Gamma-VAE-VC: Voice conversion based on VAE assuming gamma distribution for both latent variables and observation

Nanako Imaichi* and Takuya Takahashi * and Toru Nakashika*
* University of Electro-communications, Chofu, Tokyo
E-mail: {n.imaichi, takahashi, nakashika}@uec.ac.jp Tel/Fax: +81 42-443-5602

***Abstract*—In this work, we propose a voice conversion method that incorporates gamma distributions into variational autoencoders (VAEs). In recent years, voice conversion methods based on conventional VAEs, which assumes Gaussian distributions for observed data and latent variables, have achieved promising conversion performance. However, the probability distributions in VAEs should be appropriately chosen based on the characteristics of the observed data rather than being restricted to Gaussian distributions. We therefore propose a voice conversion method based on Gamma-VAE, which adopts the gamma distribution—a probability distribution defined over non-negative values—for VAEs, considering the non-negativity of amplitude spectra. Unlike approaches that seek to achieve state-of-the-art performance, this study focuses on evaluating the fundamental capabilities of VAEs when gamma distributions are used for modeling. Our experimental results demonstrated that the proposed method achieves improved reconstruction accuracy and voice conversion performance compared to conventional Gaussian-based VAEs.**

I. Introduction

Voice conversion (VC) [1] is a technique for modifying speech from a source speaker to match the vocal characteristics of a target speaker while keeping the linguistic content. This technique can be applied to a wide range of tasks, including text-to-speech (TTS) systems [2], speaking aids [3], entertainment scenarios such as singing voice conversion [4], as well as multilingual speech systems [5], accent conversion [6] [7], and emotional expression control [8], making it a valuable tool for enhancing both the accessibility and expressiveness of speech-based technologies.

In the early stages of research on VC, acoustic features extracted from mel-frequency cepstral coefficient (MFCC) are used to achieve voice conversion by exchanging features between the source and target speakers [9]. However, this method assumes a parallel corpus in which the linguistic content is aligned between different speakers. Collecting such data is costly even for one-to-one conversion, and the burden of data collection becomes even greater in practical scenarios such as many-to-one and many-to-many conversion. In recent years, neural network-based conversion frameworks have been extensively proposed, such as restricted Boltzmann machines (RBM) [10] [11] [12], feed-forward deep neural networks [13], and recurrent neural networks (RNN) [14] [15]. Most VC methods, including the above VC methods, require precisely aligned parallel data of the source and target speech. Furthermore, methods using variational autoencoder (VAE) [16] and generative adversarial nets (GAN) [17] have been proposed. However, while GAN-based voice conversion [18] [19] can handle non-parallel, many-to-many scenarios, it is not suited because the quality of converted speech degrades as the number of simultaneously trained speakers increases. On the other hand, VAE-based voice conversion [20] can adapt to many-to-many scenarios.

In recent years, speech waveforms are often generated from mel-spectrogram using methods such as WaveNet[21]and HiFiGAN[22], which are called neural vocoders, and thus non-negative mel-spectrograms are often used as acoustic features. In addition, regarding the modeling of the amplitude spectrum such as mel-spectrogram, Gamma-VAE [23] has been proposed by reconsidering the distributional assumptions in VAE, leveraging the fact that the amplitude spectrum is always non-negative. While conventional VAE assumes a Gaussian distribution for latent variables and observed data, where the domain (of the probability density function) extends from $-\infty$ to $+\infty$, Gamma-VAE instead assumes a gamma distribution, where the domain is restricted to 0 to $+\infty$. Previous study has shown that assuming a gamma distribution without normalizing the amplitude spectra achieves higher reconstruction accuracy compared to the conventional approach, which normalizes the amplitude spectra and assumes a Gaussian distribution. It has also shown that assuming the same probability distribution for latent variables and observed data is more accurate than assuming separate probability distributions for each.

Based on the above research, this paper applies Gamma-VAE to voice conversion. While this work does not aim to establish a state-of-the-art method, its primary objective is to investigate the fundamental performance of VAEs when the output distributions are modeled using gamma distributions. Specifically, the proposed Gamma-VAE assumes gamma distributions for latent vectors, and observed data, with input features represented as mel-spectrograms. This research demonstrates that Gamma-VAE can produce higher-quality converted speech than conventional VAEs under this probabilistic modeling approach.

979-8-3315-7206-8/25/$31.00

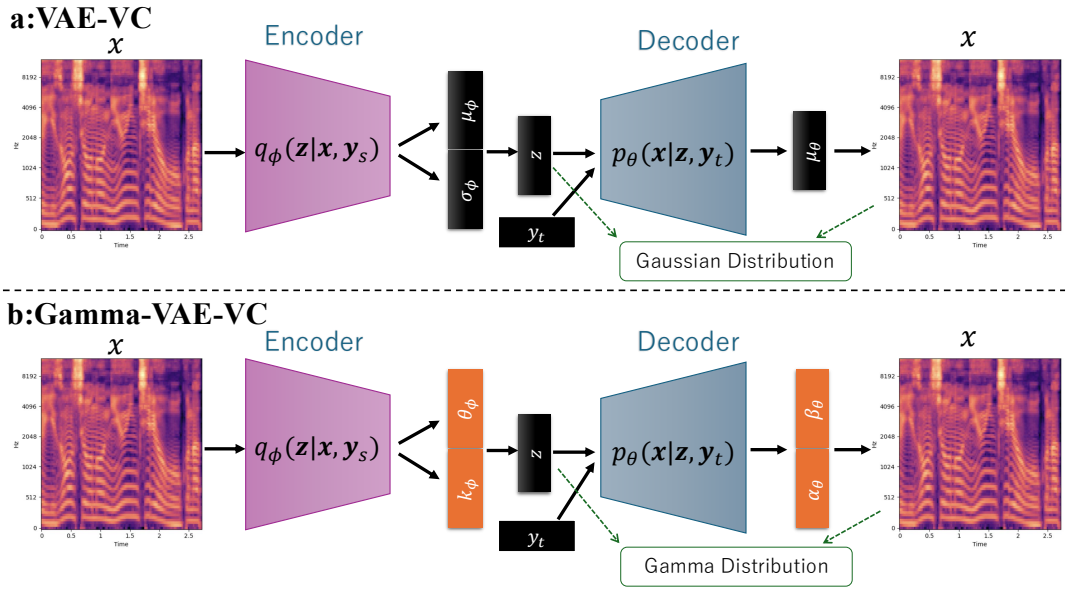

Fig. 1: Architecture of (a)VAE-VC (b)Gamma-VAE-VC

II. Conventional method: VC using Gaussian VAE

VAE-based VC aims to obtain speaker-independent latent variables (i.e., containing only linguistic information) by reconstructing speech from speaker labels and latent variables. [24] As shown in Figure 1, the conventional VAE that assumes a Gaussian distribution is trained to reconstruct the observed data $\boldsymbol{x} \in \mathbb{R}^D$ (D is the dimension of the observation) via the latent variables $\boldsymbol{z} \in \mathbb{R}^Z$ (Z is the dimension of the latent variable) and speaker labels $\boldsymbol{y} \in [0,1]^R$ (R is the number of speakers). The system consists of two components: 1) the encoder with the parameters ϕ that outputs the mean $\boldsymbol{\mu}_\phi \in \mathbb{R}^Z$ and variance $\boldsymbol{\sigma}_\phi \in \mathbb{R}^Z_{>0}$, which are the parameters of the Gaussian distribution assumed for the latent variable, and 2) the decoder with the parameters θ that outputs the mean $\boldsymbol{\mu}_\theta \in \mathbb{R}^D$, which is a parameter of the Gaussian distribution assumed for the observed data. A unit variance is often assumed for the variance of the observation.

A. Train step

VAE simultaneously learns a set of encoder parameters ϕ and a set of decoder parameters θ by maximizing the logarithmic marginal likelihood [25]. However, since the log marginal likelihood cannot generally be computed, we consider maximizing the variational lower bound, which can also be called the evidence lower bound (ELBO) [26], through latent variables. The ELBO can be computed using variational Bayes:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \phi, \theta) = -D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})||p(\boldsymbol{z})] + \mathbb{E}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})]. \tag{1}$$

This independence assumption encourages $\boldsymbol{z}$ to be speaker-independent, i.e., to represent linguistic information. The Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}[\cdot||\cdot]$ [27], the negative distance between the prior and posterior distributions of $\boldsymbol{z}$, takes on non-negative values; hence, maximizing the second term, which is the expected value of the conditional log-likelihood, maximizes the variational lower bound.

The KL divergence in the first term of the right-hand side of Eq. (1), assuming the prior distribution $p(\boldsymbol{z})$ to be the standard normal distribution, can be derived as

$$D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})||p(\boldsymbol{z})] = -\frac{1}{2}\sum_{m=1}^{Z}(\log \sigma_{\phi m}^2 - \mu_{\phi m}^2 - \sigma_{\phi m}^2). \tag{2}$$

The distribution of the second term on the right-hand side of Eq. (1) can be calculated as in (4), assuming a Gaussian distribution with a variance of 1 in each dimension.

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \exp\left(-\sum_{d=1}^{D}\frac{(x_d - \mu_{\theta d}(\boldsymbol{z}, \boldsymbol{y}))^2}{2}\right), \tag{3}$$

$$\mathbb{E}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y})] \propto -\frac{1}{2}\sum_{d=1}^{D}(\mu_{\theta d}(\boldsymbol{z}, \boldsymbol{y}) - x_d)^2. \tag{4}$$

Thus, by combining Eqs. (2) and (4), ELBO is calculated as:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \phi, \theta) = \frac{1}{2}\sum_{m=1}^{Z}(\log \sigma_{\phi m}^2 - \mu_{\phi m}^2 - \sigma_{\phi m}^2) - \frac{1}{2}\sum_{d=1}^{D}(\mu_{\theta d}(\boldsymbol{z}) - x_d)^2, \tag{5}$$

which is the objective function used to optimize the parameters.

B. Conversion step

In the conversion step with trained VAE, a converted speech is generated by decoding the speech from the speaker-

independent latent variables and target speaker labels. Specifically, the encoder $q_\phi(\boldsymbol{z_s}|\boldsymbol{x_s}, \boldsymbol{y_s})$ is first used to extract the speaker-independent latent variable $\boldsymbol{z_s}$ from the source speech $\boldsymbol{x_s}$ and source speaker label $\boldsymbol{y_s}$, where the subscript s denotes the source speaker. Voice conversion is then carried out by inputting the speaker-independent latent variable $\boldsymbol{z_s}$ and the target speaker's label $\boldsymbol{y_t}$ to the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z_s}, \boldsymbol{y_t})$.

III. Proposed method: VC using Gamma-VAE

Conventional VAE assumes a Gaussian distribution for the observed data, making it unsuitable for modeling amplitude spectra that take positive values. On the other hand, using a logarithmic amplitude spectrum as a feature fits a Gaussian distribution as a range of random variables, but the generated data are excessively smoothed. One possible approach is to make the decoder predict the variance of the data distribution to prevent over-smoothing, but the variance parameter is reportedly divergent and difficult to predict [28]. Therefore, we propose Gamma-VAE based on a gamma distribution that directly represents the amplitude spectra taking positive values[23]. In Gamma-VAE, the decoder not only outputs the parameters of the gamma distribution for the observed data, but also assumes gamma distributions for the latent variables inferred by the encoder. This is based on the idea that a variable that follows a gamma distribution and the data that follows a gamma distribution generated from that variable are compatible due to the reproducibility of the gamma distribution.

The architecture of the proposed Gamma-VAE is shown in Figure 1. The model is trained to reconstruct the observed data $\boldsymbol{x} \in \mathbb{R}^D$ (D is the feature dimension) via the latent variable $\boldsymbol{z} \in \mathbb{R}^Z$ (Z is the dimension of the latent variable). The system consists of two components: 1) the encoder with the parameters ϕ that outputs $\boldsymbol{\alpha}_\phi \in \mathbb{R}^Z$ and $\boldsymbol{\beta}_\phi \in \mathbb{R}^Z$, which are the parameters of the gamma distribution assumed for the latent variables, and 2) the decoder with the parameters θ that outputs $\boldsymbol{\alpha}_\theta \in \mathbb{R}^D$ and $\boldsymbol{\beta}_\theta \in \mathbb{R}^D$, which are the parameters of the gamma distribution assumed for the observed data.

A. Train step

As stated in II-A, the parameters of Gamma-VAE are optimized by maximizing ELBO in Eq. (1).

The objective function of the encoder is represented by the KL divergence between the posterior and prior distributions of the latent variable output by the encoder. A broad gamma distribution is adopted for $\boldsymbol{z}$, with $\mathrm{Ga}(\mathbf{1}, \mathbf{1})$ and $\mathrm{Ga}(\mathbf{2}, \mathbf{1})$ being reasonable options due to their simplicity, flexibility, and minimal prior bias, where $\mathrm{Ga}(\alpha, \beta)$ indicate the gamma distribution with a shape parameter α and a rate parameter β. In this work, we define the prior and posterior distributions as

$$p(\boldsymbol{z}) = \mathrm{Ga}(\boldsymbol{z}; \mathbf{1}, \mathbf{1}), \tag{6}$$

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Ga}(\boldsymbol{z}; \boldsymbol{\alpha}_\phi(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{\beta}_\phi(\boldsymbol{x}, \boldsymbol{y})). \tag{7}$$

With Eqs. (6) and (7), we can derive the KL divergence, which is the encoder-related term of the objective in Eq. (1), as

$$D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})||p(\boldsymbol{z})] = \sum_{m=1}^{Z} -\frac{\beta_{\phi m} - 1}{\beta_{\phi m}}\alpha_{\phi m} + \log\frac{\beta_{\phi m}}{\Gamma(\alpha_{\phi m})} + (\alpha_{\phi m} - 1)\psi(\alpha_{\phi m}), \tag{8}$$

where $\alpha_{\phi m} \in \boldsymbol{\alpha}_\phi$ and $\beta_{\phi m} \in \boldsymbol{\beta}_\phi$, and $\psi(\cdot)$ indicate a digamma function.

Assuming independent gamma distributions for all dimensions, the objective function of the decoder takes the logarithm of the probability density function of the gamma distribution when the output of the decoder is $\boldsymbol{\alpha}_\theta \in \mathbb{R}^D$ and $\boldsymbol{\beta}_\theta \in \mathbb{R}^D$. Thus, we define the conditional likelihood given latent variables $\boldsymbol{z}$ as

$$p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{y}) = \mathrm{Ga}(\boldsymbol{x}; \boldsymbol{\alpha}_\theta(\boldsymbol{z}, \boldsymbol{y}), \boldsymbol{\beta}_\theta(\boldsymbol{z}, \boldsymbol{y})) \tag{9}$$

$$= \prod_{d=1}^{D} \frac{\beta_{\theta d}^{\alpha_{\theta d}}}{\Gamma(\alpha_{\theta d})} x_d^{\alpha_{\theta d}-1} e^{-\beta_{\theta d} x_d}, \tag{10}$$

$$\log p_\theta(\boldsymbol{x}) = \sum_{d=1}^{D} -\beta_{\theta d} x_d - \log\Gamma(\alpha_{\theta d}) + \alpha_{\theta d}\log\beta_{\theta d} + (\alpha_{\theta d} - 1)\log x_d, \tag{11}$$

where $\alpha_{\theta d} \in \boldsymbol{\alpha}_\theta$ and $\beta_{\theta d} \in \boldsymbol{\beta}_\theta$, and $\Gamma(\cdot)$ denote the gamma function, respectively.

Based on the above, the combination of Eqs. (8) and (11) is the objective function:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \varphi, \theta) = \sum_{d=1}^{D} -\beta_{\theta d} x_d - \log\Gamma(\alpha_{\theta d}) + \alpha_{\theta d}\log\beta_{\theta d} + (\alpha_{\theta d} - 1)\log x_d + \sum_{m=1}^{Z} \frac{\beta_{\phi m} - 1}{\beta_{\phi m}}\alpha_{\phi m} - \log\frac{\beta_{\phi m}}{\Gamma(\alpha_{\phi m})} - (\alpha_{\phi m} - 1)\psi(\alpha_{\phi m}), \tag{12}$$

which the parameters are optimized to maximize.

B. Conversion step

In the conversion step using trained Gamma-VAE, voice conversion is carried out by decoding the speech from speaker-independent latent variables and target speaker labels. First, the encoder $q_\phi(\boldsymbol{z_s}|\boldsymbol{x_s}, \boldsymbol{y_s})$ outputs the parameter $\boldsymbol{\alpha}_\phi$ of the gamma distribution corresponding to the latent variable $\boldsymbol{z}$ from the source speech $\boldsymbol{x_s}$ and source speaker label $\boldsymbol{y_s}$. $\boldsymbol{\beta}_\phi$ are estimated and their expected value $\boldsymbol{z_s} = \boldsymbol{\alpha}_\phi/\boldsymbol{\beta}_\phi$ is obtained, where the fraction bar indicates element-wise division. Next, by inputting the obtained expected value of the latent variable $\boldsymbol{z_s}$ and the target speaker label $\boldsymbol{y_t}$ to the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z_s}, \boldsymbol{y_t})$, the parameters of the output gamma distribution $\boldsymbol{\alpha}_\phi$ and $\boldsymbol{\beta}_\phi$ are obtained. θ and $\boldsymbol{\beta}_\theta$ to generate the expected value of the speech features $\boldsymbol{\mu} = \boldsymbol{\alpha}_\theta/\boldsymbol{\beta}_\theta$, which is used as the final acoustic features after transformation.

IV. Experiments

A. Set-up

To evaluate the effectiveness of the proposed Gamma-VAE in voice conversion tasks, we conducted a series of comparative experiments with a conventional VAE-based baseline. Since naturalness has already been addressed in previous studies [23], this work focuses solely on evaluating speaker similarity.

1) Configuration: First, the mel-spectrogram or logarithmic mel-spectrogram were obtained from the speech by short-time Fourier transform, and the spectra reconstructed by each method were restored to the speech signals. As stated in I, mel-spectrograms are well-suited for waveform reconstruction using neural vocoders. Therefore, we used HiFi-GAN for recovering the phase. Spectrograms were extracted with a fast Fourier transform (FFT) size of 2,048 and a hop size of 128. The audio was resampled to 24 kHz. We set the learning rate to 0.0001 and the number of epochs to 6,000. The evaluation is carried out on the VCTK corpus [29]. The data of each speaker is then partitioned into training and test sets with a 20:1 raito.

2) Evaluation methods: We conducted subjective evaluations to assess the quality of speaker similarity. In particular, we employed the similarity mean opinion score (SMOS), a widely used subjective metric for evaluating the similarity between speakers. In the SMOS test, listeners were presented with pairs of utterances: one converted utterance and one target speaker utterance, both containing the same linguistic content. listeners evaluated each sample in a test case in accordance with the criteria: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, and 5 = Excellent. S

3) Network Architecture: By setting $Z = 320$, we use a CNN-based architecture for both the encoder and decoder. The encoder consists of five 1D convolutional layers with increasing channel sizes: $80 \rightarrow 160 \rightarrow 240 \rightarrow 320 \rightarrow 640$. Batch normalization and ReLU are applied after each layer except the last, which is split into two 320-dimensional tensors representing the gamma distribution parameters $\boldsymbol{\alpha}_\phi$ and $\boldsymbol{\beta}_\phi$, both passed through softplus to ensure positivity.

The decoder also has five convolutional blocks, each with a convolution layer, conditional batch normalization (modulated by speaker identity), and ReLU. Upsampling layers are used to increase temporal resolution. The final layer outputs shape and scale parameters $\boldsymbol{\alpha}_\theta$ and $\boldsymbol{\beta}_\theta$ of the output gamma distribution, also passed through softplus.

Note that while more complex components (e.g., residual connections, LSTM, deeper structures) could enhance capacity, we adopt this simple CNN-based design to focus on differences between the distributions modeled by the encoder and decoder.

4) Introducing a trade-off weight of the objective function: Training the model to minimize the KL divergence (Eq. (2) and Eq. (8)) implies that each dimension of the latent variable $\boldsymbol{z}$ approaches the standard exponential distribution and their respective scales become aligned and uncorrelated. However, as in the case of non-negative matrix factorization (NMF) [30], correlated underlying latent variables are sometimes desirable under non-negative constraints. Therefore, as in β-VAE [31], the degree of correlation of the underlying latent variables can be adjusted by assigning a weight $\beta > 0$ to each cost function in the variational lower bound, as

$$\mathcal{L}(\boldsymbol{x}; \phi, \theta) = (1 - \beta) \cdot \mathbb{E}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta \cdot D_{\mathrm{KL}}[q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})]. \tag{13}$$

In this case, we conducted preliminary experiment that VAE and Gamma-VAE was trained with various values of $\beta \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$ to determin β. In preliminary experiment, we use mel-cepstral distortion (MCD) to measure spectral differences and voice similarity objectively. MCD is calculated as mean squared error of mel-cepstral coefficients between converted and target speech, and lower values indicate better conversion quality and higher similarity. The results are shown in Figure 2. Although $\beta = 0.75$ achieved the lowest MCD in Gamma-VAE, we selected $\beta = 0.25$ as it provides a better trade-off between distortion and stability across different speakers. In our preliminary experiments, models trained with $\beta = 0.75$ occasionally suffered from overfitting or poor generalization. It is also worth noting that the performance scores for $\beta = 0.25$, 0.50, and 0.75 were generally comparable, with no significant differences observed. Thus, $\beta = 0.25$ was chosen as a balanced and robust setting in both VAE and Gamma-VAE.

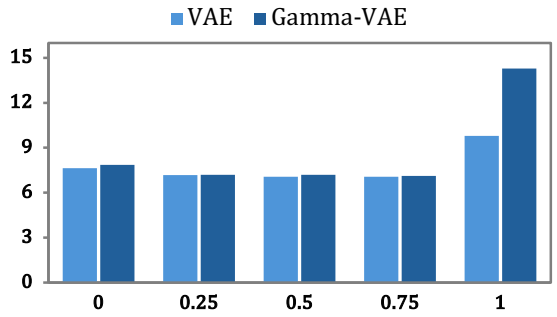

Fig. 2: Hyperparameter study on β using MCD.

TABLE I: Experimental conditions and results of subjective evaluation

Methods		M → M	F → F	M → F	F → M	Avg.
SMOS	VAE	2.35	3.14	2.38	1.48	2.26
	Gamma-VAE	2.89	2.76	2.04	1.79	2.37

B. Results and discussion

Table I presents the results of the SMOS-based subjective evaluation. On average, Gamma-VAE achieved a higher SMOS score (2.37) compared to the baseline VAE (2.26), indicating improved speaker similarity. While VAE showed relatively high scores in same-gender conversions (e.g., F→F: 3.14), its performance significantly degraded in cross-gender conversions, especially F→M (1.48). In contrast, Gamma-VAE

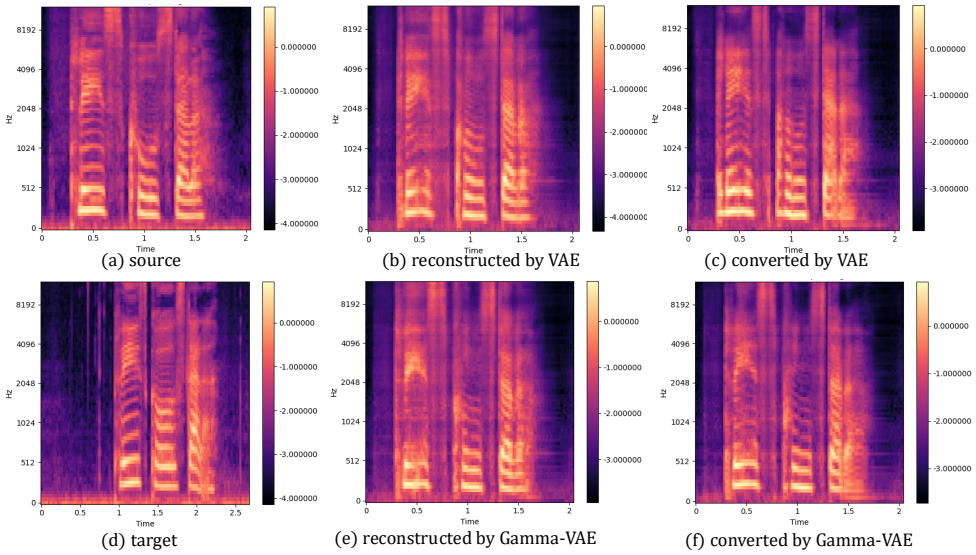

Fig. 3: Logarithm spectrograms of (a) original speech, (b) reconstructed by VAE, (c) converted by VAE, (d) target speech, (e) reconstructed by Gamma-VAE, and (f) converted by Gamma-VAE.

demonstrated more consistent performance across all conversion pairs, showing smaller variations between same-gender and cross-gender scenarios. Although the improvements in absolute values may seem modest, these results suggest that Gamma-VAE provides more stable and reliable conversion quality across diverse speaker pairs. Therefore, Gamma-VAE can be considered a more robust framework for speaker conversion tasks in terms of preserving speaker similarity. Figure 3 demonstrates that Gamma-VAE more effectively preserves spectral structures compared to the baseline VAE. The reconstructed spectrogram from Gamma-VAE (Fig. 3(e)) more closely resembles the target (Fig. 3(d)) than that of VAE (Fig. 3(b)), particularly in terms of harmonic structure and spectral envelope. In the conversion scenario, Gamma-VAE outputs exhibit clearer formant regions and more stable frequency patterns, suggesting better modeling of speaker characteristics. These visual results are consistent with the SMOS scores in Table I, reinforcing the effectiveness of Gamma-VAE in both reconstruction and speaker conversion tasks.

V. Conclusion

In this study, we proposed Gamma-VAE-VC, a voice conversion framework that assumes gamma distributions for both latent variables and observed features. Unlike conventional VAE-based methods, which rely on Gaussian assumptions, Gamma-VAE is better suited to model non-negative data such as mel-spectrograms. Through both objective (MCD) and subjective (SMOS) evaluations, the proposed method demonstrated improved stability and speaker similarity across different gender pairs. The gamma distribution's ability to represent positive-valued features resulted in better spectral reconstruction and more consistent performance, particularly in challenging cross-gender conversions. While not focused on achieving state-of-the-art results, this study validates the fundamental advantages of adopting gamma distributions in VAE-based voice conversion. Gamma-VAE thus provides a robust alternative for modeling non-negative acoustic features and highlights the importance of distributional assumptions in generative speech models.

VI. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 24H00715.

References

[1] Y. Stylianou, O. Cappe, and E. Moulines., "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, IEEE, vol. 1, 1998, pp. 285–288.

[3] Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely., "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, 2007.

[4] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, IEEE, 2012, pp. 1–6.

[5] J. Yang and L. He, "Cross-lingual text-to-speech using multi-task learning and speaker classifier joint training," *arXiv preprint arXiv:2201.08124*, 2022.

[6] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 5274–5278.

[7] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2016, pp. 1–6.

[8] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, pp. 1–13, 2017.

[9] H. Terasawa, M. Slaney, and J. Berger, "A statistical model of timbre perception.," *SAPA@ INTERSPEECH*, vol. 5, 2006.

[10] P. Smolensky *et al.*, "Information processing in dynamical systems: Foundations of harmony theory," 1986.

[11] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[12] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1403–1410, 2014.

[13] S. Desai, A. W. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 954–964, 2010.

[14] T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion.," in *Interspeech*, 2014, pp. 2278–2282.

[15] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2015, pp. 4869–4873.

[16] D. P. Kingma, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[18] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks.," in *Interspeech*, 2017, pp. 1283–1287.

[19] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," *arXiv preprint arXiv:1907.12279*, 2019.

[20] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Acvae-vc: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 9, pp. 1432–1443, 2019.

[21] A. Van Den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.

[22] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[23] N. Imaichi and T. Nakashika, "Gamma-vae: Speech representation based on vae assuming gamma distribution for both latent variables and observation," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2024, pp. 1–6.

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

[25] R. Linske, "An application of the principle of maximum information preservation to linear systems.," *Advances in Neural Information Processing Systems*, vol. 1, pp. 186–194, 1989.

[26] M.D.Hoffman and M.J.Johnson, "Elbo surgery: Yet another way to carve up the variational evidence lower bound," *in Workshop in Advances in Approximate Bayesian Inferenc*, vol. 1, p. 2, 2016.

[27] I. Csiszár, "On information-type measure of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[28] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, S. Yagi, and H. Kashima, "Student-t variational autoencoder for robust multivariate density estimation," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 36, no. 3, A–KA4, 2021.

[29] J. Yamagishi, C. Veaux, K. MacDonald, *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[30] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2000.

[31] I. Higgins, L. Matthey, A. Pal, *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.