns a marginalized distribution, $m_\theta(z)$, for the latent variable encoding, and a decoder function, $d_\theta(x|z)$. Thus the joint distribution, $p_\theta(x, z)$, can accurately model the true data distribution, $p(x)$. The decoder estimates the parameters of a distribution, such as the shape and rate parameters of

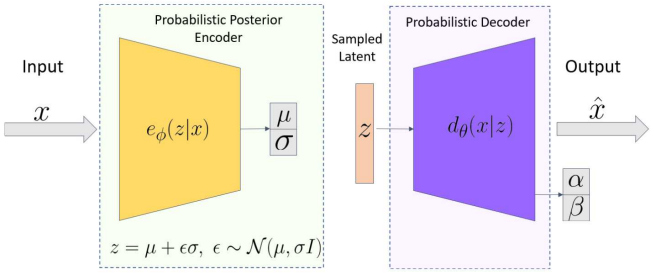


Fig. 1: VAE: Where x is complex spectrogram of original signal and $\hat{x}$ is estimated via the parameters of the decoder distribution. $\hat{x}$ is complex or real depending on decoder distribution.

a Gamma distribution, to reconstruct the input signal x. The decoding side defines a joint pdf, $q_\theta(x, z) = m_\theta(z)d_\theta(x|z)$, with everything explicit, resulting in a joint pdf, $p_\phi(x, z) = p(x)e_\phi(z|x)$ that we can readily draw samples from, although $p(x)$ is unknown.

The VAE is based on the principle of maximum likelihood estimation (MLE), where the goal is to maximize the likelihood of the training data under the model. To make the optimization process more feasible, the VAE introduces the variational reparameterization trick, which involves re-parameterizing the random noise used to sample from the latent distribution so that the resulting samples can be transformed into the latent space in a differentiable manner. By using this trick, VAEs can be trained more efficiently on large datasets using a variational lower bound on the log-likelihood, written as:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{e_\phi(z|x)}\left[\log d_\theta(x|z)\right] - \mathrm{KL}\left(e_\phi(z|x)||m_\theta(z)\right) \quad (1)$$

where x is a single training example, and θ and ϕ are the parameters of the decoder and encoder, respectively. The first term in the lower bound, $\mathbb{E}_{e_\phi(z|x)}\left[\log d_\theta(x|z)\right]$, is known as the reconstruction loss, and it measures the difference between the reconstructed data and the original data. The second term, $\mathrm{KL}\left(e_\phi(z|x)||m_\theta(z)\right)$, is known as the KL divergence, and it measures the difference between the approximate posterior distribution and the latent distribution.

IV. Understanding Decoder Distributions

One way to understand the impact of phase information on VAEs for speech modeling is to consider a general decoder that models both the magnitude and phase of speech signals. There are two main categories of decoder distributions that can be used for this purpose: (Type I) decoders that directly model the complex-valued signal $\hat{x} = [\hat{x}_r, \hat{x}_i]$, such as Gaussian or Laplace decoders, and (Type II) Decoders that model the magnitude $|\hat{x}|$ and phase $\angle\hat{x}$ information separately, such as a Positive Distribution for magnitude and Circular Distribution for phase. The Itakura-Saito VAE (IS-VAE) can be seen as a special case of the general family of decoders that only model magnitude information. However, the general family of decoders has the potential to directly generate phase information, as they are theoretically phase-aware. Since there is a clear connection between magnitude and phase information in speech signals, modeling phase information may inherently improve the modeling of the magnitude spectrum. This is in line with the assumption often made by Vocoders [10], which is that the phase of a speech signal can be approximately reconstructed from its magnitude spectrum.

In this paper, for Type II decoders we consider Log-Normal and Gamma distributions to model magnitude and the von Mises distribution to model phase information of speech. The Log Normal Distribution corresponds to the Log Spectral Distance used in speech codec [11] and the Gamma distribution has been used as a strong prior in noise suppression [12, 13] and as we will show is directly related to the IS-divergence. The von Mises distribution is a natural choice for modeling the phase of speech because it is a circular continuous and smooth distribution, which allows for more accurate modeling of the phase information in speech signals. Additionally, it is symmetric around the mean, which is useful for modeling the cyclical nature of phase information. See Table I for all decoder distributions used in this paper.

For all experiments, we consider the pdf of bins that are conditionally independent given z, in the STFT domain. We only consider one bin, $x = [x_r, x_i]$, and assume that power and phase spectra are conditionally independent given z.

TABLE I: Formulations of Type I and Type II a & b decoders such as Complex Normal & Laplace, Log Normal von Mises, Gamma von Mises, Gamma, and Itakura-Saito. These decoder distributions can be used to derive negative log-likelihoods as done in eq. (9) & eq. (12).

Decoder Distribution	Formulation	Models
Complex Normal eq. (2)	$\frac{1}{\pi\sigma^2}e^{-\lvert x-\mu\rvert^2/\sigma^2}$	Joint Complex Spectrum
Complex Laplace eq. (3)	$\frac{1}{2\pi\sigma^2}e^{-\lvert x-\mu\rvert/\sigma}$	Joint Complex Spectrum
Log Normal von Mises	$\frac{\exp\left(\kappa\frac{x^\mathsf{T}x_0}{\lvert x\rvert\lvert x_0\rvert}-\frac{(\log\lvert x\rvert/\log\lvert x_0\rvert)^2}{2\sigma^2}\right)}{(2\pi)^{1.5}\,\sigma\,I_0(\kappa)\,\lvert x\rvert^2}$	Separate Magnitude & Phase
Gamma von Mises eq. (7)	$\frac{\beta^\alpha\lvert x\rvert^{\alpha-2}e^{\kappa\frac{x^\mathsf{T}x_0}{\lvert x\rvert\lvert x_0\rvert}-\beta\lvert x\rvert}}{2\pi\,I_0(\kappa)\,\Gamma(\alpha)}$	Separate Magnitude & Phase
Gamma eq. (10)	$\frac{e^{-\frac{\lvert x\rvert^2}{2\sigma^2}}}{2\pi\sigma^2}$	Magnitude Only
Itakura-Saito eq. (11)	$\frac{1}{\Gamma(\alpha)}\frac{\lvert x\rvert^\alpha}{\lvert y\rvert^\alpha}\lvert x\rvert^{-2}e^{-\frac{\lvert x\rvert}{\lvert y\rvert}}$	Magnitude Only

A. Directly Modeling The Signal

Many decoder distributions exist that implicitly model the magnitude and phase components of speech. Here we discuss the Normal and Laplace Decoders.

Complex Normal The Gaussian decoder models a complex-valued spectrogram, implicitly modeling the magnitude & phase of speech

$$d_\theta(x|z) = \frac{1}{\pi\sigma^2}e^{-|x-\mu|^2/\sigma^2}. \quad (2)$$

Complex Laplace Similarly, the Laplace decoder models a complex-valued spectrogram, modeling the magnitude & phase of speech

$$d_\theta(x|z) = \frac{1}{2\pi\sigma^2}e^{-|x-\mu|/\sigma}. \quad (3)$$

Assuming that $\mu = 0$, the probability density functions of both decoders are related to the Gamma distribution. Specifically, $|x|^2$ and $|x|$ follow the Gamma distribution for Gaussian and Laplace decoders, respectively. Empirical results demonstrate that for speech signals, $\mu \to 0$. However, it is not currently known if it is possible to learn μ theoretically.

In conjunction, another similar decoder is the multivariate Laplace distribution. We forego this decoder since it does not correspond to either the Gamma or log-normal distribution even though it may have more favorable properties than the complex Laplace one.

B. Separately Modeling Magnitude and Phase

To explicitly capture both magnitude and phase information, we propose combining a positive distribution, such as the Log Normal or Gamma distribution, to capture magnitude information and a circular distribution, such as the von Mises distribution, which is an approximation to the Circular Normal distribution, to capture phase information.

Gamma + von Mises This decoder generates three outputs, α, β, and κ: α represents the shape parameter, β represents the rate

parameter, and $\frac{\alpha}{\beta} = |\hat{x}|$ determines the central location of the Gamma distribution. The precision of the phase spectrum of the von Mises distribution is represented by κ. The probability density function of x can be expressed as a function of its parameters, which varies depending on the specific distribution. We use a change of variables to rewrite the pdf $p(x_r, x_i)$ as $p(|x|, \angle x)$.

$$p(x_r, x_i) = \left|\frac{\partial(r,\theta)}{\partial(x_r,x_i)}\right|^{-1} p(r,\theta) = \frac{1}{|x|} p(|x|)p(\theta) \tag{4}$$

where

$$p(|x|) = \frac{\beta^\alpha |x|^{\alpha-1} e^{-\beta|x|}}{\Gamma(\alpha)}, \quad p(\theta) = \frac{e^{\kappa\cos(\theta-\theta_0)}}{2\pi I_0(\kappa)} \tag{5}$$

where Γ is a Gamma function, I_0 is a Bessel function, and $p(\theta)$ is the von Mises distribution. We can write the decoder as

$$d_\theta(x|z) = \frac{1}{|x|} \frac{\beta^\alpha |x|^{\alpha-1} e^{-\beta|x|}}{\Gamma(\alpha)} \frac{e^{\kappa \frac{x^\mathsf{T} x_0}{|x||x_0|}}}{2\pi I_0(\kappa)} \tag{6}$$

which can be rewritten with a more pleasant form as

$$d_\theta(x|z) = \frac{\beta^\alpha |x|^{\alpha-2} e^{\kappa \frac{x^\mathsf{T} x_0}{|x||x_0|} - \beta|x|}}{2\pi\, I_0(\kappa)\, \Gamma(\alpha)} \tag{7}$$

where α, β and κ are functions of z (all defined by the decoder). We constrain $\alpha > 0$ and $\beta > 0$ through reparameterization. As such we can formulate our loss function as the sum of the natural log of $d_\theta(x|z)$ and the KLD as in eq (1). The negative Log Loss becomes:

$$\begin{aligned} -\log d_\theta(x|z) = {} & \log 2\pi + \log I_0(\kappa) + \log\Gamma(\alpha) + \beta|x| \\ & -\alpha\log\beta - (\alpha-2)\log|x| - \kappa\frac{x^\mathsf{T} x_0}{|x||x_0|} \end{aligned} \tag{8}$$

When $\kappa \to 0$ the von Mises distribution becomes uniformly distributed and the phase information is not learned.

C. Phase Agnostic Decoders

Some decoders, such as the IS-VAE, do not model phase information. To generate a signal, one will need to use the ground truth oracle phase or find a magnitude-consistent phase via Griffin-Lin style methods [14]. This style of magnitude spectrum modeling is tantamount to having a von Mises distribution with infinite dispersion. As such we derive the Gamma distribution as a special case of the Gamma von Mises distribution, with $\kappa = 0$. We then derive the IS-VAE as a Gamma VAE.

Gamma Distribution Consider the Gamma von Mises Decoder which produces three outputs, α, β, and κ, where α is the shape, β is the rate, and $\alpha/\beta = |\hat{x}|$ gives the center location of the Gamma distribution. The precision of the phase spectrum of the von Mises distribution is given by κ. If we set $\kappa = 0$, the Gamma von Mises Distribution degenerates into the Gamma Distribution. The pdf of x can be written as:

$$p(x_r, x_i) = \left|\frac{\partial(r,\theta)}{\partial(x_r,x_i)}\right|^{-1} p(r,\theta) = 2\beta e^{-\beta r} p(\theta) \tag{9}$$

where $x_0 = [r^{0.5\cos(\theta)}, r^{0.5\sin(\theta)}]$, $r = x_r^2 + x_i^2$, $\theta = \tanh^{-1}(\frac{x_i}{x_r})$. The decoder is

$$d_\theta(x|z) = 2\beta e^{-\beta r} = \frac{1}{2\pi\sigma^2} e^{-\frac{|x|^2}{2\sigma^2}} \tag{10}$$

We can write $\beta = \frac{1}{2\sigma^2}$. This decoder is a special case of the 2D-normal decoder where the mean $\mu = 0$. This gives us a direct connection between Type I and Type II decoders.

Itakura-Saito Divergence We derive the Itakura-Saito-VAE as the Gamma decoder. We assume the probabilistic decoder takes the form of a Gamma distribution.

$$d_\theta(x|z) = p(|x|) = \frac{1}{\Gamma(\alpha)} \frac{|x|^\alpha}{|y|^\alpha} |x|^{-2} e^{-\frac{|x|}{|y|}} \tag{11}$$

We take the negative log-likelihood of this distribution and we get our negative log-likelihood loss of,

$$-\log d_\theta(x|z) = \frac{|x|}{|y|} - \alpha\log\frac{|x|}{|y|} + \log\Gamma(\alpha) + 2\log|x|. \tag{12}$$

In this formulation, x is the ground truth signal, y is the estimated signal, and α is a parameter of the distribution estimated by the NN.

V. Experiments

In this section we evaluate the effectiveness of different decoder distributions to learn speech by answering the following questions: (1) can we learn phase information directly in the VAE process, and does explicitly penalizing for phase lead to a better power spectrum representation (2) does modeling time dependence help model magnitude and phase of speech? (3) do decoders have different compressive capacities?

Dataset The VCTK dataset consists of 110 English speakers both male and female with various accents resampled from 48 KHz to 16 KHz. We randomly select 10% of the speakers for the training dataset and the rest are the test dataset. The time-domain speech signals are converted to spectrograms using the short-time Fourier transform (STFT) with a periodic Hann analysis window of length 64 ms (1,024 samples) and a hop size of 128 samples. For fully connected architectures, the train set is then subsampled over time, every 10 frames. This results in a training set of 1% of the full dataset. For convolution-based networks, we use a random window of 5ms and directly use 1% of the dataset. The frequency signal is then separated into log power and complex phase components before being input to the NN for both phase-aware and agnostic decoder VAEs.

Baselines In this paper we propose several decoder distributions derived above to the widely excepted baseline decoder enforced by the Itakura-Saito divergence as the reconstruction loss in the VLB. We train a VAE with two fully connected layers in the encoder and two fully connected layers for the decoder on 1% of the VCTK Corpus dataset and test on the rest of the dataset. Alternatively, we use a 1D Convolution over the time axis of the STFT as well as adding a GRU to model time dependencies in the latent space of the VAE. For all VAEs, we use a latent space of size 32 unless otherwise stated. We use the Perceptual Evaluation of Speech Quality (PESQ) score [8] and the Fréchet Audio Distance (FAD) [9] as metrics for comparing different decoders. In our experiments, we considered a few different optimizers to train the VAEs. SGD + M took too long to converge, and Adam-style optimizers often diverged due to parameter instability. We found PSGD + M to be the most stable optimizer which reliably converged to a good solution [15, 16, 17, 18].

Experimental Results Our experiments demonstrate that using ground truth or reconstructed phase via the Griffin-Lim Algorithm with Gamma or Log-Normal von Mises decoders significantly outperform the IS-VAE in terms of PESQ and FAD across various network architectures (see Table II). The Gamma decoder, which does not explicitly encode for phase, achieves the highest PESQ and lowest FAD using an FCN and 1D Convolution architecture respectively. These two models capture the spectrum information at high and low frequencies better than other architectures, including an Itakura-Saito-VAE (see 2). We omit the Complex Gaussian decoder which implicitly models the phase information from Table II as empirically

we saw that $\mu \rightarrow 0$ and it carries little information for speech reconstruction.

The precision of the decoder distribution in a VAE-based speech processing model can be utilized as a measure of the extent to which the magnitude and phase information has been captured by the model. Precision, defined as the inverse of the variance of a distribution, characterizes the tightness with which the data is concentrated around the mean.

If the precision of the decoder distribution is high, this suggests that the decoder is producing outputs that are tightly concentrated around the mean, indicating that the phase information has been effectively captured by the model. Conversely, a low precision indicates that the decoder outputs are spread out, which may indicate that the phase information has not been captured well or has been lost during the encoding-decoding process.

Figure 3 (a) illustrates that the noise power has higher precision due to its stationarity and therefore higher predictability compared to the speech power spectra. Additionally, in (b) it is shown that the phase was learned with higher precision at low frequencies during the speech frames. However, the phase learned via the von Mises distribution is still too noisy for high-quality speech reconstruction. Using ground truth or consistent phase through the Griffin-Lim Algorithm results in improved speech resynthesis compared to using the phase learned by either Type I or Type II decoders.

Overall, the proposed Gamma decoder improves PESQ by 42% (near-perfect PESQ of 4.4) and reduces FAD by 86% compared to IS-VAE.

A. Decoder Compression Capabilities

We evaluate the ability of various decoders to preserve the details of the spectrogram when compressing the data using a latent space of size 16, which results in a compression ratio of 48. As shown in Figure 4, the Gamma decoder produces the spectrogram with the most detailed features. The Log Normal + von Mises and IS decoders capture the general spectral details but are less rich in the region around 2 kHz (red horizontal line). The Gamma + von Mises decoder fails to preserve the speech features. This comparison highlights the fundamental differences in the compression capabilities among decoder distributions.

It is worth mentioning that Log Normal von Mises and Gamma von Mises decoders must encode both magnitude and phase information into the latent space of 16, whereas Gamma and IS decoders only need to encode the magnitude information.

TABLE II: Train VAEs on 1% of the VCTK dataset with STFT 3/4th overlap to train fully connected, convolutional, and recurrent architectures with different decoder distributions. Evaluate PESQ and FAD using ground truth (GT) or Griffin Lim (GL) phase information.

Decoder Loss	**PESQ** (↑ Better)		**FAD** (↓ Better)	
Phase Source	*GT*	*GL*	*GT*	*GL*
Gamma FC	**4.4**	**3.8**	0.6	1.8
Log Normal von Mises FC	3.8	3.6	1.5	2.0
Gamma von Mises FC	3.7	3.5	2.5	3.0
Gamma Conv1d	4.3	3.6	**0.4**	**0.8**
Gamma Conv1d + GRU	4.2	3.5	0.8	1.3
Gamma Conv1d + BiGRU	4.1	3.3	0.8	1.4
Itakura-Saito FC **(baseline)**	3.1	2.8	4.3	5.4

VI. Conclusion

In conclusion, this study has demonstrated that the use of a Gamma decoder, which models the magnitude spectrum, in VAEs significantly improves speech resynthesis when compared to the state-of-the-art IS-VAE. While our research has shown promising results in directly learning phase information within the VAE, further work is necessary to achieve speech quality comparable to using the ground truth phase. Our experiments have also established that VAEs with decoders that explicitly estimate the parameters of a negative log loss function perform better than those using the IS-divergence. Additionally, we have evaluated the performance of both phase-agnostic and phase-aware decoders on various types of VAEs such as fully connected, convolution, and recurrent architectures. Among these, the Gamma decoder, which does not explicitly encode for phase, achieves the highest PESQ and lowest FAD using an FCN and 1D Convolution architecture respectively. Thus, while the use of a Gamma decoder can lead to improved performance in speech resynthesis, the challenge of directly learning phase information within the VAE remains a subject of ongoing research.

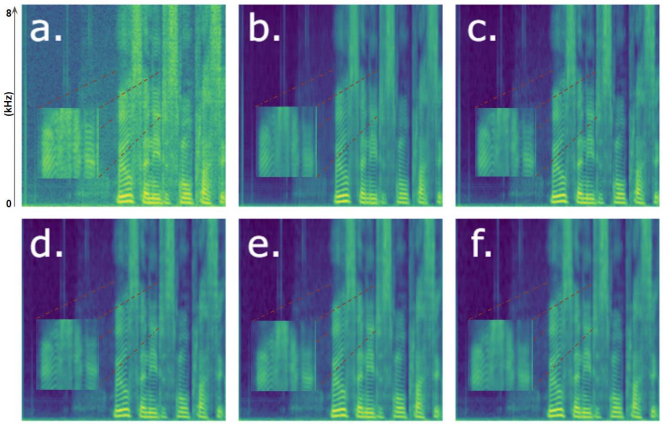
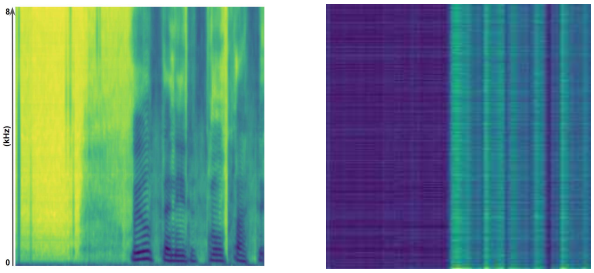


Fig. 2: Effect of architecture on Gamma Decoder vs FCN IS-VAE. a. Ground Truth, b. IS-VAE, c. Gamma FC, d. Gamma Conv1d e. Conv1d GRU, f. Conv1d BiGRU. We see that the Gamma VAE FC and Conv1d capture high-fidelity speech patterns found in the ground truth signal as compared to other architectures and the IS-VAE.



(a) **Power Spectra Precision** (b) **Phase Spectra Precision**

Fig. 3: We observe (a) that the power spectrum of speech is being learned with high precision. The phase spectra in speech frames (b) are also being learned but with insufficient detail.

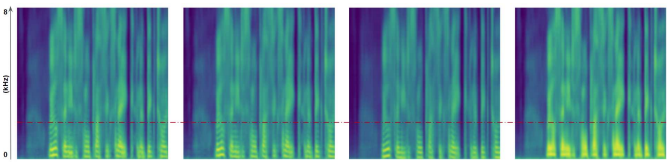


Fig. 4: From left to right we have: Gamma, Log Normal von Mises, Itakura-Saito, and Gamma von Mises decoder. We see that the Gamma distribution can capture the spectrogram at a high compression ratio with more detail. The Log Normal von Mises distribution captures speech features above the red dotted line whereas the IS decoder only captures features under it. With a high compression ratio, the Gamma von Mises fails to capture speech detail.

References

[1] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken elbo." [Online]. Available: https://arxiv.org/abs/1711.00464

[2] 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, 03 2009. [Online]. Available: https://doi.org/10.1162/neco.2008.04-08-771

[3] S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "A recurrent variational autoencoder for speech enhancement," *CoRR*, vol. abs/1910.10942, 2019. [Online]. Available: http://arxiv.org/abs/1910.10942

[4] 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 - 22nd International Conference on Digital Audio Effects*, Birmingham, United Kingdom, Sep. 2019, pp. 1–8. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02349385

[5] X. Bie, L. Girin, S. Leglaive, T. Hueber, and X. Alameda-Pineda, "A benchmark of dynamical variational autoencoders applied to speech spectrogram modeling," *CoRR*, vol. abs/2106.06500, 2021. [Online]. Available: https://arxiv.org/abs/2106.06500

[6] T. Nakashika, "Complex-valued variational autoencoder: A novel deep generative model for direct representation of complex spectra," in *INTERSPEECH*, 2020.

[7] Z. Wang, G. Wichern, and J. L. Roux, "On the compensation between magnitude and phase in speech separation," *CoRR*, vol. abs/2108.05470, 2021. [Online]. Available: https://arxiv.org/abs/2108.05470

[8] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[9] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fr\'echet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[10] J. D. Markel and A. H. Gray, *Vocoders*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1976, pp. 227–262. [Online]. Available: https://doi.org/10.1007/978-3-642-66286-7_10

[11] L. Rabiner, L. Rabiner, and B. Juang, *Fundamentals of Speech Recognition*, ser. Prentice-Hall Signal Processing Series: Advanced monographs. PTR Prentice Hall, 1993. [Online]. Available: https://books.google.com/books?id=XEVqQgAACAAJ

[12] A. Maezawa, K. Itoyama, K. Yoshii, and H. G. Okuno, "Nonparametric bayesian dereverberation of power spectrograms based on infinite-order autoregressive processes," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1918–1930, 2014.

[13] N. Dionelis and M. Brookes, "Modulation-domain kalman filtering for monaural blind speech denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 799–814, 2019.

[14] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP*, 1983.

[15] X. Li, "Online second order methods for non-convex stochastic optimizations," 2018. [Online]. Available: https://arxiv.org/abs/1803.09383

[16] ——, "Preconditioned stochastic gradient descent," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1454–1466, 2018.

[17] ——, "Preconditioner on matrix lie group for sgd," 2018. [Online]. Available: https://arxiv.org/abs/1809.10232

[18] ——, "Black box lie group preconditioners for sgd," *arXiv preprint arXiv:2211.04422*, 2022.